Scanpath Complexity: Modeling Reading Effort Using Gaze Information

Abhijit Mishra[†], Diptesh Kanojia[†,♣], Seema Nagar [⋆], Kuntal Dey[⋆], Pushpak Bhattacharyya[†]
[†]Indian Institute of Technology Bombay, India
[♣]IITB-Monash Research Academy, India
[⋆]IBM Research, India
[†]{abhijitmishra, diptesh, pb}@cse.iitb.ac.in
[⋆]{senagar3, kuntadey}@in.ibm.com

Abstract

Measuring reading effort is useful for practical purposes such as designing learning material and personalizing text comprehension environment. We propose a quantification of reading effort by measuring the complexity of eye-movement patterns of readers. We call the measure *Scanpath Complexity*. Scanpath complexity is modeled as a function of various properties of gaze fixations and saccades- the basic parameters of eye movement behavior. We demonstrate the effectiveness of our scanpath complexity measure by showing that its correlation with different measures of lexical and syntactic complexity as well as standard readability metrics is better than popular baseline measures based on fixation alone.

1 Introduction

In settings that require reading and understanding text, the effort spent by the reader is a factor of primary importance. In most scenarios, the reward associated with the task is often controlled by the effort spent on the task. For example, in education, the reading effort controls the motivation and learning experience of a student reading educational material. For text annotation that involves reading, the reading effort during annotation controls the financial incentives. Measuring reading effort reliably is, therefore, an important task. From an individual's perspective, it provides insights into one's cognitive capabilities, making it useful in designing personalized applications for learning (Sweller 1994) and optimizing learning material design (Mayer and Moreno 2003). From the perspective of Natural Language Processing, quantifying reading effort for text-annotation tasks may give rise to better *annotation-cost-models vis-à-vis* ones that rely on word and sentence counts, for incentivizing annotators (Tomanek et al. 2010).

Psychologists have attempted to create formalisms that capture the cognitive effort of reading processes using biological and psychological frameworks (Schnotz and Kürschner 2007). Exploratory work has been carried out under controlled environments, using Magnetic Resonance Imaging (MRI) (Paas et al. 2003), Electro- encephalography (Antonenko et al. 2010), etc. However, such techniques cannot be used outside laboratory settings, and are prohibitively expensive. Our method, on the other hand, relies on readers' eye-movement data which could be easily obtained using low cost eye-tracking machinery, for example, front web-cameras of hand held devices that are used to capture eye-movement behavior.

Our work is based on the eye-mind hypothesis (Just and Carpenter 1980) which states that when a subject views a word/object, he or she also processes it cognitively, for approximately the same amount of time he or she fixates on it. Though debatable (Anderson, Bothell, and Douglass 2004), the hypothesis has been considered useful in explaining theories associated with reading (Rayner and Duffy 1986; Irwin 2004; von der Malsburg and Vasishth 2011). The core idea of our work is the hypothesis that, gaze patterns indicate the conceptual difficulty the reader experiences (which, in turn, is linked with the cognitive effort (Sweller 1988)). Linear and uniform-speed gaze movement is observed over texts having simple concepts, and often non-linear movement with non-uniform speed over more complex concepts (Rayner 1998). We take a reader's eye-movement data in the form of scanpath[1] as input. The complexity of the scanpath, termed as *Scanpath Complexity*, is measured as a function of various properties of gaze fixations, saccades, and constituents of the input scanpath. Scanpath complexity is taken as a measure of reading effort.

To validate our scanpath complexity measure, we examine the correlation of scanpath complexity with different quantification of lexical and syntactic complexity and standard readability scores. For most of the participants whose eye tracking behaviour form our dataset, scanpath complexity correlates better with most of such complexity measures than does "total reading/annotation time" (or sum of fixation durations in an eye-tracking setup), which is often considered as a measure of effort (Tomanek et al. 2010; Mishra, Bhattacharyya, and Carl 2013; Joshi et al. 2014).

For our setup, we assume the reading direction to be left-to-right without loss of generality. The language under consideration for our experiments and analysis is English.

[1]Terminology: Fixation → relatively long stay of gaze on a visual object (like words in text), Sacccade → quick shifting of gaze between two positions of rest. Forward and Backward saccades are called Progressions and Regressions respectively. Scanpath → a line graph that contains fixations as nodes and saccades as edges

1.1 Feasibility of Getting Eye-tracking Data

Our method utilizes eye-movement patterns which can be reliably collected from inexpensive embedded eye-trackers. Inexpensive mobile eye-trackers are a reality now (Wood and Bulling 2014; Yamamoto et al. 2013). Leading mobile brands like Samsung have integrated eye-tracking facility on their devices enabling richer user experiences. This opens up avenues to get eye-tracking data from a large user-base non-intrusively.

