UNIVERSITY OF CALIFORNIA SAN DIEGO

Reading Comprehension Analysis and Prediction Based on EEG and Eye-Tracking Techniques

A Thesis submitted in partial satisfaction of the
requirements for the degree
Master of Science

in

Bioengineering

by

Qin Li

Committee in charge:

Professor Tzyy-Ping Jung, Chair
Professor Gert Cauwenberghs, Co-Chair
Professor Vikash Gilja

2021

The thesis of Qin Li is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2021

DEDICATION

Dedicated to Professor Tzyy-Ping Jung and Professor Gert Cauwenberghs, who taught and guided me in the research field, who encouraged me, offered invaluable suggestions when I felt lost, and who carefully and patiently observed my progress in the research projects.

Many appreciations to my family, who financially, spiritually, and emotionally supported my studies in the last five years. In the past decades, they always encouraged me to achieve my goals and taught me to take care of myself.

And lastly, I would really appreciate all my friends and every teacher in my past life, who accompanied, taught, helped, and encouraged me till today.

EPIGRAPH

For the road was so far

and so distant was my journey,

And I wanted to go up and down,

seeking my heart's dream.

– Qu Yuan

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to acknowledge Professor Tzyy-Ping Jung for his support as the chair of my committee and acknowledge Professor Gert Cauwenberghs for his support as the co-chair of my committee. Their support and guidance were essential in the completion of this work and promoted my success in future research projects.

I would like to acknowledge Professor Vikash Gilja, who taught me many useful skills for my research and offered valuable suggestions in this project.

I would like to acknowledge the SCCN members, who helped me to learn better about my research field and offered me tremendous suggestions and insights during lab meetings.

ABSTRACT OF THE THESIS

Reading Comprehension Analysis and Prediction Based on EEG and Eye-Tracking Techniques

by

Qin Li

Master of Science in Bioengineering

University of California San Diego, 2021

Professor Tzyy-Ping Jung, Chair
Professor Gert Cauwenberghs, Co-Chair

Research in reading comprehension traditionally relied on the experimental setting of word-by-word presentation which eventually revealed many neural biomarkers as well as establishing the basis of modern reading research. Since the development of brain-computer techniques and computational methods in the past decades, it has become possible to study reading comprehension in natural settings. This study used a variety of advanced technologies to analyze a dataset collected by the ZuCo group regarding reading comprehension. With natural language processing tools, we extracted the words essential for understanding sentences, and identified eye-tracking patterns that relate to these words. Using the EEG time series and frequency series,

we also looked at neural patterns associated with those words and tried to build up a statistical model and neural network model that predicted the linguistic patterns. Consequently, the study is likely to provide new insights into future cognitive linguistics and brain-computer interaction research, which may help advance reading-aid technologies.

Chapter 1

Introduction

Reading comprehension has been studied at various levels in several fields, including vision, attention, neural correlates, and lexical semantics. Because of limitations in both techniques and recording equipment, prior studies often used word-by-word presentations, which required the participants to maintain their gaze on a fixed space of the screen [1]. Many neurobiomarkers associated with reading comprehension were discovered in the 1980s, such as P300 and N400 [2]. The word-to-text integration theory attempted to explain reading comprehension through the development of a mental model in the memory space [3, 4], among many studies that examined the process of integrating single words into a complete sentence in the neural system. They assumed that every word in a sentence is used to update the working memory, which creates mental representations of words [4]. As a reader browsed through all the words, these representation units interacted and collaborated to form a mental model of the sentence [4]. Perfetti's experiment further confirmed this theory, indicating that the event-related potential (ERP), N400, occurs when the explicit hyperlink of one word is directly linked to the word in the preceding text [3]. The correlation between N400 and semantic information processing was also evident in Kutas' research [1]. Later studies further indicated a significant association between P300 and the updates of information in the working memory during reading [5]. While these studies with

word-by-word presentations provided much information concerning reading comprehension, their applications were limited due to restrictions in eye movements and reading styles.

The development of computation algorithms and brain-computer interface technologies, electroencephalography (EEG), eye-tracking, signal processing algorithms, and natural language processing (NLP) techniques enabled reading comprehension studies in a natural setting. In an experiment setting like this, the subjects were free to move their eyes freely over the sentences and paragraphs on the screen and follow their own reading style [6]. Eye motions and EEG signals were simultaneously recorded, allowing us to track eye fixation upon each word and extract neural signals corresponding to them. By integrating eye-tracking and EEG technologies, it has become possible to study the natural reading process and develop a real-time reading monitor and study aid. NLP tools allowed for greater flexibility in reading text design and analysis, which made it possible to translate everyday reading materials into computer-interpretable information.

Due to the difficulty of gathering data during the pandemic, the Zurich Cognitive Language Processing Corpus (ZuCo) dataset [6] was used in this study. Natural language processing tools, including WordNet [7], Global Vector [8], and fastText [9] to extract the significant words in sentence comprehension. Then, we tried to extract the eye-tracking and EEG biomarkers at word-level associated with reading comprehension.

1.1 ZuCo Dataset

The Zurich Cognitive Language Processing Corpus dataset collects the eye motions, and EEG signals during the natural reading settings [6]. A 128-channel EEG Geodesic Hydrocel system collected the EEG and an infrared video-based eye tracking device captured the eye position with [6]. The participants were twelve healthy adults aged 22 to 54 with their first language as English [6]. All participants approved of their participation and the disclosure of the collected data from the experiments [6] and University of Zurich Ethics Commission approved

the study [6]. Three reading tasks were divided into two sessions of 2-3 hours each at the same time of day [6].

For our study, the task-specific reading data was selected from the three reading tasks. During the entire task-specific reading, 407 sentences with 8284 words were presented. The 407 sentences were arranged from top to bottom in blocks that shared the same meaning and the sequence is shown in Table 1.1. Before every block began, the screen would display the definition of a keyword and asked the subject to determine whether the following sentence contains the relationship with the keyword and use the keyboard to enter an answer "[1]=yes" or "[2]=no" [6]. Each block would display three example sentences at the beginning ahead for practice [6]. During the practice, the eye-tracker was calibrated [6]. The sentences were displayed in the middle of the screen, with the question at the right bottom corner [6].

Table 1.1: The number of sentences per keyword.

keyword	# Sentences	# Controls
award	38	7
education	37	8
employer	38	9
founder	34	7
job title	38	8
nationality	38	8
political affiliation	36	8
visited	38	10
wife	38	7

1.2 Natural Language Processing in this Project

In this project, we started our analysis by analyzing the texts displayed to the subjects. In order to transform the texts into computationally meaningful information, we studied concepts in linguistics and applied natural language processing tools. In this study, we first identified the words that were significant in interpreting the sentence meaning, as well as the words that were

least relevant. We used stop words as our first identification method. Stop words refer to the most common words in a language, which are usually filtered out during the natural language process [10]. Stop words in English include “he”, “I”, “that”, and “run,” which are very common.

Our next step was to explore the words that are significance in the successful interpretation of the relationship between the sentence and the keyword. We used two major categories of NLP tools, the WordNet lexical databases and word vectorization representation models. Using those models, we calculated the semantic similarity between the words, since it is considered a significant measurement of the relationship between words [8].

1.2.1 WordNet Lexical Databases

A lexical database, WordNet, is used for computational analysis and machine learning, where English words are linked together based on their semantic meaning and organized into synonym sets [7]. An example of how WordNet taxonomy is organized is given in Figure 1.1. As the figure note illustrates, each taxonomy will have a root word, node words, and paths connecting the words. Many methods have been developed to interpret word relationships from this dataset. A quantifiable characteristic is semantic similarity, which indicates how similar two words are within the WordNet. The current method of calculating semantic similarity is either based on the path distance between two English words [11] or on the information content between two words [12, 13].

1.2.2 Word Vectorization Models

Recent advances in computing power and technologies have enabled neural networks to establish vector representations for English words [8, 9]. The word vector representations may differ according to the algorithms and training datasets, so when applying the models, the model bias should be carefully evaluated.

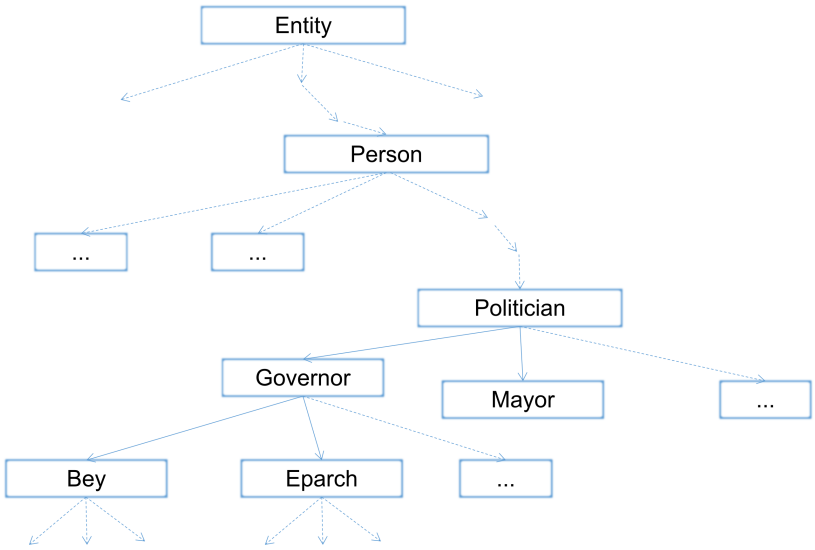

Figure 1.1: An example of WordNet connections: Words are mapped and linked by their conceptual-semantic and lexical relations [7]. The root of this taxonomy is "Entity". In this figure, each word represents a node, connected by a path. As an example "Politician" is the least common super-concept of "Eparch" and "Mayor".

The Global Vector (GloVe) model was developed by Pennington et al. in 2014 at Stanford University [8]. The vector representation was obtained with an unsupervised learning algorithm trained on large linguistic datasets, such as Wikipedia [8]. Words were mapped into the 300-dimensional vector space, based on the global word-word co-occurrence statistics [8]. By establishing vector representations of words, the distance between words will be determined by their semantic similarity [8].

FastText is another word-vector model developed by Facebook's AI Research Lab based on the studies of Bojanowski et al. [9], and Joulin et al. [14]. The algorithms from fastText enable users to apply supervised or unsupervised learning to obtain word vector representations. Both models have pre-trained word vector representations available online, so we could use these representations to extract the semantic relationships between words.

