This document provides information about the files in this directory, which are part of a project tentatively called "The History and Future of the Book: NLP Comes to the Humanities."

This part of the project focuses on the question whether many of the incomplete spellings in the texts of the Text Cration Partnership can be automatically completed or corrected through the application of machine learning techniques.

Find out about the Text Creation Partnership by looking at its website at <http://www.textcreationpartnership.org>. The TCP has over the past fifteen years produced ~65,000 SGML transcriptions of books published before 1800 in the British Isles or North America. The transcriptions were produced mainly in Asian countries via "double keyboarding." Every text is typed twice, and the results are collated for error analysis and correction. The transcribers were not necessarily native speakers, and they certainly were not experts in Early Modern English. They transcribed what they saw whether or not they knew what it meant.

The transcribers worked from digital scans of microfilm images of printed books. The microfilm images were produced over a period of thirty or more years, mainly after World War. Where the transcribers could not identify a letter, word, or phrase they marked it with code specifying as precisely as possible what is missing from the transcription: one letter, three letters, two words etc. In the files of this project the black circle (●) has been used as the symbol for a missing letter, and the term "blackdot word" is used to describe defective words. There is also a black square (▪), which stands for an indeterminate punctuation mark and nearly always occurs at the end of a word.

At <http://www.textcreationpartnership.org/docs> you can look at some of the vendor instructions for the project. Spending a few minutes with those documents is a good way of learning what the transcribers were up against.

Let's take a walk through the directory and get a sense of the logical and temporal order of the stuff in it.

## The Malcontent (A07071)

John Marston's Malcontent, a play published in 1604, is a good example of a TCP text, and it sits in the middle of the pack when it comes to transcription errors. The relevant documents about it are kept in the directory MalcontentA07071, where 'A07071' is the arbitrary filename assigned to it in the TCP project. Directories like this exist or can be readily produced for each of some 60,000 files. If you understand the relationship of files in this directory, you understand thestate of the TCP archives.

The directory 'malcontentA07071' contains the following files or directories:

1. A07071EEBO-images
2. A07071.sgm
3. A07071.xml
4. A07071.tab
5. A07071.xlsx
6. A07071short.tab

The first directory contains the 32 double page images from which the texts were transcribed. In the correction of the transcription these images are what you check the transcriptions against. Have a look at the left bottom of image 16 to see a problem where somebody needs to look at the printed original to figure out what is going on.

The file A07071.sgm is the SGML file produced by the transcribers. The file is encoded as a "TEI file" using a slightly modified version of TEI-Lite, a simplified scheme of the Text Encoding Initiative. More about the TEI at <http://tei-c.org> and more particularly its [Guidelines](http://www.tei-c.org/Guidelines/).

The file A07071.xml is the same file linguistically annotated with MorphAdorner, a NLP toolkit developed by Phil Burns (more at <http://morphadorner.northwestern.edu>

In this version some typographical textual features are thrown away, notably the long 's', but in all other respects it is a faithful but enriched representation of the source file. Specifically it does the following:

1. It recognizes each word or punctuation mark as a distinct token with an xml:id and wraps it in a <w> or <pc> element
2. It adds <c> elements to encode space, but these do not carry elements and could be thrown away
3. It provides a POS tag through the 'ana' attribute
4. It provides a lemma through the 'lemma' attribute
5. It provides a standard spelling the 'reg' attribute
6. The 'n' attribute provides a location for the transcribed on the page

With regard to #6 an attribute value like '4b-1320' means "look at the right page of image 4 and look for word the 132nd word," which means "some where in the middle of that page."

The file A04071.tab is a tabular and more representation of the same data. It is easier to find your way around it when you look at its spreadsheet version A07071.xlsx.

There you see each word token occupying a single data row, with various properties surrounding it. Thus you can put the token in the middle of a KWIC display that for humans is large enough to resolve most ambiguities. It also lets you see a token as the middle of a trigram of spellings or POS tags. The tabular display also includes a complete XPath for each token. This is helpful for focusing on or excluding stuff in parts of a document that are likely to contain special features, such as notes or speaker labels.

The file A07071short.tab is a summary of the tabular file. For every distinct and case-sensitive spelling it produces a data row like the following:

A07071 Dreames Dreams n2 dream p scene body 1

The first colum is the workid. The next four columns provide the spelling, standardspelling, POS tag, and lemma. The next three columns provide data about where the word occurs. Reading them from the back, you see that this word occurs in the "body" (as opposed to the "front" or "back" element of a text). It occurs in a <div> element with the type attribute "scene", and it occurs in a <p> or paragraph, which means that it is prose. The short tab is very useful for selecting or ignoring "paratext" of various kinds, for instance stuff that occurs in the <back> element of a text inside a <div> with the type attribute "errata".

## Lexical files

The lexical files directory contains two files:

1. tcpLowerSpell.tab
2. tcpLowerSpellblackdot.tab

The second file is just a subset of the first, extracted her for convenience. It includes all the spellings that contain one or more occurrence of the black dot that represents a missing letter.

The file tcpLowerSpell.tab was extracted from ~45,000 short tab files. It consists of 4.1 million distinct spellings in their lower-case form. Here are three rows that illustrate the structure of this file:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Spelling | colfreq | docfreq | filename | foreign |
| arbi●r●ment | 1 | 1 | N20568 |  |
| arbi●r●o | 1 | 1 | A01093 | fw-la |
| arbi●trary | 2 | 2 | more |  |

The spelling is followed by information about the collection and document frequency. If a spelling occurs only in one file, its filename is given. Otherwise the filename column says 'more'. If the word is in a foreign language, the language is marked. There is an error rate of less than 5% for getting the basic distinction between English and foreign right. It is not easy to get it right because many Early Modern English spellings could be Latin or French words.

Blackdot words account for 758,000 or 18% of all distinct spelling, but their total frequency of 2.9 million in this set texts adds up to just 0.2% of the total word count. But they cluster heavily, and where they cluster they are perceived by readers as irritating or substantively interfering with an understanding of the texts. So it matters to fix them wherever possible.

The current version of MorphAdorner has some functionalities for identifying and completing blackdot letter words. The corrections are typically in the regularized spellings or lemmata.

In this list proper names have not been identified, but that information is in the short tab files and could be readily transferred to it. False positives and negatives in the identification of names add up to about 5%.

## The SHC files directory

The SHC files directory deals more specifically with data from a subset of the TCP files, Early Modern plays that are part of a "Shakespeare His Contemporaries" project. The document SHCSpreadsheetNotes.docx cover some of the ground traversed here but goes into greater detail about the collaborative curation of that subset. The spreadsheet SHCTextCorrections.xlsx is a list of almost 45,000 corrections. The great majority of them involve "blackdot" words. The mappings of blackdot words to corrected spellings may be a useful input for machine learning techniques.