Lessons Learned from Users Reading Highlighted Abstracts in a Digital Library

Dagmar Kern, Daniel Hienert
GESIS - Leibniz Institute for the Social Science
Cologne, Germany
dagmar.kern@gesis.org,daniel.hienert@gesis.org

Katrin Angerbauer
VISUS, University of Stuttgart
Stuttgart, Germany
katrin.angerbauer@visus.uni-stuttgart.de

Tilman Dingler
University of Melbourne
Melbourne, Australia
tilman.dingler@unimelb.edu.au

Pia Borlund
Oslo Metropolitan University
Oslo, Norway
pia.borlund@oslomet.no

ABSTRACT

Finding relevant documents is essential for researchers of all disciplines. We investigated an approach for supporting searchers in their relevance decision in a digital library by automatically highlighting the most important keywords in abstracts. We conducted an eye-tracking study with 25 subjects and observed very different search and reading behavior which lead to diverse results. Some of the participants liked that highlighted abstracts accelerate their relevance decision, while others found that they disturb the reading flow. What many agree on is that the quality of highlighting is crucial for trust and system credibility.

CCS CONCEPTS

• **Human-centered computing → User studies**; • **Applied computing →** *Digital libraries and archives*;

KEYWORDS

highlighting, reading behavior, user study, relevance judgment

1 INTRODUCTION

Digital environments have generally changed reading behavior. People very often only browse or scan documents and web pages, look for keywords, read more selectively and spend less time on in-depth reading [12]. This also applies for the document triage process in digital libraries [4]. During this process, a user first decides if a given document is relevant to her information need. Most often this relevance judgment is based on the title and the abstract of a document [4, 11, 15]. Considering the importance of the abstract on the relevance judgment and the changed reading behavior, we present an approach to support keyword spotting by highlighting the most important keywords in abstracts.

Highlighting search terms in search results is already a common practice in web search and digital libraries to directly show users in which part of the document the search term has been found. In the research literature, one can find more specific approaches like dynamic highlighting of sentences based on their accordance with their salience [18], highlighting of whole sentences containing conceptual keywords [5], highlighting of automatically generated search query terms that were used to find a recommended document [2], or different highlighting of keywords in abstracts that relate to predefined concepts [17]. In our previous work, we explored the use of highlighting and other keyword summary visualizations to prime readers a priori to support reading comprehension and improve subjective impressions for which we found highlighting to be most effective [1, 7]

Eye-tracking data of a user can provide a lot of information about the search and reading behavior as well as the relevance judgment [8, 9, 13]. Therefore, we conducted a user study utilizing an eye tracker to find answers to the following research questions: (R1) How do key-term-highlighted abstracts support users in their document relevance decision? (R2) What effects have key-term-highlighted abstracts on the reading behavior?

2 KEY-TERM-HIGHLIGHTED ABSTRACTS

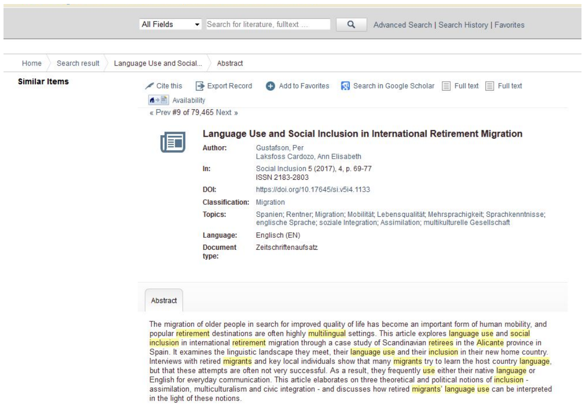
We tested the effect of key-term highlighted abstracts with the digital library system Sowiport. Sowiport [10] was a digital library for social science information with 9.7 million bibliographic records, full texts, and research projects[1]. Sowiport showed results to a search query in a ranked result list and provided a detailed view page for each result on demand with more metadata about the selected article. This included in most cases a German or English abstract. Figure 1 shows a detailed view page with a key-term-highlighted abstract.

