An investigation of the effects of n-gram length in scanpath analysis for eye-tracking research

DOI:
10.1145/3204493.3204527

Document Version
Accepted author manuscript

An investigation of the effects of n-gram length in scanpath analysis for eye-tracking research

Manuele Reani
School of Computer Science
University of Manchester
m.reani@manchester.ac.uk

Niels Peek
Division of Informatics, Imaging and Data Sciences
University of Manchester
niels.peek@manchester.ac.uk

Caroline Jay
School of Computer Science
University of Manchester
Caroline.Jay@manchester.ac.uk

ABSTRACT

Scanpath analysis is a controversial and important topic in eye tracking research. Previous work has shown the value of scanpath analysis in perceptual tasks; little research has examined its utility for understanding human reasoning in complex tasks. Here, we analyze n-grams, which are continuous ordered subsequences of participants' scanpaths. In particular we studied the length of n-grams that are most appropriate for this form of analysis. We re-use datasets from previous studies of human cognition, medical diagnosis and art, systematically analyzing the frequency of n-grams of increasing length, and compare this approach with a string alignment-based method. The results show that subsequences of four or more areas of interest may not be of value for finding patterns that distinguish between two groups. The study is the first to systematically define the parameters of the length of n-gram suitable for analysis, using an approach that holds across diverse domains.

CCS CONCEPTS

• **Human-centered computing → Interaction techniques**; **Empirical studies in HCI**;

KEYWORDS

scanpath analysis, n-gram analysis, human reasoning, eye-tracking methodology

ACM Reference Format:
Manuele Reani, Niels Peek, and Caroline Jay. 2018. An investigation of the effects of n-gram length in scanpath analysis for eye-tracking research. In *ETRA '18: ETRA '18: 2018 Symposium on Eye Tracking Research and Applications, June 14–17, 2018, Warsaw, Poland*. ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3204493.3204527

1 INTRODUCTION

Understanding human reasoning is challenging, as the workings of the mind cannot be observed directly. Techniques such as brain imaging are costly and indirect methods such as think aloud or stimulated recall techniques may introduce biases and change the reasoning process [Blondon et al. 2015]. As a valuable alternative, eye tracking methods are an unobtrusive way of investigating human cognition [Glöckner and Herbold 2011; Horstmann et al. 2009; Keller et al. 2014; Schulte-Mecklenbeck et al. 2011]. Eye movements are an explicit measure of overt visual attention that can be used as an implicit indicator of hidden cognitive activities that are difficult to study using other methods [Coco 2009]. Eye tracking methods allow us to investigate which areas of the stimulus, and thus which pieces of information the viewer attends to, and the order in which these are visited, before making a decision. Areas of interest (AOIs) are defined on the stimulus where important items of information are contained either by the experimenter or using a data-driven approach. The order in which a person's gaze is directed to different AOIs forms a scanpath, or set of ordered fixations in different areas on the screen. People taking part in a study may have very different, often unique, scanpaths. One goal of scanpath analysis is to find subsequences among the scanpaths of one or more groups of participants, such that they represent a typical gaze behavior for that group. The methods developed for such an analysis have been successfully applied to tasks such as analyzing the processes involved in face recognition [Chuk et al. 2014] and for visualizing gaze data to understand how participants interact with Web applications [Goldberg and Helfman 2010a,b]. Nevertheless, we presently lack evidence of their utility for investigating visual behaviour in more complex tasks, such as probabilistic reasoning.

2 RELATED WORK

Two of the main forms of scanpath analysis are string-alignment methods and ngram frequency-based methods. Other methods, not addressed in the present manuscript, include vector-based methods, which take into account the length and the direction of saccades, methods using the Mannan distance which only compare the spatial properties of the scanpaths, ignoring the temporal dimensions, and methods based on saliency map comparisons–for a full review of these see [Le Meur and Baccino 2013]. String alignment methods, which have been used extensively in DNA sequencing, have also been used in eye tracking research for performing scanpath analysis, especially in user-experience research conducted on Web pages. Among these methods are the Multiple Sequence Alignment technique [Hembrooke et al. 2006], the String-edit algorithm [Heminghous and Duchowski 2006], the ScanMatch algorithm [Cristino et al. 2010] and the eMINE algorithm [Eraslan et al. 2015]. The ScanMatch algorithm is an adaptation of the Needleman and Wunsch algorithm that takes into account fixation duration [Cristino et al. 2010]. Other algorithms that take into account fixation duration

ETRA '18, June 14–17, 2018, Warsaw, Poland

ACM ISBN 978-1-4503-5706-7/18/06…$15.00
https://doi.org/10.1145/3204493.3204527

exist, for instance, the Scanpath Trend Analysis algorithm [Eraslan et al. 2016]. In this research we exclude fixation duration to focus primarily on the transition order of participants' scanpaths, which is of relevance in modelling human reasoning. Considering fixation duration would add complexity to the interpretation of the results, and is thus out of scope for the current paper. As an alternative, frequency-based methods can be used to perform transition analysis comparing parts of the participants' scanpaths. To avoid ambiguity, here we distinguish transition analysis from analyses using string-alignment methods, and n-grams from subsequences as follows: transition analysis, which is the main focus of the current paper, is concerned with how often the viewer transits between two or more AOIs; this method focuses on n-gram frequency. Conversely, string-alignment methods, e.g., the analysis performed using the eMINE algorithm [Eraslan et al. 2014, 2015; Yesilada et al. 2013], are concerned with finding a subsequence, often the longest possible, that is common among a group of participants. If letters are used to name the AOIs, the scanpath of each participant can be coded using a string of these letters, which represents the order in which the participant visits the AOIs. An n-gram is a continuous sequence of n letters that are present in the scanpath in a certain order for which only adjacent letters are considered. Conversely, a subsequence is a set of letters, also part of the scanpath, for which adjacent letters may be omitted. It follows that the possible n-grams of the scanpath ABC are A, B, C, AB, BC and ABC; whereas the possible subsequences of this scanpath include all the n-grams above plus AC; thus, omitting the central letter B. Frequency-based methods count frequencies of n-grams and compare the distributions of these in different groups [Kübler et al. 2017]. A distance measure can be used to find significant differences between groups in terms of distributions of n-gram frequencies [Davies et al. 2016; Kübler et al. 2014, 2017]. These techniques have been applied more or less successfully to viewing artwork, search tasks, video game play, driving tasks, and medical image interpretation [Davies et al. 2016, 2017; Kübler et al. 2014, 2017]

