A Corpus-Based Approach for the Prediction of Language Impairment in Monolingual English and Spanish-English Bilingual Children

Keyur Gabani and **Melissa Sherman** and **Thamar Solorio** and **Yang Liu**
Department of Computer Science
The University of Texas at Dallas
keyur,mesh,tsolorio,yangl@hlt.utdallas.edu

Lisa M. Bedore and **Elizabeth D. Peña**
Department of Communication Sciences and Disorders
The University of Texas at Austin
lbedore,lizp@mail.utexas.edu

Abstract

In this paper we explore a learning-based approach to the problem of predicting language impairment in children. We analyzed spontaneous narratives of children and extracted features measuring different aspects of language including morphology, speech fluency, language productivity and vocabulary. Then, we evaluated a learning-based approach and compared its predictive accuracy against a method based on language models. Empirical results on monolingual English-speaking children and bilingual Spanish-English speaking children show the learning-based approach is a promising direction for automatic language assessment.

1 Introduction

The question of how best to identify children with language disorders is a topic of ongoing debate. One common assessment approach is based on cutoff scores from standardized, norm-referenced language assessment tasks. Children scoring at the lower end of the distribution, typically more than 1.25 or 1.5 Standard Deviations (SD) below the mean, are identified as having language impairment (Tomblin et al., 1997). This cutoff-based approach has several well-documented weaknesses that may result in both over- and under-identification of children as language impaired (Plante and Vance, 1994). Recent studies have suggested considerable overlap between children with language impairment and their typically developing cohorts on many of these tasks (e.g., (Peña et al., 2006b; Spaulding et al., 2006)). In addition, scores and cutoffs on standardized tests depend on the distribution of scores from the particular samples used in normalizing the measure. Thus, the validity of the measure for children whose demographic and other socioeconomic characteristics are not well represented in the test's normative sample is a serious concern. Finally, most norm-referenced tests of language ability rely heavily on exposure to mainstream language and experiences, and have been found to be biased against children from families with low parental education and socioeconomic status, as well as children from different ethnic backgrounds (Campbell et al., 1997; Dollaghan and Campbell, 1998).

This paper aims to develop a reliable and automatic method for identifying the language status of children. We propose the use of different lexico-syntactic features, typically used in computational linguistics, in combination with features inspired by current assessment practices in the field of language disorders to train Machine Learning (ML) algorithms. The two main contributions of this paper are: 1) It is one step towards developing a reliable and automatic approach for language status prediction in English-speaking children; 2) It provides evidence showing that the same approach can be adapted to predict language status in Spanish-English bilingual children.

2 Related Work

2.1 Monolingual English-Speaking Children

Several hypotheses exist that try to explain the grammatical deficits of children with Language Impair-

Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the ACL, pages 46–55,
Boulder, Colorado, June 2009. ©2009 Association for Computational Linguistics

ment (LI). Young children normally go through a stage where they use non-finite forms of verbs in grammatical contexts where finite forms are required (Wexler, 1994). This is referred as the optional infinitive stage. The Extended Optional Infinitive (EOI) theory (Rice and Wexler, 1996) suggests that children with LI exhibit the use of a "young" grammar for an extended period of time, where tense, person, and number agreement markers are omitted.

In contrast to the EOI theory, the surface account theory (Leonard et al., 1997) assumes that children with LI have reduced processing capabilities. This deficit affects the perception of low stress morphemes, such as *-ed, -s, be* and *do*, resulting in an inconsistent use of these verb morphemes.

Spontaneous narratives are considered as one of the most ecologically valid ways to measure communicative competence (Botting, 2002). They represent various aspects involved in children's everyday communication. Typical measures for spontaneous language samples include Mean Length of Utterance (MLU) in words, Number of Different Words (NDW), and errors in grammatical morphology. Assessment approaches compare children's performance on these measures against expected performance. As mentioned in Section 1, these cutoff based methods raise questions concerning accuracy and bias. Manually analyzing the narratives is also a very time consuming task. After transcribing the sample, clinicians need to code for the different clinical markers and other morphosyntactic information. This can take up to several hours for each child making it infeasible to analyze a large number of samples.

2.2 Bilingual Spanish-English Speaking Children

Bilingual children face even more identification challenges due to their dual language acquisition. They can be mistakenly labeled as LI due to: 1) the inadequate use of translations of assessment tools; 2) an over reliance on features specific to English; 3) a lack of appropriate expectations about how the languages of a bilingual child should develop (Bedore and Peña, 2008); 4) or the use of standardized tests where the normal distribution used to compare language performance is composed of monolingual children (Restrepo and Gutiérrez-Clellen, 2001).

