How to use and read 25,000 texts from 1470-1700

an update from

Visualising English Print

Heather Froehlich
@heatherfro
Visualising English Print 1470-1700

• A collaborative, interdisciplinary project
  – University of Wisconsin-Madison
  – University of Strathclyde
  – Folger Shakespeare Library

• http://vep.cs.wisc.edu/

ECRs involved 2016-2017:
Eric Alexander, Deidre Stuffer, Erin Winters, Erin Larson (UWisc)
Heather Froehlich, Alan Hogarth (U Strath)
“Addressing Variation at Scale in Historical Document Collections”

Eric Alexander, Deidre Stuffer and Michael Gleicher

IEEE Workshop on Visualization for the Digital Humanities

http://vis4dh.com
We want to enable literature scholars to answer questions that can only be asked at scale, such as:

- What were people writing about during the early modern era?
- How did language and topics of discussion change over time?
- Is it possible to track the evolution of particular genres?
- Is our concept of “genre” itself an accurate reflection of the types of works that were created?
- What attributes make texts similar or dissimilar from one another?
EEBO-TCP Phase I Public Release: What to expect on January 1
December 24, 2014 - Posted in News - Permalink

The Text Creation Partnership is quickly arriving at a major milestone: starting January 1, 2015, all restrictions will be lifted from EEBO-TCP Phase I, which consists of the first 25,000 texts transcribed and encoded by the TCP from 2000-2009.

These 25,000 (plus a few hundred) texts will be freely available to anyone wishing to use them, and there will no longer be any restrictions on sharing these files. They will be licensed under the Creative Commons Public Domain Dedication (CC0 1.0 Universal), which will be indicated in the header of each text.

But what does this news mean for users of the EEBO-TCP Phase I texts? Read more »

Announcing the winner of the inaugural RSA-
Today’s talk:

1. Standardise and Curate
2. Learn about the texts
3. Model stylistic difference
1. Standardise and Curate

- Machine readable vs machine actionable files
- TCP texts come as SGML/XML files (TEI-compliant)
- Incredibly rich file format, but includes TONS of extratextual stuff
SimpleText
http://graphics.cs.wisc.edu/WP/vep/simpletext/

1. Substitutes UTF and Unicode characters for their closest counterparts in ASCII

2. It does not include any metadata annotations, favoring to store those in separate metadata-specific files

3. It does not preserve physical aspects of document layout or typography, but does strive to maintain line breaks

4. It employs simple, dictionary-based spelling standardization
Spelling Standardisation

• We wanted to standardise prepositions, expand elisions, and preserve verb endings
• BUT preserving Early Modern verb endings (-st, –th) would require an overhaul of VARD’s dictionary.
## WHY NOT VARD?

<table>
<thead>
<tr>
<th>ORIGINAL</th>
<th>NORMALIZATION</th>
<th>SHOULD BE</th>
</tr>
</thead>
<tbody>
<tr>
<td>all’s</td>
<td>ell’s</td>
<td>all’s</td>
</tr>
<tr>
<td>caus’d</td>
<td>cause</td>
<td>caused</td>
</tr>
<tr>
<td>Cicilia</td>
<td>Cicely</td>
<td>Cicilia</td>
</tr>
<tr>
<td>courtesie</td>
<td>curtsy</td>
<td>courtesy</td>
</tr>
<tr>
<td>diuers</td>
<td>divers</td>
<td>diverse</td>
</tr>
<tr>
<td>hir</td>
<td>his</td>
<td>her</td>
</tr>
<tr>
<td>ile</td>
<td>isle</td>
<td>l’il</td>
</tr>
<tr>
<td>ist</td>
<td>first</td>
<td>is’t</td>
</tr>
<tr>
<td>kild</td>
<td>kilt</td>
<td>killed</td>
</tr>
</tbody>
</table>

Spelling Standardisation

• How to fix?
  - manually select some variants over others to change confidence scores
  - Mark non-variants and variants; input their standardised form
  - adding words to the dictionary
  - Use 1:1 dictionaries and python to modernise