Hidden Markov Modelling, another frequency-based method, works by generating probability distributions for sequences of AOI transitions. However, the composite probabilities generated using these models may not represent the aggregate subsequences across a group of scanpaths due to the fact that often they do not go beyond first or second order Markov chains [Goldberg and Helfman 2010a]. Nevertheless, these methods have been successful at modelling visual perception in simple tasks (e.g., face recognition) and are particularly useful when the AOIs are not defined a priori [Chuk et al. 2014]. It is worth noting that the distinctive property of Markov processes is that the next state (i.e., the target location of a saccade) is determined only by its current state (i.e., the location where a viewer is currently fixating). Although this assumption may be appropriate for simple perceptual tasks, it may not hold for tasks investigating high level cognition, such as probabilistic reasoning.

2.0.1 Eye tracking in reasoning research. A recent systematic review examining the use of eye tracking for understanding human reasoning, especially in medical settings, reported that eye tracking studies in this area of research often use visualization techniques, such as heatmaps, measure fixation duration and frequency on defined AOIs, or measure how often AOIs are re-visited after a first inspection [Blondon et al. 2015]. The authors acknowledge the importance and novelty of using eye tracking methods for understanding reasoning processes; nevertheless, they also highlight the lack of breadth in eye tracking research in this area, suggesting that future research should investigate diverse reasoning processes in different domains [Blondon et al. 2015]. A recent study investigating numerical reasoning employed an ad-hoc transition analysis technique specifically developed for exposing how people reason about rational numbers [Plummer et al. 2017]. In this study, n-grams of fixations were compared to identify whether the reasoner was counting or comparing pieces of numerical information. Three or more consecutive fixations on a single AOI were seen as an indicator of counting behavior; conversely, two or more transitions between different AOIs within three fixations were seen as an indication of comparing behavior. Although effective for its purpose, this method of analysis is domain-specific and does not transfer to reasoning tasks in general [Plummer et al. 2017]. Transition analysis can, however, be used in more general terms, especially when a number of dissimilar pieces of information are provided and when the aim of the experimenter is to discover how and in which order these pieces of information are combined during a reasoning process. This method can focus on transitions between two AOIs only, called bi-grams (2-grams), or on longer n-grams. In theory, there is no limit to the length of the n-grams that can be used for analysis. In practice, however, the number of unique n-grams found in a list of scanpaths increases dramatically as n-gram length rises, reducing the value of its analysis for scientific purposes [Kübler et al. 2017]. However, by excluding all n-grams longer than two, we may miss important associations between eye movement and a variable of interest.

2.0.2 Research Objective. The aim of the present study is to investigate whether there is a relationship between the order in which people combine items of information in different tasks and a variable of interest (e.g., performance, treatment). It examines to what extent it is useful to consider n-grams longer than two to understand the difference in cognitive processes between two experimental groups of participants. We compare the analysis of n-grams of increasing length with a string alignment method and investigate what transition analysis can tell us about human cognition. To this end we perform secondary analyses of eye tracking datasets from three diverse studies.

3 METHOD

For each dataset, we counted the occurrences of n-grams of increasing length in participants' scanpaths for two comparison groups (e.g., correct reasoners vs. incorrect reasoners). The frequencies were then normalized by the total count in a given group to obtain probabilities. This enables us to identify distinctive transitions between AOIs, of predefined variable lengths, that represent gaze behaviour in different groups of reasoners. After counting the occurrences of given n-grams, we then used an odds-ratio scale to find which n-grams have the largest relative frequency, and which are thus responsible for the largest difference in the distribution of n-gram frequency between two groups (e.g. correct vs. incorrect respondents)–see Equation 2. Odds-ratios values closer to 1 represent similarities in n-grams relative frequency (between groups),

whereas values larger or smaller than one represent differences in n-gram relative frequencies between the compared groups. The AOIs in the stimuli were coded using letters that define a specific alphabet for each stimulus. The possible n-grams with a given alphabet were then calculated using combinatorics, and their occurrence in the scanpaths of the two comparison groups counted. These permutations were calculated by collapsing consecutive fixations in the same area (e.g., for AABCD we only consider ABCD). We thus limited our investigation to transitions between different AOIs, as we are interested specifically in how people combine different items of information. Thus, for example, for an alphabet of ten AOIs, we would have $10 \times 9 = 90$ 2-grams, $10 \times 9 \times 9 = 810$ 3-grams, $10 \times 9 \times 9 \times 9 = 7,290$ 4-grams and $10 \times 9 \times 9 \times 9 \times 9 = 65,610$ 5-grams. In the case in which the order of the fixations were not of interest for the analyst, this method could be generalized to comparisons which discard the order of the AOIs within a subsequence–i.e., treating, for instance, 3-grams ABC, ACB, BAC, BCA, CAB and CBA all as equivalent. This could be accomplished by simply computing the possible combinations, instead of the permutations, of the AOIs in the alphabet. Then, one would just need to count the occurrences of these combinations, as it was done in the present study, for calculating the frequencies and deriving the distributions. We use the Hellinger distance (see Equation 1) to determine the difference between the frequency distributions of n-grams between the two groups. This metric is commonly used in non-parametric methods, and provides a stable measure of difference between two discrete distributions that is not as sensitive to zero occurrences as other distances based on entropy–e.g., the Kullback-Leibler divergence [Hellinger 1909; van Erven and Harremoës 2014].

