Left-Brained Written Word

Sushma Kumari: 3587711

Rishi Rajani: 3746976

Act 1, Scene 1 🎬

HUMAN-WRITTEN PIECES OF ENGLISH TEXT:

PROSE
POETRY
THEATRE, ETC.

DULL
DRY
UNORIGINAL
LACKING FINESSE

CONVENIENT WRITING

UNINSPIRING
LIMITED VOCAB

INCREASED USE OF CUSS AND COLLOQUIAL WORDS

HUGE TRANSITION FROM ‘THEATRE’, TO ‘FILMS’, TO ‘MOVIES’

“Fall From Grace”

Paradise Lost, by John Milton (1667)	Auction, by Quan Barry (2023)
“Whatever Hypocrites austerely talk Of **puritie** and place and innocence, Defaming as impure what God declares Pure, and commands to **som**, leaves free to all.” [1]	“The whole room instantly aroused– the men’s pants tenting, the women with their sudden secretions as happens when you are in the presence of the holiest of forms.” [2]

“Where there’s no novelty, there’s no curiosity”

Love-Letters Between a Nobleman and His Sister, by Aphra Behn (1684)	Funny Story, by Emily Henry (2024)
“Glorious woman was born for command and dominion; and though custom has usurped us the name of rule over all; we from the beginning found ourselves (in spite of all our boasted prerogative) slaves and vassals to the almighty sex. Take then my share of empire, ye gods; and give me love!” [3]	“After my request that he only smoke outside, he really must have stopped merely sticking his head out the window, because weeks pass without me smelling **weed** in the hallway.” [4]

GPT-4: Our Lord and Saviour (?)

TRAINED ON VAST CORPUS OF BOOKS

CREATIVE IN GENERATING UNSEEN TEXTS

ABLE TO TAKE UP “PERSONAS”

