Assessing the Readability of Web Search Results for Searchers with Dyslexia

Adam Fourney
Microsoft Research
Redmond, WA
adamfo@microsoft.com

Meredith Ringel Morris
Microsoft Research
Redmond, WA
merrie@microsoft.com

Abdullah Ali
University of Washington
Seattle, WA
xyleques@uw.edu

Laura Vonessen
University of Washington
Seattle, WA
laurav4@cs.washington.edu

ABSTRACT

Standards organizations, (e.g., the World Wide Web Consortium), are placing increased importance on the cognitive accessibility of online systems, including web search. Previous work has shown an association between query-document relevance judgments, and query-independent assessments of document readability. In this paper we study the lexical and aesthetic features of web documents that may underlie this relationship. Leveraging a data set consisting of relevance and readability judgments for 200 web pages as assessed by 174 adults with dyslexia and 172 adults without dyslexia, we answer the following research questions: (1) Which web page features are most associated with readability? (2) To what extent are these features also associated with relevance? And, (3) are any features associated with the differences in readability/relevance judgments observed between dyslexic and non-dyslexic populations? Our findings have implications for improving the cognitive accessibility of search systems and web documents.

CCS CONCEPTS

• **Information systems → Information retrieval**; • **Human-centered computing → Accessibility**;

KEYWORDS

dyslexia, readability, web search

1 INTRODUCTION

In the last few years, both industry and academic research have begun to place increased importance on the *cognitive accessibility* of technological systems, including traditional productivity software applications (e.g., onenote.com/learningtools), as well as web technologies such as web search [3, 9, 10]. In this latter category, the W3C defines cognitive accessibility research as that which describes the "challenges of using web technologies for people with learning disabilities or cognitive disabilities" [11]. The most prevalent of these disabilities is dyslexia, a spectrum disorder that impairs reading and spelling, and which affects up to 20% of the English-speaking population [13].

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
SIGIR '18, July 8–12, 2018, Ann Arbor, MI, USA
© 2018 Association for Computing Machinery.
ACM ISBN 978-1-4503-5657-2/18/07…$15.00
https://doi.org/10.1145/3209978.3210072

In the context of web search, prior work [10] has shown that dyslexia can negatively impact all stages of informational search, including: query formation, deciding which results to click, and locating relevant information within documents. While technologies such as query suggestions, query auto-completion, and spelling correction can mitigate problems associated with query formulation [4], dyslexic searchers continue to struggle with judging the relevance of search results and eliminating non-relevant results [8]. These latter skills are among the most important, and often differentiate successful searchers from those who struggle [1].

In previous, work we conducted a study of dyslexic and non-dyslexic adults and found that dyslexic searchers award lower relevance scores on average than searchers who do not have dyslexia [10]. Moreover, we also reported an association between document readability and document relevance, with hard-to-read documents tending to command lower relevance scores. In interviews, dyslexic searchers indicated that documents were often abandoned due to poor accessibility.

In this paper, we add to this body of research by characterizing how the aesthetic and lexical characteristics of a web page can impact its subjective readability scores and relevance judgments. Our work leverages the data set collected in [10], and answers the following specific research questions:

- **R**Q1: Which web page features are most associated with readability?
- **R**Q2: To what extent, if any, are these same features also associated with relevance?
- **R**Q3: Which features help explain the differences observed between dyslexic and non-dyslexic populations?

RQ1 and RQ2 build upon prior research on the automatic analysis of website comprehensibility (e.g., [15]), as well as on efforts to incorporate content quality into a page's static rank (e.g., [2]). However, our work is differentiated by its strong focus on cognitive accessibility and its separate consideration of dyslexic and non-dyslexic populations. Our work is most similar in spirit to that of Collins-Thompson et al., who demonstrated how web search can be made more accessible to children by re-ranking documents to better match page reading levels to the searcher's reading abilities [5]. However, we consider a broader range of features when modeling readability – most notably, the presence of images and other visual elements (e.g., lists).

The remainder of the paper is structured as follows: we begin by describing the data set and features, then address each research question in turn. We conclude by discussing the implications of our findings.

2 RATINGS DATA & WEBPAGE FEATURES

2.1 Readability and Relevance Ratings

Our analysis leverages the relevance and readability judgments collected and described in [10]. This data includes judgments from 174 dyslexic and 172 non-dyslexic adults. Each participant was assigned one of 10 web search queries, and was tasked with rating their agreement to 10 Likert statements for each of the top 20 search results (presented in random order). Likert statements were scored on a 5-point scale (strongly disagree: 1, strongly agree: 5). In this paper we focus on only two statements:

- **Readability:** Overall the website was easy to read.
- **Relevance:** The web page was relevant to the web search task.

