Pomona College
Department of Computer Science

Cross-linguistic Data-driven Measurement of Child Language Development

Shannon Lubetich

May 2, 2015

Submitted as part of the senior exercise for the degree of
Bachelor of Arts in Computer Science

Professor Dave Kauchak

Abstract

In order to solve the puzzle of child language acquisition, researchers have developed a number of metrics. All of these metrics involve examining a transcript of utterances, and producing a score that corresponds to language development. The most commonly used metrics rely on surface-level features, which are used due to computational ease and availability. An alternative approach is to, given a transcript, produce a number corresponding to the age of the child examined in the transcript. Using this approach, we can track a child's linguistic development based on age—a number easily understood without intimate knowledge of the process. This age prediction task has been approached from a Machine Learning perspective, involving the extraction of features from transcripts and using these to predict age [SS12, LS14]. This research has shown that simple features correlate well with age; however, it has only been done for English-speaking children.

In this paper, we explore this age prediction approach for language development to children speaking other languages. Using transcripts of children speaking Spanish, Japanese, and Hebrew from the CHILDES database, we examine the age prediction task using similar simple syntactic feature templates as those that were used in previous research in English. We find that approaches using only syntactic features perform at comparable levels to those using more language-specific, content-based features. Our best results show the ability to predict a child's age based on syntactic features from transcripts within two months of their actual age, with strong correlations between predicted and actual age for the data set. Additionally, we compare performance to content-based features and find no significant improvement over syntactic features. We suggest future experiments in order to determine the best feature sets for a given language, and call for increased data collection in this area. With the increased availability of child language transcripts, such a cross-linguistic data-driven approach has the ability to influence, motivate, and assist the research area of child language development.

Acknowledgments

I would like to thank everyone involved in collecting and consolidating child language data. Without their efforts, none of this would have been possible. I would additionally like to thank Professor Kenji Sagae, without whom this project wouldn't have happened. I would like to thank NSF for awarding me an REU grant that started my interest and research in this area. Finally, I would like to thank Professor Dave Kauchak for his patience, guidance, and wisdom.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The way to measure a child's language development is not uniform. Though psychologists and linguists have agreed that children follow similar developmental paths (e.g. one-word utterances, then two-word utterances, some types of phrases before others, etc.), there is no easily-accessible, reliable, and explicit way of measuring and calibrating this level of development. Several intermediary metrics have been proposed, such as Mean Length of Utterance, or MLU [Bro73]. This is a metric that is easily computed, and outputs the average number of morphemes, which are the smallest units of linguistic meaning in a word, per utterance for a transcript of a child's speech. However, the reliability of MLU has been questioned, since at a certain age, utterances stop getting longer (and thus the number of morphemes contained in them stops increasing) but the utterances still get internally more complex due to their grammatical structure.

This led to the proposal of metrics based more on internal grammatical structure, involving a predefined set of grammatical structures that correlate with linguistic development and are used to calculate a score of grammatical complexity. An example of such a metric is the Index of Productive Syntax, or IPSyn [Sca90]. This metric is calculated by examining a child's transcript and awarding points if certain structures are encountered. At the end of an 100-utterance transcript, the totaled number of points is used as a representation of that child's language development. However, such a metric requires a previously existing and agreed upon set of grammatical structures that correspond well with language development, and also the necessary amount of people trained and versed in such a metric in order to evaluate transcripts manually and use results appropriately. Additionally, the evaluation is mostly manual, though some Natural Language Processing techniques have helped to automate parts of the process [SLM05, HLI$^+$13, LS14].

This lack of an automatic, accessible, understandable, and universal metric prevents the study of child language research from making conclusions and comparisons about how exactly children acquire language. The ability to track child language development automatically and reliably could allow researchers to learn more about the human language

faculty in general. We lack knowledge of how first language acquisition works because we are not able to access the internal system that allows a child to speak. For the time being, we cannot see inside a child's mind and examine their firing synapses as a series of clearly labeled bells and whistles. Instead, we only have access to the output of this system– the actual speech they generate. To extract something meaningful from this output, it is necessary to have a general metric to apply as a measure of language development.

Instead of examining a transcript and scoring it based on some abstract and obscure point system, it might be most helpful to examine a transcript and label it with the "developmental age" corresponding to what age the child speaking is thought to be, based on the given transcript. The question then becomes *how* one predicts a developmental age for the speech of the child. After all, this is why abstract metrics have been proposed and used previously–as intermediaries in correlating the changes in score with changes in age.

In fact, such an approach has been explored. Sahakian and Snyder [SS12] and Lubetich and Sagae [LS14] demonstrated the success of data-driven approaches to predict the age of a child, given a transcript of their utterances. However, their experiments were conducted with only English-speaking children. This ignores the possibility of acquiring useful information from the study of acquisition of other languages, and limits our ability to study language acquisition in general. In this paper, we explore the possibility of using a data-driven approach to this problem in non-English languages, including Japanese, Spanish, and Hebrew. We compare the performance of a language-independent approach based on syntactic features of a transcript to approaches based more on content features of a specific language. We hope to introduce and validate an approach that can be used to measure child language development cross-linguistically, providing an easily computable and understandable metric that can guide and inform future research.

Chapter 2

Background

The way children acquire the ability to understand and produce language is not fully understood, despite a wealth of research on the subject. From a Chomskian perspective, there is a preexisting, native language faculty in the brain that requires external linguistic input to set appropriate parameters [FCH+00]. From an interactionist view, children have no preexisting language faculty, but use linguistic input from their environment to abstract and internalize patterns [Tom03]. Both of these approaches are backed by numerous arguments and supporters in the field of language acquisition research. Because we do not have access inside a child's mind, we cannot directly understand how they process language and eventually acquire the ability to speak it at an adult level. In order to gain more insight into this process, researchers have proposed several metrics of language development. Such metrics allow us to examine the abilities of children at different ages, potentially revealing information about how the language faculty develops.

