Exploring the Influence of Spelling Errors on Lexical Variation Measures

Ryo Nagata[†‡], **Taisei Sato**[†], **and Hiroya Takamura**[♭♠]
[†]Konan University
[‡]RIKEN AIP
[♭]National Institute of Advanced Industrial Science and Technology
[♠]Tokyo Institute of Technology
nagata-coling2018-k @ hyogo-u.ac.jp.

Abstract

This paper explores the influence of spelling errors on lexical variation measures. Lexical variation measures such as Type-Token Ration (TTR) and Yule's K are often used for learner English analysis and assessment. When applied to learner English, however, they can be unreliable because of the spelling errors appearing in it. Namely, they are, directly or indirectly, based on the counts of distinct word types, and spelling errors undesirably increase the number of distinct words. This paper examines the hypothesis that lexical variation measures become unstable in learner English because of spelling errors. Specifically, it tests the hypothesis on English learner corpora of three groups (middle school, high school, and college students). To be precise, it estimates the difference in TTR and Yule's K caused by spelling errors, by calculating their values before and after spelling errors are manually corrected. Furthermore, it examines the results theoretically and empirically to deepen the understanding of the influence of spelling errors on them.

Title and Abstract in French

Une exploration de l'influence des erreurs orthographiques sur les mesures de variation lexicale

Cet article explore l'influence des erreurs orthographiques sur les mesures de variation lexicale. Les mesures de variation lexicale telles que le rapport type-occurrence (*Type-Token Ratio*; TTR) et le K de Yule sont souvent employées pour analyser et évaluer l'anglais d'apprenants langue seconde. Lorsqu'on les applique a l'anglais d'apprenants, elles peuvent cependant s'avérer peu fiables du fait des erreurs orthographiques que l'on y rencontre. De fait, elles se fondent directement ou indirectement sur le décompte des types de mots diffŕents, et les erreurs orthographiques augmentent artificiellement le nombre de mots différents. Cet article examine l'hypothèse selon laquelle les mesures de variation lexicale deviennent instables sur l'anglais d'apprenants à cause des erreurs orthographiques. Il teste cette hypothèse sur des corpus d'anglais d'apprenants de trois groupes (élèves de collège, de lycée, et étudiants à l'université). Plus précisément, il estime la différence causée par les erreurs orthographiques sur le TTR et le K de Yule en calculant leurs valeurs avant et après correction manuelle des erreurs orthograpiques. Il examine de plus ces résultats théoriquement et empiriquement pour approfondir notre compréhension de l'influence des erreurs orthographiques sur ces mesures.

1 Introduction

Vocabulary richness measures are often used for learner language analysis and assessment as in the work by Arnaud (1984), Attali and Burstein (2006), Ishikawa (2015), Gregori-Signes and Clavel-Arroitia (2015), and Šišková (2012). Among a wide variety, Type-Token Ratio (TTR), which is defined as the

Proceedings of the 27th International Conference on Computational Linguistics, pages 2391–2398
Santa Fe, New Mexico, USA, August 20-26, 2018.

number of distinct words divided by the total number of words, is probably the most popular one because of its readiness; it is often used in the related research areas including the work as just mentioned above. Many modifications have been made to it, resulting in, for example, Guiraud's R (Guiraud, 1959), Herdan's C (Herdan, 1960), and Malvern and Richard's D (Malvern et al., 2004), to name a few (see Malvern et al. (2004) for the information about other measures). Another well known measure is Yule's K (Yule, 1944; Tanaka-Ishii and Aihara, 2015), which has the nice property that it is stable with respect to text length.

When applied to learner English[1], however, they might be unreliable because of spelling errors appearing in it. Typically, directly or indirectly, vocabulary richness measures including TTR and K are based on the number of distinct words in the target text. Here, it should be emphasized that the more spelling errors occur, the more (superficially) distinct words tend to appear (and thus, the greater number of distinct words). In the view of vocabulary richness, however, spelling errors derived from a canonical spelling should be treated as one word type. For example, the word *because* is mistakenly spelt as *becose*, *becouse*, and *becous* in learner English and they would unnecessarily increase the value of TTR if they were individually counted as unique word types. With the accumulation of the influence, vocabulary richness measures will likely deviate from their true values. A similar argument can be made about K that also uses the statistics on distinct words in the target text (in a different way, though).