In total, the data set includes 279 complete responses where participants rated all 20 pages, and 67 partial responses where participants rated fewer than 20 pages. On average, each ⟨page, Likert statement⟩ pair received a set of 17.3 responses. In this paper, we consider the mean responses assigned by dyslexic participants, as well as the mean responses from non-dyslexic participants. This yields a total of 800 data points (10 queries × 20 pages × 2 Likert items × 2 user populations).

2.2 Feature Extraction

We are interested in associating page-level aesthetic and lexical features with the readability and relevance scores of the above-mentioned 200 web pages. To support this analysis, we used an automated web browser [1] to retrieve and render each web page, and to execute a custom script in the page's JavaScript context. This script waited for the page to fully load and stabilize (i.e., undergo no further changes to the document object model for 10 seconds), after which it captured the features listed in Table 1. Our feature set was heavily inspired by the page elements and properties implicated in interviews with dyslexic searchers as helping or hindering readability [10]. It was also inspired by the work of Ivory et al. [7].

Of the 14 features listed in Table 1, `gunning_fog` is the most direct measure of a web page's reading difficulty. The Gunning-Fog Index is a popular readability assessment tool that estimates the number of years of education (U.S. grade levels) necessary to understand a given document [6]. It is a linear formula that considers a document's average sentence length, as well as the percentage of polysyllabic words present in the document (words with three or more syllables). To support the computation of the Gunning-Fog Index, it is necessary to divide the document into sentences. We do this by first breaking the page's visible text into distinct passages, with each passage demarcated by the transition from one HTML block element[2] to the next, when doing an in-order traversal of the document object model (DOM) tree. This isolates navigation elements, titles, and captions, rather than having them blend together into longer phrases. We then subdivide these passages into sentences using an off-the-shelf sentence chunker[3].

In addition to computing the Gunning-Fog Index, we use this segmentation to compute the `sentence_text_ratio` feature. This ratio compares the number of words contained in well-formed sentences (phrases that have sentence casing, contain at least one verb, and are terminated with the correct punctuation), to the number of words found outside of sentences. Words in the latter scenario are often found in titles, labels, lists, captions and other display text.

Feature	Description
mean_text_contrast	Average foreground-background contrast ratio of text-bearing elements [14].
mean_font_size	Average font size of text-bearing elements.
mean_line_length	Average number of characters that fit on each line of text.
mean_image_size	Average image size in pixels (measured along the diagonal).
n_fonts	Number of fonts used. (Unique combinations of family, variant, weight, etc.)
n_words	Number of words in the document.
n_images	Number of images in the document.
n_headings	Number of headings in the document.
n_links	Number of hyperlinks in the document.
n_lists	Number of bullet or numbered lists.
n_tables	Number of tables in the document.
sentence_text_ratio	Ratio of text appearing in sentences vs. not in sentences (e.g., titles, labels, etc.)
image_area_%	Percentage of the page covered by images.
gunning_fog	Gunning-Fog readability score.

Table 1: The 14 web page features considered in our study.

3 RESULTS

3.1 Features Associated with Readability

To answer our first research question, we conducted a multiple regression analysis to ascertain the degree to which the page features correlate with the average readability scores ascribed by dyslexic and non-dyslexic participants. We begin by centering the data, and scaling to unit variance. Given the large number of features, we then check for multicollinearity, but find that the average variance inflation factor is quite low (VIF: 1.46). As such, we move forward with the analysis, and present the results in the left three columns of Tables 2 and 3 for dyslexic and non-dyslexic participants, respectively.

For the participants with dyslexia (Table 2), this regression analysis was significant, $F_{14,185} = 2.46$, $p = 0.003$. $R^2 = 0.157$, indicating that 15.7% of the variance in the readability scores can be explained by these 14 features. Further inspection of the model predictors shows a significant positive coefficient for the average image size feature ($p < 0.01$), and significant negative coefficients for the Gunning-Fog reading level ($p = 0.02$), as well as for the feature that captures the ratio of text appearing in sentences vs. not appearing in sentences ($p = 0.04$). As noted above, non-sentence text often comprises titles, labels, list elements, and other display text that may help to organize content. Having more of these elements will tend to lower the `sentence_text_ratio`.