Chapter 2

Word-Level Semantic Analysis and Eye-Fixation Biomarkers

2.1 Introduction

In ZuCo's experiment, the subjects were asked to determine whether the sentences contain a relationship with the keyword in the question [6]. By ZuCo, 82.3% of the sentences contained the relationship, and the answer correctness rate was above 90% on average [6]. As a result, it was very difficult to study the cases of comprehension versus miscomprehension of the relationship. As an alternative, we would like to determine the eye-tracking and EEG features associated with detecting the relationship during sentence reading.

In the first step, we would try to identify the words that were significant in sentence comprehension. We assumed that words that are semantically related to the question are those which are significant in a correct sentence comprehension. Based on this assumption, we calculated the semantic similarity between the words in the each and the keyword in the question using NLP tools. The calculated semantic similarity is used to represent the semantic significance of the words in comprehending the sentence. There were seven different methods used to calculate

similarity scores, and we compared their selection of high similarity words. We then identified the words with the higher similarity scores.

Considering the bias in the datasets and the divergence of algorithms, we needed to find a common ground among the methods. In this case, we defined a concept, similarity level (SL), which is how many methods agreed that a word has a similarity score of one standard deviation above the mean. The statistics of eye-fixation data were then compared to determine the threshold of SL to pick up the high-similarity words shared by the majority. It was tentatively planned that if four or more methods independently indicated that a word had a similarity score of one standard deviation above the mean, it would be considered a high-similarity word. With this standard, we obtained a group of high-similarity words, and the rest will be considered less relevant to the keyword.

In addition, we examined the statistics of eye-fixation features for words with different similarity scores. Using the eye-fixation results, we can then sort the words into high-similarity and low-similarity groups.

2.2 Methods

2.2.1 WordNet-Based Word Similarity Calculations

A path-based method to measure semantic similarity was proposed by Wu and Palmer in 1994 and named Wu-Palmer similarity [11]. The Wu-Palmer method represents a word as a concept, C. Since a word could have multiple meanings, the meanings would be assigned to different conceptual domains that can be interpreted as different taxonomies. The conceptual similarity between any two concepts, C_1 and C_2 can be expressed as follows [11]:

$$Sim_{WP}(C_1, C_2) = \frac{2N_3}{N_1 + N_2 + 2N_3} \tag{2.1}$$

In this equation, we introduce a third concept, C_3, which is the least common super-concept of C_1 and C_2. In this equation, N_1 is the path length in terms of the number of nodes (or words) between C_1 and $C3$, N_2 is the path length between C_2 and C_3, N_3 is the path length from C_3 to root. For words with multiple meanings, the semantic similarity can be calculated as follows [11]:

$$WordSim(W_1, W_2) = \sum_i \alpha_i Sim_{WP}(C_{i,1}, C_{i,2}) \tag{2.2}$$

where W_1 and W_2 are two words and α_i is the weight for each conceptual domain.

Lin, in 1998, developed a method to calculate word similarity [13] based on Resnik's similarity measure [12]. The Lin similarity is determined by the information contents of each word. For two words, W_1 and W_2, the calculation can be expressed as [13]:

$$Sim_{Lin}(W_1, W_2) = \frac{2IC(W_1, W_2)}{IC(W_1) + IC(W_2)} \tag{2.3}$$

$$IC(W) = -\sum_{f \in F(W)} logP(f) \tag{2.4}$$

$$IC(W_1, W_2) = -\sum_{f \in F(W_1) \bigcap f \in F(W_2)} logP(f) \tag{2.5}$$

IC refers to the information content, which describes the amount of information in a set of features for each word. $F(W)$ represents the set of features for one word, which are defined and summarized into a feature vector by this method. The Lin similarity score would increase if the feature sets overlapped more. We would expect a Lin similarity of 1 if the features of the two words match completely.

2.2.2 Cosine Similarity

To estimate the similarity between any two vectors in a vectorized system, we can use cosine similarity:

$$Sim(\mathbf{V_1}, \mathbf{V_2}) = cos(\theta) = \frac{\mathbf{V_1}\mathbf{V_2}}{\|\mathbf{V_1}\|\|\mathbf{V_2}\|} \tag{2.6}$$

The vector representations of two words are $\mathbf{V_1}$ and $\mathbf{V_2}$. θ represents the angle between the two vectors. The similarity score will increase if two vectors are closer in the vector space. The cosine similarity between two vector representations would equal 1 when they completely overlapped.

In this study, the Global Vector (GloVe) [8] and fastText [9] algorithms are applied to extract the similarity between each word and the keyword in the question. The pre-trained GloVe model on the Common Crawl dataset was available from `https://nlp.stanford.edu/projects/glove/`. The fastText models are pre-trained on the wiki-news dataset, and Common Crawl dataset with or without considering subword information [15]. The pre-trained word vector representations of fastText are available from `https://fasttext.cc/docs/en/english-vectors.html`. These learned vector representations were used to calculate the cosine similarity between each word in the sentence and the keyword in the question.

2.2.3 Evaluation of Word Similarity

For each word in the sentence, similarity scores were calculated based on seven methods: 1) Wu-Palmer similarity based on WordNet, 2) Lin similarity based on WordNet, 3) cosine similarity based on GloVe pre-trained on Common Crawl, 4-5) cosine similarity based on fastText pre-trained on wiki-news with or without subword information, and 6-7) cosine similarity based on fastText pre-trained on Common Crawl databases with or without subword information. The seven calculation methods were compared in terms of their selection of high-similarity words. A comparison of the methods was made and we attempted to find the high-similarity words that they commonly agreed on.

Considering the datasets' biases and the algorithms' divergence, we needed to find the common space among all the methods in the word that they found similar with the question keyword. In this case, we defined a concept, similarity level (SL), which is how many methods agreed that a word has a similarity score of one standard deviation above the mean. A threshold was set up to pick up the high-similarity words that most methods agreed. The tentative plan was that if more than four methods independently indicated that a word has a similarity score of one standard deviation above the mean, it is considered a high-similarity word. Based on this standard, we obtained a group of high-similarity words, and the rest will be considered irrelevant to the keyword.

2.2.4 Eye-Tracking Data Preprocessing

The saccades were detected by measuring the velocity and acceleration of the eye movements [16]. The acceleration threshold was set at $8000°/s^2$, velocity threshold at $30°/s$, and deflection threshold at $0.1°$[6]. The periods between saccades corresponded to the fixations [16]. The eye-fixations with a duration below 100 ms were discarded since they are less likely to reveal reading-relevant information [17]. The eye-fixations were recorded as two-dimensional coordinates, which were then compared to the word boundaries to determine which word the eyes stared upon [6].

2.3 Results

Among the 8284 words, each method found five hundred to ten hundred high-similarity words, summarized in Table 2.1. Across each two methods, 20% to 89% of high-similarity words are shared. For the methods based on word vectorization models, 491 to 722 high-similarity words can be shared with a second method. Between the two WordNet-based methods, 410 words worth of high-similarity were shared. The vector-based methods and the WordNet methods shared

Table 2.1: The number of high-similarity words found by each method (in gray) and shared by each two methods (others).

#Words with a high similarity score shared per method	fastText wiki	fastText crawl	fastText wiki subword	fastText crawl subword	GloVe wiki	Wu-Palmer	Lin
fastText wiki	663	—	—	—	—	—	—
fastText crawl	512	584	—	—	—	—	—
fastText wiki subword	589	518	1003	—	—	—	—
fastText crawl subword	568	531	722	813	—	—	—
GloVe score	506	491	612	625	963	—	—
Wu-Palmer	286	234	274	280	251	919	—
Lin	207	166	200	195	199	410	594

Table 2.2: For each similarity level, the mean and standard deviation of each characteristic are displayed in the table.

	Counts	#Stop Words	Word Length	Word Length (no SWs)	#Fix	Prob. ≥ 1 Fix	t1 (sec)	t2 (sec)	t3 (sec)
SL=0	6092	2374	4.63±2.68	5.91±2.63	0.70±0.55	0.46±0.27	224±88	234±102	249±116
SL=1	1045	397	4.79±2.47	6.16±2.10	0.73±0.50	0.50±0.26	223±86	231±106	243±107
SL=2	395	121	5.40±2.74	6.64±2.38	0.76±0.48	0.52±0.26	223±87	227±93	252±114
SL=3	168	95	4.77±2.45	6.96±2.30	0.69±0.51	0.47±0.26	225±86	239±99	256±119
SL=4	99	1	7.42±2.64	7.46±2.63	1.17±0.58	0.69±0.21	228±82	236±108	258±126
SL=5	244	1	7.68±2.19	7.70±2.17	1.12±0.55	0.70±0.22	229±92	234±106	248±113
SL=6	103	0	6.70±1.72	6.70±1.72	1.11±0.53	0.66±0.19	225±79	218±84	224±68
SL=7	138	1	6.22±2.01	6.24±2.00	1.04±0.48	0.67±0.20	231±90	234±96	249±113

166 to 286 high similarity words.

Table 2.2 displayed the statistics for each similarity level. The majority of words (n=6092) had SL$=0$, and very few words (n=138) had SL$=7$. There were 1045 words with SL$=1$, which is the next largest set.

We then counted the stop words for each score calculation. Each method assigned a high similarity score to 20 to 306 stop words (mean=134, std= 120). 1 to 103 stop words were shared by two methods or more. The majority of stop words in the sentences had SL≤ 3. Each set of SL≤ 3 had $30\% - 57\%$ words that were stop words. There were no or one stop words in every set of SL≥ 4. Before and after removing the stop words, different word lengths were observed across the SLs. For SL≤ 5, we can observe an increase trend in the average word length. We will

