"Eve Eat Dust Mop" Measuring Syntactic Development in Child Language with Natural Language Processing and Machine Learning

Shannon Lubetich
Pomona College
Claremont, CA 91711
1 (360) 434-7388
shannonlubetich@gmail.com

1. INTRODUCTION

Children start understanding language before they can speak it, and then speak it without exhibiting conscious effort to dissect syntactic and morphological structures. Although this process has been the focus of much study, our understanding of first language acquisition is still limited. In attempts to measure child language development over time, several metrics have been proposed. The most commonly used metrics are quick and easy to compute, but focus on superficial aspects of language, such as length of utterance. These metrics often fail to account for the fact that, at a certain age, a child's language can become grammatically more complex without increasing in length. Several metrics based on the usage of grammatical structure have been proposed as being more sensitive to changes in language over a wider range of ages [12, 5, 3]. Using these metrics for computation of language development scores involves identification of several specific grammatical structures in child language transcripts, a process that requires linguistic expertise and is both time-consuming and error-prone. One such metric is the Index of Productive Syntax, or IPSyn, [12] an empirically validated metric based on an inventory of grammatical structures derived from child language literature.

In our research, we sought to create a fully automated version of IPSyn to allow for the quick, easy, and accurate analysis of grammatical complexity for a large number of transcripts. Additionally, we wanted to determine the possibility of measuring child language development in a fully data-driven manner based solely on the extraction of morphological and syntactic features from a transcript. We further examine a data-driven approach that can be used in absence of a specifically defined inventory like IPSyn. After demonstrating the accuracy of this approach for English-speaking children and their transcripts, we extend the research to other languages to examine if such abstract feature templates can be used to measure child language development cross-linguistically.

2. BACKGROUND AND RELATED WORK

This work is motivated by the evaluation of previously existing metrics of child language development. For example, the most commonly used metric is Mean Length of Utterance, or MLU [1], which is based on the number of morphemes per utterance. The main appeal of MLU is that it can be easily computed automatically, given machine-readable transcripts. However, MLU's ability to track language development from age four has been questioned [4, 12], and its usefulness is still the subject of debate [8].

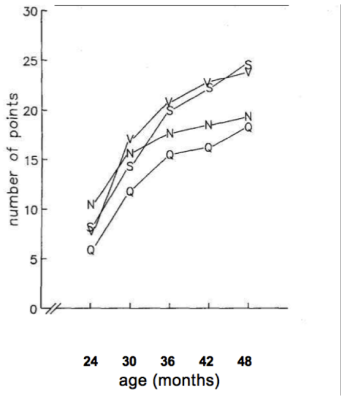
We focus on creating a program to automatically measure syntactic development based on the IPSyn scale, as presented by Hollis S. Scarborough [12]. IPSyn works by awarding points based on the presence of increasingly complex grammatical structures from a predefined inventory derived from child language literature. Previous work [6] has shown that this process can be automated using current natural language processing techniques and a carefully crafted set of patterns that can be matched to the grammatical structures in the IPSyn inventory.

After creating such a program, we were able to generate a large set of labeled data to use in a data-driven approach to this problem. Data-driven approaches have been attempted for analyzing syntactic complexity of child language [11], but these have been based on extracting language specific features, whereas our approach opts instead to be language independent. We first use a data-driven approach to predict IPSyn scores, but then move on to attempting to predict the age of a child based on features from their transcript; this framework can be used to track language development in the absence of a metric such as IPSyn. This approach is then used to explore the possibility of measuring child language development in languages other than English.

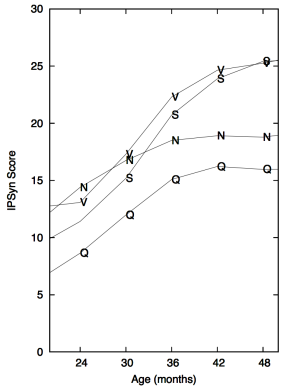
3. APPROACH AND UNIQUENESS

3.1 Automating IPSyn: Approach

To solve this problem, we first focused on writing a program to mimic the performance of a trained language researcher in analyzing a child language transcript to produce an IPSyn score. Our approach was to define the predetermined structures that are awarded points in a separate text file that is read in by the program and translated into matchable patterns, which are then searched for within individual transcripts. If these structures are encountered, the score of grammatical complexity increases until a whole transcript has been analyzed for all possible structures defined by IPSyn.