Granger and Wynne (1999) have already pointed out this problem. They actually reported on the difference in TTR between an original text and its spelling-error-corrected version. At the same time, their report is not complete in that they only targeted spelling errors whose edit distance is within one from their correct spelling. Considering that spelling errors exist whose edit distance is more than one, one should adequately take care of all spelling errors in the target text to estimate accurately their influence on vocabulary richness measures. As Arnaud (1984) show (and also we will in Sect. 2), it is not straightforward to determine which type of spelling error to correct and how.

Apart from TTR, as far as we know, no one has yet revealed their influence on other lexical variation measures such as K. It would be especially interesting to see how they affect K which is stable with respect to text length.

In view of this background, in this paper, we explore the influence of spelling errors on TTR and K. Specifically, we explore the following hypothesis:

> **Hypothesis**: Lexical variation measures become unstable in learner English because of spelling errors.

to augment the findings Granger and Wynne (1999) showed. We test this hypothesis on English learner corpora of three groups (middle school, high school, and college students). To be precise, we estimate the differences in TTR and K caused by spelling errors by calculating their values before and after spelling errors are manually corrected. In addition, we examine the results theoretically and empirically to deepen our understanding of the influence of spelling errors on TTR and K.

The rest of this paper is structured as follows. Section 2 describes the corpus data we used in this work. It also explores the spelling errors we targeted. Section 3 describes the method we used to investigate the influence of spelling errors on TTR and K. Section 4 shows the results. Section 5 investigates them to discuss the influence of spelling errors on lexical richness theoretically and empirically.

2 Data and Spelling Errors

We use for our investigation English learner corpora of three groups ranging over middle school, high school, and college students whose mother tongue is all Japanese[2]. They consist of essays written on different topics; their statistics are shown in Table 1. Note that the essays in each group are written on different topics and thus it would be difficult to make intra-group comparisons; our emphasis here is on the inter-group comparisons before and after spelling error correction.

[1] In this paper, *learner English* refers to English as a Foreign Language.

[2] Because of the copyright, not all the corpus data can be open to the public. However, we have released a part of it with the information about spelling errors and their corrections to the public. It corresponds to the Konan-JIEM (KJ) learner corpus, which is available at `http://http://www.gsk.or.jp/en/catalog/gsk2016-b/`.

Group	# essays	# words	# errors	# topics	Av. edit distance[3]
Middle school	384	21,324	583	3	1.33
High school	251	23,561	680	3	1.37
College	438	37,774	1,271	14	1.42
TOTAL	1,073	82,659	2,534	20	1.38

Table 1: Basic Statistics on the Corpus Data.

To measure the difference, one has to determine which types of spelling errors should be corrected; we identified 13 error types in the corpora. Table 2 shows them with their acronyms and short explanations (e.g., RE for *Real word spelling error*).

Based on the taxonomy, we test two ways of correcting spelling errors to measure the difference. The first is to correct all 13 error types. The second is to classify them into three groups: those corrected, not corrected, and not counted. The first group (those corrected) consists of SP, PC, OC, GC, and NM. Spelling errors in these types would unnecessarily increase the number of distinct words if they were not corrected and left as they are. The second only consists of RE as in *Their is a house.* $\rightarrow$ *There is a house.* For this type of error, the writer might not know the correct word, and thus they should be left as is. Considering this, this type of error is not corrected (and thus it is counted as a unique word type). The rest (the third group) are those that do not exist in the English language. Therefore, they are not considered in calculating TTR and K. Namely, they are not included in the number of distinct words nor in the total number.

3 Method

We first applied the following preprocessing steps to the target corpora. We respectively used the sentence splitter and the tokenizer in Stanford Parser 3.5.0 (Chen and Manning, 2014) to split the essays into sentences and then into word tokens. We converted all word tokens into lowercase. After this, we removed those tokens containing no English alphabet letter from the corpora. In addition, we removed spelling errors whose correct spellings were not identified.

