Flesch-Kincaid is Not a Text Simplification Evaluation Metric

Teerapaun Tanprasert
Pomona College
Claremont, CA
teerapaun.tanprasert@pomona.edu

David Kauchak
Pomona College
Claremont, CA
david.kauchak@pomona.edu

Abstract

Sentence-level text simplification is evaluated using both automated metrics and human evaluation. For automatic evaluation, a combination of metrics is usually employed to evaluate different aspects of the simplification. Flesch-Kincaid Grade Level (FKGL) is one metric that has been regularly used to measure the readability of system output. In this paper, we argue that FKGL should not be used to evaluate text simplification systems. We provide experimental analyses on recent system output showing that the FKGL score can easily be manipulated to improve the score dramatically with only minor impact on other automated metrics (BLEU and SARI). Instead of using FKGL, we suggest that the component statistics, along with others, be used for posthoc analysis to understand system behavior.

1 Introduction

Critical to any application area is evaluation. Evaluation is often accomplished using one or more quantifiable evaluation metrics. Evaluation metrics are the main tool for comparing and analyzing approaches (Hossin and Sulaiman, 2015) and are often used to define whether progress is being made in a field. A good evaluation metric should be a proper measure of the quality of a particular algorithm and, importantly, should not be "gameable". Specifically, an approach should not be able to obtain a better score on the evaluation metric by manipulating the algorithm or output in ways that do not improve the actual quality of the output.

In this paper, we examine evaluation for text simplification, specifically, sentence-level text simplification. Text simplification aims to transform text into a variant that is easier to understand by a broader range of people while retaining as much of the original content as possible. A range of approaches for text simplification have been proposed ranging from lexical simplification (Shardlow, 2014), where only words and phrases are changed, to fully generative approaches that leverage models from machine translation (Coster and Kauchak, 2011a; Wubben et al., 2012) and recent sequential neural networks (Nisioi et al., 2017; Zhang and Lapata, 2017; Nishihara et al., 2019). Text simplification evaluation has been done with two general approaches: human evaluation and automated metrics.

Human evaluation relies on annotators to judge the quality of the simplifications on three dimensions: fluency/grammaticality, how well the sentence represents fluent, grammatical text; adequacy, how well the content is preserved; and, simplicity, how simple the text is (Woodsend and Lapata, 2011). The first two metrics were adapted from other text generation tasks (Knight and Marcu, 2002) with the addition of simplicity for text simplification. When human evaluation is used, these three metrics have been consistently employed. Human evaluations provide concrete analysis of texts simplification systems along important dimensions, however, human evaluation is costly and is not practical for development, tuning, and other real-time uses. As such, text simplification has also relied on automated metrics for evaluation.

Automatic evaluation of text simplification has varied more across papers, though three metrics are most commonly employed: BLEU, SARI, and Flesch-Kincaid. BLEU (Papineni et al., 2001) compares the n-gram overlap via precision of a system simplification with a human reference simplification and was borrowed from machine translation. BLEU was the first metric suggested for text simplification that utilized reference simplifications (Zhu et al., 2010), however, it focuses less on simplicity and more on fluency and content preservation. To counter this, SARI was proposed as an alternate metric (Xu et al., 2016). SARI also compares

Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021), pages 1–14
August 5–6, 2021. ©2021 Association for Computational Linguistics

against human references, but also utilizes the input sentence allowing it to better capture addition and deletion of information.

Finally, a third automated metric that has been used to measure readability and fluency is Flesch-Kincaid Grade Level (FKGL). FKGL was initially proposed in the 1940s (Flesch, 1948) and since then has been used extensively in the medical domain, though it has never been shown to affect actual comprehension (Shardlow, 2014; Kauchak and Leroy, 2016). FKGL combines two text statistics to calculate the score: the average number of syllables per word and the average number of words per sentence:

$$FKGL = 0.39\frac{N_{words}}{N_{sentences}} + 11.8\frac{N_{syllables}}{N_{words}} - 15.59 \quad (1)$$

In recent text simplification papers, both BLEU and SARI are common evaluation metrics (Vu et al., 2018; Guo et al., 2018; Scarton and Specia, 2018; Qiang, 2018; Niklaus et al., 2019; Nishihara et al., 2019). FKGL is not as popular as it was before SARI was introduced, but it continues to be used as an evaluation metric in recent papers (Xu et al., 2016; Zhang and Lapata, 2017; Guo et al., 2018; Qiang, 2018; Scarton and Specia, 2018; Nassar et al., 2019; Nishihara et al., 2019).

In this paper, we argue that FKGL is not a proper evaluation metric for text simplification and should not be used to evaluate text simplification systems, i.e., alongside other metrics like BLEU and SARI. FKGL was one of the first metrics suggested for text simplification (Zhu et al., 2010) and has been used by many as an evaluation metric to compare systems. However, FKGL was not originally designed to evaluate system output (it was designed to measure human output) and, because of its simplistic nature, is very easy to game, either explicitly (as we do in this paper) or implicitly by certain model biases (e.g., text simplification algorithms that split sentences will tend to have better FKGL scores). Recent work has shown that systems with good FKGL scores are not necessarily correlated with high-quality simplifications (Martin et al., 2018; Alva-Manchego et al., 2020), however, this is the first in-depth analysis of the FKGL metric for evaluation and where specific system transformations are analyzed.

To explore how FKGL can be manipulated, we introduce six simple methods for modifying system output and examine the impact these modifications have on automated evaluation metrics. The modifications could be made explicitly by a system in an attempt to improve their score, or, more worrisome, implicitly. In addition to the FKGL scores, we also present and and discuss how BLEU and SARI respond to the modifications. We show that with some very minor modifications, FKGL can be improved dramatically with minimal effect on the other two evaluation metrics. We conclude with some recommendations on how to incorporate FKGL-like metrics into text simplification analysis.

2 History of Flesch-Kincaid

The earliest version of the Flesch-Kincaid readability formula appears in Flesch's doctoral dissertation (Flesch, 1943) and calculated based on the the average number of words per sentence, the number of affixes, and the number of references to people. The formula was derived based on the McCall-Crabbs Standard Test Lessons in Reading (McCall and Crabbs, 1926), a standardized test given to children in grades 3-7. The McCall-Crabbs tests contains 376 passages with 8 reading comprehensive questions per passage. Each lesson is labeled with its difficulty as a grade level. Based on these texts, Flesch developed the formula to predict the grade of children in grades 3-7 who answered at least 75% of the questions correctly about a given passage. The original goal of the formula was to help students track their progress.

Five years later, he published a new formula: the Reading Ease Score (Flesch, 1948). He adjusted the original formula by recomputing the coefficients and replacing previous text measurements with the ones used today, the average number of syllables and the average sentences length. Like the original study, this new formula was validated with children and was based on the same criterion, McCall-Crabbs Standard Test Lessons in Reading.