Spanish speaking children with LI show different clinical markers than English speaking children with LI. As mentioned above, English speakers have problems with verb morphology. In contrast, Spanish speakers have been found to have problems with noun morphology, in particular in the use of articles and clitics (Restrepo and Gutiérrez-Clellen, 2001; Jacobson and Schwartz, 2002; Bedore and Leonard, 2005). Bedore and Leonard (2005) also found differences in the error patterns of Spanish and related languages such as Italian. Spanish-speakers tend to both omit and substitute articles and clitics, while the dominant errors for Italian-speakers are omissions.

3 Our Approach

We use language models (LMs) in our initial investigation, and later explore more complex ML algorithms to improve the results. Our ultimate goal is to discover a highly accurate ML method that can be used to assist clinicians in the task of LI identification in children.

3.1 Language Models for Predicting Language Impairment

LMs are statistical models used to estimate the probability of a given sequence of words. They have been explored previously for clinical purposes. Roark *et al.* (2007) proposed cross entropy of LMs trained on Part-of-Speech (POS) sequences as a measure of syntactic complexity with the aim of determining mild cognitive impairment in adults. Solorio and Liu (2008) evaluated LMs on a small data set in a preliminary trial on LI prediction.

The intuition behind using LMs is that they can identify atypical grammatical patterns and help discriminate the population with potential LI from the Typically Developing (TD) one. We use LMs trained on POS tags rather than on words. Using POS tags can address the data sparsity issue in LMs, and place less emphasis on the vocabulary and more emphasis on the syntactic patterns.

We trained two separate LMs using POS tags from the transcripts of TD and LI children, respectively. The language status of a child is predicted using the following criterion:

$$d(s) = \begin{cases} \text{LI} & if\ (PP_{TD}(s) > PP_{LI}(s)) \\ \text{TD} & otherwise \end{cases}$$

where s represents a transcript from a child, and $PP_{TD}(s)$ and $PP_{LI}(s)$ are the perplexity values from the TD and LI LMs, respectively. We used the SRI Language Modeling Toolkit (Stolcke, 2002) for training the LMs and calculating perplexities.

3.2 Machine Learning for Predicting Language Impairment

Although LMs have been used successfully on different human language processing tasks, they are typically trained and tested on language samples larger than what is usually collected by clinicians for the purpose of diagnosing a child with potential LI. Clinicians make use of additional information beyond children's speech, such as parent and teacher questionnaires and test scores on different language assessment tasks. Therefore in addition to using LMs for children language status prediction, we explore a machine learning classification approach that can incorporate more information for better prediction. We aim to identify effective features for this task and expect this information will help clinicians in their assessment.

We consider various ML algorithms for the classification task, including Naive Bayes, Artificial Neural Networks (ANNs), Support Vector Machines (SVM), and Boosting with Decision Stumps. Weka (Witten and Frank, 1999) was used in our experiments due to its known reliability and the availability of a large number of algorithms. Below we provide a comprehensive list of features that we explored for both English and Spanish-English transcripts. We group these features according to the aspect of language they focus on. Features specific to Spanish are discussed in Section 5.2.

1. *Language productivity*

   (a) *Mean Length of Utterance (MLU) in words*
   Due to a general deficit of language ability, children with LI have been found to produce language samples with a shorter MLU in words because they produce grammatically simpler sentences when compared to their TD peers.

   (b) *Total number of words*
   This measure is widely used when building language profiles of children for diagnostic and treatment purposes.

   (c) *Degree of support*
   In spontaneous samples of children's speech, it has been pointed out that children with potential LI need more encouragement from the investigator (Wetherell et al., 2007) than their TD peers. A support prompt can be a question like "*What happened next*?" We count the number of utterances, or turns, of the investigator interviewing the child for this feature.

2. *Morphosyntactic skills*

   (a) *Ratio of number of raw verbs to the total number of verbs*
   As mentioned previously, children with LI omit tense markers in verbs more often than their TD cohorts. For example:

   *...the boy **look** into the hole but didn't find...*

   Hence, we include the ratio of the number of raw verbs to the total number of verbs as a feature.

