How Do Children Learn a Language? A Machine Learning Analysis of Childes Data

Delores Tang

MACSS 30200

Abstract

How children gradually acquire a language has been a popular research topic in the field of linguistics and developmental psychology for decades. However, researchers are still unsure about how and why children tend to comprehend and produce certain sentence structures before others. In order to explore more thoroughly on the pattern of language acquisition, this project applied machine learning techniques on the open-sourced Childes data, analyzing 5- to 11-year-old children's language production. The objective of this project was to find the pattern of structural change in sentences produced by children at different ages by examining their produced sentences on a word level. Simultaneously, the project also aimed to distinguish the differences in such pattern produced by children with and without language impairment. Results suggested that certain part-of-speech pairs and trigrams appeared in their language more frequently as they grew up, and this pattern differed significantly between children with and without language impairment. As the power of this current study was greatly constrained by the small data size, future research utilizing similar methodologies could use a larger corpus for analysis.

Children's language development has been a popular area of study in the field of developmental psychology and linguistics over the past decades. In psychology, language ability is examined along with many other cognitive functions such as learning and memory (Zlatev, Lundmark, Falck, & Andren, 2009). In linguistics, studying early language development offers crucial insights to how language systems first evolved in human history (Bra˚ ten, 2009). Therefore, most studies focused on a young infant population of the age 0 to 3 years old to investigate their language acquisition trajectory. According to the past literature, children with English as the first language seemed to acquire certain elements, such as nouns, before other elements, like verbs (Gentner, 1975). However, there still lacks a structural analysis of which types of words or sentences can be more easily learned and why. Therefore, this current study aims to look for the pattern of change in frequently used sentence structures across age. More specifically, I intended to analyze, computationally, how the proportional use of certain sentence structures change in frequency across age.

In addition, few studies have compared the sentence structures produced by children with typical and impaired language abilities (Pavelko & Owens, 2019). Clinical diagnosis primarily focuses on grammatical structures, sentence length, sentence complexity, and other language production and comprehension abilities. However, little clinical research went down to the word level, addressing specific types of words and sentence components. Though this word-level analysis may not be necessary for language impairment diagnosis at the current stage, it could extend clinicians' knowledge about this specific mental disorder and could point a future

direction of using machine learning techniques to assist clinicians in detecting language impairment.

Therefore, in this current study, I analyzed young children's language production, at a word level, over the age of 5 to 11. My analysis could be divided into three parts: 1) *Preliminary Analysis* - investigating the correlation between the proportional use of certain parts of speech (POS) and Age, separated by Language Impairment; 2) *Word-Based Analysis* - utilizing neural network and machine learning techniques with inputs as word embeddings and outputs as Language Impairment and Age; and 3) *POS-Based Analysis* - looking for the Age patterns of using specific part-of-speech (POS) pairs and trigrams and their differences between Language Impairment conditions.

Before getting to the description of this current study, I would like to first review the past literature on the pattern of language acquisition.

Literature Review

The noun-verb distinction in language acquisition

According to the previous research on developmental psychology and linguistics, children's ability to acquire verbs seemed to lag behind their ability to acquire nouns. Young children's abilities to learn (Nelson, 1974), to understand (Gentner, 1975), and to produce nouns (Goldin-Meadow, Seligman, & Gelman, 1976) seemed to be better than verbs at an early stage of language learning. For example, as shown by a series of studies conducted by Waxman and his colleagues (2001, 1995, 2009), infants as young as 12-24 months could successfully map novel nouns to objects, while the same ability for mapping verbs to actions and events did not come up until 24 months of age. In another experimental study (Goldin-Meadow, Seligman, & Gelman, 1976), two-year-olds showed a better ability to both comprehend and produce nouns than verbs.

This noun-verb distinction does not only persist across different stages of development, but it also seems to appear in languages other than English. For instance, in a cross-cultural study by Imai et al. (2008), 3- and 5-year-old children whose mother tongues were either Japanese, English, or Chinese all exhibited better abilities to generalize novel nouns than novel verbs despite nuanced differences in the process of learning novel vocabularies.

In order to explain this pattern in noun and verb acquisition, many hypotheses have been proposed by linguists and psychologists. In the book *Language Development: Language, Thought, and Culture* by Gentner (1982), two theories - the Natural Partitions hypothesis and the Linguistic Relativity theory - were compared. The first hypothesis claims that there is a conceptual distinction between nouns and predicate terms such as verbs and prepositions. In our perceptual world, concrete concepts, represented using nouns in general, are easier to grasp than abstract concepts of actions, change-of-state, or causal relations conveyed using verbs (Gentner, 1982). On the other hand, the Linguistic Relativity theory approaches to this phenomenon from a different direction and argues that there is no perceptual difference in viewing concrete versus abstract concepts. According to this theory, nouns are learned before verbs in English because English is a noun-centered language, and this pattern may be different on children in other languages that are verb-centered such as Chinese and Japanese, a belief that was later refuted by cross-cultural studies such as Imai et al. (2008).

Researchers inspired by this Natural Partitions idea later categorized words in a different way other than their parts of speech. For instance, Gentner (1982) summarized words to four types - nominal, interminate, predicate, and expressive, instead of simply nouns versus verbs, when studying early language learning in order to find the underlying mechanism that makes verbs easier to acquire by young children. Specifically, nominal terms have concrete object

references; predicate terms refer to actions, change of state, or other predicate notions and are mainly verbs and prepositions; expressive terms directly portray feelings and thoughts; and, finally, intermediate terms have ambiguous usage. According to Gentner (1982), nominal words were learned earliest and followed by intermediate words, while other verb-like predicate and expressive terms entered considerably later.

The categorization of words was given more fluidity by introducing a new concept of "imageability" (McDonough et al., 2011). McDonough and his colleagues suggested that, because nouns represent more concrete concepts, they are generally more imageable than verbs that represent abstract terms. Therefore, the term 'imageability' was defined to measure the ease for a concept to evoke a corresponding mental image. Findings suggested that imageability of a word was negatively correlated to the age of acquisition: the easier for a word to be mentally represented, the earlier it could be acquired by children (McDonough et al., 2011). A relevant view was proposed by Leddon et al. (2011), who argued that it would take longer for children to verbs because verbs convey more information than nouns. In other words, in order to comprehend a verb, a child does not only need to understand the action itself but also the arguments that the verb acts on and the relationship between different objects. Both claims by McDonough et al. (2011) and Leddon et al. (2011) have supported the idea that the noun-verb distinction pattern in language acquisition relates to how children perceive the world and processes information. Therefore, by exploring this learning pattern more closely, psychologists could better understand how children gradually learn language structures as well as learn to describe the world.

Challenges and Limitations of Previous Studies

Although researchers have put plenty of efforts to improve their understanding of the pattern of language acquisition from the noun-verb distinction to imageability to words, there are still certain difficulties faced by researchers of all times.

First, linguistic data are always insufficient. On the one hand, studies that recorded natural conversations in daily life returned highly unorganized data. Experimenters usually have little control over the environment and the topic of the conversation so that correlation would be impossible to draw. On the other hand, studies inviting children to laboratories would be criticized for not providing a secure, naturalistic environment for children participants to produce language. More importantly, linguistic data collection can be both labor- and time-consuming, because it often requires experimenters to cooperate with families for years, while more families would drop out from such longitudinal studies as time goes.