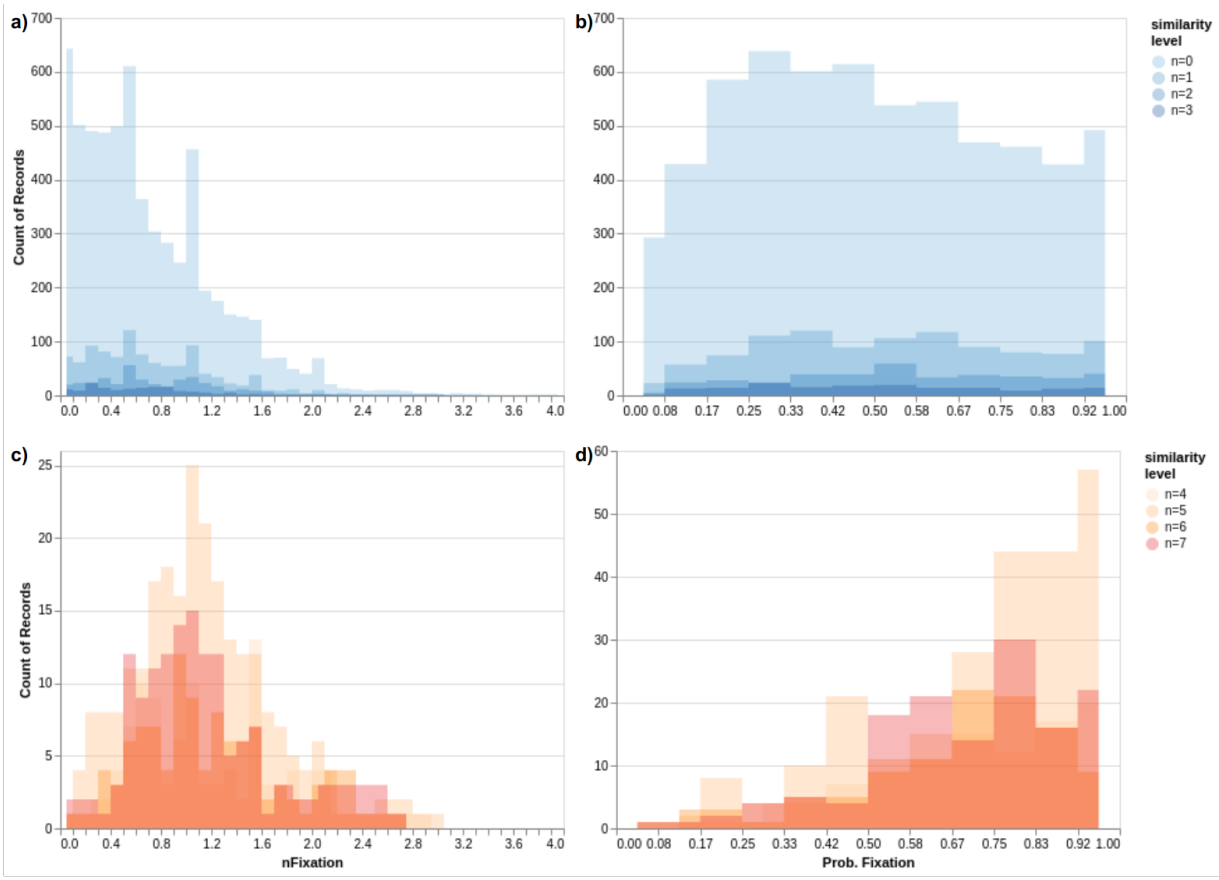

Figure 2.1: Histogram of the counts of eye-fixations for each similarity level. A) and c) are for the number of fixations per word. B) and D) are for the probability of getting one or more eye-fixation.

also discuss the statistics for the eye-fixations and durations in the following figures.

In Figure 2.1, we plotted the distribution of eye-fixation numbers and the probability of eye-fixations, grouped by the SL. Table 2.1 contained the mean and standard deviations of these distributions. Figure 2.1a) showed that the eye-fixations on a word with an SL of ≤ 3 were skewed toward the left. Averaged across all words (n=7700) of the four SL groups, the mean was 0.70, and the standard deviation was 0.54. Figure 2.1b) showed that the probability of at least one eye-fixation on a word with an SL ≤ 3 was spreading along the axis. Figure 2.1c) showed that the number of eye-fixations on a word with an SL ≥ 4 was nearly a normal distribution. For the four groups, the mean was 1.10, and the standard deviation was 0.54 across all words (n=584).

Table 2.3: P-values for comparing two similarity levels in their numbers of eye-fixations per word. Red highlights indicate those below the threshold ($\alpha = \frac{0.01}{28}$).

	SL=0	SL=1	SL=2	SL=3	SL=4	SL=5	SL=6	SL=7
SL=0	1.00E+00	—	—	—	—	—	—	—
SL=1	4.91E-02	1.00E+00	—	—	—	—	—	—
SL=2	1.20E-02	2.93E-01	1.00E+00	—	—	—	—	—
SL=3	9.83E-01	4.21E-01	1.66E-01	1.00E+00	—	—	—	—
SL=4	2.24E-12	6.20E-11	2.01E-09	2.66E-10	1.00E+00	—	—	—
SL=5	1.37E-26	7.98E-22	1.83E-16	7.48E-15	5.24E-01	1.00E+00	—	—
SL=6	5.44E-12	2.45E-10	1.29E-08	1.96E-09	4.54E-01	8.06E-01	1.00E+00	—
SL=7	1.48E-13	5.27E-11	2.02E-08	6.19E-09	7.02E-02	1.10E-01	2.87E-01	1.00E+00

Table 2.4: P-values for comparing two similarity levels in their probability to get one fixation or more per word. Red highlights indicate those below the threshold ($\alpha = \frac{0.01}{28}$).

	SL=0	SL=1	SL=2	SL=3	SL=4	SL=5	SL=6	SL=7
SL=0	1.00E+00	—	—	—	—	—	—	—
SL=1	1.96E-05	1.00E+00	—	—	—	—	—	—
SL=2	9.00E-06	1.43E-01	1.00E+00	—	—	—	—	—
SL=3	6.63E-01	1.91E-01	3.48E-02	1.00E+00	—	—	—	—
SL=4	3.84E-19	1.30E-14	4.80E-11	4.89E-13	1.00E+00	—	—	—
SL=5	2.11E-42	8.50E-30	4.25E-19	4.81E-18	8.00E-01	1.00E+00	—	—
SL=6	4.16E-17	4.90E-12	2.17E-08	1.78E-10	2.05E-01	7.74E-02	1.00E+00	—
SL=7	4.39E-22	2.17E-15	1.98E-10	3.63E-12	3.13E-01	1.34E-01	7.43E-01	1.00E+00

Table 2.3 indicated that the number of eye-fixations on a word with any SL ≤ 3 was statistically different from that of any SL ≥ 4. Also, any SL ≤ 3 did not statistically differ from each other in terms of the number of eye-fixations on a word. So did all SLs ≥ 4. Table 2.4 indicated a similar conclusion that the probability of at least one eye-fixation on a word with any SL ≤ 3 was statistically different from the words with any SL ≥ 4. Both SL $= 1$ and SL $= 2$ had a statistically significant difference from SL$= 0$.

We then checked the statistics of the eye-fixation duration and plotted the distribution in Figures2.2 a) and c). The shape of all the similarity levels looked similar. According to Table 2.5, no statistical significance existed across any two similarity levels. Considering the difference in the reading speed per subject, we plotted the distribution normalized to the average first-fixation duration for each subject in Figures 2.2 b) and d). The distribution was more centralized, and Table 2.6 indicated that SL$= 5$ was statistically different from SL$= \{0, 1, 2\}$. By performing a

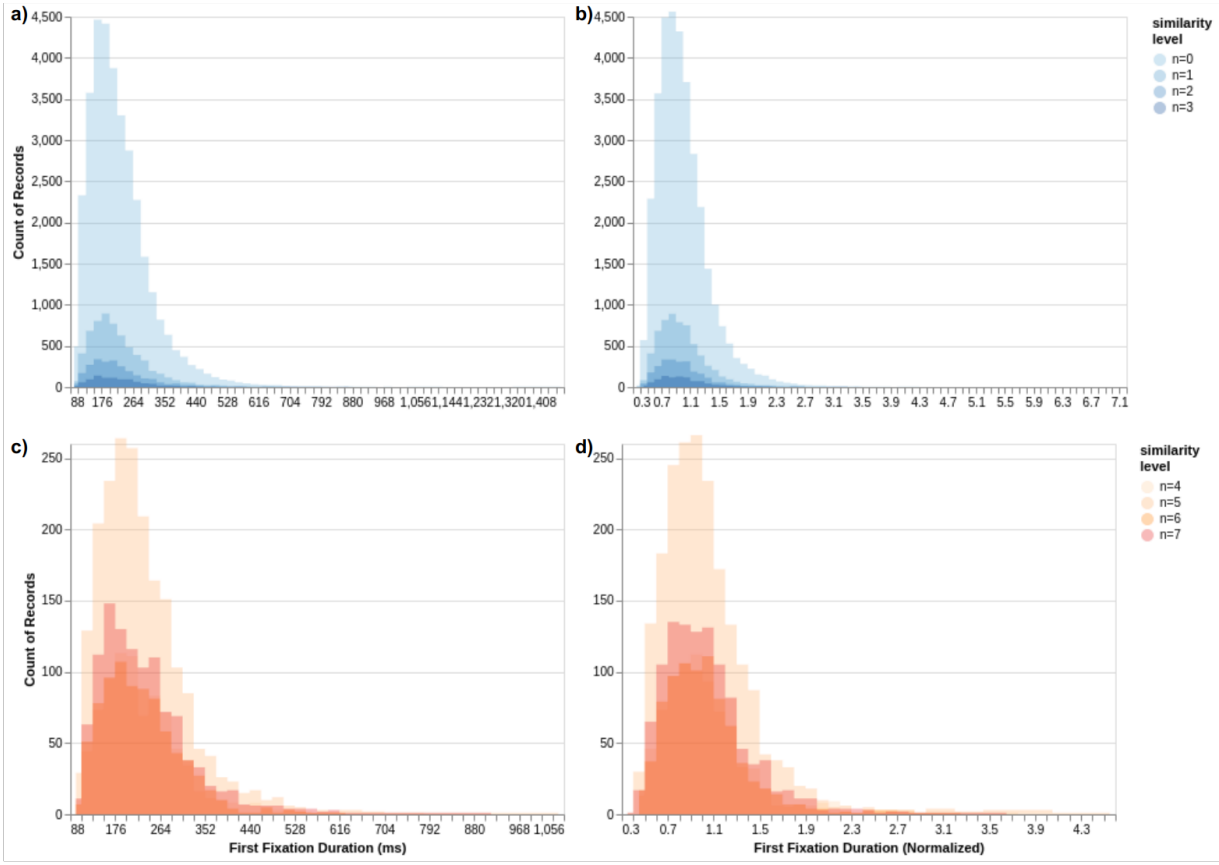

Figure 2.2: Histogram of the duration of eye-fixation for each similarity level. A) and c) are for the first eye-fixation duration per word. B) and D) are for the normalized eye-fixation duration per word.

t-test on the two groups SL$\leq$ 3 and SL$\geq$ 4, we got a p-value of 2.1×10^{-4} for the first-fixation duration and a p-value of 4.5×10^{-8} for the normalized duration. Those p-values indicated a statistical significance between SL$\leq$ 3 and SL$\geq$ 4 on the eye-fixation duration.

Table 2.5: P-values for comparing two similarity levels according to their first fixation duration per word.

