

Virtual Modernization

Using Modern Spellings to Search Old Texts

**An Update for TCP Members,
New Orleans, June 25, 2006**

Jeff Garrett, Northwestern University

About the Project I

- **Name:** “CIC CLI Virtual Modernization Project”
- **Mission:** Develop a simple tool that allows both expert and non-expert users to search early modern English texts using modern spellings only
- **Basic Strategy:** Use modern English keywords that automatically invoke all early-modern spelling variants

Legend

CIC: Committee on Institutional Cooperation, a consortium of Big Ten universities & U. of Chicago

CLI: the CIC Center for Library Initiatives

About the Project II

- **Current Virt-Mod Participants:** All 13 CIC members plus Chadwyck-Healey
- **Principal Investigator:** Prof. Martin Mueller, Northwestern University
- **Budget:** Year 1: \$49,000; Year 2: \$25,000
- **Project Duration:** January 1, 2006–December 31, 2007. Thereafter release into public domain.
- **Currently Active Sites:**
 - Northwestern University:
<http://panini.northwestern.edu/philologic/shakespearesources.html>
 - University of Chicago:
<http://www.lib.uchicago.edu/efts/EEBO/search.html>

About the Project III

TCP in Research

Virtual Standardization and the TCP Project

by Martin Mueller -
Department of English,
Northwestern University

2

The following is a brief report on a project to apply virtual orthographic standardization to early modern English texts. This project has been sponsored by the Council for Library Initiatives of the CIC Libraries. It is funded through contributions from the libraries and Proquest. The project is being carried out at Northwestern University collaboratively between Academic Technologies and the Library. The project leader is Martin Mueller (Department of English).

The goal of the project is to get to a point where a modern reader can retrieve at least 99% of Early Modern English word occurrences by putting in modern search terms. This means that on a typical page of 400 words for or fewer words would be missed by

(null)?

6. Does it belong to some
7. Is it a word that is always sometimes capitalized
8. Is it an abbreviation (or a contraction (3), a word Roman numeral?
9. The frequency of the

At the moment such information 350,000 spellings. We expect million by Labor Day 20

In carrying out this work

turn
and
we
and
less
The
'be
ma
or



- Martin Mueller's introduction to the project in the Winter 2006 issue of the *Text Creation Partnership Newsletter*, p. 2-3

The Nature of the Problem for Which Virtual Modernization is An Answer

¶ Here begynneth the lyfe of the blessed martyr saynte Thomas.

¶ Here begynneth the lyfe of the blessed martyr **saynte Thomas.**



The martir saynte Thomas was son to Gylberde Bequet a burgeys of the Cite of London. And was borne in y^e place where as now standeth the churche called saynte Tho-