In this chapter, we discuss approaches to measuring child language development, beginning with MLU, a very common metric, in Section 2.1. We then discuss metrics relying more on grammatical structure in Section 2.2, and the use of Machine Learning and Natural Language Processing approaches to automate some of these metrics in Section 2.3. Section 2.4 summarizes several other metrics that aim to measure linguistic competency.

2.1 Mean Length of Utterance

Roger Brown outlined five stages of grammatical development of children, and introduced a metric for this development known as Mean Length of Utterance, or MLU [Bro73].

MLU is based on the idea that a child's utterances increase in length as a child gets older and as their language skills develop. It measures the mean number of morphemes per utterance, where a morpheme is an instance of a smallest unit of linguistic meaning in a word. For example, "untied" contains three morphemes: the prefix "un," the root "tie," and the past tense suffix "-ed." The total number of morphemes in a transcript is divided

by the number of "utterances," which can be thought of as sentences, to yield the MLU score.

However, the effectiveness of using an MLU score to model grammatical development has been questioned, and tested in a number of papers. [KF85] found that MLU did not correlate significantly with age, and also that MLU did not show differences between children with varying profiles of grammatical development. Others have examined the application of the MLU measurement to children with language disorders [SRTF+91]. Scarborough et al.'s results showed that, for normally-developing preschoolers, MLU correlated well with language development, but tended to overpredict the linguistic ability of those with language disorders.

2.2 Grammatical Metrics

Several metrics have been proposed that use grammatical information extracted from preprocessing to categorize the development of children. Some of these have been adapted to use in languages other than English as well. The focus on grammatical structure is motivated by the lack of correlation between MLU and age. At a young age, it seems valid that longer utterances signify increased development. However, at a certain age, it has been posited that utterances no longer increase in length, but instead increase in internal complexity depending on the types of grammatical structures used [Sca90]; MLU does not account for this, but grammatical metrics can.

In the next sections, we examine two such grammatical structure metrics in detail. First, we discuss Developmental Sentence Scoring, a metric that has been adapted from English for measuring development in Japanese and Spanish-speaking children. Then, we discuss the Index of Productive Syntax, which has only been defined specifically for English, but has been shown to correlate well with age.

Developmental Sentence Scoring

One such metric is Developmental Sentence Scoring, or DSS, originally proposed as a tool to help clinicians estimate a child's language level and plan lesson strategies for improvement [LC71]. DSS filled the gap between metrics suited for spontaneous speech like MLU, and those meant for examining utterances at a syntactic level that were not built for spontaneous utterances. This scoring system relies on eight previously defined grammatical categories such as pronouns, verbs, negatives, and conjunctions, and examines only complete sentences consisting of a subject and a predicate. Points are awarded if all required syntactic and morphological rules have been observed in a given grammatical category. This does not give information on the child's explicit errors; that requires further analysis.

In part because of the small number of defined grammatical categories, DSS has been adapted to measure child language development in languages other than English. Developmental Sentence Scoring for Japanese, or DSSJ, required the manual redefinition of

DSS's grammatical categories to those corresponding to language development in Japanese [MMO+13]. The same categories could not be used because of differences between the two languages (for example, English is head-initial and Japanese is head-final), and adaptation required a great deal of knowledge of both languages. DSSJ was used to evaluate a number of subjects, and found to be a score that "reliably reflects the morpho-syntactic development of Japanese children."

The Developmental Assessment of Spanish Grammar, or DASG, is similar to DSS, but the creators stress that it is not an attempted translation of DSS, but rather incorporates knowledge of Spanish language acquisition [Tor76]. They present syntactic hierarchies for six grammatical categories: indefinite pronouns and noun modifiers, personal pronouns, primary verbs, secondary verbs, conjunctions, and interrogative words.

Index of Productive Syntax

Another proposed metric that utilizes grammatical information is the Index of Productive Syntax, or IPSyn [Sca90]. IPSyn works by processing an 100-utterance transcript of child language and awarding points based on the appearance of 60 items in a previously defined inventory of grammatical structures. IPSyn was shown to correlate well with age, but operates with a large structure inventory defined specifically for English, so is not easily adaptable to other languages.

2.3 Metric Automation

One of the appeals of MLU as a metric is that it is very easy to automate and compute: for each utterance, one must segment it and total the number of morphemes, and then average the morpheme length of the utterances across the total number of utterances in the transcript.

Other metrics prove more difficult to automate, as they are focused on explicitly defined constructions that require more manual effort to define in a way that software can recognize.

Data

Before any automation can take place, machine-readable transcripts are required. As with most research fields, considerable work cannot be done without a large, available dataset. Brian MacWhinney and Catherine Snow established the Child Language Data Exchange System, or CHILDES[1], which is the child language component of the TalkBank system used for sharing and studying conversational interactions [Mac00]. This online corpus contains transcripts of child language from numerous studies, with a wide range of types of interaction, age groups, geographic locations, socioeconomic statuses, and languages spoken. In order for a transcript to be included, it must follow certain transcription

[1] http://childes.psy.cmu.edu

and annotation formats. CHILDES additionally provides an analysis program, CLAN, involving pattern matching and statistical tools for extracting information from transcripts.

Preprocessing

Some automated approaches may require preprocessing to identify things like grammatical relations and part-of-speech tags; while there are a number of such programs out there, most are suited to a specific domain–that is, whatever corpus they were trained on. Child language data is quite different from things like United Nations transcripts and news articles, in that it involves disfluencies, nonce words, and gaps, which can prove difficult for standard preprocessing programs to handle.

