Word-Copying on a Website as a Word Complexity Indicator and the Relation to Web Users' Preferred Languages

Ilan Kirsh
kirsh@mta.ac.il
The Academic College of Tel Aviv-Yaffo
Tel Aviv, Israel

ABSTRACT

The first step toward accessibility improvement in the context of Human-Computer Interaction (HCI) is the identification of potential barriers. A recent study shows that words that are frequently copied to the clipboard by web users are relatively complex. A plausible reason is that users copy challenging words to search for a translation or more information. Accordingly, tracking word-copying operations of web users may be useful in identifying complex words. This study focuses on the users that apply word-copying. It shows significant differences in the frequency of word-copying operations among different populations of users. On the examined website, whose content is in English, users whose preferred language is not English copied single words to the clipboard significantly more frequently than users whose preferred language is English. Further analysis of the data also shows that word-copying was more frequent among users whose preferred languages have low proximity to English, such as Asian languages, compared to Western European languages. These results support the observation that word-copying indicates complexity, as it is reasonable to expect that native speakers of foreign languages (and especially of languages that are considerably different from the website's language) are more likely to need help with complex words. Word complexity is subjective and audience-dependent. This study contributes to the understanding of which users tend to use word-copying, and accordingly, in which context word-copying data can be used as a word complexity indicator. It also introduces a new practical approach for detecting language barriers in order to improve language accessibility on global websites.

CCS CONCEPTS

• **Information systems** → **Web log analysis**; **Traffic analysis**; **Browsers**; • **Human-centered computing** → **Empirical studies in HCI**; **Web-based interaction**.

KEYWORDS

Accessibility, Language Barriers, Clipboard, Copy, Browser, User Behavior, Text Simplification, Complex Word Identification (CWI), Web Analytics.

Asian CHI Symposium 2021 , May 8–13, 2021, Yokohama, Japan

ACM ISBN 978-1-4503-8203-8/21/05…$15.00
https://doi.org/10.1145/3429360.3468172

ACM Reference Format:
Ilan Kirsh. 2021. Word-Copying on a Website as a Word Complexity Indicator and the Relation to Web Users' Preferred Languages. In *Asian CHI Symposium 2021 (Asian CHI Symposium 2021), May 8–13, 2021, Yokohama, Japan.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3429360.3468172

1 INTRODUCTION

Evaluation is often the key to success in Human-Computer Interaction (HCI). Understanding the strengths and weaknesses of an existing implementation is the first step toward improving it [19]. Web analytics provides effective and scalable tools for the evaluation of website usage [2, 6, 7]. The data that can be collected automatically (and therefore at low cost) from online users' activity can be used to assess the website usability [4, 5, 12, 13]. An obvious requirement for websites is that their content should be clear and easily understood by their target audiences. Accordingly, a global website should eliminate unnecessary language barriers by presenting readable and simple-to-understand text, and in some cases also by providing a translation of the website content or parts of it.

Automatic text simplification methods can be used to highlight text complexity and suggest improvements that can increase the readability and understandability of textual content. The first step in text simplification is identifying challenging parts within the text, such as complex words [18]. Automatic Complex Word Identification (CWI) is mainly based on analyzing certain word features. For example, complex words tend to have more syllables [3, 15, 17] and more characters [1, 17]. Complex words are usually also less frequent than simple words (and therefore less familiar) [14, 17]. Various CWI implementations apply machine learning methods using combinations of indicators as features to assess word complexity [16, 21].

Users copy strings of various types to the clipboard, including words and phrases to look up elsewhere, code fragments for use in software projects, and key sentences for citations and text summaries [10, 11]. A recent study found that individual words that are copied frequently by web users (i.e. not as part of a copying operation of a text fragment consisting of multiple words) are usually relatively complex [8]. A reasonable explanation of this phenomenon is that some users copy complex words to the clipboard to search for more information (including translations) on the internet when they struggle to understand the text. Accordingly, word-copying (as well as other interactions of users with websites, such as moving the mouse slowly near words [9]) may indicate text complexity. Therefore, word-copying data could be used for the identification of complex words as part of a text simplification process.

Word complexity is subjective and audience dependent [16, 20]. Word-copying, as a word complexity indicator, may potentially have

the advantage of revealing what the real users of a website consider to be complex. However, a website audience is not necessarily homogeneous and may consist of different types of users, often from different countries and with different levels of proficiency in the website's language.

