Comparing and Predicting Eye-tracking Data in Mandarin and Cantonese

Junlin Li
The Hong Kong Polytechnic University
junlin.li@connect.polyu.hk

Bo Peng, Yu-Yin Hsu, Emmanuele Chersoni
The Hong Kong Polytechnic University
{bopeng, yyhsu, echers}@polyu.edu.hk

Abstract

Eye-tracking data in Chinese languages present unique challenges due to the non-alphabetic and unspaced nature of the Chinese writing systems. This paper introduces the first deeply-annotated joint Mandarin-Cantonese eye-tracking dataset, from which we achieve a unified eye-tracking prediction system for both language varieties. In addition to the commonly studied first fixation duration and the total fixation duration, this dataset also includes the second fixation duration, expressing fixation patterns that are more relevant to higher-level, structural processing.

A basic comparison of the features and measurements in our dataset revealed variation between Mandarin and Cantonese on fixation patterns related to word class and word position. The test of feature usefulness suggested that traditional features are less powerful in predicting the second-pass fixation, to which the linear distance to root makes a leading contribution in Mandarin. In contrast, Cantonese eye-movement behavior relies more on word position and part of speech.

1 Introduction

Eye-tracking has quickly become one of the most popular methodologies in psycholinguistic studies, as it allows researchers to measure people's real-time processing efforts during a reading task (Attardo and Pickering, 2023). Consequently, more computational models have been proposed to predict eye-fixation patterns in English and many other languages (Hollenstein et al., 2021a,b, 2022; Salicchi et al., 2022).

Chinese languages, being non-alphabetic, are considered unique in eye-tracking research, mainly due to the unspaced nature of the writing conventions, the visual complexity of the characters, and the abundance of homophonic and homographic characters (Hsu and Huang, 2000; Bai et al., 2008). The computational modeling of Chinese eye-movement patterns is still relatively limited, although several traditional psycholinguistic models have been proposed to measure Chinese reading times (Rayner et al., 2007; Li and Pollatsek, 2020; Thierfelder et al., 2020). Such models have focused on factors such as word frequency, word length, and word predictability but have not considered syntactic and semantic processes that may have an equally decisive influence on eye-movement behaviors.

To fill such research gaps, this paper first introduces a deeply-annotated eye-tracking dataset that covers Mandarin texts in simplified characters and Cantonese texts in traditional characters, thus representing two demographically important language varieties. Based on this joint dataset, we implemented a series of statistical tests to investigate the inter-linguistic variance from the perspective of fixation durations. Furthermore, we propose a feature-rich prediction model of basic eye-tracking measurements in Chinese, in addition to an ablation study of the usefulness of features. Our predictors include both traditional and new features, such as syntactic features, local lexical semantic features, and contextual semantic representations. We believe that comparing these features will further broaden our understanding of the differences between Mandarin and Cantonese. The contributions of the present study are as follows:

- we presente the first parallel Mandarin-Cantonese eye-tracking dataset. The dataset is annotated with three eye-tracking features, including the second fixation duration, which reflects higher-level, structural processing of a sentence;
- we explore the similarities and differences between Mandarin and Cantonese, two demographically important varieties within the family of the Sinitic languages, from the perspective of cognitive processing as reflected in eye-tracking measurements.

Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023), pages 121–132

- we introduce computational models to approximate and predict the fixation patterns of the two varieties. Specifically, we integrate morphosyntactic features and contextualized semantic representations with traditional lexical features into the modeling of fixation measurements.

2 Related Work

As eye-tracking data are closely linked to real-time cognitive processes, they can reveal the automatic operations in our brains that are related to different linguistic modules, such as lexical access (Clifton Jr et al., 2007), syntactic processing (Van Schijndel and Schuler, 2015), semantic processing (Hwang et al., 2011; De Groot et al., 2016), and pragmatic competence (Gironzetti, 2020). Regarding the modeling of fixation patterns, previous research has highlighted the close relationship between eye-tracking measurements and certain word properties, including word position (Just and Carpenter, 1980), word frequency (Yan et al., 2006; Liversedge et al., 2014), word predictability (Rayner et al., 2005), and word length (Li et al., 2011; Zang et al., 2018). Fixation on a particular word is also sensitive to the cognitive load from the previous word (known as a spill-over effect) (Rayner et al., 1989; Pollatsek et al., 2008).[1]

In addition to these traditional lexical features, morpho-syntactic features, such as part-of-speech categories (POS) and syntactic dependency, also impact fixation patterns. POS have been demonstrated to influence the number of fixations and the fixation duration (Blanchard, 1985). Concerning syntactic dependency, previous studies indicated that cognitive loads from syntactic structure lead to increased refixation probability and duration (Conklin and Pellicer-Sánchez, 2016; Frenck-Mestre, 2005), which is mainly related to the second fixation duration (**SFD**) in this paper and partially reflected on the total fixation duration (**TFD**). Previous research also reported that the sensitivity of first-pass processing to the syntactic agreement increases the first fixation duration (Deutsch, 1998; Deutsch and Bentin, 2001), although there is counter-evidence that syntactic parsing only increases the total fixation duration by affecting the second fixation (Pearlmutter et al., 1999). Despite being crucial for modeling fixation patterns, it should be noted that most of the research that has targeted Chinese languages has not considered POS and syntactic dependency.

Regarding the distinctiveness of the Chinese writing system, most studies have supported the view that words and characters are equally salient units in the cognitive processing of texts written in Chinese characters, as both word properties and character properties influence reading-time measurements and eye-movement behaviors (Bai et al., 2008; Li et al., 2015). Following this assertion, the word-level features widely applied in the reading-time modeling of alphabetic languages are equally applied in Chinese-specific research. Features related to higher-level processing, such as syntactic properties, are also considered in research on Chinese language processing (Lu et al., 2022; Chen and Tsai, 2015; Zang et al., 2020). However, previous studies using syntactic properties have mainly focused on syntactic complexity and the grammatical function of a word without linking the syntactic dependency of the entire sentence to eye-movement modeling.

3 Dataset

This section introduces our eye-tracking dataset's construction procedures and annotation structure.[2] We then present the results of inter-variety comparisons regarding basic eye-tracking measurements in the next section.