   (b) *Subject-verb agreement*
   Research has shown that English-speaking children with LI have difficulties marking subject-verb agreement (Clahsen and Hansen, 1997; Schütze and Wexler, 1996). An illustration of subject-verb disagreement is the following:

   *...and **he were** looking behind the rocks*

   As a way of capturing this information in the machine learning setting, we consider various bigrams of POS tags: noun and verb, noun and auxiliary verb, pronoun and verb, and pronoun and auxiliary verb. These features are included in a bag-of-words fashion using individual counts. Also, we allow a window between these pairs to capture agreement between subject and verb that may have modifiers in between.

(c) *Number of different POS tags*

This feature is the total number of different POS tags in each transcript.

3. *Vocabulary knowledge*

   We use the Number of Different Words (NDW) to represent vocabulary knowledge of a child. Although such measures can be biased against children from different backgrounds, we expect this possible negative effect to decrease as a result of having a richer pool of features.

4. *Speech fluency*

   Repetitions, revisions, and filled pauses have been considered indicators of language learning difficulties (Thordardottir and Weismer, 2002; Wetherell et al., 2007). In this work we include as features (a) the number of fillers, such as *uh, um, er*; and (b) the number of disfluencies (abandoned words) found in each transcript.

5. *Perplexities from LMs*

   As mentioned in Section 3.1 we trained LMs of order 1, 2, and 3 on POS tags extracted from TD and LI children. We use the perplexity values from these models as features. Additionally, differences in perplexity values from LI and TD LMs for different orders are used as features.

6. *Standard scores*

   A standard score, known as a z-score, is the difference between an observation and the mean relative to the standard deviation. For this feature group, we first find separate distributions for the MLU in words, NDW and total number of utterances for the TD and LI populations. Then, for each transcript, we compute the standard scores based on each of these six distributions. This represents how well the child is performing relative to the TD and LI populations. Note that a cross validation setup was used to obtain the distribution for the TD and LI children for training. This is also required for the LM features above.

4 Experiments with Monolingual Children

4.1 The Monolingual English Data Set

Our target population for this work is children with an age range of 3 to 6 years old. However, currently we do not have any monolingual data sets readily available to test our approach in this age range. In the field of communication disorders data sharing is not a common practice due to the sensitive content of the material in the language samples of children, and also due to the large amount of effort and time it takes researchers to collect, transcribe, and code the data before they can begin their analysis. To evaluate our approach we used a dataset from CHILDES (MacWhinney, 2000) that includes narratives from English-speaking adolescents with and without LI with ages ranging between 13 and 16 years old. Even though the age range is outside the range we are interested in, we believe that this data set can still be helpful in exploring the feasibility of our approach as a first step.

This data set contains 99 TD adolescents and 19 adolescents who met the LI profile at one point in the duration of the study. There are transcripts from each child for two tasks: a story telling and a spontaneous personal narrative. The first task is a picture prompted story telling task using the wordless picture book, "*Frog, Where Are You?*" (Mayer, 1969). In this story telling task children first look at the story book –to develop a story in memory– and then are asked to narrate the story. This type of elicitation task encourages the use of past tense constructions, providing plenty of opportunities for extracting clinical markers. In the spontaneous personal narrative task, the child is asked to talk about a person who annoys him/her the most and describe the most annoying features of that person. This kind of spontaneous personal narrative encourages the participant for the use of third person singular forms (*-s*). Detailed information of this data set can be found in (Wetherell et al., 2007).

We processed the transcripts using the CLAN toolkit (MacWhinney, 2000). MOR and POST from CLAN are used for morphological analysis and POS tagging of the children's speech. We decided to use these analyzers since they are customized for children's speech.

	Story telling			Personal narrative		
Method	**P (%)**	**R (%)**	**F_1 (%)**	**P (%)**	**R (%)**	**F_1 (%)**
Baseline	28.57	10.53	15.38	33.33	15.79	21.43
1-gram LMs	41.03	84.21	55.17	34.21	68.42	45.61
2-gram LMs	75.00	47.37	**58.06**	55.56	26.32	35.71
3-gram LMs	80.00	21.05	33.33	87.50	36.84	**51.85**

Table 1: Evaluation of language models on the monolingual English data set.