After this, we calculated TTR and K for the three corpora before and after spelling error correction. We tested two ways of correcting spelling errors as described in Sect. 2. Namely, we corrected all 13 types of errors and also only a part of them. We excluded mistakenly concatenated and split words from spelling errors. The former are word form errors that should be spelled with more than one word token as in *highschool* $\rightarrow$ *high school*. The latter are those that should spelled as one word token as in *grand father* $\rightarrow$ *grandfather*. We then compared the resulting TTR and K with those for the original corpora.

We used the following definitions for TTR and K. Let N be the total number of words in the target corpus, V be the number of distinct words in it. Then, TTR is defined as

$$\text{TTR} = \frac{V}{N}. \tag{1}$$

Also, let $V(m, N)$ $(1 \leq m \leq N)$[4] be the number of words appearing m times in a corpus whose total number of words is N. Then, K is defined as

$$K = c[-\frac{1}{N} + \sum_{m=1}^{N} V(m, N)(\frac{m}{N})^2] \tag{2}$$

where c is a constant enlarging of the value of K. Note that the larger the value of TTR is, the richer the vocabulary is and that the opposite holds for K. Also note that the symbols introduced here will be used again to discuss the results in Sect. 5.

[3]The average was calculated for spelling errors falling into SP error types, excluding those whose correct forms were unknown.

[4]Theoretically, m ranges between one and N. In practice, however, $V(m, N)$ tends to be zero for large values of m.

Error code	Explanation	Treatment
SP	Spelling that does not exist in English. e.g., I am a *sistem* engineer.	✓
PC	Inappropriate plural form conjugation. e.g., I didn't do *anythings*.	✓
OC	Over-regularized morphology. e.g., I *gived* her her hat.	✓
GC	Conjugation error other than the above two. e.g., I am *driveing*.	✓
NM	Spelling error in names. e.g., I went to *Desneyland*.	✓
RE	Real word spelling error (i.e., context sensitive error). e.g., *Their* is a house.	×
RO	Romanized Japanese. e.g., I ate an *omusubi*.	-
SR	Romanized Japanese that has no equivalent English expression. e.g., I went to *Hukuoka*. or that becomes proper English if transliterated. e.g., I ate *susi*.(susi → sushi).	-
CW	Coined word that is not used in English. e.g., I want to be a *nailist*.	-
FW	Foreign words other than Japanese. e.g., I have an *Arbeit*.	-
AL	Non-American (e.g. British English) spelling. e.g., It's my *favourite*.	-
AB	Improper abbreviation that is not used in English. e.g., I went to *USJ*.	-
O	Other than the above.	-

Table 2: Spelling Error Taxonomy and its Treatment: Each symbol denotes as follows: ✓: corrected; ×: not corrected; -: not counted.