3.1 Data Collection and Normalization

This study used two comparable eye-tracking corpora collected by ourselves, one in Mandarin and one in Cantonese, which were recorded using a normal reading paradigm. Each corpus included 30 participants who were native speakers of the target language; the mean age of the Mandarin group was 25.8 years old (22 females) and the Cantonese group was 21.7 years old (20 females). During the recording sessions, the participants read a translated version of *The Little Prince* by Antoine de Saint-Exupéry, in Mandarin, and in Cantonese, respectively. The Mandarin texts were presented in simplified Chinese characters and the Cantonese

[1] According to some studies, another factor affecting fixations is the semantic relatedness of a word with its context, which can be measured via Distributional Semantic Models (Pynte et al., 2008; Mitchell et al., 2010; Salicchi et al., 2023). However, the evidence for the role of semantic relatedness in predicting reading times and eye fixations is controversial (Frank, 2017).

[2] Code and datasets will be made available via Github at the following URL: `https://github.com/CN-Eyetk/MCFIX`.

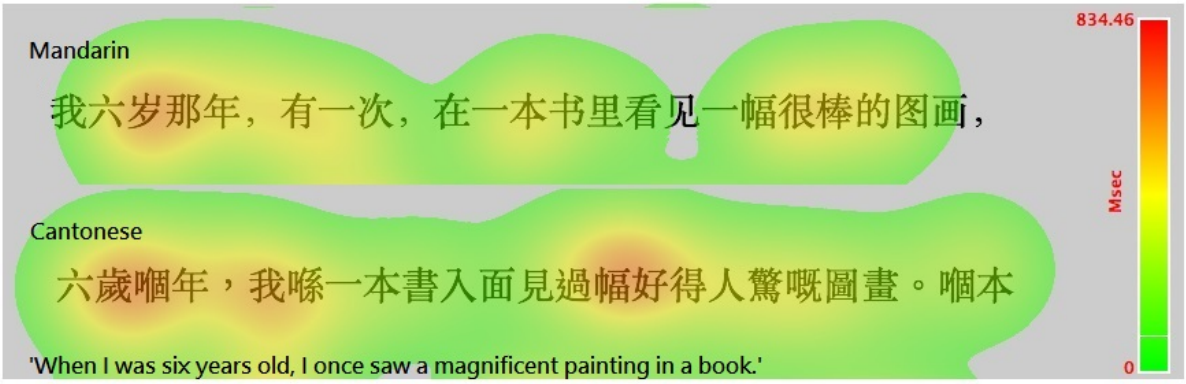

Figure 1: An example heatmaps of fixation duration by Mandarin and Cantonese readers, weighted by the duration of the individual fixations.

texts in traditional Chinese characters. Each corpus contained recordings of two reading tasks using the same texts, i.e., the natural reading (**NR**, only with a reading comprehension task) and the task-specific reading (**TSR**, with the purpose of finding specific information in a given text). Each corpus contained three eye-movement measurements and their standard deviations: first fixation duration (**FFD**), second fixation duration (**SFD**), and total fixation duration (**TFD**). Figure 1 shows the heatmaps of fixation duration recorded from one Mandarin and one Cantonese reader in our data.

We then normalized the raw data as follows: If a word, w, occurs $N_{total} = n + n_{null}$ times, where n is the number of instances with fixation values, and n_{null} is the number of instances with null values, then the normalized value equals the sum of the fixation values of n occurrences divided N_{total} times. Table 2 shows the descriptive statistics for these fixation measurements of the two language varieties in each task. Our datasets show that monosyllabic words were more dominant in Cantonese than in Mandarin, especially for content words such as verbs and nouns, as shown in Figure 2; this tendency is in line with the monosyllabic salience observed in Cantonese (Li et al., 2016).

SENT	WORD	POS	LDR	LDH	DEPTH	Freq	N_{SYL}
1	看见(see)	VERB	0	0	0	260.0	2
1	一(one)	NUM	1	5	2	8489.0	1
1	幅(clf)	DET	2	4	2	103.0	1
1	很(very)	ADV	3	1	3	1755.0	1
1	棒(good)	ADJ	4	2	2	27.0	1
1	的(de)	PART	5	1	3	77946.0	1
1	图画(figure),	NOUN	6	6	1	25.0	2

Table 1: Annotation Example

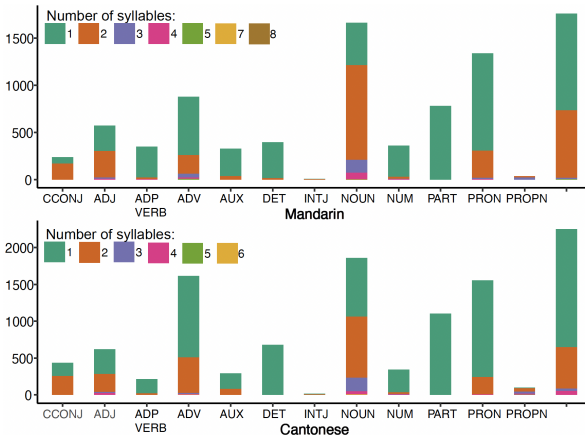

Figure 2: Two-way comparison of syllable number and part-of-speech in Mandarin and Cantonese

3.2 Annotation Structure

In addition to eye-movement measurements, we obtained several linguistic features of our dataset in the annotation: (1) **Word Segmentation**, which inherited the word segmentation marked by native speakers with a Ph.D. in linguistics during the collection of eye-tracking data; (2) **Part-of-speech**, which is derived from `jiagu` toolkit (https://github.com/ownthink/Jiagu) for the Mandarin text, and from `pycantonese` (Lee et al., 2022) for the Cantonese text; the results of which were manually checked and aligned by a Mandarin speaker and a Mandarin-Cantonese bilingual speaker; (3) Syntactic distances, including dependency depth (**DEPTH**), linear distance to Head (**LDH**), and linear distance to root (**LDR**); all of these were based on a syntactic analysis derived by the Stanford Dependency Parser (Chang et al., 2009); and (4) Traditional features in eye-tracking modeling, including **word frequency** (obtained from the `cifu` dictio-

nary (Lai and Winterstein, 2020) and the `cncorpus` word-frequency list), and the **syllable number**.