The CHILDES database provides an analysis program, CLAN, and a morphological analyzer, MOR [Mac00]. It also suggests the POST part-of-speech-tagger, which was specifically defined for the CLAN software suite [PLN00]. Additionally, MEGRASP is a parser designed for child language data that can be used to generate dependency tree structures of child utterances [SDL$^+$07].

Parsers are used to extract more information from text, including how the syntactic tree is constructed, either in a branching format involving constituents or a graph-like structure combined with grammatical relations. Because a great deal of research has been done in NLP on English, numerous accurate and precise parsers exist for English; however, they are not as common in less-studied languages. It is still possible to develop parsers for other languages, which is an area of ongoing research involving competitions to test the flexibility and adaptability of parsers on new and understudied languages [BM06]. Additionally, the parser for child language mentioned above, MEGRASP, has been expanded to parse child language data in Spanish [SDL$^+$10] and Japanese [MMO$^+$13].

Approaches

[Cha03] attempted to analyze child language in the same way a researcher would go through and score a transcript, but with the ability to score a larger volume in a shorter amount of time. They automated the Developmental Sentence Scoring (DSS) metric, achieving a correlation coefficient of .97 with manually generated scores.

Another proposed metric that utilizes grammatical information is the Index of Productive Syntax [Sca90]. This has been automated by a number of researchers. One of the approaches handles multiple transcription formats, is publicly available, and mimics the process of a human scorer with similar levels of accuracy [HLI$^+$13]. As a proof-of-concept of the applications of their dependency parser, Sagae et al. briefly presented an automated IPSyn scorer [SLM05]. Another approach achieved similar levels of accuracy with a process mimicking human scorers, but also proposed a more data-driven perspective [LS14]. In the data-driven approach, the authors selected non-content-based features from utterances, and used them to predict the IPSyn score based on a corpus of annotated data. They then

expanded this approach to the task of age prediction.

Sahakian and Snyder present another approach to measuring child language development, forgoing any attempt at automating previously defined metrics in favor of using machine learning techniques to extract features and predict child age [SS12].

The previous two techniques approached the idea of measuring child language in a different way, with age. Whereas previous metrics output a score where a higher number corresponds to greater development, these approaches output the age of a child. This avoids the issue that metrics often run into: whether or not they correlate well with age. Ultimately, the goal is to measure a child's language development path. We expect that as children age their speech will get more sophisticated and their language skills will improve. The ability to have an accurate and meaningful representation of how well they perform at different ages will help reveal patterns and trends for further research. Predicting a child's age can be easily evaluated and understood by researchers. This approach also avoids defining a specific set of grammatical categories that a transcript must be evaluated on, line-by-line, and instead predicts age of a child based on automatically extracted features of a transcript.

Feature Selection

The success of the age prediction task depends heavily on which features are chosen for the data-driven approach. Lubetich and Sagae [LS14] do not consider content features, but instead use four simple syntactic feature templates, such as part-of-speech tags and grammatical relations between words. This approach was motivated by the fact that IPSyn does not examine content, but instead the abstraction of grammatical structures. Additionally, avoiding content features means that this approach can be applied to a transcript of any language, given enough data to train a model for age prediction based on the features that correspond with development in that language. The feature templates used will be the same, but the features could be weighted differently for the prediction task, if for example, there is a certain grammatical relation that appears in one language to correspond more strongly with development than in another.

Sahakian and Snyder [SS12] define a larger feature set for their age prediction task, using a combination of features that have been shown to correlate with language development. These include MLU, mean depth of dependency trees, D-level, counts of obligatory morphemes, and Type-Token Ratio (TTR). D-level is a score for individual sentences based on "Developmental Level," which was originally proposed by Rosenberg and Abbeduto in 1987 and is a 7-point scale of complexity based on the presence of specific grammatical constructions. The feature set used by [SS12] differs from [LS14] in that it utilizes measurements that depend on carefully crafted structure inventories (D-level) and features that are vocabulary-centric (e.g. counts of certain morphemes, TTR).

Additional features can always be incorporated and tested in such a task, and there is a large variety available to choose from. In one of the earliest procedural books for clinically

assessing the language production of children, Jon Miller outlines what he considers to be a minimal set of features required to fully assess a child's language development [Mil81]. These include vocabulary and syntax-semantics in comprehension; syntax, semantics, and phonology in production; and communication functions and intentions in language use. He encourages using MLU as a helpful metric, but also suggests others like TTR and multiple procedures specifically relating to syntactic analysis.

2.4 Other Linguistic Competency Metrics

Several other research areas have approached the task of predicting the level of linguistic competency, such as classifying the reading level of a text. Various metrics have been proposed in this area, ranging from more superficial and easily computed to those requiring heavy analysis [FJHE10]. The Flesch-Kincaid Grade Level Formula examines the average number of words per sentence and the average number of syllables per word to predict readability as a grade level [KFRC75]. The Gunning FOG index uses average sentence length and the percentage of words with more than two syllables [Gun]. Others use word frequency, with the idea that more infrequent words will appear only in more difficult texts. Statistical language modeling approaches have also been applied: [PO09] uses features from n-gram language models in combination with features from syntactic parse trees and traditional approaches to predict readability level with a support vector machine. In a comparison of a number of these approaches, [FJHE10] found that average words per sentence was the best shallow feature for predicting readability level. Other well-performing features included the presence of noun part-of-speech tags, language modeling with word and part-of-speech tag pairs, and entity-density.

Chapter 3

Approach

In this chapter, we explain in detail the methodologies and experiments used in our approach. The end goal involves taking a transcript of child language, extracting features from it, and using a trained regression model to predict the age of the child from the extracted features. In order to get to that point, we must train a model using labeled data of extracted features with the corresponding actual age.