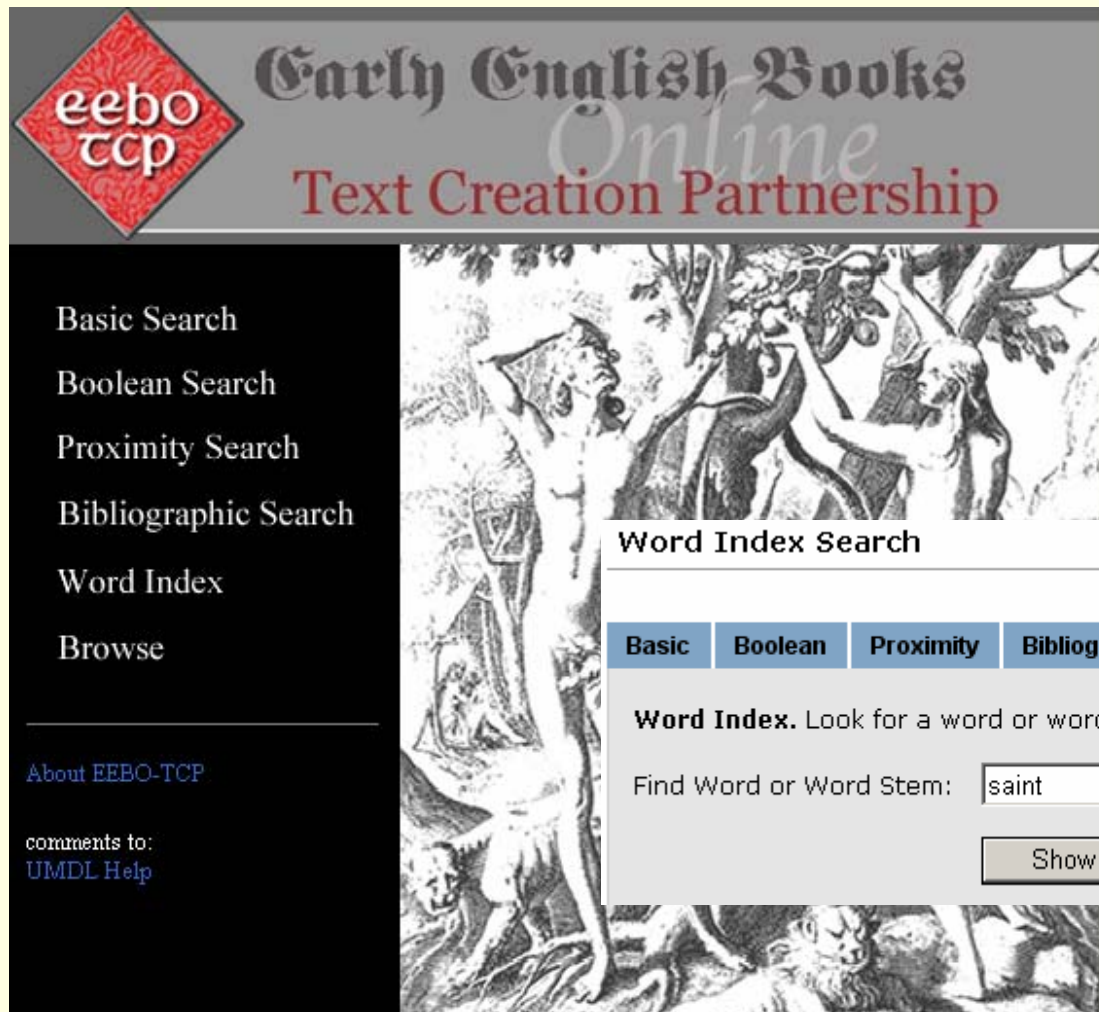
mas of Akers. And this Gylberde was a good deuote man / and toke the crosse vpon hym / and wente on pylgrymage to the holy lande / and had a seruaunt wth hym. And whan he had acomplished his pylgryma-

The martir saynte Tholmas was son to Gylberde Bequet a burgeys of the Cite of London. And was bor+ne in y^e place where as now standeth the chur+che called saynte Tholmas of Akers. And this Gylberde was a good deuote man / and toke the crosse vpon hym / and wente on pylgrymage to the holy lande / and had a seruaunt w^t hym. And whan he had acomplished his pylgryma+ge / he was taken homwarde by the hethenmen / and brought in prison of a prynce named Amerau-t where longe tyme he / and his felyshyp suffred moche peyne and sorowe. And the prynce hadde great affeccyon to+warde this Gylberd / and had oft co-munycyon with hym of the cristen feyth / & of y^e royalm of Englande by whiche conuersacion: it fortunod / that y^e daughter of this prynce had especiall loue vnto this gylberde / & was famylier with hym: and on a tyme she disclosed hir loue to hym / sayenge if he wolde promyse to wed hir / she shulde forsake frendes / heritage / and countre for his loue / and become cristen / and after longe colmunycacion betwene them / he promysed to wed hyr

[London?] : Imprinted by me Rycharde Pynson ... [1520]

