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

永田　亮† GrahamNeubig††
† 甲南大学知能情報学部 †† Carnegie Mellon University

1. はじめに

英語学習者の綴り誤りに関する研究において，基礎的なデータとなるのは綴り誤りの情報が付与されたコーパスであるが，現状では十分に整備されていない．多様な学習者コーパスが公開される中で，綴り誤り情報が付与されたものは，ほとんど存在しない．そのため，多くの研究者は，独自に構築した学習者コーパスを綴り誤りの研究に使用している（例えば，文献 [1], [5], [6], [8] など）．

綴り誤り情報が付与された学習者コーパスは言語処理にも関係が深い．学習者の英文中の綴り誤りを自動的に検出／訂正する技術 [3], [4] は言語処理の一つの重要なタスクであるが，その開発と評価には，綴り誤り情報が付与された学習者コーパスが必要不可欠である．また，綴り誤りは，学習者の英文を扱う言語処理システムの性能を低下させる要因となることが報告 [7], [9]〜[11] されており，綴り誤りとシステム性能の関係を明らかにする上でも，綴り誤りの情報は重要な役割を果たす．

著者が知る限り，綴り誤りの情報が付与された英語学習者コーパスの中で一般公開されているものは，FCE CLC コーパス [12] のみである．同コーパスは，綴り誤り情報を提供する数少ないコーパスということで，その価値は高いが，誤り全般の情報を提供することを目的としており，綴り誤りに関しては三種類の分類（非単語誤り，文脈依存誤り，米語綴り）しかない．2. で述べるように，日本人英語学習者コーパスを対象にして，我々が行った調査では，十三種類もの綴り誤りが確認されている．従って，情報の粒度という点で，FCE CLC コーパスには改善の余地がある．

以上のような背景を受け，本研究では新たな綴り誤り情報付き英語学習者コーパスを構築した．具体的には，十三種類の誤り分類に基づき，複数の英語学習者コーパスに綴り誤り情報を付与した．従って，FCE CLC コーパスに比べ，本コーパスは，より詳細な綴り誤りの情報を提供する．更に，全てのコーパスは，同時に文法誤り情報も付与されている．また，一部については，品詞，句，構文（句構造）の情報も付与されている．そのため，本コーパスは，綴り誤り，文法誤り，文法情報（品詞，句，構文）を組み合わせた多角的な分析に利用可能である．これらのコーパスのうち，著作権処理がなされた部分については，教育研究用に公開している．また，全コーパスから学習した綴り誤りモデルも教育研究用に公開している．

2. 綴り誤りの種類

綴り誤りの情報を付与する上で，何を綴り誤りとするかが重要である．英語の辞書に掲載されていない文字列を綴り誤りの候補とすることは可能である．しかしながら，その場合，綴り誤りでないものも多く含んでしまう可能性がある．例えば，日本人英語学習者の英文であれば，英語の辞書に掲載されていない人名や地名が頻出する．また，英単語の代わりにローマ字表記した日本語が使用される場合もある．逆に，英語の辞書には掲載されているが，与えられた文脈には適さない文脈依存誤りも存在する．何を綴り誤りとするかは，結局のところ，研究の目的に大きく依存する．

そこで，本研究では，綴り誤りとして扱われる可能性があるものをできるだけ幅広く綴り誤りと認める方針とした．一旦，幅広く綴り誤りを付与しておけば，研究の目的に応じて，必要な誤りだけを取捨選択することは容易であるためである．4. で説明する日本人英語学習者コーパスを調査したところ，表 1 に示す十三種類の綴り誤りを確認した．なお，日本の英語教育では，米語式の綴りが用いられることが多いことから，本研究でも米語式の綴りを採択している．

少し分かりにくいが，綴り誤りの中には，二段階に誤る可能性があるものもある．例えば,「ローマ字語」は，ローマ字表記された日本語自体に綴り誤りがある（例：“omusube”；正しくは “omusubi”）場合がある．その場合，正しいローマ字表記にしたうえで，対応する適切な英語に訂正（例：上の例の場合，“rice ball”）することができる．これに該当するものは，表 1 の「二段階誤り」の列に「○」を付している．

以上の十三種類に加えて，綴り誤りの形態という観点からも，綴り誤りを分類することができる．すなわち，綴りの分割と連結である．綴りの分割は，本来一つの単語であるべきものが複数の要素に分割された誤りである（例：“grand father”）．逆に，綴りの連結は，二語以上の単語が一つに連結された誤りである（例：“highschool”）．綴りの分割と連結は，上述の十三種類と複合的に起こることに注意が必要である．

3. 綴り誤り情報の付与方法

本研究では，XML に準拠した形式で綴り誤りに関する情報を付与する．XML タグを用いて，誤りの位置，種類，正しい綴りを学習者コーパスに付与する．具体例を示すと，

表 1： 綴り誤りの分類.