E.G., CHATGPT RESPONDS LIKE A CHARACTER IN A SHAKESPEAREAN PLAY

CAN ACCEPT AND ANALYZE DOCUMENTS

API CAPABLE

Goals

CORPUS OF TOKENS, WORDS AND VERSES FROM EARLY MODERN ENGLISH TEXTS

INITIATING A LARGE-SCALE LIBRARY

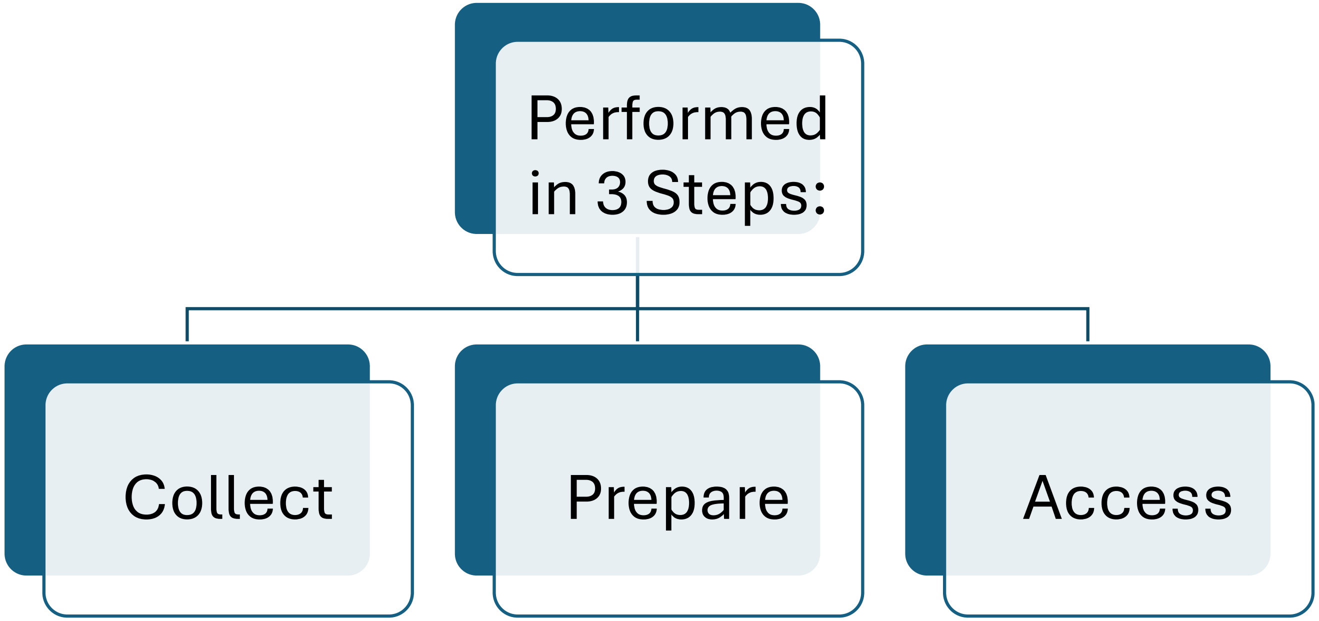
CAN BE FULLY UTILIZED BY GPT-4 TO GENERATE CREATIVE PIECES OF ART

Oxymoron: Left-Brained Written Word

Left Hemisphere of Brain	Written Word
• Logical • Analytical • Factual [5]	• Creative • Intuitive • Imaginative
Combination: Logic of creativity	

Photo Ref: Microsoft Designer's Image Creator (Powered by DALL-E 3)

Undertaking



Performed in 3 Steps:

Collect

Prepare

Access

Collect - I

- TEI XML Edition of the Helsinki Corpus (v.096), 2011 [6] (✅)
- Consists of:
  - Old English (≈ 850 – 1150 CE)
  - Middle English (≈ 1150 – 1500 CE)
  - **Early Modern English (≈ 1500 – 1700 CE) with ≈ 550,000 words**
- Well documented and encoded as per TEI guidelines
- Variety of works, spanning from categories (like dramas and poetry) to genres (like romance and comedy)
- Most foreign words and non-ASCII characters already handled

Collect - II

- Why Early Modern English:

  Regulation of language

  Reform in spelling of words

  Massive loanword vocabulary [7]

Prepare – I (WIP)

- Storage of the Corpus in BaseX Database (✅)
- Encoding using **XQuery*** on BaseX to fetch the following:
  - Tokens
  - Words
  - Verses
  - Proses
  - Complete Texts
  - Metadata, like title, author etc.
- Expected Query output format: comma-delimited (useful in the next step)

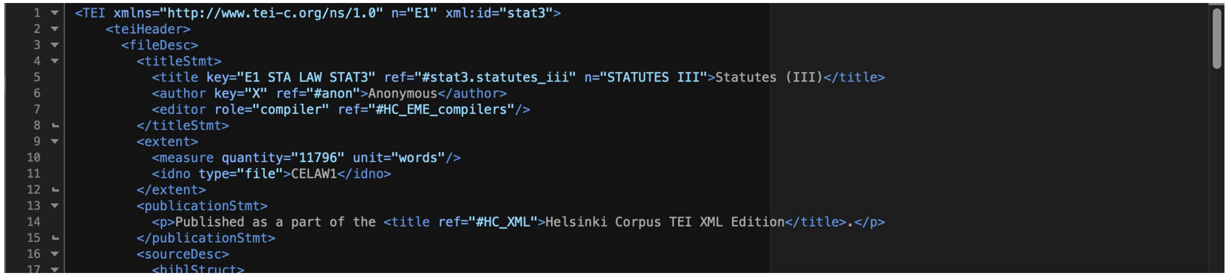
* Extension: More information on this in the upcoming slides

Prepare – II (WIP)