This study examined the effect of web users' preferred languages on word-copying. It found that on the examined website, whose content is in English, users whose preferred language is English copied words less frequently than users with other preferred languages. This supports the hypothesis that word-copying correlates with users struggle with challenging words, and accordingly, that word-copying may be effective in detecting text that is complex for the users. The study went further and compared the frequency of word-copying operations among users with different preferred languages (which are not English). The results show that word-copying was especially common among users with preferred languages that have low proximity to English. These results have several practical implications regarding the use of word-copying data for automatic detection of language barriers on websites and automatic text simplification, as discussed in this paper.

2 THE DATASET

The dataset used in this study consists of activity data of users of the ObjectDB website[1] collected over six months in 2020. The ObjectDB website contains mainly technical textual content for programmers.

Fig. 1 illustrates the process that has been used to collect data for this study. To track copy operations, a reference to a *Tracking Script* was embedded in the website pages. As a result, when a web page was loaded it triggered a request to load the Tracking Script from the Tracking Server. Once loaded, the Tracking Script recorded page views and copy operations and reported them to the *Collector* component in the Tracking Server, which stored the data in a dedicated database. For every page view and every copy operation, the script recorded the user's most preferred language, which is the first language in the browser's **navigator.languages** field[2]. Following the common practice of web analytics, and to protect user privacy, all the collected data were anonymized. The *Reporter* component was used to retrieve the collected data and analyze the results.

The dataset contains 1,295,221 page views of 437,536 unique users (estimate, based on browser fingerprint) with 89 different preferred languages (after merging dialects, i.e. using only the first two letters of the browser language codes). 2,438 word-copying operations by 670 different users have been recorded. The most frequently copied words were: 'criteria' (59 times), 'embedded' (43 times), 'transient' (38 times), 'composite' (36 times), 'embeddable' (35 times), 'persistence' (34 times), and 'redundant' (22 times), which are all relatively complex, as analyzed and discussed in a previous paper [8]. This study, however, focuses on the users that copy words and their preferred languages, rather than on the copied words.

[1] https://www.objectdb.com

[2] https://developer.mozilla.org/en-US/docs/Web/API/NavigatorLanguage/languages

3 RESULTS AND EVALUATION

As discussed above, copying single words from websites is already known to be associated with word complexity [8]. The main research question of this study is whether word-copying operations are more common among users whose preferred language is different from the website's content language. The data in this study are anonymized and the specific linguistic abilities of the users are unknown. Therefore, the user preferred language as specified in the browser settings is used. Note that this is often automatically set by the browser, for example, by following the operating system locale, as set by the user. The browser's most preferred language might not always reflect the actual preference of the user (e.g. when the same computer is used by several users), however, it is expected to be a language that the user can use, as this is the browser's UI language.

The null hypothesis is that there is no difference, regarding word-copying activity, between people whose preferred language is the website language and people whose preferred language is different. To check this hypothesis, the users in the dataset were divided into two groups: those whose preferred language is English, and all other users. For each group, the rate of users that applied word-copying at least once was calculated (counting users rather than copy operations may assist in avoiding a possible bias due to a few very active users). Table 1 shows that users whose preferred language is not English copied words more frequently (184 vs. 131 per 100,000). The difference between the two groups is highly significant (P = 0.000011, using Fisher's exact two-tailed test). Accordingly, the null hypothesis is rejected, and the conclusion is that users of this website whose preferred language is not English are generally more likely to use word-copying.

Another question is whether word-copying is particularly frequent among users of specific languages. Table 2 lists all the preferred languages in the dataset that had at least five different word-copying users (i.e. users that applied word-copying at least once). The languages are ordered by the rates of word-copying users per 100,000 users. The results show significant differences among languages. For example, the rate for Chinese is 881 word-copying users per 100,000, whereas the rate for Italian is only 46 per 100,000. Notably, there are four languages in Table 2 with rates lower than English, but for three of them, the difference is not statistically significant: the 'Significant Difference' column specifies the Fisher's exact two-tailed test values in comparison with English.