Flesch-Kincaid Grade Level is a variation of the Reading Ease formula with readjusted weights and is the formula that has been commonly used in text simplification evaluation. The formula was derived three decades later (Kincaid et al., 1975) specifically to evaluate the readability of technical materials for military personnel. 531 Navy personnel in four technical training schools at Navy bases were tested for their reading comprehension level according to the comprehension section of the Gates-McGinitie reading test as well as their comprehension of 18 passages from Rate Training

Manuals. Despite the fact that this formula was derived from Navy personnel, with military-based material, and specifically for Navy use, it has been broadly used in a range of settings to evaluate the readability of text, for example, it is commonly used to guide text generation by medical writers in the medical domain and even Microsoft Word includes both the Flesch Reading Ease and FKGL scores (Shedlosky-Shoemaker et al., 2009).

We provide this background to raise some concerns based on its origins for its application for text simplification evaluation. The inputs of the formula – sentence count, word count, and syllable count – were decided based on a study in the 1940s where modern text analysis tools were not available. Both the Flesch Reading Ease and FKGL scores were developed based on very specific corpora and very targeted populations, children grades 3-7 in the former case and Navy personnel in the latter case. Most importantly, the text passages used to collect data were always written by people and assumed to be mostly free of errors in terms of writing. These assumption cannot be made for text generated by automated systems.

3 Modifying Text Simplification Output

One of the main drawbacks of the FKGL metric is that the formula is based on fairly simplistic text statistics. Because of this, it is straightforward to manipulate the output of a text simplification to artificially improve the FKGL score. We suggest six approaches to modify the output of an automatically simplified text that aim to manipulate these statistics. We view the modifications as an explicit post-processing step, however, many of them could be incorporated into a text simplification system either explicitly as a way to improve the score, or implicitly as a side-effect of the algorithm used (e.g., sentence splitting). Each approach suggested modifies the output text on a sentence level. In the analyses we consider the effect of applying each approach to varying proportions of the sentences output by the system.

random-period: Randomly insert a period into the sentence. Adding a period to the sentence splits the sentence into two sentences which reduces the average number of words per sentence.

random-the: Randomly insert the word "the" into the sentence. This adds a short and very common word to reduce the average syllable count per word while minimizing the impact on the meaning.

replace-longest: Replace the longest word in the sentence (by character count) with the word "the". Assuming that the number of characters in a word positively correlates with the number of syllables, replacing the longest word with "the" should reduce the average syllable count per word.

replace-rand-period: Replace a random word with a period in the sentence. This is similar to *random-period*, but additionally removes a random word to reduce the number of words per sentence.

replace-rand-the: Replace a random word with "the": imitates *random-the*., but additionally removes a random word to reduce the number of words per sentence.

rand-period+ repl-longest: combine *random-period* and *replace-longest* to magnify the effects on FKGL.

4 Data

To understand the problems with FKGL, we analyzed the output from the five text simplification systems examined by Zhang and Lapata (2017), a number of which are state-of-the-art: PBMT-R (Wubben et al., 2012), a phrase-based approach based on statistical MT; Hybrid (Narayan and Gardent, 2014), a model that combines sentence splitting and deletion with PBMT-R; EncDecA, a basic neural encoder-decoder model with attention; and two deep reinforcement learning models, Dress and Dress-Ls (Zhang and Lapata, 2017).

There are two main corpora that are used to train and evaluate text simplification systems: Wikipedia (Zhu et al., 2010; Coster and Kauchak, 2011b), which consists of automatically aligned sentences between English Wikipedia and Simple English Wikipedia, and Newsela (Xu et al., 2015), which consists of news articles manually simplified at varying levels of simplicity. We present the results for the Newsela corpus since it involves explicit human simplification and has been shown to be less noisy than the Wikipedia corpus (Xu et al., 2015). We also conducted the experimental analysis on the Wikipedia corpus and saw similar results.

5 Experimental Analysis

We applied each of the modification techniques to a varied percentage of output sentences, from 10% to 100% in increments of 10%, for the five text simplification systems. The sentences to be modified were randomly selected from the system output.

We calculated FKGL[1] as well as BLEU (Papineni et al., 2001) and SARI[2] (Xu et al., 2016) to observe how the modifications affect other common text simplification evaluation metrics. To account for per-sentence variation and randomness in some of the modification approaches, we repeated the experiments 100 times and averaged the results.

5.1 Results

Figure 1 shows the trends of the effect that the modification approaches have on FKGL for Dress-Ls, and Table 1 presents more detailed experimental results for the three best performing systems (Dress-Ls, EncDecA, and Hybrid). The three methods that involve sentence splitting result in aggressive improvements in the FKGL score; replacing the longest word shows some improvement; and the other two approaches involving "the" have minimal effect. In the most extreme case, *rand-period+ repl-longest* reduces the FKGL score to almost zero when applied to all of the sentences. *With simple post-processing applied to the output, a text simplification approach can achieve an arbitrarily low FKGL score.*

Figures 2 and 3 show the effect that the modification approaches have on the BLEU and SARI scores for Dress-Ls. There is virtually no effect on the SARI scores by any of the modification techniques and none of the approaches change the score by more than 0.004, regardless of percentage of sentences modified. BLEU, on the other hand, does register some differences for the modified output. *rand-period+ repl-longest* has the most drastic effect and, in the most extreme case, for Dress-Ls it reduces the BLEU score from 0.2374 to 0.1710 when it is applied to all sentences. The other five modification techniques have more minor effects, e.g., *random-period* drops the score to 0.1953, when applied to all sentences.

Using multiple evaluation metrics partially mitigates the gameability of FKGL since BLEU is affected. However, the effect on BLEU is significantly smaller than the effect on FKGL. While the Dress-Ls system did originally have the highest BLEU and SARI scores, it did not have the highest FKGL score. However, if we randomly inserted a period into just 10% of the sentences of the Dress-Ls output, the FKGL score would improve to 4.543, the BLEU score would drop slightly to 0.233 and there is no significant change in SARI score. After the transformation, the system would still be the best performing model with respect to BLEU and SARI, but now it would also be the best performing model with respect to FKGL. *With a*

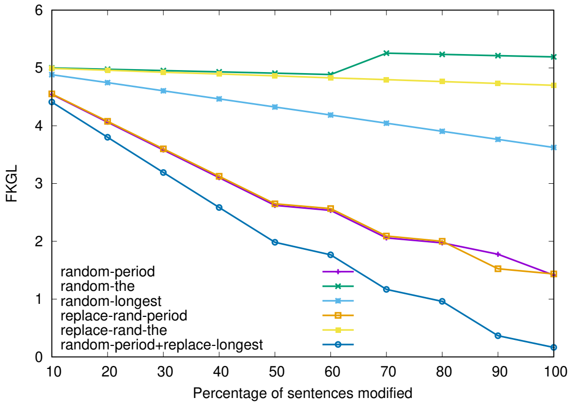

Figure 1: FKGL scores (smaller is better) from the experiments on the Dress-Ls test output, averaged over 100 runs.

