Automatic Complex Word Identification Using Implicit Feedback From User Copy Operations

Ilan Kirsh[0000−0003−0130−8691]

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

Abstract. Complex Word Identification (CWI) is one of the key components of lexical text simplification. This paper proposes a new approach to CWI on websites, based on tracking what web users copy to their clipboards. Users may copy to the clipboard words that they are not familiar with or that make the text difficult to understand, in order to search for more information on the internet. Accordingly, this study examines the hypothesis that word copying on a website is an indicator of word complexity. Copied words on a sample website are compared to uncopied words using three simple word complexity indicators: number of syllables, number of characters, and general word frequency. The results show that copied words are more likely to be evaluated as complex than uncopied words and words that are copied more frequently are more likely to be evaluated as complex than words that are copied less frequently, by all three indicators. Consequently, word copying on a website can be considered a novel CWI indicator. Unlike traditional CWI indicators, which are based on static word features, this new indicator provides a different approach by targeting complex words based on dynamic user behavior. Therefore, simplifying these complex words might be particularly helpful to the readers. Further work should evaluate using this word copying indicator in complete CWI and text simplification implementations.

Keywords: Complex Word Identification (CWI) · Lexical Text Simplification · Clipboard Copy and Paste · Web Usage Mining · Web Pages

1 Introduction

Shardlow defined text simplification as "the process of modifying natural language to reduce its complexity and improve both readability and understandability" [17]. Similarly to text translation and text summarization, text simplification can also benefit from automation. But translation and summarization could be more tolerant of errors than simplification, at least in some applications. Errors in text simplification could lead to output text that is more complex than the input, making the "simplification" result unusable [17].

Automatic text simplification is a challenging task. Many studies focus on the more modest goal of lexical text simplification, which involves replacing individual complex words with simpler words with similar meanings, without changing sentence structures and grammar [6, 14, 19, 20]. Lexical text simplification can

In Proceedings of the 21st International Conference on Web Information Systems Engineering (WISE 2020), LNCS, Springer, 2020. The final authenticated version is available online at https://doi.org/10.1007/978-3-030-62008-0_11

be performed in stages, where the first stage is Complex Word Identification (CWI) [17]. Any word that reduces the readability or understandability of the text may be considered complex. Word complexity is audience dependent. For example, a word can be simple for native speakers of a language and complex for non-native speakers. Words that are identified as complex (for the prospective audience) are candidates for substitution with simpler words in the next stages of the text simplification process.

Simple features of words can be used as indicators of word complexity. Three of the most commonly used word complexity indicators are:

1. **Syllable Count** - Complex words tend to be longer and have more syllables. Readability tests such as the Gunning fog index [5] and the SMOG grade [12] classify words with three syllables or more as complex.
2. **Character Count** - An alternative readability test, the Coleman–Liau index [2], uses the number of characters in a word as a complexity measure. On average, complex words tend to have more characters.
3. **Frequency** - Less frequent words are usually less familiar and therefore more complex than more frequent words [11].

These three indicators have been found to be strongly correlated with word complexity [16]. Other indicators, sense count and synonym count, which may indicate potential word ambiguity and therefore complexity, have been found to have weaker correlations with word complexity [16].

Many CWI implementations use a combination of indicators, including the indicators described above, as features in machine learning models. Various machine learning methods, including SVM classifiers, Random Forests, Neural Networks, and Bayesian Ridge classifiers, have been examined in the SemEval 2016 task 11 [15] and the CWI 2018 shared task [21]. Taking into account the context in which words appear in the text can improve the results [4].

Eye gaze tracking is commonly used in research on reading behaviors. It may be used in the context of CWI to identify complex words, because encountering complex words may be reflected in the user eye gaze, for example, as extended reading time [1]. Identifying words that are complex for real users, using eye tracking methods, could be more reliable than using static word complexity indicators. However, collecting eye tracking data requires special equipment and user collaboration, so the scope of this approach is limited, and it is usually impractical to collect eye gaze data from ordinary users on public websites.

