Language dominance prediction in Spanish-English bilingual children using syntactic information: a first approximation

Gabriela Ramirez-de-la-Rosa[1], Thamar Solorio[1], Manuel Montes-y-Gómez[1,2], Yang Liu[3], Aquiles Iglesias[4], Lisa Bedore[5], and Elizabeth Peña[5]

[1] Department of Computer and Information Sciences, University of Alabama at Birmingham, gabyrr,solorio@cis.uab.edu
[2] Coordinación de Ciencias Computacionales, INAOE, mmontesg@inaoep.mx
[3] Department of Computer Science, The University of Texas at Dallas, yangl@hlt.utdallas.edu
[4] Department of Communication Sciences and Disorders, Temple University, iglesias@temple.edu
[5] Department of Communication Sciences and Disorders, The University of Texas at Austin, lbedore,lizp@mail.utexas.edu

Abstract. This paper presents results on a preliminary study using syntactic information to predict language dominance in Spanish-English bilingual children. Our approach uses a bag of syntactic grammar rules taken from narratives in English and Spanish. We then measure prediction accuracy of categorizing children into Spanish-dominant, English-dominant, and Balanced Bilingual. The results are competitive to previous work using a much larger and diverse set of features with shallow syntactic analysis. This paper shows the potential benefit of adding a deeper syntactic analysis for modeling language in young children, even in the case of having mixed language samples.

1 Introduction

In the field of communication disorders, the analysis of spontaneous language samples is a common practice to determine language status of children. Typically, this involves a very expensive process of manually coding and analyzing these samples to find patterns that are known to be good clinical markers. For the analysis of language from monolingual children, especially English-speaking children, there is a vast amount and breath of research that supports the use of these clinical markers. However, for bilingual populations the literature is not as extensive, although it is steadily growing. One task considered critical by clinical researchers when analyzing language from bilingual children is identification of language dominance. That is, in order to make final recommendations or diagnosis, it has been found to be critical to know which language, if any of the two, is more developed in the child. Recent research in communication disorders presents two approaches for determining language dominance in bilingual children, one based on measures of language exposure [1] and the other one based on measures of

language productivity [8], although the former seems to be more widely accepted. However, robust determination of language exposure is very expensive, as it requires parents and teachers to keep track of the amount of input and output of children over a period of time, typically a week.

Previous work by Solorio et al. from the Natural Language Processing (NLP) community has looked at a corpus driven approach for this problem of determining language dominance [12]. They framed this problem as a text classification task, where the classes are the three potential language dominance categories: English dominant (ED), Spanish dominant (SD), and balanced bilingual (BB), and they extracted a large variety of features from the language samples to train a machine learning classifier. In this paper we follow the idea of using a machine learning algorithm, but the set of features we explore here are purely syntactic, and were not explored in the work mentioned above. Our results show that deeper syntactic information carries rich relevant content for the task of determining the language dominance of Spanish-English bilingual children. We extract features from the parse trees generated by off the shelf syntactic parsers for English and Spanish, then we train a learning algorithm using a the set of syntactic rules founds in each transcripts as feature, we called it bag of rules (BOR). The accuracy results obtained by our simple syntactic based approach are higher than several of the features presented in previous work. We speculate that combining this information with that in Solorio et al.'s paper can lead to even higher accuracies.

2 Work Related

Previous work has used NLP techniques to help in the areas of communication disorders. In [3], in order to predict language impairment in monolingual English and Spanish-English bilingual children, they used six sets of features to build a computational model: language productivity, morphosyntactic skills, vocabulary knowledge, speech fluency, perplexities from LMs and standard scores. In this previous work the best result reported was around 60% of F-measure. In a more recent work, an addition of 3 sets of features to previous features was proposed. In particular, demographic information, syntactic complexity, and POS n-grams, were included to predict the dominant language in bilingual children [12]. This more recent work added some syntactic information as features but only at the level of part of speech tags. The best result obtained in this work was 72% of accuracy.

On the other hand, NLP techniques have also been explored in the detection of mild cognitive impairment [11], where features such as Yngve and Frazier scores, together with features derived from automated parse trees are explored in that work to model syntactic complexity. And similar features are used in classification of language samples as belonging to children suffering autism, language impairment, or none of the above [10].

The last two approaches inspired us to explore the use of information generated by automatically parsing the language samples. The features, as they are proposed here, have not been used in previous work. In this sense, the novelty of our study is the use of a representation analogous to bag of words that used syntactic patterns as extracted from parse trees. The next section describes our proposed method in more detail.