3.1 Age Prediction as Regression

The problem of predicting a child's age from a transcript of their speech can be set up as a regression task in the following manner. First, given transcripts of a child's speech annotated with the age of that child, we extract features from the transcript and pair them with the age in months. This annotated data is used to train a regression model using a support vector machine. Once trained, the model can be used to predict the age of a child given a new transcript. The same feature templates are used to extract features from the new transcript, and then the trained model is used to predict an age in months from the extracted features.

3.2 Feature Selection

As discussed in Section 2, previous approaches to the data-driven age prediction task have utilized a variety of features. We take a similar approach to Lubetich and Sagae, relying on simple syntactic templates [LS14]. In order to access information relating to the grammatical structure of an utterance, we first need a way of annotating utterances in a child language transcript with morphological and syntactic information. Here, we turn to several programs to assist in preprocessing. We used the CLAN tools for morphology analysis (MOR) [Mac00], part-of-speech tagging (POST) [PLN00], and parsing (MEGRASP) [SDL$^+$10], since it is straightforward to process CHILDES transcripts using these, and they

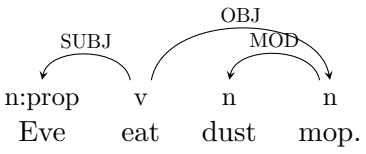

Figure 3.1: A dependency tree generated with part-of-speech and grammatical relation information.

provide high-accuracy analyses for child language transcripts. This results in transcripts with utterances annotated with part-of-speech tags and dependency parse trees, a pictographic example of which can be seen in Figure 3.1[1]. Here, "Eve" is a proper noun and the dependent of its head word "eat," which is a verb. The grammatical relation between these two words is labeled "SUBJ," meaning that the dependent is the subject of the head word; in this case, "Eve" is the subject of "eat."

A number of different of combinations of these grammatical annotations can be used to form feature templates.

Bag-of-words (BOW): This consists of using the actual words in the transcript. This is heavily language-dependent and context-specific, and we use this as a comparison of how more abstract feature templates perform in our regression task.

Part-of-speech tags (POS): Here, we simply use the part-of-speech tag of each word. This relies on morphological information and can be thought of as a bag of part-of-speech tags.

Grammatical relations (GR): Another feature template is a bag of dependency labels, where each label corresponds 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 [SDL$^+$10].

Head-dependent part-of-speech pairs (POSPOS): Our third syntactic 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-dependent grammatical relation pairs (GRGR): This feature class is based on pairs of grammatical relations, where each pair corresponds to a two-level dependency in the parse tree. That is, this pairs the label corresponding to a grammatical relation between a dependent word and its head word in the dependency tree and the label

[1]For ease of understanding, this tree involves an English utterance. However, similar grammatical information and syntactic annotations are extracted for each language.

of the grammatical relation between that head word and its head word. This can be thought of as including the relationship between child node, parent node, and grandparent node.

Head-relation-dependent triples (POSGRPOS): The last feature class is similar to the head-dependent pairs described above, but also includes 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.

For example, extracting the above feature sets from the utterance and its annotations shown in Figure 3.1 would give us the following features:

BOW	`Eve eat dust mop`
POS	`n:prop v n n`
GR	`SUBJ OBJ MOD`
POSPOS	`v_n:prop v_n n_n`
GRGR	`OBJ_MOD`
POSGRPOS	`v_n:prop_SUBJ v_n_OBJ n_n_MOD`

These features were chosen specifically for this approach. Bag-of-words will be used as a baseline example of using a content-specific feature in the age prediction task. Our goal is not to model how the content of a child's speech changes and develops, but to model how the grammatical complexity increases. This requires the use of feature patterns that can extract relevant and representative information from morphological and syntactic annotations, which we attempt to do with the other patterns outlined above.

Chapter 4

Experiments

In this chapter, we present the data used, the experimental setup, and the evaluation metrics for performance on the age prediction task.

4.1 Data

The CHILDES database aggregates transcripts of child language from around the world [Mac00]. The preprocessing programs that generate morphological and syntactic annotations (as discussed in Section 3.2) are only available in a small number of languages. This naturally focused our experiments on three languages: Japanese, Hebrew, and Spanish. A number of transcripts for these languages are available in the database, but for this approach, they must meet specific criteria.

First, because we treat age prediction as a child-specific task, we need enough data for a single child to train a regression model across a number of age intervals. This requires us to use children involved in longitudinal studies. Additionally, we seek spontaneous child vocalizations. That is, we want to exclude any utterances that are imitations or repetitions of things said by other conversational participants, as well as any reading-type tasks. Because the author does not actually speak the languages being studied, we had to trust in the descriptions of the data collection tasks as "spontaneous" and "natural play" settings. Additionally, the transcription scheme used for recording child speech includes markers for things such as imitations, and any utterances marked as such were excluded from our experiments.

In order to normalize the amount of data available for a single transcript, we mimic the approach of calculating IPSyn [Sca90] by splitting transcripts into separate files such that each includes exactly 100 utterance lines from the child under study.

This resulted in a data set of over 400 transcripts of Japanese children from the Miyata and Ishii corpora [Miy95, Ish99] (Ryo and Jun, covering ages from 9 months to 3 years 8 months), over 200 transcripts of Hebrew children from the Berman longitudinal corpus (Hagar and Leor, including ages from 1 year 7 months to 3 years 3 months), and over 100

files of Spanish children from the Llinàs-Grau corpus (Irene) and the Vila corpus (Emilio, covering ages from 1 year 8 months to 4 years [Vil90]). Additionally, we included corpora from two English speaking children, using a data set of 350 transcripts from the Suppes and Sachs corpora [Sup74, Sac83] (Nina and Naomi, ages 1 year 6 months to 4 years 9 months). These summaries are displayed in Table 4.1, with details of the individual child corpora found in Table 4.2.