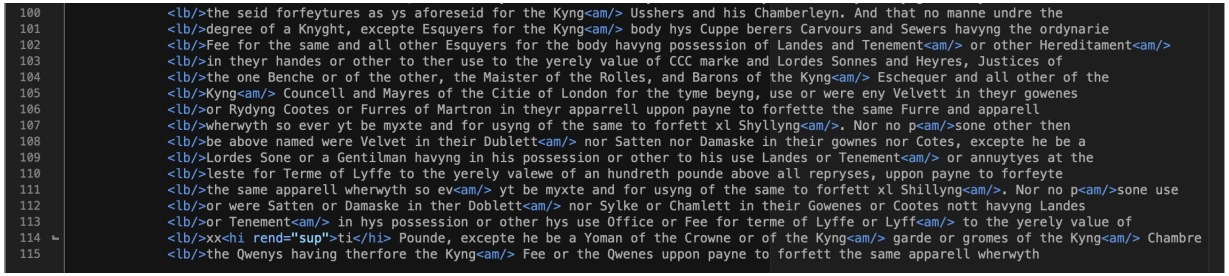
Early Analysis:

Performance of Data Exploration

Period	Text Count
-850	10
850-950	18
950-1050	49
1050-1150	6
1150-1250	14
1250-1350	17
1350-1420	42
1420-1500	33
1500-1570	44
1570-1640	48
1640-1710	46



```
<TEI xmlns="http://www.tei-c.org/ns/1.0" n="E1" xml:id="stat3">
    <teiHeader>
      <fileDesc>
        <titleStmt>
          <title key="E1 STA LAW STAT3" ref="#stat3.statutes_iii" n="STATUTES III">Statutes (III)</title>
          <author key="X" ref="#anon">Anonymous</author>
          <editor role="compiler" ref="#HC_EME_compilers"/>
        </titleStmt>
        <extent>
          <measure quantity="11796" unit="words"/>
          <idno type="file">CELAW1</idno>
        </extent>
        <publicationStmt>
          <p>Published as a part of the <title ref="#HC_XML">Helsinki Corpus TEI XML Edition</title>.</p>
        </publicationStmt>
        <sourceDesc>
          <biblStruct>
```