Second, due to the lack of data available, most linguists studied this pattern of language acquisition on children with normal language abilities. Though some neuroscientists studied noun-verb processing on patients with mental disorders such as aphasia and amnesia (Alyahya, 2018), most were conducted on adults. The literature on how language impaired children learn and use language is largely missing in the field of developmental psychology. However, knowing the differences between the language learning process of typical versus language impaired children can give crucial insights into the understanding of language impairment. It could help clinicians to better explain what hinders language impaired children to acquire a complete language system and, thus, come up with efficient treatments to facilitate their language learning. Moreover, linguists rarely adopt deep learning in their research despite the fast-improving techniques of using this method for social science data analysis and natural language processing.

Attempting to overcome these two limitations, I would use the Gillam corpus (Gillam et al., 2005) obtained from the open-sourced database, Childes (https://childes.talkbank.org/), and perform neural network models to examine the patterns of early language acquisition at a word level.

In conclusion, I reviewed the literature on the pattern of children's early language acquisition. In general, consistent results were reported to suggest that infants across culture tend to learn nouns before verbs. Two hypotheses emerged to explain this distinction - Natural Partitions and Linguistic Relativity, and the latter was refuted by cross-cultural empirical studies. I claim that a word-level analysis is needed to better our understanding of what types of words and sentence structures could be more easily learned across age and how that pattern differs between children with and without impaired language abilities.

Methods

In this section, I would describe the data I used in this experiment and the three stages of analyses I did (Preliminary Analysis, Word-Based Analysis, and POS-Based Analysis). This current study aims to extract features of words produced by children of different ages and with different language abilities. Transcripts would be analyzed in word level, and the input variables at each stage mainly involve: 1) the proportional use of parts of speech (POS), 2) word embeddings (as inputs to the neural networks), 3) and vectorized POS tags. There were two major output variables: 1) Language Impairment (Impaired vs. Normal), and Age (ranges from 5 years 0 months to 11 years 11 months).

Dataset

For this project, I used the Gillam corpus (Gillam & Pearson, 2004) from the open-sourced database, Childes (MacWhinney, 2000). Childes stands for the Child Language Data

Exchange System, coming from a project started by MacWhinney and his colleagues in 1984 in order to establish a language acquisition database for researchers studying in various languages and all around the world.

Specifically, the Gillam corpus contained linguistic data produced by 250 children with language impairment and 520 with normal language abilities. The original purpose of this corpus was to build the Test of Narrative Language (TNL; Gillam & Pearson, 2004), a test that aims to evaluate the quality of narratives produced by children and, therefore, to assist the diagnosis of language disorders. Linguistic data were obtained in a naturalistic way, in which children participants were asked to tell a story in a given scene (in McDonald's in this experiment). Children were also audiotaped during storytelling, and the audio files were transcribed later to CHAT format and uploaded to Childes under Gillam corpus.

Data preprocessing. In order to apply neural network models and other data analytic tools to the dataset, I preprocessed the data so that separate transcripts were combined into one file. First, I utilized childes-db (Sanchez et al., 2018), an API built in R that helps to transform CHAT files to more accessible and reproducible tabular format. I further processed the obtained tabular format to CSV format and kept only columns that were relevant to the current study: *id* (the identification number for each sentence), *transcript_id* (the identification number for each transcript/child), *gloss* (one gloss is a sentence produced by a child), *part_of_speech* (POS tags as defined by Childes), *Impaired* (whether a child has impaired = 1 or normal = 0 language abilities), *Age* (5 to 11), and *Gender* (female = 0 and male = 1).

After renaming and organizing columns of the dataset, each row represents one sentence produced by a child. Rows with empty gloss were dropped. As Gillam corpus dropped some children's data due to incompleteness, only 668 children's data were analyzed. There are 497

children who had normal language abilities, and 171 had impaired abilities; 323 were female and 345 were male. Children were ranged from age 5 to 11 years old, and specific statistics of age are included in Figure 1.

Preliminary Analysis

Before applying any deep learning models to the data, I first conducted some preliminary analysis using the frequencies of verbs and nouns children produced during storytelling as the input variable and their language abilities and Age as dependent variables. One way ANOVA, t-test, and Ordinary Least Square (OLS) regression were used in the Preliminary Analysis in Python. The proportional use of nouns, verbs, and nouns + verbs was computed for each participant by dividing the number of times of use by the total number of words. Besides verbs and nouns, frequencies of other POS tags were also computed in order to examine if there were patterns of the proportional use of other parts of speech that might be predicting features of the two output variables, Age and Language Impairment.

Word-Based Analysis

The word-based analysis was conducted using different neural networks with the column "gloss" (transformed into word embeddings) as input and either Language Impairment or Age as output.

Word embeddings. After processing the Gillam data, words in "gloss" was transformed into embeddings in three ways: GloVe, Word2Vec using a continuous Bag-of-Words (CBOW) architecture, and Word2Vec using a Skip-Gram (SG) architecture.

The Global Vectors for Word Representation (GloVe) is an unsupervised algorithm that represents words using pre-trained vectors (Pennington, Socher, & Manning, 2014). In this study,

I used the word vector file “Glove.6B.zip”, which was trained on Wikipedia 2014 and Gigaword to obtain the embedding matrix.

Word2Vec is another word vectorizing approach developed by Mikolov and colleagues (2013). Different from other word vectorizers, it produces vectors for a target word based on its context, specifically, words surrounding the target word. CBOW- and Skip-Gram-based Word2Vec are slightly different in how they define and predict the context of a given word.

Three different embedding matrices were computed in this study for further word-based analysis.

Vocabulary. Each row in gloss was a sentence produced by a child, and it was regarded as one sample. In total, there are totally 24609 examples in the preprocessed data. The data was divided into a training set (68% of the data), a validation set (12% of the data), and a test set (20% of data). A vocabulary of 3958 words was obtained from the dataset and was used to create three different embedding matrices.

Model Construction. Several deep learning models were built and manually tuned using word embeddings as input and either Language Impairment or Age as output.

Naive Simple RNN model as a baseline. During analysis, I used a simple recurrent neural network model without pre-trained embedding matrices as a baseline model. One baseline model was trained and evaluated for each of the output variables - Age and Language Impairment.

Two models are of the same architecture. Each model has one embedding layer, which transforms vocabulary index to word embeddings, as no pre-trained matrix was used. The model has three hidden layers, each with 32 neurons and an activation function of ReLU (Rectified Linear Unit). For Age, the output layer used Softmax function to perform a multinomial

classification, while for Language Impairment, the output layer used a sigmoid function to perform binary classification.

LSTM-based RNN. Three Recurrent Neural Network (RNN) models were then trained using three embedding matrices obtained respectively using GloVe, CBOW-based Word2Vec, and Skip-Gram-based Word2Vec for word-based analysis. By comparing the initial models, I selected the GloVe embedding based on the reported validation loss and accuracy. I then built an RNN model stacked with a Long Short-Term Memory (LSTM) layer with modified numbers of hidden layers and neurons per layer. Because no overfitting was observed for word-based analysis, I did not use any dropout or regularizers.