Table 1 provides an illustration of the annotation of linguistic features.

4 Cross-variety Comparison

4.1 One-way Comparison

Concerning the cross-variety variance between the data in the two corpora, we fit a linear model against FFD, SFD, and TFD (all in a log scale). The fixed effects included *LanguageTypes* (Cantonese vs. Mandarin, the former as the treatment), and *WordFrequency*, *POS* and *SyllableCount* ($N_{Syllable}$). The estimation was implemented by `lm` function in `RStudio` (Allaire, 2012).

The result shown in Table 3 (With $N_{syllable}$) highlights the effect of writing system simplification. The tendency indicates that the Mandarin readers consistently had significantly shorter fixation durations for all the FDs of the TSR task and the FFD of the NR task. This finding was consistent with the expectation that lower visual complexity may reduce cognitive effort; thus, Mandarin readers who encountered simplified Chinese texts showed significantly shorter first-past fixation times than Cantonese readers who processed traditional Chinese texts. This tendency extended to all the fixation measures in task-specific reading.

Nonetheless, the tendency caused by the writing system's simplification could be weakened by the fact that Cantonese has more monosyllabic words, thus simpler words, as shown in Figure 2. This was demonstrated by the finding that (1) the exclusion of syllable count from random effect neutralized the significance (See without $N_{syllable}$ in Table 3), and (2) the descriptive statistics of FD levels did not show a significant difference between the two variables (See Table 2).

4.2 FD Variance by POS and Word Position

On par with the general effects of language variety on the word-level fixation duration, this research also implemented a Tukey post hoc test to investigate the FD differences of each POS between the two language varieties. Figure 5 (in the appendix) shows that pronoun fixation and noun fixation (excluding proper names) had significant cross-variety differences, as Mandarin readers tended to fixate more on nouns in both reading tasks, while Cantonese readers were inclined to fixate more on pronouns in TSR. This consistent tendency concerning noun fixation presumably arises from the different distribution of syllabic length between Mandarin and Cantonese, as nouns in Mandarin are more likely to be disyllabic than monosyllabic (see Figure 2). The tendency for pronoun fixation, we assume, arises from the fact that Cantonese pronouns are more ambiguous than those in Mandarin. For example, the singular third-person pronouns of masculine gender "ta1"(他), feminine gender "ta1"(她), and neutral gender "ta1"(它) in Mandarin all correspond to the only singular third-person pronoun "keoi5"(佢) in Cantonese, which may cause Cantonese readers to spend more time on processing pronominal reference. In addition, Cantonese demonstrative pronouns have high-frequency homographs (or pseudo-homographs). The demonstrative pronoun "ni1/nei1" (呢"this") is homographic with the sentence-final particle "ne1" (呢). The demonstrative pronoun "go2" is pseudo-homographic with the classifier "go3" (個). This property presumably induces more efforts for Cantonese readers in the lexical access for demonstrative pronouns.

Apart from POS, we also investigated the similarities and differences between the two language varieties in terms of the effect of word position on fixation durations. For this, we fit the correlation between the word position in the sentence (normalized by sentence length) and the fixation duration with the third-degree polynomial formula (to capture non-linearity). Non-overlapping contours of a confidence interval indicate statistically significant differences. As shown in Figure 3, the final part of each sentence showed significant differences between Mandarin and Cantonese. Cantonese consistently tended to involve a descent of fixation durations in the final quarter of a sentence, while Mandarin was almost the opposite in such a local span, except for TSR's first fixation duration.

5 Methodology

This section introduces the features and the regressors used in the prediction of eye-tracking measurements, derived from the results of the cross-variety comparison drawn from both psycholinguistic and computational studies.

5.1 Prediction Targets

The prediction targets include the subject-wise normalized level (below referred to as *mean level*) of **FFD**, **SFD**, and **TFD** and the *standard deviations*

Mode	Variety	Word Count	FFD_{avg}	FFD_{std}	SFD_{avg}	SFD_{std}	TFD_{avg}	TFD_{std}
NR	Cantonese	5050	108.53	55.56	37.32	41.97	171.37	144.47
	Mandarin	3939	108.98	55.42	40.27	44.91	180.21	160.61
TSR	Cantonese	5047	101.22	52.75	30.29	34.10	145.89	102.89
	Mandarin	3941	101.26	54.45	30.82	34.27	147.30	106.41

Table 2: Descriptive statistics of fixation durations

Mode	Y	With $N_{syllable}$			Without $N_{syllable}$		
		estimates	Pval	Sig	estimates	Pval	Sig
TSR	FFD	+0.031	0.004	**	0.007	0.544	
	SFD	+0.060	0.056		-0.015	0.663	
	TFD	+0.046	0.000	**	0.009	0.515	
NR	FFD	+0.009	0.003	**	-0.01	0.400	
	SFD	+0.002	0.952		-0.064	0.061	
	TFD	+0.017	0.222		-0.016	0.282	

Table 3: Estimates of the effect of *Cantonese* on FDs, with $N_{syllable}$ not placed in random effect (on the left) and placed (on the right).

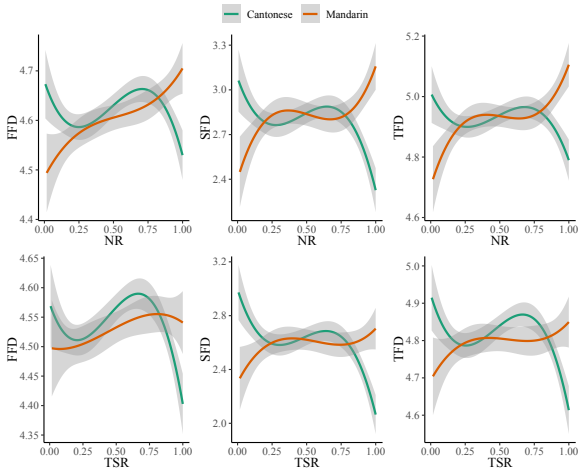

Figure 3: Polynomial contours and their 95% confidence interval of the correlation between normalized word position and FDs (in log scale) in Mandarin and Cantonese.

of **FFD**, **SFD**, and **TFD**, for both Mandarin and Cantonese. We believe that it is important to include the standard deviations in our gold standard: eye-tracking metrics prediction is an example of a task in which predicting only the mean value from a set of measurements typically excludes a large amount of variation existing in the data. For this reason, in the spirit of paving the way for NLP systems that can better deal with human label variation (Plank, 2022), we added this additional challenge to our dataset.