	Japanese	Hebrew	Spanish	English
Number of files	426	204	132	354
Start age	0y;9m	1y;7m	1y;8m	1y;6m
End age	3y;8m	3y;3m	4y;0m	4y;9m
Age range (months)	35	20	28	39

Table 4.1: Summary of corpora used per language from the CHILDES database. Ages are formatted in "ay;bm" where the child is a years and b months old.

	Japanese	Hebrew	Spanish	English
	Ryo	Hagar	Irene	Nina
Number of files	56	90	87	271
Start age	1y;10m	1y;7m	1y;10m	1y;11m
End age	3y;0m	3y;3m	3y;1m	3y;3m
	Jun	Leor	Emilio	Naomi
Number of files	370	114	45	83
Start age	0y;9m	1y;9m	1y;8m	1y;6m
End age	3y;8m	3y;0m	4y;0m	4y;9m

Table 4.2: Details of files from individual children used from the CHILDES database. Ages are formatted in "ay;bm" where the child is a years and b months old.

4.2 Experimental Setup

Following the example of previous studies involving the data-driven age prediction task, we chose to approach this as a child-specific learning problem [SS12, LS14]. We take the transcripts from a single child, train a model on them and their extracted features, and then use this model to predict age on transcripts from the same child (but that were withheld from the training set). Because of limited data, we used leave-one-out cross-validation. In this experiment, our datapoints are equivalent to transcripts annotated with the age of child at the time the transcripts were collected. The leave-one-out approach requires that

for every datapoint d for a single child, we train a model using all other data points, and then predict the age for d, evaluating our performance by examining the average difference between our model's predicted age and the actual age for this datapoint.

We used the SVM Light[1] implementation of support vector regression with default parameter settings [Joa99].

Our content-based feature experiment involved using bag of words as our only feature template, extracting all words from a child's utterances in a transcript and using those to train a model and predict age.

Then, we ran numerous experiments involving different combinations of more abstract, syntactic feature templates, the building blocks of which were defined in Section 3.2. We used POS, POSPOS, GR, GRGR, POSGRPOS, and then finally a combination of POS, GR, POSPOS, and POSGRPOS that had been shown to perform best in English for the age prediction task by Lubetich and Sagae [LS14].

4.3 Evaluation Metrics

To evaluate the performance of our various approaches to the age prediction task, we examined accuracy and correlation. Because we used leave-one-out cross-validation, we have a numerous instances of resulting age predictions in months, as well as the corresponding actual age for that prediction. In order to measure performance, we take the difference of these two numbers (the absolute value of the difference between actual and predicted age in months) and average this value over all datapoints. Additionally, we take these sets of predicted age and corresponding actual age, and determine the Pearson r for this data[2]. Calculating the Pearson r is a way of representing how well two variables linearly correlate. A perfect correlation between actual and predicted age would result in a Pearson r of 1.

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

[2] We use the python scipy.stats module for this calculation. More information can be found at http://www.statsoft.com/textbook/glosp.html#Pearson%20Correlation

Chapter 5

Results

In this chapter, we first summarize our overall findings, highlighting the features that yielded the best performance on the age prediction task. We then present results for each language, and the specific children within that language.

We preface this chapter with an observation about the claim that such a data-driven, feature dependent approach is language independent. This 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. Thus, we cannot directly compare the results of performance in one language to that in another. Our experiments are meant to validate the possibility of using syntactic feature templates for age prediction *within a language*, and to explore how a variety of templates and combinations perform.

5.1 Best Features

Though we cannot directly compare one language to another, our experiments found similar intra-language results. The best feature templates consistently included part-of-speech information, and often excluded grammatical relations. The head-dependent pair grammatical relations (GRGR) perform the worst for many of the children, which we conjecture is due to the level of intricacy and detail encoded by this feature. GRGR includes the grammatical relation between a child node and its parent in the parse tree, but also the grammatical relation between the parent node and its parent (the original child node's grandparent) in the parse tree. This could become too specific as to no longer encode relevant information. Additionally, due to the simplistic nature of child speech, a single utterance might have no, or very few, GRGR features to extract. This problem can be further compounded by data scarcity, giving us even fewer instances of GR and GRGR features that correspond well with language development.

Part-of-speech information proves to be a functional middleman between the language-specific words and the extremely abstract grammatical relations. Part-of-speech information includes salient facts about the words themselves without getting too specific, and also can encode internal complexity when used in combination (such as with head-dependent part-of-speech pairs or by including grammatical relations).

For these reasons, we find features including part-of-speech information to consistently perform best on our age prediction task. In most cases, such features perform at similar levels to the bag of words feature template. This demonstrates the validity of choosing to use abstract, grammatical features instead of content-based features in an age prediction task. This supports the hypothesis that language development is characterized not just by the complexity of words, but also by the complexity of structures. We have shown that this complexity can be accessed using simple feature templates involving morphological and syntactic information.

In the following pages, we present performance results for our age prediction task in each language, followed by an in-depth discussion.

5.2 Result Evaluation

First, we considered the data-driven approach in a child specific manner. This means that the accuracy of performance on the task was measured for a single child's predicted and actual age. Each table of results is arranged such that if feature patterns are in the same sectioned off row, then there is no significant difference in their performance on the task based on a two-tailed t-test of significance. Feature sets that are separated by a horizontal line in the table do perform significantly differently.

In order to observe something meaningful about the performance of this data-driven age-prediction approach for each language in general, we then combined the predicted and actual age datapoints for the children within a language, and recalculated the average differences and correlation coefficients for each feature pattern. In this analysis, we excluded feature templates that performed statistically significantly worse for both children individually.