The results in Table 2 highlight an interesting pattern. Languages that are linguistically more closely related to English, such as German, Spanish, French, Portuguese, and Italian (Germanic and Romance languages) have lower rates of word-copying users relative to languages that are considered far from English, such as Chinese, Vietnamese, and Japanese. The 'Lexical Distance' column specifies the lexical distance of a language as calculated by elinguistics[3]. Interestingly, the lexical distance values are above 60 for the seven languages with the highest rates of word-copying users in Table 2 and below 60 for the six languages (including English itself) with the lowest rates of word-copying users.

An additional indication regarding the significant distance between the languages at the top of Table 2 and English is their writing

[3] http://www.elinguistics.net/Compare_Languages.aspx

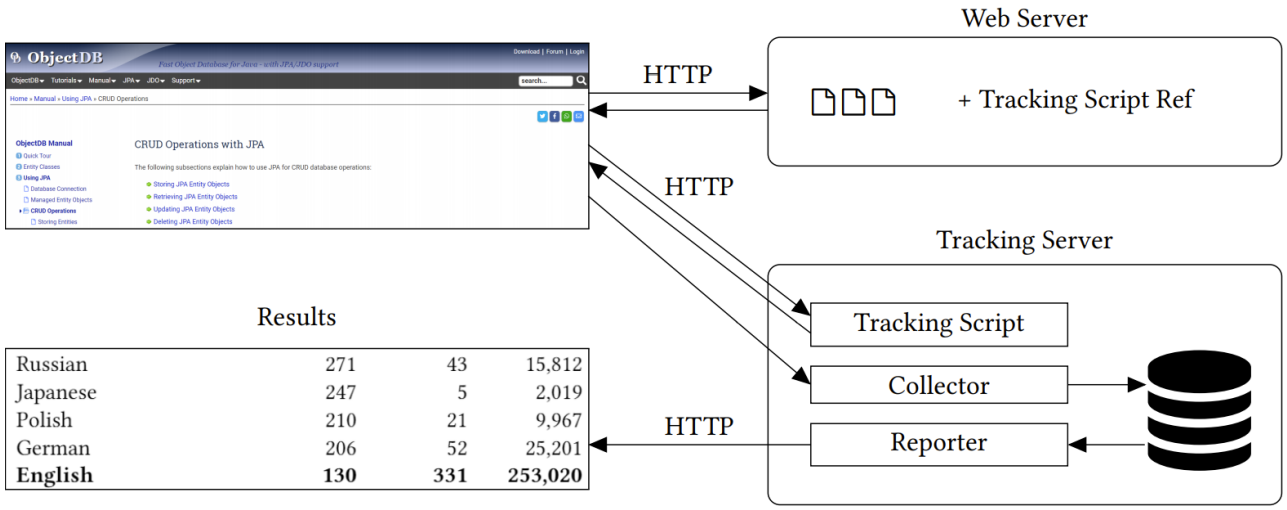

Figure 1: Tracking word-copying operations and preferred languages

Table 1: Comparison of word-copying by preferred language: English vs. non-English

Preferred Language	Word-Copying Users per 100,000	Word-Copying Users	Total Users
English	131	331	253,020
Non-English	184	339	184,516
Any	153	670	437,536

Table 2: Word-copying among users of different preferred languages

Preferred Language	Word-Copying Users per 100,000 Users	Word-Copying Users	Total Users	Significant Difference (P)*	Lexical Distance*	Latin Alphabet
Chinese	881	78	8,857	0.0000	82.4	✗
Ukrainian	762	7	919	0.0003	60.3	✗
Greek	754	5	663	0.0020	69.9	✗
Vietnamese	484	14	2,891	0.0000	96.1	✗
Russian	272	43	15,812	0.0000	60.3	✗
Japanese	248	5	2,019	0.1994	88.3	✗
Polish	211	21	9,967	0.0482	66.9	✓
German	206	52	25,201	0.0032	30.8	✓
English	**131**	**331**	**253,020**	**1.0000**	**0.0**	✓
Spanish	120	44	36,611	0.6413	57.0	✓
French	108	27	24,887	0.4041	48.7	✓
Portuguese	82	17	20,818	0.0544	59.8	✓
Italian	46	5	10,782	0.0122	47.8	✓

* compared to English

systems. The top rates of word-copying users in Table 2 are associated with languages that use non-Latin alphabets, i.e. use different alphabets than English.