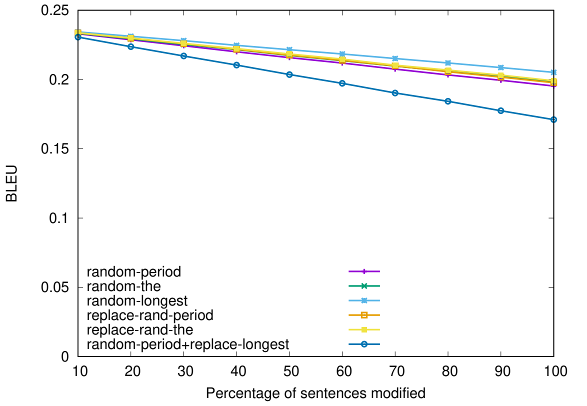

Figure 2: BLEU scores (larger is better) from the experiments on the Dress-Ls test output, averaged over 100 runs.

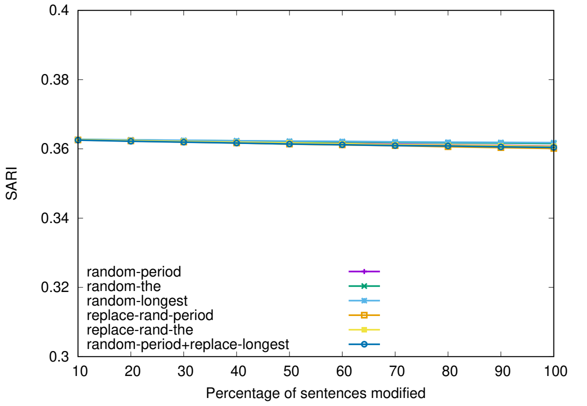

Figure 3: SARI scores (larger is better) from the experiments on the Dress-Ls test output, averaged over 100 runs.

[1] https://github.com/mmautner/readability

[2] We used the implementation for BLEU and SARI from the Joshua Simplification System.

FKGL	**Dress-Ls**				**EncDecA**				**Hybrid**			
Approach	**0%**	**10%**	**50%**	**100%**	**0%**	**10%**	**50%**	**100%**	**0%**	**10%**	**50%**	**100%**
random-period	5.024	4.5426	2.6223	1.4154	5.757	5.2902	3.4309	1.9016	4.775	4.2706	2.6543	1.3512
random-the		5.0006	4.9095	5.1919		6.1273	6.0509	5.9596		4.7434	4.6204	4.8678
replace-longest		4.8837	4.3242	3.6244		5.6408	5.1763	4.5984		4.6108	3.9492	3.1241
replace-rand-period		4.5510	2.6494	1.4359		5.2959	3.4474	1.9173		4.2884	2.7283	1.4524
replace-rand-the		4.9915	4.8636	4.7003		5.8014	5.8058	5.8001		4.7282	4.5449	4.3104
rand-period+ repl-longest		4.4098	1.9831	0.1643		5.1806	2.8913	0.8477		4.1234	1.9268	-0.0665

BLEU	**Dress-Ls**				**EncDecA**				**Hybrid**			
Approach	**0%**	**10%**	**50%**	**100%**	**0%**	**10%**	**50%**	**100%**	**0%**	**10%**	**50%**	**100%**
random-period	0.237	0.2330	0.2158	0.1953	0.212	0.2086	0.1954	0.1794	0.108	0.1069	0.1004	0.0898
random-the		0.2334	0.2174	0.1985		0.2088	0.1963	0.1814		0.1071	0.1015	0.0919
replace-longest		0.2343	0.2215	0.2052		0.2097	0.2008	0.1895		0.1069	0.1016	0.0948
replace-rand-period		0.2336	0.2176	0.1977		0.2088	0.1965	0.1808		0.1063	0.0984	0.0883
replace-rand-the		0.2337	0.2184	0.1991		0.2088	0.1965	0.1808		0.1063	0.0984	0.0879
rand-period+ repl-longest		0.2306	0.2036	0.1710		0.2067	0.1871	0.1621		0.1059	0.0957	0.0806

SARI	**Dress-Ls**				**EncDecA**				**Hybrid**			
Approach	**0%**	**10%**	**50%**	**100%**	**0%**	**10%**	**50%**	**100%**	**0%**	**10%**	**50%**	**100%**
random-period	0.363	0.3626	0.3618	0.3608	0.360	0.3598	0.3593	0.3586	0.347	0.3470	0.3468	0.3465
random-the		0.3627	0.3621	0.3616		0.3599	0.3596	0.3593		0.3471	0.3471	0.3473
replace-longest		0.3627	0.3622	0.3618		0.3600	0.3598	0.3597		0.3471	0.3472	0.3474
replace-rand-period		0.3626	0.3614	0.3601		0.3598	0.3590	0.3579		0.3470	0.3466	0.3462
replace-rand-the		0.3626	0.3617	0.3607		0.3599	0.3593	0.3586		0.3470	0.3469	0.3468
rand-period+ repl-longest		0.3625	0.3614	0.3604		0.3598	0.3591	0.3587		0.3470	0.3471	0.3471

Table 1: Experimental results (FKGL, BLEU and SARI scores) for 10%, 50% and 100% of the sentences being modified on three systems: Dress-Ls, EncDecA and Hybrid.

simple modification to the system output, the best performing model could be changed with respect to FKGL without affecting the other two metrics significantly.

For the sake of brevity, we only include detailed experimental analysis of the output of Dress-Ls, however, the results were similar across all systems[3]. To provide some additional examples, Table 1 shows the FKGL, BLEU, and SARI scores for Dress-Ls, EncDecA, and Hybrid where 10%, 50%, and 100% of the sentences were modified. We chose EncDecA and Hybrid as additional systems to include since they performed well on at least one of the automated metrics and represent fairly different approaches to the text simplification problem. The trends seen for Dress-Ls are also seen with the other two systems: FKGL can be aggressively improved, BLEU is slightly impacted, and SARI is not affected. Regardless of the type of system, because of the simplicity of FKGL, the results can be arbitrarily improved.

5.2 Understanding BLEU and SARI

Although the focus of this paper was on FKGL, we also analyzed BLEU and SARI further to understand why the modification approaches affected those metrics. The BLEU score is calculated as the average of the n-gram precisions of size 1 to 4, where precision is the proportion of n-grams in the system output that are found in the corresponding reference simplification. The SARI score is an average of F1 scores based on three operations relative to the reference text: added n-grams, kept n-grams, and deleted n-grams.

Table 2 shows each of the individual component calculations for the Dress-Ls system when the six modifications are applied to 100% of the sentences. Since the approaches rely on randomization, the results shown are an average of 100 trials. For conciseness, we only include the results for Dress-Ls, though all systems showed very similar trends. Full results, including 2-gram and 3-gram F1 and precision scores for SARI, for all systems are provided in the appendix.