5.2 Features

We used two sets of features in our prediction experiment: Linguistic features and GPT word embeddings.

5.2.1 Linguistic Features

Given the annotation structure in our dataset, we selected nine linguistic features as shown below.

Traditional features based on previous studies included **Frequency** (of the current word and its previous word), **Syllable Number** (of the current word and its previous word), **Word Position**, and **POS**. Specifically, Frequency was extracted from `cifu` dictionary (Lai and Winterstein, 2020) and `cncorpus` word-frequency list[3] and was projected to a log scale; the previous word frequency and syllable number are specified as "-1" for sentence-initial words; word position is the order of a word in a sentence divided by the sentence length (by word).

In addition, we proposed five new features, of which four had not been used in modeling fixation patterns of Chinese languages in natural language processing, and one feature that has recently been shown to be useful in predicting eyemovement patterns: they are **DEPTH**, **LDH** and **LDR**, which are summarized in section 3, **Word Predictability** measured by `GPT2 Surprisal` (Salicchi et al., 2022), and **Orthographic Neighborhood**. The orthographic neighborhood refers to how likely a character cooccurs with other characters in a compound-word, inferring a given character's ambiguity level. We calculated this based on the `cifu` dictionary and `xinhua` wordlist for Cantonese and Mandarin, respectively. To calculate the Orthographic Neighborhood of each word, we divide the word into characters and sum up the number of words containing each single character, treating the summation as the value of the Orthographic Neighborhood. The Surprisal of Mandarin and Cantonese was computed with a simplified Chinese GPT2

[3] https://github.com/bedlate/cn-corpus

trained on CLUE Corpus Small[4] (hereafter referred to as clue), while the Surprisal of Cantonese was additionally calculated with a traditional-Chinese GPT2 finetuned on cantonese-wikipedia for 10 epochs [5], which is referred to as jed351 below. For each round of Cantonese FD modeling, we fed one of the two Surprisals and finally reported the better performance. On account of the character-base tokenization of both clue model and jed351 model, we sum up the Surprisal score of each character c_k^i (the i-th character of the k-th word w_k in a sentence) to represent the exact score of the whole word. More in detail:

suppose $w_k = [c_k^1, c_k^2, ..., c_k^m]$), which means that the k-th word in the current sentence has m characters. Then the surprisal of the whole k-th word is represented as:

$$Surprisal(w_k) = \sum_{i=0}^{m} Surprisal(c_k^i) \quad (1)$$

suppose c_n is the n-th character in the whole sentence, then the surprisal of each character is:

$$Surprisal(c_n) = -\log(P(c_n|c_0, c_1, ..., c_{n-1})) \quad (2)$$

5.2.2 GPT Contextual Word Embeddings

To explore the effectiveness of contextualized word representation in improving eye-tracking prediction, we extracted the last hidden state of each word input from the GPT2 architecture to be concatenated with the linguistic features mentioned above. We used the clue model to extract GPT word embedding for both Mandarin and Cantonese. Since clue is basically trained on the Mandarin corpus, we equally used the jed351 model to extract embedding for Cantonese. We separately try one of the two types of Cantonese GPT embedding for each regressor.

All compositions of features tried in this research are summarized below. For each feature composition, we tried both interactions (using the PolynomialFeatures module in scikit-learn) and non-interaction between linguistic features and reported the best results.

[4] https://huggingface.co/uer/gpt2-chinese-cluecorpussmall.

[5] https://huggingface.co/jed351/gpt2_tiny_zh-hk-wiki

	Gpt Embedding	Other Features
Mandarin	noGpt	Linguistic Features (with clue Surprisal)
	clue	Linguistic Features (with clue Surprisal)
Cantonese	noGpt	Linguistic Features (with clue Surprisal)
		Linguistic Features (with jed351 Surprisal)
	clue	Linguistic Features (with clue Surprisal)
	jed351	Linguistic Features (with jed351 Surprisal)

Table 4: All possible composition of features for Mandarin and Cantonese.

5.3 Regressors

To propose an optimal prediction system, we utilized several regression models to approximate the eye-movement measurements concerned, using the implementations in the scikit-learn Python package and catboost package (Dorogush et al., 2018) (for GradientBoostDecisionTree only, due to its slow implementation without GPU acceleration). Below in Table 5 we listed the main hyper-parameters.

Regressors	Hyper-Parameters
BRR (BayesianRidge)	alpha=1.0, normalized=True
ELAST (ElastRegressor)	alpha=1.0 , l1_ratio = 0.5 , selection="cyclic"
GBDT (CatBoostRegressor)	num_leaves = 31 , learning_rate =0.03
LGB (LGBMRegressor)	objective='regression' , num_leaves = 31 , learning_rate =0.05
LR (LinearRegression)	fit_intercept=True
MLP (MLPRegressor)	hidden_layer_size=5, activation = identity, solver = adam
PLSR (PLSRegression)	n_components = 5
RF (RandomForestRegressor)	min_samples_split=2, min_samples_leaf =1
RR (Ridge)	alpha=1.0, normalize =True

Table 5: Regressor Parameter Settings

5.4 Metrics

To evaluate and compare the performance of the participating systems, we used the mean absolute error (MAE) in the 5-fold cross-evaluation as the main metric in the **Results and Discussion** section, as it increments linearly with the increases in the error. To complement, the mean squared error (MSE), the R-Square ($R2$), the Pearson correlation ($Pears.$), and the Spearman correlation ($Spear.$) for the 5-fold cross-evaluation are jointly reported for the best prediction system for each of

