The SHCTextCorrections Spreadsheet

# Introduction

This memo accompanies the spreadsheet SHCTextcorrections .xlsx and explains what it is about.

The spreadsheet is a list of ~45,000 textual changes made to the the TCP transcriptions of 520 non-Shakespearean plays written between 1550 and 1650--a very large majority of all the plays that have survived from that period. This is part of a project that I call "SHC" or "Shakespeare His Contemporaries", where "contemporaries" is defined broadly as playwrights working from the generation before to the generation after Shakespeare's active carer from ~1588 to 1612.

The list is a cumulative record of curatorial labour stretching over several years. Can that cumulative record become a useful input for machine learning to routines so that some of the very tedious work of basic error correction can be "downsourced" to a machine? My informal analysis suggests that the answer to this is "yes". About a quarter of simple textual defects in the much larger corpus of the 50,000 EEBO-TCP texts can be fixed algorithmically. That's well over a million little things, and as the Scots say "many a mickle makes a muckle."

The TCP transcriptions were created by commercial vendors employing individuals, mostly in Asian countries, to produce accurate transcription through double keyboarding. The data entry clerks worked from digital scans of microfilm images of printed texts. Many things can and do go wrong in that chain of transmission. The texts may be poorly printed to begin with (too much or not enough ink). The microfilm image may "skew" text at the inner margins of a page. And so on.

# Known unknowns

The transcribers entered what they saw, and they used special codes to describe what they could not decipher. Call these the "known unknowns." The representation of those codes in the spreadsheet is as follows:

1. The black circle ● (\u25cf) represents a single missing letter
2. The black small square ▪ (\u25aa) represents a punctuation mark that could not be further identified as a comma , period, etc.
3. The white diamond inside angle brackets 〈◇〉(\u3008, \u25c7, \ u3009) represents a missing word
4. The elision sign inside angle brackets 〈…〉(\u3008, \u2026, \ u3009) represents a short span of text of indeterminate length, which may begin or end inside a word

The transcribers were meticulous in their marking of "known unknowns," but they were not always right in gussing the number of missing letters. A single black circle is by far the most error code in the corpus of some 50,000 texts, and in ~95% of cases it accurately identifies the presence of a single undeciphipherable character.

In 2013 five Northwestern undergraduates on summer internships spent eight weeks working their way through the plays, comparing the transcriptions with the digital scans and seeking to decipher the "known unknowns" by looking more closely at the scans and bringing to that task their growing and non-trivial knowledge of the conventions of Early Modern drama.[[1]](#footnote-1) I reviewed their work and have continued it. The spreadsheet is the cumulative record of that work.

Whereas the transcribers were instructed to record what they saw on the page, the undergraduates were instructed to record what should have been on the printed page to begin with. Early Modern printing was a very error-prone business, and the printers were well aware of it, as witnessed by the many "errata" pages and their often rueful and whimsical prefaces. Thus in the string "a baskit m●●er" , it does not matter whether the two missing letters were illegible on the digital scan or on the printed page. What matter is that it should have been a "a baskit maker," and about that fact there can be little doubt. The students' task was not to improve the writer or modernize the spelling, but it was to record the printer's intention.

# Unknown unknowns

In addition to known unknowns, the transcriptions include unknown unknowns or defects, whether in the printed source or the transcription, that went undiscovered and unmarked. The known unknowns are distributed very unevenly across the texts and are almost entirely a function of the legibility of the digital scan from which the transcriber worked. Something like 90% of the errors occur in 10% of the pages.

A high percentage of known unknowns on a given page is evidence that the text was difficult to transcribe, and it is more likely to have unknown defects as well. Students were asked to correct such unknown unknowns and employed various routines for spotting them. Modern spell checkers are of no use for that purpose, but corpus specific lists o unique spellings are. Endemic errors in early modern print involve the confusion of long 's' with 'f', and 'n' with 'u'.