For BLEU, all levels of precision drop for all three modification approaches. The 1-gram precision is the least affected, while larger n-gram

[3] Complete experimental results are included in the appendix.

Percent Modified	0	100					
Approach	**none**	**random-period**	**random-the**	**replace-longest**	**replace-rand-period**	**replace-rand-the**	**rand-period+ repl-longest**
BLEU							
1-gram	0.4590	0.4300	0.4394	0.4468	0.4340	0.4428	0.4186
2-gram	0.2638	0.2289	0.2301	0.2339	0.2289	0.2276	0.2026
3-gram	0.1896	0.1496	0.1509	0.1581	0.1511	0.1497	0.1249
4-gram	0.1384	0.0997	0.1003	0.1074	0.1016	0.1003	0.0763
SARI							
1-gram							
Add F1	0.0382	0.0382	0.0518	0.0505	0.0371	0.0504	0.0505
Keep F1	0.1181	0.1181	0.1169	0.1181	0.1186	0.1174	0.1181
Delete P	0.9740	0.9740	0.9741	0.9722	0.9718	0.9717	0.9722
4-gram							
Add F1	0.0189	0.0155	0.0145	0.0150	0.0154	0.0143	0.0112
Keep F1	0.0450	0.0446	0.0450	0.0463	0.0448	0.0447	0.0455
Delete P	0.9885	0.9876	0.9879	0.9878	0.9874	0.9874	0.9869

Table 2: Breakdown of the components making up BLEU and SARI scores for the original Dress-Ls output and the modified texts.

precisions show increasingly larger effects. This intuitively makes sense since randomly inserting/replacing a word in an originally correct sequence of words should affect multiple n-grams of larger size. None of the decreases are large in magnitude, but they are all in the same direction and contribute to the slight drop in BLEU scores.

For SARI, at the 1-gram level, the Add F1 score actually improves for both *random-the* and *replace-longest* since they add a common word ("the") that has a high likelihood of matching with a word in the reference simplification. However, for longer n-grams the Add F1 score drops for similar reasons to the BLEU score precisions drop. Besides the Add F1 score, however, the other scores remain virtually unchanged. In aggregate, the Add effect tends to balance out between increases in smaller n-grams and decreases in larger n-grams and because the other components do not change much, the overall SARI score remains unaffected.

The effects of the modifications on BLEU and SARI are minimal, especially compared to the effects on FKGL. While this helps illustrate how a manipulation of FKGL could be done, it does not necessarily imply that BLEU and SARI are sufficiently reliable. Even though both metrics are relatively resilient against our modification approaches, these approaches were designed specifically to manipulate the FKGL score and, thus, do not serve as evidence against the concerns that have been raised about their robustness (Callison-Burch et al., 2006; Sulem et al., 2018).

System	**Average length**	**Average syllables**	**% split**
Original	23.08	1.346	0
Reference	12.741	1.263	1.857
Dress-Ls	14.392	1.284	1.207
EncDecA	16.986	1.280	0.557
Hybrid	12.382	1.329	0.000
Dress	14.222	1.276	1.207
PBMT-R	22.933	1.304	1.300

Table 3: Post-hoc statistics for original and reference data from the test corpus and five system outputs.

6 A Better Approach

FKGL should not be used as an evaluation metric. Instead, it can be used for post-hoc analysis to understand the behavior of the systems. Even better, rather than reporting the FKGL score, which can be affected by multiple types of changes in the system, papers can report the individual components of FGKL, i.e., the average sentence length and the average number of syllables. This demystifies the readability score and provides concrete information about the types of changes that are being made by the systems. A comparative analysis of 30 metrics showed that these features are better correlated with human judgement than FKGL (Martin et al., 2018), and some recent papers have reported the average sentence length statistic already (Kriz et al., 2019; Kumar et al., 2020; Maddela et al., 2021). These two metrics can be supplemented with other corpus statistics that also help understand what changes the systems are making, e.g., the proportion of sentences that are split.

Table 3 shows these three statistics for the five text simplification approaches. These statistics allow for a concrete analysis of what the different approaches are doing. All the models reduce the sentence length, except for PBMT-R. Hybrid is the most aggressive at creating short sentences, though it does not do any sentence splitting, so it accomplishes this through deletion, which may explain the low BLEU score. All of the models are selecting words with less syllables, except for Hybrid. Finally, all models except Hybrid are doing sentence splitting, with the EncDecA doing the least splitting. These statistics paint a much more vivid picture of what the different approach are doing than a single readability score.

7 Conclusions

In this paper, we have provided an experimental analysis of the FKGL score on state-of-the-art text simplification systems. We find that very basic post-processing techniques can drastically improve the FKGL score of a system with negligible effects on two other metrics, BLEU and SARI. Based on these findings, we argue that FKGL should no longer be used as a text simplification evaluation metric. Instead, the components of FKGL and other related statistics should be used to help understand what different systems are doing. If this analysis is not compelling enough and FKGL continues to be used, then we propose concrete methods for improving FKGL, with minimal work and only minor effects on the other automated metrics.

References

Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. 2020. Data-driven sentence simplification: Survey and benchmark. *Computational Linguistics*, 46(1):135–187.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluation the role of bleu in machine translation research. In *Proceedings of European Association for Computational Linguistics*.

William Coster and David Kauchak. 2011a. Learning to simplify sentences using wikipedia. In *Proceedings of the workshop on monolingual text-to-text generation*.

William Coster and David Kauchak. 2011b. Simple english wikipedia: a new text simplification task. In *Proceedings of Assication for Computational Linguistics*.

Rudolf Flesch. 1943. Marks of readable style; a study in adult education. *Teachers College Contributions to Education*.

Rudolph Flesch. 1948. A new readability yardstick. *Journal of applied psychology*, 32(3):221.

Han Guo, Ramakanth Pasunuru, and Mohit Bansal. 2018. Dynamic multi-level multi-task learning for sentence simplification. In *Proceedings of International Conference on Computational Linguistics*, pages 462–476.

Mohammad Hossin and MN Sulaiman. 2015. A review on evaluation metrics for data classification evaluations. *International Journal of Data Mining & Knowledge Management Process*.

David Kauchak and Gondy Leroy. 2016. Moving beyond readability metrics for health-related text simplification. *IT professional*, 18(3):45–51.

J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel.

Kevin Knight and Daniel Marcu. 2002. Summarization beyond sentence extraction: A probabilistic approach to sentence compression. *Artificial Intelligence*, 139(1):91–107.

Reno Kriz, Joao Sedoc, Marianna Apidianaki, Carolina Zheng, Gaurav Kumar, Eleni Miltsakaki, and Chris Callison-Burch. 2019. Complexity-weighted loss and diverse reranking for sentence simplification. In *Proceedings of NAACL-HLT*, pages 3137–3147.