The rest of the paper is organized as follows. Section 2 summarizes literature related to this work, primarily touching upon eye-tracking for reading research and scanpath analysis. In section 3, we explain our approach to model scanpath complexity. Section 4 describes various scanpath attributes, the constituents of our scanpath complexity model. Our experiment setup is detailed in section 5. Sections 6 is devoted to detailed evaluation of scanpath complexity. We discuss our results in Section 7 before concluding the paper in section 8

2 Related Work

Analyzing gaze data to gain insights into reading processes is a mature area of research (refer Rayner (1998) for an overview). A number of successful models of eye-movement control for reading include the one from Reichle and Laurent (2006), the E-Z Reader (Reichle, Rayner, and Pollatsek 2003; Reichle, Pollatsek, and Rayner 2006), SWIFT (Engbert et al. 2005) and Bayesian inference based models (Bicknell and Levy 2010; Engbert and Krügel 2010). Eye-movement in reading has also been analyzed to study the correlation of eye-movement parameters derived from fixations and saccades with the lexical and syntactic complexities of text. Rayner and Duffy (1986) show how fixation time is associated with different forms of lexical complexity in the form of word frequency, verb complexity, and lexical ambiguity. Demberg and Keller (2008) relate complex eye-movement patterns to the syntactic complexity present in the text. von der Malsburg and Vasishth (2011) show that complex saccadic patterns (with higher degree of regression) are related to syntactic re-analysis arising from various forms of syntactically complex structures (*e.g.,* garden-path sentence).

Scanpath analysis has been used in literature to evaluate users' perceived difficulty in contexts such as computer interfaces (Goldberg and Kotval 1999) and complex digital images on the Internet (Josephson and Holmes 2002). Works such as Underwood et al. (2003) and Williams et al. (1999) highlight the applications of scanpath analysis. Looking at the recent advancements, one can sense the growing importance of analyzing scanpath as a whole entity for *reading research*, instead of considering eye-movement attributes like fixations and saccades independently (Coco and Keller 2012; Holsanova, Holmberg, and Holmqvist 2009; von der Malsburg and Vasishth 2011). Methods have been proposed to compare scanpaths such as *ScanMatch* (Cristino et al. 2010) and the *ScaSim* similarity score (von der Malsburg and Vasishth 2011), and scanpath multi-match by Dewhurst et al. (2012). From the scanpath perspective, in an approach similar to ours, Malsburg, Kliegl, and Vasishth (2015) also propose a method to determine scanpath regularity, and observe that sentences with short words and syntactically more difficult sentences elicited more irregular scanpaths.

Eye-tracking has also been used to quantify annotation effort that involves reading. Tomanek et al. (2010) propose a cognitive cost for annotation based on eye-tracking data. Mishra, Bhattacharyya, and Carl (2013) measure translation annotation difficulty of a given sentence based on gaze input of translators who label the training data. Joshi et al. (2014) develop a method to measure the sentiment annotation complexity using cognitive evidence from eye-tracking. However, these methods are too simplistic in the sense that they take total annotation time (measured by summing fixation and/or saccade duration) as a measure of annotation effort. We believe that a deeper analysis of eye-tracking data is needed for measuring annotation effort than simply considering the total reading/annotation time. While it seems quite intriguing and realistic to apply our measure in these settings, it goes beyond the scope of this paper.

3 Modeling Scanpath Complexity

Scanpath complexity denoted as $ScaComp$ is proposed as a function of several attributes of the scanpath, that are derived from two basic properties: fixations and saccades. Mathematically,

$$ScaComp = f(X, \theta) \tag{1}$$

where X corresponds to a set of N attributes $x_1, x_2, x_3, ..., x_N$ (that we explain later in section 4) and θ corresponds to model parameters.

Now, the function f can be- (i) heuristically defined or (ii) learned automatically using supervised statistical techniques. The problem with designing a predefined function is that it is extremely difficult to know the dependencies between scanpath attributes and, hence, coming up with the most suitable f is difficult. On the other hand, in the supervised learning paradigm, one would need data-points capturing dependencies: in our setting this would mean obtaining reading effort scores from human readers. We propose two simple ways to model scanpath complexity following the two paradigms above.

3.1 Heuristic *ScaComp*

We assume scanpath complexity to be linearly proportional to each scanpath attribute. A *ScaComp* measure can then be given as,

$$ScaComp = \theta \times \prod_{i=1}^{N} x_i + C \tag{2}$$

where x_i is the value of the i^{th} attribute of the scanpath. θ is the constant of proportionality and C is another constant. Setting ($C = 0$ and $\theta = 1$), $ScaComp$ becomes a product[2] of the value of each attribute. For the rest of the paper, we represent this heuristic with the term $\boldsymbol{ScaComp\_H}$.

[2]To get a non-zero product, attributes with values as zero are discarded

3.2 Supervised *ScaComp*

Scanpath complexity can also be designed as a weighted sum of constituents. In the simplest form, thus,

$$ScaComp = \sum_{i=1}^{N} w_i x_i + C \tag{3}$$

with w_i representing the weight estimate for attribute x_i and C representing the intercept of the regression line. To estimate the model parameters (w and C), we rely on example data-points for which the dependent variable $ScaComp$ is available through manual annotation. For the rest of the paper, we use the term $\boldsymbol{ScaComp\_L}$ to present scanpath complexity measured following this approach. We now explain various attributes we have considered for modeling scanpath complexity.