5.3 Tables of Results

5.3.1 Japanese

Feature Template	Average Age Difference (months)	Pearson r
POSPOS	2.02	0.77
POSGRPOS	2.04	0.77
BOW	2.15	0.75
GRGR	2.35	0.71
POS	2.43	0.72
COMBO	2.45	0.71
GR	2.54	0.67

Table 5.1: Results of the prediction task for Ryo from the Miyata corpus. All horizontal groupings are significantly different at a level of $p < .05$.

Feature Template	Average Age Difference (months)	Pearson r
BOW	2.12	0.88
POSGRPOS	2.2	0.86
POSPOS	2.25	0.85
COMBO	2.45	0.85
POS	2.61	0.83
GRGR	3.1	0.77
GR	3.36	0.80

Table 5.2: Results of the prediction task for Jun from the Ishii corpus. All horizontal groupings are significantly different at a level of $p < 0.005$.

Feature Template	Average Age Difference (months)	Pearson r
BOW	2.13	0.87
POSGRPOS	2.18	0.85
POSPOS	2.22	0.85
COMBO	2.45	0.84

Table 5.3: Results of the prediction task for Japanese. All horizontal groupings are significantly different at a level of $p < 0.005$.

5.3.2 Hebrew

Feature Template	Average Age Difference (months)	Pearson r
POSPOS	3.04	0.75
POSGRPOS	3.05	0.74
POS	3.2	0.72
POS[1]	3.2	0.72
BOW	3.34	0.62
COMBO	3.62	0.74
GR	3.92	0.67
GRGR	3.92	0.55

Table 5.4: Results of the prediction task for Hagar from the Berman corpus. All horizontal groupings are significantly different at a level of $p < 0.05$.
[1]This line is repeated here because it is not significantly different from BOW, but also not significantly different from the feature templates in the above grouping.

Feature Template	Average Age Difference (months)	Pearson r
BOW	1.69	0.88
POS	1.96	0.86
POSPOS	1.99	0.85
POSGRPOS	2.04	0.84
COMBO	2.53	0.85
GRGR	2.97	0.82
GR	3.96	0.81

Table 5.5: Results of the prediction task for Leor from the Berman corpus. All horizontal groupings are significantly different at a level of $p < 0.05$.

Feature Template	Average Age Difference (months)	Pearson r
BOW	2.42	0.76
POSPOS	2.45	0.79
POSGRPOS	2.48	0.78
POS	2.51	0.78
COMBO	3.01	0.76

Table 5.6: Results of the prediction task for Hebrew. All horizontal groupings are significantly different at a level of $p < 0.001$.

5.3.3 Spanish

Feature Template	Average Age Difference (months)	Pearson r
BOW	2.99	0.85
POSPOS	3.4	0.85
POS	3.43	0.81
POSGRPOS	3.45	0.85
COMBO	4.03	0.78
GRGR	4.32	0.67
GR	4.37	0.61

Table 5.7: Results of the prediction task for Irene from the Llinàs-Grau corpus. All horizontal groupings are significantly different at a level of $p < 0.001$.

Feature Template	Average Age Difference (months)	Pearson r
POS	5.67	0.81
POSPOS	7.39	0.87
BOW	7.48	0.86
COMBO	7.77	0.91
POSGRPOS	7.86	0.84
POSGRPOS[2]	7.86	0.84
GR	8.06	0.90
GRGR	8.08	0.77

Table 5.8: Results of the prediction task for Emilio from the Vila corpus. All horizontal groupings are significantly different at a level of $p < 0.005$.
[2]This feature template is included in this grouping because it is not significantly different from the features in this grouping or the one above.

Feature Template	Average Age Difference (months)	Pearson r
POS	4.20	0.81
BOW	4.52	0.79
POSPOS	4.76	0.80

Table 5.9: Results of the prediction task for Spanish. All horizontal groupings are significantly different at a level of $p < 0.05$.

5.3.4 English

Feature Template	Average Age Difference (months)	Pearson r
BOW	1.83	0.90
POSGRPOS	2.39	0.85
POSPOS	2.4	0.84
POSPOS[3]	2.4	0.84
COMBO	2.54	0.84
POS	2.55	0.84
GRGR	2.8	0.80
GR	2.86	0.76

Table 5.10: Results of the prediction task for Nina from the Suppes corpus. All horizontal groupings are significantly different at a level of $p < 0.05$.
[3]This feature template is repeated here because it is not significantly different from others in this group or the group above.

Feature Template	Average Age Difference (months)	Pearson r
POSPOS	5.52	0.78
POSGRPOS	5.56	0.78
BOW	5.67	0.80
POS	5.9	0.80
POS[4]	5.9	0.80
COMBO	6.18	0.76
GR	6.75	0.73
GRGR	6.95	0.64

Table 5.11: Results of the prediction task for Naomi from the Sachs corpus. All horizontal groupings are significantly different at a level of $p < 0.001$.
[4]This feature template is repeated here because it is not significantly different from others in this group or the group above.

Feature Template	Average Age Difference (months)	Pearson r
BOW	2.72	0.78
POSPOS	3.12	0.77
POSGRPOS	3.13	0.77

Table 5.12: Results of the prediction task for English. All horizontal groupings are significantly different at a level of $p < 0.01$.

5.4 Data Analysis

5.4.1 Japanese

This section discusses the application of our approach for Japanese, an East Asian language.

Examining results individually, we see in Table 5.1 that for Ryo, our syntactic templates predict age within two months of actual age. This is a closer prediction than for Jun, shown in Table 5.2, but the prediction task for Jun results in stronger correlations between the predicted and actual age overall. It is interesting to note that for both Japanese children, the top performance on the prediction task included the same sets of features – POSPOS, POSGRPOS, and BOW.