This study proposes a new approach to automatic identification of complex words on web pages by tracking web users' copy operations. Users copy strings of various types to the clipboard [8], including words and phrases to look up elsewhere [10], key sentences for citations and text summaries [9], and programming code fragments [10]. This paper shows that copying words to the clipboard is a word complexity indicator. Consequently, tracking word copying on websites should be considered a new technique in CWI and text simplification. Similarly to eye gaze tracking, it has the benefit of tracking real users and finding their real needs, but it is not constrained by the limitations that make eye tracking impractical on most websites.

2 Implementation

The architecture of the CWI implementation that was developed and explored as part of this study is shown in Figure 1.

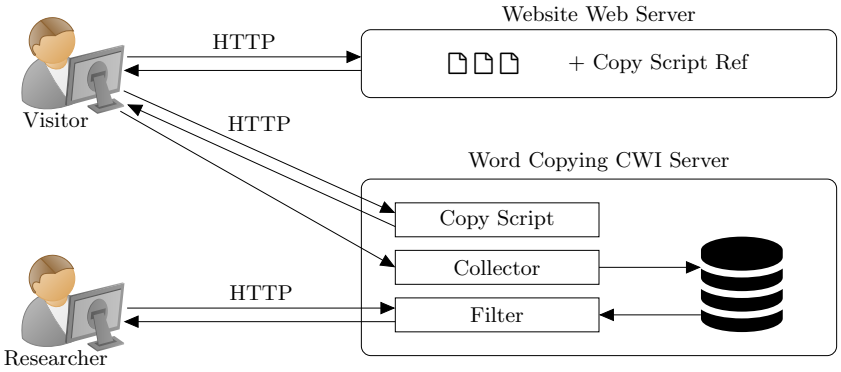

Fig. 1. Architecture of the Word Copying CWI Implementation

The pages of the website are modified to include a reference to a *Copy Script*. When these pages are loaded into a visitor's browser, the browser follows the reference and loads the Copy Script. The Copy Script includes JavaScript code that tracks copy events and reports them to the *Collector* component in the server, which stores the data (following anonymization) in a dedicated database.

The *Filter* component is responsible for distinguishing copied words from all the other copied strings, and its output is the copied words along with their corresponding copying frequencies. Only copy operations of single text words are counted, copy operations of multiple words are filtered out, as well as copy operations of text strings contained in HTML PRE elements, which are often code fragments. A valid word is defined in this context (of CWI for texts in English), as a string consisting of lower case letters, except for the first character that can also be an upper case letter. All copied strings are converted to lower case for case-insensitive counting.

This basic Filter implementation is not perfect. For example, it rejects some possible types of complex words (e.g. abbreviations), as well as complex phrases made up of several words. It is also adjusted for the specific website being tested. Nevertheless, this basic implementation is satisfactory for the purposes of this study. Further work is required in order to develop more advanced and general filtering methods that can be used on other types of websites.

This CWI implementation can be used as a standalone system, where the most frequently copied words (above a given threshold) are identified as complex, or in combination with other word complexity indicators, providing an additional source of information and indication of word complexity in a multi-indicator CWI system.

3 Experiments and Results

The CWI implementation was run on 231 documentation pages of the ObjectDB website (www.objectdb.com). Usage data were collected for a period of several months, ending in March 2020. 654,399 page views of 241,644 unique visitors (estimated) and 53,131 copy operations were recorded and used as the input dataset for the experiment.

Table 1 shows that the most frequently copied strings on this website are code fragments. Programmers may copy code fragments to their clipboards in order to paste them in their IDEs [8, 10].

