Data-driven Measurement of Child Language Development with Simple Syntactic Templates

Shannon Lubetich
Pomona College
Claremont, CA 91711
shannon.lubetich@pomona.edu

Kenji Sagae
Institute for Creative Technologies
University of Southern California
Los Angeles, CA 90089
sagae@ict.usc.edu

Abstract

When assessing child language development, researchers have traditionally had to choose between easily computable metrics focused on superficial aspects of language, and more expressive metrics that are carefully designed to cover specific syntactic structures and require substantial and tedious labor. Recent work has shown that existing expressive metrics for child language development can be automated and produce accurate results. We go a step further and propose that measurement of syntactic development can be performed automatically in a completely data-driven way without the need for definition of language-specific inventories of grammatical structures. As a crucial step in that direction, we show that four simple feature templates are as expressive of language development as a carefully crafted standard inventory of grammatical structures that is commonly used and has been validated empirically.

1 Introduction

Although child language 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 metric is Mean Length of Utterance, or MLU (Brown, 1973), 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. Although MLU values may not be meaningful across languages, the general approach is suitable for analysis within different languages. However, MLU's ability to track language development from age four has been questioned (Klee and Fitzgerald, 1985; Scarborough, 1990), and its usefulness is still the subject of debate (Rice et al., 2010).

Several metrics based on the usage of grammatical structure have been proposed as more sensitive to changes in language over a wider range of ages (Scarborough, 1990; Lee and Canter, 1971; Fletcher and Garman, 1988). These metrics continue to show score increases where MLU plateaus, but their increased expressivity is typically associated with two severe drawbacks. The first is that their use 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. This issue has been addressed by recent work that shows that current natural language processing techniques can be applied to automate the computation of these metrics, removing the bottleneck of manual labor (Sagae et al., 2005; Roark et al., 2007; Sahakian and Snyder, 2012). The second drawback is that these measures are language-specific, and development of a measure for a specific language requires deep expertise and careful design of an inventory of grammatical structures that researchers believe to be indicative of language development. Going beyond previous work, which addressed the first drawback of traditional metrics for child language development, we address the second, paving the way for a language-independent methodology for tracking child language development that is as expressive as current language-specific alternatives, but without the need for carefully constructed inventories of grammatical structures.

Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, pages 2151–2160, Dublin, Ireland, August 23-29 2014.

The specific hypothesis we address in this paper is whether a fully-data driven approach that uses only a few simple feature templates applied to syntactic dependency trees can capture the same information as the well-known Index of Productive Syntax, or IPSyn (Scarborough, 1990). In contrast to previous work that showed that the computation of IPSyn scores can be performed automatically by encoding each of the 60 language structures in a language-specific inventory (e.g. wh-questions with auxiliary inversion, propositional complements, conjoined sentences) as complex patterns over parse trees, we propose that child language development can instead be measured automatically in a way that is fully data-driven and can be applied to many languages for which accurate dependency parsers are available, without relying on carefully constructed lists of grammatical structures or complex syntactic patterns in each language. Specifically, we examine two hypotheses: (1) counts of features extracted from syntactic parse trees using only simple templates are at least as expressive of changes in language development as the Index of Productive Syntax (Scarborough, 1990), an empirically validated metric based on an inventory of grammatical structures derived from the child language literature; and (2) these parse tree features can be used to model language development without the use of an inventory of specific structures, assuming only the knowledge that in typically developing children the level of language development is correlated with age. We emphasize that the goal of this work is not to develop yet one more way to compute IPSyn scores automatically, but to show empirically that lists of grammatical structures such as those used to compute IPSyn are not essential to measure syntactic development in children.

In this paper, we start by reviewing IPSyn and previous work on automatic IPSyn scoring based on manually crafted syntactic patterns in section 2. Using a similar approach, we validate the language development curves observed by Scarborough (1990) in the original IPSyn study. In section 3 we show how IPSyn scores can be computed in an entirely different, fully data-driven way, using a support vector regression. In section 4 we examine how this data-driven framework can be used to track language development in the absence of a metric such as IPSyn, which allows for application of this approach to languages other than English. We discuss related work in section 5, and conclude in section 6.

