# Exploring word class n-grams to measure language development in children

**Gabriela Ramírez de la Rosa** and **Thamar Solorio**
University of Alabama at Birmingham
Birmingham, AL 35294, USA
gabyrr,solorio@cis.uab.edu

**Manuel Montes-y-Gómez**
INAOE
Sta. Maria Tonantzintla, Puebla, Mexico
mmontesg@ccc.inaoep.mx

**Yang Liu**
The University of Texas at Dallas
Richardson, TX 75080, USA
yangl@hlt.utdallas.edu

**Aquiles Iglesias**
Temple University
Philadelphia, PA 19140, USA
iglesias@temple.edu

**Lisa Bedore** and **Elizabeth Peña**
The University of Texas at Austin
Austin, TX 78712, USA
lbedore,lizp@mail.utexas.edu

## Abstract

We present a set of new measures designed to reveal latent information of language use in children at the lexico-syntactic level. We used these metrics to analyze linguistic patterns in spontaneous narratives from children developing typically and children identified as having a language impairment. We observed significant differences in the z-scores of both populations for most of the metrics. These findings suggest we can use these metrics to aid in the task of language assessment in children.

## 1 Introduction

The analysis of spontaneous language samples is an important task across a variety of fields. For instance, in language assessment this task can help to extract information regarding language proficiency (e.g. is the child typically developing or language impaired). In second language acquisition, language samples can help determine if a child's proficiency is similar to that of native speakers.

In recent years, we have started seeing a growing interest in the exploration of NLP techniques for the analysis of language samples in the clinical setting. For example, Sahakian and Snyder (2012) propose a set of linguistic measures for age prediction in children that combines three traditional measures from language assessment with a set of five data-driven measures from language samples of 7 children. A common theme in this emerging line of research is the study of the syntax in those language samples. For instance, to annotate data to be used in the study of language development (Sagae et al., 2005), or to build models to map utterances to their meaning, similar to what children do during the language acquisition stage (Kwiatkowski et al., 2012). In addition, language samples are also used for neurological assessment, as for example in (Roark et al., 2007; Roark et al., 2011) where they explored features such as Yngve and Frazier scores, together with features derived from automated parse trees to model syntactic complexity and surprisal. Similar features are used in the classification of language samples to discriminate between children developing typically and children suffering from autism or language impairment (Prud'hommeaux et al., 2011). In a similar line of research, machine learning and features inspired by NLP have been explored for the prediction of language status in bilingual children (Gabani et al., 2009; Solorio et al., 2011). More recent work has looked at the feasibility of scoring coherence in story narratives (Hassanali et al., 2012a) and also on the inclusion of coherence



*Proceedings of the 2013 Workshop on Biomedical Natural Language Processing (BioNLP 2013)*, pages 89–97,
Sofia, Bulgaria, August 4-9 2013. 

as an additional feature to boost prediction accuracy of language status (Hassanali et al., 2012b).

The contribution of our work consists on new metrics based on n-grams of Part of Speech (POS) tags for assessing language development in children that combine information at the lexical and syntactic levels. These metrics are designed to capture the lexical variability of specific syntactic constructions and thus could help to describe the level of language maturity in children. For instance, given two lists of examples of the use of determiner + noun: ⟨the dog, the frog, the tree⟩ and ⟨this dog, a frog, these trees⟩ we want to be able to say that the second one has more lexical variability than the first one for that grammatical pattern.

Our approach to compute these new metrics does not require any special treatment on the transcripts or special purpose parsers beyond a POS tagger. On the contrary, we provide a set of measures that in addition to being easy to interpret by practitioners, are also easy to compute.

## 2 Background and Motivation

To establish language proficiency, clinical researchers and practitioners rely on a variety of measures, such as number of different words, type-token ratio, distribution of part-of-speech tags, and mean length of sentences and words per minute (Lu, 2012; Yoon and Bhat, 2012; Chen and Zechner, 2011; Yang, 2011; Miller et al., 2006), to name a few. Most of these metrics can be categorized as low-level metrics since they only consider rates of different characteristics at the lexical level. These measures are helpful in the solution of several problems, for example, building automatic scoring models to evaluate non-native speech (Chen and Zechner, 2011). They can also be used as predictors of the rate of growth of English acquisition in specific populations, for instance, in typically developing (TD) and language impaired (LI) bilingual children (Rojas and Iglesias, 2012; Gutiérrez-Clellen et al., 2012). Among the most widely used metrics are mean length of utterance (MLU), a measure of syntactic complexity (Bedore et al., 2010), and measures of lexical productivity, such as the number of different words (NDW) and the child's ratio of functional words to content words (F/C) (Sahakian and Snyder, 2012).