Table 1. Most Frequently Copied Strings

#	Text (long strings are truncated)	Count
1	@SequenceGenerator(name="seq", initialValue=1, allocationSi...	1,143
2	@GeneratedValue(strategy=GenerationType.AUTO)	1,069
3	@GeneratedValue(strategy=GenerationType.SEQUENCE, gene...	686
4	@GeneratedValue	464
5	@Embeddable	454
6	@IdClass(ProjectId.class)	391
7	ParameterExpression<Integer> p = cb.parameter(Integer.class);	364
8	@Transient	358
9	CriteriaBuilder cb = em.getCriteriaBuilder();	350
10	@GeneratedValue(strategy=GenerationType.IDENTITY)	315

Table 2 shows the most frequently copied text strings, obtained by filtering out copy operations of code fragments that are wrapped in HTML PRE tags.

Table 2. Most Frequently Copied Text Strings

#	Text (long strings are truncated)	Count
1	JPQL	109
2	Composite Primary Key	85
3	EntityManager	67
4	The IDENTITY strategy also generates an automatic value ...	63
5	ObjectDB	58
6	EntityManagerFactory	49
7	persistence.xml	47
8	The sequence strategy consists of two parts - defining a named ...	46
9	Marking a field with the @GeneratedValue annotation specifies ...	39
10	Embedded Primary Key	37

Most of the strings in Table 2 are technical terms, code fragments (embedded in the text), and long sentences, rather than complex words. Users may copy technical terms to the clipboard in order to search for more information about them [8] and complete sentences for citations and text summaries [9]. None of these strings is a valid word according to the definition in section 2, so they are all filtered out by the Filter component.

Table 3 shows the 30 most frequently copied text words, produced by the Filter. For a proficient English speaking user, the words in Table 3 may not seem very complex, but they may be complex relative to the website vocabulary level, which consists of mainly simple English words. Very simple and frequent words (e.g. "the", "of", etc.) are not on the list, even though they are very common and appear many times on every page. This is the first indication that users are more likely to copy complex words than simple words. Section 4 presents statistical evidence that supports this hypothesis.

Table 3. Most Frequently Copied Words

#	Word	Count	#	Word	Count	#	Word	Count
1	criteria	36	11	entity	10	21	detached	7
2	transient	24	12	persistable	10	22	allocation	6
3	embeddable	21	13	hollow	9	23	retrieves	6
4	embedded	20	14	explicitly	9	24	query	6
5	persistence	17	15	pessimistic	9	25	identity	6
6	composite	16	16	polymorphic	9	26	equivalent	6
7	redundant	14	17	persistent	8	27	sequence	6
8	retrieved	14	18	dangling	8	28	persist	6
9	explicit	11	19	instantiation	8	29	ascending	6
10	cascaded	11	20	retrieval	7	30	orphan	6

The significant differences in copying frequency (the "Count" columns) in tables 1, 2, and 3 show that on this website code fragments are copied much more frequently than text words. Due to these differences, complex words are only exposed in Table 3, after filtering out the other elements.

Figures 2 and 3 show some of the resulting complex words in the context of the website text. The words that were copied by users are framed.

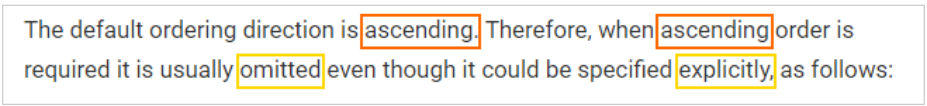

Fig. 2. Copy Visualization: "ascending" and Other Words

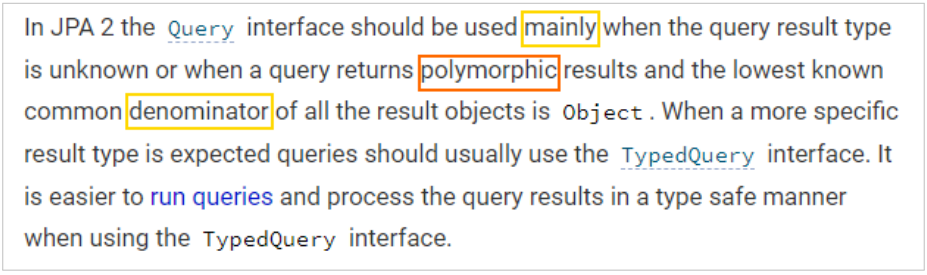

Fig. 3. Copy Visualization: "polymorphic" and Other Words

The word "ascending" (#29 in Table 3) was copied 6 times, and the word "polymorphic" (#16 in Table 3) was copied 9 times. Each of the other framed words (with the yellow border) was copied once in the text shown above (the word "explicitly" was copied 9 times in total on all the tracked web pages).