	SL=0	**SL=1**	**SL=2**	**SL=3**	**SL=4**	**SL=5**	**SL=6**	**SL=7**
SL=0	1	—	—	—	—	—	—	—
SL=1	4.41E-01	1	—	—	—	—	—	—
SL=2	4.17E-01	7.84E-01	1	—	—	—	—	—
SL=3	6.07E-01	4.29E-01	3.73E-01	1	—	—	—	—
SL=4	1.69E-01	1.09E-01	1.03E-01	5.25E-01	1	—	—	—
SL=5	1.42E-02	8.81E-03	1.40E-02	2.90E-01	7.51E-01	1	—	—
SL=6	6.62E-01	4.71E-01	4.09E-01	9.53E-01	4.86E-01	2.57E-01	1	—
SL=7	8.46E-03	5.19E-03	6.84E-03	1.35E-01	4.04E-01	5.20E-01	1.18E-01	1

Table 2.6: P-values for comparing two similarity levels according to their normalized first fixation duration per word. Red highlights indicate those above the threshold ($\alpha = \frac{0.01}{28}$).

	SL=0	**SL=1**	**SL=2**	**SL=3**	**SL=4**	**SL=5**	**SL=6**	**SL=7**
SL=0	1	—	—	—	—	—	—	—
SL=1	8.35E-01	1	—	—	—	—	—	—
SL=2	4.30E-01	5.66E-01	1	—	—	—	—	—
SL=3	6.29E-01	5.90E-01	3.98E-01	1	—	—	—	—
SL=4	2.73E-02	2.94E-02	1.91E-02	1.99E-01	1	—	—	—
SL=5	7.30E-05	2.07E-04	2.70E-04	4.51E-02	6.30E-01	1	—	—
SL=6	3.33E-01	3.18E-01	2.05E-01	7.33E-01	3.37E-01	1.03E-01	1	—
SL=7	9.04E-04	1.30E-03	1.03E-03	4.66E-02	5.28E-01	8.14E-01	9.76E-02	1

2.4 Conclusion

In this chapter, we analyzed all the words using NLP tools and examined eye-fixation statistics. We first calculated the semantic similarity between the words in the sentences and the keyword in the question. Each method assigned a high similarity score to around six hundred words or more. However, only 1.7% of words (n=138) obtained a high similarity score by every method. This number is very small, compared to 407 sentences in total. Therefore, we need to determine looser criteria for a high similarity.

A large percentage of stop words is present in every set of SL≤ 3. Despite 616 stop words getting a high similarity score by one or more methods, only 219 of them got a high score by two methods or more. Interestingly, each calculation method tended to assign tens or hundreds of stop words a high similarity score, but the other methods usually disagreed. As SL increases, the word length in the set increases. After SL goes above 4, there is no statistically significant difference between the word lengths in each set of SL.

Based on our analysis of eye-fixation statistics, we observed statistically significant differences between any set of SL≤ 3 and any set of SL≥ 4. Compared to the sets of SL≤ 3, a higher number of fixations and a higher probability of at least one eye fixation were observed in the sets of SL$/geq4$. First, second, and third fixation durations were not statistically significant among any two sets of SLs. After we tried normalizing the fixation duration to the average fixation duration of each subject, we discovered that the set of SL$= 5$ was different from SL$= 0,1,2$.

Thus, we found that the words in the sets of SL≤ 3 were statistically different from the words in the sets of SL≥ 4. Prior studies have also shown that the content that is critical for comprehension attracts the most attention and thus more eye-fixations [18]. Thus, it was probably that words with SL≥ 4 have a significant impact on sentence comprehension. This is evident that a threshold of SL ≥ 4 could be used to select the words with a higher relevance to the sentence comprehension.

Next, we separated the words into two groups, and the statistically significant features of each group are shown in Table 2.7. Among the 8284 words, 584 words were classified as high SL (SL$\geq$ 4) and the rest as low SL (SL$\leq$ 3) to the keyword in the question. There were significantly more eye-fixations for high-SL words (1.11 vs. 0.70) in the table. These words were more likely to get eye-fixations upon them (0.68 vs. 0.47). These words also had a longer first eye-fixation duration after normalizing the time with each subject's average fixation duration (1.03 vs. 1.00). We were going to explore more features of the two groups of words in the next chapter.

Table 2.7: The table summarizes the significant differences between high similarity-level words and low similarity-level words.

	Counts	**#Stop Words**	**string length**	**string length (no SW)**	**#Fixation**	**Prob. Eye-Fix**	**t1-norm**
Low SL	7700	2987	4.69±2.65	6.00±2.56	0.70±0.54	0.47±0.27	1.00±0.37
High SL	584	3	7.12±2.24	7.14±2.23	1.11±0.54	0.68±0.21	1.03±0.38
Total	8284	2990	4.86±2.70	6.13±2.55	0.73±0.55	0.49±0.27	1.00±0.37
p-vale			3.54E-99	4.14E-28	7.14E-57	8.42E-89	4.52E-08

Chapter 3

EEG-biomarker Analysis

3.1 Introduction

Following the low-SL and high-SL group division in Chapter 2, we analyzed the EEG signal time-locked to the eye-fixations upon each group of words, in both the time domain and in the frequency domain. Then, we used statistical models and a simple convolutional neural network to learn and predict the two groups of words based on the EEG signals.

3.2 Methods

3.2.1 EEG Data Preprocessing

ZuCo collected the EEG signals from the subjects with a 128-channel EEG Geodesic Hydrocel system [6]. The EEG signals was recorded at a sampling rate of 500Hz with a bandpass set at 0.1 to 100 Hz and all electrodes were referenced to channel Cz [6]. 105 scalp electrodes were retained in the dataset, while those for EOG, facial, and neck signals were used for artifact removal and were discarded from the preprocessed dataset [6]. In their paper [6], the ZuCo dataset was already preprocessed using the Automagic protocol (version: 1.4.6) available at

https://github.com/methlabUZH/automagic. They identified and replaced bad electrodes based on an EEGLab script *clean_rawdata.m* from http://sccn.ucsd.edu/wiki/Plugin_list_process [6, 19]. Using EEGlab script *pop_eegfiltnew.m* [19], a 0.5 Hz high-pass filter was applied, followed by a 49-51 Hz notch filter [6]. They performed artifact removal using linear regression, ICA, and multiple artifact rejection algorithms [6].

Prior to our analysis, we discarded EEG signals associated with eye-fixations short than 100 ms since previous studies have shown that fixations shorter than this duration does not reveal reading-relevant information [17]. We also extracted 200 ms of signals before fixation onset and 800 ms after for each fixation-locked EEG signal. A statistical analysis was performed to compare the difference between the two groups.

Independent frequency bandpass filters were performed on the entire EEG recording to extract the time-series for the frequency bands of theta (4-7 Hz), alpha (7.5-13 Hz), beta (13.5-30 Hz), and gamma (30.5-50 Hz) with the EEGLab script *pop_eegfiltnew.m* [17]. To each of these time series, a Hilbert transformation was applied. For each frequency, we averaged the Hilbert series amplitude within the eye-fixation time window as the band power. This method of frequency band power calculation was equivalent to the fast Fourier transformation based calculation [20]. To compare the overall and dynamic EEG frequency patterns time-locked to eye fixations, the band power under the following conditions was calculated: 1) within the entire fixation window; 2) within each 20-ms window from the fixation onset to 300 ms later. 300 ms was selected since 86.7% of the total fixations were between 100 ms and 300 ms. In each time window, scalp topography was plotted as a quad-layered topographic image of 32x32 pixels for all four frequency bands combined. Machine learning methods and neural networks would be used to analyze the images.

3.2.2 Feature Space Projection

In order to learn the representation of the high-dimensional data, dimension reduction techniques were applied to extract representative features in the lower dimensions. To balance out the high amplitude of alpha and gamma bands, the band power for each frequency band was normalized using the standard score. Principal component analysis (PCA) and factor analysis (FA) were independently applied to project the band power for all the channels to 50 components, which preserves 95.5% of the variance in the PCA space. By linear discriminant analysis (LDA), those components were then projected into the subspace that characterizes two classes (high SL vs. low SL). This LDA model was trained with 4-fold validation within and across subjects and established a linear decision boundary between the two groups. The LDA model was used to classify the two classes based on Bayes' rule, and the test results were reported.

3.2.3 Simple Convolutional Neural Network

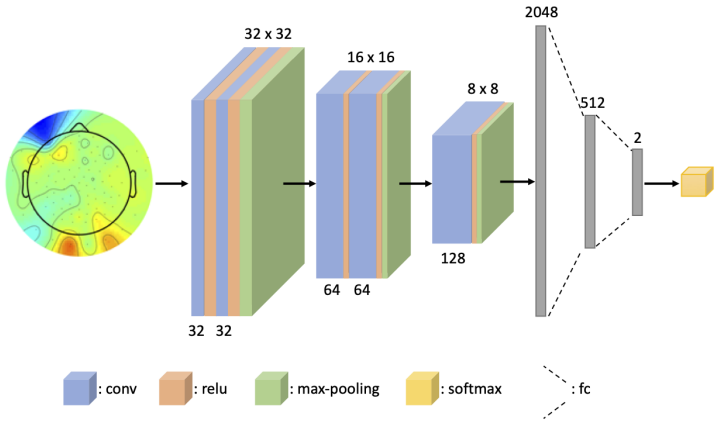

Figure 3.1: The simple convolutional neural network structure: Convolutional layers (conv), rectified linear units (ReLU), max-pooling layers, fully connected layers (fc) and softmax were used in this model.

The structure of the simple convolutional neural network was visualized in Figure 3.1, which was modified from EEGLearn's CNN model [21]. A topographic image comprised of

four layers of four colors (i.e., four frequency bands) was the input to the neural network. A 2D convolutional neural network was constructed using five convolutional layers, three max-pooling layers, and three fully connected layers, as shown in Figure 3.1. Tue CNN model would output the binary classification result of either high SL or low SL

3.2.4 Handling Class Imbalance

In Chapter 2, we extracted 584 words in the high-SL group, whereas there were 7700 in the low-SL group, which caused the class imbalance problem. In our preliminary study, we also found that the total fixations on the low-SL words were around ten times greater than those on the high-SL words. The two classes to be put into the training set were highly imbalanced. As a result, we examined multiple candidate strategies for the class imbalance problem and selected to under-sample the larger class. In order to decrease the sample size for the low-SL group, we first excluded all low-SL words with more than one fixation. We then used the random under-sampling method to decrease the number of samples in the larger class. The low-SL class was down-sampled to an equal size of the high-SL class.

3.3 Results

Figure 3.2 showed the fixation-related potentials (FRP) associated with high-SL and low-SL words. From this figure, we can see that some channels showed a lower voltage at 90 ms, 160 ms and 230 ms since the fixation onset for the fixation onto high-SL words. However, no time point in any channel passed a t-test with a threshold $\alpha = \frac{0.01}{105 \times 501} = 1.90 \times 10^{-7}$.