Dhruv Kumar, Lili Mou, Lukasz Golab, and Olga Vechtomova. 2020. Iterative edit-based unsupervised sentence simplification. In *Proceedings of Association for Computational Linguistics*, pages 7918–7928.

Mounica Maddela, Fernando Alva-Manchego, and Wei Xu. 2021. Controllable text simplification with explicit paraphrasing. In *NAACL-HLT*, Online. Association for Computational Linguistics.

Louis Martin, Samuel Humeau, Pierre-Emmanuel Mazare, Éric Villemonte de la Clergerie, Antoine Bordes, and Benoît Sagot. 2018. Reference-less quality estimation of text simplification systems. In *Proceedings of the Workshop on Automatic Text Adaptation*, pages 29–38.

William Anderson McCall and Lelah Mae Crabbs. 1926. *Standard Test Lessons in Reading...* 5. Teachers College, Columbia University, Bureau of Publications.

Shashi Narayan and Claire Gardent. 2014. Hybrid simplification using deep semantics and machine translation. In *Proceedings Association for Computational Linguistics*.

Islam Nassar, Michelle Ananda-Rajah, and Gholamreza Haffari. 2019. Neural versus non-neural text simplification: A case study. In *Proceedings of the Workshop of the Australasian Language Technology Association*, pages 172–177.

Christina Niklaus, Matthias Cetto, André Freitas, and Siegfried Handschuh. 2019. Transforming complex sentences into a semantic hierarchy. In *Proceedings of Association for Computational Linguistics*, pages 3415–3427.

Daiki Nishihara, Tomoyuki Kajiwara, and Yuki Arase. 2019. Controllable text simplification with lexical constraint loss. In *Proceedings of Association for Computational Linguistics: Student Research Workshop*, pages 260–266.

Sergiu Nisioi, Sanja Štajner, Simone Paolo Ponzetto, and Liviu P Dinu. 2017. Exploring neural text simplification models. In *Proceedings of Association for Computational Linguistics*, pages 85–91.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. BLEU. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL*. Association for Computational Linguistics.

Jipeng Qiang. 2018. Improving neural text simplification model with simplified corpora. *CoRR*, abs/1810.04428.

Carolina Scarton and Lucia Specia. 2018. Learning simplifications for specific target audiences. In *Proceedings of Association for Computational Linguistics*, pages 712–718.

Matthew Shardlow. 2014. A survey of automated text simplification. *International Journal of Advanced Computer Science and Applications*.

Randi Shedlosky-Shoemaker, Amy Curry Sturm, Muniba Saleem, and Kimberly M Kelly. 2009. Tools for assessing readability and quality of health-related web sites. *Journal of genetic counseling*, 18(1):49.

Elior Sulem, Omri Abend, and Ari Rappoport. 2018. Bleu is not suitable for the evaluation of text simplification. In *Proceedings of Empirical Methods in Natural Language Processing*, pages 738–744.

Tu Vu, Baotian Hu, Tsendsuren Munkhdalai, and Hong Yu. 2018. Sentence simplification with memory-augmented neural networks. In *Proceedings of North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 79–85.

Kristian Woodsend and Mirella Lapata. 2011. Learning to simplify sentences with quasi-synchronous grammar and integer programming. In *Proceedings of Empirical Methods in Natural Language Processing*, pages 409–420.

Sander Wubben, Antal van den Bosch, and Emiel Krahmer. 2012. Sentence simplification by monolingual machine translation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1015–1024.

Wei Xu, Chris Callison-Burch, and Courtney Napoles. 2015. Problems in current text simplification research: New data can help. *Transactions of the Association of Computational Linguistics*.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.

Xingxing Zhang and Mirella Lapata. 2017. Sentence simplification with deep reinforcement learning. In *Proceedings of Empirical Methods in Natural Language Processing*, pages 584–594.

Zhemin Zhu, Delphine Bernhard, and Iryna Gurevych. 2010. A monolingual tree-based translation model for sentence simplification. In *Proceedings of ICCL*.

Appendix

A Experimental Results for All Systems

Tables 4-8 show the complete FKGL, BLEU and SARI scores for the modified outputs of all five systems: Dress-Ls, EncDecA, Hybrid, Dress and PBMT-R.

B BLEU n-gram Score Breakdown

Table 9 shows the precision scores for the individual n-grams (1-4) of the unmodified system output and output with all sentences modified (100%) for each of the six modification approaches on outputs of all five systems.

C SARI n-gram Score Breakdown

Table 10 shows the SARI component scores for the unmodified system output and with all sentences modified (100%) for each of the six modification approaches on all five systems.

Dress-Ls										
Approach/ % modified	10	20	30	40	50	60	70	80	90	100
FKGL										
random-period	4.5426	4.0609	3.5802	3.1014	2.6223	2.5358	2.0595	1.9742	1.7763	1.4154
random-the	5.0006	4.9772	4.9543	4.9319	4.9095	4.8870	5.2557	5.2346	5.2130	5.1919
replace-longest	4.8837	4.7464	4.6050	4.4644	4.3242	4.1857	4.0442	3.9038	3.7647	3.6244
replace-rand-period	4.5510	4.0765	3.6007	3.1251	2.6494	2.5670	2.0910	2.0005	1.5256	1.4359
replace-rand-the	4.9915	4.9607	4.9259	4.8955	4.8636	4.8288	4.7985	4.7638	4.7324	4.7003
random-period +replace-longest	4.4098	3.8000	3.1911	2.5864	1.9831	1.7665	1.1681	0.9604	0.3671	0.1643
SARI										
random-period	0.3626	0.3624	0.3622	0.3619	0.3618	0.3616	0.3613	0.3612	0.3610	0.3608
random-the	0.3627	0.3626	0.3624	0.3623	0.3621	0.3620	0.3619	0.3618	0.3617	0.3616
replace-longest	0.3627	0.3626	0.3625	0.3623	0.3622	0.3622	0.3620	0.3620	0.3619	0.3618
replace-rand-period	0.3626	0.3623	0.3620	0.3617	0.3614	0.3612	0.3609	0.3606	0.3604	0.3601
replace-rand-the	0.3626	0.3624	0.3622	0.3619	0.3617	0.3615	0.3612	0.3611	0.3609	0.3607
random-period +replace-longest	0.3625	0.3622	0.3619	0.3617	0.3614	0.3612	0.3609	0.3608	0.3606	0.3604
BLEU										
random-period	0.2330	0.2287	0.2243	0.2200	0.2158	0.2119	0.2075	0.2033	0.1994	0.1953
random-the	0.2334	0.2293	0.2253	0.2216	0.2174	0.2136	0.2097	0.2059	0.2022	0.1985
replace-longest	0.2343	0.2312	0.2281	0.2247	0.2215	0.2184	0.2151	0.2120	0.2086	0.2052
replace-rand-period	0.2336	0.2297	0.2258	0.2218	0.2176	0.2138	0.2099	0.2057	0.2017	0.1977
replace-rand-the	0.2337	0.2300	0.2261	0.2224	0.2184	0.2148	0.2104	0.2068	0.2032	0.1991
random-period +replace-longest	0.2306	0.2237	0.2170	0.2104	0.2036	0.1972	0.1903	0.1843	0.1775	0.1710