• ‘heede’ > head (unless w ‘to’: to heed)

http://graphics.cs.wisc.edu/WP/vep/tag/spelling-standardization/
<table>
<thead>
<tr>
<th>CHARACTERS</th>
<th>CHANGE TO</th>
<th>LOCATION IN WORD</th>
</tr>
</thead>
<tbody>
<tr>
<td>cyon</td>
<td>tion</td>
<td>End</td>
</tr>
<tr>
<td>lie</td>
<td>ly</td>
<td>End</td>
</tr>
<tr>
<td>shyp</td>
<td>ship</td>
<td>End</td>
</tr>
<tr>
<td>t’</td>
<td>to_</td>
<td>Start</td>
</tr>
<tr>
<td>th’</td>
<td>the_</td>
<td>Start</td>
</tr>
<tr>
<td>tiue</td>
<td>tive</td>
<td>End</td>
</tr>
<tr>
<td>vn</td>
<td>un</td>
<td>Start</td>
</tr>
<tr>
<td>vs</td>
<td>us</td>
<td>Anywhere</td>
</tr>
<tr>
<td>ynge</td>
<td>ing</td>
<td>End</td>
</tr>
</tbody>
</table>

Some other things we changed

- **doe > do**
- **bee > be**
- Replaced reserved XML characters (<, >, %) with at-signs (@)
- Replaced ampersands (&) with the word “and”
- A dash (—) becomes two hyphens (–)
- TCP illegible characters (bullet: •) became carets (^)
- TCP unrecognizable punctuation (small black square: □) became asterisks (*)
- Replaced non-ASCII characters not assigned ASCII equivalents (e.g., pilcrow: ¶) with at signs (@)
- TCP missing word symbol (lozenge in brackets: ◊) became ellipses in parentheses ((...))
- Deletes TCP end-of-line hyphen characters supplied during transcription (vertical bar: |, broken vertical bar: ‖)

http://graphics.cs.wisc.edu/WP/vep/pipeline-2/
TEI-compliant XML vs VEP SimpleText

AN Easy, certain, and perfect method, to cure and prevent the Spanish sickness. Whereby the learned and skilful Surgeon may heal a great many other diseases. Compiled by Peter Lowe, Arellian: Doctor in Chirurgery in Paris; Surgeon ordinary to Henry the most Christian King of France and Navarre.

AT LONDON, Printed by James Roberts. Anno Domini. 1596.

TO THE RIGHT Honourable Robert Deorarx, Earl of Essex, Vicount Hereford, Knight of the most noble order of the Garter. &c. Peter Lowe wishes good success in all his worthy enterprises, and high attempts.

Although (right honourable) this Treatise be far dissimilant from your studies, yet I have embellished my selfe to let it passe, under your honourable patronage, for by Gods assistance, I intend hereafter to publish divers other Books of Chirurgie, all which shall be shrouded under your honourable shield, and protection, in regard of your manifold courtesies, and liberality to all Strangers bountifully extended, whereof you cease not every day to give experiment, but especially in regard of your honourable vallant prowess, and rare Martial exploits in aiding my dread Sovereigne, i.e. Master, the most victorious King of France, so that in honouring him, I must love you, and for your herocall vertues, admire your wisdom, and courage, in prosecuting all venturous, and hard attempts: as lately hath beene experimented to the terror of the Enemy, peaceable welfare of this Realm, and immortal fame unto you, &c. your postillar for ever. Then forasmuch as the dedication hereof, cannot bring, any addition of credit unto you, but rather, be the more acceptable, under the title of your honourable patronage, most humbly with my self, I present the same unto your honours good liking, as one, that daily intemitteth not, to wish the advancement of your honourable estate.

Your Honours in all duty most affectionate Peter Lowe.
All files have undergone the same process → build corpora
Corpora

- We offer 5 collections of corpora:
  1. VEP TCP Collection
  2. VEP Early Modern Drama Collection
  3. VEP Early Modern Science Collection
  4. VEP Early Modern 1080
  5. VEP Shakespeare Collection

Each corpus from these collections are available in 2 forms: Unrestricted and All
VEP TCP Collection

• EEBO-TCP Phase 1 corpus: 25,368 texts
• ECCO-TCP corpus: 2,473 texts
• EVANS-TCP corpus: 5,012 texts

All of our TCP collections are available in either Standardised or Unstandardised SimpleText format.

http://graphics.cs.wisc.edu/WP/vep/vep-tcp-collection/
VEP Early Modern Drama Collection

• Core Drama 1660 corpus
  – 554 total plays; 471 unrestricted plays

• Expanded Drama 1660 corpus
  – 666 total plays; 569 unrestricted plays

• Expanded Drama 1700 corpus
  – 1,244 total plays; 1,009 unrestricted

VEP Early Modern Science Collection

• **Super Science Corpus**
  – 1,979 total texts; 1,130 unrestricted texts

• **Big Names of Science Corpus**
  – 329 total texts; 272 unrestricted texts

Early Modern 1080 Corpus

• **1080 texts**
  – Selected from EEBO-TCP phase I and ECCO-TCP
  – Randomly sampled at a rate of 40 texts / decade