Finally, the best model for each output variable was selected and evaluated on the held-out test set.

POS-Tag-Based Analysis

Since the purpose of the study is to examine specific types of words and sentence structures, POS tags were more important than original words in the analysis.

POS bigrams and trigrams. POS tags were not only analyzed in a single-tag manner but grouped into bigrams and trigrams. Each bigram/trigram was indexed and vectorized using the TfidVectorizer provided by the Sklearn package in Python. This tool vectorizes each input, character, word, or bigram, according to its TF-IDF (Term Frequency-Inverse Document Frequency) score. This approach enables researchers to take the frequency of words appeared in documents into account while also offset the naturalistic imbalance of using certain words more frequently than others.

Model construction. Several machine learning models were tested and manually tuned using POS vectors as input and either Language Impairment or Age as output. After excluding

those models that did not converge, I further tuned five models - Logistic Regression, Naive Bayes, Linear Support Vector Machine Classifier (LinearSVC), Ridge, and Random Forest Classifier. See Figure 2 for their best validation mean squared errors (MSE). The ridge model was selected as the best model for both outputs (Language Impairment and Age).

Feature selection. The model was tested on the test set. Feature importance plots were generated for both Impairment and Age to evaluate what factors tended to be most influential. Individual Conditional Expectation (ICE) plots were created for the five most influential factors in order to interpret the direction of effects.

Results

Preliminary Results

Noun-Verb Patterns. In order to test whether the early hypothesis of the noun-verb distinction continues to children aged 5 to 10, I analyzed whether the ratio of noun versus verb (N / V; N = the total number of nouns produced by a participant, V = the total number of verbs produced by a participant) was correlated to Age. The result was neither significant on the whole sample nor on the subsamples separated by language impairment. However, when analyzing the correlation between the proportion of nouns and verbs together across all words produced ($\frac{N+V}{T}$; T = the total number of words produced by a participant), both groups, Impaired and Typical, showed a decrease ($p < .05$) in the proportion of nouns and verbs (Figure 3). Moreover, this decreasing trend differed significantly between two groups ($F = 5.13$, $p < .05$), where the Impaired group had a sharper decrease in noun and verb use than the controlled group.

Other Parts of Speech. Besides of nouns and verbs, I also calculated the proportions of the use of all the other 42 different types of part-of-speech (POS) in children's language production ($\frac{P}{V}$; P = the number of a certain type of part of speech, P, produced by a participant).

Only proportions of the interrogative (e.g., “what”) and relative (e.g., “when”) pronouns were significantly correlated to Age (See Figure 4). I also separately analyzed by language impairment conditions. Although children with language impairment used fewer interrogative and relative pronouns than children with typical language abilities throughout the development, the differences tended to converge as they grow up (refer to Table 1 for specific test statistics).

Word-Based Results

The baseline sequential model returned low maximum validation accuracy (*val_acc* = 0.24) and high minimum validation cost (*val_cost* = 2.23) in predicting Age, but it returned relatively better results for Language Impairment (*val_acc* = 0.78, *val_loss* = 0.50).

Modified models gave similar results predicting Age, suggesting that either a change at the word level does not exist across ages or the training data size was too small for the models to learn.

Modified RNN models with LSTM predicting Language Impairment returned better results. Among all the three embeddings (Table 2 for statistics of all three embeddings), the model using GloVe embedding was selected to be the best according to its performance (see Figure 5 for performance over epochs). When testing on the test set, the test loss (0.7706) was slightly higher than the validation loss (0.5322), indicating that there might be overfitting. The test accuracy (0.7709) was similar to the validation accuracy (0.7731), implying that it could be possible to use words as inputs to predict whether children have impaired abilities using deep learning.

POS-Based Results

As mentioned earlier, the ridge model was selected as the best model for both Impairment and Age as outputs (see Figure 6 for the feature importance plot for language impairment and Age).

For Impairment, the following five bigrams appeared to be most important among all bigrams and trigrams: conjunction + article (determiner), subjective pronoun + infinitive,coordinator + article (determiner), conjunction + subjective pronoun, and coordinator + subjective pronoun. The coordinator + article (determiner) pair was used less frequently by children with typical language abilities but more frequently by those with impaired abilities. Other four bigrams were all more frequently used by language impaired children (see Figure 7 for the ICE plots of the five most important bigrams/trigrams for Impairment).

For Age, interjection + subjective pronoun, conjunction + subjective pronoun, preposition + article (determiner), coordinator + subjective pronoun, and coordinator + proper noun were found as the most important features. The coordinator + proper noun pair was used more frequently as Age increased. The interjection + subjective pronoun, conjunction + subjective pronoun, and preposition + article (determiner) bigrams all appeared less frequently in sentences produced by older children. The coordinator + subjective pronoun pair seemed to first increase and then decrease in its frequency of use as children grow older (see Figure 8 for the ICE plots of the five most important bigrams/trigrams for Age).

Discussion

According to the results of this analysis, I could draw some important conclusions about children's produced sentence structures at the age of 5 to 11.

First, the Preliminary Analysis showed that older children seemed to produce more clauses led by interrogative (e.g., "who") and relative (e.g., "when") pronouns. At all age levels

children with language impairment tended to produce fewer interrogative and relative pronouns. This finding supported the Sampling Utterances and Grammatical Analysis Revised (SUGAR) Measures to identify Specific Language Impairment (SLI) mentioned in the article by Pavelko and Owens (2019), in which clauses use was considered to be a diagnostic criterion for language impairment. However, the results also suggested that the difference in proportional use of these two pronouns between two impairment conditions converged as children grew up. This finding implies that language impaired children may be slow to acquire certain sentence structures when they were young, but they could still catch up with children with typical abilities later in their childhood.

The Word-Based Analysis provides a potential way to use deep learning to help identify language impairment for clinical psychologists. Specifically, I found that neural network models returned a fairly good prediction of language impairment using word embeddings as inputs. However, although the model reported a relatively high validation and test accuracy, all models I built based on this dataset stopped improving at $val\_acc = 0.7731$, indicating that the training data was too small for further training. In the future, researchers could apply similar neural network techniques on larger datasets, which might give better results in using deep learning to identify language impairments.

The POS-Based Analysis gave more complex results that require linguistic analysis to interpret. First, language impaired children seemed to prefer certain sentence structures over others. For example, they tended to use more structures such as pairs of a coordinator/conjunction with a subjective pronoun (e.g., “and I”, “and she”, “so she”) than children with typical language abilities. One possible explanation is that conjunctions followed with another sentence, compared with clauses, can be more easily learned for children with

impaired language abilities. Therefore, when language impaired children are attempting to express complex meanings, they would be more comfortable to use conjunctions rather than clauses. Interestingly, the difference between children with and without language impairment seemed to match the difference between old and young children. Younger children tended to use more conjunction/coordinator + subjective pronoun pairs than the olders. It suggests that children with language impairment may be lagged in language development.