In absence of prior baselines that address how the model attributes can be combined to get the most effective model possible, we considered two rudimentary functions (linear sum and product) prima facie. We draw inspirations from general science where it is very standard in case of modeling of physical phenomena to take product of all influencing factors or their inverses- as the case may be (e.g., in laws relating pressure, temperature and volume), and in case of statistical phenomena to use linear regression like expressions. We thought it is quite important to gain a first level insight, and most importantly, creating a baseline for future research. More data and more observations will refine the expression to capture reality more closely, we hope.

4 Scanpath Attributes

Various attributes corresponding to fixations and saccades combine to form scanpath complexity. We divide these attributes into two categories - Fixational attributes and Saccadic attributes, as explained in Table 1. Except for the last attribute (*negative saccade log likelihood*), all attributes are well known to the psycholinguistic community and have been used in a number of works (Holmqvist et al. 2011). The motivation behind why these attributes may be used to model reading behavior is well documented. Hence, we give a detailed explanation for the last attribute only. Also note that we do not normalize the attributes by text length assuming that reading effort is often associated with the length of the text, hence, normalization would rule out its effect.

4.1 Saccade (Un)likelihood

We first propose a saccade transition model that is based on an ideal reading behavior. It is often believed that (see discussion in the next paragraph), readers ideally perform saccades approximately half the time to the next word of the currently fixated word, the rest of the saccadic transitions are distributed amongst the words following the next word, the previous word and the currently fixated word.

Malsburg, Kliegl, and Vasishth (2015) find that in the Potsdam Sentence Corpus (Kliegl et al. 2004), 50% of the saccades target the next word in a sentence; in 19% of the saccades, the next word is skipped; 17% of the saccades result in refixations of the current word; and 8% are regressive saccades landing on the word directly preceding the current word. Other saccade targets are rare. Reading models like E-Z reader and SWIFT are based on such saccadic distributions as well.

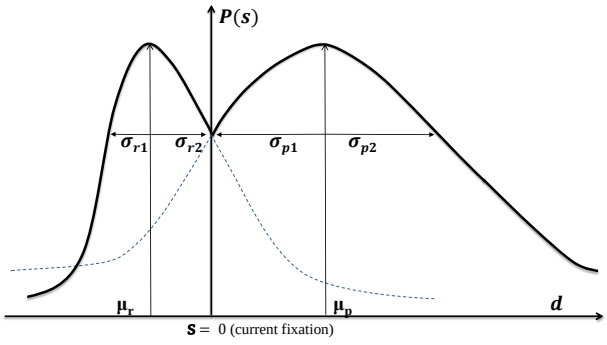

Figure 1: Distribution of saccade transition during reading. Subscripts p and r correspond to progressions and regressions respectively.

Based on these details, we propose a bi-modal ideal saccade transition distribution which comprises two asymmetric Gaussian[3] distribution (denoted by $\mathcal{N}_{assym}$); one for progressions and the other for regressions. The distribution is depicted in Figure 1.

At any point of time during reading, the probability of the next saccade of length s can be given as,

$$\begin{aligned} P(s) &= \psi * \mathcal{N}_{assym}(\mu_p, \sigma_{p1}, \sigma_{p2}) \\ &+(1-\psi) * \mathcal{N}_{assym}(\mu_r, \sigma_{r1}, \sigma_{r2}) \end{aligned} \tag{4}$$

where ψ is the probability of performing a progressive saccade, $1-\psi$ is the probability of performing a regressive saccade. μ_p , σ_{p1} and σ_{p2} are mean and standard deviation associated with the left part and the right part of the asymmetric Gaussian distribution for the progressive saccades. μ_r,σ_{r1} and σ_{r2} are mean and standard deviation associated with the left part and the right part of the asymmetric Gaussian distribution for the regressive scaccades.

The distribution $\mathcal{N}_{assym}$ with parameters μ , σ_1 and σ_2 can be described as,

$$\mathcal{N}_{assym}(\mu, \sigma_1, \sigma_2) = \frac{1}{Z} exp(-\frac{(s-\mu)^2}{2\sigma^2})$$

and,

$$\sigma = \begin{cases} \sigma_1, & s < \mu \\ \sigma_2, & s \geq \mu \end{cases}$$

and the normalization constant Z is given [4] by,

$$Z = \sqrt[2]{\frac{\pi}{2}}(\sigma_1 + \sigma_2)$$

[3]We chose asymmetric Gaussian over other similar distribution since it is easy to control the shape of the left and right part of the distribution

[4]Integrating $P(x_{t_{i+1}})$ from $-\infty$ to ∞ & equating to 1 yields Z

Attributes	Intent
Basic Fixational Attributes	
Total Fixation Duration *(FD)*	Sum of all fixation duration
Total First-Fixation Duration *(FFD)*	Sum of duration of fixations during the first pass reading of words
Total Regression-Fixation Duration *(RFD)*	Sum of duration of fixation on a regressed word
Total Fixation Count *(FC)*	Count of all fixations
Skipped Word Percentage (*SKIP*)	Fraction of words which have no fixation on them (or skipped)
Basic Saccadic Attributes	
Total regression count *(RC)*	Count of regressions
Total saccade distance *(SD)*	Sum of saccadic distance in terms of character count.
Total regression distance *(RD)*	Sum of regression distance in terms of character count
Complex Saccadic Attributes (Introduced by us)	
Negative Saccade log-likelihood *(NLL)*	Negative of the log-likelihood of saccade transitions with respect to an ideal saccade transition model (refer to section 4.1)

Table 1: Scanpath attributes considered as components of scanpath complexity

Such a hypothetical model should assign high probability to trivial saccades (*i.e.,* small progressions) and low probabilities to both short and large regressions (beyond one word) and extremely long progressions, which are highly improbable except in the scenario where simple skimming of text is done instead of attentive reading.