MLU, NDW, F/C and some other low-level measures have demonstrated to be valuable in the assessment of language ability considering that practitioners often only need to focus on productivity, diversity of vocabulary, and sentence organization. Although useful, these metrics only provide superficial measures of the children's language skills that fail to capture detailed lexico-syntactic information. For example, in addition to knowing that a child is able to use specific verb forms in the right context, such as, third person singular present tense or regular past tense, knowledge about what are the most common patterns used by a child, or how many different lexical forms for *noun + verb* are present in the child's speech is needed because answering these questions provides more detailed information about the status of grammatical development. To fill in this need, we propose a set of measures that aim to capture language proficiency as a function of lexical variability in syntactic patterns. We analyze the information provided by our proposed metrics on a set of spontaneous story retells and evaluate empirically their potential use in language status prediction.

## 3 Proposed measures

To present the different metrics we propose in this study we begin with the definition of the following concepts:

A *syntactic pattern* $p$ is an $n$-gram of part-of-speech tags denoted as $p = \langle t_1\ t_2\ ...\ t_n \rangle$, where $t_i$ indicates the part-of-speech tag corresponding to the word at position $i$. For simplicity we use $t_i^p$ to indicate the tag at position $i$ from pattern $p$. Two examples of syntactic patterns of length two are 'DT NN' and 'DT JJ' [1].

A *lexical form* $f$ is an $n$-gram of words. It is defined as $f = \langle w_1\ w_2\ ...\ w_n \rangle$, where $w_i$ is the word at position $i$. Similarly to the previous definition, we use $w_i^f$ to indicate the word at position $i$ in a lexical form $f$.

A lexical form $f$ corresponds to a syntactic pattern $p$ if and only if $|f|$ is equal to $|p|$ and $\forall_k tag(w_k^f) = t_k^p$, where $tag()$ is a function that returns the part-of-speech of its argument. The set of lexical forms in a given transcript corresponding to a syntactic pattern $p$ is denoted by $LF^p$. Two examples of lexical forms from the syntactic pattern 'DT NN' are 'the cat' and 'the frog'.

[1] We use the Penn Treebank POS tagset

| | |
|---|---|
| **DT** | the (62), a (17), all (8), no(2), that (1) |
| **NN** | frog (16), boy(7), dog (6), boat (4), name (3), place (2), house (2), water (2), rabbit (2), noise (2), stick (1), tree (1), bye(1), floor (1), um (1), baby (1), forest (1), room (1), foot (1), rock (1), squirrel (1), back (1), rabb (1), card (1), one (1), present (1), dress (1), box (1), family (1) |
| **VBD** | saw (7), dropped (4), said (4), started (4), looked (3), kicked (3), called (3), found (2), took (2), got (2), jumped (2), heard (2), thought (1), turned (1), fell (1), waked (1), stood (1), wa (1), touched (1), told (1), scared (1), tur (1), haded (1), opened (1), shh (1) |
| **DT NN** | the frog (3), the dog (2), the place (2), the water (2), the boat (2), a noise (2), the forest (1), the rock (1), a tree (1), a present (1), a um (1), the card (1), the box (1), the rabb (1), the floor (1), the back (1), no one (1) |
| **DT VBD** | all started (2), all heard (1) |

Table 1: Example of 5 syntactic patterns with their lists of lexical forms and the number of repetitions of each of them. This information corresponds to an excerpt of an example transcript. DT is the part-of-speech tag for determiner, NN for noun, and VBD for verb in past tense.

The *bag-of-words* associated to a syntactic pattern $p$ is denoted as $W^p$. This set is composed of all the words from the lexical forms that correspond to the syntactic pattern $p$. It is formally defined as follows: $W^p = \{w|w \in f, f \in LF^p\}$. For example, the bag-of-words of the syntactic pattern 'DT NN' with lexical forms 'the cat' and 'the frog' is $\{the, cat, frog\}$.

Table 1 shows five syntactic patterns of a transcript's fragment. For each syntactic pattern in the transcript we show the list of its lexical forms and their frequency. We will use this example in the description of the measures in the following subsections.