	Story telling			Personal narrative		
Algorithm	**P (%)**	**R (%)**	**F_1 (%)**	**P (%)**	**R (%)**	**F_1 (%)**
Naive Bayes	38.71	63.16	48.00	34.78	42.11	38.10
Bayesian Network	58.33	73.68	65.12	28.57	42.11	34.04
SVM	76.47	68.42	**72.22**	47.06	42.11	44.44
ANNs	62.50	52.63	57.14	50.00	47.37	48.65
Boosting	70.59	63.16	66.67	69.23	47.37	**56.25**

Table 2: Evaluation of machine learning algorithms on the monolingual English data set.

4.2 Results with Monolingual English-Speaking Children

The performance measures we use are: precision (P), recall (R), and F-measure (F_1). Here the LI category is the positive class and the TD category is the negative class.

Table 1 shows the results of leave-one-out-cross-validation (LOOCV) obtained from the LM approach for the story telling and spontaneous personal narrative tasks. It also shows results from a baseline method that predicts language status by using standard scores on measures that have been associated with LI in children (Dollaghan, 2004). The three measures we used for the baseline are: MLU in words, NDW, and total number of utterances produced. To compute this baseline we estimate the mean and standard deviation of these measures using LOOCV with the TD population as our normative sample. The baseline predicts that a child has LI if the child scores more than 1.25 SD below the mean on at least two out of the three measures.

Although LMs yield different results for the story telling and personal narrative tasks, they both provide consistently better results than the baseline. For the story telling task the best results, in terms of the F_1 measure, are achieved by a bigram LM (F_1 = 58.06%) while for the personal narrative the highest F_1 measure (51.85%) is from the trigram LM. If we consider precision, both tasks have the same increasing pattern when increasing LM orders. However for recall that is not the case. In the story telling task, recall decreases at the expense of higher precision, but for the personal narrative task, the trigram LM reaches a better trade-off between precision and recall, which yields a high F_1 measure. We also evaluated 4-gram LMs, but results did not improve, most likely because we do not have enough data to train higher order LMs.

The results for different ML algorithms are shown in Table 2, obtained by using all features described in Section 3.2. The feature based approach using ML algorithms outperformed using only LMs on both tasks. For the story telling task, SVM with a linear kernel achieves the best results (F_1 = 72.22%), while Boosting with Decision Stumps provides the best performance (F_1 = 56.25%) for the personal narrative task.

4.3 Feature and Error Analysis

The ML results shown above use the entire feature set described in Subsection 3.2. The next question we ask is the effectiveness of different features for this task. The datasets we are using in our evaluation are very small, especially considering the number of positive instances. This prevents us from having a separate subset of the data for parameter tuning or feature selection. Therefore, we performed additional experiments to evaluate the usefulness of individual features. Figure 1 shows the F_1 measures

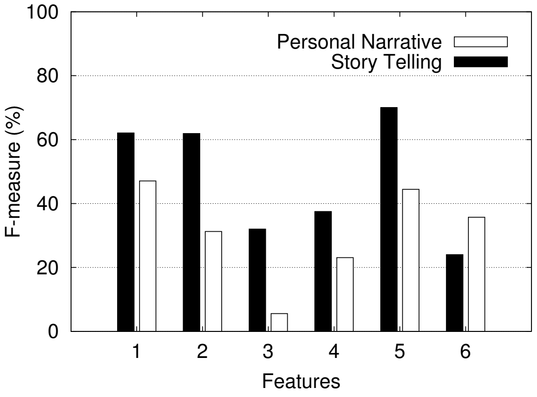

Figure 1: Discriminating power of different groups of features. The numbers on the x-axis correspond to the feature groups in Section 3.2.

when using different feature groups. The numbers on the x-axis correspond to the feature groups described in Section 3.2. The F_1 measure value for each of the features is the highest value obtained by running different ML algorithms for classification.

We noticed that for the story telling task, using perplexity values from LMs (group 5) as a feature in the ML setting outperforms the LM threshold approach by a large margin. It seems that having the perplexity values as well as the perplexity differences from all the LMs of different orders in the ML algorithm provides a better estimation of the target concept.

Only the standard scores (group 6) yield a higher F_1 measure for the personal narrative task than the story telling one. The majority of the features (5 out of 6 groups) provide higher F_1 measures for the story telling task, which explains the significantly better results on this task over the personal narrative in our learning approach. This is consistent with previous work contrasting narrative genre stating that the restrictive setting of a story retell is more revealing of language difficulties than spontaneous narratives, where the subjects have more control on the content and style (Wetherell et al., 2007).