Considering an observed scanpath of N saccades, one indicator of irregularity/complexity of saccades can be given by how improbable the saccade transitions are with respect to the saccade transition model. This is captured by the cumulative negative log-likelihood (NLL) of all the saccades in a scanpath with respect to the saccade transition model. Mathematically,

$$NLL = -\sum_{i=1}^{N} log(P(s_i)) \tag{5}$$

where s_i is the length of the i^{th} saccade.

5 Experimental Setup

We compute scanpath complexity in two ways, by following equations 2 and 3. Our technique requires scanpath data to be available. To combine scanpath attributes using supervised statistical techniques (equation 3), we need data annotated with scores representing reading/annotation effort. Even though there exist a number of eye-movement datasets for reading, we could not find any dataset that has such annotation available. We, hence, create an eye-movement dataset which we briefly describe below.

5.1 Creation of Eye-movement Database

We collected 32 paragraphs of $50 - 200$ words on 16 different topics belonging to the domains of history, geography, science and literature. For each topic, two comparable paragraphs were extracted from **Wikipedia**[5] and **simple Wikipedia**[6]. This diversifies our dataset with respect to different dimensions- length, domains and linguistic complexity. The dataset can be freely downloaded[7] for academic use.

The documents are annotated by 16 participants. 13 of them are graduate/post-graduate students with science and engineering background in the age group of $20 - 30$ years, with English as the primary language of academic instruction. The other 3 are expert linguists and they belong to the age group of $47 - 50$. To ensure that they possess good English proficiency, a small English comprehension test was carried out before the start of the experiment. Once they cleared the comprehension test, they were given a set of instructions beforehand and were advised to seek clarifications before they proceeded further. The instructions mention the nature of the task, annotation input method, and necessity of head movement minimization during the experiment.

The eye-tracking experiment is conducted by following the standard norms in eye-movement research (Holmqvist et al. 2011). The task given to the participants is to read one document at a time and assign the paragraph with a "reading difficulty" score of 1 to 10. Higher scores indicate higher degree of difficulty. During reading, eye movement data of the participants (in terms of fixations, saccades and pupil-size) are tracked using an SR-Research Eyelink-1000 Plus eye-tracker. The eye-tracking device is calibrated at the start of each reading session. Participants are allowed to take breaks after two reading sessions to prevent fatigue.

5.2 Choice of NLL Model Parameters

Humans obtain useful information in reading (English text) from about 19 characters, more from the right of fixation than the left (Rayner 1998). For experimental purposes the parafoveal range[8] is often considered to be 7 characters to the left and 12 characters to the right of the current fixation (Bicknell and Levy 2010). Assuming that the probability of the next progressive/regressive saccade will, at maximum, be near the parafoveal boundaries, we fix the value of μ_r and μ_p to be -8 and 13 respectively. The shape parameters σ_{p1}, σ_{p2}, σ_{r1} and σ_{r2} (equation 4) are empirically set to $22, 18, 3, 13$ respectively by trial and error, plotting the distribution. Probability of regression $(1 - \psi)$ is kept as

[5]https://en.wikipedia.org/

[6]https://simple.wikipedia.org/

[7]http://www.cfilt.iitb.ac.in/cognitive-nlp

[8]Parafovea or the parafoveal belt is a region in the retina, that captures information within 2 degrees (approximately 6-8 characters) of the point of fixation being processed in foveal vision.

Attributes	Intent
Basic Properties	
Word Count *(W)*	
Sentence Count *(S)*	
Characters per Word *(C/W)*	
Syllables per Word *(S/W)*	
Words per Sentence *(W/S)*	
Readability Scores	
Flesch-Kincaid *(FK)* (Kincaid et al. 1975)	
Gunning-Fog *(GF)* (Gunning 1969)	
SMOG *(SMOG)* (Mc Laughlin 1969)	
LEXILE *(LEX)* (Stenner et al. 1988)	
Lexical Complexity	
Total Degree of Polysemy *(DP)*	Sum of number of Wordnet senses of all content words.
Lexical Sophistication *(LS)*	Lexical Sophistication Index proposed by Lu (2012)
Lexical Density *(LD)*	Ratio of content words to total number of words
Out-of-vocabulary Words *(OOV)*	Ratio of words not present in GSL (jbauman.com/gsl.html) and AWL (victoria.ac.nz/lals/resources/academicwordlist) to total words.
Syntactic Complexity	
Dependency Distance *(DD)*	Average distance of all pairs of dependent words in sentence (Lin 1996)
Non-terminal to Terminal ratio *(NN)*	Ratio of the number of non-terminals to the number of terminals in the constituency parse of a sentence
Clause per Sentence *(CL/S)*	
Complex Nominal per Clause *(CN/C)*	
Semantic Properties	
Discourse Connectors *(DC)*	Number of Discourse Connectors
Co-reference Distance *(CD)*	Sum of token distance b/w co-referring entities of anaphora in sentence
Perplexity *(PP)*	Trigram perplexity using language models trained on a mixture of sentences from the Brown corpus