Y	Lang	Gptvec	brr	elast	gbdt	lgb	lr	mlp	plsr	rf	rr
FFD	Cantonese	-	38.86	38.87	36.19	38.82	38.48	38.99	38.78	**35.64**	38.47
		clue	35.19	36.81	**33.14**	37.81	35.62	35.72	36.27	33.76	35.60
		jed351	35.78	36.46	**33.20**	37.95	35.78	36.22	36.08	33.86	35.78
	Mandarin	-	36.49	36.56	**34.37**	37.20	36.48	36.79	36.69	34.49	36.36
		clue	34.34	35.46	**32.53**	36.64	35.33	35.13	35.02	33.80	35.28
SFD	Cantonese	-	24.54	24.53	**23.48**	24.81	24.49	24.63	24.49	24.98	24.49
		clue	23.39	24.05	**22.58**	24.53	24.06	23.78	23.92	23.73	24.05
		jed351	23.70	23.89	**22.59**	24.56	23.93	23.93	23.86	24.19	23.93
	Mandarin	-	24.38	24.58	**23.81**	25.22	24.38	24.81	24.38	25.38	24.38
		clue	23.84	24.30	**23.39**	24.96	25.08	24.35	24.08	24.64	25.06
TFD	Cantonese	-	74.17	74.03	**70.77**	74.24	74.13	74.55	74.11	74.76	74.11
		clue	68.96	70.31	**66.32**	72.87	71.34	71.34	71.03	68.30	71.29
		jed351	70.27	70.45	**66.41**	72.89	70.91	71.98	70.65	70.22	70.90
	Mandarin	-	73.50	74.07	**71.62**	75.63	73.57	74.46	73.55	77.35	73.56
		clue	70.71	71.60	**69.11**	74.82	74.90	73.61	72.05	73.22	74.84

Table 6: Performance (By MAE, lower is better) of different regressors (with and without GPT2 embeddings) on subject-normalized FFD, SFD, and TFD levels

the 6 FD measurements.

6 Results and Discussion

6.1 Regressor Performance

Table 6 presents the optimal MAE of each regressor in the prediction of the *mean levels* of **FFD**, **SFD**, and **TFD**. Table 7 shows all the metrics for the best system with and without GPT embedding. For the regressor selection, the **GBDT** regressor was dominantly the optimal choice for predicting eye-tracking data for the two Sinitic language varieties. In general, our prediction system is most helpful in approximating a human's first-pass eye-movement behavior, as the best *R2* scores were 44% and 41% for the Mandarin and Cantonese first fixation predictions, respectively (see Table 7). The correlation scores listed in Table 7 ranged between 0.57 and 0.66 for the **FD** mean value prediction and between 0.26 and 0.46 for the **FD** standard deviation prediction, demonstrating the predictability of the **FD** measurements in our dataset and the effectiveness of the features proposed in this research.

The utility of GPT embeddings was evaluated in this study, with Table 7 indicating that they are particularly effective in predicting FFD. Specifically, the performance (by $R2$) on mean level prediction for FFD in Mandarin and Cantonese was reinforced by 6% and 10%, respectively. However, GPT embeddings were found to be less helpful in predicting the mean level of SFD and TFD for both varieties. These results suggest that contextual semantics play a relatively marginal role in predicting non-initial fixation behavior for Mandarin and Cantonese.

The Pears correlation scores listed in Table 7 show a moderate correlation (0.4 - 0.6 for psychology) for most measurements between the ground truth and prediction, except for the standard deviation of Mandarin FFD (Akoglu, 2018).

6.2 Feature Usefulness

To investigate the usefulness of the linguistic features, we performed a series of ablation analyses against each feature in relation to the 6 measurements under discussion and found the change in *MAE* to be a metric suitable for measuring the usefulness. Intending to identify the pure usefulness of each feature, we restricted our ablation analyses to non-interaction **GBDT** regressors to avoid potential confusion due to cross-module interactions and regressor differences. In this paper, we mainly discuss the contribution of each feature to the *MAE* reduction of the mean level prediction.

Figure 4 presents each feature's usefulness (corresponding to positive values and highlighted in color) to the prediction of the mean level of each measurement. To facilitate the discussion, we divided the features into (1) Traditional Features utilized in psycholinguistic research, including **Frequency** (Word Frequency), $\mathbf{N}_{syl}$ (Syllable Count), **POS** (Part-of-speech), **Word Position**, **Prev Freq** (Previous Word Frequency), and **Prev**N_{syl} (Previous Syllable Number) (2) Newly-introduced features in this research, including **DEPTH**, **LDR** and **LDH**, **Surprisal**, and the **Neighbor** (Orthographical Neighborhood).

6.2.1 Traditional Features

The traditional features widely used in psycholinguistic research indicated the usefulness of all types

Y	Variety	+GPT Embedding						-GPT Embedding				
		Mapper	Gptvec	MAE	R2	Pears	Spear	Mapper	MAE	R2	Pears	Spear
FFD	Cantonese	gbdt-	clue	33.14	0.41	0.64	0.62	rf-	35.64	0.31	0.57	0.53
	Mandarin	gbdt+	clue	32.53	0.44	0.66	0.65	gbdt+	34.37	0.38	0.62	0.60
FFD_{std}	Cantonese	gbdt+	jed351	21.82	0.13	0.37	0.36	gbdt+	22.46	0.08	0.28	0.27
	Mandarin	lgb-	clue	22.03	0.06	0.26	0.28	lgb+	22.22	0.04	0.22	0.23
SFD	Cantonese	gbdt+	clue	22.58	0.32	0.57	0.52	gbdt+	23.48	0.28	0.53	0.45
	Mandarin	gbdt+	clue	23.39	0.33	0.58	0.56	gbdt-	23.81	0.31	0.56	0.54
SFD_{std}	Cantonese	gbdt+	clue	33.09	0.20	0.46	0.46	gbdt-	34.26	0.16	0.40	0.40
	Mandarin	gbdt+	clue	33.21	0.18	0.44	0.48	gbdt-	33.23	0.20	0.45	0.47
TFD	Cantonese	gbdt-	clue	66.32	0.36	0.60	0.60	gbdt-	70.77	0.31	0.56	0.51
	Mandarin	gbdt-	clue	69.11	0.37	0.62	0.66	gbdt-	71.62	0.36	0.60	0.61
TFD_{std}	Cantonese	gbdt+	clue	56.99	0.17	0.43	0.47	gbdt-	58.47	0.16	0.41	0.39
	Mandarin	brr-	clue	62.33	0.20	0.45	0.47	gbdt-	62.50	0.20	0.45	0.47

Table 7: The best model for each language variety on each fixation measurement. The "+" on the mapper denotes the introduction of interaction between linguistic features, while the "-" denotes the contrary.