$$H(P,Q) = \frac{1}{\sqrt{2}} \sqrt{\sum_{x \in X} \left(\sqrt{p(x)} - \sqrt{q(x)}\right)^2} \quad (1)$$

A permutation test was used for performing statistical inference using the Hellinger distance–see Equation 1–where P and Q are the two distributions to be compared and x is the vector of possible n-grams. This metric is a measure of how much two distributions overlap, and this was used as the dependent variable of the test. The distance between two distributions of n-grams (e.g. an n-grams distribution for the correct group Vs an n-grams distribution for the incorrect group) is a representation of the difference between two comparable samples. We wished to determine whether the difference between two groups was large enough to reject the null hypothesis that the two n-grams distributions came from the same population. This could show an association between a variable of interest (e.g., correctness) and eye-movement (e.g., n-grams distribution). This approach has two main advantages. Firstly, permutation tests are known to be robust against type one error [Wilcox 2010]. Secondly, using a permutation test allows us to obtain an estimate of the sampling distribution, the shape of which can then be analyzed to better understand the underlying behavior. This method was also used to find which n-grams were responsible for the largest difference between groups. Here we examined n-grams with a length of two to five. Longer n-grams were not compared as the potential number of n-grams rises exponentially, and the chance of a given n-gram occurring, therefore, drops dramatically beyond an n-gram length of three. It follows that the analysis of longer n-grams ceases to be useful if the aim of the analysis is to find repetitive patterns in the data that discriminate between groups [Kübler et al. 2017]. Furthermore, if certain n-grams were present in both groups with similar frequency, these would not be useful for finding differences between, for example, correct and incorrect respondents.

For comparison, we also performed scanpath analysis using a string alignment method (the eMINE algorithm) from Eraslan et al. [2015]. The eMINE method combines the Levenshtein Distance with the longest common subsequence technique. This algorithm is relatively easy to implement, is flexible without constrains on subsequence length, is not computationally intensive and appears to be efficient at finding the average common subsequence of a list of scanpaths in tasks often employed in user-experience research [Eraslan et al. 2014, 2015; Yesilada et al. 2013]. However, this method has not yet been applied to more complex tasks, such as probabilistic reasoning, and does not offer techniques for performing inferential statistics, such as comparing two or more groups of participants, assigned to different treatments. Nevertheless, this method should theoretically find a common subsequence with a non-predefined length within a list of scanpaths.

3.0.1 Datasets. We applied these methods to three datasets from different studies that had been conducted within our HCI research group; each one is a subset of the data obtained in the original study, selected for specific characteristics of the stimuli, which are explained below. We used datasets from studies investigating probabilistic reasoning and medical diagnosis, a form of expert reasoning. To show the versatility of our method, we also use a dataset from a study investigating the visual perception of paintings.

The first study ("Weather") asked 49 participants to estimate the probability of rain given information about barometric pressure and historical weather data in a fictional city; in this study, participants had to solve a problem, which required Bayesian reasoning [Reani et al. 2018]. The data were presented with a graph (a tree diagram), and the experiment examined associations between eye movement and probabilistic reasoning ability. For this form of reasoning, expertise in a specific domain is not required. There was another subset in this study, which presented the same problem using Venn diagrams. We chose the "Tree" subset because this representation shows complete information, whereas the Venn representation showed only part of the data; thus, the former was more suitable for our purpose of investigating how people combine different pieces of information during reasoning tasks. The second dataset ("ECG") was taken from a study investigating the relationship between ECG (electrocardiogram) interpretation and eye movements [Davies et al. 2016]. A group of 43 medically trained participants were asked to look at an ECG of a patient and infer the underlying medical condition. We used data from the "Anterior Lateral Stemi" stimulus of the original study. There were eleven other subsets related to other stimuli, which are omitted here. The Anterior Lateral Stemi was the only subset with fairly equal groups; i.e., the number of participants who answered correctly was similar to the number of participants who answered incorrectly. As the aim of the study was to compare groups, these groups needed to have similar, or at least comparable, numbers of participants in them. In these studies the AOIs were defined by the experimenter, in a top down manner, over meaningful items of information (see Figure 1).

In both studies, the variable used to define participants was accuracy, i.e., whether they correctly answered the reasoning problem, or correctly inferred the ECG condition. The third dataset ("Art") was taken from a study which looked at the perception of paintings [Davies et al. 2017]. In this study 40 participants were asked simply to look at a series of paintings for ten seconds each. One group read a textual description of each painting before viewing it; the other group did not. The aim of the study was to determine whether the description influenced the way people subsequently viewed the painting. We used the subset related to the painting of "Sir Gregory Page-Turner" by Pompeo Batoni, as the other subsets were related to paintings showing scenery or abstract images, which were less suited to the aim of our study, which was to investigate how people combine distinctive (separated) pieces of information in laboratory tasks. The AOIs in this study were generated using a clustering algorithm (DBSCAN) over the fixation data, and were thus entirely data driven [Sander et al. 1998]. This dataset was used to assess whether the method proposed could be generalized to contexts for which participants are not asked to perform any particular task, and in which the AOIs are generated in a bottom-up fashion. In this study the groups were defined by the treatment–i.e., whether or not they read a description.