Figure 3.3 illustrated the FRPs associated with high-SL and stop words. Many channels experienced a decrease in voltage of 0.6 μV in the time ranges 150 ms - 170 ms and 210 ms - 240 ms for the high-SL group. Figure 3.4 showed the t-test result at each time point between the two groups. We zoomed in to the region of statistically significant points. For six or more time points,

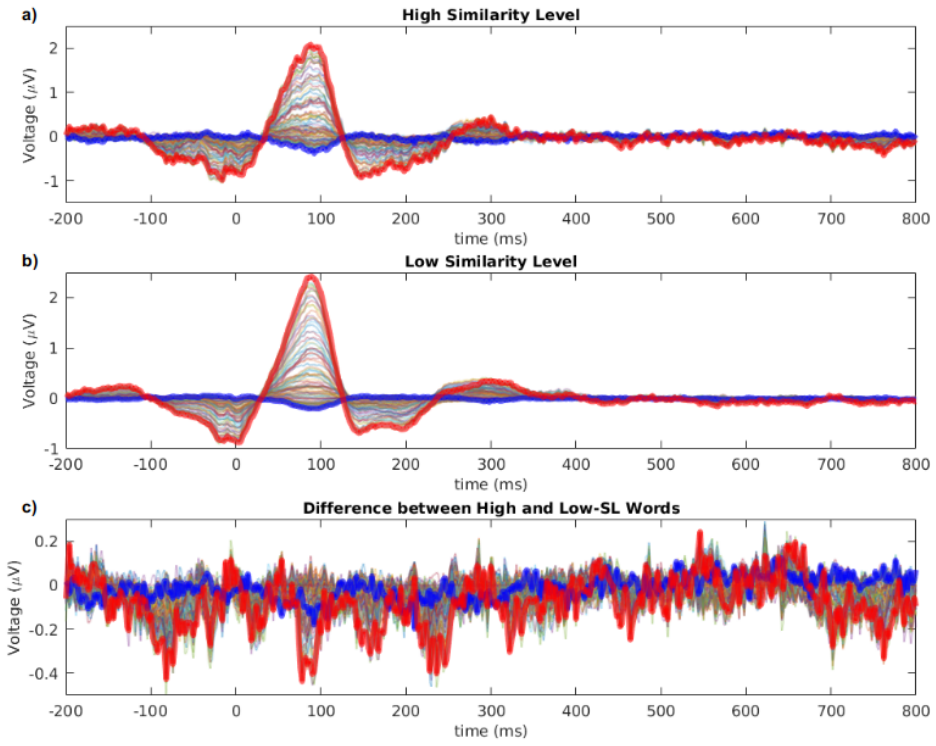

Figure 3.2: EEG waves time-locked to the fixation onset on **a)** high-similarity-level words **b)** low-similarity-level words and **c)** the difference between them.

eight channels had statistical significance: E69, E70, E74, E75, E82, E83, E89, and E100. All of these channels resided in the occipital lobe. For one to five time points, sixteen channels had statistical significance: E45, E47, E50, E51, E57, E58, E59, E64, E65, E66, E71, E76, E90, E95, E96, and E101. Those channels were either from the occipital lobe or the parietal lobe. Figure 3.4 plotted the electrodes with statistical significance on the scalp map. As we can see, most statistically significant channels were found in the occipital lobe, and some were found in the parietal lobe.

In Figure 3.5, we plotted the power scalp topographic maps for the high-SL and low-SL words averaged across all subjects in order to visualize the frequency band power distribution. The occipital lobe displayed strong power in all four frequency bands for both groups. A few channels in the occipital lobe showed an increase in alpha, beta, and gamma band power, as the

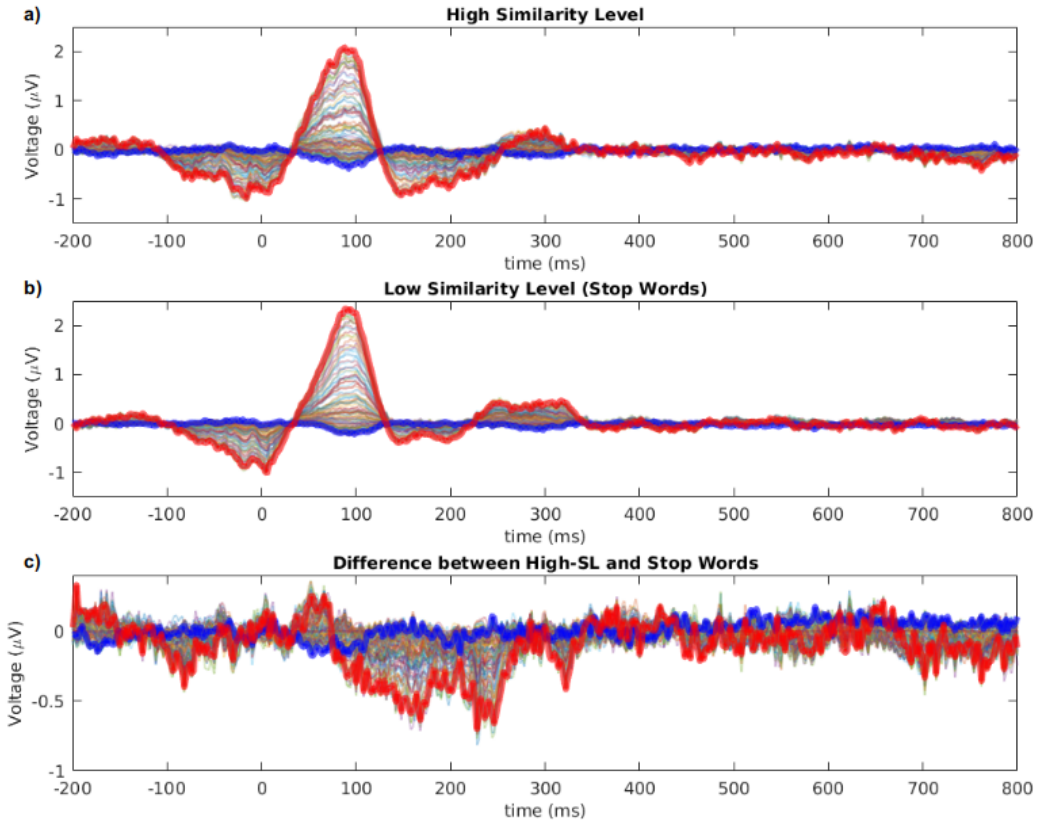

Figure 3.3: EEG wave time-locked to the fixation onset on **a)** high-similarity-level words **b)** stop words and **c)** the difference between them.

third row showed. After performing the t-test, we did not find any statistically significance for any frequency at any channel.

As shown in Figures 3.6, we visualized the frequency band power distribution every 20 seconds starting from the eye-fixation onset. Through the 300 ms, some occipital lobe channels showed higher gamma band power in words with high similarity. Beta power also increased in the occipital lobe from 0 ms to 100 ms and from 160 ms to 240 ms. A decrease in beta power was observed between 120 and 160 ms. We then performed the t-test for each time-frequency series at each channel, which was shown in Figure 3.7. The electrode E74 showed statistical significance for the gamma band between 60 ms and 240 ms. The electrode E89 showed statistical significance

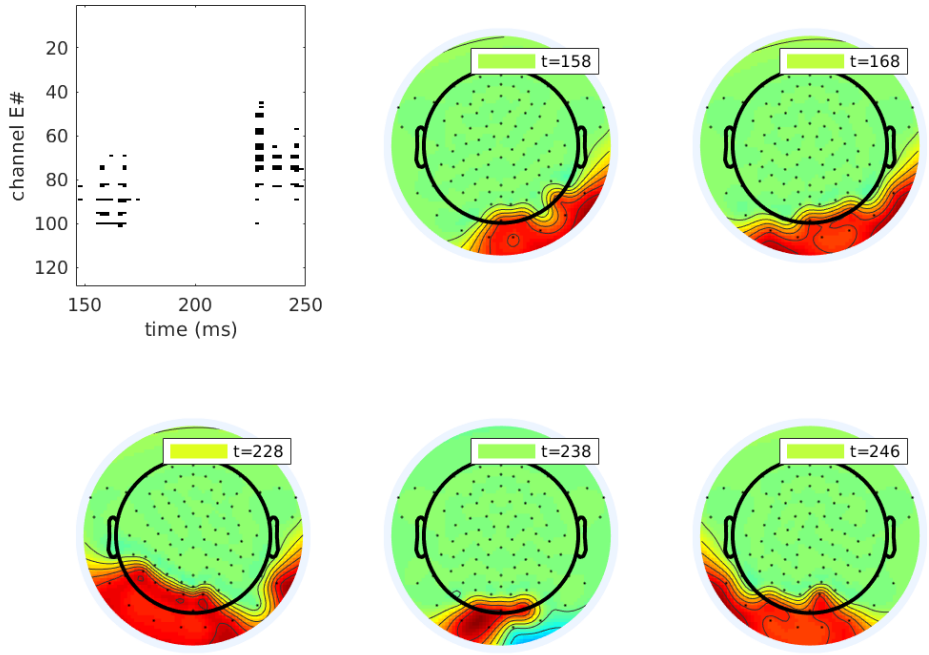

Figure 3.4: T-test results for 105 channels at 501 time points between -200 ms and 800 ms, with a threshold of $\alpha = 1.90 \times 10^{-7}$. The upper left plot shows when and what channels passed t-test in black. Red color indicates significant channels at each time points.

for the gamma band between 60-80 ms and 280-300 ms.