3 Proposed Approach

The goal of the task is the prediction of language dominance of a child into one of three core categories: BB (balanced bilingual), ED (English dominant), and SD (Spanish dominant). Since we want to streamline the process of language analysis as much as possible, we restrict the feature set to features that can be automatically extracted from the transcripts. Moreover, since previous work for automated language dominance prediction has not explored the use of parse trees, or features derived from parse trees, we study in this work their contribution to developing an accurate model for this task. We expect that children at similar stages of language acquisition will have mastered a similar set of grammatical constructions and that this can be exploited by a learning algorithm. An interesting twist in this classification task is the fact of having information, language samples, in each of the two languages. While it is widely accepted that in a bilingual population is important to assess language ability on both languages, it is less clear how to do this in a machine learning scenario. Here, we explore different ways to combine the observed samples in both languages.

The idea of this study is very simple. It consists of the following steps:

1. **Automatically parsing the transcripts.** In this step we generate a set of parse trees for each transcript using trained monolingual parsers. Because we lack gold standard parse trees of bilingual child language, we are assuming that a parser trained on mostly adult language will not have a major negative effect in our proposed solution. However, it should be noted here that the noise from the parse trees is not only coming from the differences between adult language constructs and those from children, but also from the mixed language input. As explained in the following section, children are prompted to elicit the language samples in one target language, but very frequently these children code switched between their two languages. Our assumption is that because this noise is systematic, the parser will make consistent decisions when unexpected tokens appear during analysis, it will not have a major effect on classification accuracy into language dominance. But we do recognize that if careful analysis will be performed on the parse trees, then adaptation of the parsers, to both child language, and mixed language input, might be needed.
2. **Finding rules**. Using every parse tree for a transcript, we find each rule of the form of $\alpha \rightarrow \beta$, where α is the root of a subtree and β is the set of children in that particular subtree. Because we are more interested in grammatical structure than in the actual vocabulary, we only add to the list those rules not involving a lexicon entry.
3. **Creating the representation of transcripts**. Once we gather the lexicon of grammar rules fired in the training set, we used them as features to represent each transcript. This representation is analogous to BOW (bag of words), but instead of words we have rules, thus we refer to this representation as BOR (bag of rules). We also use standard Boolean weights for the rules. The intuition is that it is enough to observe a syntactic construct once to assume the child masters that construction.
4. **Training a model for language dominance prediction.** Each transcript in the training set is transformed into a BOR vector. Then we use a standard machine learning algorithm to train a model. We assume then, that this problem of language

dominance prediction can be cast as a classification problem. Very similar to text classification, except that in this case a semantic classification is not the ultimate goal. But the general framework is the same.

5. **Classifying a child**. To classify the language dominance of a new child, we transform the transcript to a vector of n dimensions, where n is the number of elements in the BOR. Then we can use the trained model generated in the previous step to make a prediction for the new sample.

In following section we describe the data set used to evaluate our proposed representation.

4 Data

The data set used in this paper contains transcripts gathered as part of an on-going longitudinal study of language impairment in Spanish-English speaking children [9]. The children in this study were enrolled in kindergarten with a mean age of about 6 years and 1 month. A total of 180 children participated in this study, however, we only worked with 52 bilingual children since the data for the rest of the children was not available for analysis at this point. Table 1 shows the distribution of our data.

Category	Children
Balanced Bilingual (BB)	19
English Dominant (ED)	11
Spanish Dominant (SD)	22

Table 1. Distribution of our dataset into the three categories

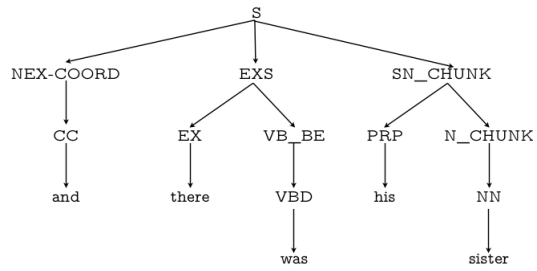
The transcripts were gathered following standard procedures for collection of spontaneous language samples in the field of communication disorders. For each child in the sample 4 transcripts of story narratives were collected, two in each language. Children are shown a wordless picture book and are asked to narrate the story behind the book. The story narratives are based on Mayer's wordless picture books. The books used for English were *A boy, A dog, and a frog* [4] and *Frog, where are you?* [6]. The books used for Spanish were *Frog on his own* [7] and *Frog goes to dinner* [5].