Despite the findings of this analysis, this study suffers from many limitations. First, when using neural networks in the word-based analysis, I used each sentence as one input example for predicting Impairment and Age. However, there were in total 24,609 sentences produced by children participants while 3809 words in the vocabulary as input features. The data size was, indeed, too small compared with such a large set of features. Although my results exhibited some trend that word-level natural language models could be used to identify language impairment, the small data set hindered my models to return good training results. Therefore, future studies might consider to use a larger corpus for deep learning analyses. Second, children's produced sentences were drastically different in length. This could partially affect the training process. I have not figured out an effective approach in preprocessing the data to avoid the noise created by imbalanced sentence lengths. Therefore, future research may search for other means to control for sentence lengths. Third, I recognized that categorizing children to language-impaired versus typical might be greatly imprudent. The severity of language impairment was not measured so that correlations between language impairment severity and certain sentence structures could not be drawn. Therefore, more detailed clinical reports would be needed if researchers want to explore more closely to language impairment.

Overall, this study aimed to examine the pattern of sentence structures produced by children at different ages and with different language abilities. Three types of analyses were conducted: the Preliminary Analysis (using proportional use of POS tags), the Word-Based Analysis (using word embeddings as inputs and neural network models for analysis), and the POS-Based Analysis (using vectorized POS tags as inputs and machine learning models for analysis). Although the results gave some insightful findings regarding language development across Age, many neural network models built in the study were largely limited in accuracy and learning abilities due to the small data size. However, the study pointed out a potential future direction for linguists and psychologists by adopting machine learning techniques to analyze language production and to identify language impairment.

References

Alyahya, R. S. W., Halai, A. D., Conroy, P., & Lambon Ralph, M. A. (2018). Noun and verb processing in aphasia: Behavioural profiles and neural correlates. *NeuroImage: Clinical*, *18*, 215–230. doi: 10.1016/j.nicl.2018.01.023.

Bra˚ ten, S. (2009). *The Intersubjective Mirror in Infant Learning and Evolution of Speech.* Amsterdam: John Benjamins Publishing Co. Retrieved from http://search.ebscohost.com.proxy.uchicago.edu/login.aspx?direct=true&db=nlebk&AN=284010&site=eds-live&scope=site.

Gentner, D. (1975). Evidence for the psychological reality of semantic components: The verbs of possession. *Explorations in Cognition*, *35*, 211-246.

Gentner, D. (1982). Why nouns are learned before verbs: Linguistic Relativity versus Natural Partitioning. In S. Kuczaj II (Ed), *Language Development, Volume 2: Language, Thought, and Culture* (pp.301-334). Hillsdale, N.J.: Lawrence Erlbaum.

Gillam, R. B., & Pearson, N. (2004). Test of Narrative Language. Austin, TX: Pro-Ed Inc.

Goldin-Meadow, S., Seligman, M., & Gelman, R. Language in the two-year-old. *Cognition, 4*. 189-202.

Imai, M., Li, L., Haryu, E., Okada, H., Hirsh-Pasek, K., Golinkoff, R. M., & Shigematsu, J. (2008). Novel Noun and Verb Learning in Chinese-, English-, and Japanese-Speaking Children. *Child Development*, *79*, 979–1000. doi: 10.1111/j.1467-8624.2008.01171.x.

Leddon, E., Arunachalam, S., Waxman, S., Fu, X., Gong, H., & Wang, L. (2011). Noun and Verb Learning in Mandarin-acquiring 24-month-olds. *Online Proceedings Supplement of the 35th Annual Boston University Conference on Language Development*.

MacWhinney, B. (2000). *The CHILDES Project: Tools for analyzing talk. Third Edition.* Mahwah, NJ: Lawrence Erlbaum Associates.

McDonough, C., Song, L., Hirsh-Pasek, K., Golinkoff, R. M., & Lannon, R. (2011). An image is worth a thousand words: why nouns tend to dominate verbs in early word learning. *Developmental Science*, *14*, 181–189. doi: 10.1111/j.1467-7687.2010.00968.x.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., & Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. *NIPS.*

Nelson, K. (1974). Concept, word, and sentence: Interrelations in acquisition and development. *Psychological Review, 81*, 267-285. doi: 10.1037/h0036592.

Pavelko, S. L., & Owens Jr., R. E. (2019). Diagnostic Accuracy of the Sampling Utterances and Grammatical Analysis Revised (SUGAR) Measures for Identifying Children With Language Impairment. *Language, Speech & Hearing Services in Schools, 50*, 211–223. https://doi-org.proxy.uchicago.edu/10.1044/2018_LSHSS-18-0050.

Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global Vectors for Word Representation.

Sanchez, A., Meylan, S., Braginsky, M., MacDonald, K. E., Yurovsky, D., & Frank, M. C. (2018, April 23). childes-db: a flexible and reproducible interface to the Child Language Data Exchange System. Retrieved from psyarxiv.com/93mwx.

Waxman, S. R., & Booth, A. E. (2001). Seeing pink elephants: Fourteen-month-olds' interpretations of novel nouns and adjectives. *Cognitive Psychology, 43,* 217-242.

Waxman, S. R., Lidz, J. L., Braun, I. E., & Lavin, T. (2009). 24-month-old children's interpretations of novel nouns and verbs in dynamic scenes. *Cognitive Psychology, 59,* 67-95.

Waxman, S.R., & Markow, D.B. (1995). Words as invitations to form categories: Evidence from 12-month-old infants. *Cognitive Psychology, 29*, 257-302.

Zlatev, J., Lundmark, C., Falck, M. J., & Andren, M. (2009). *Studies in Language and Cognition.* Newcastle upon Tyne: Cambridge Scholars Publishing. Retrieved from http://search.ebscohost.com.proxy.uchicago.edu/login.aspx?direct=true&db=nlebk&AN=553533&site=eds-live&scope=site.

Figure 1. Descriptive statistics of Age in Gillam corpus data.

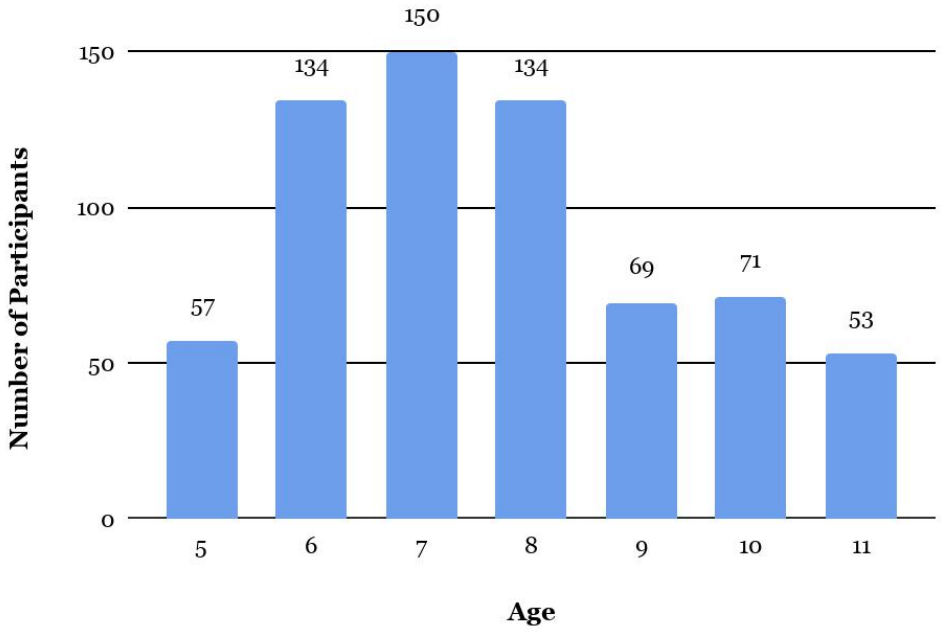


Figure 2. Mean Squared Errors (MSEs) for all machine learning models. *The left* is the error rates for Language Impairment, and *the right* is the error rates for Age.