Table 4: Metric scores of 10-100% modified outputs of Dress-LS

EncDecA										
Approach/ % modified	10	20	30	40	50	60	70	80	90	100
FKGL										
random-period	5.2905	4.8237	4.3576	3.8938	3.4304	2.9668	2.8942	2.4334	2.3618	1.9012
random-the	6.1272	6.1077	6.0884	6.0696	6.0507	6.0319	6.0138	5.9956	5.9777	5.9600
replace-longest	5.6413	5.5258	5.4092	5.2943	5.1792	5.0623	4.9459	4.8306	4.7143	4.5984
replace-rand-period	5.2958	4.8351	4.3718	3.9104	3.4496	2.9878	2.9127	2.4525	2.3804	1.9200
replace-rand-the	5.8045	5.8418	5.7950	5.7942	5.8163	5.8146	5.8060	5.7801	5.7838	5.7498
random-period +replace-longest	5.1811	4.6057	4.0323	3.4621	2.8906	2.3232	2.1496	1.5827	1.4098	0.8463
SARI										
random-period	0.3598	0.3597	0.3595	0.3594	0.3592	0.3591	0.3590	0.3588	0.3587	0.3586
random-the	0.3599	0.3598	0.3597	0.3596	0.3596	0.3595	0.3595	0.3594	0.3593	0.3593
replace-longest	0.3600	0.3599	0.3599	0.3598	0.3598	0.3597	0.3598	0.3597	0.3597	0.3597
replace-rand-period	0.3598	0.3596	0.3593	0.3591	0.3590	0.3587	0.3585	0.3583	0.3582	0.3580
replace-rand-the	0.3599	0.3597	0.3596	0.3594	0.3593	0.3591	0.3589	0.3588	0.3587	0.3585
random-period +replace-longest	0.3598	0.3597	0.3595	0.3593	0.3592	0.3590	0.3590	0.3589	0.3588	0.3587
BLEU										
random-period	0.2085	0.2052	0.2019	0.1987	0.1954	0.1921	0.1891	0.1858	0.1827	0.1796
random-the	0.2087	0.2056	0.2024	0.1994	0.1964	0.1935	0.1905	0.1875	0.1844	0.1815
replace-longest	0.2098	0.2075	0.2053	0.2031	0.2007	0.1986	0.1964	0.1941	0.1918	0.1895
replace-rand-period	0.2088	0.2058	0.2025	0.1994	0.1966	0.1932	0.1903	0.1871	0.1841	0.1808
replace-rand-the	0.2089	0.2057	0.2027	0.1995	0.1965	0.1933	0.1900	0.1870	0.1837	0.1805
random-period +replace-longest	0.2068	0.2019	0.1967	0.1917	0.1869	0.1817	0.1768	0.1721	0.1670	0.1621

Table 5: Metric scores of 10-100% modified outputs of EncDecA

Hybrid										
Approach/ % modified	10	20	30	40	50	60	70	80	90	100
FKGL										
random-period	4.2706	3.7659	3.2634	3.1522	2.6543	2.5450	2.0501	1.9458	1.4523	1.3512
random-the	4.7434	4.7118	4.6808	4.6503	4.6204	4.5907	4.9520	4.9236	4.8950	4.8678
replace-longest	4.6108	4.4474	4.2792	4.1167	3.9492	3.7866	3.6211	3.4546	3.2903	3.1241
replace-rand-period	4.2884	3.7986	3.3114	3.2161	2.7283	2.5598	2.1379	2.0364	1.5465	1.4524
replace-rand-the	4.7282	4.6833	4.6363	4.5903	4.5449	4.4997	4.4539	4.4070	4.3613	4.3104
random-period +replace-longest	4.1234	3.4704	2.8198	2.5699	1.9268	1.6807	1.0422	0.7992	0.1686	-0.0665
SARI										
random-period	0.3470	0.3469	0.3469	0.3468	0.3468	0.3467	0.3467	0.3466	0.3466	0.3465
random-the	0.3471	0.3471	0.3471	0.3471	0.3471	0.3472	0.3472	0.3472	0.3472	0.3473
replace-longest	0.3471	0.3471	0.3472	0.3472	0.3472	0.3473	0.3473	0.3473	0.3474	0.3474
replace-rand-period	0.3470	0.3469	0.3468	0.3467	0.3466	0.3466	0.3465	0.3464	0.3463	0.3462
replace-rand-the	0.3470	0.3470	0.3470	0.3470	0.3469	0.3469	0.3469	0.3469	0.3468	0.3468
random-period +replace-longest	0.3470	0.3470	0.3471	0.3471	0.3471	0.3471	0.3471	0.3471	0.3471	0.3471
BLEU										
random-period	0.1069	0.1054	0.1042	0.1026	0.1004	0.0981	0.0959	0.0939	0.0917	0.0898
random-the	0.1071	0.1059	0.1047	0.1033	0.1015	0.0994	0.0975	0.0956	0.0938	0.0919
replace-longest	0.1069	0.1055	0.1043	0.1028	0.1016	0.1002	0.0989	0.0975	0.0962	0.0948
replace-rand-period	0.1063	0.1043	0.1025	0.1006	0.0984	0.0965	0.0944	0.0926	0.0904	0.0883
replace-rand-the	0.1063	0.1042	0.1022	0.1004	0.0984	0.0962	0.0941	0.0921	0.0898	0.0879
random-period +replace-longest	0.1059	0.1036	0.1012	0.0986	0.0957	0.0924	0.0895	0.0866	0.0836	0.0806