4 Evaluation

The dataset contains 53,131 copy operations resulting from 654,399 page views, though most of the copy operations are related to code. Only 823 copy operations are accepted by the Filter as related to valid text words (based on the strict definition of valid words in section 2), and these copy operations relate to 326 different words. For evaluation purposes, words not included in the list of the 333,333 most frequent words in the Google's Trillion Word Corpus [13] (e.g. "persistable") were excluded, so the evaluation focused on 316 distinct copied words in 801 copy operations.

These relatively small numbers may not be sufficient for a complete evaluation of the proposed CWI approach as a standalone implementation (e.g. compared to other CWI methods), but as shown in this section, they are sufficient to conclude that words that are copied by users are more likely to be complex words than words that are not copied. In other words, copying words to the clipboard on a website can be considered a CWI indicator.

The null hypothesis is that there is no difference in complexity between copied and uncopied words. To test the null hypothesis we can use the three CWI indicators that have been described in section 1: syllable count, character count, and frequency. In addition to simplicity, using these three indicators rather than testing against human tagging of complex words, which is often used in full evaluation and comparison of CWI implementations (e.g. in [15, 21]), has the advantage of objectivity and avoiding biases. Human complex word tagging was proved to be subjective and inconsistent among different taggers [15, 18].

Google's Trillion Word Corpus can be used to estimate word frequency. Given the list of 333,333 most frequently used words in this corpus ordered by decreasing frequency [13], we can define the frequency rank of a word as its position in the list (e.g. #1, the highest, for the word "the", which is the most frequent word

in the corpus) and expect words that are less frequent to be generally more complex. Syllables in words have been counted using a Java library[1], which although not 100% accurate, is sufficient for the purpose.

Table 4 shows the values of these three indicators for the most frequently copied words in the dataset (words with at least 8 copy operations). We can reject the null hypothesis by showing that word complexity as evaluated by these indicators is significantly different for copied and uncopied words, i.e. copied words are more likely to be evaluated as complex than uncopied words, with a significant statistical difference.

Table 4. Complexity Indicators Values for the Most Frequently Copied Words

#	Word	Copies	Syllables	Characters	Frequency Rank
1	criteria	36	3	8	2,468
2	transient	24	2	9	12,548
3	embeddable	21	4	10	89,240
4	embedded	20	3	8	5,356
5	persistence	17	3	11	14,474
6	composite	16	3	9	6,414
7	redundant	14	3	9	12,423
8	retrieved	14	2	9	7,609
9	explicit	11	3	8	6,371
10	cascaded	11	3	8	70,361
11	entity	10	3	6	4,067
12	hollow	9	2	6	9,566
13	explicitly	9	4	10	8,551
14	pessimistic	9	4	11	32,253
15	polymorphic	9	4	11	37,407
16	persistent	8	3	10	9,645
17	dangling	8	2	8	25,694
18	instantiation	8	5	13	42,186

For the evaluation, the words on the website are divided into two sets: 3,234 uncopied words, which were not copied at all in a single-word copy operation, and 311 copied words, which were copied at least once. Table 5 presents a comparison of these two sets using eight binary criteria for word complexity, based on the three CWI indicators with ranges of values that indicate complexity (relative to the average values for the website's words). Each row in the table shows how many words in these two sets meet each criterion, and so are identified as complex.