Using feature space projection methods, we analyzed the frequency patterns with results shown in Figure 3.8 and Figure 3.9. Based on a 5-fold train-test split, the training accuracy was 56.7%, and the test accuracy was 55.7% when using LDA modeling and PCA augmentation. The training accuracy was 56.7%, and test accuracy was 56.3% when using LDA modeling and FA augmentation, as shown in Table 3.1. Our next step was to train both LDA models for data within individual subjects. The average test accuracy for LDA with PCA was 59.1% with a standard deviation of 6.0%. The average test accuracy for LDA with FA was 59.4%, with a standard deviation of 4.7%, as shown in Figure 3.10 and Table 3.1. Both models achieved an accuracy above the chance level. We also tried to classify the words for each block of sentences within

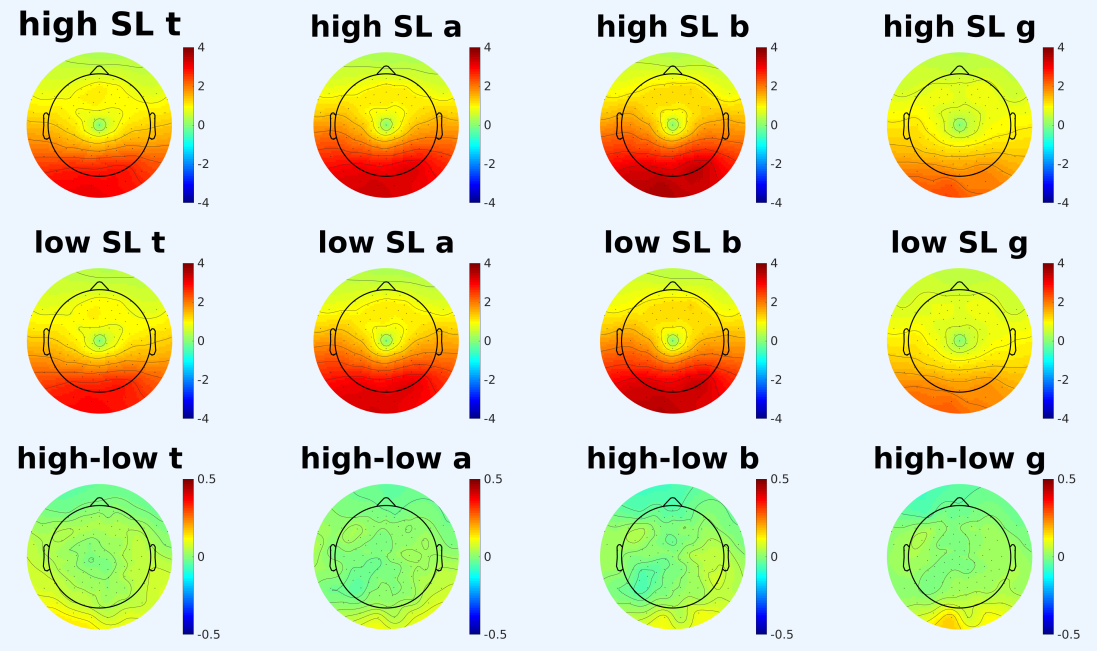

Figure 3.5: The band power distribution averaged across all subjects for each band frequency in the entire fixation window. Each letter represents a frequency band (t: theta; a: alpha; b: beta; g: gamma). The first row contained high-SL words, the second row contained low-SL words, and the third row is the difference between them.

or across subjects, but the results were chance accurate. With the data from the eight channels instead of all channels, the average test accuracy drops to 51%-52% for both models.

Afterwards, we trained and tested the CNN model for 300 epochs with a learning rate of 5×10^{-5}. The model was first fitted and tested for the samples across all subjects (4745 samples for the high-SL group vs. 4745 samples for the low-SL single-fixation group). The accuracy at every epoch of time was calculated, as shown in Figure 3.11. This figure shows that the test accuracy reached the maximum after one hundred and twenty epochs, which is 59.3%.

We also fitted and tested models for single subjects, and the accuracy of training and testing was displayed in Figure 3.12. Table 3.1 displayed the best accuracy in the test. Each subject has a class size of 395 ± 103 EEG pieces associated with the high-SL words. Low-SL sets were randomly under-sampled to the same size. Overall, the best test accuracy was 63.3% with a standard deviation of 5.6%.

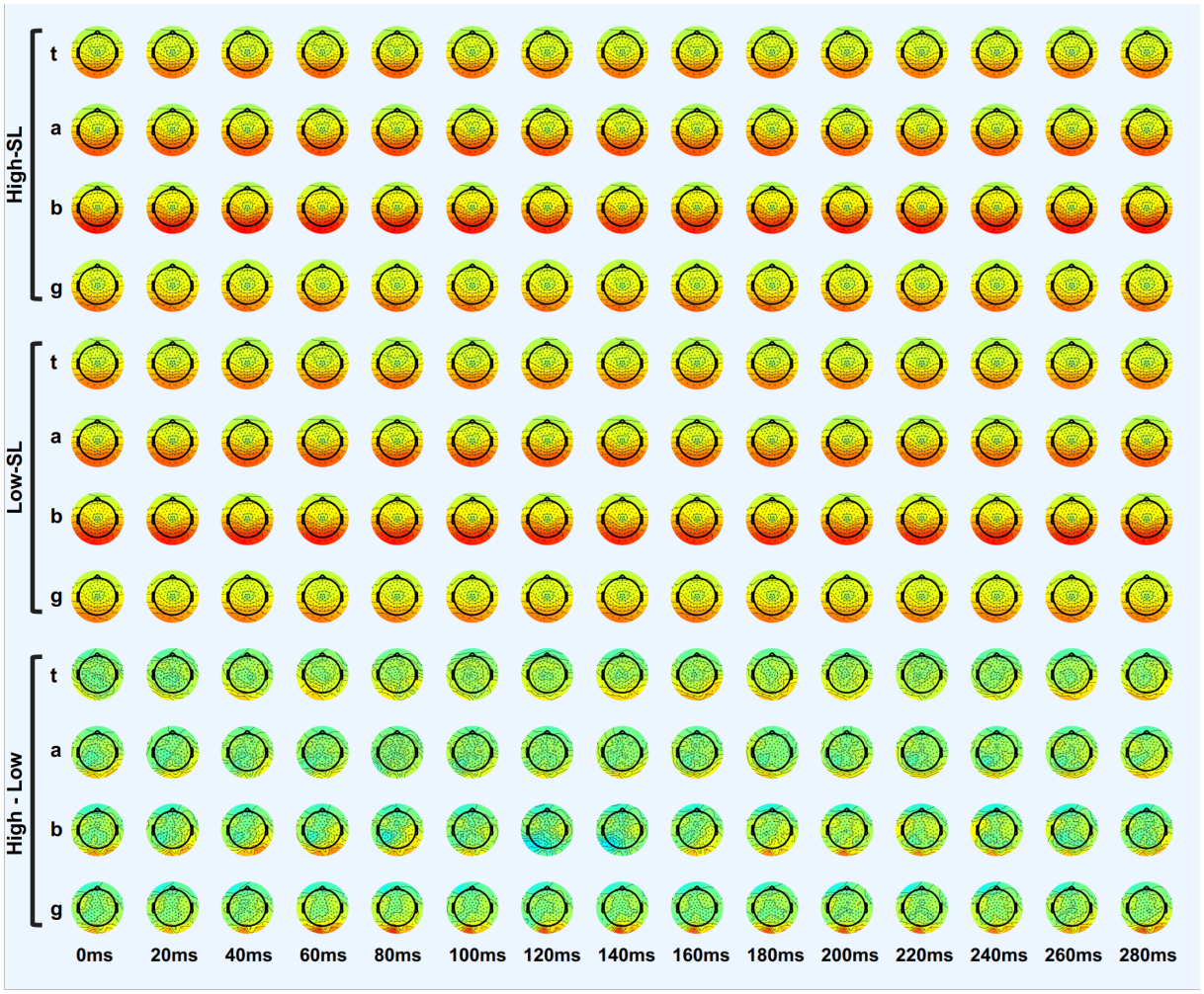

Figure 3.6: The band power distribution averaged across all subjects for each band frequency per 20-ms window. The first four rows are for high similarity level words, the second four rows are for the low similarity level words, and the rest four row are the difference between them. Each column is a 20-ms time window.

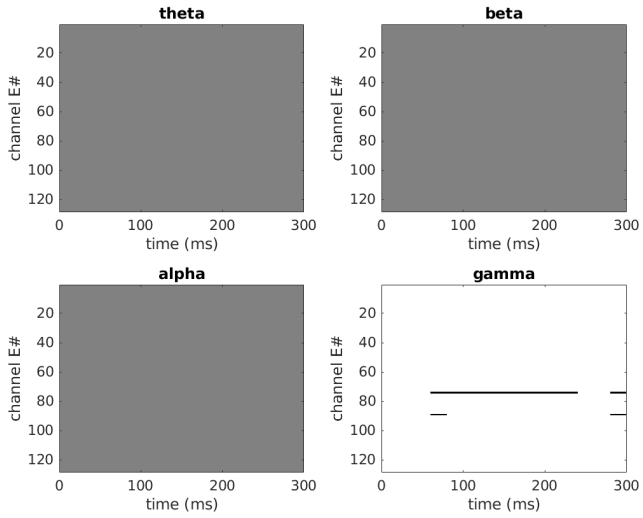

Figure 3.7: This plot showed the band power distribution averaged across all subjects for each band frequency at each time window. The first four rows contained high-SL words, the second four rows contained low-SL words, and the third four rows were the difference between them.

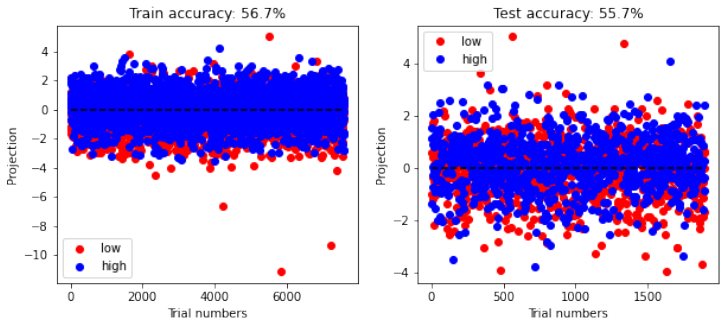

Figure 3.8: The test result of the LDA with PCA data-augmentation across all subjects.

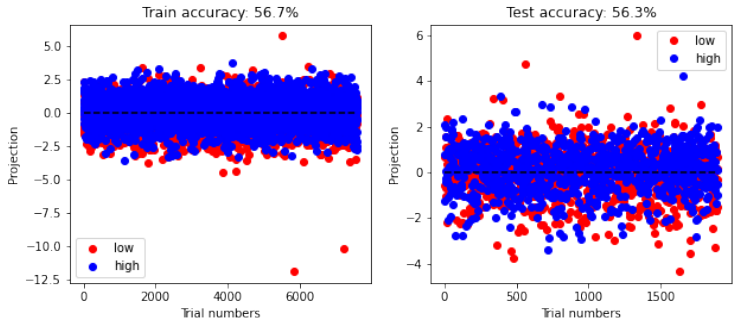

Figure 3.9: The test result of the LDA with FA data-augmentation across all subjects.