Existing Strategies I: The TCP Word Index

- EEBO-TCP currently uses a word index for users to assemble



The screenshot displays the EEBO-TCP website interface. At the top left is the EEBO-TCP logo, a red diamond with the text 'eebo tcp'. To its right, the text 'Early English Books Online' is written in a stylized font, with 'Online' in a lighter, larger font. Below this, 'Text Creation Partnership' is written in a red serif font. A vertical navigation menu on the left side lists: Basic Search, Boolean Search, Proximity Search, Bibliographic Search, Word Index, and Browse. Below the menu are links for 'About EEBO-TCP' and 'comments to: UMDL Help'. The main content area features a large illustration of Adam and Eve in the Garden of Eden. Overlaid on the right side of the illustration is a search interface titled 'Word Index Search'. This interface includes a horizontal menu with tabs for 'Basic', 'Boolean', 'Proximity', 'Bibliographic', 'Word Index Search' (which is selected), and 'History'. Below the menu, the text reads 'Word Index. Look for a word or word stem in the word index.' There is a text input field labeled 'Find Word or Word Stem:' containing the word 'saint'. Below the input field is a button labeled 'Show Word Index'.

Existing Strategies I: The TCP Word Index

- This strategy can overlook non-adjacent spelling variants, some of which can be significant:

<input type="checkbox"/> s'aint	1
<input checked="" type="checkbox"/> sa int	3
<input checked="" type="checkbox"/> sa+int	3
<input checked="" type="checkbox"/> saint	81761
<input type="checkbox"/> saint▪	18
<input type="checkbox"/> santa	2
<input type="checkbox"/> saint-adoration	1

<input type="checkbox"/> sayns-park	1
<input type="checkbox"/> saynst	1
<input checked="" type="checkbox"/> sa+ynt	35
<input checked="" type="checkbox"/> saynt	13399
<input checked="" type="checkbox"/> sa ynt	62
<input type="checkbox"/> sāynt	1
<input type="checkbox"/> saynt-augustin	1
<input type="checkbox"/> saynt	1

<input checked="" type="checkbox"/> sa+ynte	1
<input checked="" type="checkbox"/> sayn+te	7
<input checked="" type="checkbox"/> saynte	1709
<input checked="" type="checkbox"/> sayn te	5
<input type="checkbox"/> saynted	2
<input type="checkbox"/> sayntefye	3

A Note about Chadwyck-Healey Databases

- Chadwyck-Healey is still the principal supplier of tagged full-text versions of early modern English texts
 - [English Poetry Database](#)
 - English Drama (3,900 Plays) – late 13th c. to early 20th
 - Early English Prose Fiction – from 1500 to 1700
 - . . . numerous others
- Like EEBO-TCP, Chadwyck-Healey uses an alphabetical word index to allow users to construct their own lemmatized search groups
- Demonstration:

A Note about Chadwyck-Healey Databases

Keyword in Play List

Engli

[<<BACK](#)

Enter the first few letters of a word in the box below and click **Look For** to jump to the nearest match in the list of words or terms below. Alternatively, select from the list using the checkboxes and click **Select** to take your selections back to the search screen.

HOME PAGE

SEARCH

COMPLETE CO

INFORMATION

LITERATURE C

Keyword in Play:

Look For

Up

Select Keyword in Play

- saynge (16)
- saynges (2)
- sayngs (1)
- sayngys (1)
- saynt (179)
- saynte (12)**
- saynted (2)
- sayntes (21)

Select

Existing Strategies II: Gale's "Fuzzy Searching"

Test drive the next generation of InfoTrac®
Now.

CIC Northwestern University

Eighteenth Century Collections Online

Help
Search Tips
Gale Databases

Basic Search Advanced Search Browse Authors Browse Works

Advanced Search

Enter search term(s) and select index type(s).
Indicate choice of Boolean operators (AND, OR, NOT)

<input type="text"/>	in	Full Text	AND	Fuzzy search Level None
<input type="text"/>	in	Full Text	AND	None
<input type="text"/>	in	Full Text	AND	None
<input type="text"/>	in	Full Text	AND	None
<input type="text"/>	in	Full Text		None

SEARCH Clear Form

Limit Your Search:
by Year(s) of Publication: (yyyy-yyyy)

Existing Strategies II: Gale's "Fuzzy Searching"

Advanced Search

Enter search term(s) and select index type(s).