2 Index of Productive Syntax (IPSyn)

The Index of Productive Syntax (Scarborough, 1990) evaluates a child's linguistic development by analyzing a transcript of utterances and awarding points when certain syntactic and morphological structures are encountered. The end result is a number score ranging from 0 to 120, with a higher score corresponding to the presence of more complex grammatical structures, and thus further linguistic development. IPSyn was designed to be more sensitive to language changes after age 3 than the more common Mean Length of Utterance (MLU) (Brown, 1973), which fails to account for the fact that children's speech increases in complexity even after utterances stop increasing in length.

IPSyn scores are calculated by analyzing a transcript of 100 utterances of a child's speech, and awarding points to specific language structures encountered. There are 60 forms in total from four categories of noun phrases, verb phrases, questions and negations, and sentence structures. Each form is awarded 0 points if not encountered, 1 point if found once in a transcript, and 2 points if found at least twice. This sums to a total ranging between 0 and 120 points. Scarborough (1990) motivates the use of this specific inventory of 60 forms by stating that they "have been shown to occur in preschool language production in innumerable studies of language acquisition during the past 25 years," highlighting that the task of generating such an inventory and performing empirical validation for additional languages requires considerable expertise and is far from trivial.

2.1 Automating IPSyn

In support of empirical testing of our first hypothesis–that features extracted from parse trees using only simple feature templates are as expressive of child language development as the carefully constructed inventory of grammatical structures in IPSyn–we first implemented an automated version of IPSyn following Sagae et al. (2005), who showed that this task can be performed nearly at the level of trained human experts. This allows us to generate IPSyn scores for a large set of child language transcripts. Our implementation differs from previous work mainly in that it uses only the tools provided in the CLAN

software suite (MacWhinney, 2000), which were designed specifically for analysis of child language transcripts, instead of the Charniak (2000) parser, which was used by Sagae et al. and later by Hassanali et al. (2014) in a more recent implementation of the same general approach.

We evaluated our implementation using the set of 20 manually scored transcripts described by Sagae et al. as Set A, and subsequently used to evaluate the implementation of Hassanali et al. Three transcripts were used as development data, following Sagae et al. The mean absolute difference between manually generated and automatically generated scores was 3.6, which is very similar to what has been reported by Hassanali et al. and by Sagae et al. (3.05 and 3.7, respectively) for the same set of transcripts. Given the possible score differences in manual scoring reported by Scarborough (1990) and the small number of transcripts used for testing, the differences observed among the automatic systems are not meaningful. In fact, in examining our development data, we found multiple errors in the manual coding, causing point discrepancies when our system produced correct results. This highlights the difficulty of performing this scoring task manually, and raises the question of whether automatic scoring has in fact surpassed the reliability of manual scoring. That three different implementations of IPSyn appear to perform comparably suggests this might be the case. We leave an empirical investigation of this question to future work.

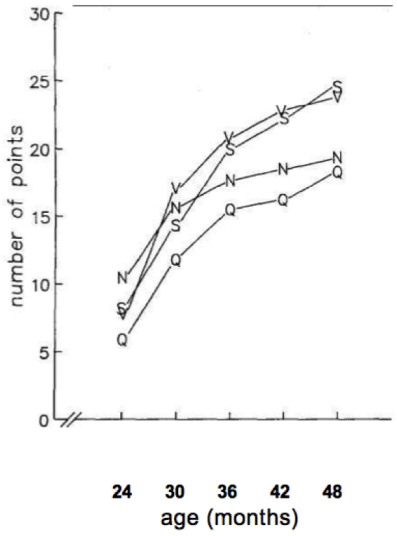
3 From automatic IPSyn to data-driven IPSyn

The fully automatic way of computing IPSyn scores described above in section 2.1, paired with a sufficiently large amount of child language transcript data, gives us a way to test the hypothesis mentioned in the beginning of section 2.1, that simple features of parse trees are as expressive as the hand-crafted IPSyn language structure inventory. We did this by first creating several 100-utterance transcripts from existing child language transcripts, then automatically assigning them IPSyn scores, and using these scores as targets to be learned from features extracted from the corresponding 100-utterance transcripts. Details of the data and learning approach used for this experiment, as well as empirical results, are described in the remainder of this section.