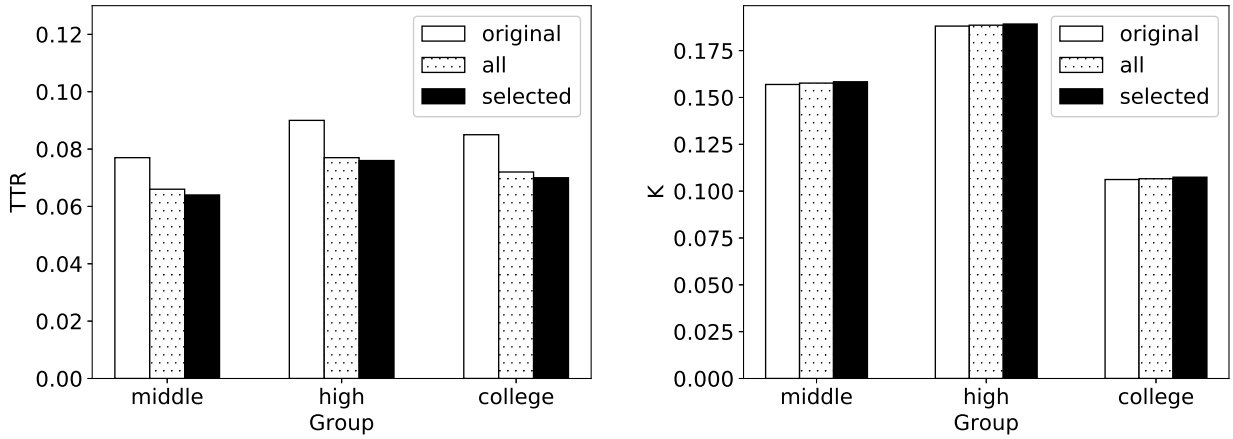
4 Results

Figure 1 shows the results. Figures 1–(a) and 1–(b) show the differences in TTR and K, respectively. The horizontal and vertical axes correspond to the three groups of the writers and the values of TTR and K, respectively. The labels *original*, *all*, and *selected* refer to the original corpus, the one where all 13 types of spelling errors are corrected, and the one where spelling errors are partly corrected, respectively.

Figure 1–(a) reveals that across the three groups, the difference in TTR is relatively large. For example, even the smallest difference, which is between *original* and *all* in *college*, approximately amounts to 16%. In *selected*, because the total number of words are slightly different (because of the not-counted spelling errors), one should be careful about its comparison with the other two. Having said that, it would be safe to say that there are at least some differences in TTR between them. These results imply that one would get varying values of TTR regardless of how he or she corrects spelling errors.

Here, note that although one might expect the values of TTR to decrease in order of *original*, *selected*, and *all*, it is not the case in the results. This is because there exist not-counted spelling errors in *selected*, which decreases both the number of distinct words and the total number of words.

Contrary to our expectation, Fig. 1–(b) reveals that almost no difference is observable in K. This applies to all groups with both ways of spelling error treatment; the difference is not more than 1% in all cases. This empirically shows that K is highly stable with respect to spelling errors.



(a): Difference in TTR.

(b): Difference in K.

Figure 1: Differences in TTR and K caused by Spelling Errors.

To sum up, the results give double answers to the hypothesis introduced in Sect. 1. TTR becomes unstable in learner English because of spelling errors. In contrast, this is not the case with K. In the next section, we explore the results theoretically and empirically to deepen our understanding of these phenomena.

5 Discussion

In this section, we explore why the influence of spelling errors is large in TTR and is very small in K. We begin with TTR, which is relatively easy to analyze, and then move on to K.

To begin with, let us assume the following situation. Suppose we have a spelling error-free corpus. Further suppose it undergoes a certain filter (or noise), which replaces a certain amount of words with misspellings (just as in the noisy channel model). The filter could be the learners' language device in their brain. Hereafter, we will refer to the error-free corpus and the erroneous corpus as *original corpus* and *error corpus*, respectively.

Now let us introduce some more symbols in addition to those defined in Sect. 3 to discuss TTR and K in more detail. Let $\mathbb{W}$ be the set of words appearing in the original corpus, $f(w)$ be the frequency of the word $w \in \mathbb{W}$, and n be the number of new spellings (words) occurring from the spelling errors in the error corpus. Here, it should be emphasized that the total number of words N is the same for the original and error corpora[5].

With these symbols, we can denote the number of distinct words (or spellings) in the error corpus as $V + n$. Strictly, words that occur only once in the corpus in question (i.e., hapax legomena) do not increase the number of distinct words or in other words do not contribute to $V + n$. Excluding spelling errors in these hapax legomena, it follows that the number of distinct spellings newly occurring from spelling errors directly affects TTR. For example, if the number of distinct spelling doubles, the increase in the number of distinct words also doubles. Also, $V + n$ shows that the frequencies of each spelling error do not affect TTR at all.

The discussion so far can be extended to other lexical richness measures that are based on the number

[5]This is because we assume that correctly spelt words in the original corpus are replaced with spelling errors.

of distinct words (i.e., V). For instance, one can apply the exact same discussion as above to Guiraud's R, which is defined as $\frac{V}{\sqrt{N}}$. The difference becomes relatively small in the measures that take the logarithm of V, but it is still relatively large compared to K (as we will show in the next paragraphs). For instance, Herdan's C, which is defined as $\frac{\log V}{\log N}$, suffers from the influence of n in the logarithm.