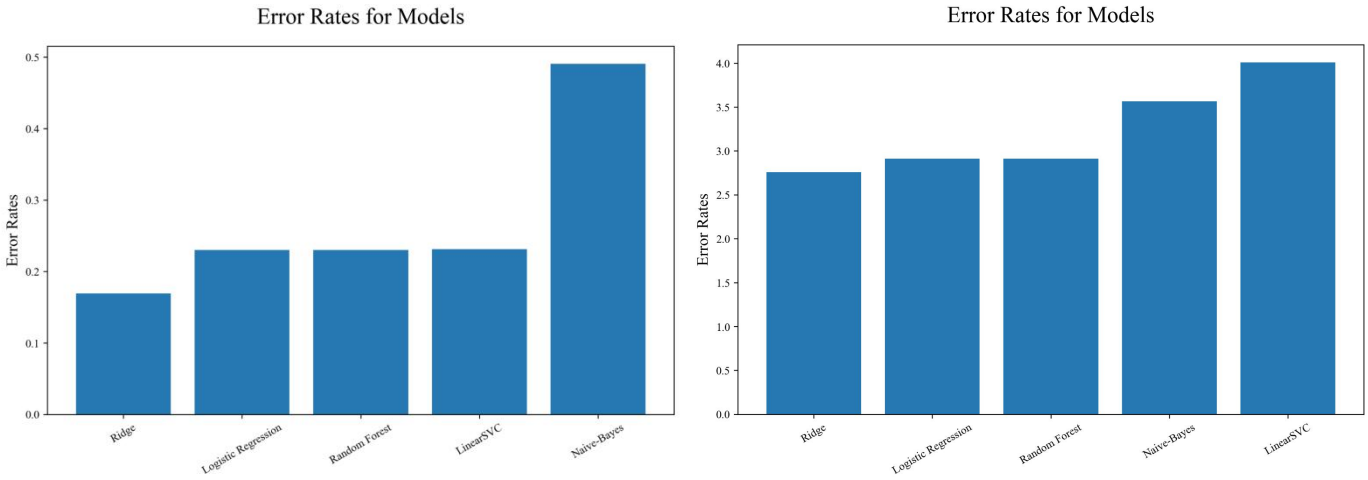


Figure 3. The proportional use of nouns and verbs across Age, separately analyzed by Language Impairment conditions.

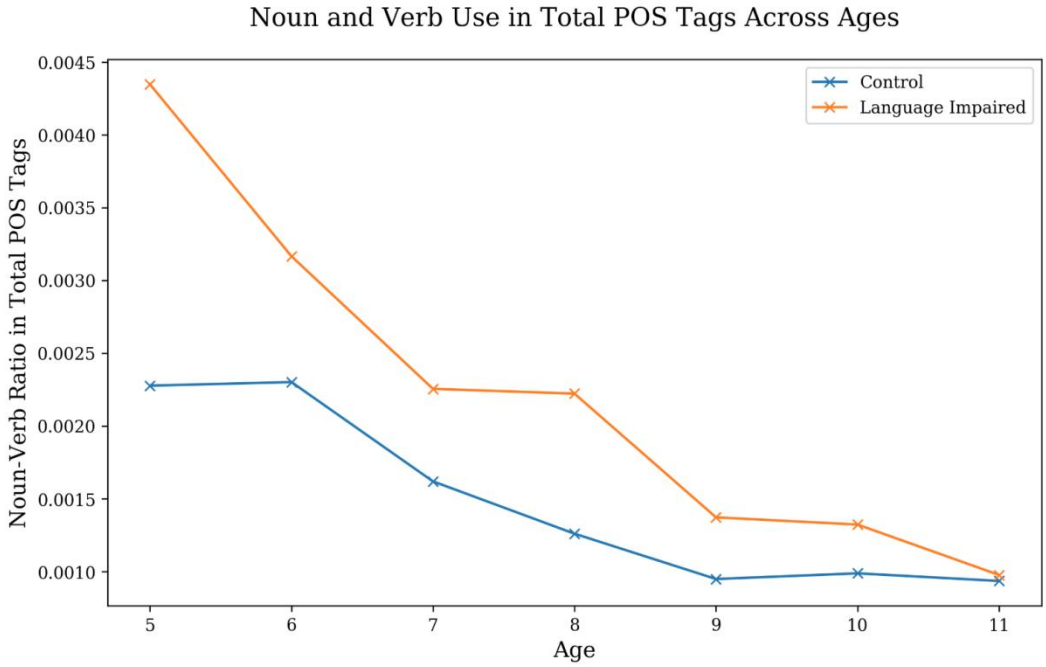


Figure 4. The proportional use of interrogative and relative pronouns across Age, separately analyzed by Language Impairment conditions.

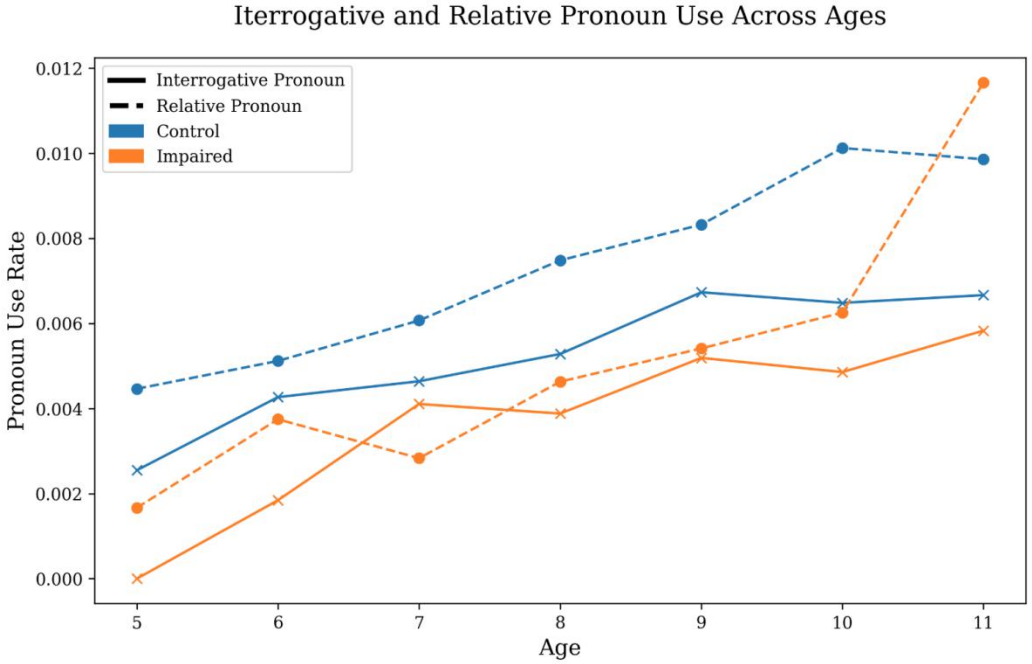


Figure 5. Validation accuracy and loss of the simple RNN model stacked with a LSTM layer over epochs, early stopped at epoch = 11.