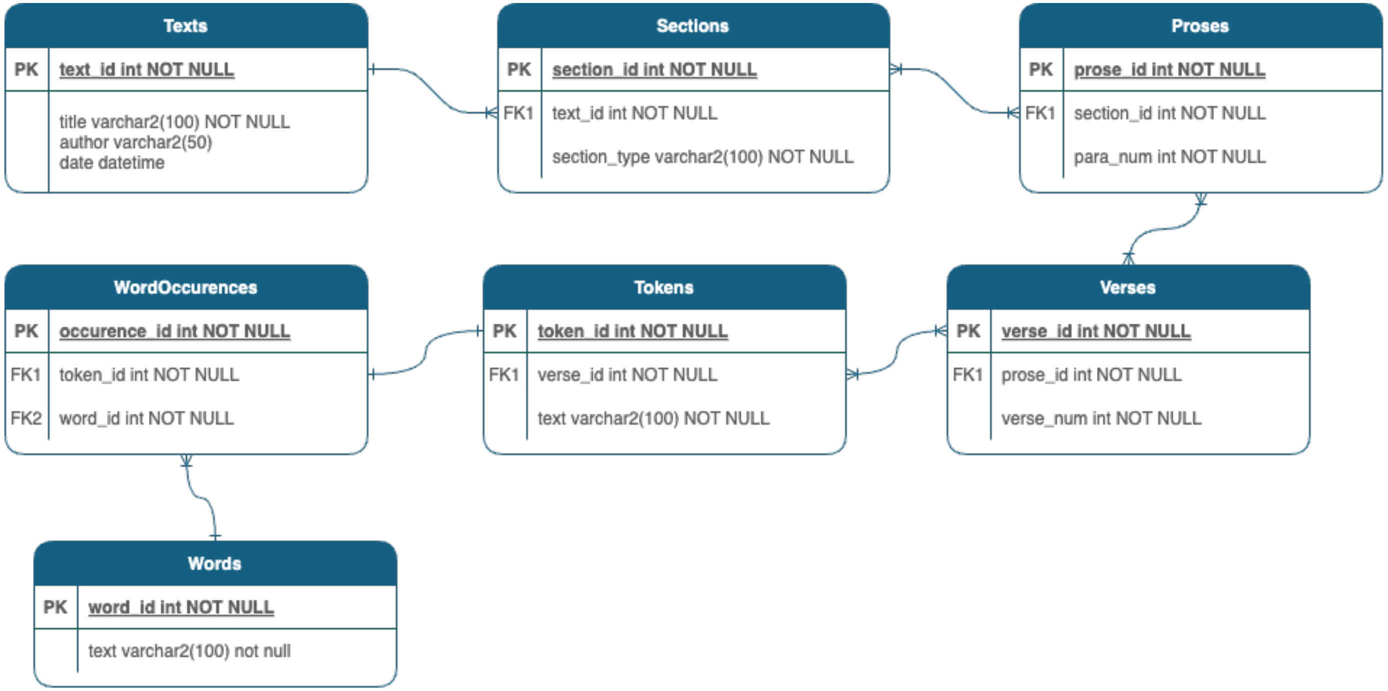
```
            <lb/>the seid forfeytures as ys aforeseid for the Kyng<am/> Usshers and his Chamberleyn. And that no manne undre the
            <lb/>degree of a Knyght, excepte Esquyers for the Kyng<am/> body hys Cuppe berers Carvours and Sewers havyng the ordynarie
            <lb/>Fee for the same and all other Esquyers for the body havyng possession of Landes and Tenement<am/> or other Hereditament<am/>
            <lb/>in theyr handes or other to ther use to the yerely value of CCC marke and Lordes Sonnes and Heyres, Justices of
            <lb/>the one Benche or of the other, the Maister of the Rolles, and Barons of the Kyng<am/> Eschequer and all other of the
            <lb/>Kyng<am/> Councell and Mayres of the Citie of London for the tyme beyng, use or were eny Velvett in theyr gowenes
            <lb/>or Rydyng Cootes or Furres of Martron in theyr apparrell uppon payne to forfette the same Furre and apparell
            <lb/>wherwyth so ever yt be myxte and for usyng of the same to forfett xl Shyllyng<am/>. Nor no p<am/>sone other then
            <lb/>be above named were Velvet in their Dublett<am/> nor Satten nor Damaske in their gownes nor Cotes, excepte he be a
            <lb/>Lordes Sone or a Gentilman havyng in his possession or other to his use Landes or Tenement<am/> or annuytyes at the
            <lb/>leste for Terme of Lyffe to the yerely valewe of an hundreth pounde above all repryses, uppon payne to forfeyte
            <lb/>the same apparell wherwyth so ev<am/> yt be myxte and for usyng of the same to forfett xl Shillyng<am/>. Nor no p<am/>sone use
            <lb/>or were Satten or Damaske in ther Doblett<am/> nor Sylke or Chamlett in their Gowenes or Cootes nott havyng Landes
            <lb/>or Tenement<am/> in hys possession or other hys use Office or Fee for terme of Lyffe or Lyff<am/> to the yerely value of
            <lb/>xx<hi rend="sup">ti</hi> Pounde, excepte he be a Yoman of the Crowne or of the Kyng<am/> garde or gromes of the Kyng<am/> Chambre
            <lb/>the Qwenys having therfore the Kyng<am/> Fee or the Qwenes uppon payne to forfett the same apparell wherwyth
```

Access – I (WIP)

- Storage of encoded texts in relational schema using PostgreSQL
- CSV files from Prepare easier to read by Postgres to create tables
- Relational schema -> efficient storage, faster retrieval
- Additional support for ‘to_tsvector’ , ‘to_tsquery’, complex XPath querying

Access – II (WIP)

Potential Outlook of the Relational Schema:



Texts

PK	text_id int NOT NULL
	title varchar2(100) NOT NULL author varchar2(50) date datetime

Sections

PK	section_id int NOT NULL
FK1	text_id int NOT NULL
	section_type varchar2(100) NOT NULL