Table 2: Textual properties,linguistic complexities and readability measures considered for evaluation

0.08, considering that around 8% of the total saccade transitions are regressions. While these parameters could be further tuned, we believe our choice of parameters is sufficient to provide a first-level insight. Note that our eye-movement data does not contain re-fixation information; we do not treat re-fixation as a separate gaze event. However, our NLL model, by its design, is capable of handling re-fixations.

5.3 Computing Scanpath Complexity

From the eye-tracking experiment, we obtain 512 unique scanpaths from 16 participants, each reading 32 paragraphs. Scanpath attributes are calculated using Python NUMPY and SCIPY libraries. As expected, annotation scores (which are to be taken as measures of scanpath complexity) obtained from participants are highly subjective and vary from person to person. We normalize these scores across all the documents for each individual by scaling them down to a range of [0,1]. Scanpath attributes are also normalized for computational suitability.

Baseline: As discussed earlier in section 1, in many reading and/or annotation settings, *total reading time* has been considered as a measure of effort. In eye-tracking setup, *total annotation time* often amounts to total fixation duration or total gaze duration. We consider total fixation duration as a measure of total annotation time which serves as a baseline in our setting.

To see how our gaze-attributes contribute to the $ScaComp\_L$ model, we perform a series of univariate linear regression tests where the cross correlation between each attribute and the dependent variable are measured and are converted to ANOVA F-scores and p-values. The F-scores for all the attributes considered in the linear regression model, taking the re-scaled human annotated score as the dependent variable are [FD: 146.14, FFD: 138.87, RFD: 84.10, FC: 154.62, SKIP: 4.71, RC: 85.92, SD: 159.32, RD: 138.12, NLL: 155.94]. It is worth noting that all the attributes turn out to be significant predictors in *99%* confidence interval.

We also perform a 10-fold cross validation to check how effective our complete set of gaze attributes are, as opposed to basic fixational and saccadic attributes alone. The average Mean Absolute Error for 10-folds turns out to be 0.1912 with all attributes, 0.1938 for basic fixational attributes and 0.1981 for basic saccadic attributes, showing that minimum error (statistically significant) is achieved when the complete set of gaze attributes is used in the linear regression setting.

6 Evaluation

Reading difficulties can broadly be related to two factors: (1) Linguistic complexity, textual attributes, readability of the given text *etc.* (2) Individual factors (age, domain knowledge and language skills). While the former is measurable through traditional NLP tools and techniques, the latter is hard to quantify. So we evaluate scanpath complexity using the various measures presented in Table 2, pertaining to linguistic complexity, textual attributes and readability. These textual properties are computed using Python NLTK API

	$Baseline$	$SComp\_H$	$SComp\_L$	p
	Basic Properties			
W	0.84	0.92	**0.94**	0.0001
S	0.41	0.50	**0.51**	0.0007
C/W	**0.56**	0.51	0.52	0.032
S/W	**0.46**	0.42	0.44	0.32
W/S	**0.55**	0.51	0.52	0.03
	Readability			
FK	**0.60**	0.56	0.58	0.02
GF	0.56	0.54	**0.58**	0.04
SMOG	0.57	0.56	**0.59**	0.03
LEX	0.58	0.58	**0.59**	0.008
	Lexical Complexity			
DP	0.61	0.70	**0.72**	0.0001
LS	**0.41**	0.35	0.33	0.008
LD	**0.30**	0.23	0.22	0.0004
OOV	**0.08**	0.03	-0.01	0.003
	Syntactic Complexity			
DD	0.56	0.55	**0.57**	0.008
NN	-0.05	-0.04	**-0.08**	0.1
CL/S	0.30	0.29	**0.32**	0.2669
CN/C	**0.69**	0.65	0.63	0.002
	Semantic Complexity			
DC	0.46	**0.53**	**0.53**	0.005
CD	0.30	0.30	**0.33**	0.13
PP	-0.02	-0.11	**-0.17**	0.0001

Table 3: Spearman's rank correlation coefficients between baseline and two forms of scanpath complexities, and linguistic complexities, textual attributes and readability measures. The coefficients are averaged over all the participants. p represents the two-tailed p-value of the t-test done between the best and the second best models, showing if the difference in the coefficients are significant with $p < 0.05$.

(Bird 2006), Stanford Core NLP tool (Manning et al. 2014) and tools facilitated by authors of referred papers.