Additionally, we find a marginally significant ($p = 0.06$) positive coefficient for the feature capturing the average number of characters that fit on each line of text. On the surface, this finding appears

[1] http://phantomjs.org/
[2] Block elements: ⟨p⟩, ⟨div⟩, ⟨li⟩, etc.
[3] http://www.nltk.org/api/nltk.tokenize.html

	Readability		Relevance	
	$R^2 = 0.157$		$R^2 = 0.186$	
	$p = 0.003$		$p < 0.001$	
	$F_{14,185} = 2.46$		$F_{14,185} = 3.02$	
Feature	β coeff	p	β coeff	p
intercept	**3.089**	**< 0.01**	**3.181**	**< 0.01**
mean_text_contrast	-0.018	0.51	-0.014	0.61
mean_font_size	0.027	0.32	0.019	0.50
mean_line_length	0.052	0.06	**0.087**	**< 0.01**
mean_image_size	**0.095**	**< 0.01**	**0.096**	**< 0.01**
n_fonts	0.018	0.47	0.021	0.42
n_words	0.021	0.58	0.055	0.16
n_images	0.005	0.86	0.054	0.06
n_headings	-0.021	0.48	0.000	0.99
n_links	0.004	0.93	-0.073	0.08
n_lists	-0.023	0.42	0.016	0.59
n_tables	0.009	0.74	-0.009	0.75
sentence_text_ratio	**-0.057**	**0.04**	-0.016	0.56
image_area_%	-0.047	0.12	-0.018	0.56
gunning_fog	**-0.062**	**0.02**	-0.037	0.18

Table 2: Results of the multiple regression analysis of readability scores (left) and relevance judgments (right), as judged by dyslexic participants. Statistically significant coefficients are labeled in bold text.

	Readability		Relevance	
	$R^2 = 0.167$		$R^2 = 0.127$	
	$p = 0.002$		$p < 0.026$	
	$F_{14,185} = 2.65$		$F_{14,185} = 1.93$	
Feature	β coeff	p	β coeff	p
intercept	**3.396**	**< 0.01**	**3.593**	**< 0.01**
mean_text_contrast	-0.018	0.62	-0.078	0.13
mean_font_size	**0.070**	**0.05**	0.086	0.11
mean_line_length	0.045	0.24	**0.127**	**0.02**
mean_image_size	**0.135**	**< 0.01**	0.080	0.18
n_fonts	-0.021	0.54	0.013	0.81
n_words	-0.008	0.87	0.103	0.18
n_images	-0.027	0.46	0.081	0.14
n_headings	-0.036	0.36	-0.006	0.92
n_links	-0.033	0.55	-0.134	0.10
n_lists	**0.101**	**0.01**	0.002	0.98
n_tables	0.056	0.12	0.047	0.40
sentence_text_ratio	**-0.073**	**0.05**	**-0.110**	**0.04**
image_area_%	-0.047	0.23	-0.027	0.64
gunning_fog	-0.008	0.82	-0.070	0.18

Table 3: Results of the multiple regression analysis of readability scores (left) and relevance judgments (right), as judged by non-dyslexic participants. Statistically significant coefficients are labeled in bold text.

to contradict the commonly held view that long lines of text are more difficult to read [12]. Inspection of the data set shows that sidebars and navigation elements represent regions with very short line-lengths, and it is likely these regions that the model penalizing.

For the participants who do not have dyslexia (Table 3), the multiple regression analysis was again significant, $F_{14,185} = 2.65$, $p = 0.002$. $R^2 = 0.167$. As before, we find a significant positive coefficient for the mean image size feature ($p < 0.01$), and a significant negative coefficient for the sentence text ratio feature ($p = 0.05$). We also find significant positive coefficients for the mean font size ($p = 0.05$), and for the number of lists contained in the document ($p = 0.01$). This agrees with the intuition that larger fonts facilitate reading, and that lists can help organize information.

3.2 Features Associated with Relevance

As noted earlier, we reported finding that readability scores were significantly correlated with relevance scores for both dyslexic users ($r = 0.439, p \ll 0.001$) and non-dyslexic users ($r = 0.454, p \ll 0.001$). We investigate if the above-mentioned features may underlie this relationship. To answer this question, we repeat the multiple regression analysis, this time for the mean relevance scores. The results are presented in the right two columns of Tables 2 and 3 for dyslexic and non-dyslexic participants, respectively.

For dyslexic participants (Table 2), this regression analysis was highly significant, $F_{14,185} = 3.02$, $p \ll 0.001$. $R^2 = 0.186$, indicating that nearly 18.6% of the variance in relevance scores can be explained by these 14 features. As a point of comparison, recall that the readability-relevance correlation that motivated this investigation has $r = 0.439 \Rightarrow R^2 = 0.192$. Inspecting the multiple regression predictors reveals that the coefficients for `mean_line_length`, and `mean_image_size` are both positive and highly significant (in both cases $p \ll 0.001$). These same predictors play a similar role, and have similar signs and magnitudes, in the multiple regression model of readability (left 3 columns). This may partially explain the observed correlation between readability and relevance. Moreover, we find that the coefficient for the `n_images` feature is also positive and marginally significant ($p = 0.06$). Taken together with the importance of `mean_image_size`, these findings suggest that dyslexic searchers may especially value visual content when assessing a document's relevance.