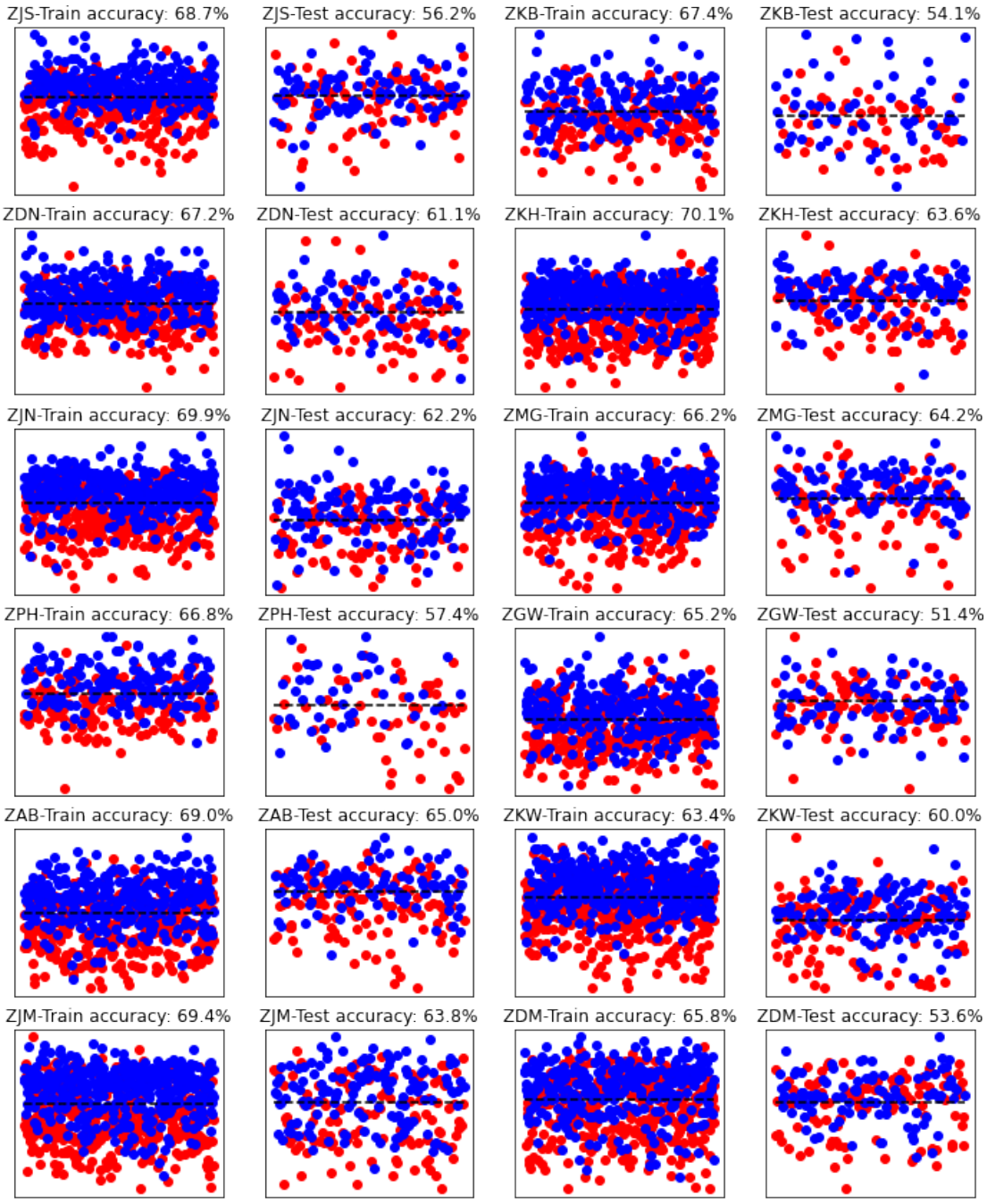

Figure 3.10: The results from the LDA with FA data-augmentation for within-subject training and testing.

Figure 3.11: The train and test accuracy per epochs of training across all subjects.

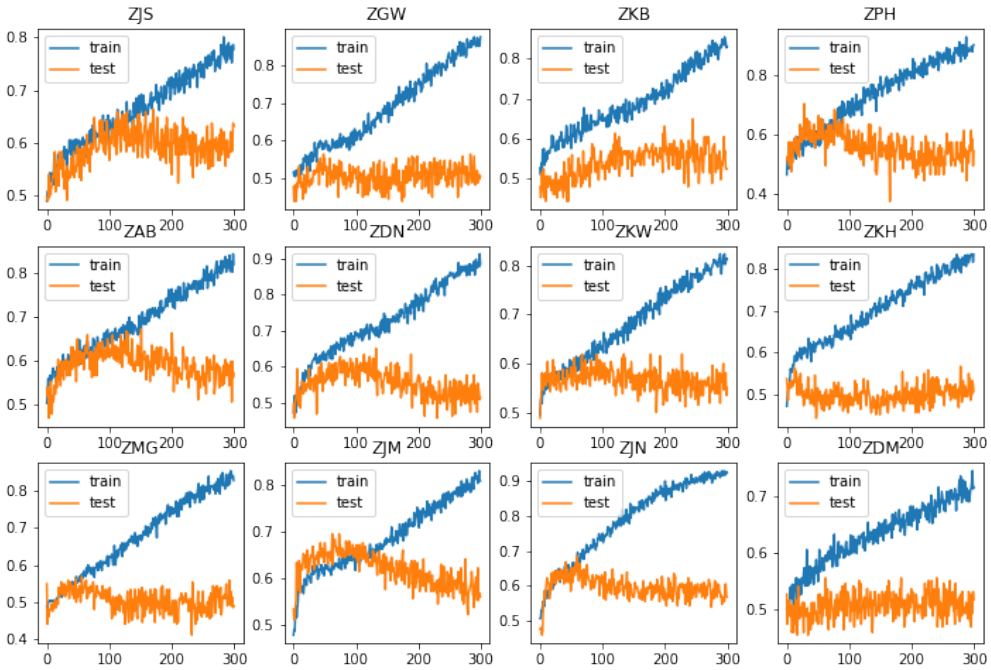

Figure 3.12: The train and test accuracy per epochs of training for each single subject.

Table 3.1: The table lists the number of words with eye-fixations per category. The training result of each method is also displayed.

Subject	ZJS	ZDN	ZJN	ZPH	ZAB	ZJM	ZKB	ZKH	ZMG	ZGW	ZKW	ZDM	Average
#high-SL words with fixation(s)	369	350	475	201	428	519	221	472	408	366	519	417	395±103
# low-SL words with 1 fixation	2348	2203	3411	1614	2603	3434	1358	2909	2827	1963	3059	2529	2521±659
# all words with fixation(s)	3371	3294	5557	2372	3916	5849	2143	4554	4122	3149	5531	3782	3970±1216
PCA+LDA test accuracy	63.2%	57.7%	65.1%	56.4%	65.4%	65.8%	51.4%	65.3%	51.5%	51.4%	62.3%	54.1%	59.1%±6.0%
FA+LDA test accuracy	56.2%	61.1%	62.2%	57.4%	65%	63.8%	54.1%	63.6%	64.2%	51.3%	60%	53.6%	59.4%±4.7%
CNN best test accuracy	66.5%	65.1%	68.5%	70.3%	67.3%	69.6%	64.9%	56.8%	56.9%	56.3%	61.9%	55.5%	63.3% ± 5.6%

3.4 Conclusion

In this chapter, we examined the EEG signals associated with the high-SL and low-SL groups of words. Time-series and time-frequency analyses suggested a couple of features which were confirmed statistically. We observed a voltage decrease of 0.4-0.7 μV in the occipital lobe between 100 ms and 270 ms for the high-SL group. In our comparison of high-SL words and stop words, the voltage decreased statistically significantly for eight channels of the occipital cortex at two time periods, 158 ms - 168 ms and 228 ms - 248 ms. This decline in voltage could be explained by N170, a event-related potential that reflects neural processing of familiar objects and words [22]. It is likely that this finding is indicative of neural biomarkers associated with the eye-fixations upon significant words within a sentence.

The ZuCo group found that the EEG power would continue decreasing after 100 ms and would remain below zero after 120 ms until the second fixation onset [6]. In addition, word length has a positive correlation with fixation duration [6]. The stop words, which were often shorter than the high-similarity words, may get a shorter fixation duration, and the voltage may rise earlier due to the second fixation onset. There were no statistically significant EEG patterns associated with word length in our current study. Yet, we cannot rule out the possibility that the high-SL words' lower voltage at around 160 ms and around 240 ms is due to their longer fixation

duration. This possibility needs to be verified or rejected by further experiments.

We observed a very mild increase in beta and gamma band power in the occipital lobe for words with high SL, but the difference did not reach statistical significance. Several findings can be drawn from time-frequency analysis of fixation-locked EEG. The gamma band power in the occipital lobe was stronger for the high-SL words between 60 ms and 300 ms, while the t-test showed two occipital channels were statistically significant.

By modeling the We achieved a test accuracy above the chance accuracy with the LDA model with either PCA or FA data augmentation. Considering the inter-subject discrepancies, we also trained the LDA model within single subjects. The test accuracy for cross-subject and within-subject conditions was both above the chance level. Using data from just one subject, both LDA models performed better in a test than when they were trained across subjects.

We also explored the possibility of using CNN to learn the difference between the two classes and predict the class based on the EEG features. Test accuracy was 59.3% for cross-subject conditions and 63.3% for within-subject conditions. The best test accuracy was often achieved between 100 epochs and 150 epochs when the training accuracy rose to around 70%. In terms of testing accuracy, CNN model that fits within single subjects typically performed better than the CNN model that fits across all samples. The CNN models performed slightly better than the LDA models as they achieved a higher test accuracy either within or between subjects.

Chapter 4

Discussions and Future Directions

As part of this project, we looked at the semantics of the ZuCo Task-Specific-Reading dataset and tried to identify biomarkers that are associated with reading different words. With regards to both EEG signals and eye-tracking data, some biomarkers were found to be statistically significant. We also tried to train algorithms to learn the differences in the EEG responses. In this chapter, we will discuss our significant findings and our innovations. Afterwards, we discussed the disadvantages of this study and future directions.

4.1 Major Conclusions

1. We propose a method combining the current NLP models to evaluate the significance of a word in interpreting the entire sentence meaning based on semantic similarity measurement. We found that the underlying calculations for each NLP model were prone to biases due to differences in assumptions, algorithm design, and language corpus embedding. Thus, we needed to consider these biases and did not rely on one calculation alone. Based on the results of each calculation, we calculated a similarity level, which is the number of methods that consider the word with a high similarity score. To evaluate and confirm the effectiveness of the similarity level, we examined eye-fixation probability and average fixation times.