We also performed some error analysis for some of the transcripts that were consistently misidentified by different ML algorithms. In the story telling task, we find that some LI transcripts are misclassified as TD because they (1) have fewer fillers, disfluencies, and degree of support; (2) are similar to the TD transcripts, which is depicted by the perplexity values for these transcripts; or (3) contain higher MLU in words as compared to their LI peers. Some of the reasons for classifying transcripts in the TD category as LI are shorter MLU in words as compared to other TD peers, large number of fillers, and excessive repetitions of words and phrases unlike the other TD children. These factors are consistent with the effective features that we found from Figure 1.

For the personal narrative task, standard scores (group 6) and language productivity (group 1) have an important role in classification, as shown in Figure 1. The TD transcripts that are misidentified have lower standard scores and MLU in words than those of their TD peers.

We believe that another source of noise in the transcripts comes from the POS tags themselves. For instance, we found that many verbs in present tense for third person singular are tagged as plural nouns, which results in a failure to capture subject-verb agreement.

Lastly, according to the dataset description, children in the LI category met the LI criteria at one stage in their lifetime and some of these children also had, or were receiving, some educational support in the school environment at the time of data collection. This support for children with LI is meant to improve their performance on language related tasks, making the automatic classification problem more complicated. This also raises the question about the reference label (TD or LI) for each child in the data set we used. The details about which children received interventions are not specified in the dataset description.

5 Experiments with Bilingual Children

In this section we generalize the approach to a Spanish-English bilingual population. In adapting the approach to our bilingual population we face two challenges: first, what shows to be promising for a monolingual and highly heterogeneous population may not be as successful in a bilingual setting where we expect to have a large variability of exposure to each language; second, there is a large difference in the mean age of the monolingual setting and that of our bilingual one. This age difference will result in different speech patterns. Younger children produce more ill-formed sentences since they are still in a language acquisition phase. Lastly, the clinical markers in adolescents are geared towards problems at the pragmatic and discourse levels, while at younger ages they focus more on syntax and morphology.

For dealing with the first challenge we are extracting language-specific features and hope that by looking at both languages we can reach a good discrimination performance. For the second challenge, our feature engineering approach has been focused on younger children from the beginning. We are aiming to capture the type of morphosyntactic patterns that can identify LI in young children. In addition, the samples in the bilingual population are story retells, and our feature setting showed to be a good match for this task. Therefore, we expect our approach to capture relevant classification patterns, even in the presence of noisy utterances.

5.1 The Bilingual Data Set

The transcripts for the bilingual LI task come from an on-going longitudinal study of language impairment in Spanish-English speaking children (Peña et al., 2006a). The children in this study were enrolled in kindergarten with a mean age of about 70 months. Of the 59 children, 6 were identified as having a possible LI by an expert in communication disorders, while 53 were identified as TD. Six of the TD children were excluded due to missing information, yielding a total of 47 TD children.

Each child told a series of stories based on Mercer Mayer's wordless picture books (Mayer, 1969). Two stories were told in English and two were told in Spanish, for a total of four transcripts per child. The books used for English were *"A Boy, A Dog, and A Frog"* and *"Frog, Where Are You?"* The books used for Spanish retelling were *"Frog on His Own"* and *"Frog Goes to Dinner."* The transcripts for each separate language were combined, yielding one instance per language for each child.

An interesting aspect of the bilingual data is that the children mix languages in their narratives. This phenomenon is called code-switching. At the beginning of a retelling session, the interviewer encourages the child to speak the target language if he/she is not doing so. Once the child begins speaking the correct language, any code-switching thereafter is not corrected by the interviewer. Due to this, the English transcripts contain Spanish utterances and vice versa. We believe that words in the non-target language help contribute to a more accurate language development profile. Therefore, in our work we decided to keep these code-switched elements. A combined lexicon approach was used to tag the mixed-language fragments. If a word does not appear in the target language lexicon, we apply the POS tag from the non-target language.

5.2 Spanish-Specific Features

Many structural differences exist between Spanish, a Romance language, and English, a Germanic language. Spanish is morphologically richer than English. It contains a larger number of different verb conjugations and it uses a two gender system for nouns, adjectives, determiners, and participles. A Spanish-speaking child with LI will have difficulties with different grammatical elements, such as articles and clitics, than an English-speaking child (Bedore and Peña, 2008). These differences indicate that the Spanish feature set will need to be tailored towards the Spanish language.

To account for Spanish-specific patterns we included new POS bigrams as features. To capture the use of correct and incorrect gender and number marking morphology, we added noun-adjective, determiner-noun, and number-noun bigrams to the list of morphosyntactic features.