We evaluate our techniques using Spearman's rank correlation coefficients between scanpath complexity and the linguistic complexity, basic textual and readability measures. This evaluation criterion is chosen to gain insights into whether any variation in such textual properties is related to the way scanpath is formed on the text. Since scanpath complexity is considered as a personalized measure, we compute the correlation coefficients for each participant to demonstrate the effectiveness of our technique. But, due to space limitations, we report correlation coefficients averaged across participants along with the standard deviations. Table 3 shows the averaged correlation coefficients. For measures pertaining to lexical complexity; the baseline method correlates well with the complexity measures. $ScaComp\_L$, on the other hand is better correlated with syntactic, semantic complexity measures and readability. $ScaComp\_H$ does not perform better the baseline for our dataset. However, we believe it can still be used as an alternative method for cases where manual annotation of cognitive effort becomes impracticable.

We perform a series of ablation tests to see how each scanpath component described in Table 1 affect our scanpath complexity measures. Ablation of one scanpath component

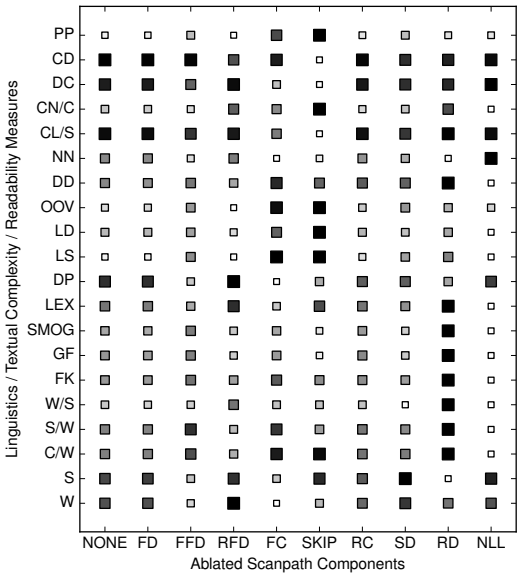

Figure 2: Results of ablation tests obtained by removing one scanpath attribute at a time in the $ScaComp\_L$ setup. Size and color intensity of the rectangles represents the average correlations (uniformly scaled) between the ablated $ScaComp\_L$ and linguistic complexities, textual attributes and readability measures presented in the y-axis. The ablated features are presented in X-axis. $NONE \rightarrow$ no ablation.

at a time, largely results in a reduction of correlation coefficients observed in both $ScaComp\_H$ and $ScaComp\_L$ settings. Due to space constraints, we report the ablation results only for $ScaComp\_L$ in Figure 2. It is worth noting that ablation of components like FD and RC, which are often used in psycholinguistic literature, results in a slight degradation of correlation values, whereas our proposed NLL measure proves to be very important, as its ablation results in a significant degradation.

We also tried ablating FD, RC and NLL together and observed a great reduction of correlation values. On the other hand, considering only these three components makes the model as good as the one with all components. Yet, in some cases, the "all-component" combination beats the "$FD - RC - NLL$" combination by a good margin.

7 Discussion

We now explain our observations (following Table 3) on the scanpath complexity measure and its relationship with various forms of textual nuances.

1. **Scanpath Complexity and Lexical Properties:** Fixation duration has been associated with lexical properties (*viz.* nature of words, number of syllables, their frequencies and predictability of words in a sentence) (Kliegl et al. 2004). This is probably why some measures of lexical complexities (*viz.,* lexical density, lexical sophistication and basic word level measures like characters per word, syllable per word) have better correlations with the total

	ID	Source	Baseline (Mean)	p_b	ScaComp_L (Mean)	p_l
1	27	Wikipedia	0.181	0.002	0.248	**6.28e-0.6**
	28	Simple Wikipedia	0.145		0.194	
2	01	Wikipedia	0.312	0.0002	0.495	**1.4e-0.5**
	02	Simple Wikipedia	0.227		0.409	

Table 4: Example cases from the dataset. ID→ID of the document in the released dataset, $Baseline$ → Average reading effort across all participants measured using baseline method. p_b p-value of a paired t-test between baseline scores obtained for all participants for Wikipedia and Simple Wikipedia documents. $ScaComp\_L$ →Average reading effort across all participants measured using supervised scanpath complexity method. p_l p-value of a paired t-test between baseline scores obtained for all participants for Wikipedia and Simple Wikipedia documents.

fixation duration, as compared to a combination of various fixation and saccadic attributes.

2. **Scanpath Complexity and Readability:** Scanpath complexity measures (especially $ScaComp\_L$) correlate better with simple readability measures like *SMOG* and *Lexile* scores. This shows the efficacy of scanpath complexity measures in capturing nuances causing reading difficulties and demanding more effort.
3. **Scanpath Complexity and Syntactic Properties:** We observe a stronger correlation between scanpath complexity and syntactic properties like dependency distance based structural complexity and clauses per sentence. This signifies the importance of saccadic attributes in the scanpath complexity formulation. After all, saccades have been quite informative about syntactic processing (Liversedge and Findlay 2000; von der Malsburg and Vasishth 2011).
4. **Scanpath Complexity and Semantic Properties:** While coreference distance is not significantly better correlated with scanpath complexity than total fixation duration, the correlation between scanpath complexity is stronger with the count of discourse connectors. It is believed that the presence of discourse connectors may increase the need of revisiting the constituent discourse segments, thereby, increasing regressive saccades. This is perhaps captured well by our scanpath complexity models.