For non-dyslexic participants (Table 3), this regression analysis was again significant, $F_{14,185} = 1.92$, $p \ll 0.001$. $R^2 = 0.127$; however, the resultant model explains considerably less of the variance. Inspection of the model coefficients reveals that `mean_line_length`, and `sentence_text_ratio` are important predictors ($p = 0.02$ and $p = 0.04$, respectively) – the latter of which was also an important predictor for readability among both groups. Again, the common role of these features may partially explain the observed correlation between readability and relevance.

3.3 Features Associated with Group Differences

Finally, we are interested in discovering if the above-mentioned features may help explain group differences in the readability and/or relevance scores. We begin by confirming the presence of group differences: the readability scores of dyslexic and non-dyslexic participants are significantly correlated ($r = 0.470, p \ll 0.001$), but dyslexic participants assign significantly lower readability scores on average ($t_{199} = 9.14, p \ll 0.001$). The same holds true for relevance scores ($r = 0.674, p \ll 0.001$, and $t_{199} = 11.00, p \ll 0.001$).

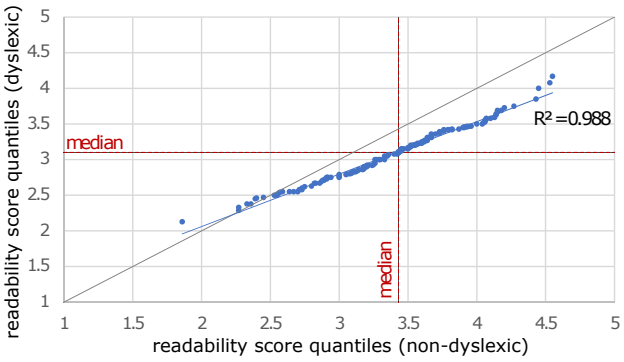

Figure 1: Q-Q plot comparing the readability score quantiles of both groups. (Dyslexic: vertical, non-dyslexic: horizontal)

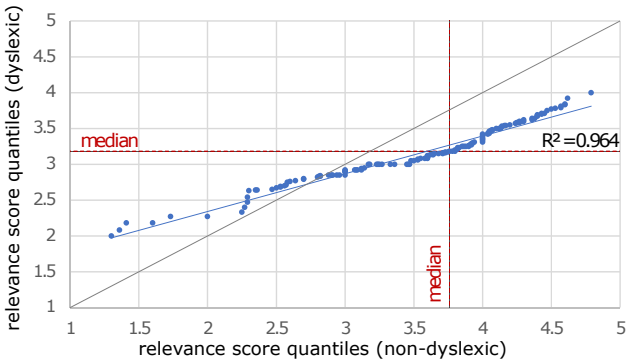

Figure 2: Q-Q plot comparing the relevance score quantiles of both groups. (Dyslexic: vertical, non-dyslexic: horizontal)

To investigate if these differences can be explained by page features, we again apply multiple regression, this time modeling the difference in scores between dyslexic and non-dyslexic participants. We do this for readability scores, and separately for relevance scores. In both cases, the analysis fails to find a significant linear relationship between page features and observed differences: For readability, the model achieves $R^2 = 0.086, p = 0.25$. For relevance, it achieves $R^2 = 0.071, p = 0.45$. We conclude that the features are insufficient for explaining group differences. However, we did observe other group differences, which we explore below.

3.4 Group Differences in Rating Scale Use

To further understand how the scores assigned by each group may differ, we generated the Q-Q plot of readability score quantiles (Figure 1). We find a strong linear relationship $R^2 = 0.988$ that is oblique to the line $y = x$, suggesting that the underlying distributions are distinct but linearly related. Furthermore, we observe that dyslexic and non-dyslexic participants make comparable use of the lower regions of the ratings scale, but this quickly diverges as the quantile increases. As a result, scores assigned by dyslexic participants are simultaneously lower on average, and are confined to a smaller region of the ratings scale. This significantly compresses the variance in the distribution of readability scores, as observed by Levene's test of equal variances ($s^2_{dys} = 0.129$, $s^2_{non} = 0.236$, $W = 9.14$, $p \ll 0.001$). This range compression is even more pronounced in the Q-Q plot of relevance scores (Figure 2). Again, Levene's test finds that the variances are unequal ($s^2_{dys} = 0.143$, $s^2_{non} = 0.498$, $W = 41.7, p \ll 0.001$). This central tendency bias may indicate that dyslexic participants are failing to sufficiently differentiate between relevant and non-relevant documents, as was observed in [8].