Indicate choice of Boolean operators (AND, OR, NOT)

			Fuzzy search Level	
<input type="text" value="franklin"/>	in	Author	AND	Low
<input type="text" value="richard"/>	in	Author	AND	None
<input type="text"/>	in	Full Text	AND	None
<input type="text"/>	in			
<input type="text"/>	in			

SEARCH Clear Form

- Francklin, Richard. [A short state of Francklin, bookseller, on David M](#)
1754. 7pp. Law
[Full Citation](#)
- Franklin, Richard. [Dissertatio de vai](#)
[persuasive for inoculation. By Ric](#)
Technology
[Full Citation](#) | [eTable of Contents](#)

Strategy III: Gather All Spelling Variants Once and For All

- Over the last year, an army of underpaid NU English students has gathered over 350,000 variant spellings from ca. 8,500 EEBO-TCP texts and assigned them to modern-English headwords.
- The goal is 500,000 spellings by September 1.
- The 350,000 spellings gathered thus far account for 98% of all word occurrences in EEBO-TCP and 96.5% in Chadwyck-Healey databases.
- Contextual note: There are 2.5 million distinct spellings in the EEBO-TCP database with 400 million word occurrences.

Example: “saint”

- Occurs in 8,500 EEBO-TCP texts in this exact spelling 82,579 times.
- Twelve additional spelling variants have been identified for a total of 106,725 occurrences.
- As a lemma, “saint” covers a total of 32 forms with 193,527 individual occurrences.

What Different Search Types Actually Search For

- saint (“original,” i.e. search this literal spelling)
 - saint
- saint (“modern,” i.e. search by modern headword)
 - saint | saincte | saint | sainte | sait | saite | saynct | sayncte | saynt | saynte | sayntte | sayt | seynt
- saint (“lemma,” i.e. search all forms, including plural, gerunds, etc., though not adj. or adv. derivatives)
 - saint | saincte | saincted | sainctes | saincts | saint | sainted | saintes | sainteth | sainting | saints | sait | saite | saynct | sayncte | saynctes | sayncts | saynt | saynte | saynted | sayntes | saynting | saynts | sayntte | sayt | seint | seints | seit | seynt | seyntes | seynts | seyt

When the Sayntes Go Marching In

- 923. A033
- 924. A033
- 925. A033
- 926. A033
- 927. A033
- 928. A033
- 929. A033
- 930. A033
- 931. A033
- 932. A033
- 933. A033
- 934. A03319 (bib:p.ij)iss
- 935. A03319 (bib:p.na)so
- 936.
- 937.
- 938.
- 939.
- 940.
- 941.
- 942.
- 943.
- 944.
- 945.
- 946. A03319 (bib:p.xxix)

Author, etc.: [Higden, Ranulf, d. 1364](#)
 Uniform title: [\[Polycronicon. English\]](#)
 Title: [Prolicionycion \[sic\] \[microform\]](#)
 Publisher: [Westminster : Printed by William Caxton,
 Date: after 2 July 1482]
 Type of material: Book
 Microfilm reel
 Microfilm reel
 Internet access
 Description: [20], CCxxv, [1], CCxxxi-CCCCxxviiij [i.e. 430] leaves
 Microfilm. Ann Arbor, Mich. : UMI, 1938. 1 microfilm reel ; 35 mm. (Early English books, 1475-1640 ; 13:6).
 Series: [Early English books, 1475-1640 ; 13:6](#).
 Access method: http://gateway.proquest.com/openurl?ctx_ver=Z39.88-2003&res_id=xri:eebo&rft_val_fmt=&rft_id=xri:eebo:image:6945 Online version

Link to resource(s) by clicking here: [Online version](#)

Notes: By Ranulf Higden. An English translation by John Trevisa. Edited and with a continuation by William Caxton. All three individuals are named on a3v.