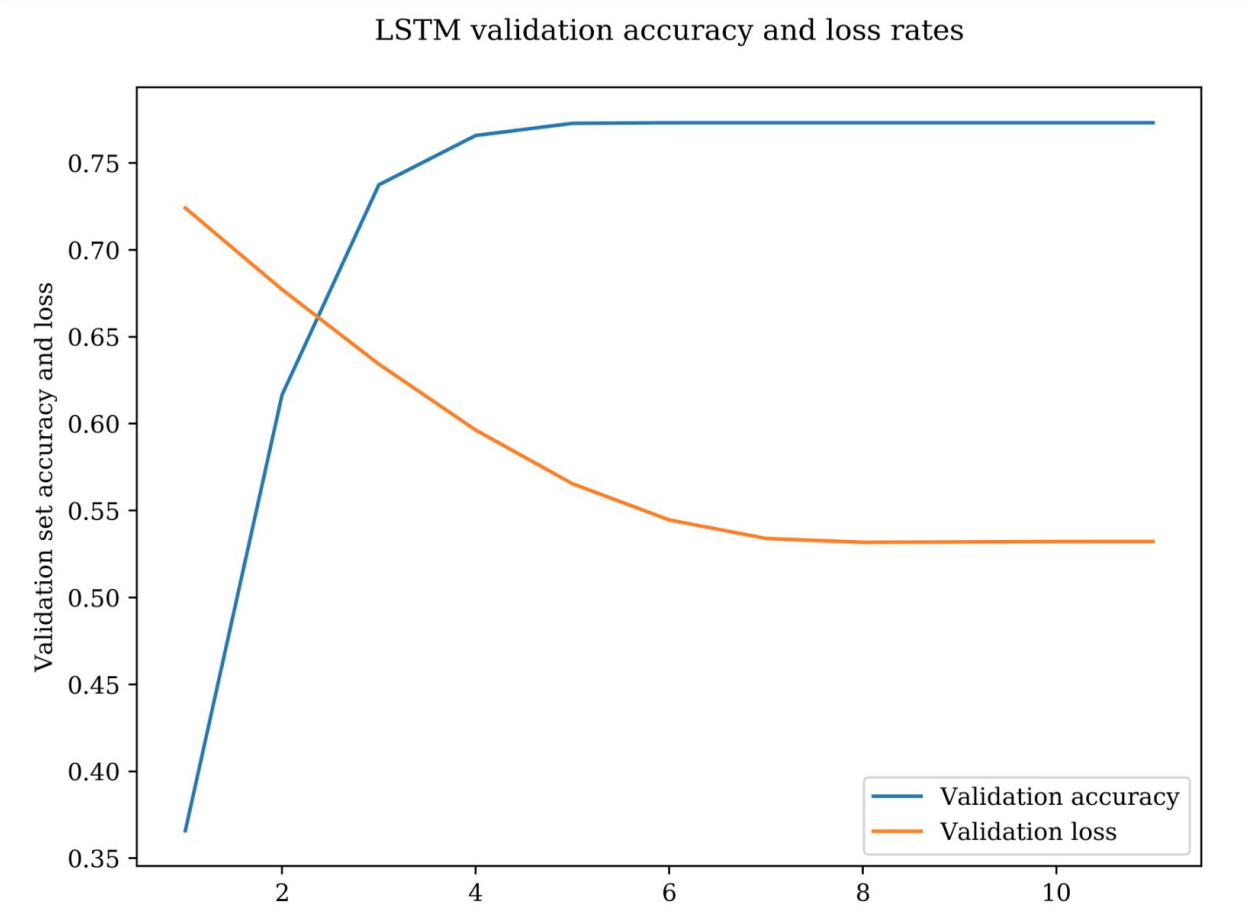
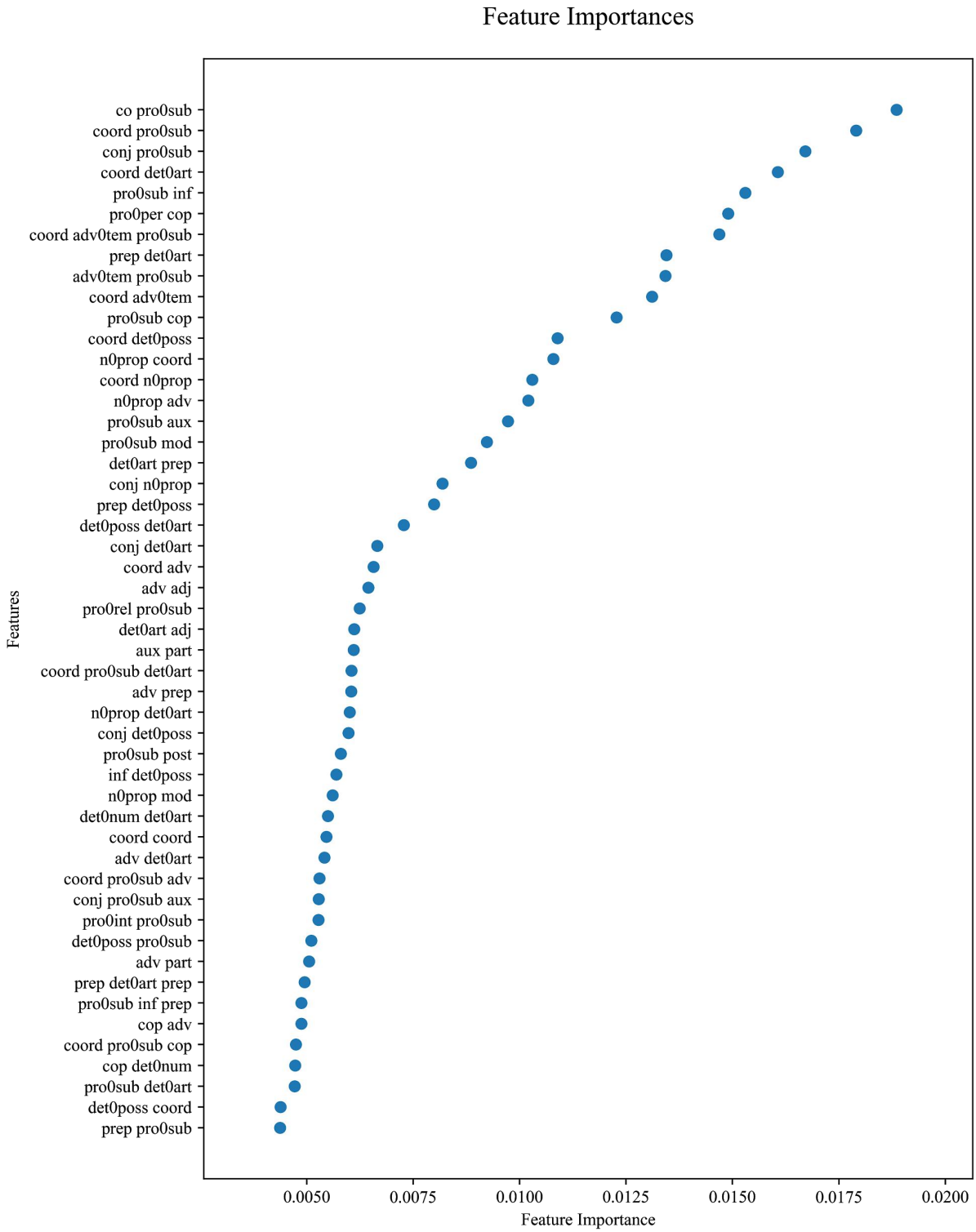


Figure 6. Feature importance plot of the 50 most important POS bigrams, *the first plot* is for predicting Language Impairment, and *the second* is for predicting Age.



