綴り誤りが語彙の豊富さの指標に与える影響の分析

佐藤 太清† 永田 亮†,†† 高村 大也†††,††††
† 甲南大学知能情報学部 †† 理化学研究所 AIP センター ††† 産業技術総合研究所人工知能研究センター
†††† 東京工業大学科学技術創成研究院
E-mail: †s1571060@s.konan-u.ac.jp, ††nagata-nlp2018@ hyogo-u.ac.jp., †††takamura@pi.titech.ac.jp

1. はじめに

語学学習支援や第二言語習得に関する研究では，語彙の豊富さを表す指標がよく用いられる．代表的なものに，Type/Token Ratio（TTR）がある．TTR は文書中の異なり語数と総語数の比として定義される．応用例として，エッセイの自動採点 [1] や自由英作文の評価と語彙の指標の関係性の調査 [6] などがある．

しかしながら，学習者コーパスを分析の対象にした場合には注意が必要である．なぜなら，学習者が算出した言語データには綴り誤りが多く含まれることがあり，TTR により語彙の豊富さが適切に表されない可能性があるからである．綴り誤りが増えると，表層上の異なり語数も増える傾向にある．したがって，異なり語数に基づく TTR の値も大きくなる傾向となる．一方で，語彙の豊富さという観点からすると，綴り誤りは対応する正しい綴りに集約して異なり語数を計数することが好ましい．例えば，“because” という単語には，“becose”，“becouse”，“becous” などの綴り誤りが日本人英語学習者コーパスでは観測されるが，本来は全て “because” を表すことから，異なり語数は 1 とすべきである．このような影響が積み重なることで，TTR が本来の値よりも大きく見積もられる可能性が高い．

この問題は，既に Granger ら [3] により指摘されている．実際に，Granger らは，綴り誤りによる TTR の値の変動も報告している．ただし，その報告では，正しい綴りに対して編集距離が 1 となる綴り誤りのみを対象にしており正確な変動は明らかにされていない．編集距離が 2 以上となる綴り誤りも存在することを考慮すると，TTR の変動を正確に推定するためには，全ての綴り誤りを適切に修正し，TTR を算出する必要がある (注 1)．2. 節で述べるように，これはそれほど自明なことではない．

以上のような背景を考慮し，本稿では，綴り誤りが語彙の豊富さの指標に与える影響を明らかにする．具体的には，3 種類のグループ（中高大学生）の日本人英語学習者コーパスを対象にして，綴り誤りの修正前後で，TTR にどの程度の差異が生じるかを測定する．また，同じことを別の語彙の豊富さの指標である Yule の K [5]（以下，単に K と表記する）についても行う．K は，TTR と同様に文書から得られる統計量に基づくが，文書長に対して安定である指標として知られている [4]．この特徴のため，TTR より K の方が語彙の豊富さの指標としてより適切であるといえる．このような特徴をもつ K が，綴り誤りからどの程度の影響を受けるかを明らかにする．

結論から述べると，検証結果は次のように要約される．TTR については，書き手のグループを問わず，綴り誤りに起因して相対的に大きな変動が生じる．一方，K については，全くといってよいほど影響を受けない．更に，本稿では，TTR および K の変動を分析することで，このような現象が起こる理由を示す．

2. 使用データと綴り誤り

分析対象として日本人英語学習者コーパスを使用する．このコーパスは書き手にトピックを与え，そのトピックについて自由記述してもらったものである．書き手は中高大学生の 3 グループに分かれる．表 1 に，このコーパスに関する統計量を示す．各グループでトピック数が異なるため，グループ間で TTR および K を比較にすることにはそれほど意味がない．本稿では，あくまでも綴り誤りによる TTR，K のグループ内の変動に焦点をあてる．

綴り誤りの影響を調べるためには何を綴り誤りとするか決めなければならない．本稿では，従来研究 [7] で定義される綴り誤り 13 種に基づいて，2 つの修正方法を試みる．綴り誤りの種類については表 2 に示す．