2.1 Approach

To identify the relevant key-terms we used the term-frequency inverse document frequency (tf-idf) measure. The corpora used to train the tf-idf models consisted of approximately 10,000 English and 21,000 German full texts on topics of social science. These texts were obtained via the Social Science Open Access Repository (SSOAR)[2] to train our German and English tf-idf models. Stop words were filtered in both languages. Furthermore, we considered only nouns and adjectives for scoring and thus highlighting, as we considered them to be the most informative content words. The preprocessing of the abstract (tokenization and part of speech-tagging) and the scoring was done during run-time. To calculate the scores of the abstracts' terms the preprocessed sentences were passed to the pre-trained models. The number of terms highlighted in the abstract was determined in relation to the abstract's length.

[1] Sowiport was discontinued at the end of 2017

[2] https://www.gesis.org/ssoar/home/, SSOAR provides an Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) interface for metadata access



Language Use and Social Inclusion in International Retirement Migration

Author:	Gustafson, Per Laksfoss Cardozo, Ann Elisabeth
In:	Social Inclusion 5 (2017), 4, p. 69-77 ISSN 2183-2803
DOI:	https://doi.org/10.17645/si.v5i4.1133
Classification:	Migration
Topics:	Spanien; Rentner; Migration; Mobilität; Lebensqualität; Mehrsprachigkeit; Sprachkenntnisse; englische Sprache; soziale Integration; Assimilation; multikulturelle Gesellschaft
Language:	Englisch (EN)
Document type:	Zeitschriftenaufsatz

The migration of older people in search for improved quality of life has become an important form of human mobility, and popular retirement destinations are often highly multilingual settings. This article explores language use and social inclusion in international retirement migration through a case study of Scandinavian retirees in the Alicante province in Spain. It examines the linguistic landscape they meet, their language use and their inclusion in their new home country. Interviews with retired migrants and key local individuals show that many migrants try to learn the host country language, but that these attempts are often not very successful. As a result, they frequently use either their native language or English for everyday communication. This article elaborates on three theoretical and political notions of inclusion - assimilation, multiculturalism and civic integration - and discusses how retired migrants' language use can be interpreted in the light of these notions.

Figure 1: Detailed view page of Sowiport.

We set the number of terms to be highlighted to ten percent. If a term occurred multiple times, it was highlighted in every occurrence.

2.2 Implementation

For the highlighting, the documents of the corpora and the abstracts were preprocessed with TreeTagger [16] to obtain the lemmas and the corresponding inflected forms of the words as well as their parts of speech. To model the corpora and to calculate the tf-idf models and to do the scoring of the abstracts' tf-idf values we used the gensim library [14] for Python. For the presentation of the highlighted key-terms we used basic HTML and CSS and modified their background color. Through a config file, we could easily turn on or off the highlighting functionality for the abstracts.

3 EVALUATION

We conducted an eye-tracking study with 25 participants to examine what effect highlighting of the most important keywords in abstracts has on the relevance decision and reading behavior in the digital library Sowiport.

3.1 Experimental Setup

The study took place in our usability lab in single sessions under controlled conditions without distraction. Participants were seated alone in the usability lab in front of a 22-inch monitor. The SMI iView RED 250 mobile eye tracker was attached to the lower frame of it. The monitor and the eye tracker were connected to a laptop that ran the eye tracking software. The eye gazes were recorded with a sampling rate of 60Hz. The participant used a keyboard and a mouse to interact with the system. The experimenter sat next to the usability lab in an observation room in front of a monitor showing the screen the participant saw. The experimenter captured everything on the participant's screen and the gaze behavior with the capturing software Camtasia[3] to be able to conduct a stimulated recall interview after the task performance.

3.2 Participants

As the digital library comprised documents in the field of social science, we needed participants with a background in social science or a similar field of study. We recruited participants through mailing lists and posters at the local university. Thirty participants took part in our study. Due to technical problems with the eye tracker we had to exclude the results of 5 participants from the data analysis. Sixteen of the remaining participants were female and nine male. Their age ranged from 23 to 45 years (mean: 28.6, SD: 4.12). Fourteen hold a Bachelor degree, ten a Master degree and one was a postdoctoral researcher. Twenty-four of them had a background in Sociology while one was a psychologist.

3.3 Simulated Work Tasks

As we were interested in how the highlighting influenced users' usual information seeking behavior in a digital library, we created two simulated work task scenarios (according to [3]), which the participants had to perform.