As shown in the table, copied words are more likely than uncopied words to be classified as complex by all of the examined complexity criteria. Note

[1] https://github.com/wfreitag/syllable-counter-java

that these complexity criteria are based on simple and generic indicators, which cannot determine complexity precisely. For example, the words "Wikipedia" and "electricity" are not necessarily complex, despite the numbers of syllables and characters. In addition, the division into copied and uncopied words is highly dependent on the dataset (uncopied words could become copied words if more copy operations were recorded for more users). Therefore, the exact percentages of words that are classified as complex by these complexity criteria are not expected to be accurate or important on their own. In fact, it might be reasonable to expect the new word copying indicator to be more accurate than the three indicators that are used in this evaluation, as discussed in section 5.

Table 5. Uncopied Words vs. Copied Words

Complexity Criterion		Uncopied Words	Copied Words	p-value
Syllables	≥ 3	1,378/3,234 (42.6%)	184/311 (59.2%)	0.000000
	≥ 4	536/3,234 (16.6%)	79/311 (25.4%)	0.000160
Characters	≥ 8	1,420/3,234 (43.9%)	198/311 (63.7%)	0.000000
	≥ 9	979/3,234 (30.3%)	140/311 (45.0%)	0.000000
	≥ 10	623/3,234 (19.3%)	85/311 (27.3%)	0.001046
	≥ 11	356/3,234 (11.0%)	48/311 (15.4%)	0.024497
Frequency Rank	$\geq$ 5,000	1,369/3,234 (42.3%)	166/311 (53.4%)	0.000198
	$\geq$ 10,000	782/3,234 (24.2%)	105/311 (33.8%)	0.000347

The results in Table 5 reject the null hypothesis, as they show a significant statistical difference between copied words and uncopied words for each of the eight word complexity criteria. Such differences are not expected under the null hypothesis. For each complexity criterion (a row in the table, representing a 2x2 contingency table) the p-value is calculated using the two-tailed Fisher's exact test.

Table 6. Words Copied Once vs. Words Copied at Least 8 Times

Complexity Criterion		Number of Times Copied		p-value
		Exactly Once	At Least 8 Times	
Syllables	≥ 3	88/167 (52.7%)	14/18 (77.8%)	0.048170
	≥ 4	38/167 (22.8%)	6/18 (33.3%)	0.381191
Characters	≥ 8	93/167 (55.7%)	16/18 (88.9%)	0.005636
	≥ 9	64/167 (38.3%)	11/18 (61.1%)	0.077752
	≥ 10	39/167 (23.4%)	7/18 (38.9%)	0.158121
	≥ 11	26/167 (15.6%)	4/18 (22.2%)	0.500274
Frequency Rank	$\geq$ 5,000	80/167 (47.9%)	16/18 (88.9%)	0.000879
	$\geq$ 10,000	49/167 (29.3%)	9/18 (50.0%)	0.105954

To examine if the frequency of copying each word matters, Table 6 compares two subsets of the set of copied words: the words that have been copied exactly once and the words that have been copied at least 8 times (shown in Table 4). Table 6 shows that words in the second subset are more likely than words in the first subset to be classified as complex by each of the eight complexity criteria. Due to the sizes of the sets, high statistical significance for these differences (p-value ≤ 0.05) is obtained only for the more inclusive criteria, $Syllables \geq 3$, $Characters \geq 8$, and $FrequencyRank \geq 5,000$.

5 Discussion

Section 4 shows that copied words are more likely to be evaluated as complex than uncopied words and words that are copied more frequently are more likely to be evaluated as complex than words that are copied less frequently, by three different CWI indicators. This has been shown for one website, and further experiments on other websites are required in order to establish these findings.

A reasonable explanation of the connection between complex words and the user behavior of copying words to the clipboard is that users search for definitions and translations of complex words. It is also possible that some users copy familiar but complex words in order to paste them instead of typing them while writing text. It is difficult to think of other convincing reasons as to why users copy to the clipboard regular words, such as "composite", "redundant", "explicit", and "ascending". Copying of very simple words (e.g. "the", "of", etc.) was not observed in the dataset.

Consequently, word copying on a website can be considered a novel CWI indicator. It may be used in a standalone CWI implementation, as described in section 2, where the most frequently copied words are identified as complex, or in combination with other CWI indicators in a multi-indicator CWI implementation. In both cases, the output of the CWI implementation can be used for text simplification (automatic or manual).