ラベル	分類名	説明	例	二段階誤り
sp	非単語誤り	英語には存在しない綴り.	I am a *sistem* engineer.	
re	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.	*Their* is a house	
rom	ローマ字語	ローマ字表記された日本語.	I ate an *omusubi*.	○
romsp	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.	I went to *Hukuoka*.	
		または英語化した日本語における綴り誤り.	I ate *susi*.	
wa	和製英語	英語には存在しない和製英語.	I want to be a *nailist*.	○
for	外国語	英語でも日本語でもない外国語.	I have an *Arbeit*.	○
conj_pl	複数形活用誤り	単複の活用誤りに起因する綴り誤り.	I didn't do *anythings*.	○
conj_or	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.	I *gived* her her hat.	○
conj	活用誤り一般	conj_pl と conj_or 以外の活用誤り.	I am *driveing*.	○
name	名前綴り誤り	名前における綴り誤り.	I went to *Desneyland*.	○
alt	代替綴り	米語式綴り以外.	It's my *favourite*.	
abb	略語誤り	英語では使用されない略語.	I went to *USJ*.	○
ur	その他	上記に分類されない綴り誤り.	It is *Univer*.	

I am a <sp crr="system">sistem</sp> engineer.

のようになる.

この例からわかるように，綴り誤りの位置は，綴り誤りがある文字列を XML タグで囲うことで示す．綴りの分割がある場合には，複数のトークンを囲う.

誤りの種類については，タグのラベル（例：sp）で表す．使用するラベルの一覧は，既に，表 1 に示した通りである.

正しい綴りについては，属性（crr=""）内に記す．正しい綴りが不明である場合は crr="???"とする．具体的に，正しい綴りとして記入する情報を表 2 に示す.

2. で説明したように，綴り誤りの種類によっては，誤りが二段階となることがあるため，特別な記述方法が必要となる．本研究では，二種類の属性（crr=""と spcrr=""）を用いて，二段階の誤りの情報を記述する．例えば，

He has an <for spcrr="Arbeit" crr="part-time job">Albito</for>.

のようになる．この例では，spcrr="Albeit"で，まず，ドイツ語の正しい綴りに訂正し，crr="part-time job"で，対応する適切な英語に訂正している.

表 2： 綴り誤りに対する訂正情報.

ラベル	分類名	訂正情報
sp	非単語誤り	対応する正しい綴り.
re	文脈依存誤り	文脈に適した綴り.
rom	ローマ字語	対応する英語の訳語.
romsp	ローマ字語特殊	対応する正しい綴り.
		訳語がない場合は，crr="???".
wa	和製英語	対応する英語の訳語.
for	外国語	対応する英語の訳語.
conj_pl	複数形活用誤り	正しく活用した語.
conj_or	過剰一般化活用誤り	正しく活用した語.
conj	活用誤り一般	正しく活用した語.
name	名前綴り誤り	正しく綴られた名前.
alt	代替綴り	対応する米語式の綴り.
abb	略語誤り	対応する英語の略語でない表記.
ur	その他	通常は crr="???".

4. 綴り誤り情報付きコーパス

対象コーパスとして，Konan-JIEM corpus [10]（以下，KJ と略記）を選択した．同コーパスは，著者らが 2007 年より構築を続けている日本人英語学習者コーパスである．既に，品詞，句，統語（句構造），文法誤りの情報を付与している．全ての英文について著作権処理をしており，教育研究用に公開している.

KJ に加えて，独自に収集，構築した別の日本人英語学習者コーパス（以下，JSC と略記）も付与対象とした．KJ の書き手は全て大学生であるが，JSC では，中学生，高校生，大学生の三グループにわたる.

これら二種類のコーパスに綴り誤り情報を付与した．前処理として，文分割を施した（一行一文形式とした）．一方，トークン分割は行わなかった．各文を目視で確認し，綴り誤りと正しい綴りを同定した．その後，オリジナルのコーパスと綴り誤り情報を付与したコーパスの差分を求め，付与結果の再確認を行った．更に，別の作業者が，コーパス全体のチェックをもう一度目視で行った.