### 3.1 Number of different lexical forms (NDLF)

Analogous to the number of different words (NDW), where words in the transcript are considered atomic units, we propose a metric where the atomic units are lexical forms. Then, we measure the number of different lexical forms used for each syntactic pattern in the transcript. Formally, given a syntactic pattern $p$ and its set of lexical forms $LF^p$, the *number of different lexical forms* is computed as follows:

$$NDLF(p) = |LF^p| \quad (1)$$

This measure gives information about the number of different ways the child can combine words in order to construct a fragment of a speech that corresponds to a specific grammatical pattern. Research in language assessment has shown that when children are in the early acquisition stages of certain grammatical constructions they will use the patterns as "fixed expressions". As children master these constructions they are able to use these grammatical devices in different contexts, but also with different surface forms. Thereby, we could use this measure to discriminate the syntactic patterns the child has better command of from those that might still be problematic and used infrequently or with a limited combination of surface forms. For example, from the information on Table 1 we see that $NDLF(DT\ NN) = 17$, and $NDLF(DT\ VBD) = 2$. This seems to indicate that the child has a better command of the grammatical construction *determiner + noun* (DT NN) and can thus produce more different lexical forms of this pattern than *determiner + verb* (DT + VBD). But also, we may use this measure to identify rare patterns, that are unlikely to be found in a typically developing population.

### 3.2 Lexical forms distribution (LFdist)

Following the idea of lexical forms as atomic units, *NDLF* allows to know the different lexical forms present in the transcripts. But we do not know the distribution of use of each lexical form for a specific syntactic pattern. In other words, NDLF tells us the different surface forms observed for each syntactic pattern, but it does not measure the frequency of use of each of these lexical forms, nor whether each of these forms are used at similar rates. We propose to use LFdist to provide information about the distribution of use for $LF^p$, the set of lexical forms observed for the syntactic pattern $p$. We believe that uniform distributions can be indicative of syntactic structures that the child has mastered, while uneven distributions can reveal structures that the child has only memorized (i.e. the child uses a fixed and small set of lexical forms). To measure this distribution we use the entropy of each syntactic pattern. In particular, given a syntactic pattern $p$ and its set of lexical forms $LF^p$, the *lexical form distribution* is computed as follows:

$$LFdist(p) = -\sum_{f_i \in LF^p} prob(f_i) \log prob(f_i) \quad (2)$$

where

$$prob(f_i) = \frac{count(f_i)}{\sum_{f_k \in LF^p} count(f_k)} \quad (3)$$

and $count()$ is a function that returns the frequency of its argument. Larger values of *LFdist* indicate a greater difficulty in the prediction of the lexical form that is being used under a specific grammatical pattern. For instance, in the example of Table 1, $LFdist(DT\ VBD) = 0.91$ and $LFdist(DT\ NN) = 3.97$. This indicates that the distribution in the use of lexical forms for *determiner + noun* is more uniform than the use of lexical forms for *determiner + verb*, which implies that for *determiner + verb* there are some lexical forms that are more frequently used than others[2]. Syntactic patterns with small values of *LFdist* could flag grammatical constructions the child does not feel comfortable manipulating and thus might still be in the acquisition stage of language learning.

### 3.3 Lexical variation (LEX)

Until now we are considering lexical forms as atomic units. This could lead to overestimating the real lexical richness in the sample, in particular for syntactic patterns of length greater than 1. To illustrate this consider the syntactic pattern $p = \langle DT\ NN \rangle$ and suppose we have the following set of lexical forms for $p = \{$'the frog', 'a frog', 'a dog', 'the dog'$\}$. The value for $NDLF(p) = 4$. But how many of these eight words are in fact different? That is the type of distinction we want to make with the next proposed measure: LEX, that is also an adaptation of type-token ratio (Lu, 2012) used in the area of communication disorders but computed over each grammatical pattern. For this example, we want to be able to find that the lexical variation of $\langle DT\ NN \rangle$ is 0.5 (because there are only four different words out of eight). Formally, given a syntactic pattern $p$, its set of lexical forms $LF^p$, and the bag-of-words $W^p$, the *lexical variation* is defined as shown in Equation 4.