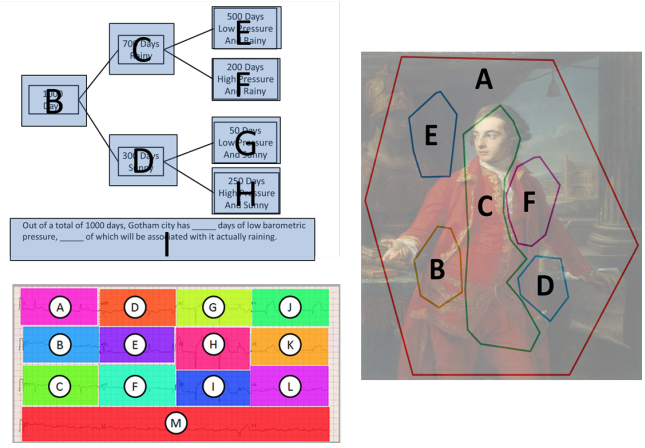

Figure 1: AOIs for the Weather problem (top left), the ECG problem (bottom left) and the Art study (right). The right image is the painting Sir Gregory Page-Turner held by the Manchester Art Gallery (image courtesy Manchester Art Gallery).

Figure 1 shows the AOIs of the three datasets. In the first two studies (top left and bottom left) the AOIs were defined by the experimenter, and represented specific items of information that were used by the participants to answer the problem. In the first study in which participants had to estimate the probability of rain (top right corner), B is the total number of days in a given period of time, C is the number of rainy days, D is the number of sunny days, E is the number of rainy days with low barometric pressure, F is the number of rainy days with high barometric pressure, G is the number of sunny days with low barometric pressure, H is the number of sunny days with high barometric pressure and I is the question. In the case of the ECG study, AOIs were defined over the ECG leads. AOIs in the painting study were defined using a clustering algorithm based on the concentration of fixations across the image, when considering all the participant data at once.

4 RESULTS

As an example, Figure 2 shows the fixation and the 2-gram heatmaps for the Weather dataset. On the left are the heatmaps auto-generated by the eye tracker, on the right are the heatmaps representing the frequency of 2-grams, for correct and incorrect groups. Some differences can be observed in the distribution of fixations across the screen, and in the 2-gram distributions between groups.

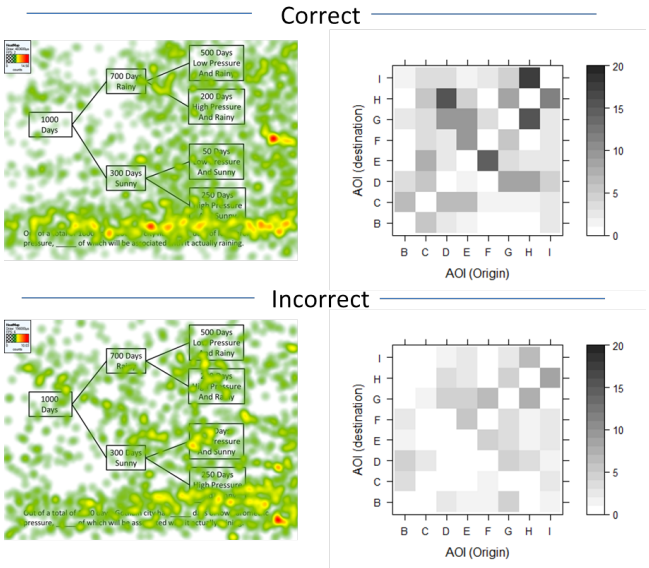

Figure 2: Heatmaps for fixations (left) and 2-gram frequency (right) for Correct (top) and Incorrect (bottom) group. On the left, the intensity of the colours represents an increased number of fixations' frequency. On the right, the increased intensity of the grey-scale represents higher transition frequency.

4.0.1 Analysis based on n-grams. The frequency of n-gram occurrences in the list of participants' scanpaths drops exponentially as n-gram length increases (see Figure 3). The number of n-grams with zero hits in the 4-gram and 5-gram analysis was very large, with only a few n-grams occurring with a frequency higher than one. The frequency distribution for 2-grams and for 3-grams were much wider; they included several n-grams that were found up to 30 times across participants' scanpaths.

Twelve permutation tests, with 10,000 permutations each, were performed to find significant differences (measured by the Hellinger distance) between two distributions of n-grams (correct vs. incorrect or description vs. no-description) for the three datasets, from 2-grams to 5-grams. These results are reported in Table 1. An example of the output is shown in Figure 4. This represents, for the Weather dataset, the distributions of distances of randomly sampled groups

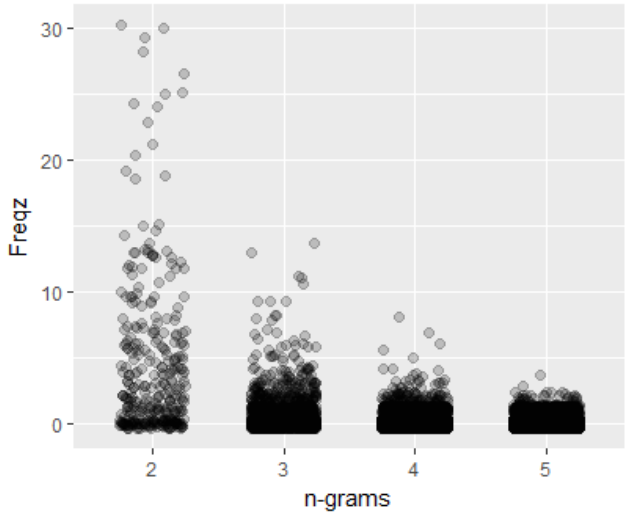

Figure 3: Frequency of n-grams by length.