5.3 Results on Bilingual Children

Results are shown for the baseline and LM threshold approach for the bilingual data set in Table 3. The baseline is computed from the same measures as the monolingual dataset (MLU in words, NDW, and total utterances).

Compared to Table 1, the values in Table 3 are generally lower than on the monolingual story telling task. In this inherently difficult task, the bilingual transcripts are more disfluent than the monolingual ones. This could be due to the age of the children or their bilingual status. Recent studies on psycholinguistics and language production have shown that bilingual speakers have both languages active at speech production time (Kroll et al., 2008) and it is possible that this may cause interference, especially in children still in the phase of language acqui-

	English			Spanish		
Method	**P (%)**	**R (%)**	**F_1 (%)**	**P (%)**	**R (%)**	**F_1 (%)**
Baseline	20.00	16.66	18.18	16.66	16.66	16.66
1-gram LMs	40.00	33.33	36.36	17.64	50.00	**26.08**
2-gram LMs	50.00	33.33	40.00	33.33	16.66	22.22
3-gram LMs	100.00	33.33	**50.00**	0.00	0.00	-

Table 3: Evaluation of language models on Bilingual Spanish-English data set.

sition. In addition, the LMs in the monolingual task were trained using more instances per class, possibly yielding better results.

There are some different patterns between using the English and Spanish transcripts. In English, the unigram models provide the least discriminative value, and the bigram and trigram models improve discrimination. We also evaluated higher order n-grams, but did not obtain any further improvement. We found that the classification accuracy of the LM approach was influenced by two children with LI who were consistently marked as LI due to a greater perplexity value from the TD LM. A further analysis shows that these children spoke mostly Spanish on the "English" tasks yielding larger perplexities from the TD LM, which was trained from mostly English. In contrast, the LI LM was created with transcripts containing more Spanish than the TD one, and thus test transcripts with a lot of Spanish do not inflate perplexity values that much.

For Spanish, unigram LMs provide some discriminative usefulness, and then the bigram performance decreases while the trigram model provides no discriminative value. One reason for this may be that the Spanish LMs have a larger vocabulary. In the Spanish LMs, there are 2/3 more POS tags than in the English LM. This size difference dramatically increases the possible bigrams and trigrams, therefore increasing the number of parameters to estimate. In addition, we are using an "off the shelf" POS tagger (provided by CLAN) and this may add noise in the feature extraction process. Since we do not have gold standard annotations for these transcripts, we cannot measure the POS tagging accuracy. A rough estimate based on manually revising one transcript in each language showed a POS tagging accuracy of 90% for English and 84% for Spanish. Most of the POS tagger errors involve verbs, nouns and pronouns. Thus while the accuracy might not seem that low, it can still have a major impact on our approach since it involves the POS categories that are more relevant for this task.

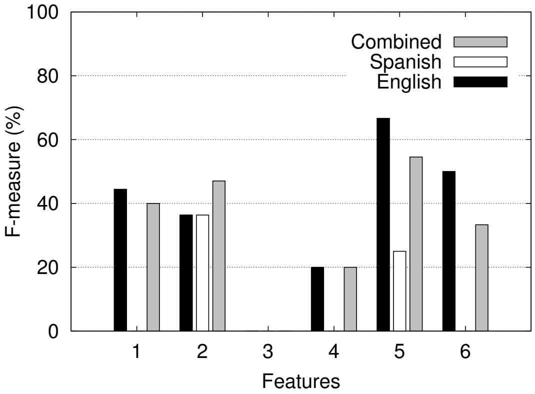

Figure 2: Discriminating power of different groups of features for the bilingual population. The numbers on the x-axis correspond to the feature groups in Section 3.2.

Table 4 shows the results from various ML algorithms. In addition to predicting the language status with the English and Spanish samples separately, we also combined the English and Spanish transcripts together for each child, and used all the features from both languages in order to allow a prediction based on both samples. The best F_1 measure for this task (60%) is achieved by using the Naive Bayes algorithm with the combined Spanish-English feature set. This is an improvement over both the separate English and Spanish trials. The Naive Bayes algorithm provided the best discrimination for the English (54%) and Combined data sets and Boosting and SVM provided the best discrimination for the Spanish set (18%).

