Amazon Mechanical Turk を利用した キーストロークログからのスペルミスの収集と分析

馬場 雪乃*
東京大学大学院 情報理工学系研究科
ybaba@nii.ac.jp

鈴木 久美
Microsoft Research
hisamis@microsoft.com

1　はじめに

コンピュータで文字入力をする際，人間はたくさんの打ち間違い（タイポ）を生み出してしまう．入力中にタイポに気がついた場合には BACKSPACE (BS) キーなどにより該当箇所を削除・修正すればいいが，気づかれなかったタイポは最終出力文字列に残ってしまう．

スペル訂正に関する研究は広く行われているが，訓練データとして用いられているのはニュース記事や検索クエリなどの最終出力文字列に残っているタイポとその修正後（だと想定される）文字列である [4, 3]．しかし，最終出力文字列に残らない，入力中に修正されるタイポも存在する．このようなタイポデータを用いることでスペル訂正エンジンの精度を向上させられると我々は考えた．

しかし，「入力中に修正されたタイポ」が含まれる公開データは存在しない．そこで本研究では，まずクラウドソーシングサービス Amazon Mechanical Turk（MTurk）を利用して，複数のユーザに文章を入力させ，BS キー操作を含むユーザの入力文字列（キーストローク）を収集した（3 章）．対象言語は英語と日本語とした．図 1 にキーストロークの例を示す．

次に，収集したキーストローク中に含まれるタイポと，最終出力文字列に現れることの多い「一般的なタイポ」を分析・比較し，キーストロークには独自の傾向のタイポが現れることを示した（4.3 節）．この事実により，特にユーザの入力中に訂正候補を提示するオンラインスペル訂正エンジンにおいて，キーストローク利用による精度向上が期待できる．また，日英のキーストロークに含まれるタイポを比較し，いくつかの面で日本語独自のタイポ傾向があることを示した（4.4 節）．さらに，既存のタイポ分析では取り上げられていないいくつかのタイポ要因を指摘した（4.5 節）．

2　関連研究

タイポの分析はいくつか行われている．荒牧らは Twitter から収集したデータ中で，編集距離 1 となる低頻度語と高頻度語のペアをそれぞれタイポ候補・原型候補として，タイポの要因と思われる 5 つの要素のうち，どれが主要な要因であるかを分析した [7]．この研究で対象としているのは Twitter 上の最終出力文字列であり，本研究のような入力中のタイポは対象としていない．また，本研究はユーザによって修正されたタイポを獲得しており高頻度語もタイポ候補にできるが，この研究では低頻度語しかタイポ候補にできない．また，対象言語は英語のみである．

図 1: キーストロークの例

Zheng らは，中国語 IME である Sogou の入力ログを収集し，入力された漢字と BS キー操作を獲得して中国語におけるタイポの分析を行った [6]．また英語の一般的なタイポに関しても分析をした．彼らは本研究と同じく BS キー操作に着目しているが，ピンインに関してはキーストロークを直接取得するのではなく，入力された漢字から対応するピンインに戻して分析を行なっている．よって，実際のキーストロークに着目した分析は本研究が初めてだと言える．

3　キーストローク収集

今回キーストローク収集で用いた MTurk は，コンピュータにとっては難しいタスクを人間（ワーカーと呼ばれる）に依頼するためのウェブサービスである．近年，様々な自然言語処理タスクにおいてアノテーションのために広く利用されていて，MTurk からのデータ収集自体が研究対象となっている [5]．2010 年には国際会議 NAACL において MTurk からのデータ収集を対象としたワークショップが開催された [1]．しかし，MTurk からユーザのキーストロークを収集した既存研究はなく，この点でも本研究は新しい問題に取り組んだと言える．表 1 に，収集したデータの概要を示す．

3.1　タスクデザイン

キーストロークを収集するためのタスクとして，ワーカーに画像を提示し「画像への説明文を記述するタス

*本研究は，筆頭著者の Microsoft Research でのインターンシップ中に行われた

図 2: タスクで用いた画像と得られた文章例

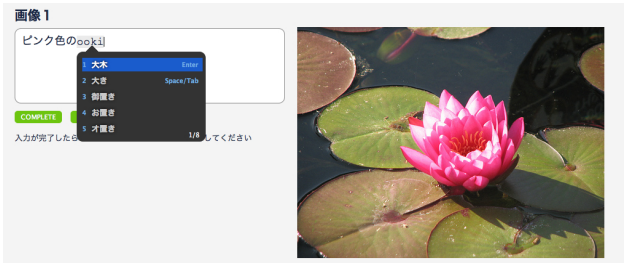

図 3: タスクインターフェースのスクリーンショット

ク」と「画像中の登場人物のセリフを記述するタスク」の二種類を用意した．対象言語を気にせず利用できるという利便性から，今回は画像を文章入力のトリガーとして選択した．利用する画像は，写真共有サービス Flickr の Your Best Shot 2009/2010[1][2] グループから選択した．図 2 に，それぞれのタスクで使用した画像の例と，ワーカーによって記述された英語・日本語の文章例を示す．

3.2 タスクインターフェース

図 3 に今回用いたタスクのインターフェースを示す．ワーカーにはキーストロークを取得していることは隠して，画像に対する文章を記述させた．キーストローク中で入力を修正した箇所の特定を容易にするため，インターフェースにおいてはマウスや矢印キーでのカーソル移動・文字列選択を受け付けないようにし，修正する際には BS キーを必要回数押させるようにした．また，スペル訂正機能は利用できないようにした．

日本語タスクにおいて，ワーカーが用いる IME にサジェスト機能がついていると，入力する文字全てのキーストロークを取得できない．そこで，Universal Text Input[8] を利用してワーカーが利用する IME をサジェスト機能のないものに統一した．

3.3 クオリティコントロール

MTurk では，短時間で多くの報酬を得るためにチート行為を行うワーカーが存在する．単純なチート行為を防ぐため，一定文字数以上入力しないとタスクを完了させられないようにし，またペースト操作を利用できないようにした．タスク実行結果は人手で確認をし，単語のみの入力などは却下した．

また，質の高いデータを多く集めるための工夫として Chen らが用いた二段階報酬制を採用した [2]．誰でもアクセスできるタスクとは別に，高額の報酬が支払われる招待制のタスクを用意し，長い文章や語彙の豊富な文章を書いたワーカーを招待した．

表 1: 収集データ情報．文の数は，一画像に対して一文が入力されたとして集計した．日本語の単語の数は，IME の変換区切りを単語区切りとして集計した．

	英語	日本語
文の数	50,000	5,985
BS 操作を含む文の数	34,008	4,157
単語の数	546,339	34,576
BS 操作を含む単語の数	75,011	8,195
編集距離 2 以内の修正前後単語ペア数	44,378	4,838
結果が採用されたユニークワーカー数	677	11

4 タイポ分析

本節では，英語・日本語キーストローク中のタイポ及び一般的な英語タイポがどのような発生傾向を持っているのか分析をし，各データ同士の比較を行う．

4.1 分析対象タイポ

英語・日本語キーストロークタイポ (en_keystroke, ja_keystroke)
各言語のキーストロークから，単語入力中に BS キーが押されたものだけを獲得した．キーストロークから修正前後の単語ペアを取得し，全て小文字化した．修正前後の編集距離が 2 以内の，英語 44,378 ペア，日本語 4,838 ペアを分析対象とした．

一般的な英語タイポ (en_common)
既存研究 [6] と同様に，Wikipedia[3] と SpellGood[4] から修正前後の単語ペアを取得した．2 つのデータ中で重複しているものは省き，アルファベット以外の文字やスペースが含まれているものは除去した．全て小文字化し，修正前後の編集距離が 2 以内の 10,609 ペアを分析対象とした．

4.2 分析項目

既存研究 [6, 7] では，(1) エラータイプ (Deletion：削除, Insertion：挿入, Substitution：他の文字への置換, Transposition：単語内での文字順序の入れ替り)，(2) 単語内でのタイポ発生位置，(3) 単語内で文字が複数回現れているか，(4) キーを押す指／手，(5) キー同士の距離，(6) 文字が母音／子音のいずれであるか，(7) 文字同士の視覚的類似性，(8) 文字同士の音韻的類似性，といった観点で分析が行われている．

本研究では，音韻的類似性以外の各項目について分析を行った．本稿では特に，en_keystroke と en_common を比較した結果，en_keystroke と ja_keystroke を比較した結果について述べる．また，3 つのデータに共通する，これまで着目されていなかったタイポ発生要因を指摘する．

[1] http://www.flickr.com/groups/yourbestshot2009/
[2] http://www.flickr.com/groups/yourbestshot-2010/
[3] http://en.wikipedia.org/wiki/Wikipedia:Lists_of_common_misspellings/For_machines
[4] http://www.spellgood.net/sitemap.html

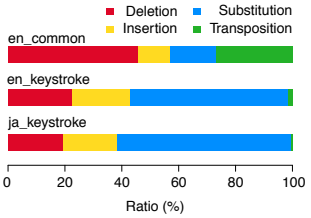

図 4: 各データセット内のエラーカテゴリ

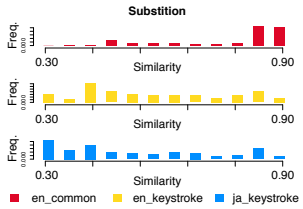

図 5: 文字同士の視覚的類似度ごとの Substitution 発生頻度

なお，以降ではある事象の発生頻度を元に議論を行なっていくが，頻度は「タイポにおける出現回数／データ中での出現回数」の割合として取り扱う．たとえば，Deletion エラーにおける子音の発生頻度は，「削除された文字が子音である回数／データ中での子音の出現回数」とする．各分析結果の図では，頻度を各データ中での頻度最大値で割った値を示した．

4.3 英語のキーストロークタイポと一般的な英語タイポの比較

英語キーストロークタイポと一般的な英語タイポを比較することで，キーストローク独自の傾向を明らかにしていく．この二つのデータを比較することで，どのようなタイポはユーザによって修正されやすいのか，または修正されずに残ってしまうのかがわかるため，この点についても言及する．

4.3.1 エラータイプの出現割合（図 4）

en_keystroke では Substitution が多いが，en_common では Deletion が多い[5]．この結果からは，Substitution は気づきやすいタイポでありユーザが入力中に修正するが，Deletion は気づきにくいタイポで最終出力文字列に残りやすいということが予想される．

4.3.2 単語内でのエラーの発生位置（図 6）

en_keystroke については，Deletion が語の先頭で多く，Insertion と Substitution が語の先頭・末尾の両方で多い．en_common では，どのエラータイプでも語の中頃での発生頻度がやや大きい．

以上から，語の先頭での Deletion，語の末尾での Insertion はユーザが気がつきやすくその場で修正されるが，語の中頃での Deletion や Insertion は気づきにくく最終文書まで残るという傾向がわかる[6]．キーストロークでの語の先頭の Substitution は，別の語に打ち直している場合もあるために多く発生しているとも考えられる．

4.3.3 Substitution での母音／子音の傾向（図 7）

en_keystroke では子音 → 子音の Substitution，母音 → 母音の Substitution の頻度に大きな差はないが，

[5]キーストロークでの，単語冒頭での Substitution は別の単語への打ち直しの可能性もあるが，単語冒頭での Substitution を集計から除いても Substitution が一番多いことに変わりはなかった．

[6]なお，キーストロークでの語の末尾での Insertion は，スペースの打ち忘れもあるが，`seas→sea` のような単複数形の修正，`nervouse→nervous` のような通常のスペル修正も多く発生している．

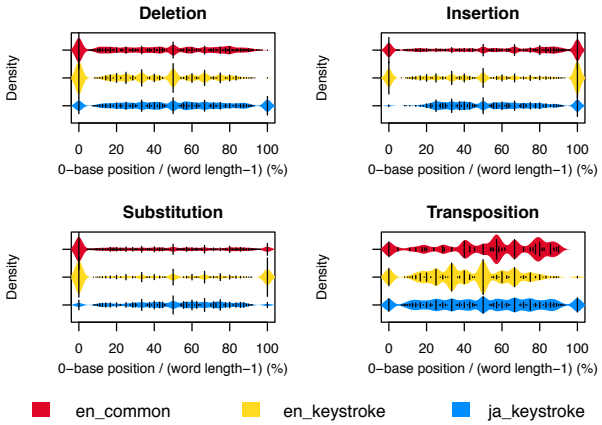

図 6: 単語内での各エラー発生位置

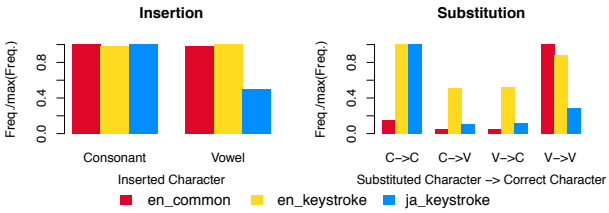

図 7: 母音／子音での Insertion, Substitution 発生頻度

en_common では母音 → 母音の Substitution が多い．子音同士の Substitution（例．`eazy→easy`）には気がつきやすいが，母音が他の母音に変わった場合（例．`visable→visible`）には気づきづらいという傾向が伺える．

4.3.4 文字が連続する箇所での Deletion（図 8）

例えば，`tomorow→tomorrow` は，連続する箇所での Deletion である．en_keystroke では，連続する箇所と非連続箇所での Deletion の頻度にあまり差がないが，en_common では，連続する箇所での Deletion の方が明らかに頻度が高い．連続する箇所での Deletion には気づきづらいと予想できる．

4.3.5 視覚的類似度と Substitution（図 5）

視覚的類似度は荒牧らと同様に，「2 ×文字 A と文字 B のオーバーラップする面積／（文字 A の面積＋文字 B の面積）」で算出した [7]．en_common では視覚的類似度が高い文字同士の Substitution が高頻度となっているが en_keystroke ではこの傾向はない．見た目が似ている文字同士（例．`yoqa→yoga`）の Substitution は最終出

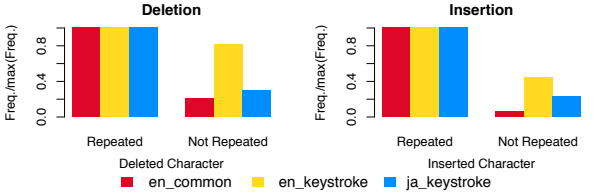

図 8: 文字が連続／非連続の箇所での Deletion 発生頻度，同じ文字を連続で入力する Insertion の発生頻度

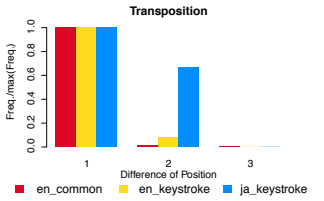

図 9: 文字同士の単語内位置差による Transposition 発生頻度

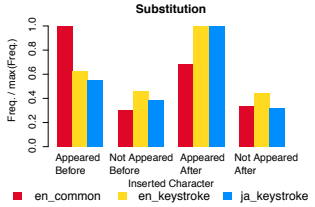

図 10: 置き換わった文字が単語内の前／後に登場している場合の Substitution 発生頻度

力文書で残りやすいと言える．

4.4 英語・日本語キーストロークタイポの比較

4.4.1 Transposition で，入れ替わる文字の単語内での位置の差（図 9）

差が 1 の場合，単語内で隣り合った文字同士での入替となる．en_keystroke では隣り合った文字同士の Transposition が圧倒的に多いが，ja_keystroke では 1 の場合と 2 の場合で大きな差がない．日本語では `airodu→aidoru` のような，ひらがな又はカタカナ単位での入れ替りが発生しやすい．特に，`kotoro→tokoro` のように母音が一致するかな同士での子音の入れ替りが多く発生しており，ja_keystroke で単語内位置の差が 2 の Transposition のうち，73%がこのケースに当てはまる．

4.4.2 Insertion での母音／子音（図 7）

ja_keystroke では子音の Insertion が母音よりも発生しやすいという結果が得られた．原因としては，日本語入力において子音を連続で入力することの方が，母音を連続で入力することよりも多いことが考えられる．ja_keystroke 中では，子音の連続は母音の連続の 4.03 倍の回数出現しており，日本語入力者にとって子音の連続入力の方が母音の連続よりも慣れた動作であると言える（なお，en_keystroke では子音の連続の出現回数は母音の連続の 1.84 倍である）．図 8 に示した通り，同じ文字を連続で打ってしまう Insertion 発生頻度は高く，日本語において子音の方が連続しやすいという特徴から子音での Insertion 頻度が高くなっていると考えられる．

4.4.3 Substitution での母音／子音（図 7）

ja_keystroke では子音 → 子音のみ他のケースよりも頻度が高いが，en_keystroke では，子音 → 子音と母音 → 母音が共に高頻度である．ja_keystroke で母音 → 母音の頻度が高くない要因として，日本語では母音によってひらがなが変わり母音を間違えにくいことが挙げられる．

4.5 既存研究では指摘されていなかったタイポの傾向

4.5.1 左手／右手の切り替えと Deletion, Insertion

Deletion, Insertion の発生箇所とその前後のキーを押す手を L, R で表現すると LLR, RRR が多いという傾向がいずれのデータでも観測された．この現象は，Deletion であれば左手から右手に入力する手を切り替える際に左手で打つ最後のキーを飛ばしてしまう傾向と，右手で連続でキーを打つときに間のキーを打ち忘れてしまう傾向だと解釈できる．また，Insertion であれば，左手から右手に切り替えるときに余分なキーを左手で押してしまうこと，右手で連続で押すときに右手で入力する他のキーを誤って打ってしまうこととなる．特に LLR のケースでは，LL が同じ指であることが多い．

4.5.2 Substitution での子音の先取り／後取り（図 10）

Substitution において，`puclic→public` のように，単語内でこれから出現する子音（先取り），または以前に出現した子音を間違えて打ってしまう（後取り）という現象が多く観測された．特に，先取りの発生頻度は先取りではない場合と比べてかなり大きい．この現象は，人間が文字入力をする際に，現在打っている文字だけではなく，後に打つ文字のことを頭に浮かべながら入力を行なっているために発生すると考えられる．

5 おわりに

MTurk を利用して BS キー操作を含むユーザのキーストロークを英語・日本語について収集した．収集した英語キーストロークと，最終出力文字列に現れることの多い「一般的なタイポ」を比較し，キーストロークが持つ独自の傾向について明らかにした．また日英のキーストロークを比較し，日本語タイポがいくつかの面で独自の傾向を持つことを示した．さらに，既存研究では指摘されていなかったタイポ発生要因を指摘した．

参考文献

[1] C. Callison-Burch and M. Dredze. Creating speech and language data with amazon's mechanical turk. In *Proc. of NAACL HLT 2010 Workshop*, 2010.

[2] D. L. Chen and W. B. Dolan. Collecting highly parallel data for paraphrase evaluation. In *Proc. of ACL '11*, 2011.

[3] J. Gao, X. Li, D. Micol, C. Quirk, and X. Sun. A large scale ranker-based system for search query spelling correction. In *Proc. of the COLING '10*, 2010.

[4] M. Kernighan, K. Church, and W. Gale. A spelling correction program based on a noisy channel model. In *Proc. of COLING '90*, 1990.

[5] R. Snow, B. O'Connor, D. Jurafsky, and A. Ng. Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proc. of EMNLP '08*, 2008.

[6] Y. Zheng, L. Xie, Z. Liu, M. Sun, Y. Zhang, and L. Ru. Why press backspace? understanding user input behaviors in chinese pinyin input method. In *Proc. of ACL '11*, 2011.

[7] 荒牧英治, 宇野良子, 岡瑞起. Typo writer: ヒトはどのように打ち間違える のか？ 言語処理学会第 16 回年次大会 (NLP2010), 2010.

[8] 鈴木久美, P. Choudhury, C. Quirk, C. Wendt, C. Yu, A. Mohammed, V. Dendi. 入力支援機能を統合した多言語入力システム「universal text input」. 言語処理学会第 18 回年次大会 (NLP2012), 2012.