[2]We recognize that this is an oversimplification of the entropy measure since the number of outcomes will most likely be different for each syntactic pattern.

$$LEX(p) = \frac{|W^p|}{|LF^p| * n} \quad (4)$$

Note that $|LF^p| = NDLF(p)$, and $n$ is the length of the syntactic pattern $p$. In Table 1 the lexical variation of the pattern '*determiner + noun*' (DT+NN) is equal to 0.58 ($\frac{20}{17*2}$), and for *determiner + verb* (DT+VBD) is equal to 0.75 ($\frac{3}{2*2}$). That means 58% of total words used under the pattern 'DT+NN' are different, in comparison with the 75% for 'DT+VBD'. In general, the closer the value of *LEX* is to 1, there is less overlap between the words in the lexical forms for that pattern. Our hypothesis behind this measure is that for the same syntactic pattern TD children may have less overlap of words than children with LI, e.g. less overlap indicates the use of a more diverse set of words.

### 3.4 Lexical use of syntactic knowledge (LexSyn)

With LEX we hope to accomplish the characterization of lexical richness of syntactic patterns assuming that each part-of-speech has a similar number of possible lexical forms. We assume as well that less overlap in the words used for the same grammatical pattern represents a more developed language than that with more overlap. However the definition of LEX overlooks a well known fact about language: different word classes have a different range of possibilities as their lexical forms. Consider open class items, such as nouns and verbs, where the lexicon is large and keeps growing. In contrast, closed class items, such as prepositions and determiners are fixed and have a very small number of lexical forms. Therefore it seems unfair to assign equal weight to the overlap of words for these different classes. To account for this phenomenon, we propose a new measure that includes the information about the syntactic knowledge that the child shows for each part of speech. That is, we weigh the level of overlap for specific grammatical constructions according to the lexicon for the specific word classes involved. Since we limit our analysis to the language sample at hand, we define the ceiling of the lexical richness of a specific word class to be the total number of different surface forms found in the transcript. In particular, given a syntactic pattern $p = \langle t_1\ t_2\ ...\ t_n \rangle$, with its set of lexical forms $LF^p$, the lexical use of syntactic knowledge is defined as:

$$\textit{LexSyn}(p) = \frac{1}{n}\sum_{i=1}^{n}\frac{|w_i^f | f \in LF^p|}{\textit{NDLF}(t_i^p)} \qquad (5)$$

where the numerator is the size of the set of words in the $i$-th position in all the lexical forms. Note that this measure does not make sense for syntactic patterns of length $< 2$. Instead, syntactic patterns of length 1 were used to identify the syntactic knowledge of the child by using the NDLF of each POS in $p$. In the example of Table 1, $\textit{LexSyn}(\textit{DT NN}) = 0.59$. This value corresponds to the sum of the number of different determiners used in position 1 for $LF^p$ divided by the total number of different determiners that this child produced in the sample (for this case, the number of determiners that this child produced is given by *NDLF*(*DT*), that is 5), plus the number of different nouns used under this syntactic pattern over the total number of nouns produced by the child (*NDLF*(*NN*)=29). The complete calculation of $LexSyn(DT\ NN) = \frac{1}{2} * (\frac{3}{5} + \frac{17}{29}) = 0.59$. This contrasts with the value of *LexSyn* for the pattern '*determiner + verb*', $LexSyn(\textit{DT VBD}) = \frac{1}{2} * (\frac{1}{5} + \frac{2}{25}) = 0.14$ that seems to indicate that the child has more experience combining determiners and nouns than determiners and verbs. Perhaps this child has had limited exposure to other patterns combining determiner and verb, or this pattern is at a less mature stage in the linguistic repertoire of the child.

Children with LI tend to exhibit a less developed command of syntax than their TD cohorts. Syntactic patterns with large values of LexSyn show a high versatility in the use of those syntactic patterns. However, since the syntactic reference is taken from the same child, this versatility is relative only to what is observed in that single transcript. For instance, suppose that the total number of different determiners observed in the child's transcript is 1. Then any time the child uses that determiner in a syntactic pattern, the knowledge of this class, according to our metric, will be 100%, which is correct, but this might not be enough to determine if the syntactic knowledge of the child for this grammatical class corresponds to age expectations for a typically developing child. In order to improve the measurement of the lexical use of syntactic knowledge we propose the measure **LexSynEx**, that instead of using the information of the same child to define the coverage of use for a specific word class, it uses the information observed for a held out set of transcripts from TD children. This variation allows the option of moving the point of reference to a specific cohort, according to what is needed.