Proses

PK	prose_id int NOT NULL
FK1	section_id int NOT NULL
	para_num int NOT NULL

WordOccurences

PK	occurence_id int NOT NULL
FK1	token_id int NOT NULL
FK2	word_id int NOT NULL

Tokens

PK	token_id int NOT NULL
FK1	verse_id int NOT NULL
	text varchar2(100) NOT NULL

Verses

PK	verse_id int NOT NULL
FK1	prose_id int NOT NULL
	verse_num int NOT NULL

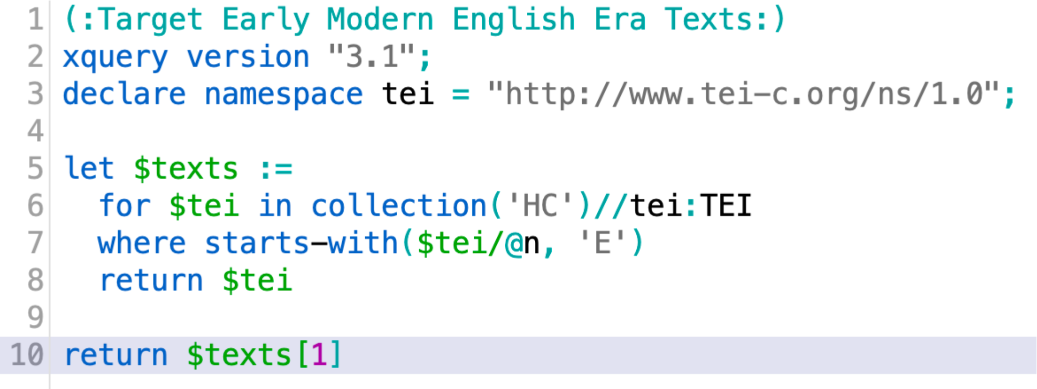
Words

PK	word_id int NOT NULL
	text varchar2(100) not null

Extension

- XQuery: SQL-like query access to XML documents for extracting text and aggregations
- Built on XPath

  Tree-like document structure with simplified access
- Ability to generate CSV files



```
(:Target Early Modern English Era Texts:)
xquery version "3.1";
declare namespace tei = "http://www.tei-c.org/ns/1.0";

let $texts :=
  for $tei in collection('HC')//tei:TEI
  where starts-with($tei/@n, 'E')
  return $tei

return $texts[1]
```

General Challenges

- Lack of Lemmatization
  - Important for Morphological Regeneration of texts
- Lack of Sentiment Analysis
  - Punctuation discarded
- Complex Architecture
  - Migration from one database to another
- Suboptimal Storage
  - Duplication of Tokens

Bibliography – I

[1] *Paradise Lost*. 1684. John Milton.
https://www.poetryfoundation.org/poems/45740/paradise-lost-book-4-1674-version

[2] *Auction: Poems*, 2023. Quan Barry
https://books.google.de/books?id=VgHaEAAAQBAJ&pg=PT5&source=gbs_selected_pages&cad=1 - v=onepage&q&f=false

[3] *Love-Letters Between a Nobleman and His Sister*. 1700. Aphra Behn
https://www.gutenberg.org/cache/epub/8409/pg8409-images.html

[4] *People We Meet on Vacation*. 2021. Emily Henry
https://books.google.de/books?id=Ov3MEAAAQBAJ&pg=PT9&source=gbs_toc_r&cad=1#v=onepage&q&f=false

Bibliography – II

[5] Left Brain vs. Right Brain: What Does This Mean for Me? 2024. Ann Pietrangelo
https://www.healthline.com/health/left-brain-vs-right-brain

[6] Helsinki Corpus TEI XML Edition. 2011. First edition. Designed by Alpo Honkapohja, Samuli Kaislaniemi, Henri Kauhanen, Matti Kilpiö, Ville Marttila, Terttu Nevalainen, Arja Nurmi, Matti Rissanen and Jukka Tyrkkö. Implemented by Henri Kauhanen and Ville Marttila. Based on The Helsinki Corpus of English Texts (1991). Helsinki: The Research Unit for Variation, Contacts and Change in English (VARIENG), University of Helsinki.

[7] Early Modern English: an Overview. Edmund Weiner, OED Deputy Chief Editor
https://www.oed.com/discover/early-modern-english-an-overview/?tl=true

Q&A