It may be perceived that any weighted combination of enough variables will give a good correlation with the dependent variable. This is why we have reported several correlations between our model predictions and a number of lexical, syntactic, semantic and readability attributes. Since, these attributes are not considered as dependent variables in our model while fitting, better correlation values between our model and these variables should mean that our predicted values indeed capture the essence of reading effort with same or more accuracy than our baseline.

We present a few example cases from our dataset in Table 4 to justify the merit of $ScaComp\_L$ measure. Case 1 represents two paragraphs collected from Wikipedia (ID 27) and Simple Wikipedia (ID 28), covering the same topic. The paragraphs differ in terms of syntactic complexity though they exhibit similar lexical complexity. Similarly, for case 2, paragraphs from Wikipedia (ID 1) and Simple Wikipedia (ID 2) vary considerably in terms of Flesch Kincaid Readability as opposed to lexical and syntactic complexities.

For both the cases mentioned above, we compute the average baseline scores based on total reading time and $ScaComp\_L$ score for all 16 participants. As expected, the average scores for Simple Wikipedia paragraphs are lower than those of the Wikipedia ones for both baseline and $ScaComp\_L$. For each case, we performed a paired t-test to see if the difference between the measured values for Wikipedia and Simple Wikipedia documents are significant. As shown in Table 4, for both the cases and for both baseline and $ScaComp\_L$, the differences are statistically significant under 99% confidence interval (with hypothesized mean difference set to 0). However, the p-values for $ScaComp\_L$ are much lower (and hence, more significant) for both the cases than the baseline. This suggests that our proposed measure is more sensitive to linguistic complexities than the baseline.

8 Conclusion

Our work tries to model readers eye-movement behavior to quantify the cognitive effort associated with reading processes. We showed that the measurement of complexity of scanpaths leads to better cognitive models that explain nuances in the reading better than total annotation time, a popular measure of cognitive effort. We have validated scanpath complexity by obtaining correlation between the measure and various levels of linguistic complexities associated with the text.

Our work does not yet address effects of individual factors *(viz.* age, domain expertise and language skills) on scanpath complexity, studying which is on our future agenda. In future, we would also like to jointly model fixations and saccades for scanpath complexity measurement, instead of treating these attributes separately.

Acknowledgment

We thank the members of CFILT Lab and the students of IIT Bombay for their help and support.

References

Anderson, J. R.; Bothell, D.; and Douglass, S. 2004. Eye movements do not reflect retrieval processes limits of the eye-mind hypothesis. *Psychological Science* 15(4):225–231.

Antonenko, P.; Paas, F.; Grabner, R.; and van Gog, T. 2010. Using electroencephalography to measure cognitive load. *Educational Psychology Review* 22(4):425–438.

Bicknell, K., and Levy, R. 2010. A rational model of eye movement control in reading. In *Proceedings of the 48th annual meeting of the ACL*, 1168–1178. ACL.

Bird, S. 2006. Nltk: the natural language toolkit. In *Proceedings of the COLING/ACL on Interactive presentation sessions*, 69–72. Association for Computational Linguistics.

Coco, M. I., and Keller, F. 2012. Scan patterns predict sentence production in the cross-modal processing of visual scenes. *Cognitive Science* 36(7):1204–1223.

Cristino, F.; Mathôt, S.; Theeuwes, J.; and Gilchrist, I. D. 2010. Scanmatch: A novel method for comparing fixation sequences. *Behavior research methods* 42(3):692–700.

Demberg, V., and Keller, F. 2008. Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition* 109(2):193–210.

Dewhurst, R.; Nyström, M.; Jarodzka, H.; Foulsham, T.; Johansson, R.; and Holmqvist, K. 2012. It depends on how you look at it: Scanpath comparison in multiple dimensions with multimatch, a vector-based approach. *Behavior research methods* 44(4):1079–1100.

Engbert, R., and Krügel, A. 2010. Readers use bayesian estimation for eye movement control. *Psychological Science* 21(3):366–371.

Engbert, R.; Nuthmann, A.; Richter, E. M.; and Kliegl, R. 2005. Swift: a dynamical model of saccade generation during reading. *Psychological review* 112(4):777.

Goldberg, J. H., and Kotval, X. P. 1999. Computer interface evaluation using eye movements: methods and constructs. *International Journal of Industrial Ergonomics* 24(6):631–645.

Gunning, R. 1969. The fog index after twenty years. *Journal of Business Communication* 6(2):3–13.

Holmqvist, K.; Nyström, M.; Andersson, R.; Dewhurst, R.; Jarodzka, H.; and Van de Weijer, J. 2011. *Eye tracking: A comprehensive guide to methods and measures*. Oxford University Press.

Holsanova, J.; Holmberg, N.; and Holmqvist, K. 2009. Reading information graphics: The role of spatial contiguity and dual attentional guidance. *Applied Cognitive Psychology* 23(9):1215–1226.

Irwin, D. E. 2004. Fixation location and fixation duration as indices of cognitive processing. *The interface of language, vision, and action: Eye movements and the visual world* 105–134.

Josephson, S., and Holmes, M. E. 2002. Visual attention to repeated internet images: testing the scanpath theory on the world wide web. In *Proceedings of the 2002 symposium on Eye tracking research & applications*, 43–49. ACM.