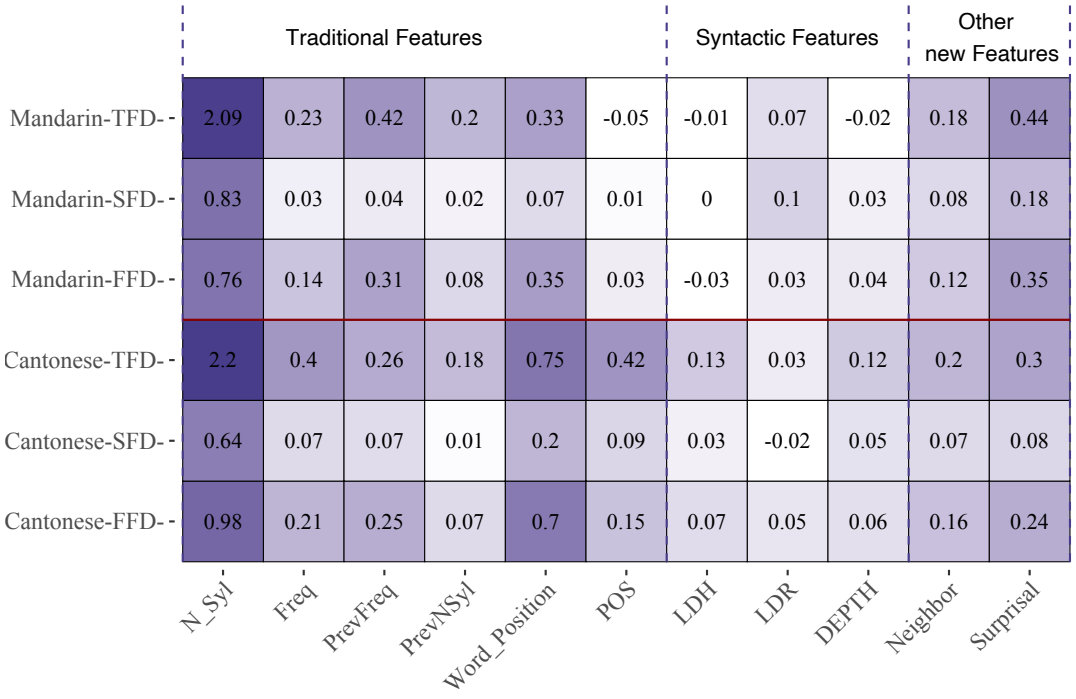

Figure 4: The usefulness of each feature based on ablation analyses of non-interaction models with no GPT embeddings.

of FDs in the two language varieties. In addition, most traditional features showed conspicuously less effectiveness in **SFD** prediction than in **FFD** and **TFD**prediction, except for the current syllabic length ($\mathbf{N}_{syl}$) for Mandarin, which showed more effectiveness in **SFD** than in **FFD**.

For the cross-variety comparison under discussion, it is worth mentioning that **Word Position** and **POS** are consistently more useful to Cantonese FD predictions. The stronger usefulness of word position in Cantonese is in line with the well-acknowledged typological statement that Cantonese exhibits a more robust canonical SVO order than Mandarin, whose word order shows the property of both SOV and SVO languages (Dryer, 1992, 2003; Liu, 2000).

6.2.2 Newly-introduced Features

Comparing with traditional features (except $\mathbf{N}_{syl}$), syntactic properties (**Depth**, **LDH**, **LDR**) are a bundle of features whose utility does not bleach as much in second fixation duration, which is consistent with suggestions from psycholinguistic re-

search that second-pass fixation is less dependent on lexical access than syntactic processing(Conklin and Pellicer-Sánchez, 2016)). Specifically, **LDR** stands out as the third most contributive feature in modeling Mandarin's second fixation duration, following **Surprisal** and $\mathbf{N}_{syl}$.

Neighbor and **Surprisal** also display overall effectiveness on all FDs. Specifically, **Surprisal** is the second most useful feature in the prediction of Mandarin **FFD** and **TFD**, following $\mathbf{N}_{syl}$. The finding that the **Surprisal** tends to be more beneficial to Mandarin can be attributed to the specific properties of the GPT models that we applied in this research, both of which take Mandarin text as their dominant training data (to the best of our knowledge, there are no publicly available GPT-like autoregressive Transformer models trained purely on Cantonese texts).

7 Conclusions

In this paper, we introduced an extensively annotated dataset of Mandarin and Cantonese eye-tracking data and shed light on their differences by features, such as word formation, word class, and word order. We also proposed a prediction system of fixation behaviors accompanied by new features from different modules, such as dependency features, the orthographic neighborhood, and GPT word embeddings, which were introduced with the goal of the computational prediction of Chinese eye-tracking data.

Based on a comparison of the regressor performance under different feature compositions, we investigated the usefulness of GPT vectors and linguistic features in reducing prediction errors. The results highlighted the effectiveness of our newly introduced features in modeling fixation patterns in representative Chinese language varieties and the importance of word order, part-of-speech, and syntax in addressing how Mandarin and Cantonese differ in language comprehension.

The findings in our study identify a few possible topics for future studies on language processing and regional syntactic variation of Chinese languages, such as how the syllabic structure, the visual complexity of different writing systems, pronominal resolution, syntactic relations, word order interact with gazing patterns and reading times of Chinese language speakers, especially for native speakers of different varieties. In addition to the varieties we studied here, we also plan to enlarge the dataset by including Mandarin processed through traditional Chinese characters, which is the standard system used in Taiwan. Finally, for future psychological and computational modeling studies, possible refinements of the representations in our experiment could be features targeting orthographic complexity and lexical ambiguity.

Limitations

The current study still has some limitations. For feature introduction, the GPT-based features are probably biased toward Mandarin text due to the position of Cantonese as a low-resource language. For the design of the prediction system, our approach is blind to the sequential properties of word-level fixation measurements. For future exploration, it would be promising to explore a sequential modeling approach.