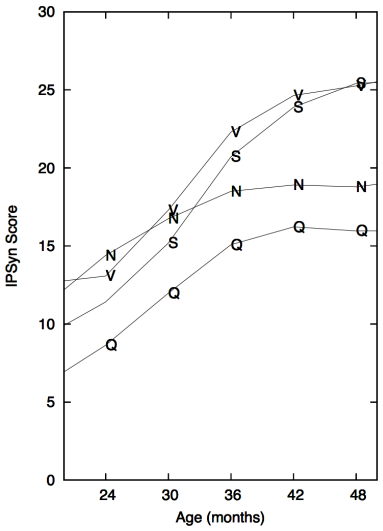
3.1 Generating IPSyn data

To obtain enough child language transcripts in a wide range of ages to test our hypothesis, we turned to the CHILDES database. To generate training and development sets for our experiments, we used transcripts from CHILDES of 14 different children with ages ranging from 1 year 1 month to 8 years. Because each application of IPSyn requires only 100 child utterances, transcripts were split, producing a total of 593 transcripts, each containing 100 utterances. The 14 children in our dataset came from the following CHILDES corpora: Brown, MacWhinney, Sachs and Warren. The reason for choosing these corpora is that they were quickly identified as containing spontaneous natural interactions, as opposed to reading or specific games and activities designed to elicit a certain kind of language production. It is likely that other corpora in CHILDES would also suit our purposes, but the data in these four corpora was sufficient for our experiments. Each of the 593 transcripts was assigned an IPSyn score automatically. From the Brown, MacWhinney and Sachs corpora, we used transcripts from a total of four children (Adam from Brown, Mark and Ross from MacWhinney, and Naomi from Sachs), from whom language data was collected over several years. Transcripts from these three corpora, 572 in total, served as our training set. The Warren corpus includes data from ten children with ages ranging from 1;6 to 6;2 (that is, 1 year and 6 months to 6 years and 2 months, using the commonly accepted age notation for this type of data), from which we created 21 transcripts that served as our development set.

The complete set of 593 transcripts with IPSyn scores gives us the opportunity to verify whether the language development curves observed by Scarborough (1990) averaged over 75 transcripts in the original IPSyn study matches curves produced from averaging results from 593 transcripts from entirely different subjects. Figure 1 shows a side-by-side comparison between the original figure from (Scarborough, 1990) and a corresponding figure generated with our automatically scored transcripts. Although not identical, the two figures are remarkably similar, reflecting that aspects of the emergence of grammar in child language development are shared across children, and that IPSyn captures some of these aspects.



(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.

Finally, we used the Garvey corpus to generate a test set. This corpus includes data from 48 different children with ages ranging from 2;10 to 5;7, from which we extracted 60 transcripts covering all 48 children and the full range of ages in the corpus. No data from the 48 children in the Garvey corpus, which we used as a test set, were used for training or development of the models used in our experiments.

3.2 A regression model for IPSyn

Given 593 pairs of transcript and IPSyn score, we approached the task of learning a data-driven model for IPSyn scoring as one of regression. For each transcript, a set of features is extracted, and the IPSyn score is associated with that feature vector. The features extracted from the transcripts followed four templates, described in the next subsection. If an accurate function for predicting IPSyn scores from these feature vectors can be learned, our hypothesis that these features are at least expressive enough to track child language development as well as the inventory of IPSyn structures is confirmed. To learn our model, we used the SVM Light[1] implementation of support vector regression (Drucker et al., 1997).

