Corpus Analysis & Observations (Prepare Stage)

In order to perform the encoding, it is important to know the datasource one works with. To do that we must ask the datasource the right questions, investigate, outliers, and know where in the source is what we are seeking from it.

Following is an insight into the lengthy Analysis performed during the Prepare Stage of the project:

1. Searching for Early Modern English (EME) Texts

   Since the goal was to look solely for EME texts, a way to identify them had to be discovered.

   Questions asked (in the form of Queries):
   Q i) Are there any categorizations to tell them apart?
   Yes. There are several ways to distinguish and one of them is to look for the Language Category ID and Text Language in the taxonomy of the document's header.

   Q ii) How to use this categorization to search for the EME texts?
   Text Language

   Text Language = 'Early Modern English' can be found in the language usage element of a text's header

   Observation: Some texts seem to have the Text Language repeated more than once.

   Q iii) What is the reason for the above observation and how can it potentially impact the project?

   These texts from the observation have multiple Text Class elements. The reason is that in cases where errata corrections have been made to the parameter values of the original version, the default Text Class element contains the corrected values, while the original values are preserved for reference in a separate Text Class element identified as the 'old' version.

   Observation: the 'new' Text Class element is identified with a default="true" attribute-value pair

   Q iv) Now that EME texts have been identified, which is the shortest text that can be used for a sample of the tokenization?

   A text called "Letter (to his mother)" with 122 words

   Observation: Multiple instances of the title "Boethius" have been found

Q v) Which title names are repeating?

41 such texts, including 3 instances of “Boethius” and several different letters.

Q vi) What is the reason for this?

To answer this, the document itself had to be looked at. In the case of “Boethius”, which is actually an eponymy of Boethius’ work *“On the Consolation of Philosophy”*, the Corpus contains the translation of his original work performed by different 3 authors in 3 different periods during the EME:
- “Boethius” by George Colville (1500-1570/ E1 Period)
“Hetherto it suffyseth that I haue shewed the maner and forme, of false felicite or blessednes, which if thou beholdeste perfetlye, it restythe to declare from henceforthe, whyche is the very true felicitie.”
- “Boethius” by Elizabeth I (1570-1640/ E2 Period)
“Hitherto hit sufficeth to shewe the forme of gileful felicitie, wiche if you Clirely beholde, the ordar than must be to shewe you the true.”
- “Boethius” by Viscount Preston (1640-1710/ E3 Period)
“Let it suffice that I have hitherto described the Form of counterfeit Happiness: So that if thou considerest well, my Method will lead me to give to thee a perfect Draught of the true.”

Observation: In the case of the E3 Period (1640-1710) “Boethius”, clearly, the spelling and the grammatical structure are presentably closer to Modern (present-day) English, the text is interpretable, and can be easily understood by most modern readers. Hence, the corpus of texts is filtered further down to the E3 period.

Q vii) How to identify texts from the E3 period?

Very easily, the TEI element (that contains each text) has an attribute “n” to identify the period (E3 in this case).

Q viii) Which, now, is the shortest text?

“A Letter by the Privy Council (to Lord Rochester)”, which is a letter written to Lord Laurence Hyde, 1st Earl of Rochester, by members of the Privy Council.

Observations: The letter consists of formal ‘court speak’, and makes use of superscript in the closing of the letter like Yo^{r} (meaning ‘Your') and Lo^{ps} (meaning ‘Lordships’). Similar observations have been made for other letters, such as w^{ch} (meaning ‘which’) and y^{r} (also meaning ‘your’), although there have been instances where in the same letters where the complete word ‘your’ can be found.

Decision: For the sake of brevity, encoding letters shall be avoided, to circumvent the abbreviations and contractions of words.

Q ix) Which is the next shortest text after filtering out letter text types?

“My Great Journey to Newcastle and to Cornwall”, a travelogue, by Celia Fiennes, with 5,151 words available in the Corpus

2. Encoding task

To summarize, the encoding task was not straightforward, especially since the objective was to convert a complex XML tree into a flat relational schema.

Observations:

- The element being referenced for fetching the Text’s title was incorrect, as it turns out that the Texts were fetched from larger works (like say a Journal is a work and an article may be a text found in the corpus). That led to an added relation in the schema, called works, which stores the title of the Work and its author
- The text body contained multiple divisions, with each division having a possibly different type from the other in a single body, as well as having the possibility of different titles (because, as mentioned in the previous observation, texts in the corpus have been selected from larger works, and hence, they can their own individual titles
- A single division can have multiple headers (for example chapter and subchapter names)
- Page numbers are indicated with </pb> elements, having an attribute-value pair (for e.g., n=“141”), indicating the actual page number from the original works
- Each paragraph is embedded within elements
- In case of a Drama text type, paragraphs are additionally embedded within <sp></sp> elements, to indicate the paragraphs spoken by different speakers
- From the previous observation, the speaker name is also found in the <sp></sp> elements, following which, the spoken paragraphs are identified.
- Each line from the original works is indicated with </lb> elements
- In each line (after every </lb>), there is a possibility of two different types of elements: <note></note> which denotes a possible change in the Corpus while it was converted to XML format, and; <hi></hi>, which indicates if a set of words has been highlighted. Both of these have to be sensitively handled

For all of the above observations, complex queries were coded to retrieve metadata, and tokens. For example, for the last observation, a function was generated that carefully replaced the elements itself with the texts within the elements, because they were hindering the tokenization (texts appearing after these elements were not being tokenized).
Though the queries were complex, it was ensured that all queries run with a low run-time complexity, and it was managed to bring the most complex ones to under 1.5 seconds.