(a) Original IPSyn study.



(b) Automatically generated.

Figure 1: Comparison between the IPSyn development curves for the four subscales in (a) the 75 transcripts in the original IPSyn study (reproduced from (Scarborough, 1990)), and (b) our set of 593 transcripts scored automatically.

3.2 Automating IPSyn: Evaluation

To evaluate our approach, we ran this program on 593 transcripts, and extracted scores for the individual subsections defined by IPSyn. These subsections are categories of syntactic structures, and consist of Noun Phrases, Verb Phrases, Questions and Negations, and Sentence Structures. We then compared our results of how these scores change over time, and thus how a child's language changes over time, to the original analysis of just 75 transcripts done by Scarborough [12] and found extreme similarities, as can be seen in Figure 1.

Additionally, we compared the performance of our automatic scorer on 17 transcripts that had been manually scored for IPSyn. We found a mean absolute difference of 5.8 points on the 100 point IPSyn scale. The point-to-point difference between any two human scorers is around 5 points, so our program performs to the level of accuracy of trained child language researchers.

3.3 Data-driven Approach: Predicting IPSyn

We attempted to produce scores of grammatical complexity without searching for and awarding points to specific syntactic structures, but instead extracting morphological and syntactic features of transcripts. Figure 2 shows an example sentence and its corresponding features. The arrows define grammatical relations between words, and the words are individually marked with their parts-of-speech. Essentially, we extracted a combination of these features that correspond to components within the parse tree of an utterance, and then trained a classifier on these extracted features and their corresponding scores, using the SVM Light[1] implementation of support vector regression [2]. The features we found worked best for this task were as follows: part-of-speech tags, grammatical relations, combined part-of-speech tags of a word in the parse tree and its dependent, and combined part-of-speech tags of a word, its dependent, and the grammatical relation between them.

3.4 Data-driven Evaluation: Predicting IPSyn

We calculated the accuracy of our learned regression model by comparing our generated scores to manually-computed scores of 17 transcripts. We obtained a mean absolute error of 6.7 points on the 100 point IPSyn scale. We additionally used 10-fold cross validation on our set of 593 transcripts, and found a mean absolute error of the score predictions obtained using regression of 6.8 points. (See Figure 3.)

[1] http://svmlight.joachims.org/

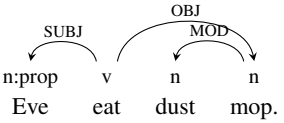


Figure 2: A dependency tree generated with part-of-speech and grammatical relation information.

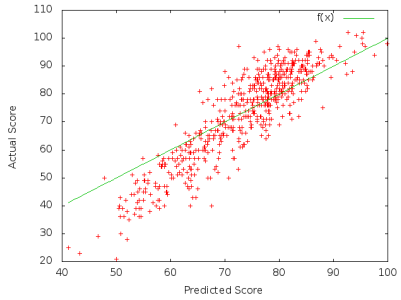


Figure 3: Our classifier's predicted score vs. previously determined IPSyn score. The plotted line depicts where all the points would fall if predicted matched actual.

3.5 Data-driven Approach: Predicting Age

To determine whether our data-driven regression approach can model the development of individual children at the level where accurate age predictions can be made, we used the same feature templates described in Section 3.3, but trained a regression model to predict age in months, rather than IPSyn scores. Because this is a child-specific prediction task, we train separate regression models for each child.

3.6 Data-driven Evaluation: Predicting Age

We tested our age predictions using 10-fold cross-validation for three children from three different CHILDES corpora (Adam from

Brown, Ross from MacWhinney and Naomi from Sachs) [6] for whom enough data was available over a wide enough range of ages. In each case the regression approach performed well. Table 1 shows the mean absolute error in months for each child, and the Pearson r for the correlation between predicted age and actual age.