1 つ目の修正方法は綴り誤り 13 種全てを修正するというものである．綴り誤りの例として非単語誤りや文脈依存誤りなどがある．例えば，非単語誤りは 1. 節で示したようなものである．ただし，綴り誤りには複数の単語からなる誤りと

表 1： 日本人英語学習者コーパスに関する統計量．

種類	文書数	総単語数	綴り誤り数	トピック数
中学生	384	21,324	583	3
高校生	251	23,561	680	3
大学生	438	37,774	1,271	14

(注 1)：綴り誤りを訂正した際に，異なり語数に影響がある場合とない場合があることに注意が必要である．訂正後の単語が，コーパスに元々なければ，その綴り誤りによる異なり語数の変動はない．

表 2： 綴り誤りの分類と修正方法.

分類名	説明	修正方法 (注 2)
非単語誤り	英語には存在しない綴り.	○
複数形活用誤り	単複の活用誤りに起因する綴り誤り.	○
過剰一般化活用誤り	活用を過剰一般化したことによる綴り誤り.	○
活用誤り一般	上 2 つ以外の活用誤り.	○
名前綴り誤り	名前における綴り誤り.	○
文脈依存誤り	英語の綴りではあるが与えられた文脈では正しくない綴り.	×
ローマ字語	ローマ字表記された日本語.	-
ローマ字語特殊	対応する英語がないローマ字語における綴り誤り.	-
	または英語化した日本語における綴り誤り.	-
和製英語	英語には存在しない和製英語.	-
外国語	英語でも日本語でもない外国語.	-
代替綴り	米語式綴り以外.	-
略語誤り	英語では使用されない略語.	-
その他	上記に分類されない綴り誤り.	-

1 つの単語が複数の単語に分かれる誤りがある．これらの誤りは綴り誤りとしない．

2 つ目の修正方法では，より正確に分析するために，英語の語彙の豊富さを測るという観点から修正する綴り誤りの種類を取捨選択する．具体例には，13 種の綴り誤りのうち，修正するもの，修正しないもの，除外するものに分ける．修正する綴り誤りは非単語誤り，複数形活用誤り，過剰一般化活用誤り，活用誤り一般，名前綴り誤りとする．これらの綴り誤りは，正しい綴りに集約しなければ異なり語数が増え見かけ上，語彙が豊富になる誤りである．修正しないものは前述の分割・統語誤りと文脈依存誤りである．文脈依存誤りには，例えば，“there’s” を “their’s” と間違えるものがある．この例では書き手が前者の単語 “there” を知らない可能性がある．この綴り誤りを修正すると書き手が知らない可能性がある単語を計数することになるため，修正すべきではない．上述以外の綴り誤りはローマ字語，外国語など英語ではない単語であり，英語の語彙に含めるのは妥当ではないと考えられる．これらの綴り誤りは除外する．すなわち総語数などの統計量に計数しない．

3. 検 証 手 順

前処理として Stanford Parser 3.5.0 [2] を用いてコーパス内の単語の同定を行った．また全てのアルファベットは小文字に変換した．さらに TTR と K の厳密な値を求めるために次の条件の単語をコーパス内から削除した．条件はアルファベットを含まない，修正後の単語が “???”(注 3) の 2 つである．

(注 2)：選択して修正する場合の修正方法である．各記号の意味は次の通りである．○：修正する　×：修正しない　-：除外.

(注 3)：従来研究 [7] では，正しい綴りが不明な誤りは正しい綴りを “???” としている.