表 3 に両コーパスの統計を示す．表 3 の誤り数は，十三種類全てを対象にした値である.

表 4 に，両コーパスにおける誤りの分布を示す．誤りの分類は，KJ での割合が多い順にソートしている.

表 4 より，両コーパスとも「非単語誤り（sp）」が最も多いことがわかる．これは，過去の知見 [5] とも一致する．し

表 3： 各コーパスに関する統計量.

コーパス	KJ	JSC
トークン数	30,586	66,955
綴り誤り数	933	1,621
綴り誤り率（%）	3.1	2.4

かしながら，KJ と JSC では，その割合は大きく違う．これは，KJ では書き手が大学生のみであるのに対し，JSC では書き手が中学生，高校生，大学生であることに起因する可能性がある．また，JSC では,「その他（ur）」の割合が KJ に比べ非常に高い．この理由の一つとして，JSC では，一部の英文については紙ベースで収集したことを挙げることができる（転記時に判読できなかった単語を綴り誤りに含めている）．

表 4： 綴り誤りの分布（%）.

ラベル	綴り誤り分類	KJ	JSC
sp	非単語誤り	54.7	76.1
re	文脈依存誤り	15.5	5.7
rom	ローマ字語	10.7	5.7
romsp	ローマ字語特殊	5.5	1.5
wa	和製英語	4.5	0.4
for	外国語	3.0	3.2
conj_pl	複数形活用誤り	1.8	1.5
conj_or	過剰一般化活用誤り	1.4	2.7
conj	活用誤り一般	1.1	0.2
name	名前綴り誤り	0.9	0.4
alt	代替綴り	0.6	0.7
abb	略語誤り	0.2	0.4
ur	その他	0.1	1.5

5. 綴り誤りの傾向分析

構築した綴り誤り情報付き学習者コーパスを利用して，綴り誤りを定量的，定性的に分析することが可能である．例えば，よく知られた「日本人英語学習者は，“l” と “r” を頻繁に混同する」という綴り誤りなどをコーパスデータから得られる統計量を用いて定量的に確認することができる．更には，典型例として知られていない綴り誤りの傾向を発見できる可能性もある．

ここでは，ケーススタディとして，本コーパスから，綴り誤りモデルを学習し，綴り誤りの分析に利用した．綴り誤りモデルとして，本研究では，ある文字列が別の文字列に誤る条件付き確率を用いる．すなわち，綴り誤りモデルを $\Pr(e|c)$ とする．ただし e と c は，それぞれ誤りである文字列とそれを正しく訂正した文字列を表す．

この確率の推定には，綴り誤りと正しい綴りのペアだけではなく，文字単位のアライメントが必要となる．文字単位のアライメントは，次のヒューリスティクスで求める（例中の ϕ は，空文字を表す）:

（1） 綴り誤りと正しい綴りの間で，編集距離が最小となる文字マッチングを得る．

例 綴り誤り：hirery，正しい綴り：hillary

h i r ϕ e r y

h i l l a r y

（2） マッチしない文字をグルーピングする．

h i re r y

h i lla r y

（3） 同じ文字タイプ（母音，子音，数字，その他）で始まり，別の文字タイプが続くグループを分割.

h i r e l y

h i ll a r y

このアライメントを施した結果を対象にして，modified Kneser-Ney smoothing [2] を用いて条件付き確率を推定した．

表 5 に，条件付き確率が高い順に上位 20 件を示す(注 1)．表中の $c \to e$ は，ある文字列 c が e に誤るという意味を表す．すなわち，$\Pr(e|c)$ に対応する．なお，表中の $\Pr(e|c)$ は%表示であることに注意されたい．

表 5 より，予想通り，日本では区別しない音（“l → r” や “v → b” など）や類似した音（“z → g” や “c → s”）で誤りが多いことがわかる．ただし，典型例とされる “l” と “r” の混同でも，その確率は 1%未満であり，大部分は正しく綴られていることになる．最も確率が高いのは，“z → g” であが，“magagine” のような綴り誤りで見られた．これは，正しい綴りの単語（例：“engine” など）からの類推による誤りと想像される．

また，母音の区別に関する誤りも多い．例えば，“stady（正しくは，study）” のように,（日本人）英語で,「あ」と発音されるが “a” 以外の文字で表記される単語は綴り誤りが頻出する．また，コーパスデータ中には，逆の影響（“a → u”）もみられた（例：“buttle”；正しくは，“battle”）．