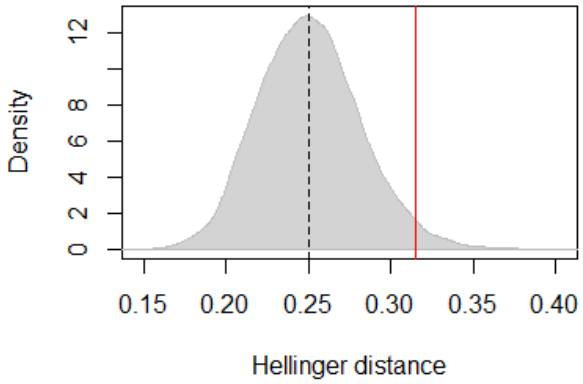

Figure 4: Distribution of distances for 2-grams in the Weather dataset. The red vertical line represents the value of the Hellinger distance between the 2-gram distribution for the correct group and the 2-gram distribution for the incorrect group.

of 2-grams–i.e., two groups of comparable sizes were created 10,000 times, and the Hellinger distance was calculated for each pair, by randomly selecting samples (with replacement) from the full dataset. In the graph, the vertical continuous red line represents the distance between the Correct and the Incorrect groups.

Table 1: Hellinger distances between the distributions of n-grams and related p-values

	Weather		ECG		Art	
n-gram	H.dist	P.value	H.dist	P.value	H.dist	P.value
2-gram	0.32	***0.02***	0.34	0.24	0.18	0.69
3-gram	0.65	0.29	0.67	0.32	0.48	***0.10***
4-gram	0.88	0.55	0.88	0.35	0.72	0.32
5-gram	0.96	0.71	0.97	0.26	0.88	0.64

From Table 1, it can be noted that the distance between the two distributions tends to increase when the length of the n-grams increases–i.e., as we go from 2-gram to 5-gram this distance gets closer to one, indicating a very large difference between the two distributions, for all three datasets. This is in line with the fact that as the n-gram length increases the uniqueness of these n-grams also increases. A significant p-value, for an alpha = 0.05, is found only for the 2-grams analysis in the Weather dataset. Another low p-value is found in the Art dataset, for the 3-grams analysis. The p-values in the present context are indicative of whether a difference between the n-grams distributions between two comparable groups was likely to be due by chance. Smaller p-values therefore indicate that this difference was unlikely to be caused by random errors, and was likely associated with the manipulation of the variable of interests (e.g., correctness). For these two permutation tests we determine which n-grams are responsible for the difference between the distributions by using the odds-ratio scale which is a representation of the differences in n-gram relative frequency between two groups. This is defined in Equation 2 in which p (for group one) and q (for group two) are the probabilities associated with each n-gram, normalized by the total of each of the groups. Their confidence intervals are also reported to demonstrate the precision of the odds-ratio scale.

$$OR = \frac{\frac{p}{1-p}}{\frac{q}{1-q}} \tag{2}$$

For the Weather datasets, the analysis of 2-grams shows that 2-gram CE (OR = 7.49, 95% CI [0.97, 57.57]) and 2-gram CH (OR = 6.314, 95% CI [0.81, 49.16]) had the largest odds ratios and 2-gram FB (OR = 0.19, 95% CI [0.019, 1.8]) and 2-gram GB (OR = 0.08 95% CI [0.01, 0.64]) had the smallest. For the Art datasets, the analysis of 3-grams shows that the 3-gram DCD (OR = 5.28, 95% CI [0.63, 44.11] and 3-gram ADA (OR = 4.49, 95% CI [0.51, 37.77]) had the largest odd-ratios and 3-grams FAC (OR = 0.14, 95% CI [0.02, 1.2]) and 3-gram CAF (OR = 0.24, 95% CI [0.05, 1.18]) had the smallest. These n-grams are the locations where the distributions of n-grams differ most and they highlight important links between eye movement and the variable of interest.

The present method focuses mainly on drawing statistical inference from eye-tracking data which is used to explore cognitive processes in three different tasks. For this reason, one needs to extract enough data to be able to generalize the findings. As demonstrated by the current results, the use of longer n-grams would produce very different subsequences that are unique to some individuals. Nevertheless, as shown in Figure 3, there are some 4-grams and 5-grams with a frequency of five or more hits. These may expose some interesting patterns. However, most of these longer n-grams have odds-ratios close to one. This means that they were found in the scanpaths of participants from different groups with equal or

very similar frequency, and therefore did not serve to distinguish between the groups. The hit frequency of most of the 4-grams and 5-grams with odds-ratios large (or small) enough to show any difference was less than four, with only one exception. The 4-gram HDHG in the Weather dataset (OR = 5.62, 95% CI [0.68, 46.72]) is the only long n-gram with an odds-ratio deviating from 1, thus indicating a difference between groups, and a hit frequency higher than four.