3.3 Features

An important step in learning a regression model for IPSyn is choosing what features to use. To support our goal of language independence, we decided not to consider language specific features that have been shown to be useful in this task but are language dependent[2], and opted instead to see whether the use of only simple parse tree features would be sufficient. The only prerequisite for extraction of our feature set is that each transcript must be parsed to produce a syntactic dependency tree. We used the CLAN tools for morphology analysis (MOR), part-of-speech tagging (POST) and parsing (MEGRASP)[3], since it is straightforward to process CHILDES transcripts using those, and they provide high-accuracy analyses for child language transcripts. The accuracy of the MEGRASP dependency parser for child utterances in English is estimated to be close to 93% (Sagae et al., 2010).

All of the features used in our model are extracted from parse trees according to four simple classes that target the following information:

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

[2] This is in contrast to, for example, the related work of Sahakian and Snyder (2012), which we discuss in section 5.

[3] Models for MOR and POST are available for a wide variety of languages. Models for MEGRASP are available only for English and Japanese, but our data-driven approach is not tied to any specific tagger or parser.

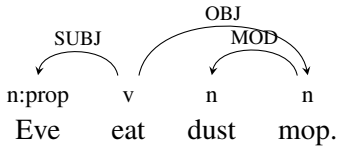


Figure 2: A dependency tree generated with part-of-speech and grammatical relation information.

Part-of-speech tags: The first type of feature we used is simply the part-of-speech tag of each word. This can be thought of as a bag of part-of-speech tags. We intentionally avoided the commonly used bag-of-words, because our goal is to obtain a model that tracks changes in syntax structure, not content. Although it is highly likely that lexical features would be very informative in this learning task, they would be useful for the wrong reason: our model is intended to target the emergence of *syntax*, and not what children talk *about* at different ages. We note, however, that as with the Penn Treebank tagset, the tags used by MOR and POST also reflect morphology, so that information is accounted for. The full tag set is listed in (MacWhinney, 2000).

Grammatical relations: The second feature class in our model is a bag of dependency labels, where each label correspond to a grammatical relation that holds between two words in the dependency tree (the head word and the dependent word). The full set of grammatical relations is listed in (Sagae et al., 2010).

Head-dependent part-of-speech pairs: Our third feature class is based on pairs of part-of-speech tags, where each pair corresponds to a bilexical dependency relation in the parse tree, and one of the tags comes from the head in the dependency, and the other tag comes from the dependent.

Head-relation-dependent triples: The last feature class is similar to the head-dependent pairs described above, but also including the dependency label that indicates the grammatical relation that holds between the head and dependent words. Features in this class are then triples composed of a head part-of-speech tag, a dependent part-of-speech tag, and a dependency label.

As an example, given the parse tree shown in Figure 2, the following features would be extracted:

```
n:prop v n n
SUBJ OBJ MOD
v_n:prop v_n n_n
v_n:prop_SUBJ v_n_OBJ n_n_MOD
```

Features are extracted for every tree in each transcript. Because our goal is to measure grammatical development in child language, these four feature templates were designed to capture the grammatical relations represented in dependency trees, while leaving out the content reflected in specific lexical items. While the content of what is said may be related to language development, our features are intended to focus on syntactic information, covering exactly each of the labeled arcs and the part-of-speech tags in a dependency tree (Figure 2) with the words removed. We also experimented with part-of-speech tag bigrams (pairs of adjacent part-of-speech tags), and dependency chains formed by two dependency arcs. The final choice of the four templates described above was based on results obtained on development data.

3.4 Data-driven IPSyn evaluation

We trained a support vector regression model using our training set of 572 transcripts, using a polynomial kernel and tuning the degree d and the regularization metaparameter C on the development set. While the default C and d values resulted in a mean absolute error of 6.6 points in the score predictions in the development set, setting $C = 1$ and $d = 3$ resulted in a mean absolute error of 4.1 on the development set. We used these values for the rest of our experiments. The mean absolute error obtained on our