Title: [Prolicionycion \[sic\]](#).
 Author: [Higden, Ranulf, d. 1364](#).
 Collection: [Early English Books Online](#)
[Table of contents](#) | [Add to bookbag](#)

... fithes feyn pete. And they that overcome cytres and townees & see the preeftee and defoible clerkes and holy place and trete for see to wuntes of holy sayntes that shal bifall for Wyckednes curd lyping of Cristen may. **¶** This doynge semeth fulfilled in : last tyme of Ezechius thempour. When that false prophete iohannetus occupied Persida and made Egypt and affeyn sub tre and Wote and brought in the false laue and secte of Sa fene. as it is ynnemore plain Wotton after Iohannetus tyme. **¶** Umely seynt Gregory seyth that thylke men haue no

¶ Liber primus

and is double. the ouer galystra, and the nether galystra, and Jeynes to gyrene. And also to Siria and to Fenicia. In eyther galystra is good lond and grete plente of corn and of frucht, good fische and swete wyne. And the fithte and sixthe. And southe take is

nerfor **seynt** austyn libro p confessionu~ sayth that f holy **seyntes** that myght helpe ayenst theyr enemye thought **seynt** helene the holy crosse that our lord C che of **seynt** Iohan that heyght seynt Iohans chirche ad Ianiculum. About the

Complex Searches

Search in Texts or Find Documents

Search for:

Show Occurrences: [In Context](#) [Line by Line](#) [Similarity](#)

[Single Term and Phrase Search](#) (default)

[Phrase separated](#) by words

[Proximity Searching](#) in the same Sentence or ir

Orthography: Original Modern Lemma

saint sinner.

5 occurrences. Please follow the link(s) at the bottom of the page to view the text found.

lif that neuer dyeth. Seynt Girgorye seith. That the felon **synners** shal be punished for theyr synnes. As **saynt** Austen sayth / the **synner** soroweth for synning, nor none of the **sayntes**, but the beste men were **synners**, lette it be saide to reders, in what wyse **saynt** Austayne wolde a **synner** sholde knowe his synne, he shal sayeth thys noble clerke **saynt** Hierom) of **synners**, yt he myght wipe away his synne, nor none of the **sainctes**, but the beste men were **synners**, lette it be saide with pictures, yt these **sayntes** are moste **synners**, and these wyse, moste of all we thynke that the **sainctes** are more mercifull in hearing **synners**, than the iust, the Moone to the **Saints**, the faithful wwtnesse to **synners**, the aungels: no men: nor among ye **Saintes**; for they were all **synners**: neither among the liuing, nor with the dead, the holy **sainctes** with vs poore **synners**, Angells with men, the soules both of the **sainctes** and of **synners** doe either perishe with synne, nor shal they iudge. This man is a **Saint** and that man a **sinner**; he the seruant of God, and that man Christ in the 20. of **Saint** Iohn, to reconcile **synners** by that forme of penitence mented by the force of **Saint** Pauls power of binding **synners**, giuen by Christ, to be the punishment that **Saint** Paule had inflicted vpon this **sinner**, Sain

- “Search all texts for phrases with any form of the word ‘saint’ followed within five words by any form of the word ‘sinner’.”

Yet to come . . .

- 150,000 new spellings added by September 2006, bringing total spelling variants to 500,000
- POS (part of speech) tagging or a different disambiguation routine for homonyms, e.g. “grene” (color) vs. “grene” (grain)
- Better treatment of Latin, French, and other foreign-language words and passages
- Application of virtual modernization to early modern foreign-language corpora, e.g. Chadwyck-Healey’s German collections
- More answers to the question: What questions is Virtual Modernization the answer to?
- Promotion/demonstration projects outside English departments

New Members are Welcome

- The CIC CLI Virtual Modernization Project will accept associate members from outside the CIC for balance of two-year pilot period.
- Benefits:
 - Access to all Virt-Mod enabled sites
 - Opportunity to co-develop Virt-Mod
 - Participation in Virt-Mod listserv
 - Help with in-house Virt-Mod implementations
- One-time Cost: \$1,500
- Remember: Virtual Modernization becomes open access in January 2008!

Questions?

- Jeff Garrett
Northwestern University Library
Evanston, IL 60208
- Email: jgarrett@northwestern.edu
- Phone: 847-467-5675