5 Experimental Setting

For extracting the parse trees we used FreeLing[6]. This parser comes with trained models for English and Spanish. The output of FreeLing is a set of parse trees. We break down the parse trees into grammar rules by traversing each tree in a breath first fashion. We only add rules to the BOR representation those composed of a root and its immediate

[6] FreeLing is available in the website: http://nlp.lsi.upc.edu/freeling

children. In Table 2 we show an example of a parse tree generate by FreeLing and the rules we extracted from it. Once we have the BORs we use them as features to represent the test transcripts. The value assigned to each rule in the vector is a boolean weight $w_{i,j}$, one if the rule i appears in the transcript j, and zero otherwise.



S → NEX-COORD EXS SN-CHUNK
NEX-COORD → CC
EXS → EX VB-BE
VB-BE → VBD
SN-CHUNK → PRP N-CHUNK
N-CHUNK → NN

Table 2. Parse tree generated by FreeLing for the sentence *and there was his sister* in one of the transcripts from our dataset and the rules we extracted from it.

As we mentioned in the previous section, we have 4 transcripts per child, but since our data set is small and we are using a corpus driven approach, we decided to duplicate the number of instances by separating the 4 sets of transcripts per child into 2 pairs. We realize that we are reducing by half how much information we observe per child to train our model and to test prediction accuracy. However in this case we believe it is more important to have more data samples to both train and evaluate. Moreover, clinicians and clinical researchers use one transcript per language for the most part, so this is also aligned with current practices. Despite this separation of transcripts per story, we were careful to put in the same partition (training or test) all transcripts of the same child. That way we avoid confounding the ultimate goal of the task.

To decide the language dominance of a particular child or instance we consider 2 transcripts, thus $I = \{T_1 \cup T_2\}$. Because we have 4 transcripts per child, we consider the following options for combining the transcripts:

- One in English and one in Spanish
- Both in the same language (English or Spanish)

These two combination are selected to answer one question: what is more helpful for analyzing language ability in bilingual children, using information from two languages, or more input in a single language? We already know the answer to this question from the point of view of communication disorders, and we speculate that in this case as

well the most beneficial scenario will be when using information from both languages. But it is interesting to explore if this pattern will hold when using a machine learning algorithm to predict language dominance.

To evaluate the performance of our method we used 5x2 cross fold validation, following recommendations in [2] for small sample sets. This means, we did 5 replications of 2-fold cross validation, in each repetition the available data were randomly partitioned into two equal-sized sets. In all our experiments we used the Weka [13] implementation of the machine learning algorithms.

6 Experimental Results

In our first experiment we wanted to determine whether by taking into account language samples only in one language is possible learn to distinguish between the three categories. However, to provide a fair comparison to that of using samples from each language, we took the two samples in the same language from each child. Thus we have two scenarios in this experiment: English-English and Spanish-Spanish. Table 3 shows the accuracy using five of the most common classification methods used in NLP problems: Naive Bayes, Support Vector Machines, C4.5, and k-Nearest Neighbors with $k = 1$ and $k = 5$.

	NB	**SVM**	**C4.5**	**1-NN**	**5-NN**
English	45.9	**49.62**	43.7	45.2	45.9
Spanish	**58.5**	55.6	48.1	44.4	45.9

Table 3. Accuracy of BOR representation over 5 classification methods: Naive Bayes (NB), Support Vector Machine (SVM), Decision tree C4.5 and k-Nearest Neighbors with $k = 1$ (1-NN) and $k = 5$ (5-NN). Using transcripts in one language: English or Spanish.

The results shown are rather poor, but are comparable to results reported in [12] on the same data set when using individual sets of features even though they are using information on both languages. Their reported accuracy ranges from 40%, when using only demographic information, to 72%, when using different metrics of syntactic complexity. However, direct comparisons are not possible since they used a leave one out cross validation setting.

Now we want to show that our hypothesis of combining information from both languages is better than looking only at one language. In this setting we used two transcripts per child, one for English and one for Spanish. Table 4 shows the results of this setting over the same 5 classification methods used in the previous experiment. The results improve accuracy by up to 10% in relation to the first experiment.