別の傾向として，表 5 から，文字の余剰（例:“ϕ → e”）よりも，文字の脱落（例:“r → ϕ）のほうが多いことが見て取れる．文字の脱落は，日本語で対応する音がない（例：“forward”），ローマ字表記とは異なる発音である（例：“because”），そもそも英語で発音されない文字（例：“designer”）などに分類される（いずれの例も，下線部が脱落）．

以上の通り，綴り誤り情報付き学習者コーパスから得られる綴り誤りモデルを通じて，綴り誤りを定量的に分析することができる．また，綴り誤りモデルを通して得られた知見を

（注 1）：ただし，学習に用いた綴り誤り情報付きコーパスでの頻度が 5 以上の e と c の組み合わせのみを対象とした．また，$e = c$，すなわち誤りでないものもは対象外とした．

表 5： 日本人英語学習者コーパスで確率の高い綴り誤り．

$\Pr(e\|c)$ (%)	$c \to e$
5.26	z → g
0.55	ll → r
0.52	l → r
0.42	u → a
0.38	r → ϕ
0.34	e → ϕ
0.33	c → s
0.31	r → l
0.30	v → b
0.29	m → n
0.27	g → ϕ
0.25	h → ϕ
0.24	o → a
0.23	u → ϕ
0.23	l → ϕ
0.21	ϕ → e
0.19	n → ϕ
0.18	a → e
0.17	p → ϕ
0.17	c → ϕ

コーパスデータ中の用例と照らし合わせることで，定性的な分析も可能となる．このような分析を，綴り誤り情報付き学習者コーパスと綴り誤りモデルなしで行うことは困難であろう．

6. おわりに

本稿では，著者らが構築した綴り誤り情報付き学習者コーパスについて述べた．本コーパスの構築にあたり，綴り誤りの分類を行い，その分類に基づき綴り誤りの情報を日本人英語学習者コーパスに付与した．また，構築したコーパスを，綴り誤りに関する統計量と共に紹介した．本コーパスは，著作権処理がなされている部分については，教育研究用に公開している(注 2)．また，本コーパスから学習した綴り誤りモデルも公開している．

参考文献

[1] Y. Bestgen and S. Granger, “Categorizing spelling errors to assess L2 writing,” International Journal of Continuing Engineering Education and Life-Long Learning, vol.21, no.2/3, pp.235–252, 2011.

[2] S.F. Chen and J. Goodman, “An empirical study of smoothing techniques for language modeling,” Proc. of 34th ACL, pp.310–318, 1996.

[3] M. Flor, “Four types of context for automatic spelling correction,” TAL, vol.53, no.3, pp.61–99, 2014.

[4] M. Flor and Y. Futagi, “On using context for automatic correction of non-word misspellings in student essays,” Proc. of 7th BEA Workshop, pp.105–115, 2012.

[5] M. Flor and Y. Futagi, “Patterns of misspellings in L2 and L1 English: a view from the ETS spelling corpus,” Bergen Language and Linguistic Studies, vol.6, pp.107–132, 2015.

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

[7] N.R. Han, M. Chodorow, and C. Leacock, “Detecting errors in English article usage by non-native speakers,” Natural Language Engineering, vol.12, no.2, pp.115–129, 2006.

[8] K. Kukich, “Techniques for automatically correcting words in text,” ACM Computing Surveys, vol.24, no.4, pp.377–439, 1992.

[9] R. Nagata and K. Sakaguchi, “Phrase structure annotation and parsing for learner English,” Proc. of 54th ACL, pp.1837–1847, 2016.

[10] R. Nagata, E. Whittaker, and V. Sheinman, “Creating a manually error-tagged and shallow-parsed learner corpus,” Proc. of 49th ACL, pp.1210–1219, 2011.

[11] J.Z. Sukkarieh and J. Blackmore, “c-rater: Automatic content scoring for short constructed responses,” Proc. of 2nd International FLAIRS Conference, pp.290–295, 2009.

[12] H. Yannakoudakis, T. Briscoe, and B. Medlock, “A new dataset and method for automatically grading esol texts,” Proc. of ACL, pp.180–189, 2011.

(注 2)：本コーパスの利用を希望する方は，第一著者まで連絡されたい．なお，近日中に，言語資源協会からも公開予定である．