4 CONCLUSION

In conclusion we find that a number of web page features associate both with readability and with relevance. Features that were repeatedly implicated include: `mean_line_length`, `mean_image_size`, and `sentence_text_ratio`. These features are important for both groups, suggesting that factoring readability into search engine ranking and/or making improvements to readability in the design of web pages would likely improve cognitive accessibility to the user population overall, not merely to people with dyslexia. We envision using these features to re-rank documents in a manner similar to [5]. Alternatively, these features might be distilled into cues that could be presented on results pages to indicate document accessibility. This may be an especially reasonable strategy for features associated with image content. Finally, we report that dyslexic searchers may have a strong central tendency bias when providing explicit relevance judgments. Future studies may want to consider this effect, and perhaps work to actively mitigate its impact.

REFERENCES

[1] Mikhail Ageev, Qi Guo, Dmitry Lagun, and Eugene Agichtein. 2011. Find It if You Can: A Game for Modeling Different Types of Web Search Success Using Interaction Data. In *Proceedings of SIGIR '11*. ACM, 345–354.
[2] Michael Bendersky, W. Bruce Croft, and Yanlei Diao. 2011. Quality-biased Ranking of Web Documents. In *Proceedings of WSDM '11*. ACM, 95–104.
[3] Gerd Berget and Frode Eika Sandnes. 2015. Searching Databases without Query-Building Aids: Implications for Dyslexic Users. *Information Research* 4, 20 (12 2015).
[4] Gerd Berget and Frode Eika Sandnes. 2016. Do Autocomplete Functions Reduce the Impact of Dyslexia on Information-searching Behavior? The Case of Google. *J. Assoc. Inf. Sci. Technol.* 67, 10 (Oct. 2016), 2320–2328.
[5] Kevyn Collins-Thompson, Paul N. Bennett, Ryen W. White, Sebastian de la Chica, and David Sontag. 2011. Personalizing Web Search Results by Reading Level. In *Proceedings of CIKM '11*. ACM, 403–412.
[6] Impact Information. 2004. Robert Gunning's Fog Readability Formula: Plain Language at Work Newsletter #8. (2004). http://www.impact-information.com/impactinfo/newsletter/plwork08.htm
[7] Melody Y. Ivory, Rashmi R. Sinha, and Marti A. Hearst. 2000. Preliminary Findings on Quantitative Measures for Distinguishing Highly Rated Information-Centric Web Pages. In *Proceedings of 6th Conference on Human Factors and the Web*. 15.
[8] A. MacFarlane, A. Albrair, C. R. Marshall, and G. Buchanan. 2012. Phonological Working Memory Impacts on Information Searching: An Investigation of Dyslexia. In *Proceedings of IIIX '12*. ACM, 27–34.
[9] Andrew MacFarlane, George Buchanan, Areej Al-Wabil, Gennady Andrienko, and Natalia Andrienko. 2017. Visual Analysis of Dyslexia on Search. In *Proceedings of CHIIR '17*. ACM, 285–288.
[10] Meredith Ringel Morris, Adam Fourney, Abdullah Ali, and Laura Vonessen. 2018. Understanding the Needs of Searchers with Dyslexia. In *Proceedings of CHI '18*. ACM.
[11] Lisa Seeman and Michael Cooper. 2015. Cognitive Accessibility User Research: W3C First Public Working Draft 15 January 2015. (2015). https://www.w3.org/TR/coga-user-research/
[12] Dmitry A. Tarasov, Alexander P. Sergeev, and Victor V. Filimonov. 2015. Legibility of Textbooks: A Literature Review. *Procedia - Social and Behavioral Sciences* 174 (2015), 1300 – 1308. International Conference on New Horizons in Education, INTE 2014, 25-27 June 2014, Paris, France.
[13] The International Dyslexia Association. 2017. Dyslexia Basics. (2017). https://dyslexiaida.org/dyslexia-basics/
[14] World Wide Web Consortium. 2015. Contrast (Minimum): Understanding SC 1.4.3. (2015). https://www.w3.org/TR/UNDERSTANDING-WCAG20/visual-audio-contrast-contrast.html
[15] Ping Yan, Zhu Zhang, and Ray Garcia. 2007. Automatic Website Comprehensibility Evaluation. In *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence (WI '07)*. IEEE Computer Society, 191–197.