test set of 48 children (60 transcripts) not used in training or tuning of the system was 3.9. When applying our regression model to the manually scored set of 20 transcripts used by Sagae et al. (2005), the mean absolute difference was 4.2 from the scores computed automatically using the approach in section 2.1, and 5.4 from the manually computed scores, which we consider our gold standard target. Compared to these manually computed scores, the absolute difference of 5.4 is higher than what we obtained using carefully designed templates based on the IPSyn inventory, but still within the range of variability expected for trained human scorers (Scarborough, 1990). It is important to keep in mind that the goal of this experiment was not to improve on the accuracy of previous automatic scoring programs, which work quite differently by listing manually crafted patterns over parse trees, but to show that a scoring function can be learned in a data-driven way, without manually crafted patterns. The results obtained with our regression model do confirm our hypothesis that simple features extracted from parse trees are enough for tracking child language development in the same way as the much more complex patterns included in IPSyn.

4 Age prediction

Given the ability of our data-driven approach to approximate IPSyn scores, confirming that a regression approach with parse tree features is capable of capturing the progression of language development, we now turn to the question of whether the same type of data-driven framework can be used to track child language development without the need for a metric such as IPSyn.

Assuming only that language acquisition progresses monotonically over time, we can apply the same data-driven regression approach to predict a child's age given a language sample. This task was approached recently by Sahakian and Snyder (2012), who used an ensemble of existing metrics with a few additional features. Unlike in our approach, Sahakian and Snyder do include lexical features and hand-selected patterns in the form of an existing metric (D-level). They make the reasonable argument that the task of age prediction is child-dependent, and that prediction across children would not make sense due to individual variation in the rate of language development. Following Sahakian and Snyder, we first approach age prediction as a child-specific task, but then discuss the application of our regression models for other children than those used for training.

4.1 Child-specific age prediction

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. 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) 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, confirming our second hypothesis.

4.2 Pilot experiment with Japanese language data

A great advantage of using a data-driven framework based on simple feature templates rather than a traditional approach for measuring syntactic development with manually crafted lists of grammatical structures is that the data-driven approach is, in principle, language-independent. The same features described in section 2.1 could be extracted from dependency parse trees in any language, assuming only that these dependency trees can be produced automatically. Syntactic dependency parsers and treebanks are in fact available for a variety of languages (Buchholz and Marsi, 2006; Nivre et al., 2007). Although the availability of treebanks that include child language samples is certainly desirable, it is not clear whether it is strictly required in order to generate the syntactic structures used in our approach. While Sagae et al. (2005) and Hassanali et al. (2014) obtained high levels of accuracy in IPSyn scoring using the Charniak (2000) parser with a model trained on the Wall Street Journal portion of the Penn Treebank (Marcus et al., 1993), we have not verified the effects of parser errors in our data-driven approach. Of course, the language independence claim applies only to the ability to measure syntactic development within different languages, and direct numerical comparisons across languages are not meaningful, since the available syntactic annotations for different languages follow different conventions and syntactic theories.

Although a full empirical validation of our regression approach in other languages is left as future work, we performed a pilot experiment with a single Japanese child that suggests our findings may be robust across languages. We used transcripts from the child Ryo, from the Miyata corpus of the Japanese section of the CHILDES database[4]. We extracted 80 transcripts of 100 utterances each, covering ages 1;10 (22 months) to 3;0 (36 months). These transcripts were analyzed with the Japanese version of the MEGRASP parser for CHILDES transcripts at an estimated accuracy of 93% (Miyata et al., 2013). Using the exact same experimental settings and feature templates as for English, we performed a 10-fold cross-validation for age prediction using the Japanese data. We obtained a strong correlation between predicted age and actual age, with $r = 0.82$ ($p < 0.0001$). Although this value is slightly lower than the values in Table 1 for English, the range of target values (age in months) is more compressed. Although this experiment included only one child, it does suggest that our approach may work well for Japanese.

5 Related work

Within the literature on assessment of child language development, the metric most closely related to our work is the Index of Productive Syntax (Scarborough, 1990), which we discussed in more detail in

[4]http://childes.psy.cmu.edu/data/EastAsian/Japanese/Miyata/

section 2, and used as a target for data-driven learning. Other traditional metrics include Developmental Sentence Scoring (Lee and Canter, 1971), Language Assessment Remediation and Screening Procedure (Fletcher and Garman, 1988), and D-level (Parisse and Le Normand, 1987) all of which share with IPSyn the reliance on a hand-crafted inventory of grammatical structures meant to be identified manually in transcribed child language samples.