## 4 Data set

The data used in this research is part of an ongoing study of language impairment in Spanish-English speaking children (Peña et al., 2003). From this study we used a set of 175 children with a mean age of about 70 months. Language status of these children was determined via expert judgment by three bilingual certified speech-language pathologists. At the end of the data collection period, the experts reviewed child records in both languages including language samples, tests protocols, and parent and teacher questionnaire data. They made independent judgments about children's lexical, morphosyntactic, and narrative performance in each language. Finally, they made an overall judgment about children's language ability using a 6 point scale (severely language impaired to above normal impairment). If at least two examiners rated children's language ability with mild, moderate or severe impairment they were assigned to the LI group. Percent agreement among the three examiners was 90%. As a result of this process, 20 children were identified by the clinical researchers as having LI, while the remaining 155 were identified as typically developing (TD).

The transcripts were gathered following standard procedures for collection of spontaneous language samples in the field of communication disorders. Using a wordless picture book, the children were asked to narrate the story. The two books used were 'A boy, a dog, and a frog' (Mayer, 1967) and 'Frog, where are you?' (Mayer, 1969). For each child in the sample, 4 transcripts of story narratives were collected, 2 in each language. In this study we use only the transcripts where English was the target language.

## 5 Procedure

The purpose of the following analysis is to investigate the different aspects in the child's language that can be revealed by the proposed metrics. All our measures are based on POS tags. We used the Charniak parser (Charniak, 2000) to generate the POS tags of the transcripts. For all the results reported here we removed the utterances from the interrogators and use all utterances by the chil-

dren. From the 155 TD instances, we randomly selected 20, that together with the 20 instances with LI form the test set. The remaining 135 TD instances were used as the normative population, our training set.

After the POS tagging process, we extracted the set of syntactic patterns with length equal to 1, 2, 3 and 4 that appear in at least 80% of the transcripts in the training set. The 80% threshold was chosen with the goal of preserving the content that is most likely to represent the TD population.

## 6 Analysis of the proposed measures and implications

Figure 1 shows 5 plots corresponding to each of our proposed measures. Each graph shows a comparison between the average values of the TD and the LI populations. The x-axis in the graphs represents all the syntactic patterns gathered from the training set that appeared on the test data, and the y-axis represents the difference in the z-score values of each measure from the test set. The x-axis is sorted in descending order according to the z-score differences between values of TD and LI.

The most relevant discovery is that *NDFL*, *LFdist*, *LexSyn* and *LexSynEx* show a wider gap in the z-scores between the TD and LI populations for most of the syntactic patterns analyzed. This difference is easy to note visually as most of the TD patterns tend to have larger values, while the ones for children with LI have lower scores. Therefore, it seems our measures are indeed capturing relevant information that characterizes the language of the TD population.