VEP Shakespeare Collection

• **Shakespeare TCP (A11954)**
  – 36 Shakespeare plays, taken from the First Folio in from EEBO-TCP phase I (TCPID A11954)

• **VEP Shakespeare Folger**
  – Our plain-text version of the Folger Digital Texts corpus

2. Learn about the texts

- [http://vep.cs.wisc.edu/metadataBuilder/](http://vep.cs.wisc.edu/metadataBuilder/)
  - A way of combining several different spreadsheets’ worth of metadata into ONE MEGA SPREADSHEET
The Visualizing English Print (VEP) project has access to a wide range of metadata about the documents in our various corpora. This metadata comes from a variety of sources, including the TCP master metadata sheet, Docuscope tagging (as performed by Ubie), and information about the documents that we have curated ourselves. Across all of these sources (and more!), there are many more columns of metadata than most researchers would actually want.

The VEP Metadata Builder gives researchers a way to build their own metadata files that include the information most relevant to them. To get your metadata, simply pick the documents with which you are working, pick the metadata fields that are of interest, browse the resultant table to confirm your selection, and finally save the metadata in one of a number of possible formats (as well as an accompanying README file).

**Step 1:** Pick the documents

- Full TCP (Unrestricted only)

A list of all unrestricted TCP texts (EEBO-TCP Phase I, EEBO-TCP Phase II, ECCO-TCP, and Evans-TCP). A list of free TCP texts only.

**Step 2:** Pick the metadata fields

- **TCP Metadata (source)** - All selected (10)
- **Ubie Categories (source)** - None selected
- **Non-English Language Metadata (source)** - None selected
- **Figures-Per-Text Metadata (source)** - None selected
- **Derived Date Metadata (source)** - None selected
- **Text Links (source)** - None selected

**Step 3:** Examine and confirm the metadata

<table>
<thead>
<tr>
<th>TCP</th>
<th>EEBO</th>
<th>VID</th>
<th>STC</th>
<th>Status</th>
<th>Author</th>
<th>Date</th>
<th>Title</th>
<th>Pages</th>
</tr>
</thead>
<tbody>
<tr>
<td>A00002</td>
<td>995650634</td>
<td>15849</td>
<td>STC 1000.6; ESTC S115415</td>
<td>Free</td>
<td>Aylett, Robert, 1583-1555?</td>
<td>1625</td>
<td>The bride's ornaments viz. five...</td>
<td>134</td>
</tr>
<tr>
<td>A00005</td>
<td>99642406</td>
<td>7058</td>
<td>STC 10000; ESTC S115560</td>
<td>Free</td>
<td>Higden, Ranulf d</td>
<td>1515</td>
<td>Here begins a short...</td>
<td>202</td>
</tr>
</tbody>
</table>
Available metadata differs by corpus

Step 1: Pick the documents

Full TCP (Unrestricted only)

A list of all unrestricted TCP texts (EEBO-TCP Phase I, EEBO-TCP Phase II, ECCO-TCP, and Evans-TCP). A list of free TCP texts only.

Step 2: Pick the metadata fields

TCP Metadata

Ubiq Categories

Non-English Language Metadata

Figures-Per-Text Metadata

Derived Date Metadata

Text Links

Build Metadata

Step 3: Examine and confirm the metadata
Available metadata differs by corpus

Step 1: Pick the documents

Early Modern 1080

A corpus of 1080 digitized texts built from the EEBO-TCP Phase I and the ECCO-TCP used to generate a topic model for Serendip. Texts originally were published between 1530 and 1799.

Step 2: Pick the metadata fields

Master Metadata (source)
Text Links (source)
Ubiq Categories (source)
TCP Metadata (source)
Non-English Language Metadata
Figures-Per-Text Metadata
Derived Date Metadata

Build Metadata!

Step 3: Examine and confirm the metadata

Step 4: Save the metadata
3. Model Stylistic Difference

http://vep.cs.wisc.edu/ubiq/
Ubiqu+Ity FAQ

1. What text format/s can I upload and how many?

2. What is the DocuScope dictionary and where can I download it?

3. How do I specify my own rules?

4. How does the chunk feature work?

5. How does the blacklist feature work?

6. What does the CSV output look like, and how can I read it?

7. What does the SlimTV text viewer look like?

8. Is the source code for Ubiqu+Ity available?

9. To whom should I email comments, ideas, and/or bug reports?

10. How many LATs (linguistic categories) does a specific version of Docuscope have?

11. What files are the defect statistics option intended for?

12. Why replace the HTML viewer files with SlimTV tokens files?
Super Science Corpus
Philosophy of Science
Thank you!

http://vep.cs.wisc.edu/