Each of these metrics for child language development, along with the Mean Length of Utterance (Brown, 1973), can be computed semi-automatically using the Computerized Profiling system (Long et al., 2004). Although fully automatic computation with Computerized Profiling produces levels of reliability lower than that of manual scoring, the system can be used with human intervention to produce results of higher quality. More closely related is the work of Sagae et al. (2005) on automating IPSyn using patterns extracted from automatic parse trees. The work we describe in section 2.1 is closely based on that of Sagae et al., which we use as a way to validate our data-driven approach.

Roark et al. (2007) examined the ability of several automatically computed syntactic complexity metrics to discriminate between healthy and language impaired subjects. Among other metrics, Roark et al. used Frazier scoring (Frazier, 1985) and Yngve scoring (Yngve, 1960), which are more commonly associated with processing difficulty than with emergence of syntax in child language development, but are related to our approach in that they are based on straightforward generic features of parse trees (depth, node count), like our counts of grammatical relation labels. Finally, Sahakian and Snyder (Sahakian and Snyder, 2012) have also approached the problem of learning automatic metrics for child language development using a regression approach. Their focus, however, was on the combination of the existing metrics MLU, mean depth of tree (similar to Yngve scoring mentioned above) and D-level, along with a few hand-picked features (counts of certain closed-class words, ratio of function words to content words, and average word frequency), to achieve better discrimination than any of these metrics or features alone. A key difference between our approach and that of Sahakian and Snyder is that their approach builds on and assumes the existence of a metric such as D-level, which, like IPSyn, includes a carefully designed language-dependent inventory of language structures, while we use only simple feature templates applied to parse trees. In addition, they include vocabulary-centric features, while we explicitly avoid vocabulary features, focusing on structural features. It is possible that Sahakian and Snyder's approach would benefit from the parse tree features of our approach, either by using the features directly, or by taking a score obtained by our approach as an additional feature in theirs.

6 Conclusion and future work

We presented a framework for assessment of syntactic development in child language that is completely data-driven, and unlike traditional metrics such as IPSyn, LARSP and D-level, does not rely on a language-dependent inventory of language structures chosen specifically for the task. Instead, our approach is based on the application of support vector regression with simple features extracted from syntactic parse trees. In our experiments we used dependency parses produced by the MEGRASP parser for CHILDES transcripts, but it is likely that other modern dependency and constituent parsers would provide similar results. We showed that our framework is capable of learning IPSyn scores, and that for individual children it can model syntactic development well after MLU and IPSyn scores fail to correlate with age.

Having shown that the feature templates described in section 2.1 are as expressive as the inventory of grammatical structures in IPSyn at tracking language development, and that syntactic development of individual children can be modeled using our data-driven framework in complete absence of an existing metric such as IPSyn, it is interesting to consider the applicability of this framework to different languages for which child language development metrics have not been developed or are not widely used. One possible way to do this is to train several age regression models representing different development profiles. In most practical scenarios, the child's age is known and would not need to be predicted by a model. By predicting age with several different models and selecting the one that most closely predicts the child's actual age, a language development profile matching the child can be found. This could be used, for example, in diagnosis of language impairment. In this paper we established only the expressive

power of regression using simple syntactic features, and the application of this approach to practical tasks is left as an interesting direction for future work.

A related direction for future work is the application of this method for assessment of syntactic development in languages other than English. Given the availability of child language data in various languages (MacWhinney, 2000) and recent progress in syntactic analysis for many of these languages (Buchholz and Marsi, 2006; Nivre et al., 2007), we are optimistic about the applicability of our approach to other languages. Preliminary results using data from one Japanese child suggest that the same set of simple feature templates can be used to track language development in Japanese.