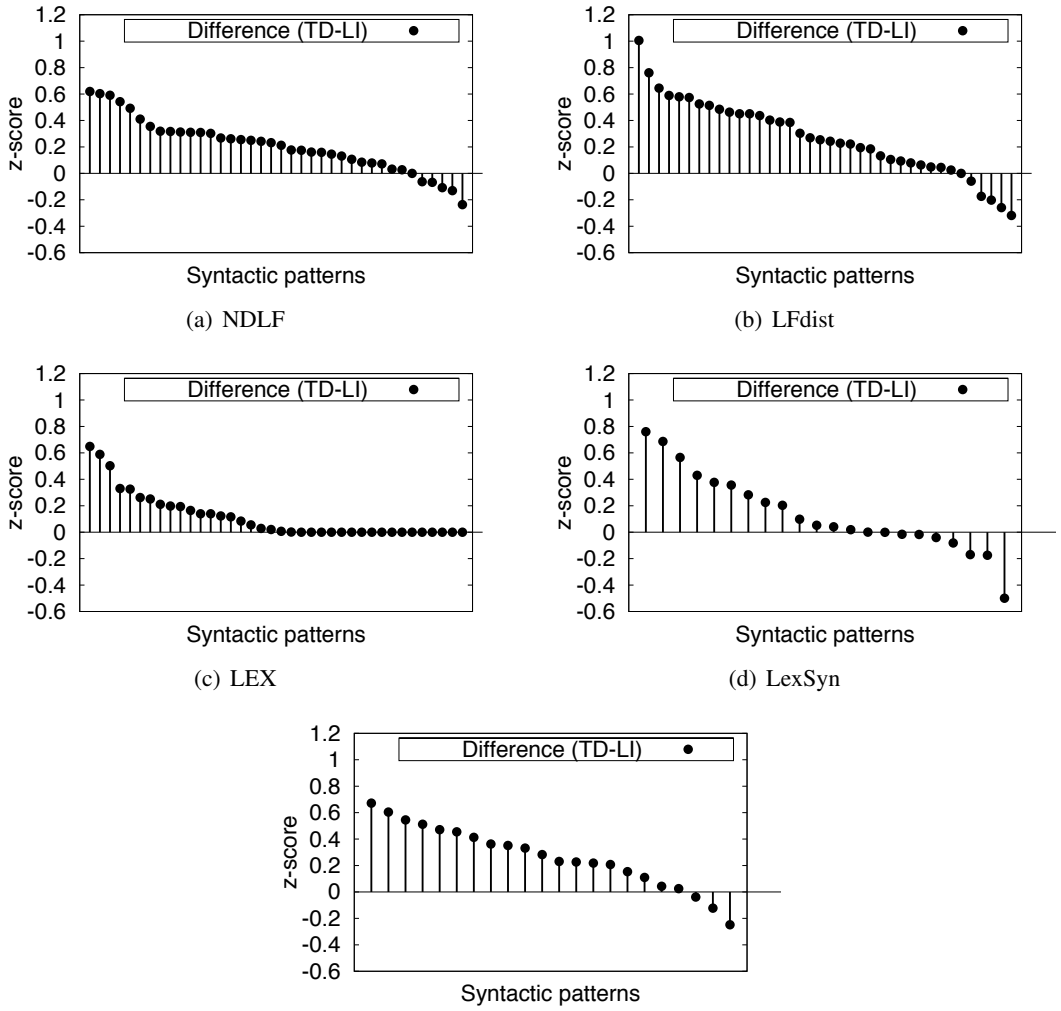
Analyzing *LEX* from Figure 1, we see that most of the *LEX* values are positive, for both TD and LI instances, and we cannot observe marked differences between them. That might be a consequence of assuming all word classes can have an equivalent number of different lexical forms. Once we weigh each POS tag in the pattern by the word forms the child has used (as in *LexSyn* and *LexSynEx*), noticeable differences across the two groups emerge. When we include syntactic knowledge of a group of children (as in *LexSynEx*), those similarities disappear. This behavior highlights the need for a combined lexico-syntactic measure that can describe latent information about language usage in children.

For building an intervention plan that helps to improve child language skills, practitioners could use the knowledge of specific grammatical constructions that need to be emphasized –those that seem to be problematic for the LI group. These structures can be identified by pulling the syntactic patterns with the largest difference in z-scores from the TD population. Table 2 shows a list of syntactic patterns with small values for LI and the largest differences between LI and TD instances in the test set. As the table indicates, most of the syntactic patterns have length greater than 1. This is not surprising since we aimed for developing measures of higher-order analysis that can complement the level of information provided by commonly used metrics in language assessment (as in the case of MLU, NDW or F/C). The table also shows that while each measure identifies a different subset of syntactic patterns as relevant, some syntactic patterns emerge in all the metrics. For instance, *personal pronoun + auxiliary verb + adverb* and *there + auxiliary verb*. This repetition highlights the importance of those grammatical constructions. But the differences also show that the metrics complement each other. In general, the syntactic patterns in the list represent complex grammatical constructions where children with LI are showing a less advanced command of language use.

| **LFdist** |
|---|
| verb (3rd person singular present) |
| verb (past tense) + personal pronoun |
| personal pronoun + auxiliary verb + adverb |
| verb (gerund) |
| **NDLF** |
| there + auxiliary verb |
| personal pronoun + auxiliary verb + adverb |
| adjective + noun |
| verb (3rd person singular present) |
| **LexSyn** |
| verb (past tense) + personal pronoun |
| personal pronoun + verb (past tense) + personal pronoun |
| personal pronoun + auxiliary verb + adverb |
| there + auxiliary verb |
| **LexSynEx** |
| personal pronoun + auxiliary verb + adverb |
| personal pronoun + verb (past tense) + personal pronoun |
| verb (past tense) + personal pronoun |
| there + auxiliary verb |

Table 2: List of syntactic patterns with the biggest difference between LI and TD in 4 measures: *LFdist*, *NDLF*, and *LexSyn* and *LexSynEx*.