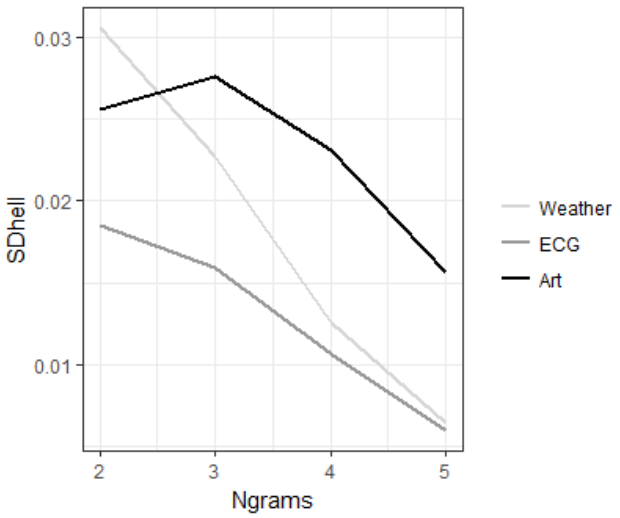

Figure 5: Standard deviation values of the sampling distribution of the permutation tests using the Hellinger distance, across the 3 datasets for different n-grams.

When the length of the n-grams increases, the standard deviation of the distributions of the distances in the permutation tests tends to reduce (see Figure 5). This can also be seen in Figure 6, which represents the entire shape of the distributions across datasets. An extreme reduction in variance is not beneficial here because it introduces a narrowness bias, which results in an unrepresentative estimate of the sampling distribution [Hesterberg 2015]. This phenomenon suggests that the analysis of long n-grams may not provide any useful information as a large Hellinger distance and a narrow variance are unlikely to show any discriminative patterns of eye movement between groups of scanpaths. This provides evidence that 2-gram analysis, which considers transitions only from one location to another, is likely to be the most appropriate for most datasets. However, as shown in Figure 5 and 6, the variance in the distribution actually increases for the Art dataset when moving from 2-grams to 3-grams, and then it decreases slightly for longer n-grams. For this dataset the analysis of 3-grams may be more meaningful (see discussion). This is also consistent with the small p-value found in the permutation tests when analyzing 3-grams for this dataset.

4.0.2 Analysis based on string alignment. The scanpath length across all the conditions and datasets was very diverse, ranging from one to 130 characters (see Table 2).

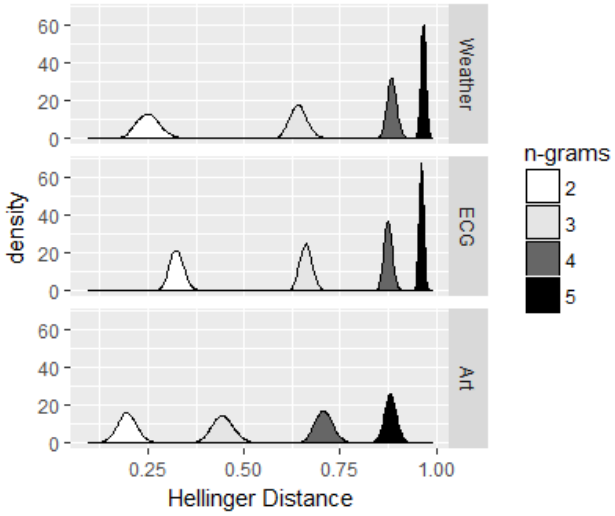

Figure 6: Shape of the sampling distributions of the permutation tests using the Hellinger distance for the 3 datasets for different n-gram lengths. The narrowness of the distribution represents lack of variability in distances between the randomly sampled groups.

Table 2: Statistics for scanpath lengths across datasets representing the minimum, maximum, median, interquartile interval and 1st quartile.

Data	Min	Max	Median	IQI	Q1
Weather	1	43	14	12	8
ECG	1	130	24	17.5	15
Art	1	24	13	5	10

String alignment methods such as the eMINE algorithm can be reductionist (i.e., they may produce a single short subsequence that is unable to capture an average behavior across the participants in a group) especially if there is a large diversity in participants' scanpaths [Eraslan et al. 2015]. For instance, if the scanpath of a participant of one group is just one letter, this would collapse all the comparisons to that letter. To minimize this problem, we excluded from the analysis all the participants who had a scanpath length of less than eight characters, which was the smallest value of the first quartiles (Q1) for the datasets (see Table 2). This left 38 participants for the Weather dataset, 39 for the ECG dataset and 32 for the Art dataset. After analyzing the data with the eMINE algorithm we found that for the Weather and the ECG datasets the algorithm was unable to identify a common subsequence for either group. For the Art dataset, the algorithm identified the subsequence CCC for the group with the description of the painting and CA for the group without the description.

5 DISCUSSION

The analysis of n-grams shows that although the frequency of occurrence drops exponentially with an increase in n-gram length,

2-gram analysis may not be an optimal choice for every type of dataset. It appears that, at least for the Weather and the Art datasets, eye movement is related to performance/treatment. This is found only for 2-gram analysis in the Weather dataset and for 3-gram analysis in the Art dataset. The n-grams with the largest and smallest odds-ratios show the most distinctive gaze strategies adopted by different groups and the confidence intervals of these odds-ratios provide information about the reliability of these results. We also found a narrowness bias for the n-gram analysis with longer n-grams (4-grams and 5-grams). The variance of the sampling distribution decreases when moving from 2-grams to 5-grams and the rate of this decrease is dependent on the alphabet size. As shown in Figures 5 and 6, the variance decreases steeply for the ECG dataset for which the alphabet size is larger and less steeply for the Weather dataset for which the alphabet size is smaller. For the Art dataset, the variance increases slightly from 2-grams to 3-grams and then decreases more gradually towards longer n-grams. These results, in line with Kübler et al. [2017], indicate that longer n-grams may not provide useful information because as n-gram length increases participants' eye movement patterns diverge further. However, when the alphabet size is small enough, increasing the length of the n-grams from two to three seems to be beneficial. Indeed, with an increase in the alphabet size from six AOIs (e.g., the Art dataset) to eight AOIs (e.g., the Weather dataset) we found a decrease in variance in the sampling distribution and a larger p-value as produced by the permutation tests. A heuristic for choosing the right length of n-gram for analysis could be to choose the length that yields the largest variance in the estimate of the sampling distribution. In our study, the length that yielded the largest variance was 2-gram for the Weather dataset and 3-gram for the Art dataset. This result points to the high reliability of short subsequences in scanpath analysis which is in line with past results obtained from analyses using HMM which, often, do not go beyond first or second order Markov chain, as there is seldom enough data to estimate a growing number of higher order probabilities [Goldberg and Helfman 2010a; Hayes et al. 2011].