2. We found statistically significant eye-fixation features for the words with different similarity levels (SL). First, the word with a high SL had a higher probability of getting at least one eye-gazing and more times of eye-fixations upon it. The absolute duration of first fixations was not significantly different between the high-SL words and the low-SL words. However, after normalized with the average first fixation duration for each subject, we found that high-SL words had a 3% longer duration of the first fixation upon them. To conclude, the words with a higher SL to the question keyword often attract a higher probability of eye-gazing, more times of eye-gazing, and a slightly longer duration of eye-gazing.

3. For the low-SL words, fixation-locked EEG potential decreased dramatically from 100ms to 270ms, possibly related to the manifestation of neural biomarkers N170. The comparison between high-SL and low-SL words did not indicate statistical significance, but the comparison between high-SL words and stop words did. Compared to stop words, high-SL words would have a significantly lower voltage by 0.6 μV in the occipital lobe between 158 ms - 168 ms and between 228 ms - 248 ms. The channels indicating statistical significance were often found in the occipital and parietal regions, as shown in Figure 3.4.

4. We found greater gamma-band power in the occipital lobe for the high-SL words than for those associated with the low-SL words. Upon evaluating the time frequency decomposition of the EEG signal, we observed a greater gamma band power in the occipital lobe from 60 ms to 240 ms. We found two channels in the occipital lobe to be statistically significant, which were E74 and E89.

5. The EEG frequency band power was used to train an LDA model and a CNN model to predict the high-SL or low-SL for the words. CNN and LDA models were both able to differentiate between two groups, showing an accuracy level higher than a chance level. This analysis showed that machine learning models can be used to detect comprehension-related EEG features.

4.2 Innovations

The previous studies often utilized a word-by-word paradigm /cite[Kutas1980, Perfetti2008, POLICH2007], which was extremely restrictive in terms of eye position and reading speed. Through the development of eye-tracking and high-resolution EEG recording techniques, studies of neural features during natural reading settings have become more feasible in recent decades. Although the recent studies explored eye movement across words within a normal sentence, they didn't fully investigate neural signals that accompany eye-gazing upon each word within the sentence. We also had other tools for decoding any neural features found during reading comprehension processes since complex NLP models were developed in the recent ten years. Thus, the integration of the new approaches into traditional reading comprehension studies would be a worthwhile endeavor.

The ZuCo group also attempted to integrate the popular NLP models with EEG, eye-tracking, and magnetoencephalography datasets to explore the relationship between word meaning and biological signals [23, 24]. According to the authors, the eye-tracking and EEG data correlated to some extent when fitted to the NLP models [23]. The studies were established at the sentence level and did not explore the use of NLP tools to decode the neural responses. Therefore, we aimed to use NLP tools in studies of eye-tracking and EEG regarding reading comprehension.

Based on those previous studies, this project explored and demonstrated a potential application of NLP tools to assist in the analysis of the experiment texts and the exploration of the eye-tracking and EEG biomarkers related to word-level comprehension. In this study, we examined word-level reading comprehension biomarkers during natural sentence reading, as a unique direction in the field of reading comprehension research. Using NLP tools, we identified the significant words in interpreting the relationship contained in the sentence, followed by a statistical analysis of eye-tracking and EEG signals in relation to those words. In this paper, the statistically significant features for those words were summarized, which were unique from the

ZuCo group's findings. We hoped that our findings would contribute to future research in this area.

In this study, we tested the feasibility of using a neural network model to predict the word similarity level from the EEG signal. In this attempt, the aim was to try to expand the applications of neural networks to predict the effects of word stimuli, which have not been fully investigated within the current field. The results of this study confirmed that machine learning algorithms and neural networks are capable of decoding complex neural signals.

4.3 Reflections and Future Directions

During the pandemic season, we were unable to conduct in-person experiments and collect data of interest due to various difficulties. We were limited in what analysis we could perform and had to adapt existing datasets for our purpose. We did not have enough data on miscomprehension cases in this dataset, since the answer correctness rate was above 90% /cite[Hollenstein2018]. We may, in the future, design a modified experiment and adapt the questions and sentence presentations in order to compare the cases of true comprehension and miscomprehension. It is also important to record the subjects' uncertainty level in understanding the sentences.

This project could better address the data imbalance problem. Our model was not sufficiently general for those non-involved data since we chose the random under-sampling method. Methods such as sample duplication of the minor class and class-specific learning rates were explored, but results were not improved as expected. Only the under-sampling method worked. In the future, we may use random forest, which should be resilient to class imbalances.

The EEG frequency-domain features were learned using the CNN model based on previous studies on application of image-learning neural networks to EEG data [21, 25]. Time-series data and time-frequency decomposition of the EEG signals were not fully analyzed with neural network models. We explored deeper neural networks, but a successful implementation required

more time and more in-depth knowledge preparation. We may continue to explore the possibility of applying recurrent neural networks to these data.

Bibliography

[1] M. Kutas and S. A. Hillyard, "Event-related brain potentials to semantically inappropriate and surprisingly large words," *Biological Psychology*, vol. 11, no. 2, pp. 99–116, Sep. 1980. [Online]. Available: https://doi.org/10.1016/0301-0511(80)90046-0

[2] M. Kutas and K. D. Federmeier, "Thirty years and counting: Finding meaning in the n400 component of the event-related brain potential (ERP)," *Annual Review of Psychology*, vol. 62, no. 1, pp. 621–647, Jan. 2011. [Online]. Available: https://doi.org/10.1146/annurev.psych.093008.131123

[3] C. Perfetti, C.-L. Yang, and F. Schmalhofer, "Comprehension skill and word-to-text integration processes," *Applied Cognitive Psychology*, vol. 22, no. 3, pp. 303–318, 2008. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/acp.1419

[4] D. N. Rapp and H. A. Taylor, "Interactive dimensions in the construction of mental representations for text." *Journal of Experimental Psychology: Learning, Memory, and Cognition*, vol. 30, no. 5, pp. 988–1001, 2004. [Online]. Available: https://doi.org/10.1037/0278-7393.30.5.988

[5] J. Polich, "Updating p300: An integrative theory of p3a and p3b," *Clinical Neurophysiology*, vol. 118, no. 10, pp. 2128–2148, 2007. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1388245707001897

[6] N. Hollenstein, J. Rotsztejn, M. Troendle, A. Pedroni, C. Zhang, and N. Langer, "ZuCo, a simultaneous EEG and eye-tracking resource for natural sentence reading," *Scientific Data*, vol. 5, no. 1, Dec. 2018. [Online]. Available: https://doi.org/10.1038/sdata.2018.291

[7] G. A. Miller, "WordNet: a lexical database for English," *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, Nov. 1995. [Online]. Available: https://doi.org/10.1145/219717.219748

[8] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1532–1543. [Online]. Available: http://www.aclweb.org/anthology/D14-1162

[9] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 135–146, Dec. 2017. [Online]. Available: https://doi.org/10.1162/tacl_a_00051

[10] A. Rajaraman and J. D. Ullman, "Data mining," in *Mining of Massive Datasets*. Cambridge University Press, pp. 1–17. [Online]. Available: https://doi.org/10.1017/cbo9781139058452.002

[11] Z. Wu and M. Palmer, "Verbs semantics and lexical selection," in *Proceedings of the 32nd annual meeting on Association for Computational Linguistics* - . Association for Computational Linguistics, 1994. [Online]. Available: https://doi.org/10.3115/981732.981751

[12] P. Resnik, "Using information content to evaluate semantic similarity in a taxonomy," *CoRR*, vol. abs/cmp-lg/9511007, 1995. [Online]. Available: http://arxiv.org/abs/cmp-lg/9511007

[13] D. Lin, "An information-theoretic definition of similarity," in *In Proceedings of the 15th International Conference on Machine Learning*. Morgan Kaufmann, 1998, pp. 296–304.

[14] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, "Bag of tricks for efficient text classification," *CoRR*, vol. abs/1607.01759, 2016. [Online]. Available: http://arxiv.org/abs/1607.01759

[15] T. Mikolov, E. Grave, P. Bojanowski, C. Puhrsch, and A. Joulin, "Advances in pre-training distributed word representations," in *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

[16] D. D. Salvucci and J. H. Goldberg, "Identifying fixations and saccades in eye-tracking protocols," in *Proceedings of the symposium on Eye tracking research & applications - ETRA '00*. ACM Press, 2000. [Online]. Available: https://doi.org/10.1145/355017.355028

[17] S. C. Sereno and K. Rayner, "Measuring word recognition in reading: eye movements and event-related potentials," *Trends in Cognitive Sciences*, vol. 7, no. 11, pp. 489–493, Nov 2003. [Online]. Available: https://doi.org/10.1016/j.tics.2003.09.010

[18] S.-C. Chen, H.-C. She, M.-H. Chuang, J.-Y. Wu, J.-L. Tsai, and T.-P. Jung, "Eye movements predict students' computer-based assessment performance of physics concepts in different presentation modalities," *Computers & Education*, vol. 74, pp. 61–72, May 2014. [Online]. Available: https://doi.org/10.1016/j.compedu.2013.12.012

[19] A. Delorme and S. Makeig, "EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis," *Journal of Neuroscience Methods*, vol. 134, no. 1, pp. 9–21, Mar. 2004. [Online]. Available: https://doi.org/10.1016/j.jneumeth.2003.10.009

[20] A. Bruns, "Fourier-, hilbert- and wavelet-based signal analysis: are they really different approaches?" *Journal of Neuroscience Methods*, vol. 137, no. 2, pp. 321–332, Aug. 2004. [Online]. Available: https://doi.org/10.1016/j.jneumeth.2004.03.002

[21] P. Bashivan, I. Rish, M. Yeasin, and N. Codella, "Learning representations from eeg with deep recurrent-convolutional neural networks," 2016.

[22] B. Rossion, C. A. Joyce, G. W. Cottrell, and M. J. Tarr, "Early lateralization and orientation tuning for face, word, and object processing in the visual cortex," *NeuroImage*, vol. 20, no. 3, pp. 1609–1624, Nov. 2003. [Online]. Available: https://doi.org/10.1016/j.neuroimage.2003.07.010

[23] N. Hollenstein, A. de la Torre, N. Langer, and C. Zhang, "Cognival: A framework for cognitive word embedding evaluation," 2019.

[24] N. Hollenstein, M. Barrett, M. Troendle, F. Bigiolli, N. Langer, and C. Zhang, "Advancing nlp with cognitive language processing signals," 2019.

[25] S. Siddharth, T.-P. Jung, and T. J. Sejnowski, "Utilizing deep learning towards multi-modal bio-sensing and vision-based affective computing," *IEEE Transactions on Affective Computing*, pp. 1–1, 2019. [Online]. Available: https://doi.org/10.1109/taffc.2019.2916015