前処理の後，綴り誤りの修正前後で TTR と K を算出した．修正方法は，2. 節で述べた 2 つの方法を用いた．すなわち従来研究 [7] の綴り誤り 13 種全てを修正する方法と，部分的に修正する方法である．検証は TTR，K をそれぞれ修正前後で比較することで行った．

TTR と K は以下の定義を用いて算出した．いま，単語の異なり語数を V，総語数を N で表すことにする．このとき TTR は，

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

で定義される．また，m をコーパス中の単語の頻度とする．更に，$V(m, N)$ を総語数 N であるコーパス中で頻度が m である単語の種類数とする．このとき K は，

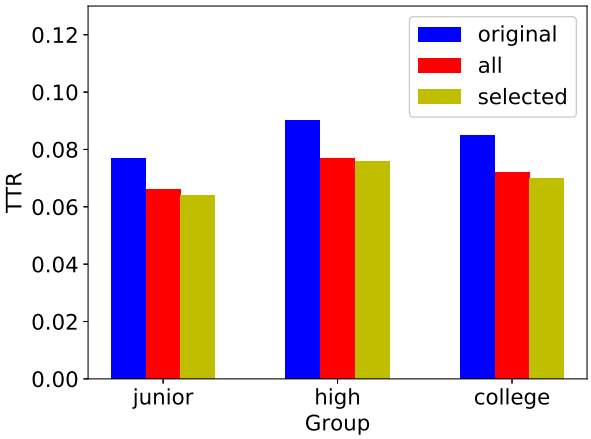
$$K = c[-\frac{1}{N} + \sum_{m=1}^{N} V(m, N)(\frac{m}{N})^2] \tag{2}$$

で定義される．ただし，c は，K の見た目の大きさを調整する定数で本質的な意味はない．なお TTR では値が大きい方が，K では逆に値が小さい方が語彙が豊富であることを意味することに注意されたい．

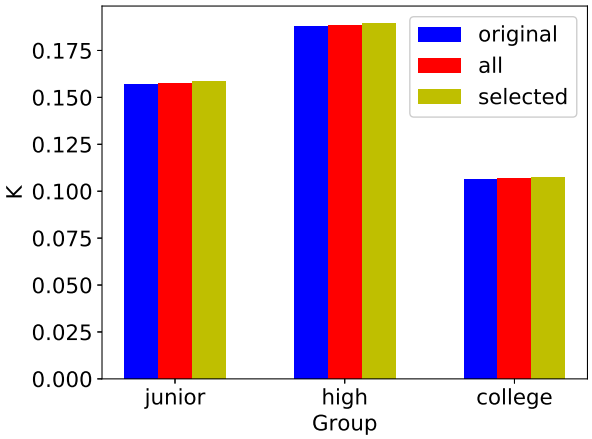
4. 検 証 結 果

図 1 に結果を示す．図 1–a と図 1–b は，それぞれ修正前後の TTR の値と K の値を棒グラフにしたものである．グラフの横軸はグループを表す．縦軸は算出した TTR，または K に対応する値である．各ラベル “original”，“all”，“selected” はそれぞれ綴り誤りを修正しなかったもの，全ての綴り誤りを修正したもの，一部の綴り誤りだけを修正したものを指す．

図 1–a の TTR では，どのグループでも綴り誤りの修正前後で大きな変動がある．“original” と “all” の変動を計算すると最小でも 16%ある．TTR では予想通り，綴り誤りの影



1–a：TTR における変動



1–b：K における変動

図 1：　検証結果：綴り誤りに起因する TTR および K の変動.

響が大きいことがわかる．“selected” は対象外の綴り誤りがあり、他の 2 つと総語数が異なるため，比較には慎重になるべきであるが，“selected” と “all” との間にも若干の変動があることがわかる．このことは，何を綴り誤りとしと修正するかにより，TTR に変動がある可能性を示唆する．

また，直感的には，“original”，“selected”，“all” の順に TTR の値が小さくなると予想されるが，図 1–a では “original”，“all”，“selected” の順になっており，予想に反する．これは “selected” では，対象外とした綴り誤りがあり，異なり語数，総語数と共に減少したことに起因する．

一方，図 1–b に示されるとおり，K についてはどのグループでも変動がほとんど観測されない．変動は，どのグループでも 1%以下である．K では予想に反して，綴り誤りの影響がほとんどないことがわかる．また，綴り誤りを取捨選択した “selected” は “all” より大きいが，その変動はほとんどないため，手間をかけて修正する綴り誤りを取捨選択する必要はないといえる．いずれにせよ，K は，綴り誤りに対して非常に安定な指標であることが実験的に示されたといえる．