Acknowledgments

We would like to thank the reviewers for their insightful feedback. This research was made possible by the start-up research funds (1-BE3F, and 1-BD8S) at the Hong Kong Polytechnic University. We also thank Deran Kong, Wenxi Fei, and Ka Keung Leon Lee for assisting with corpus-data collection.

References

Haldun Akoglu. 2018. User's Guide to Correlation Coefficients. *Turkish Journal of Emergency Medicine*, 18(3):91–93.

Joseph J Allaire. 2012. RStudio: Integrated Development Environment for R. *Boston, MA*, 770(394):165–171.

Salvatore Attardo and Lucy Pickering. 2023. *Eye Tracking in Linguistics*. Bloomsbury Publishing.

Xuejun Bai, Guoli Yan, Simon P Liversedge, Chuanli Zang, and Keith Rayner. 2008. Reading Spaced and Unspaced Chinese Text: Evidence from Eye Movements. *Journal of Experimental Psychology: Human Perception and Performance*, 34(5):1277.

Harry E Blanchard. 1985. A Comparison of some Processing Time Measures Based on Eye Movements. *Acta Psychologica*, 58(1):1–15.

Pi-Chuan Chang, Huihsin Tseng, Dan Jurafsky, and Christopher D Manning. 2009. Discriminative Reordering with Chinese Grammatical Relations Features. In *Proceedings of the NAACL-HLT Workshop on Syntax and Structure in Statistical Translation (SSST-3)*.

Po-Heng Chen and Jie-Li Tsai. 2015. The Influence of Syntactic Category and Semantic Constraints on Lexical Ambiguity Resolution: An Eye Movement Study of Processing Chinese Homographs. *Language and Linguistics*, 16(4):555–586.

Charles Clifton Jr, Adrian Staub, and Keith Rayner. 2007. Eye Movements in Reading Words and Sentences. *Eye Movements*, pages 341–371.

Kathy Conklin and Ana Pellicer-Sánchez. 2016. Using Eye-tracking in Applied Linguistics and Second Language Research. *Second Language Research*, 32(3):453–467.

Floor De Groot, Falk Huettig, and Christian NL Olivers. 2016. When Meaning Matters: The Temporal Dynamics of Semantic Influences on Visual Attention. *Journal of Experimental Psychology: Human Perception and Performance*, 42(2):180.

Avital Deutsch. 1998. Subject-predicate Agreement in Hebrew: Interrelations with Semantic Processes. *Language and Cognitive Processes*, 13(5):575–597.

Avital Deutsch and Shlomo Bentin. 2001. Syntactic and Semantic Factors in Processing Gender Agreement in Hebrew: Evidence from ERPs and Eye Movements. *Journal of Memory and Language*, 45(2):200–224.

Anna Veronika Dorogush, Vasily Ershov, and Andrey Gulin. 2018. CatBoost: Gradient Boosting with Categorical Features Support. *arXiv preprint arXiv:1810.11363*.

Matthew S Dryer. 1992. The Greenbergian Word Order Correlations. *Language*, 68(1):81–138.

Matthew S Dryer. 2003. Word Order in Sino-Tibetan Languages from a Typological and Geographical Perspective. *The Sino-Tibetan Languages*, pages 43–55.

Stefan L Frank. 2017. Word Embedding Distance Does not Predict Word Reading Time. In *Proceedings of CogSci*.

Cheryl Frenck-Mestre. 2005. Eye-movement Recording as a Tool for Studying Syntactic Processing in a Second Language: A Review of Methodologies and Experimental Findings. *Second Language Research*, 21(2):175–198.

Elisa Gironzetti. 2020. Eye-tracking Applications for Spanish Pragmatics Research. In *The Routledge Handbook of Spanish Pragmatics*, pages 517–531. Routledge.

Nora Hollenstein, Emmanuele Chersoni, Cassandra L Jacobs, Yohei Oseki, Laurent Prévot, and Enrico Santus. 2021a. CMCL 2021 Shared Task on Eye-tracking Prediction. In *Proceedings of the NAACL Workshop on Cognitive Modeling and Computational Linguistics*.

Nora Hollenstein, Emmanuele Chersoni, Cassandra L Jacobs, Yohei Oseki, Laurent Prévot, and Enrico Santus. 2022. CMCL 2022 Shared Task on Multilingual and Crosslingual Prediction of Human Reading Behavior. In *Proceedings of the ACL Workshop on Cognitive Modeling and Computational Linguistics*.

Nora Hollenstein, Federico Pirovano, Ce Zhang, Lena Jäger, and Lisa Beinborn. 2021b. Multilingual Language Models Predict Human Reading Behavior. In *Proceedings of NAACL*.

Sheng-Hsiung Hsu and Kuo-Chen Huang. 2000. Effects of Word Spacing on Reading Chinese Text from a Video Display Terminal. *Perceptual and Motor Skills*, 90(1):81–92.

Alex D Hwang, Hsueh-Cheng Wang, and Marc Pomplun. 2011. Semantic Guidance of Eye Movements in Real-world Scenes. *Vision Research*, 51(10):1192–1205.

Marcel Adam Just and Patricia A. Carpenter. 1980. A Theory of Reading: From Eye Fixations to Comprehension. *Psychological Review*, 87(4):329–354.

Regine Lai and Grégoire Winterstein. 2020. Cifu: A Frequency Lexicon of Hong Kong Cantonese. In *Proceedings of LREC*.

Jackson L. Lee, Litong Chen, Charles Lam, Chaak Ming Lau, and Tsz-Him Tsui. 2022. PyCantonese: Cantonese Linguistics and NLP in Python. In *Proceedings of LREC*.

David C. S. Li, Cathy S. P. Wong, Wai Mun Leung, and Sam T. S. Wong. 2016. Facilitation of Transference: The Case of Monosyllabic Salience in Hong Kong Cantonese. *Linguistics*, 54(1):1–58.

Xingshan Li, Pingping Liu, and Keith Rayner. 2011. Eye Movement Guidance in Chinese Reading: Is there a Preferred Viewing Location? *Vision Research*, 51(10):1146–1156.