Task A: "Please use Sowiport to find publications on a topic you are well familiar with, for example, your main research topic. Please bookmark (using browser bookmarks) documents that seem to be useful to you."

Task B: "Please use Sowiport to find publications on a topic you are not familiar with, but you are interested in learning more about it. Please bookmark (using browser bookmarks) all documents that seem to be useful to you." The simulated work tasks provide context for realistic searching and form the basis for reliable reading and relevance assessment behavior.

3.4 Methodology

Data collection took place from August 2017 to January 2018 at our research institute. We invited participants to single sessions. We followed a within-subject design approach with the highlighting and simulated work tasks (A and B) being the independent variables in two conditions: (1) abstracts with key-term-highlights (named highlighting) and (2) abstracts without key-term-highlights (named non-highlighting). The condition sequence was counterbalanced among participants.

3.5 Procedure

All participants followed the same procedure: after greeting the participant, the experimenter explained the purpose of the study and asked to sign a consent form as well as to fill out a demographic survey. The experimenter provided a short introduction to Sowiport and gave the participant time to familiarize with the system. Afterwards, the eye tracker was calibrated and the first task was given. The participant was asked to read the instruction carefully. Then, she started searching for documents. There was no announced time-constraint so that the participant searched until she considered the information need was satisfied. After the search in the highlighting condition, the participants filled out a questionnaire to provide feedback on the highlighting. Afterwards, the stimulated recall interview took place. For each bookmark or closed detailed view page, the experimenter asked how useful this document was to solve the task and why. The second run with the other condition and the other task was conducted following the

[3] https://www.techsmith.com/video-editor.html

	highlighting		Task A		Task B		more-readers		less-readers	
	non-highlighting	highlighting	non-highlighting	highlighting	non-highlighting	highlighting	non-highlighting	highlighting	non-highlighting	highlighting
Number of examined documents	284	265	178	112	106	153	118	132	166	133
(1) Average fixation time a on word in abstract	13.38 (14.01)	13.02 (12.4)	12.44 (14.45)	13.19 (12.10)	14.96 (13.17)	12.89 (12.66)	**21.63 (16.36)** *	15.81 (12.25) *	7.52 (7.99)	10.25 (11.97)
(2) Relevance judgment	3.86 (1.29)	3.63 (1.38)	4.01 (1.21)	3.68 (1.44)	3.59 (1.39)	3.6 (1.33)	3.53 (1.48)	3.59 (1.47)	**4.09 (1.9)** *	3.68 (1.28) *
(3) Ratio bookmarked and examined documents	**0.82 (0.44)** *	0.73 (0.44) *	**0.88 (0.42)** *	0.71 (0.46) *	0.72 (0.46)	0.75 (0.43)	0.74 (0.55)	0.68 (0.47)	**0.87 (0.33)** *	0.78 (0.41) *
(4) Jumps in abstract	0.50 (0.50) *	**0.68 (0.47)** *	0.55 (0.5) *	**0.68 (0.47)** *	0.43 (0.5) *	**0.68 (0.46)** *	0.46 (0.50) *	**0.62 (0.49)** *	0.53 (0.50) *	**0.74 (0.44)** *
(5) Sequentially read	0.79 (0.41)	0.73 (0.44)	0.76 (0.43)	0.74 (0.44)	**0.84 (0.37)** *	0.73 (0.44) *	0.87 (0.34)	0.85 (0.36)	**0.73 (0.44)** *	0.62 (0.48) *
(6) Percentage read	0.44 (0.4)	0.48 (0.42)	0.39 (0.4)	0.52 (0.44)	0.53 (0.4)	0.44 (0.40)	0.67 (0.39)	0.63 (0.4)	0.27 (0.3)	0.33 (0.38)

Table 1: Results for each dependent variable separated by the different subgroups, *p<0.05.

same procedure as in the first run. In the end, the participant filled out the post-questionnaire asking which of the two conditions she would prefer and which advantages and disadvantages she saw in the highlighting. We thanked the participants by compensating their effort with 30€. Altogether, one session took about 2 hours.

