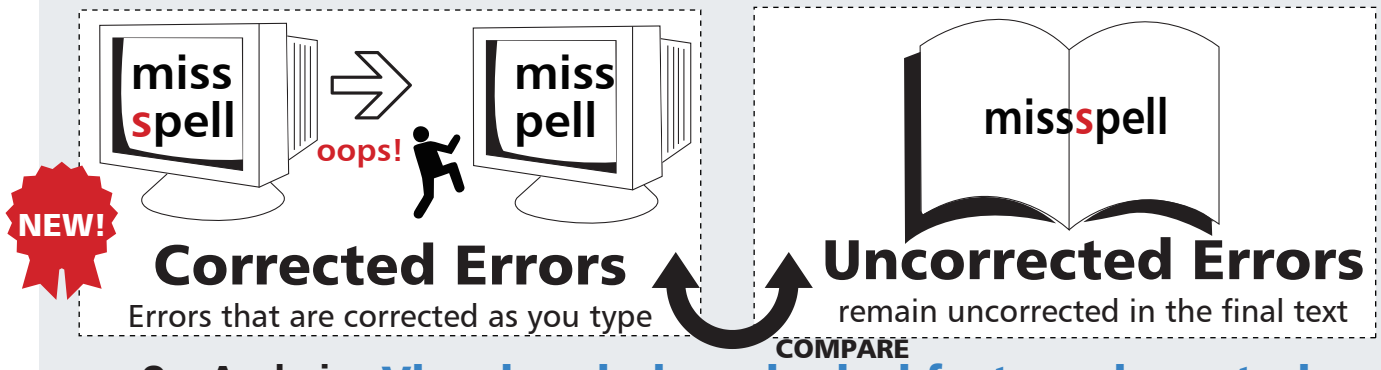
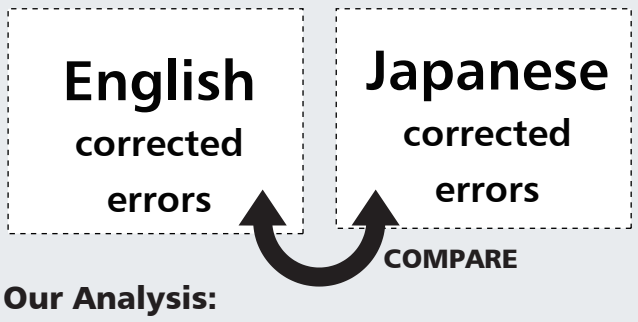
How Are Spelling Errors Generated and Corrected? A Study of Corrected and Uncorrected Spelling Errors Using Keystroke Logs

Yukino Baba (The University of Tokyo)
Hisami Suzuki (Microsoft Research)

SUMMARY



Our Analysis: **Visual and phonological factors characterize the differences**

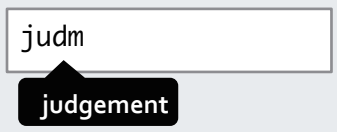


Our Analysis:
Phonological differences in EN and JP affects the characteristics of the errors

MOTIVATION

Understanding how people generate and correct spelling errors will enable us to build...

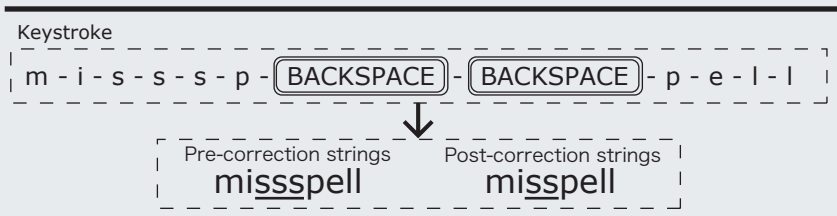
Online spelling correction



Fuzzy Transliteration based text input method

DATA COLLECTION

Corrected Errors



1. Collected users' key-strokes including the use of backspace keys via MTurk
2. Obtained pre- and post-correction strings from keystrokes

Uncorrected Errors

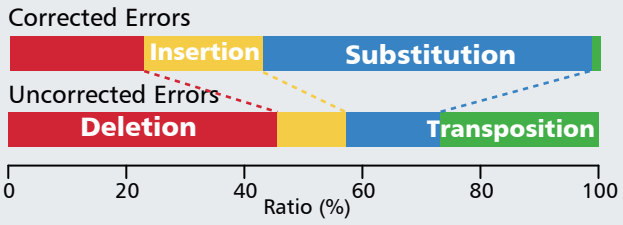
Obtained from Wikipedia: Lists of common misspellings and Spellgood.net (10K pairs)

44K for EN and 5K for JP, available online!

ANALYSIS

Corrected vs. Uncorrected Errors in EN

1. Error Types

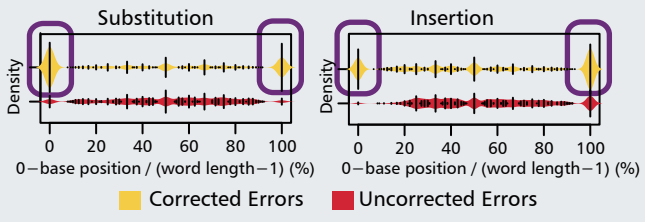


The most common error type

Corrected Errors	Uncorrected Errors
Substitution	**Deletion**

Substitution errors are easy to catch, **Deletion** errors are not

2. Position of error



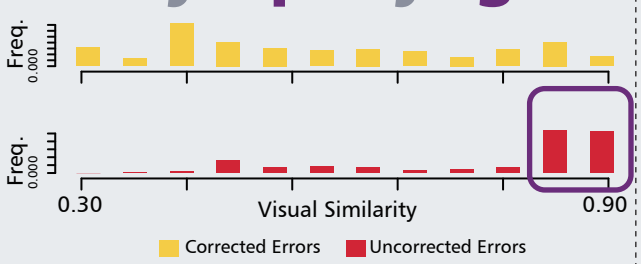
Common positions of error

Corrected Errors	Uncorrected Errors
at word edges	in the middle

Errors both at the begining and the end of a word are corrected more often

3. Visual similarity in Substitution

e.g. **yoqa→yoga**



Corrected Errors	Uncorrected Errors
no such tendency	Substitution Errors of **visually similar characters** are common

4. Effect of character repetition

e.g. **tomor ow→ tomorrow**

Uncorrected Errors:
Deletion Errors where **characters are repeated** is observed frequently

5. Phonological similarity in Substitution errors

V→V vowel-to-vowel is more common than **C→C** consonant-to-consonant **in Uncorrected Errors**

e.g. **visable→visible** e.g. **eazy→easy**

Errors in EN and JP

1. Syllable-based transposition errors

EN **teh→the**

transpositions of adjacent characters are common

JP **kotoro→tokoro**

Syllable-based transposition errors occur commonly

2. Errors in consonants/vowels

JP **V→V** errors are relatively uncommon than C→C

EN no notable difference

Look-ahead and look-behind errors

puclic→public Look-ahead

gigl→girl Look-behind

ACKNOWLEDGMENTS

This work was conducted during the internship of the first author at Microsoft Research