Table 3 shows some statistics about the lexical forms present under *pronoun + verb (3rd person singular present) + verb (gerund or present participle)* (PP VBZ VBG) in all our data set. The last




(a) NDLF

(b) LFdist

(c) LEX

(d) LexSyn

(e) LexSynEx

Figure 1: Performance comparison of the proposed measures for the TD and LI groups. Each data point represents the difference in z-scores between the average values of the TD and LI instances in the test set.

row in that table presents an example of the lexical forms used by two children. Note that for the child with LI, there is only one lexical form: *he is touching*. On the other hand, the TD child is using the grammatical pattern with six different surface forms. Clinical practitioners can take this information and design language tasks that emphasize the use of 'PP VBZ VBG' constructions.

| | **TD** | **LI** |
|---|---|---|
| number of PP | 6 | 5 |
| number of VBZ | 3 | 2 |
| number of VBG | 7 | 4 |
| Example (instances: td-0156 and li-3022) | she is putting<br>she is going<br>he is pushing<br>she is looking<br>she is carrying<br>she is playing | he is touching |

Table 3: Statistics of the surface forms for the grammatical pattern *PP VBZ VBG*.

### 6.1 Analysis of correlations among measures

To analyze the level of overlap between our measures we computed correlation coefficients among them. The results are shown in Table 4.

The results from the correlation analysis are not that surprising. They show that closely related measures are highly to moderately correlated. For instance, LEX and *eLEX* have a correlation of 0.69, and *LexSynEx* and *LexSyn* have a correlation of 0.61. *NDLF* and *LFdist* showed a positive correlation score of 0.81. This high correlation hints to the fact that as the number of lexical forms increases, so does the gap between their fre-

| | LFdist | NDLF | LEX | eLEX | LexSyn | LexSynEx |
|---|---|---|---|---|---|---|
| LFdist | 1.00 | | | | | |
| NDLF | 0.81 | 1.00 | | | | |
| LEX | -0.53 | -0.31 | 1.00 | | | |
| eLEX | -0.54 | -0.43 | 0.69 | 1.00 | | |
| LexSyn | 0.07 | 0.02 | -0.23 | -0.10 | 1.00 | |
| LexSynEx | -0.02 | -0.03 | -0.08 | -0.03 | 0.61 | 1.00 |

Table 4: Correlation matrix for the proposed metrics.

quency of use. While this may be a common phenomenon of language use, it does not have a negative effect since the same effect will be observed in both groups of children and we care to see the differences in performance between a TD and an LI population.

For all other pairs of measures, the correlation scores were in the range of $[-0.5, 0.1]$. It was interesting to note that *LexSyn* showed the lowest correlation with the rest of the measures (between $[-0.11, 0.01]$).

Correlation coefficients between our metrics and MLU, NDW, and F/C were computed separately for syntactic patterns of different lengths. However all the different matrices showed the same correlation patterns. We found a *high* correlation between MLU and NDW, but low correlation with all our proposed measures, except for one case: NDW and LexSyn seemed to be highly correlated ($\sim$-0.7). Interestingly, we noted that despite the high correlation between MLU and NDW, MLU and LexSyn showed weak correlation ($\sim$-0.4). Overall, the findings from this analysis support the use of our metrics as complimentary measures for child language assessment.

## 7 Conclusions and future work

We proposed a set of new measures that were developed to characterize the lexico-syntactic variability of child language. Each measure aims to find information that is not captured by traditional measures used in communication disorders.

Our study is still preliminary in nature and requires an in depth evaluation and analysis with a larger pool of subjects. However the results presented are encouraging. The set of experiments we discussed showed that TD and LI children have significant differences in performance according to our metrics and thus these metrics can be used to enrich models of language trajectories in child language acquisition. Another potential use of metrics similar to those proposed here is the design of targeted intervention practices.

The scripts to compute the metrics as described in this paper are available to the research community by contacting the authors. However, the simplicity of the metrics makes it easy for anyone to implement, and it certainly makes it easy for clinical researchers to interpret.

Our proposed metrics are a contribution to the set of already known metrics for language assessment. The goal of these new metrics is not to replace existing ones, but to complement what is already available with concise information about higher-order syntactic constructions in the repertoire of TD children.

We are interested in evaluating the use of our metrics in a longitudinal study. We believe they are a promising framework to represent language acquisition trajectories.

## Acknowledgments

This research was partially funded by NSF under awards 1018124 and 1017190. The first author also received partial funding from CONACyT.

## References

Lisa M. Bedore, Elizabeth D. Peña, Ronald B. Gillam, and Tsung-Han Ho. 2010. Language sample measures and language ability in Spanish-English bilingual kindergarteners. *Journal of Communication Disorders*, 43:498–510.