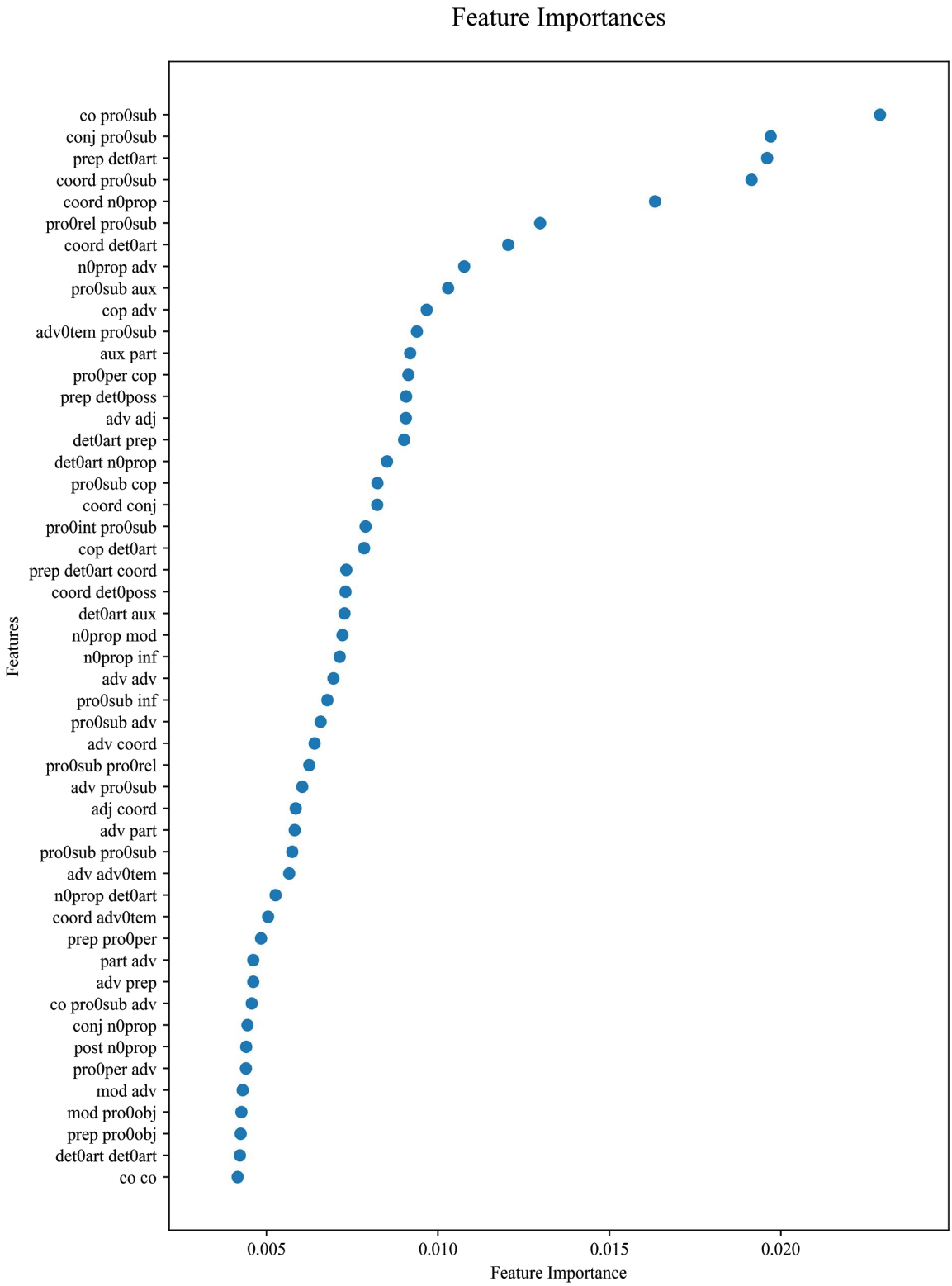

Feature Importances
co pro0sub
conj pro0sub
prep det0art
coord pro0sub
coord n0prop
pro0rel pro0sub
coord det0art
n0prop adv
pro0sub aux
cop adv
adv0tem pro0sub
aux part
pro0per cop
prep det0poss
adv adj
det0art prep
det0art n0prop
pro0sub cop
coord conj
pro0int pro0sub
cop det0art
prep det0art coord
coord det0poss
det0art aux
n0prop mod
n0prop inf
adv adv
pro0sub inf
pro0sub adv
adv coord
pro0sub pro0rel
adv pro0sub
adj coord
adv part
pro0sub pro0sub
adv adv0tem
n0prop det0art
coord adv0tem
prep pro0per
part adv
adv prep
co pro0sub adv
conj n0prop
post n0prop
pro0per adv
mod adv
mod pro0obj
prep pro0obj
det0art det0art
co co
Features
0.005
0.010
0.015
0.020
Feature Importance

Figure 7. The Independent Conditional Expectation (ICE) plots of the five most important bigrams/trigrams for Impairment.