The results from the analysis based on string alignment, using the eMINE algorithm [Eraslan et al. 2015], show that even when eliminating participants with scanpaths of less than eight characters, the algorithm was able to find a common subsequence only for the Art dataset, possibly due to the fact that the alphabet for this dataset was smaller (only six letters) than the other datasets (8 letters for the Weather and 13 letters for ECG datasets). With a larger alphabet size the diversity in the list of scanpaths can increase dramatically. These results suggest that this method might not be able to identify representative common subsequences in a group of participants' scanpaths if there is a large number of AOIs. However, as mentioned by Eraslan et al. [2014], this phenomenon can be attributed to the hierarchical structure of the eMINE algorithm; by adding a constraint to prevent losing the AOIs present in all individual scanpaths in the intermediate steps performed by the algorithm, some representative subsequences can be obtained. Future research could explore the limitations of string alignment methods for eye tracking data analysis in reasoning research based on these results.

The limitation of our approach is that we cannot prove that the results from the n-gram analysis are accounted for by the size of the alphabet alone. The Weather dataset comes from an experiment in which the AOIs were generated top-down and the task investigated reasoning abilities, whereas the Art dataset comes from an experiment in which the AOIs were generated bottom-up and the task did not involve any reasoning. It is possible that these factors may have influenced the relationship between n-gram length and the narrowness of the sampling distribution. Future studies may be needed to clarify this. Nevertheless, this study is the first to address the issue of determining appropriate n-gram length in scanpath analysis, comparing this to a string alignment method, and provides evidence that scanpath analysis using n-gram based methods can provide insights into the differences in gaze behaviour between two groups.

ACKNOWLEDGMENTS

This work was supported by the Engineering and Physical Sciences Research Council (EPSRC).

REFERENCES

Katherine Blondon, Rolf Wipfli, and Christian Lovis. 2015. Use of eye-tracking technology in clinical reasoning: a systematic review. *Studies in Health Technology and Informatics* 210 (2015), 90–94. https://doi.org/10.3233/978-1-61499-512-8-90

Tim Chuk, Antoni B. Chan, and Janet H. Hsiao. 2014. Understanding eye movements in face recognition using hidden Markov models. *Journal of Vision* 14, 11 (2014), 1–14. https://doi.org/10.1167/14.11.8

Moreno I Coco. 2009. The statistical challenge of scan-path analysis. In *Human System Interactions, 2009 (HSI' 09)*. IEEE, 372–375. https://doi.org/10.1109/HSI.2009.5091008

Filipe Cristino, Sebastiaan Mathôt, Jan Theeuwes, and Iain D. Gilchrist. 2010. ScanMatch: A novel method for comparing fixation sequences. *Behavior Research Methods* 42, 3 (2010), 692–700. https://doi.org/10.3758/BRM.42.3.692

Alan Davies, Gavin Brown, Markel Vigo, Simon Harper, Laura Horseman, Bruno Splendiani, Elspeth Hill, and Caroline Jay. 2016. Exploring the relationship between eye movements and electrocardiogram interpretation accuracy. *Scientific Reports* 6 (2016), 38227. https://doi.org/10.1038/srep38227

Alan Davies, Manuele Reani, Markel Vigo, Simon Harper, Martin Grimes, Clare Gannaway, and Caroline Jay. 2017. Does descriptive text change how people look at art? A novel analysis of eye-movements using data-driven Units of Interest. *Journal of Eye Movement Research* 10, 4 (2017). https://doi.org/10.16910/jemr.10.4.4

Sukru Eraslan, Yeliz Yesilada, and Simon Harper. 2014. Identifying patterns in eyetracking scanpaths in terms of visual elements of Web pages. In *Web Engineering (Lecture Notes in Computer Science)*. Springer, Cham, 163–180. https://doi.org/10.1007/978-3-319-08245-5_10

Sukru Eraslan, Yeliz Yesilada, and Simon Harper. 2015. Eye tracking scanpath analysis techniques on web pages: A survey, evaluation and comparison. *Journal of Eye Movement Research* 9, 1 (2015), 1–19. https://doi.org/10.16910/jemr.9.1.2

Sukru Eraslan, Yeliz Yesilada, and Simon Harper. 2016. Eye tracking scanpath analysis on Web pages: how many users?. In *Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications (ETRA '16)*. ACM, New York, NY, USA, 103–110. https://doi.org/10.1145/2857491.2857519

Andreas Glöckner and Ann-Katrin Herbold. 2011. An eye-tracking study on information processing in risky decisions: Evidence for compensatory strategies based on automatic processes. *Journal of Behavioral Decision Making* 24, 1 (2011), 71–98. https://doi.org/10.1002/bdm.684