Table 6: Metric scores of 10-100% modified outputs of Hybrid

Dress										
Approach/ % modified	10	20	30	40	50	60	70	80	90	100
FKGL										
random-period	4.4416	3.9367	3.4778	2.9987	2.5223	2.4322	1.9566	1.8709	1.3976	1.3138
random-the	4.9011	4.8782	4.8557	4.8336	4.8115	4.7899	4.7686	5.1378	5.1166	5.0966
replace-longest	4.7838	4.6432	4.5021	4.3628	4.2182	4.0781	3.9378	3.7971	3.6582	3.5147
replace-rand-period	4.3679	3.7817	3.4981	3.0221	2.5450	2.4596	1.9872	1.8958	1.4210	1.3311
replace-rand-the	4.8922	4.8580	4.8276	4.7936	4.7614	4.7301	4.6980	4.6648	4.6344	4.5964
random-period +replace-longest	4.3096	3.3390	3.0860	2.4810	1.8723	1.6601	1.0603	0.8498	0.2588	0.0536
SARI										
random-period	0.3621	0.3618	0.3616	0.3614	0.3612	0.3610	0.3608	0.3607	0.3605	0.3603
random-the	0.3622	0.3620	0.3619	0.3617	0.3616	0.3615	0.3614	0.3613	0.3612	0.3611
replace-longest	0.3622	0.3621	0.3620	0.3620	0.3619	0.3618	0.3618	0.3617	0.3617	0.3617
replace-rand-period	0.3620	0.3617	0.3614	0.3612	0.3609	0.3607	0.3605	0.3601	0.3599	0.3597
replace-rand-the	0.3621	0.3619	0.3617	0.3615	0.3613	0.3612	0.3609	0.3608	0.3607	0.3605
random-period +replace-longest	0.3620	0.3618	0.3615	0.3613	0.3612	0.3609	0.3608	0.3606	0.3604	0.3603
BLEU										
random-period	0.2230	0.2187	0.2145	0.2104	0.2062	0.2021	0.1979	0.1941	0.1902	0.1864
random-the	0.2233	0.2193	0.2156	0.2116	0.2078	0.2041	0.2005	0.1969	0.1931	0.1895
replace-longest	0.2243	0.2214	0.2183	0.2156	0.2124	0.2095	0.2066	0.2034	0.2004	0.1974
replace-rand-period	0.2234	0.2196	0.2156	0.2121	0.2080	0.2041	0.2005	0.1964	0.1925	0.1889
replace-rand-the	0.2234	0.2198	0.2158	0.2120	0.2080	0.2043	0.2003	0.1964	0.1926	0.1886
random-period +replace-longest	0.2208	0.2142	0.2078	0.2015	0.1954	0.1887	0.1826	0.1761	0.1700	0.1638

Table 7: Metric scores of 10-100% modified outputs of Dress

PBMT-R										
Approach/ % modified	10	20	30	40	50	60	70	80	90	100
FKGL										
random-period	7.5360	7.0897	6.2541	5.8091	5.3639	4.9187	4.4746	4.4210	3.9773	3.5354
random-the	8.7462	8.7303	8.7147	8.6992	8.6838	8.6684	8.6535	8.6384	8.6233	8.6087
replace-longest	8.2773	8.1807	8.0855	7.9908	7.8946	7.7988	7.7028	7.6075	7.5115	7.4150
replace-rand-period	7.6177	7.0975	6.2632	5.8203	5.3775	4.9330	4.4944	4.3366	3.9970	3.5487
replace-rand-the	8.3526	8.3343	8.3227	8.3330	8.3251	8.2865	8.3119	8.3015	8.3143	8.3201
random-period +replace-longest	7.4441	6.9062	5.9773	5.4449	4.9073	4.3749	3.8442	3.7010	3.1695	2.6411
SARI										
random-period	0.3568	0.3566	0.3565	0.3563	0.3562	0.3560	0.3559	0.3557	0.3556	0.3555
random-the	0.3568	0.3568	0.3567	0.3566	0.3565	0.3565	0.3564	0.3564	0.3563	0.3562
replace-longest	0.3568	0.3566	0.3564	0.3563	0.3562	0.3560	0.3559	0.3558	0.3557	0.3556
replace-rand-period	0.3566	0.3564	0.3561	0.3559	0.3556	0.3554	0.3553	0.3550	0.3548	0.3546
replace-rand-the	0.3567	0.3565	0.3564	0.3561	0.3560	0.3558	0.3557	0.3555	0.3554	0.3553
random-period +replace-longest	0.3566	0.3564	0.3561	0.3558	0.3556	0.3554	0.3552	0.3549	0.3549	0.3546
BLEU										
random-period	0.1751	0.1730	0.1709	0.1689	0.1668	0.1647	0.1628	0.1608	0.1588	0.1567
random-the	0.1752	0.1732	0.1711	0.1692	0.1674	0.1655	0.1637	0.1617	0.1598	0.1580
replace-longest	0.1754	0.1736	0.1718	0.1700	0.1682	0.1664	0.1647	0.1628	0.1611	0.1592
replace-rand-period	0.1751	0.1732	0.1710	0.1691	0.1670	0.1650	0.1631	0.1610	0.1590	0.1571
replace-rand-the	0.1752	0.1732	0.1713	0.1691	0.1673	0.1651	0.1632	0.1611	0.1590	0.1571
random-period +replace-longest	0.1736	0.1701	0.1664	0.1628	0.1593	0.1559	0.1523	0.1487	0.1454	0.1418

Table 8: Metric scores of 10-100% modified outputs of PBMT-R

Percent Modified	0	100					
Approach		random-period	random-the	replace-longest	replace-rand-period	replace-rand-the	random-period+replace-longest
Dress-Ls							
1-gram	0.4590	0.4300	0.4394	0.4468	0.4340	0.4428	0.4186
2-gram	0.2638	0.2289	0.2301	0.2339	0.2289	0.2276	0.2026
3-gram	0.1896	0.1496	0.1509	0.1581	0.1511	0.1497	0.1249
4-gram	0.1384	0.0997	0.1003	0.1074	0.1016	0.1003	0.0763
EncDecA							
1-gram	0.4156	0.4300	0.4394	0.4468	0.4340	0.4428	0.4186
2-gram	0.2373	0.2281	0.2300	0.2339	0.2291	0.2275	0.2037
3-gram	0.1686	0.1495	0.1518	0.1581	0.1516	0.1501	0.1265
4-gram	0.1212	0.0990	0.1014	0.1074	0.1019	0.1005	0.0787
Hybrid							
1-gram	0.3708	0.4300	0.4394	0.4468	0.4339	0.4432	0.4186
2-gram	0.1328	0.2281	0.2298	0.2339	0.2286	0.2275	0.2038
3-gram	0.0710	0.1494	0.1517	0.1581	0.1509	0.1501	0.1268
4-gram	0.0442	0.0991	0.1015	0.1074	0.1012	0.1007	0.0794
Dress							
1-gram	0.4517	0.4300	0.4394	0.4468	0.4336	0.4432	0.4186
2-gram	0.2537	0.2282	0.2299	0.2339	0.2286	0.2281	0.2038
3-gram	0.1800	0.1499	0.1516	0.1581	0.1507	0.1500	0.1266
4-gram	0.1292	0.0998	0.1016	0.1074	0.1010	0.1005	0.0790
PBMT-R							
1-gram	0.3577	0.4300	0.4394	0.4468	0.4340	0.4428	0.4186
2-gram	0.2020	0.2280	0.2299	0.2339	0.2289	0.2274	0.2039
3-gram	0.1392	0.1492	0.1518	0.1581	0.1514	0.1500	0.1270
4-gram	0.0979	0.0990	0.1014	0.1074	0.1016	0.1008	0.0796

Table 9: BLEU score breakdown (1-, 2-, 3- and 4-gram scores) for all combination of systems and modification approaches