Using copy operations for CWI requires large amounts of web usage data, as shown by the demonstration of the filtering process in section 3. The dataset used in this study is based on web usage data collected over several months from a medium traffic website. On low traffic websites, copy operations may be less effective for CWI. On high traffic websites (e.g. Wikipedia) they could be significantly more effective. Collecting data for longer periods may help.

Tracking copy operations is related to *session recording*, which is a common practice in modern web analytics where user activity on websites, including mouse movements and keystrokes, is recorded. It raises interesting questions regarding user privacy and personal data protection, due to the risk of collecting sensitive personal information intentionally or unintentionally [3]. However, session recording does not necessarily require prior user consent under personal data protection regulations, such as GDPR (under certain terms, as discussed by the IT and privacy lawyer Arnoud Engelfriet [3]). Sensitive personal data, which are not required for CWI, should not be collected. If the data collected

are completely anonymized, which is a standard practice in web analytics, then they are no longer considered personal data (e.g. according to GDPR). In some sense, counting copy operations is similar to counting page-views, which has always been considered a legitimate web analytics practice.

The evaluation in section 4 shows significant statistical evidence that word copying is a CWI indicator and that this indicator is stronger for words that are copied more frequently. Note that the three indicators that have been used for evaluation in section 4 are not very accurate and cannot be used to evaluate the effectiveness of other CWI indicators. It is reasonable to expect that the word copying indicator (assuming it reflects users' need for assistance with complex words) is more accurate than the three indicators that have been used to test it. However, being well known tested, objective word complexity indicators, they are useful for showing that word copying is also a CWI indicator. Further work is required to assess the effectiveness of the word copying indicator, both as a standalone indicator and in combination with other CWI indicators.

The word copying indicator reflects the collective experience of the website audience by considering copy operations as implicit user "votes". Words identified by these votes as complex might be the biggest barriers to understanding the text and may mostly appear in paragraphs that are read more frequently. Therefore, simplifying these complex words might be particularly helpful to the readers. Accordingly, this new approach might have the potential of being more effective, reliable, and accurate than other CWI indicators, and it might also be more accurate than human tagging of complex words, which is subjective and inconsistent among different experts [15, 18]. Further evaluation is required in order to assess this potential.

Other user activities on websites might also indicate word complexity. An interesting hypothesis that has to be tested in this context, is that slower mouse movements near words also indicate word complexity. Some users move the mouse cursor during reading to mark the reading position, so slowing or stopping near words might indicate difficulties in reading or understanding [7].

6 Conclusions and Further Work

This study introduces a new approach to automatic CWI on websites, based on tracking copy operations of users. An experiment on a sample website shows that copied words are more likely to be evaluated as complex than uncopied words and words that are copied more frequently are more likely to be evaluated as complex than words that are copied less frequently, by three different word complexity indicators. Consequently, word copying on a website can be considered a novel CWI indicator, which targets complex words based on real user behavior.

Further work should investigate using this word copying indicator in complete CWI and text simplification implementations, and evaluate the effectiveness of using copy operations in CWI on various types of websites.