Acknowledgments

We thank the anonymous reviewers for insightful suggestions. This work was partly supported by the National Science Foundation under grants 1263386 and 1 219253 and by the U.S. Army. Any opinion, content or information presented does not necessarily reflect the position or the policy of the United States Government, and no official endorsement should be inferred.

References

Roger Brown. 1973. *A first language: The early stages*. George Allen & Unwin.

Sabine Buchholz and Erwin Marsi. 2006. Conll-x shared task on multilingual dependency parsing. In *Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL-X)*, pages 149–164, New York City, June. Association for Computational Linguistics.

Eugene Charniak. 2000. A maximum-entropy-inspired parser. In *Proceedings of the first conference on North American chapter of the Association for Computational Linguistics*, pages 132–139, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Harris Drucker, Chris, Burges L. Kaufman, Alex Smola, and Vladimir Vapnik. 1997. Support vector regression machines. In *Advances in Neural Information Processing Systems 9*, volume 9, pages 155–161.

Paul Fletcher and Michael Garman. 1988. LARSPing by numbers. *British Journal of Disorders of Communication*, 23(3):309–321.

L. Frazier. 1985. Syntactic complexity. In D. R. Dowty, L. Karttunen, and A. M. Zwicky, editors, *Natural Language Parsing: Psychological, Computational, and Theoretical Perspectives*, pages 129–189. Cambridge University Press, Cambridge.

Khairun-Nisa Hassanali, Yang Liu, Aquiles Iglesias, Thamar Solorio, and Christine Dollaghan. 2014. Automatic generation of the index of productive syntax for child language transcripts. *Behavior Research Methods*, 46:254–262.

Thomas Klee and Martha Deitz Fitzgerald. 1985. The relation between grammatical development and mean length of utterance in morphemes. *Journal of Child Language*, 12:251–269, 6.

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

Steven H. Long, Marc E. Fey, and Ron W. Channell. 2004. Computerized profiling (version 9.6.0).

Brian MacWhinney. 2000. *The CHILDES Project: Tools for Analyzing Talk*. Lawrence Erlbaum Associates, 3rd edition edition.

Mitchell P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of english: The penn treebank. *Comput. Linguist.*, 19(2):313–330, June.

Susanne Miyata, Kenji Sagae, and Brian MacWhinney. 2013. The syntax parser GRASP for CHILDES (in Japanese). *Journal of Health and Medical Science*, 3:45–62.

Joakim Nivre, Johan Hall, Sandra Kubler, Ryan McDonald, Jens Nilsson, Sebastian Riedel, and Deniz Yuret. 2007. The CoNLL 2007 shared task on dependency parsing. In *Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL 2007*, pages 915–932, Prague, Czech Republic, June. Association for Computational Linguistics.

Christophe Parisse and Marie-Thrse Le Normand. 1987. Indicators of linguistic competence in the peer group conversational behavior of mildly retarded adults. *Applied Psycholinguistics*, 8:19–32.

Mabel L. Rice, Filip Smolik, Denise Perpich, Travis Thompson, Nathan Rytting, and Megan Blossom. 2010. Mean length of utterance levels in 6-month intervals for children 3 to 9 years with and without language impairments. *Journal of Speech, Language, and Hearing Research*, 53:1–17.

Brian Roark, Margaret Mitchell, and Kristy Hollingshead. 2007. Syntactic complexity measures for detecting mild cognitive impairment. In *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, BioNLP '07, pages 1–8, Stroudsburg, PA, USA. Association for Computational Linguistics.

Kenji Sagae, Alon Lavie, and Brian MacWhinney. 2005. 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. Association for Computational Linguistics.

Kenji Sagae, Eric Davis, Alon Lavie, Brian MacWhinney, and Shuly Wintner. 2010. Morphosyntactic annotation of CHILDES transcripts. *Journal of Child Language*, 37:705–729.

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

Hollis S. Scarborough. 1990. Index of productive syntax. *Applied Psycholinguistics*, 11:1–22, 3.

Victor H. Yngve. 1960. A model and an hypothesis for language structure. *Proceedings of the American Philosophical Society*, 104:444–466.