5. 考　　察

本節では，綴り誤りからの影響が，TTR では大きく，K では小さいことを考察する．まずは，比較的取り扱いが容易である TTR から始め，次に K について議論する．

議論を始めるにあたり，次のような場面を想定する．いま，綴り誤りが全く存在しないコーパスを考える．このようなコーパスに対して，なんらかのノイズが加わり綴り誤りを含んだ別のコーパスができたとする．ここでは仮想的なノイズを想定しているが，例えば，学習者（の脳）というフィルタを通して付加されたノイズを具体的に考えることができる．以降では，綴り誤りのないコーパスを**原コーパス**，綴り誤りが付加されたコーパスを**誤りコーパス**と表記する．なお，4. 節で見たように，綴り誤りを全て修正するか選択して修正するかで，得られる TTR と K の値が変わるが，修正前に比べて変動が相対的に小さいため，以降では全て修正する場合を想定して議論を進める．

さて，考察を進めるにあたり，3. 節で定義した記号に加え，次の記号を導入する．いま，原コーパスに出現した単語の集合を $\mathbb{W}$ で表すことにする．したがって，3. 節の記号を用いると $|\mathbb{W}|{=}V$ である．また，綴り誤りにより，誤りコーパスに n 個の新しい綴り（単語）が生じたとする．ここで，上の想定に基づくと，総語数 N については，原コーパスと誤りコーパスで変化がないことに注意が必要である (注 4)．更に，単語 $w \in \mathbb{W}$ の原コーパス中での頻度を $f(w)$ で表すことにする．

これらの記号を用いると，誤りコーパスにおける異なり語数は，$V+n$ となる (注 5)．したがって，新しく生じた綴りの種類数の影響が，直接，TTR に加わることになる．例えば，新しく生じた綴りの種類数が 2 倍になれば，異なり語数の増加も 2 倍になる．また，新しく生じた綴りの頻度は関係ないということもわかる．

次に，K について議論する．式 (2) で定義される K は，少し表現を変え，

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

(注 4)：ある語が綴り誤りに置き換わるという想定をしているため総語数に変化は生じない．

(注 5)：条件として，綴り誤りの影響で正しい綴りの単語がコーパス中から消えるということが起こらない場合に限る．

と表すこともできる．式 (2) では，頻度 m についての和を考えているのに対して，この式では，単語の集合についての和を考えていることに注意されたい．

式 (3) の表現から，総語数 N が同じ場合，K は，$f(w)^2$，すなわち，各単語の頻度の二乗に基づいて決定されることがわかる．したがって，TTR とは異なり，新しく生じた綴りの種類数 n の値が 2 倍になっても，K に対する影響は必ずしも 2 倍にならないこともわかる．また，$f(w)^2$ に基づくことから，K に対する影響は，高頻度な w について単語の影響が支配的になることもわかる．

更に，1 つの単語についても定性的な分析が可能である．いま，ある単語 $w \in \mathbb{W}$ から，1 種類の綴り誤りが生まれたとしよう．また，その綴り誤りは w の全出現のうち，$100r\%$ に生じたとする．このとき，その影響は，一般に，

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

で表される．したがって，K への影響は，綴り誤りが生じた割合 r に依存することになる．最も変動が大きいのは $r = \frac{1}{2}$ のときであるが，多くの単語では，r は，通常，非常に小さい値であると考えられる．言い換えれば，大部分については正しく綴られ，綴り誤りの起こる確率は相対的に小さいということである (注 6)．この仮定が正しいとすると，