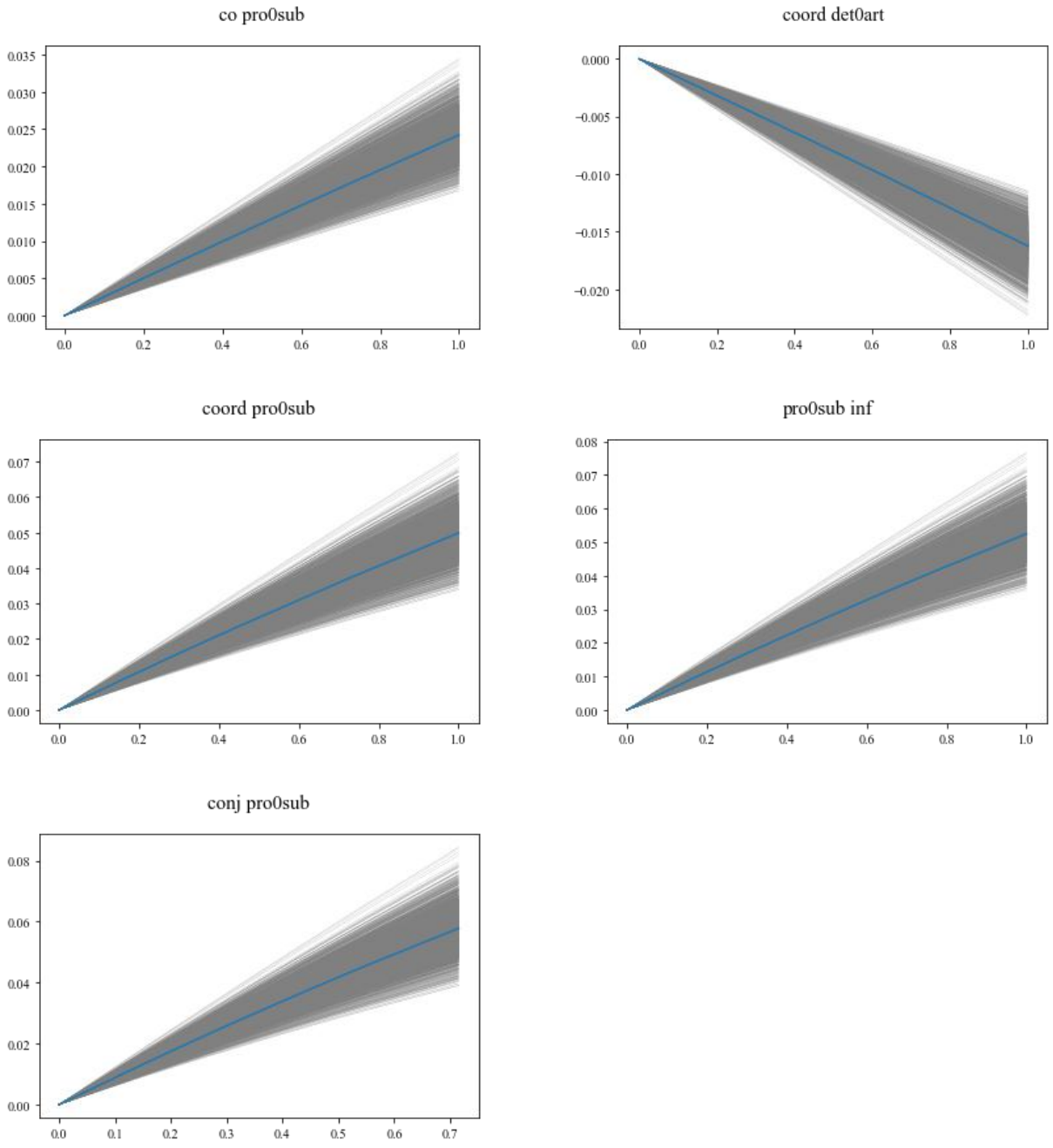


Figure 8. The Independent Conditional Expectation (ICE) plots of the five most important bigrams/trigrams for Age.

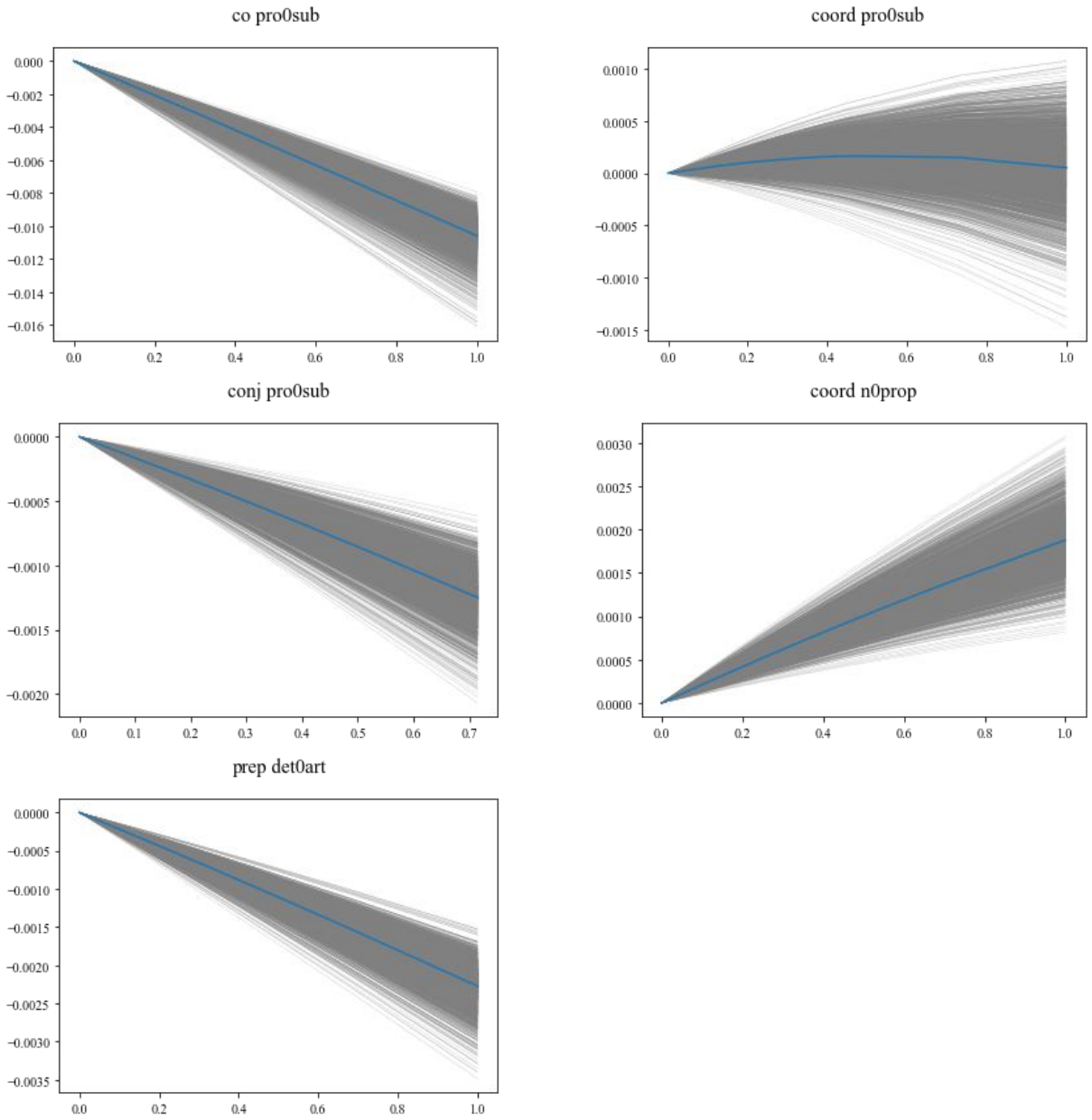


Table 1. Test statistics for the correlation between the proportional use of interrogative and relative pronouns and Age, separatedly analyzed by Language Impairment conditions.

	Coefficient	Standard error	t statistics	p-value
	Impaired Language Abilities			
Interrogative pronouns	67.71	21.93	3.09	.002**
Relative pronouns	59.46	20.29	2.93	.004**
	Typical Language Abilities			
Interrogative pronouns	80.88	15.88	5.09	.000***
Relative pronouns	83.41	12.41	6.72	.000***

Table 2. Validation loss and accuracy using three models with LSTM architecture for predicting Language Impairment.

	Model 1 - GloVe	Model 2 - CBOW	Model 3 - SG
Parameters	LSTM architecture with 1 hidden layer Number of nodes per layer = 64 Number of epochs = 20 Batch size = 512		
Maximum validation accuracy	0.78	0.78	0.78
Minimum validation loss	0.61	0.66	0.63