Child (corpus)	Mean Abs Err	Pearson (r)
Adam (Brown)	2.5	0.93
Ross (MacWhinney)	3.7	0.84
Naomi (Sachs)	3.1	0.91

Table 1: Regression results for single corpus age prediction ($p < 0.0001$ for all r values.)

Perhaps more interesting than the strong correlations between actual age and predicted age for each of the individual corpora is a comparison of these correlations to correlations between age and MLU, and age and IPSyn score. One main general criticism of MLU is that it fails to correlate well with age for older children (around three to four years old). More detailed metrics such as IPSyn are believed to have better correlation with age after that point. We do observe this situation in our data. Interestingly, our predicted age scores have much stronger correlations to actual age for older children, which suggests that our regression approach with simple syntactic features is more expressive in tracking syntactic development in older children than either MLU or IPSyn. This is shown in Table 2, which contains Pearson r correlation coefficients for age and MLU, age and IPSyn, and age and predicted age using our regression approach.

Child (corpus)	MLU r	IPSyn r	Regression r
Adam (Brown)	$0.37^{\dagger}$	$0.53^{\dagger}$	$0.85^{\dagger}$
Ross (MacW)	0.19	0.34^{*}	$0.79^{\dagger}$
Naomi (Sachs)	0.27	0.52	$0.82^{\dagger}$

Table 2: Pearson correlation coefficients between actual age and MLU, actual age and IPSyn score, and actual age and predicted age, for children at least three years and four months old. $^{\dagger}p < 0.0001$. $^{*}p < 0.05$.

The results shown in Table 2 confirm that features extracted from parse trees alone can offer substantially better prediction of age for individual children than MLU or even IPSyn scores. This is not surprising, given that weights for these features are optimized to predict age using data from the specific child and discriminative learning, but it does show that these features offer enough resolution to track syntactic development in child language.

3.7 Cross-linguistic Approach: Predicting Age

As discussed in Section 3.3, we extracted features from the parts-of-speech and dependency parse trees of an utterance. The programs used in this preprocessing are available for English, but also for Japanese, Spanish, and Hebrew [7], [9]. Thus, our data-driven approach outlined in Section 3.4 can be applied to transcripts of children speaking these languages. We again turned to CHILDES corpora [6], including transcripts from two children from each of the languages; Japanese (Ryo and Ishii from Miyata), Spanish (Irene from Llinàs-Grau and Emilio from Vilo), and Hebrew (Hagar and Leor from Berman longitudinal). These children were chosen due to the availability of a number of transcripts across a wide range of ages, since they all came from longitudinal, individually-focused studies of children. We again approached the regression task as child-specific, training and predicting using files from one child. Since we were unsure the same feature patterns would result in the best performance on the prediction task in this approach, our main goal was to determine if any sort of simple syntactic feature templates would result in adequate prediction accuracy. Essentially, we tested a number of combinations of simple syntactic features, comparing their performance on the prediction task to a baseline of using only bag-of-words as our regression features. Bag-of-words simply selects the exact words in the transcript, which is not language independent, but is understood to correlate with age as children use more complex words as they get older.

3.8 Cross-linguistic Evaluation: Predicting Age

We tested our age predictions using leave-one-out cross-validation, due to the limited amount of data, for each child in each language discussed above. In evaluating our models, we averaged the performance of the models for the two children in a single language, so that we could have some idea of performance in a single language. Though bag-of-words performed in the top set of feature templates for each language, there were a number of feature templates that did not perform at a significantly different level as using bag-of-words. Table 3 shows the mean absolute error in months from predicted age and actual age, and the Pearson r for the correlation between predicted age and actual age (both averaged over the two children for that language).

Language	Mean Abs Err	Pearson (r)
Japanese	2.18	0.85
Hebrew	2.42	0.76
Spanish	4.2	0.81

Table 3: Regression results for single language age prediction, using a simple syntactic feature template that does not perform worse than the bag-of-words approach ($p < 0.001$).

The results shown in Table 3 confirm that, in languages other than English, features extracted from parse trees alone can offer a comparable prediction of age as content-based, language-specific features (in this case, bag-of-words). The slightly weaker performance than in English could be due to smaller available data sets, and more research is needed to fine tune parameters in training the model and selecting feature combinations.