As we mentioned in related work, the closer work that predicted language dominance and used the same datasets of transcripts [12] shows an accuracy of 72%. However, they used 9 types of features measuring different dimensions of language combined with some demographic information, and the only type of syntactic information

	NB	SVM	C4.5	1-NN	5-NN
Using English and Spanish	63.3	**67.8**	49.3	55.6	57.0

Table 4. Accuracy of BOR representation over 5 classification methods: Nave Bayes (NB), Support Vector Machines (SVM), C4.5, and k-Nearest Neighbors with $k = 1$ (1-NN) and $k = 5$ (5-NN). Using transcripts in both languages: English and Spanish.

used in that work was at the level of POS n-grams. In this paper we used only the syntactic information extracted from parsing the transcripts in a BOR representation. While our results are a little bit below previous results, they are still relevant in that they show how this syntactic information is valuable, and can outperform other feature types from previous work, including speech fluency measures, language productivity measures, demographic information, morphosyntactic features, speaking rate, and n-grams of POS. We believe that combining this BOR representation with those features used in [12] can boost accuracy further.

7 Conclusions and Future Work

We proposed a representation based on bag of rules from parse trees for the problem of predicting language dominance in Spanish-English children. Our results show that combining information from transcripts in both languages yields the best results. This study also shows that syntactic information is important for language analysis, even though there could be a considerable amount of noise in the parse trees from having mixed language, as well as child language.

The results obtained are comparable to the recent work looking at the same problem, but different from them we only look at one dimension of language. We only extract features derived from syntactic trees, while previous work looks at vocabulary, language production, fluency, and measures of readability, among others. We predict that adding this dimension to previous work will help achieve higher prediction accuracy.

As future work we want to explore other syntactic information that can also be extracted from the parse trees to build a more robust language model that can improve the results achieved so far. Adding other features we would be able to reduce the impact of noise produced by the code switching. Other things we are working on include the use of different weighting schemes for the rules, such as TF-IDF, and entropy of the grammar rules.

Acknowledgments

This research was partially supported by the National Science Foundation under grants 1018124 and 1017190, and by grant R01DC007439 from the National Institute on Deafness and Other Communication Disorders (NIDCD). This work was also supported in part by the UPV, award 1932, under the program Research Visits for Renowned Scientists (PAID-02-11) and by the European Commission as part of the WIQ-EI project (project no. 269180) within the FP7 People Programme.

References

1. Bedore, L.M., Peña, D., E., Gillam, B., R., Ho, T.: Language sample measures and language ability in spanish english bilingual kindergarteners. Journal of Communication Disorders 43(6), 498–510 (Nov-Dec 2010)
2. Dietterich, T.G.: Approximate statistical tests for comparing supervised classification learning algorithms. Neural Computation 10(7), 1895–1924 (1998)
3. Gabani, K., Sherman, M., Solorio, T., Liu, Y., Bedore, L.M., Peña, E.D.: A corpus-based approach for the prediction of language impairment in monolingual English and Spanish-English bilingual children. In: North American Chapter of the Association for Computational Linguistics - Human Language Technologies (NAACL-HLT) 2009. pp. 46–55. Association for Computational Linguistics, Boulder, Colorado (June (2009))
4. Mayer, M.: A boy, a dog, and a frog. Dial Press, New York, NY ((1967))
5. Mayer, M.: Frog goes to dinner. Dial Press, New York, NY ((1969))
6. Mayer, M.: Frog, where are you? Dial Press, New York, NY ((1969))
7. Mayer, M.: Frog on his own. Dial Press, New York, NY ((1973))
8. Paradis, J., Crago, M., Genesee, F., Rice, M.: French-english bilingual children with SLI: How do they compare with their monolingual peers? Journal of Speech, Language, and Hearing Research 46, 113–127 (2003)
9. Peña, E.D., Bedore, L.M., Gillam, R.B., Bohman, T.: Diagnostic markers of language impairment in bilingual children. Grant awarded by the NIDCD, NIH (2006)
10. Prud'hommeaux, E.T., Roark, B., Black, L.M., van Santen, J.: Classification of atypical language in autism. In: Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics. pp. 88–96. Association for Computational Linguistics, Portland, Oregon, USA (June 2011)
11. Roark, B., Mitchell, M., Hollingshead, K.: Syntactic complexity measures for detecting mild cognitive impairment. In: BioNLP 2007: Biological, translational, and clinical language processing. pp. 1–8. ACL, Prague (June 2007)
12. Solorio, T., Sherman, M., Liu, Y., Bedore, L., Peña, E., Iglesias, A.: Analyzing language samples of Spanish-English bilingual children for the automated prediction of language dominance. Natural Language Engineering 17, 367–395 (2011)
13. Witten, I.H., Frank, E.: Data Mining, Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufmann, San Francisco, CA ((1999))