In the combined analysis found in Table 5.3, we find that the use of the best feature templates results in a prediction that is just over two months off of the actual age (on average), with a reasonably strong correlation. In their preliminary test of this task using Japanese data, Lubetich and Sagae found that their approach resulted in a Pearson r value of 0.82 [LS14]. Whereas these authors used the "COMBO" feature set that they found worked best for English, our technique used a variety of feature templates, allowing us to explore their performance and find feature templates that resulted in a statistically significant increase in correlation between predicted age and actual age.

Though BOW (bag-of-words) is a content-based, language-specific feature and performs the best at our prediction task, we additionally find two other feature templates with differences in performance that are not statistically significant from that of BOW. This supports the hypothesis that there do exist more abstract, syntax-based, language-independent features that can perform at similar rates to a language-specific feature on this task.

5.4.2 Hebrew

Here we analyze results of our approach on children speaking Hebrew, a Semitic language.

Examining the individual children's data, we find that the top features for both include BOW, POS, POSPOS, and POSGRPOS. Similarly, in Japanese, we also found that the BOW, POSPOS, and POSGRPOS feature templates resulted in the most accurate and well correlated age predictions. Preliminarily, we observe that feature templates involving part-of-speech tags and head-dependent part-of-speech tag pairs perform the best at our age prediction task, across at least two languages.

In Table 5.6, we show the combined analysis, finding a lower correlation than other experiments here and also cited approaches, but observe that there is a smaller available dataset for Hebrew, which can affect performance on a data-driven machine learning task.

5.4.3 Spanish

In this section, we examine performance on Spanish, a Romance language.

The result of training and testing a model for Irene (see Table 5.7) shows a separation in performance between feature templates using part-of-speech information and those that do not. Both the average age difference and correlation coefficient get statistically significantly worse without any such part-of-speech features.

In observing the performance for Emilio (see Table 5.8), we see a larger month difference in predicted and actual age than in any previous experiments. Our current explanation for this returns to the need for extensive data, distributed equally over a certain age range. Our data for Emilio covered ages 1 year 8 months to 4 years old, but only included 45 files, whereas Irene's data included twice as many transcripts over a 12 month smaller age range.

The combined analysis in Table 5.9 shows that, in fact, bag-of-words does not perform better than any other feature pattern, supporting our hypothesis that abstract features depending on morphological information can perform similarly to language-dependent features.

5.4.4 English

The goal of this research is to compare and validate the performance of simple syntactic feature templates in the data-driven age prediction task for children speaking languages other than English. Though performance in English has been outlined in previous research [SS12, LS14], these approaches used different feature templates and parameters in the regression task. To provide a baseline for comparison we also ran our experiments on a small English data set similar to the data sets of the other languages.

As discussed in the beginning of this chapter, it is not meaningful to compare performances in one language to another. This experiment is meant to validate that simple syntactic feature templates again, work for English, and additionally to explore a variety of templates and combinations and examine how they perform within a language.

The individual performance of this approach for both children seen in Tables 5.10 and 5.11 shows that POSPOS, POGRPOS, and BOW are the top-performing feature templates. This is similar to what we've seen in previous sections for other languages. However, we find that performance for Naomi from the Sachs corpus has a lower correlation and greater difference in months from the predicted and actual age than our other models. Again, a lack of sufficient data may contribute to this level of error. We acquired 83 files for Naomi, covering an age range of 39 months; whereas Nina's corpus generated 271 transcripts over 16 months. Our lack of available data for Naomi across a broader range of ages can hinder the accuracy of the learned model and performance on the prediction task.

Even though Table 5.12 combines the data from Nina and Naomi found in Tables 5.10 and 5.11 respectively, it produces average differences and correlations much better than those found on the experiments with Naomi. Again, we conjecture that this is due to the small number of transcripts used from Naomi over a wide range of ages; the number of transcripts used in Nina's prediction task is approximately three times more than those

available for Naomi's, covering a smaller range of ages. Thus, when all paired datapoints are combined, the average performance is much better than the performance seen just for Naomi.

Chapter 6

Future Work

We have presented an approach to measuring child language development that is data-driven and language-independent. We accomplish this by using machine learning techniques and extracting morphological and syntactic features from child language transcripts. We examined the performance of this approach in Japanese, Hebrew, and Spanish, demonstrating its viability as a metric.

This project is one of the first to explore the possibility of applying morphological and syntactic analyzers to languages other than English. Additionally, this application was done on transcripts of child language.

Using the resulting morphological and syntactic information, we demonstrated the possibility of using a data-driven regression model to predict developmental age of a child.

The real limitation in this research is the availability of data. In evaluating our approach on English-speaking children, we found that the performance of a data-driven model is extremely reliant upon the availability of data to train that model. In the case where there is not an abundance of training data, the predictions are understandably inaccurate. Our results in Chapter 5 are limited by this factor. Because of the limited data, we have only shown the application to two specific children within each language examined. The further validation of this data-driven approach to measuring child language development requires applying it to more children. Such experiments could help determine the best feature templates per language, as well as parameter settings for support vector regression. Additionally, it would be interesting to examine what features are most heavily weighted in the trained regression model, and see how this differs between languages. This could reveal information about types of words or grammatical structures that correspond with linguistic development within a language, and potentially cross-linguistically.

Given enough data, we believe such an easily computable metric could help influence and further the field of child language development and language acquisition research.

For a single language, if we have determined the best features to extract, and train on a large corpus of children over a variety of ages, then, given the transcript of a new child, we can predict what their expected "developmental speaking age" would be, based

on the data of their peers. This predicted age could be used to target language areas for improvement, or identify speech disorders early in development.