5.4 Feature Analysis

Similar to the monolingual dataset, we performed additional experiments exploring the contribution of different groups of features. We tested the six

	English			Spanish			Combined		
Algorithm	**P (%)**	**R (%)**	**$\mathbf{F_1}$ (%)**	**P (%)**	**R (%)**	**$\mathbf{F_1}$ (%)**	**P (%)**	**R (%)**	**$\mathbf{F_1}$ (%)**
ANNs	66.66	33.33	44.44	0.00	0.00	-	100.00	16.66	28.57
SVM	14.28	16.66	15.38	20.00	16.66	18.18	66.66	33.33	44.44
Naive Bayes	60.00	50.00	**54.54**	0.00	0.00	-	75.00	50.00	**60.00**
Logistic Regression	25.00	16.66	20.00	-	0.00	-	50.00	33.33	40.00
Boosting	50.00	33.33	40.00	20.00	16.66	18.18	66.66	33.33	44.44

Table 4: Evaluation of machine learning algorithms on the Bilingual Spanish-English data set.

groups of features described in Section 3.2 separately. Overall, the combined LM perplexity values (group 5) provided the best discriminative value ($F_1 = 66\%$). The LM perplexity values performed the best for English. It even outperformed using all the features in the ML algorithm, suggesting some feature selection is needed for this task.

The morpohsyntactic skills (group 2) provided the best discriminative value for the Spanish language features, and performed better than the complete feature set for Spanish. Within group 2, we evaluated different POS bigrams for the Spanish and English sets and observed that most of the bigram combinations by themselves are usually weak predictors of language status. In the Spanish set, out of all of the lexical combinations, only the determiner-noun, noun-verb, and pronoun-verb categories provided some discriminative value. The determiner-noun category captured the correct and incorrect gender marking between the two POS tags. The noun-verb and pronoun-verb categories covered the correct and incorrect usage of subject-verb combinations. Interestingly enough, the pronoun-verb category performed well by itself, yielding an F_1 measure of 54%. There are also some differences in the frequencies of bigram features in the English and Spanish data sets. For example, there is no noun-auxiliary POS pattern in Spanish, and the pronoun-auxiliary bigram appears less frequently in Spanish than in English because in Spanish the use of personal pronouns is not mandatory since the verb inflection will disambiguate the subject of the sentence.

The vocabulary knowledge feature (group 3) did not provide any discriminative value for any of the language tasks. This may be because bilingual children receive less input for each language than a monolingual child learning one language, or due to the varied vocabulary acquisition rate in our bilingual population.

6 Conclusions and Future Work

In this paper we present results on the use of LMs and ML techniques trained on features representing different aspects of language gathered from spontaneous speech samples for the task of assisting clinicians in determining language status in children. First, we evaluate our approach on a monolingual English-speaking population. Next, we show that this ML approach can be successfully adapted to a bilingual Spanish-English population. ML algorithms provide greater discriminative power than only using a threshold approach with LMs.

Our current efforts are devoted to improving prediction accuracy by refining our feature set. We are working on creating a gold standard corpus of children's transcripts annotated with POS tags. This data set will help us improve accuracy on our POS-based features. We are also exploring the use of socio-demographic features such as the educational level of parents, the gender of children, and enrollment status on free lunch programs.

Acknowledgments

This work was supported by NSF grant 0812134, and by grant 5 UL1 RR024982 from NCRR, a component of NIH. We also thank the three NAACL reviewers for insightful comments on the submitted version of this paper.

References

Lisa M. Bedore and Laurence B. Leonard. 2005. Verb inflections and noun phrase morphology in the spontaneous speech of Spanish-speaking children with specific language impairment. *Applied Psycholinguistics*, 26(2):195–225.

Lisa M. Bedore and Elizabeth D. Peña. 2008. Assessment of bilingual children for identification of language impairment: Current findings and implications for practice. *International Journal of Bilingual Education and Bilingualism*, 11(1):1–29.

Nicola Botting. 2002. Narrative as a tool for the assessment of linguistic and pragmatic impairments. *Child Language Teaching and Therapy*, 18(1):1–21.

Thomas Campbell, Chris Dollaghan, Herbert Needleman, and Janine Janosky. 1997. Reducing bias in language assessment: Processing-dependent measures. *Journal of Speech, Language, and Hearing Research*, 40(3):519–525.

Harald Clahsen and Detlef Hansen. 1997. The grammatical agreement deficit in specific language impairment: Evidence from therapy experiments. In Myrna Gopnik, editor, *The Inheritance and Innateness of Grammar*, chapter 7. Oxford University Press, New York.