Eugene Charniak. 2000. A maximum-entropy-inspired parser. In *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference*, NAACL 2000, pages 132–139, Stroudsburg, PA, USA. Association for Computational Linguistics.

Miao Chen and Klaus Zechner. 2011. Computing and evaluating syntactic complexity features for automated scoring of spontaneous non-native speech. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 722–731, Stroudsburg, PA, USA. Association for Computational Linguistics.

Keyur Gabani, Melissa Sherman, Thamar Solorio, Yang Liu, Lisa M. Bedore, and Elizabeth D. Peña. 2009. A corpus-based approach for the prediction of language impairment in monolingual English and Spanish-English bilingual children. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL '09, pages 46–55, Stroudsburg, PA, USA. Association for Computational Linguistics.

V. Gutiérrez-Clellen, G. Simon-Cereijido, and M. Sweet. 2012. Predictors of second language acquisition in Latino children with specific language impairment. *American Journal of Speech Language Pathology*, 21(1):64–77.

Khairun-nisa Hassanali, Yang Liu, and Thamar Solorio. 2012a. Coherence in child language narratives: A case study of annotation and automatic prediction of coherence. In *Proceedings of 3rd Workshop on Child, Computer and Interaction (WOCCI 2012)*.

Khairun-nisa Hassanali, Yang Liu, and Thamar Solorio. 2012b. Evaluating NLP features for automatic prediction of language impairment using child speech transcripts. In *Interspeech*.

Tom Kwiatkowski, Sharon Goldwater, Luke Zettlemoyer, and Mark Steedman. 2012. A probabilistic model of syntactic and semantic acquisition from child-directed utterances and their meanings. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 234–244, Avignon, France. Association for Computational Linguistics.

Xiaofei Lu. 2012. The relationship of lexical richness to the quality of ESL learners' oral narratives. *The Modern Language Journal*, 96(2):190–208.

Mercer Mayer. 1967. *A boy, a dog, and a frog*. Dial Press.

Mercer Mayer. 1969. *Frog, where are you?* Dial Press.

Jon F. Miller, John Heilmann, Ann Nockerts, Aquiles Iglesias, Leah Fabiano, and David J. Francis. 2006. Oral language and reading in bilingual children. *Learning Disabilities Research and Practice*, 21:30–43.

Elizabeth D. Peña, Lisa M. Bedore, Ronald B. Gillam, and Thomas Bohman. 2003. Diagnostic markers of language impairment in bilingual children. Grant awarded by the NIDCH, NIH.

Emily T. Prud'hommeaux, Brian Roark, Lois M. Black, and Jan van Santen. 2011. Classification of atypical language in autism. In *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics*, pages 88–96, Portland, Oregon, USA, June. Association for Computational Linguistics.

Brian Roark, Margaret Mitchell, and Kristy Hollingshead. 2007. Syntactic complexity measures for detecting mild cognitive impairment. In *Biological, translational, and clinical language processing*, pages 1–8, Prague, Czech Republic, June. Association for Computational Linguistics.

Brian Roark, Margaret Mitchell, John-Paul Hosom, Kristy Hollingshead, and Jeffrey Kaye. 2011. Spoken language derived measures for detecting mild cognitive impairment. *IEEE Transcations on Audio, Speech, and Language Processing*, 19(7):2081–2090, September.

Raúl Rojas and Aquiles Iglesias. 2012. The language growth of Spanish-speaking English language learners. *Child Development*.

Kenji Sagae, Alon Lavie, and Brian MacWhinney. 2005. Automatic measurement of syntactic development in child language. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, ACL '05, pages 197–204, Stroudsburg, PA, USA. Association for Computational Linguistics.

Sam Sahakian and Benjamin Snyder. 2012. Automatically learning measures of child language development. In *ACL*, pages 95–99. The Association for Computational Linguistics.

Thamar Solorio, Melissa Sherman, Y. Liu, Lisa Bedore, Elizabeth Peña, and A. Iglesias. 2011. Analyzing language samples of Spanish-English bilingual children for the automated prediction of language dominance. *Natural Language Engineering*, pages 367–395.

Charles Yang. 2011. A statistical test for grammar. In *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics*, pages 30–38, Portland, Oregon, USA, June. Association for Computational Linguistics.

Su-Youn Yoon and Suma Bhat. 2012. Assessment of ESL learners' syntactic competence based on similarity measures. In *EMNLP-CoNLL*, pages 600–608. Association for Computational Linguistics.