References

1. Bingel, J., Barrett, M., Klerke, S.: Predicting misreadings from gaze in children with reading difficulties. In: Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications. pp. 24–34. Association for Computational Linguistics, New Orleans, Louisiana (Jun 2018). https://doi.org/10.18653/v1/W18-0503, https://www.aclweb.org/anthology/W18-0503
2. Coleman, M., Liau, T.L.: A computer readability formula designed for machine scoring. Journal of Applied Psychology **60**, 283–284 (08 1975)
3. Gilliam Haije, E.: Are session recording tools a risk to internet privacy (03 2018), https://mopinion.com/are-session-recording-tools-a-risk-to-internet-privacy/
4. Gooding, S., Kochmar, E.: Complex word identification as a sequence labelling task. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 1148–1153. Association for Computational Linguistics, Florence, Italy (07 2019). https://doi.org/10.18653/v1/P19-1109, https://www.aclweb.org/anthology/P19-1109
5. Gunning, R.: The Technique of Clear Writing. McGraw-Hill, New York, New York (1952)
6. Horn, C., Manduca, C., Kauchak, D.: Learning a lexical simplifier using Wikipedia. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). pp. 458–463. Association for Computational Linguistics, Baltimore, Maryland (Jun 2014). https://doi.org/10.3115/v1/P14-2075, https://www.aclweb.org/anthology/P14-2075
7. Kirsh, I.: 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, forthcoming. Association for Computing Machinery, New York, NY, USA (10 2020)
8. Kirsh, I.: 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), forthcoming. INSTICC, SciTePress, Setúbal, Portugal (11 2020)
9. Kirsh, I., Joy, M.: 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 (07 2020)
10. Kirsh, I., Joy, M.: 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. pp. 33–43. Association for Computing Machinery, New York, NY, USA (06 2020). https://doi.org/10.1145/3405962.3405984, https://doi.org/10.1145/3405962.3405984
11. Leroy, G., Kauchak, D.: The effect of word familiarity on actual and perceived text difficulty. Journal of the American Medical Informatics Association : JAMIA **21**, 169–172 (10 2013). https://doi.org/10.1136/amiajnl-2013-002172
12. McLaughlin, G.H.: Smog grading: A new readability formula. Journal of Reading **12**, 639–646 (08 1969)
13. Norvig, P.: Data accompanying chapter 14, Natural Language Corpus Data, in Beautiful Data by Toby Segaran and Jeff Hammerbacher, pp. 219–242. O'Reilly Media, Inc., USA (2009), https://norvig.com/ngrams/

14. Paetzold, G., Specia, L.: Benchmarking lexical simplification systems. In: Chair), N.C.C., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S. (eds.) Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016). European Language Resources Association (ELRA), Paris, France (may 2016)
15. Paetzold, G., Specia, L.: SemEval 2016 task 11: Complex word identification. In: Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016). pp. 560–569. Association for Computational Linguistics, San Diego, California (jun 2016). https://doi.org/10.18653/v1/S16-1085, https://www.aclweb.org/anthology/S16-1085
16. Shardlow, M.: 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. pp. 103–109. Association for Computational Linguistics, Sofia, Bulgaria (08 2013), https://www.aclweb.org/anthology/P13-3015
17. Shardlow, M.: A survey of automated text simplification. International Journal of Advanced Computer Science and Applications(IJACSA), Special Issue on Natural Language Processing 2014 **4**(1), 58–70 (2014). https://doi.org/10.14569/SpecialIssue.2014.040109, http://dx.doi.org/10.14569/SpecialIssue.2014.040109
18. Specia, L., Jauhar, S.K., Mihalcea, R.: Semeval-2012 task 1: English lexical simplification. In: Proceedings of the First Joint Conference on Lexical and Computational Semantics. p. 347–355. SemEval '12, Association for Computational Linguistics, USA (2012)
19. Stajner, S., Saggion, H., Ponzetto, S.: Improving lexical coverage of text simplification systems for spanish. Expert Systems with Applications **118**, 80–91 (08 2019). https://doi.org/10.1016/j.eswa.2018.08.034
20. Swain, D., Tambe, M., Ballal, P., Dolase, V., Agrawal, K., Rajmane, Y.: Lexical text simplification using wordnet. In: Singh, M., Gupta, P., Tyagi, V., Flusser, J., Ören, T., Kashyap, R. (eds.) Advances in Computing and Data Sciences. pp. 114–122. Springer Singapore, Singapore (2019)
21. Yimam, S.M., Biemann, C., Malmasi, S., Paetzold, G., Specia, L., Štajner, S., Tack, A., Zampieri, M.: 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. pp. 66–78. Association for Computational Linguistics, New Orleans, Louisiana (06 2018). https://doi.org/10.18653/v1/W18-0507, https://www.aclweb.org/anthology/W18-0507