4 RESULTS

Altogether, the 25 participants judged the relevance of 912 documents. To be able to answer our research questions we focused on the detailed view pages that had an abstract because only those were different in the highlighting and non-highlighting conditions. From these documents, we further looked at just the documents which had at least one fixation in the abstract because in the other cases the other metadata was responsible for the relevance decision like the title or author and not the abstract. That left us with 549 abstracts for analysis. A total of 265 abstracts were with highlighting and 284 were without highlighting. On average, each participant looked at 10.6 (SD=3.34) detailed view pages in the highlighting condition and 11.36 (SD=5.23) detailed view pages in the non-highlighting condition. To determine the statistical significance of our results, we conducted Wilcoxon Signed Rank tests with $\alpha = 0.05$. Table 1 shows an overview of the results, which we will report on in the following.

4.1 Dependent variables

Our analysis is based on the following dependent variables: (1) Average fixation time on a word in abstract: value calculated by dividing the fixation time per abstract by the number of words in the abstract. In doing so, we address the different abstract lengths. (2) Relevance judgment of document: For each bookmarked or rejected document participants judged how useful the document for their task was on a Likert-style scale from 1 (not at all useful) to 5 (very useful). (3) Ratio between documents bookmarked and documents examined. (4) Jumps in abstract: For each document, we identified if a user jumped between different parts of the abstract (boolean value). (5) Sequentially read: whether the abstract or part of the abstract had been sequentially read (boolean value). (6) Percentage read of the abstract: How much of the abstract relative to its length had been read. This value is calculated by the number of rows a participant had read and the abstract's total number of rows.

4.2 Data recording and preparation

Sowiport has a logging component which logs every user action within the whole session. These are user actions like typed-in search query, the resulting list of search results, which detailed view has been shown, which link has been clicked, etc. Additionally to the logging data, we recorded the eye gaze data and used the SMI analyzing software BeGaze to calculate the fixation times on the abstract. To be able to do that we had to draw for every abstract the Area of Interest on the detailed view page. Through questionnaires, we collected the subjective assessments. The audio recording of the stimulated recall interview had been transcribed.

4.3 How do key-term-highlighted abstracts support users' document relevance decision?

4.3.1 Average fixation time on word in abstract. When we started with our research, we assumed that the key-term-highlights allow users to skim the text faster and to extract the gist of the text just by reading the highlighted terms. We thought that based on that, they could decide faster if a document is relevant or not. However, the results do not confirm our assumption. We did not find any significant differences regarding the average fixation time comparing the documents in the highlighting and non-highlighting condition. We wanted to know if there is a factor that influences this result in any way. Therefore we looked at different subgroups. First, we checked if the task might play a role, but again there are no significant differences in the highlighting condition, but we found significant differences in the non-highlighting condition. Participants were quicker in their relevance decision when looking for documents to a familiar topic (task A) (mean: 12.44, SD: 14.45) than while looking for documents to an unfamiliar topic (mean: 14.96, SD: 13.17) (p=0.022). Secondly, we divided the group of participants into those who read much of the abstract (more-readers (n=13) read on average >=33% of the abstract) and those who read little of the abstract (less-readers (n=12) read on average <33% of the abstract). More readers spent significantly less time reading the abstract in the highlighting condition (mean: 15.81, SD:12.25) than in the non-highlighting condition (mean: 21.63, SD: 16.36) (p=0.003). The results of the less-readers were not statistically significant.

4.3.2 Relevance judgment of document. There were no significant differences comparing highlighting vs. non-highlighting. When looking at the two tasks, it is remarkable that participants judged the relevance of bookmarked documents in the non-high-lighting with task A (familiar topic) higher (mean 4.01, SD:1.21) than with task B (mean 3.59, SD: 1.39) (p=0.008). We found that less-readers judged the relevance of bookmarked documents higher in the non-high-lighting condition (mean: 4.09, SD: 1.9) than in the highlighting condition (mean: 3.68, SD: 1.28) (p= 0.004).

4.3.3 Bookmarked documents. Here, we found significant differences in the highlighting and non-highlighting condition. Participants bookmarked more documents in the non-highlighting condition (mean: 0.82, SD: 0.44) than in the highlighting condition (mean: 0.73, SD: 0.44) (p=0.033). This is also true for task A (familiar topic) and the less-readers.