Now let us move on to K. The definition of K given by Eq. (2) can be rewritten as

$$K = c[-\frac{1}{N} + \sum_{w \in \mathbb{W}} (\frac{f(w)}{N})^2]. \tag{3}$$

Here, the summation is taken over the word set $\mathbb{W}$ in Eq. (3) whereas it is over the frequency of m in Eq. (2).

From Eq. (3), we can see that K is based on $f(w)^2$, i.e., the second power of each word frequency. This implies that even if n doubles in the original corpus because of spelling errors, that does not necessarily mean that their influence will double. Also, the influence from highly frequent words is dominant in the difference in K because of the second power.

We can further discuss the influence of each word. Suppose that a spelling error was created from $w \in \mathbb{W}$ and that $100r\%$ of all occurrences of w underwent the noise and were replaced with it. Then, the contribution of w to the entire value of K decreases by:

$$\frac{1}{N^2}\{(1-r)f(w)\}^2. \tag{4}$$

On the other hand, the newly created spelling increases the value of K by:

$$\frac{1}{N^2}\{rf(w)\}^2. \tag{5}$$

In total, the difference caused by a spelling error can generally be written as:

$$\frac{1}{N^2}[\{(1-r)f(w)\}^2 + \{rf(w)\}^2] = \frac{1}{N^2}\{(1-r)^2 + r^2\}f(w)^2. \tag{6}$$

Accordingly, it follows that the influence is only dependent on r. The difference becomes a maximum when $r = \frac{1}{2}$. At the same time, the values of r are expected to be small for most words. In other words, it is expected that a word is spelt correctly most of the time and only a few are misspelled[6]. If this is true, the following approximation holds:

$$\frac{1}{N^2}\{(1-r)^2 + r^2\}f(w)^2 = \frac{1}{N^2}(1 - 2r + 2r^2)f(w)^2 \tag{7}$$

$$\approx \frac{1}{N^2}(1-2r)f(w)^2. \tag{8}$$

This expression immediately shows that the difference caused by one type of spelling error is negligible when r is small. Even if r is relatively large, its influence becomes relatively small with respect to the whole value of K considering the term has the coefficient $\frac{1}{N^2}$. The discussion can be extended to the general situation where t spellings are created from $w \in \mathbb{W}$. The influence becomes maximum when $r = \frac{1}{2}$ (or generally, $r = \frac{1}{t+1}$). However, one can safely expect that it will rarely occur in frequent words in learner corpora (or corpora in general). For example, it would be rare to encounter a situation where half of the occurrences of a frequent word (e.g., *the*) are spelt correctly and the rest mistakenly (e.g., *hte*). It may happen to infrequent words, but again they scarcely affect the whole value of K because the influence of highly frequent words is dominant in K as we have just discussed at the beginning of this section.

We actually estimated r from the three corpora. It turned out that its mean and standard deviation were 0.06 and 0.10 for words where $10 \leq f(w) < 100$; similarly, 0.01 and 0.024 for words where

[6]Strictly, it can be restated that every word has, correctly or incorrectly, its representative spelling. Namely, it is assumed that $1 - r \ll 1$ or $r \ll 1$ holds.

$f(w) \geq 100$. These observations agree with our expectation and empirically explain why K is stable with respect to spelling errors.

The discussion above reconfirms the empirical findings that the difference caused by spelling errors is large in TTR and is negligible in K. Whether or not the difference in TTR can be problematic depends on the purpose and/or the way it is used. However, one should always keep in mind the fact that its difference is (at least superficially) large when one uses it. For example, the difference might cause a problem to a machine learning-based classifier when it is used as a feature; actually, some researchers (Pilán et al., 2016; Tack et al., 2017) are aware of this and thus they correct spelling errors as a preprocessing in their applications. In contrast, there is much less concern about the use of K. Besides, it has a nice property that it is also stable with respect to text length as Kimura and Tanaka-Ishii (2011) show. Consequently, it follows that it would be more accurate to use K instead of TTR when one analyzes learner English.

6 Conclusions

In this paper, we have addressed the influence of spelling errors on lexical variation measures. Specifically we have explored the following hypothesis:

> **Hypothesis**: Lexical variation measures become unstable in learner English because of spelling errors.

to augment the findings Granger and Wynne (1999) showed. We have tested it for TTR and Yule's K using three groups of learner English.