4. RESULTS AND CONTRIBUTIONS

4.1 Results

As demonstrated above in Section 3.2, we were successful in creating a program to automatically analyze a transcript for grammatical complexity using the IPSyn scale with levels of accuracy similar to that of trained manual scorers. Section 3.4 also showed the possibility of a data-driven approach to this problem, using machine learning to train a regression model on morphological and syntactic features of the utterances within a transcript as corresponding with scores of grammatical complexity. Furthermore, Section 3.6 and 3.8 applied this data-driven approach to avoid predicting a metric of grammatical complexity, and instead choosing to predict age of a child based on simple syntactic features of a transcript. Section 3.6 showed the results of high correlation of predicted age and actual age for English-speaking children, demonstrating that this approach out performs correlation between MLU and age, or IPSyn and age, as children get older. Section 3.7 explored the possibility of this age prediction task for children speaking languages

other than English, finding that simple syntactic feature templates can perform at similar levels as more language-specific metrics.

4.2 Contributions

The contributions of this work are:

1. We created a fully automatic implementation of the IPSyn scoring system that could be used by child language researchers to quickly and accurately analyze the syntactic complexity of a transcript of child language. This can be used to generate IPSyn scores for numerous children, create a baseline for what score is expected at a given age, and then allow for the analysis and comparison of a new child. Determining the level of language development of this new child in relation to averages could help identify language-learning disorders or differences in language learning due to a number of factors such as socioeconomic status, bilingualism, and more.

2. In our data-driven approach, we presented a framework for an assessment of syntactic development in child language that is completely data-driven and just relies on morphological and syntactic features. Additionally, this framework is completely language independent, so can be applied to any number of languages as long as there is information available from syntactic analysis. This allows for the prediction of score of grammatical complexity, as well as other information such as age of child, based solely on features extracted from syntactic analysis.

5. REFERENCES

[1] R. Brown. *A first language: The early stages*. George Allen & Unwin, 1973.

[2] H. Drucker, B. L.Burges, A. Smola, and V. Vapnik. Support vector regression machines. *Advances in Neural Information Processing Systems 9*, 9:155–161, 1997.

[3] P. Fletcher and M. Garman. LARSPing by numbers. *British Journal of Disorders of Communication*, 23(3):309–321, 1988.

[4] T. Klee and M. D. Fitzgerald. The relation between grammatical development and mean length of utterance in morphemes. *Journal of Child Language*, 12:251–269, 6 1985.

[5] L. L. Lee and S. M. Canter. Developmental sentence scoring: A clinical procedure for estimating syntactic development in childrenÍs spontaneous speech. *Journal of Speech and Hearing Disorders*, 36(3):315–340, 1971.

[6] B. MacWhinney. *The CHILDES Project: Tools for Analyzing Talk*, 3rd edition. Lawrence Erlbaum Associates, 2000.

[7] S. Miyata, K. Sagae,and B. MacWhinney. The Syntax Parser GRASP for CHILDES (in Japanese). *Journal of Health and Medical Science*, 3:45–62, 2013.

[8] M. L. Rice, F. Smolik, D. Perpich, T. Thompson, N. Rytting, and M. Blossom. Mean length of utterance levels in 6-month intervals for children 3 to 9 years with and without language impairments. *Journal of Speech, Language, and Hearing Research*, 53:1–17, 2010.

[9] K. Sagae, E. Davis, A. Lavie, B. MacWhinney and S. Wintner. Morphosyntactic annotation of CHILDES transcripts. *Journal of Child Language*, 37:705–729, 2010.

[10] K. Sagae, A. Lavie, and B. MacWhinney. Automatic measurement of syntactic development in child language. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 197–204, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.

[11] S. Sahakian and B. Snyder. Automatically learning measures of child language development. In *ACL (2)*, pages 95–99. The Association for Computer Linguistics, 2012.

[12] H. S. Scarborough. Index of Productive Syntax. *Applied Psycholinguistics*, 11(Peer Reviewed Journal):1–22+, 1990.