Joshi, A.; Mishra, A.; Senthamilselvan, N.; and Bhattacharyya, P. 2014. Measuring sentiment annotation complexity of text. In *ACL (Daniel Marcu 22 June 2014 to 27 June 2014)*. ACL.

Just, M. A., and Carpenter, P. A. 1980. A theory of reading: from eye fixations to comprehension. *Psychological review* 87(4):329.

Kincaid, J. P.; Fishburne Jr, R. P.; Rogers, R. L.; and Chissom, B. S. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, DTIC Document.

Kliegl, R.; Grabner, E.; Rolfs, M.; and Engbert, R. 2004. Length, frequency, and predictability effects of words on eye movements in reading. *European Journal of Cognitive Psychology* 16(1-2):262–284.

Lin, D. 1996. On the structural complexity of natural language sentences. In *Proceedings of the 16th conference on Computational linguistics-Volume 2*, 729–733. ACL.

Liversedge, S. P., and Findlay, J. M. 2000. Saccadic eye movements and cognition. *Trends in cognitive sciences* 4(1):6–14.

Lu, X. 2012. The relationship of lexical richness to the quality of esl learners oral narratives. *The Modern Language Journal* 96(2):190–208.

Malsburg, T.; Kliegl, R.; and Vasishth, S. 2015. Determinants of scanpath regularity in reading. *Cognitive science* 39(7):1675–1703.

Manning, C. D.; Surdeanu, M.; Bauer, J.; Finkel, J. R.; Bethard, S.; and McClosky, D. 2014. The stanford corenlp natural language processing toolkit. In *ACL (System Demonstrations)*, 55–60.

Mayer, R. E., and Moreno, R. 2003. Nine ways to reduce cognitive load in multimedia learning. *Educational psychologist* 38(1):43–52.

Mc Laughlin, G. H. 1969. Smog grading-a new readability formula. *Journal of reading* 12(8):639–646.

Mishra, A.; Bhattacharyya, P.; and Carl, M. 2013. Automatically predicting sentence translation difficulty. In *ACL*. ACL.

Paas, F.; Tuovinen, J. E.; Tabbers, H.; and Van Gerven, P. W. 2003. Cognitive load measurement as a means to advance cognitive load theory. *Educational psychologist* 38(1):63–71.

Rayner, K., and Duffy, S. A. 1986. Lexical complexity and fixation times in reading: Effects of word frequency, verb complexity, and lexical ambiguity. *Memory & Cognition* 14(3):191–201.

Rayner, K. 1998. Eye movements in reading and information processing: 20 years of research. *Psychological bulletin* 124(3):372.

Reichle, E. D., and Laurent, P. A. 2006. Using reinforcement learning to understand the emergence of" intelligent" eye-movement behavior during reading. *Psychological review* 113(2):390.

Reichle, E. D.; Pollatsek, A.; and Rayner, K. 2006. E–z reader: A cognitive-control, serial-attention model of eye-movement behavior during reading. *Cognitive Systems Research* 7(1):4–22.

Reichle, E. D.; Rayner, K.; and Pollatsek, A. 2003. The ez reader model of eye-movement control in reading: Comparisons to other models. *Behavioral and brain sciences* 26(04):445–476.

Schnotz, W., and Kürschner, C. 2007. A reconsideration of cognitive load theory. *Educational Psychology Review* 19(4):469–508.

Stenner, A.; Horabin, I.; Smith, D. R.; and Smith, M. 1988. The lexile framework. *Durham, NC: MetaMetrics*.

Sweller, J. 1988. Cognitive load during problem solving: Effects on learning. *Cognitive science* 12(2):257–285.

Sweller, J. 1994. Cognitive load theory, learning difficulty, and instructional design. *Learning and instruction* 4(4):295–312.

Tomanek, K.; Hahn, U.; Lohmann, S.; and Ziegler, J. 2010. A cognitive cost model of annotations based on eye-tracking data. In *Proceedings of the 48th Annual Meeting of the ACL*, 1158–1167. ACL.

Underwood, G.; Chapman, P.; Brocklehurst, N.; Underwood, J.; and Crundall, D. 2003. Visual attention while driving: sequences of eye fixations made by experienced and novice drivers. *Ergonomics* 46(6):629–646.

von der Malsburg, T., and Vasishth, S. 2011. What is the scanpath signature of syntactic reanalysis? *Journal of Memory and Language* 65(2):109–127.

Williams, L. M.; Loughland, C. M.; Gordon, E.; and Davidson, D. 1999. Visual scanpaths in schizophrenia: is there a deficit in face recognition? *Schizophrenia research* 40(3):189–199.

Wood, E., and Bulling, A. 2014. Eyetab: Model-based gaze estimation on unmodified tablet computers. In *Proceedings of the Symposium on Eye Tracking Research and Applications*, 207–210. ACM.

Yamamoto, M.; Nakagawa, H.; Egawa, K.; and Nagamatsu, T. 2013. Development of a mobile tablet pc with gaze-tracking function. In *Human Interface and the Management of Information. Information and Interaction for Health, Safety, Mobility and Complex Environments*. Springer. 421–429.