Joseph H. Goldberg and Jonathan I. Helfman. 2010a. Scanpath clustering and aggregation. In *Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications (ETRA '10)*. ACM, New York, NY, USA, 227–234. https://doi.org/10.1145/1743666.1743721

Joseph H. Goldberg and Jonathan I. Helfman. 2010b. Visual scanpath representation. In *Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications (ETRA '10)*. ACM, New York, NY, USA, 203–210. https://doi.org/10.1145/1743666.1743717

Taylor R Hayes, Alexander A Petrov, and Per B Sederberg. 2011. A novel method for analyzing sequential eye movements reveals strategic influence on Raven's Advanced Progressive Matrices. *Journal of Vision* 11, 10 (2011), 1–11. https://doi.org/10.1167/11.10.10

Ernst Hellinger. 1909. Neue begründung der theorie quadratischer formen von unendlichvielen veränderlichen. *Journal für die reine und angewandte Mathematik* 136 (1909), 210–271. http://www.digizeitschriften.de/dms/img/?PID=GDZPPN002166941

Helene Hembrooke, Matt Feusner, and Geri Gay. 2006. Averaging scan patterns and what they can tell us. In *Proceedings of the 2006 Symposium on Eye Tracking*

Research & Applications (ETRA '06). ACM, New York, NY, USA, 41–41. https://doi.org/10.1145/1117309.1117325

John Heminghous and Andrew T. Duchowski. 2006. iComp: a tool for scanpath visualization and comparison. In *Proceedings of the 3rd Symposium on Applied Perception in Graphics and Visualization (APGV '06)*. ACM, New York, NY, USA. https://doi.org/10.1145/1140491.1140529

Tim C. Hesterberg. 2015. What teachers should know about the bootstrap: resampling in the undergraduate statistics curriculum. *The American Statistician* 69, 4 (2015), 371–386. https://doi.org/10.1080/00031305.2015.1089789

Nina Horstmann, Andrea Ahlgrimm, and Andreas Glöckner. 2009. *How distinct are intuition and deliberation? An eye-tracking analysis of instruction-induced decision modes*. SSRN Scholarly Paper ID 1393729. Social Science Research Network, Rochester, NY. https://papers.ssrn.com/abstract=1393729

Carmen Keller, Christina Kreuzmair, Rebecca Leins-Hess, and Michael Siegrist. 2014. Numeric and graphic risk information processing of high and low numerates in the intuitive and deliberative decision modes: An eye-tracker study. *Judgment and Decision Making; Tallahassee* 9, 5 (2014), 420–432. https://search.proquest.com/docview/1585893892/abstract/758527BBEDAD46B5PQ/1

Thomas C. Kübler, Enkelejda Kasneci, and Wolfgang Rosenstiel. 2014. SubsMatch: scanpath similarity in dynamic scenes based on subsequence frequencies. In *Proceedings of the Symposium on Eye Tracking Research and Applications (ETRA '14)*. ACM, New York, NY, USA, 319–322. https://doi.org/10.1145/2578153.2578206

Thomas C. Kübler, Colleen Rothe, Ulrich Schiefer, Wolfgang Rosenstiel, and Enkelejda Kasneci. 2017. SubsMatch 2.0: scanpath comparison and classification based on subsequence frequencies. *Behavior Research Methods* 49, 3 (2017), 1048–1064. https://doi.org/10.3758/s13428-016-0765-6

Olivier Le Meur and Thierry Baccino. 2013. Methods for comparing scanpaths and saliency maps: strengths and weaknesses. *Behavior research methods* 45, 1 (2013), 251–266. https://doi.org/10.3758/s13428-012-0226-9

Patrick Plummer, Melissa DeWolf, Miriam Bassok, Peter C. Gordon, and Keith J. Holyoak. 2017. Reasoning strategies with rational numbers revealed by eye tracking. *Attention, Perception, & Psychophysics* 79, 5 (2017), 1426–1437. https://doi.org/10.3758/s13414-017-1312-y

Manuele Reani, Niels Peek, and Caroline Jay. 2018. *Eye-tracking analysis for Bayesian reasoning*. Technical Report. Zenodo. https://zenodo.org/record/1155584 DOI: 10.5281/zenodo.1155584.

Jörg Sander, Martin Ester, Hans-Peter Kriegel, and Xiaowei Xu. 1998. Density-based clustering in spatial databases: the algorithm GDBSCAN and its applications. *Data Mining and Knowledge Discovery* 2, 2 (1998), 169–194. https://doi.org/10.1023/A:1009745219419

Michael Schulte-Mecklenbeck, Anton Kühberger, and Rob Ranyard. 2011. The role of process data in the development and testing of process models of judgment and decision making. *Judgment and Decision Making; Tallahassee* 6, 8 (2011), 733. https://search.proquest.com/docview/1011297089/abstract/81E9929F24F41D0PQ/1

Tim van Erven and Peter Harremoës. 2014. Rényi divergence and Kullback-Leibler divergence. *IEEE Transactions on Information Theory* 60, 7 (2014), 3797–3820. https://doi.org/10.1109/TIT.2014.2320500 arXiv: 1206.2459.

Rand R. Wilcox. 2010. *Fundamentals of modern statistical methods: substantially improving power and accuracy*. Springer. https://doi.org/10.1007/978-1-4419-5525-8

Yeliz Yesilada, Simon Harper, and Sukru Eraslan. 2013. Experiential transcoding: an eyetracking approach. In *Proceedings of the 10th International Cross-Disciplinary Conference on Web Accessibility (W4A '13)*. ACM, New York, NY, USA, 1–4. https://doi.org/10.1145/2461121.2461134