Xingshan Li and Alexander Pollatsek. 2020. An Integrated Model of Word Processing and Eye-movement Control during Chinese Reading. *Psychological Review*, 127(6):1139.

Xingshan Li, Chuanli Zang, Simon P Liversedge, and Alexander Pollatsek. 2015. The Role of Words in Chinese Reading. *The Oxford Handbook of Reading*, page 232.

Danqing Liu. 2000. The Typological Properties of Cantonese Syntax . *Asia Pacific Journal of Language in Education* .

Simon P Liversedge, Chuanli Zang, Manman Zhang, Xuejun Bai, Guoli Yan, and Denis Drieghe. 2014. The Effect of Visual Complexity and Word Frequency on Eye Movements during Chinese Reading. *Visual Cognition*, 22(3-4):441–457.

Zijia Lu, Ying Fu, Manman Zhang, Chuanli Zang, and Xuejun Bai. 2022. Parafoveal Processing of Part-of-speech Information in Chinese Reading. *Acta Psychologica Sinica*, 54(5):441.

Jeff Mitchell, Mirella Lapata, Vera Demberg, and Frank Keller. 2010. Syntactic and Semantic Factors in Processing Difficulty: An Integrated Measure. In *Proceedings of ACL*.

Neal J Pearlmutter, Susan M Garnsey, and Kathryn Bock. 1999. Agreement Processes in Sentence Comprehension. *Journal of Memory and language*, 41(3):427–456.

Barbara Plank. 2022. The 'Problem' of Human Label Variation: On Ground Truth in Data, Modeling and Evaluation. In *Proceedings of EMNLP*.

Alexander Pollatsek, Barbara J Juhasz, Erik D Reichle, Debra Machacek, and Keith Rayner. 2008. Immediate and Delayed Effects of Word Frequency and Word Length on Eye Movements in Reading: A Reversed Delayed Effect of Word Length. *Journal of Experimental Psychology: Human Perception and Performance*, 34(3):726.

Joel Pynte, Boris New, and Alan Kennedy. 2008. On-line Contextual Influences During Reading Normal Text: A Multiple-Regression Analysis. *Vision Research*, 48(21):2172–2183.

Keith Rayner, Xingshan Li, Barbara J Juhasz, and Guoli Yan. 2005. The Effect of Word Predictability on the Eye Movements of Chinese Readers. *Psychonomic Bulletin & Review*, 12:1089–1093.

Keith Rayner, Xingshan Li, and Alexander Pollatsek. 2007. Extending the E-Z Reader Model of Eye Movement Control to Chinese Readers. *Cognitive Science*, 31(6):1021–1033.

Keith Rayner, Sara C. Sereno, Robin K. Morris, A. Réne Schmauder, and Charles Clifton Jr. 1989. Eye Movements and On-line Language Comprehension Processes. *Language and Cognitive Processes*, 4(3-4):SI21–SI49.

Lavinia Salicchi, Emmanuele Chersoni, and Alessandro Lenci. 2023. A Study on Surprisal and Semantic Relatedness for Eye-Tracking Data Prediction. *Frontiers in Psychology*, 14.

Lavinia Salicchi, Rong Xiang, and Yu-Yin Hsu. 2022. HkAmsters at CMCL 2022 Shared Task: Predicting Eye-tracking Data from a Gradient Boosting Framework with Linguistic Features. In *Proceedings of the ACL Workshop on Cognitive Modeling and Computational Linguistics*.

Philip Thierfelder, Gautier Durantin, and Gillian Wigglesworth. 2020. The Effect of Word Predictability on Phonological Activation in Cantonese Reading: a Study of Eye-fixations and Pupillary Response. *Journal of Psycholinguistic Research*, 49:779–801.

Marten Van Schijndel and William Schuler. 2015. Hierarchic Syntax Improves Reading Time Prediction. In *Proceedings of NAACL-HLT*.

Guoli Yan, Hongjie Tian, Xuejun Bai, and Keith Rayner. 2006. The Effect of Word and Character Frequency on the Eye Movements of Chinese Readers. *British Journal of Psychology*, 97(2):259–268.

Chuanli Zang, Hong Du, Xuejun Bai, Guoli Yan, and Simon P Liversedge. 2020. Word Skipping in Chinese Reading: The Role of High-frequency Preview and Syntactic Felicity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 46(4):603.

Chuanli Zang, Ying Fu, Xuejun Bai, Guoli Yan, and Simon P Liversedge. 2018. Investigating Word Length Effects in Chinese Reading. *Journal of Experimental Psychology: Human Perception and Performance*, 44(12):1831.

A Appendix

In this appendix, Figure 5 presents each FD's difference between Mandarin and Cantonese ($FD_{Mandarin} - FD_{Cantonese}$) grouped by part-of-speeches. FDs are in log scale. Differences above zero denote longer FD for Mandarin. Part-of-speeches involving significant differences are colored.

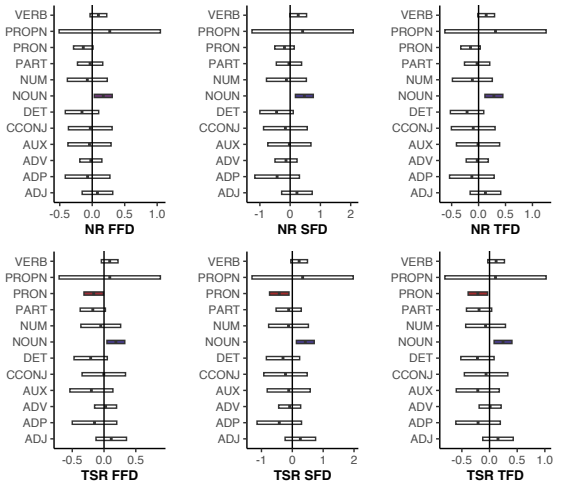

Figure 5: Tukey post hoc test of FD difference between paired part-of-speech in Mandarin and Cantonese (reporting 95%-level confidence intervals of the difference of"Mandarin-Cantonese"). FDs are in log scale. Word classes involving significant variance are colored. Positive difference means longer FD for Mandarin for the corresponding word category.