Bibliography

[BM06] Sabine Buchholz and Erwin Marsi. 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 2006. Association for Computational Linguistics.

[Bro73] Roger Brown. *A first language: The early stages*. George Allen & Unwin, 1973.

[Cha03] Ron W. Channell. Automated developmental sentence scoring using computerized profiling software. *American Journal of Speech-Language Pathology*, 12(3):369–375, 08 2003.

[FCH+00] V Fromkin, S Curtiss, B.P. Hayes, N Hyams, Keating P.A., H Koopman, P Munro, D Sportiche, E.P. Stabler, D Steriade, T Stowell, and A Szabolscsi. Principles and parameters. In V Fromkin, editor, *Linguistics: An Introduction to Linguistic Theory*, pages 338–348. Blackwell Publishing Ltd., 2000.

[FJHE10] Lijun Feng, Martin Jansche, Matt Huenerfauth, and Noémie Elhadad. A comparison of features for automatic readability assessment. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 276–284. Association for Computational Linguistics, 2010.

[Gun] R Gunning.

[HLI+13] Khairun-nisa Hassanali, Yang Liu, Aquiles Iglesias, Thamar Solorio, and Christine Dollaghan. Automatic generation of the index of productive syntax for child language transcripts. 2013.

[Ish99] Takeo Ishii. The jun corpus, unpublished. 1999.

[Joa99] T. Joachims. Making large-scale SVM learning practical. In B. Schölkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods - Support Vector Learning*, chapter 11, pages 169–184. MIT Press, Cambridge, MA, 1999.

[KF85] Thomas Klee and Martha Deitz Fitzgerald. The relation between grammatical development and mean length of utterance in morphemes. *Journal of Child Language*, 12:251–269, 6 1985.

[KFRC75] J. Peter Kincaid, Robert P. Fishburne, Richard L. Rogers, and Brad S. Chissom. Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel. Technical report, February 1975.

[LC71] Laura L. Lee and Susan M. Canter. Developmental sentence scoring: A clinical procedure for estimating syntactic development in children's spontaneous speech. *Journal of Speech and Hearing Disorders*, 36:315 – 340, 1971.

[LS14] Shannon Lubetich and Kenji Sagae. Data-driven measurement of child language development with simple syntactic templates. Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, pages 2151–2160. Dublin City University and Association for Computational Linguistics, 2014.

[Mac00] Brian MacWhinney. *The CHILDES Project: Tools for Analyzing Talk*. Lawrence Erlbaum Associates, 3rd edition, 2000.

[Mil81] Jon F. Miller. *Assessing Language Production in Children: Experimental Procedures*. University Park Press, 1981.

[Miy95] Susanne Miyata. The aki corpus longitudinal speech data of a japanese boy aged 1.6-2.12. *Bulletin of Aichi Shukutoku Junior College*, 34:183–191, 1995.

[MMO+13] Susanne Miyata, Brian MacWhinney, Kiyoshi Otomo, Hidetosi Sirai, Yuriko Oshima-Takane, Makiko Hirakawa, Yasuhiro Shirai, Masatoshi Sugiura, and Keiko Itoh. Developmental sentence scoring for japanese. *First Language*, 33(2):200–216, 2013.

[PLN00] Christophe Parisse and Marie-Thrse Le Normand. Automatic disambiguation of the morphosyntax in spoken language. *Behavior Research Methods, Instruments, and Computers*, 32:468–481, 2000.

[PO09] Sarah E. Petersen and Mari Ostendorf. A machine learning approach to reading level assessment. *Computer Speech & Language*, 23(1):89 – 106, 2009.

[Sac83] J Sachs. Talking about the there and then: The emergence of displaced reference in parentchild discourse. In K. E. Nelson, editor, *Childrens language*, volume 4. Lawrence Erlbaum Associates, 1983.

[Sca90] H. S. Scarborough. Index of Productive Syntax. *Applied Psycholinguistics*, 11(Peer Reviewed Journal):1–22+, 1990.

[SDL+07] K. Sagae, E. Davis, A. Lavie, B. MacWhinney, and S. Wintner. High-accuracy annotation and parsing of CHILDES transcripts. In *Proceedings of the Workshop on Cognitive Aspects of Computational Language Acquisition*, pages 25–32. Association for Computational Linguistics, 2007.

[SDL+10] Kenji Sagae, Eric Davis, Alon Lavie, Brian MacWhinney, and Shuly Wintner. Morphosyntactic annotation of CHILDES transcripts. *Journal of Child Language*, 37:705729, March 2010.

[SLM05] Kenji Sagae, Alon Lavie, and Brian 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.

[SRTF+91] Hollis S. Scarborough, Leslie Rescorla, Helen Tager-Flusberg, Anne E. Fowler, and Vicki Sudhalter. The relation of utterance length to grammatical complexity in normal and language-disorded groups. *Applied Psycholinguistics*, 12(Peer Reviewed Journal):23–45, 1991.

[SS12] Sam Sahakian and Benjamin Snyder. Automatically learning measures of child language development. In *ACL (2)*, pages 95–99. The Association for Computer Linguistics, 2012.

[Sup74] P Suppes. The semantics of childrens language. *American Psychologist*, 29:103–114, 1974.

[Tom03] M Tomasello. Origins of language. In *Constructing a language: a usage-based theory of language acquisition*, pages 9–42. Harvard University Press, 2003.

[Tor76] Allen S. Toronto. Developmental assessment of spanish grammar. *Journal of Speech and Hearing Disorders*, 41(2):150–171, 05 1976.

[Vil90] I Vila. *Adquisición y desarrollo del lenguaje*. Gra, 1990.