Christine A. Dollaghan and Thomas F. Campbell. 1998. Nonword repetition and child language impairment. *Journal of Speech, Language, and Hearing Research*, 41(5):1136–1146.

Christine A. Dollaghan. 2004. Taxometric analyses of specific language impairment in 3- and 4-year-old children. *Journal of Speech, Language, and Hearing Research*, 47(2):464–475.

Peggy F. Jacobson and Richard G. Schwartz. 2002. Morphology in incipient bilingual Spanish-speaking preschool children with specific language impairment. *Applied Psycholinguistics*, 23(1):23–41.

Judith F. Kroll, Chip Gerfen, and Paola E. Dussias. 2008. Laboratory designs and paradigms: Words, sounds, sentences. In L. Wei and M. G. Moyer, editors, *The Blackwell Guide to Research Methods in Bilingualism and Multilingualism*, chapter 7. Blackwell Pub.

Laurence B. Leonard, Julia A. Eyer, Lisa M. Bedore, and Bernard G. Grela. 1997. Three accounts of the grammatical morpheme difficulties of English-speaking children with specific language impairment. *Journal of Speech, Language, and Hearing Research*, 40(4):741–753.

Brian MacWhinney. 2000. *The CHILDES project: Tools for analyzing talk*. Lawrence Erlbaum, Mahwah, NJ.

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

Elizabeth D. Peña, Lisa M. Bedore, Ronald B. Gillam, and Thomas Bohman. 2006a. Diagnostic markers of language impairment in bilingual children. Grant awarded by the NIDCD, NIH.

Elizabeth D. Peña, Tammie J. Spaulding, and Elena Plante. 2006b. The composition of normative groups and diagnostic decision making: Shooting ourselves in the foot. *American Journal of Speech-Language Pathology*, 15(3):247–254.

Elena Plante and Rebecca Vance. 1994. Selection of preschool language tests: A data-based approach. *Language, Speech, and Hearing Services in Schools*, 25(1):15–24.

María Adelaida Restrepo and Vera F. Gutiérrez-Clellen. 2001. Article use in Spanish-speaking children with specific language impairment. *Journal of Child Language*, 28(2):433–452.

Mabel L. Rice and Kenneth Wexler. 1996. Toward tense as a clinical marker of specific language impairment in English-speaking children. *Journal of Speech and Hearing Research*, 39(6):1239–1257.

Brian Roark, Margaret Mitchell, and Kristy Hollingshead. 2007. Syntactic complexity measures for detecting mild cognitive impairment. In *Proceedings of the Workshop on BioNLP 2007*, pages 1–8. ACL.

Carson T. Schütze and Kenneth Wexler. 1996. Subject case licensing and English root infinitives. In *Proceedings of the 20th Annual Boston University Conference on Language Development*. Cascadilla Press.

Thamar Solorio and Yang Liu. 2008. Using language models to identify language impairment in Spanish-English bilingual children. In *Proceedings of the Workshop on BioNLP 2008*, pages 116–117. ACL.

Tammie J. Spaulding, Elena Plante, and Kimberly A. Farinella. 2006. Eligibility criteria for language impairment: Is the low end of normal always appropriate? *Language, Speech, and Hearing Services in Schools*, 37(1):61–72.

Andreas Stolcke. 2002. SRILM – an extensible language modeling toolkit. In *Proceedings of the International Conference on Spoken Language Processing*, volume 2, pages 901–904.

Elin T. Thordardottir and Susan Ellis Weismer. 2002. Content mazes and filled pauses on narrative language samples of children with specific language impairment. *Brain and Cognition*, 48(2-3):587–592.

J. Bruce Tomblin, Nancy L. Records, Paula Buckwalter, Xuyang Zhang, Elaine Smith, and Marlea O'Brien. 1997. Prevalence of specific language impairment in kindergarten children. *Journal of Speech, Language, and Hearing Research*, 40(6):1245–1260.

Danielle Wetherell, Nicola Botting, and Gina Conti-Ramsden. 2007. Narrative in adolescent specific language impairment (SLI): a comparison with peers across two different narrative genres. *International Journal of Language and Communication Disorders*, 42:583–605(23).

Kenneth Wexler. 1994. Optional infinitives. In David Lightfoot and Norbert Hornstein, editors, *Verb Movement*. Cambridge University Press.

Ian H. Witten and Eibe Frank. 1999. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann.