How Are Spelling Errors Generated and Corrected? A Study of Corrected and Uncorrected Spelling Errors Using Keystroke Logs

Yukino Baba
The University of Tokyo
yukino.baba@gmail.com

Hisami Suzuki
Microsoft Research
hisamis@microsoft.com

Abstract

This paper presents a comparative study of spelling errors that are corrected as you type, vs. those that remain uncorrected. First, we generate naturally occurring online error correction data by logging users' keystrokes, and by automatically deriving pre- and post-correction strings from them. We then perform an analysis of this data against the errors that remain in the final text as well as across languages. Our analysis shows a clear distinction between the types of errors that are generated and those that remain uncorrected, as well as across languages.

1 Introduction

When we type text using a keyboard, we generate many spelling errors, both typographical (caused by the keyboard layout and hand/finger movement) and cognitive (caused by phonetic or orthographic similarity) (Kukich, 1992). When the errors are caught during typing, they are corrected on the fly, but unnoticed errors will persist in the final text. Previous research on spelling correction has focused on the latter type (which we call **uncorrected errors**), presumably because the errors that are corrected on the spot (referred to here as **corrected errors**) are not recoded in the form of a text. However, studying corrected errors is important for at least three reasons. First, such data encapsulates the spelling mistake and correction by the author, in contrast to the case of uncorrected errors in which the intended correction is typically assigned by a third person (an annotator), or by an automatic method (Whitelaw et al., 2009; Aramaki et al., 2010)[1]. Secondly, data on corrected errors will enable us to build a spelling correction application that targets correction on the fly, which directly reduces the number of keystrokes in typing. This is crucial for languages that use transliteration-based text input methods, such as Chinese and Japanese, where a spelling error in the input Roman keystroke sequence will prevent the correct candidate words from appearing in the list of candidates in their native scripts, thereby preventing them from being entered altogether. Finally, we can collect a large amount of spelling errors and their corrections by logging keystrokes and extracting the pre- and post-correction strings from them. By learning the characteristics of corrected and uncorrected errors, we can expect to use the data for improving the correction of the errors that persisted in the final text as well.

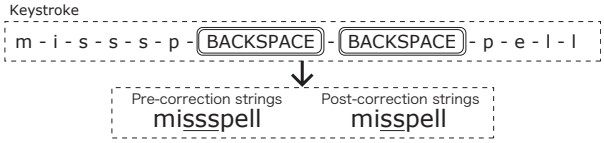

Figure 1: Example of keystroke

In this paper, we collect naturally occurring spelling error data that are corrected by the users online from keystroke logs, through the crowdsourcing infrastructure of Amazon's Mechanical Turk (MTurk). As detailed in Section 3, we display images to the worker of MTurk, and collect the descriptions of these images, while logging their keystrokes including the usage of backspace keys, via a crowd-based text input service. We collected logs for two typologically different languages, English and Japanese. An example of a log along with the extracted pre- and post-correction strings is shown in Figure 1. We then performed two comparative analyses: corrected vs. uncorrected errors in English (Section 4.3), and English vs. Japanese corrected errors (Section 4.4). Finally, we remark on an additional cause of spelling errors observed in all the data we analyzed (Section 4.5).

2 Related Work

Studies on spelling error generation mechanisms are found in earlier work such as Cooper (1983). In particular, Grudin (1983) offers a detailed study of the errors generated in the transcription typing scenario, where the subjects are asked to transcribe a text without correcting the errors they make. In a more recent work, Aramaki et al. (2010) automatically extracted error-correction candidate pairs from Twitter data based on the assumption that these pairs

[1] Using web search query logs is one notable exception, which only targets spelling errors in search queries (Gao et al., 2010)

Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, pages 373–377,
Jeju, Republic of Korea, 8-14 July 2012. ©2012 Association for Computational Linguistics

fall within a small edit distance, and that the errors are not in the dictionary and substantially less frequent than the correctly spelled counterpart. They then studied the effect of five factors that cause errors by building a classifier that uses the features associated with these classes and running ablation experiments. They claim that finger movements cause the spelling errors to be generated, but the uncorrected errors are characterized by visual factors such as the visual similarity of confused letters. Their experiments however target only the persisted errors, and their claim is not based on the comparison of generated and persisted errors.

Outside of English, Zheng et al. (2011) analyzed the keystroke log of a commercial text input system for Simplified Chinese, and compared the error patterns in Chinese with those in English. Their use of the keystroke log is different from ours in that they did not directly log the input in pinyin (Romanized Chinese by which native characters are input), but the input pinyin sequences are recovered from the Chinese words in the native script (hanzi) after the character conversion has already applied.

3 Keystroke Data Collection

Amazon's Mechanical Turk (MTurk) is a web service that enables crowdsourcing of tasks that are difficult for computers to solve, and has become an important infrastructure for gathering data and annotation for NLP research in recent years (Snow et al. 2008). To the extent of our knowledge, our work is the first to use this infrastructure to gather user keystroke data.

3.1 Task design

In order to collect naturally occurring keystrokes, we have designed two types of tasks, both of which consist of writing something about images. In one task type, we asked the workers to write a short description of images (image description task); in the other, the workers were presented with images of a person or an animal, and were asked to guess and type what she/he was saying (let-them-talk task). Using images as triggers for typing keeps the underlying motivation of keystroke collection hidden from the workers, simultaneously allowing language-independent data collection. For the image triggers, we used photos from the Flickr's Your Best Shot 2009/2010 groups . Examples of the tasks and collected text are given in Figure 2.

Figure 2: Examples of tasks and collected text (Translated text: "A flock of penguines are marching in the snow." and "Mummy, my feet can't touch the bottom.")

3.2 Task interface

For logging the keystrokes including the use of backspaces, we designed an original interface for the text boxes in the MTurk task. In order to simplify the interpretation of the log, we disabled the cursor movements and text highlighting via a mouse or the arrow keys in the text box; the workers are therefore forced to use the backspace key to make corrections. In Japanese, many commercially available text input methods (IMEs) have an auto-complete feature which prevents us from collecting all keystrokes for inputting a word. We therefore used an in-house IME that has disabled this feature to collect logs. This IME is hosted as a web service, and keystroke logs are also collected through the service. For English, we used the service for log collection only.

4 Keystroke Log Analysis

4.1 Data

We used both keystroke-derived and previously available error data for our analysis.

Keystroke-derived error pairs for English and Japanese (en_keystroke, ja_keystroke): from the raw keystroke logs collected using the method described in Section 3, we extracted only those words that included a use of the backspace key. We then recovered the strings before and after correction by the following steps (Cf. Figure 1):

- To recover the post-correction string, we deleted the same number of characters preceding a sequence of backspace keys.
- To recover the pre-correction string, we compared the prefix of the backspace usage (`misssp` in Figure 1) with the substrings after error correction (`miss`, `missp`, ⋯, `misspell`), and considered that the prefix was spell-corrected into the substring which is the longest and with the smallest edit distance

(in this case, `misssp` is an error for `missp`, so the pre-correction string is `missspell`).

We then lower-cased the pairs and extracted only those within the edit distance of 2. The resulting data which we used for our analysis consists of 44,104 pairs in English and 4,808 pairs in Japanese[2].

Common English errors (en_common): following previous work (Zheng et al., 2011), we obtained word pairs from Wikipedia[3] and SpellGood[4]. We lower-cased the entries from these sources, removed the duplicates and the pairs that included non-Roman alphabet characters, and extracted only those pairs within the edit distance of 2. This left us with 10,608 pairs.

4.2 Factors that affect errors

Spelling errors have traditionally been classified into four descriptive types: Deletion, Insertion, Substitution and Transposition (Damerau, 1964). For each of these types, we investigated the potential causes of error generation and correction, following previous work (Aramaki et al., 2010; Zheng et al., 2011). Physical factors: (1) motor control of hands and fingers; (2) distance between the keys; Visual factors: (3) visual similarity of characters; (4) position in a word; (5) same character repetition; Phonological factors: (6) phonological similarity of characters/words.

In what follows, our discussion is based on the frequency ratio of particular error types, where the frequency ratio refers to the number of cases in spelling errors divided by the total number of cases in all data. For example, the frequency ratio of consonant deletion is calculated by dividing the number of missing consonants in errors by the total number of consonants.

4.3 Corrected vs. uncorrected errors in English

In this subsection, we compare corrected and uncorrected errors of English, trying to uncover what factors facilitate the error correction.

Error types (Figure 3) Errors in en_keystroke are dominated by Substitution, while Deletion errors are the most common in en_common, indicating that

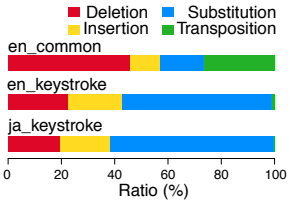

Figure 3: Ratios of error types

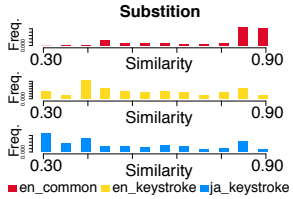

Figure 4: Visual similarities of characters in substitution errors

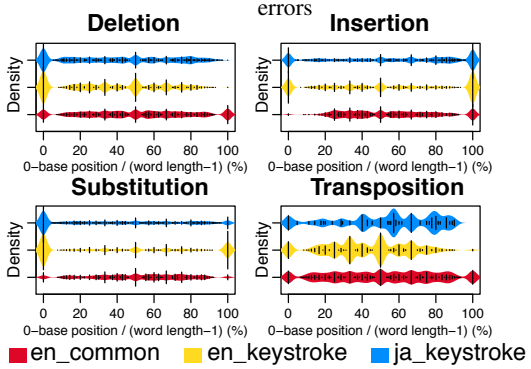

Figure 5: Positions of errors within words

Substitution mistakes are easy to catch, while Deletion mistakes tend to escape our attention. Zheng et al. (2011) reports that their pinyin correction errors are dominated by Deletion, which suggests that their log does in fact reflect the characteristics of corrected errors.

Position of error within a word (Figure 5) In en_keystroke, Deletion errors at the word-initial position are the most common, while Insertion and Substitution errors tend to occur both at the beginning and the end of a word. In contrast, in en_common, all error types are more prone to occur word-medially. This means that errors at word edges are corrected more often than the word-internal errors, which can be attributed to cognitive effect known as the bathtub effect (Aitchison, 1994), which states that we memorize words at the periphery most effectively in English.

Effect of character repetition (Figure 6) Deletion errors where characters are repeated, as in `tomorow`→`tomorrow`, is observed significantly more frequently than in a non-repeating context in en_common, but no such difference is observed in en_keystroke, showing that visually conspicuous errors tend to be corrected.

Visual similarity in Substitution errors (Figure 4) We computed the visual similarity of characters by $2\times(the\ area\ of\ overlap\ between\ character\ A\ and\ B)/(area\ of\ character\ A+area\ of\ character\ B)$ follow-

[2]The data is available for research purposes under `http://research.microsoft.com/research/downloads/details/4eb8d4a0-9c4e-4891-8846-7437d9dbd869/details.aspx`

[3]`http://en.wikipedia.org/wiki/Wikipedia:Lists_of_common_misspellings/For_machines`

[4]`http://www.spellgood.net/sitemap.html`

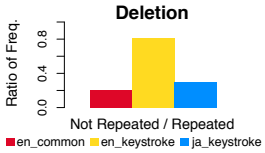

Figure 6: Effect of character repetition in Deletion

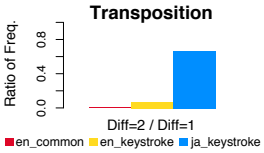

Figure 7: Difference of positions within words in Transposition

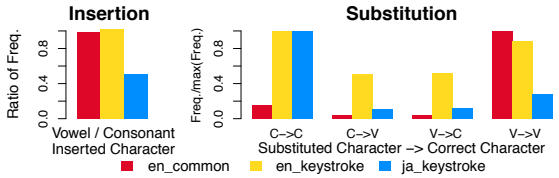

Figure 8: Consonants/vowels in Insertion and Substitution

ing Aramaki et al. (2010)[5]. Figure 4 shows that in en_common, Substitution errors of visually similar characters (e.g., `yoqa→yoga`) are in fact very common, while in en_keystroke, no such tendency is observed.

Phonological similarity in Substitution errors (Figure 8) In en_keystroke, there is no notable difference in consonant-to-consonant (C→C) and vowel-to-vowel (V→V) errors, but in en_common, V→V errors are overwhelmingly more common, suggesting that C→C can easily be noticed (e.g., `eazy→easy`) while V→V errors (e.g., `visable→visible`) are not. This tendency is consistent with the previous work on the cognitive distinction between consonants and vowels in English: consonants carry more lexical information than vowels (Nespor et al., 2003), a claim also supported by distributional evidence (Tanaka-Ishii, 2008). It may also be attributed to the fact that English vowel quality is not always reflected by the onthography in the straightforward maner.

Summarizing, we have observed both visual and phonological factors affect the correction of errors. Aramaki et al. (2010)'s experiments did not show that C/V distinction affect the errors, while our data shows that it does in the correction of errors.

4.4 Errors in English vs. Japanese

From Figure 3, we can see that the general error pattern is very similar between en_keystroke and ja_keystroke. Looking into the details, we discovered some characteristic errors in Japanese, which are phonologically and orthographically motivated.

Syllable-based transposition errors (Figure 7) When comparing the transposition errors by their distance, 1 being a transposition of adjacent characters and 2 a transposition skipping a character, the instances in en_keystroke are mostly of distance of 1, while in ja_keystroke, the distance of 2 also occurs commonly (e.g., `kotoro→tokoro`). This is interesting, because the Japanese writing system called kana is a syllabary system, and our data suggests that users may be typing a kana character (typically CV) as a unit. Furthermore, 73% of these errors share the vowel of the transposed syllables, which may be serving as a strong condition for this type of error.

[5]We calculated the area using the Courier New font which we used in our task interface.

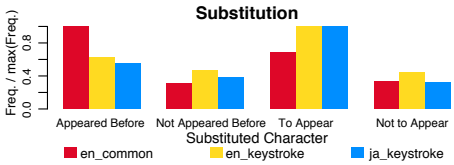

Figure 9: Look-ahead and Look-behind in Substitution

Errors in consonants/vowels (Figure 8) Errors in ja_keystroke are characterized by a smaller ratio of insertion errors of vowels relative to consonants, and by a relatively smaller ratio of V→V substitution errors. Both point to the relative robustness of inputting vowels as opposed to consonants in Japanese. Unlike English, Japanese only has five vowels whose pronunciations are transparently carried by the orthography; they are therefore expected to be less prone to cognitive errors.

4.5 Look-ahead and look-behind errors

In Substitution errors for all data we analyzed, substituting for the character that appeared before, or are to appear in the word was common (Figure 9). In particular, in en_keystroke and ja_keystroke, look-ahead errors are much more common than non-look-ahead errors. Grudin (1983) reports cases of permutation (e.g., `gib→big`) but our data includes non-permutation look-ahead errors such as `puclic→public` and `otigaga→otibaga`.

5 Conclusion

We have presented our collection methodology and analysis of error correction logs across error types (corrected vs. uncorrected) and languages (English and Japanese). Our next step is to utilize the collected data and analysis results to build online and offline spelling correction models.

Acknowledgments

This work was conducted during the internship of the first author at Microsoft Research. We are grateful to the colleagues for their help and feedback in conducting this research.

References

Aitchison, J. 1994. *Words in the Mind*. Blackwell.

Aramaki, E., R. Uno and M. Oka. 2010. TYPO Writer: ヒトはどのように打ち間違えるのか？ (TYPO Writer: how do humans make typos?). In *Proceedings of the 16th Annual Meeting of the Natural Language Society (in Japanese)*.

Cooper, W. E. (ed.) 1983. *Cognitive Aspects of Skilled Typewriting*. Springer-Verlag.

Damerau, F. 1964. A technique for computer detection and correction of spelling errors. *Communications of the ACM 7(3): 659-664*.

Gao, J., X. Li, D. Micol, C. Quirk and X. Sun. 2010. A large scale ranker-based system for search query spelling correction. In *Proceedings of COLING*.

Grudin, J. T. 1983. Error patterns in novice and skilled transcription typing. In Cooper, W.E. (ed.), *Cognitive Aspects of Skilled Typewriting*. Springer-Verlag.

Kukich, K. 1992. Techniques for automatically correcting words in text. In *ACM Computing Surveys, 24(4)*.

Nespor, M., M. Peña, and J. Mehler. 2003. On the different roles of vowels and consonants in speech processing and language acquisition. *Lingue e Linguaggio*, pp. 221—247.

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

Tanaka-Ishii, K. 2008. 単語に内在する情報量の偏在 (On the uneven distribution of information in words). In *Proceedings of the 14th Annual Meeting of the Natural Language Society (in Japanese)*.

Whitelaw, Casey, Ben Hutchinson, Grace Y. Chung, and Gerard Ellis. 2009. Using the web for language independent spellchecking and autocorrection. In *Proceedings of ACL*.

Zheng, Y., L. Xie, Z. Liu, M. Sun. Y. Zhang and L. Ru 2011. Why press backspace? Understanding user input behaviors in Chinese pinyin input method. In *Proceedings of ACL*