These results are compatible with the observation that word-copying signals word complexity [8]: if word-copying is associated with seeking help in the understanding of text then users whose native language is considerably different from English may be more likely to need such help on a website with texts in English. Consequently, the results support the concept of automatic CWI, based on tracking interactions of users with websites. Moreover, as word complexity is subjective and audience dependent [16, 20], these results indicate that word-copying may be particularly effective in the identification of complex words for specific users: those with preferred languages with low proximity to the text's language, due to their specific needs and the higher availability of word-copying data for these populations. The correlation that was found between word-copying rates and preferred languages also indicates that word-copying data may be used for the identification of language barriers in specific populations.

4 CONCLUSIONS AND FUTURE WORK

This study examined the relation between word-copying activity on a website and the users' preferred languages, as specified in the browser settings. The results show that on the examined website, whose content is in English, word-copying was applied more frequently by users whose preferred language is not English. Moreover, there were also significant differences in word-copying rates among users of different languages. Word-copying was especially frequent among users whose preferred language has low proximity to English, such as Asian languages. These results are compatible with the observation that word-copying can be used as a word complexity indicator, as users whose preferred language is considerably different from English may require more help (on average) to understand text in English. As there may also be other major affecting factors, the results cannot determine a cause-effect relation.

Despite this limitation, the correlations that were found in this study are important. First of all, the results support the hypothesis that word-copying correlates with word complexity [8], by using a different perspective (examining the users that apply word-copying rather than the words that they copy), as we can expect that users for whom the text language is challenging would seek clarification by copying words more frequently. This strengthens the argument in favor of using word-copying as an HCI solution for text complexity evaluation and simplification based on tracking users' interactions with websites.

Word complexity is subjective and audience-dependent. The results indicate that word-copying could be particularly effective in detecting complexity and simplifying text for the benefit of non-native speakers of the text language, due to the higher need for support as well as the increased availability of word-copying data related to these populations. The differences in rates of word-copying users among different populations can also be useful for prioritizing investment in language assistant tools for certain audiences. For example, a very-high word-copying rate among users with a particular preferred language may justify providing a fully translated version of the text to that language. Medium levels of rates of word-copying users may justify other measures, including text simplification, providing a glossary of complex terms, or presenting tooltips with explanations of complex words. In adaptive websites, such tools can be applied selectively, for example, only for populations or individual users with high word-copying rates.

This study is based on anonymized web usage data, thus the linguistic background of the users is unknown, and the results are based on the preferred language setting of the users' browsers. Future work may further increase the knowledge about word-copying and specifically who applies it, by tracking users for which additional data are available, including more precise details regarding their language proficiency. The results of this study, which highlight differences among user populations, can contribute to further work on CWI and text simplification based on HCI and web analytics data such as word-copying. They can also promote new web analytics tools for the detection of language barriers in specific populations, based on user behavior on websites.

REFERENCES

[1] M. Coleman and T. L. Liau. 1975. A Computer Readability Formula Designed for Machine Scoring. *Journal of Applied Psychology* 60 (08 1975), 283–284. Issue 2.

[2] Brent Dykes. 2014. *Web Analytics Kick Start Guide: A Primer on the Fundamentals of Digital Analytics.* Peachpit, Pearson Education, USA.

[3] Robert Gunning. 1952. *The Technique of Clear Writing.* McGraw-Hill, New York, New York.

[4] Layla Hasan, Anne Morris, and Steve Probets. 2009. Using Google Analytics to Evaluate the Usability of E-Commerce Sites. In *Proceedings of the 1st International Conference on Human Centered Design.* Springer Berlin Heidelberg, Berlin, Heidelberg, 697–706. https://doi.org/10.1007/978-3-642-02806-9_81

[5] Razib Iqbal, Matthew Scott, and Tarah Cleveland. 2016. Measuring Actual Visitor Engagement in News Websites. In *4th International Workshop on News Recommendation and Analytics (INRA 2016).* Association for Computing Machinery, Halifax, Canada, 4 pages.

[6] Avinash Kaushik. 2007. *Web Analytics: An Hour a Day.* SYBEX Inc., USA.

[7] Avinash Kaushik. 2010. *Web Analytics 2.0.* SYBEX Inc., USA.