4.4 What effects have key-term-highlighted abstracts on the reading behavior in digital libraries?

4.4.1 "Jumps" in abstract. The highlights affected the reading behavior in so far that participants jumped more in the abstracts in the highlight condition (mean: 0.68, SD: 0.47) than in the non-highlighting condition (mean: 0.5, SD: 0.50) (p<0.0001). This is also evident in all subgroups. This result is not surprising as the idea of highlights is to guide the user's attention to the highlighted terms. In analyzing the gaze behavior of all participants, however, we observed that there were three different types of interactions: (1) participants first skimmed the highlighted terms and then decided to read more or not, (2) participants started reading the abstract and after a while stopped sequential reading and checked the remaining highlighted key-terms, and (3) the highlighted key-terms were fully ignored by participants.

4.4.2 Sequentially read. Most often jumps were combined with sequential reading. Participants read significantly more sequentially in the non-highlighting condition than in the highlighting condition while performing task B (unfamiliar topic) (non-highlighting mean: 0.84, SD:0.37, highlighting mean: 0.73, SD: 0.37, p=0.04) and when they were part of the subgroup less-readers (non-highlighting mean: 0.74, SD:0.44, highlighting mean: 0.62, SD: 0.49, p=0.041).

4.4.3 Percentage read. We could not find any significant differences in the amount of text participants read in the highlighting and non-highlighting condition for all subgroups. There were, however, significant differences in the non-highlighting condition considering each subgroup: participants read more of the abstract while searching for a document on an unfamiliar (task B) (mean: 0.53, SD:0.4) than on a familiar topic (task A) (mean 0.39, SD: 0.39) (p=0.003). More-readers read more from the abstract (mean: 0.67, SD: 0.39) than less-readers (mean: 0.27, SD: 0.33) (p<0.0001) and in this case, this was also true for the highlighting condition.

4.5 Subjective Feedback

In the post-questionnaire, we asked participants which version of the system they preferred. The answers were almost equally distributed (14 in favor of highlighting condition vs. 11 preferring non-highlighting). Answers given to 7-point-Likert-style scales revealed that the highlighting distracted participants rather less while reading the abstract (mean: 2.88, SD: 1.87, 1=not at all distracting to 7=very distracting) but the usefulness of the highlighting was rated mediocre (mean: 4.12, SD: 0.94, 1=not at all useful to 7=very useful to 1=not at all useful). Participants mentioned three main benefits of the highlighting: "faster relevance decision" (n=5), "quick overview" (n=5) and "relevant keywords at a glance" (n=4). The three most often stated disadvantages of the systems were: "Disturb the reading flow" (n=8), "wrong/unimportant keywords were highlighted" (n=6), "risk of overlooking important terms when they are not highlighted" (n=4).

4.6 Discussion and Conclusion

To summarize the insights from our experiment: (1) Highlighting not being beneficial for all: for some of the participants, highlights had a positive effect on relevance judgment and reading behavior but not for all. If a digital library plans to offer such a function, we recommend an opt-out option. (2) Task A vs. Task B: With the simulated work tasks A and B we intended to evoke a similar search behavior for both conditions. However, the comparison of the dependent variables of task A and task B in the non-highlighting conditions revealed that there were already some differences in search behavior (see table 1). (3) Versatile search and reading behavior: we observed very different search and reading behavior in both conditions. Some participants used a lot of search queries, visited a high number of detailed view pages, took their time for solving the tasks and read the abstract carefully. Others finished their tasks quite soon with only using a few search terms. Most often they just skimmed the abstract or ignored it at altogether. This is also reflected in the rather large standard deviation of our results (see 1). (4) Quality of highlighting: The quality of the highlighted keywords is crucial and needs to be very high so that users can trust that really the most important keywords are highlighted.

In future work, we plan to analyze the seeking and reading behavior further (similar as proposed in [6]) to find similar searching or reading patterns that give us more indications for which user group or which task a simple key-term-highlighting based on tf-idf might be helpful for supporting document relevance judgment. We further intend to address the quality issue by investigating different approaches to highlighting most important keywords.

5 ACKNOWLEDGEMENTS

We thank the participants of our study.