Percent Modified	0	100					
Approach		random-period	random-the	replace-longest	replace-rand-period	replace-rand-the	random-period +replace-longest
Dress-Ls							
1-gram							
Add F1	0.0382	0.0382	0.0518	0.0505	0.0371	0.0504	0.0505
Keep F1	0.1181	0.1181	0.1169	0.1181	0.1186	0.1174	0.1181
Delete P	0.9740	0.9740	0.9741	0.9722	0.9718	0.9717	0.9722
2-gram							
Add F1	0.0370	0.0345	0.0322	0.0319	0.0323	0.0311	0.0285
Keep F1	0.0742	0.0739	0.0740	0.0751	0.0736	0.0735	0.0746
Delete P	0.9805	0.9798	0.9800	0.9794	0.9788	0.9787	0.9784
3-gram							
Add F1	0.0263	0.0229	0.0215	0.0215	0.0221	0.0211	0.0173
Keep F1	0.0573	0.0570	0.0573	0.0588	0.0569	0.0569	0.0582
Delete P	0.9850	0.9841	0.9844	0.9843	0.9837	0.9836	0.9832
4-gram							
Add F1	0.0189	0.0155	0.0145	0.0150	0.0154	0.0143	0.0112
Keep F1	0.0450	0.0446	0.0450	0.0463	0.0448	0.0447	0.0455
Delete P	0.9885	0.9876	0.9879	0.9878	0.9874	0.9874	0.9869
EncDecA							
1-gram							
Add F1	0.0382	0.0382	0.0518	0.0505	0.0372	0.0511	0.0505
Keep F1	0.1181	0.1181	0.1169	0.1181	0.1188	0.1174	0.1181
Delete P	0.9740	0.9740	0.9741	0.9722	0.9719	0.9718	0.9722
2-gram							
Add F1	0.0387	0.0343	0.0317	0.0319	0.0333	0.0316	0.0289
Keep F1	0.0744	0.0738	0.0739	0.0751	0.0736	0.0736	0.0748
Delete P	0.9812	0.9798	0.9800	0.9794	0.9788	0.9788	0.9785
3-gram							
Add F1	0.0293	0.0228	0.0217	0.0215	0.0223	0.0215	0.0174
Keep F1	0.0576	0.0570	0.0571	0.0588	0.0570	0.0568	0.0586
Delete P	0.9859	0.9841	0.9843	0.9843	0.9837	0.9836	0.9833
4-gram							
Add F1	0.0219	0.0154	0.0148	0.0150	0.0150	0.0147	0.0113
Keep F1	0.0454	0.0449	0.0450	0.0463	0.0448	0.0448	0.0459
Delete P	0.9893	0.9877	0.9879	0.9878	0.9874	0.9874	0.9870
Hybrid							
1-gram							
Add F1	0.0382	0.0382	0.0518	0.0505	0.0365	0.0511	0.0505
Keep F1	0.1181	0.1181	0.1169	0.1181	0.1186	0.1174	0.1181
Delete P	0.9740	0.9740	0.9741	0.9722	0.9718	0.9717	0.9722
2-gram							
Add F1	0.0387	0.0339	0.0319	0.0319	0.0324	0.0312	0.0286
Keep F1	0.0744	0.0739	0.0740	0.0751	0.0732	0.0734	0.0744
Delete P	0.9812	0.9798	0.9800	0.9794	0.9787	0.9787	0.9784
3-gram							
Add F1	0.0293	0.0225	0.0215	0.0215	0.0215	0.0210	0.0177
Keep F1	0.0576	0.0571	0.0571	0.0588	0.0563	0.0567	0.0579
Delete P	0.9859	0.9841	0.9843	0.9843	0.9835	0.9836	0.9832
4-gram							
Add F1	0.0219	0.0153	0.0145	0.0150	0.0144	0.0142	0.0116
Keep F1	0.0454	0.0451	0.0449	0.0463	0.0440	0.0446	0.0452
Delete P	0.9893	0.9877	0.9879	0.9878	0.9873	0.9874	0.9869

Percent Modified	0	100					
Approach		random-period	random-the	replace-longest	replace-rand-period	replace-rand-the	random-period +replace-longest
Dress							
1-gram							
Add F1	0.0382	0.0382	0.0518	0.0505	0.0369	0.0511	0.0505
Keep F1	0.1181	0.1181	0.1169	0.1181	0.1187	0.1174	0.1181
Delete P	0.9740	0.9740	0.9741	0.9722	0.9718	0.9717	0.9722
2-gram							
Add F1	0.0387	0.0340	0.0324	0.0319	0.0324	0.0317	0.0287
Keep F1	0.0744	0.0738	0.0739	0.0751	0.0735	0.0735	0.0745
Delete P	0.9812	0.9797	0.9800	0.9794	0.9788	0.9787	0.9784
3-gram							
Add F1	0.0293	0.0224	0.0215	0.0215	0.0218	0.0216	0.0173
Keep F1	0.0576	0.0568	0.0571	0.0588	0.0567	0.0568	0.0579
Delete P	0.9859	0.9841	0.9843	0.9843	0.9836	0.9836	0.9832
4-gram							
Add F1	0.0219	0.0151	0.0147	0.0150	0.0146	0.0147	0.0113
Keep F1	0.0454	0.0446	0.0450	0.0463	0.0445	0.0445	0.0452
Delete P	0.9893	0.9876	0.9878	0.9878	0.9874	0.9873	0.9869
PBMT-R							
1-gram							
Add F1	0.0382	0.0382	0.0518	0.0505	0.0368	0.0509	0.0505
Keep F1	0.1181	0.1181	0.1169	0.1181	0.1187	0.1172	0.1181
Delete P	0.9740	0.9740	0.9741	0.9722	0.9718	0.9716	0.9722
2-gram							
Add F1	0.0387	0.0337	0.0320	0.0319	0.0327	0.0311	0.0288
Keep F1	0.0744	0.0739	0.0740	0.0751	0.0736	0.0731	0.0746
Delete P	0.9812	0.9798	0.9800	0.9794	0.9788	0.9786	0.9784
3-gram							
Add F1	0.0293	0.0223	0.0216	0.0215	0.0220	0.0207	0.0177
Keep F1	0.0576	0.0571	0.0572	0.0588	0.0568	0.0564	0.0581
Delete P	0.9859	0.9842	0.9843	0.9843	0.9837	0.9835	0.9832
4-gram							
Add F1	0.0219	0.0148	0.0145	0.0150	0.0151	0.0137	0.0116
Keep F1	0.0454	0.0447	0.0449	0.0463	0.0447	0.0442	0.0454
Delete P	0.9893	0.9877	0.9878	0.9878	0.9874	0.9873	0.9869

Table 10: SARI score breakdown (F1 and precision scores used in the score calculation for 1-, 2-, 3- and 4-gram) for all combination of systems and modification approaches (long table spanning two pages)