[8] Ilan Kirsh. 2020. Automatic Complex Word Identification Using Implicit Feedback From User Copy Operations. In *Proceedings of the 21st International Conference on Web Information Systems Engineering (WISE 2020), Lecture Notes in Computer Science.* Springer International Publishing, Cham, 155–166. https://doi.org/10.1007/978-3-030-62008-0_11

[9] Ilan Kirsh. 2020. Using Mouse Movement Heatmaps to Visualize User Attention to Words. In *Proceedings of the 11th Nordic Conference on Human-Computer Interaction (NordiCHI 2020), Tallinn, Estonia.* Association for Computing Machinery, New York, NY, USA, 117:1–5. https://doi.org/10.1145/3419249.3421250

[10] Ilan Kirsh. 2020. What Web Users Copy to the Clipboard on a Website: A Case Study. In *Proceedings of the 16th International Conference on Web Information Systems and Technologies (WEBIST 2020).* INSTICC, SciTePress, Setúbal, Portugal, 303–312. https://doi.org/10.5220/0010113203030312

[11] Ilan Kirsh and Mike Joy. 2020. An HCI Approach to Extractive Text Summarization: Selecting Key Sentences Based on User Copy Operations. In *Proceedings of the 22nd HCI International Conference (HCII 2020), Communications in Computer and Information Science.* Springer International Publishing, Cham, 335–341. https://doi.org/10.1007/978-3-030-60700-5_43

[12] Ilan Kirsh and Mike Joy. 2020. Splitting the Web Analytics Atom: From Page Metrics and KPIs to Sub-Page Metrics and KPIs. In *Proceedings of the 10th International Conference on Web Intelligence, Mining and Semantics (WIMS 2020), Biarritz, France.* Association for Computing Machinery, New York, NY, USA, 33–43. https://doi.org/10.1145/3405962.3405984

[13] Lakhwinder Kumar, Hardeep Singh, and Ramandeep Kaur. 2012. Web Analytics and Metrics: A Survey. In *Proceedings of the International Conference on Advances in Computing, Communications and Informatics* (Chennai, India) *(ICACCI '12).* Association for Computing Machinery, New York, NY, USA, 966–971. https://doi.org/10.1145/2345396.2345552

[14] Gondy Leroy and David Kauchak. 2013. The effect of word familiarity on actual and perceived text difficulty. *Journal of the American Medical Informatics Association : JAMIA* 21 (10 2013), 169–172. https://doi.org/10.1136/amiajnl-2013-002172

[15] G. H McLaughlin. 1969. SMOG grading: A new readability formula. *Journal of Reading* 12 (08 1969), 639–646. Issue 8.

[16] Gustavo Paetzold and Lucia Specia. 2016. SemEval 2016 Task 11: Complex Word Identification. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016).* Association for Computational Linguistics, San Diego,

California, 560–569. https://doi.org/10.18653/v1/S16-1085

[17] Matthew Shardlow. 2013. A Comparison of Techniques to Automatically Identify Complex Words". In *51st Annual Meeting of the Association for Computational Linguistics Proceedings of the Student Research Workshop*. Association for Computational Linguistics, Sofia, Bulgaria, 103–109. https://www.aclweb.org/anthology/P13-3015

[18] Matthew Shardlow. 2014. A Survey of Automated Text Simplification. *International Journal of Advanced Computer Science and Applications(IJACSA), Special Issue on Natural Language Processing 2014* 4, 1 (2014), 58–70. https://doi.org/10.14569/SpecialIssue.2014.040109

[19] Helen Sharp, Jennifer Preece, and Yvonne Rogers. 2019. *Interaction Design. Beyond Human-Computer Interaction, 5th Edition.* John Wiley & Sons, Inc., USA.

[20] Lucia Specia, Sujay Kumar Jauhar, and Rada Mihalcea. 2012. SemEval-2012 Task 1: English Lexical Simplification. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics* (Montréal, Canada) *(SemEval '12)*. Association for Computational Linguistics, USA, 347–355.

[21] Seid Muhie Yimam, Chris Biemann, Shervin Malmasi, Gustavo Paetzold, Lucia Specia, Sanja Štajner, Anaïs Tack, and Marcos Zampieri. 2018. A Report on the Complex Word Identification Shared Task 2018. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*. Association for Computational Linguistics, New Orleans, Louisiana, 66–78. https://doi.org/10.18653/v1/W18-0507