REFERENCES

[1] Katrin Angerbauer, Tilman Dingler, Dagmar Kern, and Albrecht Schmidt. 2015. Utilizing the Effects of Priming to Facilitate Text Comprehension. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA '15)*. ACM, New York, NY, USA, 1043–1048. https://doi.org/10.1145/2702613.2732914

[2] Daniel Billsus, David M. Hilbert, and Dan Maynes-Aminzade. 2005. Improving Proactive Information Systems. In *Proceedings of the 10th International Conference on Intelligent User Interfaces (IUI '05)*. ACM, New York, NY, USA, 159–166. https://doi.org/10.1145/1040830.1040869

[3] Pia Borlund. 2003. The IIR evaluation model: a framework for evaluation of interactive information retrieval systems. *Information Research. An International Electronic Journal* 8, 3 (2003).

[4] George Buchanan and Fernando Loizides. 2007. Investigating Document Triage on Paper and Electronic Media. In *Research and Advanced Technology for Digital Libraries*, László Kovács, Norbert Fuhr, and Carlo Meghini (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 416–427.

[5] Ed H. Chi, Michelle Gumbrecht, and editor="Jacko Julie A. Hong, Lichan". 2007. Visual Foraging of Highlighted Text: An Eye-Tracking Study. In *Human-Computer Interaction. HCI Intelligent Multimodal Interaction Environments*. Springer Berlin Heidelberg, Berlin, Heidelberg, 589–598.

[6] Michael J. Cole, Jacek Gwizdka, Chang Liu, Ralf Bierig, Nicholas J. Belkin, and Xiangmin Zhang. 2011. Task and user effects on reading patterns in information search. *Interacting with Computers* 23, 4 (2011), 346–362. https://doi.org/10.1016/j.intcom.2011.04.007

[7] Tilman Dingler, Dagmar Kern, Katrin Angerbauer, and Albrecht Schmidt. 2017. Text Priming - Effects of Text Visualizations on Readers Prior to Reading. In *Human-Computer Interaction – INTERACT 2017*, Regina Bernhaupt, Girish Dalvi, Anirudha Joshi, Devanuj K. Balkrishan, Jacki O'Neill, and Marco Winckler (Eds.). Springer International Publishing, Cham, 345–365.

[8] Laura A. Granka, Thorsten Joachims, and Geri Gay. 2004. Eye-tracking Analysis of User Behavior in WWW Search. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '04)*. ACM, New York, NY, USA, 478–479. https://doi.org/10.1145/1008992.1009079

[9] Jacek Gwizdka. 2014. Characterizing Relevance with Eye-tracking Measures. In *Proceedings of the 5th Information Interaction in Context Symposium (IIiX '14)*. ACM, New York, NY, USA, 58–67. https://doi.org/10.1145/2637002.2637011

[10] Daniel Hienert, Frank Sawitzki, and Philipp Mayr. 2015. Digital library research in action–supporting information retrieval in sowiport. *D-Lib Magazine* 21, 3/4 (2015).

[11] Joseph W Janes. 1991. Relevance judgments and the incremental presentation of document representations. *Information Processing & Management* 27, 6 (1991), 629–646.

[12] Ziming Liu. 2005. Reading behavior in the digital environment: Changes in reading behavior over the past ten years. *Journal of documentation* 61, 6 (2005), 700–712.

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

[14] Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. ELRA, Valletta, Malta, 45–50. http://is.muni.cz/publication/884893/en.

[15] Tefko Saracevic. 1969. Comparative effects of titles, abstracts and full texts on relevance judgments. *Proceedings of the American Society for Information Science* 6, 1 (1969), 293–299.

[16] H. Schmid. 1999. *Improvements in Part-of-Speech Tagging with an Application to German*. Springer Netherlands, Dordrecht, 13–25".

[17] Chih-Hsuan Wei, Hung-Yu Kao, and Zhiyong Lu. 2013. PubTator: a web-based text mining tool for assisting biocuration. *Nucleic acids research* 41, W1 (2013), W518–W522.

[18] Qian Yang, Gerard de Melo, Yong Cheng, and Sen Wang. 2017. HiText: Text Reading with Dynamic Salience Marking. In *Proceedings of the 26th International Conference on World Wide Web Companion (WWW '17 Companion)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 311–319. https://doi.org/10.1145/3041021.3054168