$$\frac{1}{N^2}[\{(1-r)f(w)\}^2+\{rf(w)\}^2] \approx \frac{1}{N^2}(1-2r)f(w)^2$$

となり，綴り誤りの影響はほぼ無視できることがわかる．仮に，仮定が成り立たず，r が大きな値となっても，$\frac{1}{N^2}$ がかかることを考慮すると K 全体への影響は非常に小さい．ここまで，1 つの単語から 1 種類の綴り誤りが生じる場面を想定していたが，一般に t 種類の綴り誤りが生じる場合も，同様の議論が成り立つ．綴り誤りの影響が最大になるのは，$r = \frac{1}{2}$（一般には，$r = \frac{1}{t+1}$）のときであるが，高頻度な単語では起こりにくいと予想される．例えば，高頻度な単語（例えば，“the”）の半分は正しく綴られ，残りの半分は綴り誤り（例えば “hte”）となるような状況は稀であろう．実際，今回用いたコーパスで r を推定 (注 7) したところ，頻度 10 以上，100 未満の単語では平均 0.06（標準偏差 0.10），頻度 100 以上では平均 0.01（標準偏差 0.024）となり r が小さいことが確かめられた．

以上により，綴り誤りによる変動が，TTR では大きく，K では小さいことが定量的にも定性的にも確かめられた．TTR の変動が大きいことが問題となるかどうかは，使い方や目的に依存するため一概にはいえない．しかしながら，綴り誤りのあり／なしで，少なくとも見た目上は TTR の値は 15〜20%程度変動する．TTR を用いる場合は，このような変動を常に念頭におく必要がある．したがって，TTR は，学習者コーパスを対象にした場合，扱いが難しい指標であるといえる．一方，K については，そのような心配は少ない．更に，木村ら [4] が示しているように，K は文章長に対しても安定な指標である．これらのことから，学習者コーパスを対象にした分析では，TTR ではなく K を積極的に使用すべきであると結論付けられる．

6. おわりに

本稿では綴り誤りが語彙の豊富さの指標に与える影響について調査した．3 種類の（中高大学生）の日本人英語学習者コーパスを対象にして，綴り誤りの修正前後で TTR および K の値にどの程度の変動が生じるかを測定した．その結果，TTR は綴り誤りの影響を受けやすく，綴り誤りの修正前後で変動が大きい傾向にあることを明らかにした．一方，K については綴り誤りの影響がほとんどないことも明らかにした．また，定性的な分析により，K が綴り誤りに対して非常に安定である理由を示した．この結果は，語彙の豊富さを調査する場合，TTR ではなく，綴り誤りの影響を受けにくい K を用いることが適切であることを示唆している．

(注 6)：より正確には，ある単語において，正しい綴りとその綴り誤りの中には，代表的な綴りが存在し，その他の綴りの頻度は相対的に低いと述べられる．すなわち，$1-r \ll 1$ または $r \ll 1$ ということを仮定している．

(注 7)：中高大学生 3 グループのサブコーパスをまとめたものを使用した．また，r は，ある 1 つの単語について，全体（正しい綴りと綴り誤り）の頻度に対する綴り誤りの頻度と定義して求めた．上述の平均値は，この値の平均値である．

参考文献

[1] Y. Attali and J. Burstein, “Automated essay scoring with E-rater v.2.0,” The Journal of Technology, Learning, and Assessment, vol.4, no.3, pp.3–30, 2006.

[2] D. Chen and C.D. Manning, “A fast and accurate dependency parser using neural networks,” Proceedings of EMNLP 2014, pp.740–750, 2014.

[3] S. Granger and M. Wynne, “Optimising measures of lexical variation in EFL learner corpora,” Corpora Galore, pp.249–257, 1999.

[4] D. Kimura and K. Tanaka-Ishii, “A study on constants of natural language texts,” Journal of Natural Language Processing, vol.18, no.2, pp.119–137, 2011.

[5] G.U. Yule, The Statistical Study of Literary Vocabulary, Cambridge, 1944.

[6] 水本 篤, “自由英作文における語彙の統計指標と選定者の総合的評価の関係,” 統計数理研究所共同研究リポート, no.215, pp.15–28, 2008.

[7] 永田 亮，Graham Neubig, “綴り誤り研究のための日本人英語学習者コーパスの構築,” 言語処理学会第 23 回年次大会発表論文集, pp.1030–1033, 2017.