# The structure of the spreadsheet

The spreadsheet is derived from a MySQL database that includes a tabular representation of the approximately 11 million word tokens (including punctuation) that are found in the EMD corpus.

Here are the first three rows of the spreadsheet (in a slightly condensed version)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| location | TCPText | KWICL | correction | KWICR |
| A00456-1-b-0210 | Chast●●●e | expressed a rare example of the vertue of | Chastitie | , by Virginias constancy , in wishing rather |
| A00456-1-b-1020 | Kichard | Imprinted at London , by William How , for | Richard | Ihones . 1575. |
| A00456-2-b-0050 | cup●● | Qui | cupis | aethereas & summas scandere sedes Vim simul |

The location column combines a workid with a reference to the page image: 1-b-0210 means "word 21 on the right page of image 1" (the word counter increments by 1-). The spreadsheet also includes the unique xml:id for the token in the text. The "TCPText" columm shows the transcriber's reading. The "correction" column shows the corrected text, flanked by left and right context, usually 35 characters, unless the text occurs at the beginning or end of an XML element that marks a clear discursive boundary.

For human readers, 35 characters on either side usually provides enough information to solve the puzzle. In many cases you do not need to look at the digital scan or page image. Thus in the first row " Chast●●●e" must be "Chastitie", and you don't need to know much about early modern orthography to know that '-ie' is the common spelling for words that we now spell with a terminal 'y'. There is no "Kichard", and the only possible completion for 'ichard' is 'Richard'. But general speaking "headless" words are much harder to guess than words that lack something in the middle. Words without a tail are also hard, especially in Early Modern English, where a final 'e' is optional.

The third row reminds you that there is a lot of foreign stuff in Early Modern texts. In the XML data from which this spreadsheet is ultimately drawn, the texts were tokenized and linguistically annotated with Phil Burns' MorphAdorner, which is very good at catching foreign words. Which means that from a machine-learning perspective it is relatively easy to exclude the 'noise' from non-English passages.

# Common defects

Most of the textual defects fall into one of the following classes:

1. A missing letter: s●tale => **stale** wit
2. missing letter, wrong count: ●●ght => from my **sight**
3. missing word: 〈◇〉: not to **indure** the sight
4. two words wrong joined: awronge => **a wronge** way
5. one word wrongly split: a biectly => thinke As **abiectly** of thee
6. 'f' for 's'(easy): Cofsen => my **Cossen** Wagtaile
7. 'f' for 's'(hard): loft => haue not **lost** their lives
8. 's' for 'f' (easy): sortunate => such a **fortunate** dame
9. 's' for 'f' (hard): wise => run away from his **wife**
10. 'n' for 'u': peenishnes => **peeuishnes**

You notice that except for the missing word all the puzzles can be solved definitively from a consideration of the word in its context. That is probably true of between half and two thirds of textual defects in the TCP corpora. If you are concerned with finding out what the printer wanted you to read, machine based correction can take you quite a ways. The interesting question is how much human labour (of a not very interesting type) you can replace and how you can target human oversight with the best effect for the least effort.

Machine learning techniques are probably useless for resolving the remaining 12,000 known defects. They involve spellings that five very bright undergraduates couldn't figure out, and many of them involve missing words that will be in most cases beyond a machine to begin with. But if the "error maps" from the Early Modern plays can be turned into helpful training data for a first pass through the much larger corpus of Early Modern English and Early American texts.

The data in the spreadsheet can be rearrange and expanded in various ways. For the purpose of human-based curation, the 35 character window is a useful compromise. If you add more you may get better results in a few cases, but you slow down the work. It would be easy to extract different surrounding data: every thing in

1. the sentence
2. the speech or <sp> element
3. the previous and next speech

Ditto for the test data

1. The students were Nayoon Ahn, Hannah Bredar, Madeline Burg, Nicole Sheriko, Melina Yeh, and except for Nicole, they were freshmen or sophomores at the time. [↑](#footnote-ref-1)