As a result, we have found that the hypothesis holds for TTR. To be precise, we have revealed that the difference in its value caused by spelling errors is relatively high throughout the tree groups, observing not less than 16% difference.

In contrast, this is not the case with K, which shows no more than 1% difference throughout the three groups. In other words, it is highly stable with respect to spelling errors.

We have investigated the results in detail to reveal why such a counterintuitive phenomenon occurs. We have shown theoretical and empirical reasons why new spellings derived from correct word forms by spelling errors directly affect the value of TTR, but not the value of K. Based on the results and the discussion, we have concluded that it would be more accurate to use K instead of TTR when one measures the lexical variation of learner English.

It will be interesting to test the present hypothesis on English texts written by learners other than Japanese. Also, it will be interesting to do the same for other languages than English. While our findings suggest that it will likely hold for other learner Englishes and also for other languages, mother tongue interference or the nature of other languages might affect the results. In particular, in certain languages (e.g., Japanese), it becomes much harder to identify spelling errors. Accordingly, the future work should include how to identify spelling errors in such languages. We will investigate these research questions in our future work.

Acknowledgments

We would like to thank Pierre Zweigenbaum for proofreading the abstract in French. We would also like to thank the three anonymous reviewers for their valuable comments on our paper.

References

Pierrre J.L. Arnaud. 1984. The lexical richness of L2 written productions and the validity of vocabulary tests. In *Proc. of International Symposium on Language Testing*, pages 14–28.

Yigal Attali and Jill Burstein. 2006. Automated essay scoring with E-rater v.2.0. *The Journal of Technology, Learning, and Assessment*, 4(3):3–30.

Danqi Chen and Christopher Manning. 2014. A fast and accurate dependency parser using neural networks. In *Proc. of 2014 Conference on Empirical Methods in Natural Language Processing*, pages 740–750.

Sylviane Granger and Martin Wynne, 1999. *Optimising measures of lexical variation in EFL learner corpora*, pages 249–257. Corpora Galore. Rodopi.

Carmen Gregori-Signes and Begoñ Clavel-Arroitia. 2015. Analysing lexical density and lexical diversity in university. In *Proc. of 7th International Conference on Corpus Linguistics: Current Work in Corpus Linguistics: Working with Traditionally-conceived Corpora and Beyond*, pages 546–556.

Pierre Guiraud. 1959. *Problèmes Et Méthodes De La Statistique Linguistique*. D. Reidel Publishing Company, Dordrecht.

Gustav Herdan. 1960. *Type-Token Mathematics: A Textbook of Mathematical Linguistics*. The Hague: Mouton, Amsterdam.

Shin'ichiro Ishikawa. 2015. Lexical development in L2 English learners' speeches and writings. In *Proc. 7th International Conference on Corpus Linguistics: Current Work in Corpus Linguistics: Working with Traditionally-conceived Corpora and Beyond*, pages 202–210.

Daisuke Kimura and Kumiko Tanaka-Ishii. 2011. A study on constants of natural language texts. *Journal of Natural Language Processing*, 18(2):119–137.

David D. Malvern, Brian J. Richards, Ngoni Chipere, and Pilar Duràn. 2004. *Lexical Diversity and Language Development*. Palgrave Macmillan, London.

Ildikó Pilán, Elena Volodina, and Torsten Zesch. 2016. Predicting proficiency levels in learner writings by transferring a linguistic complexity model from expert-written coursebooks. In *Proc. of 26th International Conference on Computational Linguistics*, pages 2101–2111.

Anaïs Tack, Thomas François, Sophie Roekhaut, and Cédrick Fairon. 2017. Human and automated CEFR-based grading of short answers. In *Proc. of 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 169–179.

Kumiko Tanaka-Ishii and Shunsuke Aihara. 2015. Computational constancy measures of texts–Yule's K and Rèyi's entropy. *Computational Linguistics*, 1(3).

Zdislava Šišková. 2012. Lexical richness in EFL students' narratives. *Language Studies Working Papers*, 4:26–36.

G. U. Yule. 1944. *The Statistical Study of Literary Vocabulary*. Cambridge.