Shakespeare Project

**Robert Terbach, Simon Vetter,
Lukas Fischer, Thomas Döring,
Lukas Kairies**

Current Topics in Digital Philology

Shakespeare's words

- from Oxford English Dictionary

(http://www.oed.com/)

- Texts therefore unnecessary

Project Gutenberg

- Originally we wanted Project Gutenberg
- Varying headers and footers → need to fix manually

Alternatives

- Found other corpora with different kinds of text, not only books
- Two sources: KU Leuven, University of Oxford Text Archive

The English language of the north-west in the late Modern English period: a Corpus of late 18c Prose

(http://www.ota.ahds.ac.uk/desc/2468)

- Text types: Prose, Letters
- Years: 1761-1790
- Texts: 1800 (very short each)
- Words: 300,000
- License: restricted, free for non-commercial, but needs permission

A Corpus of English Dialogues

(http://www.ota.ahds.ac.uk/desc/2507)

- Text types: dialogues, constructed and authentic
- Years: 1560-1760
- Texts: 177
- Words: 1,200,000
- License: restricted, free for non-commercial, but needs permission

The Lampeter Corpus of Early Modern English Tracts

(http://www.ota.ahds.ac.uk/desc/3193)

- Text types: tracts (e.g. political)
- Years: 1641-1739
- Texts: 120
- Words: > 1,000,000
- License: CC-BY-SA 3.0 (credits, free, but result must also be CC-BY-SA 3.0)

Corpus of English Novels

(https://perswww.kuleuven.be/~u0044428/cen.htm)

- Text types: novels, british & north american
- Years: 1881-1922
- Texts: 290
- Words: > 26,000,000
- License: unknown, request free userid/password

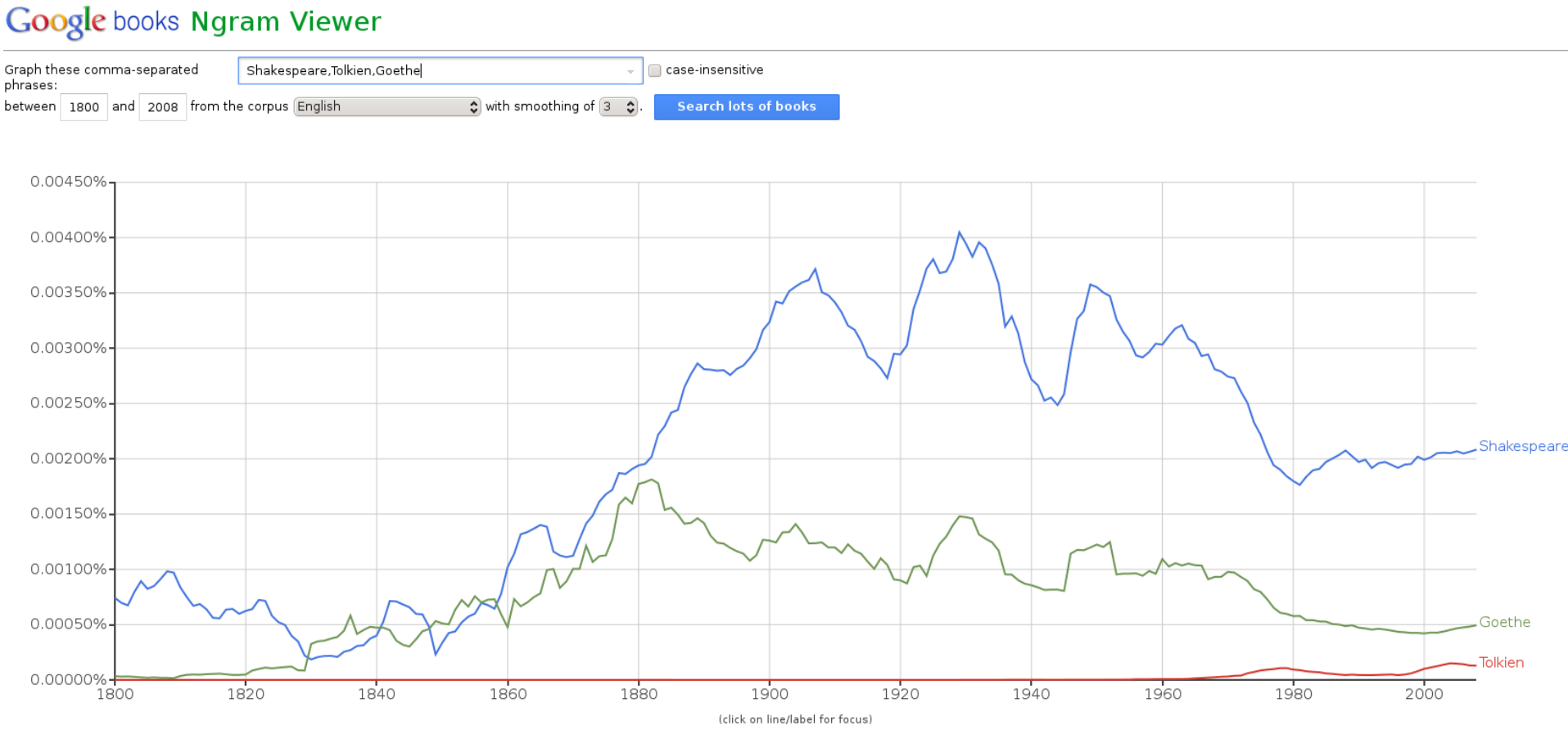
The Corpus of Late Modern English Texts (CLMET)

(https://perswww.kuleuven.be/~u0044428/clmet.htm)

- Text types: various (e.g. letters, novels, drama, journals)
- Years: 1710-1920
- Texts: 333
- Words: > 34,300,000
- License: unknown, request free userid/password

Our questions

- How often does each word occur per year



Our questions

- Which author used most shakespeare words?
- In which book are most shakespeare words?
- Rank which shakespeare words occurred most per year

Why are we asking this

- Why are these questions interesting?
  - Because there is literature about Shakespeare created without the use of the internet
  - Get as much information as possible from the gathered data
  - Can we find new interesting facts

Why are we asking this

Shakespeare-Wörterbuch von 1900: https://archive.org/details/shakespearewordb00fostuoft

- Opensource Shakespeare - http://www.opensourceshakespeare.org/info/paper_toc.php
- Welche Wörter kannte Schakespeare, die er nicht aufschrieb? http://biomet.oxfordjournals.org/content/63/3/435.short
- Coined by Shakespeare: Words and meanings first used by the Bard, http://www.getcited.org/pub/100261442

Why are we asking this

- Why should they be of interest to others?
  - Shakespeare is an important writer of his time

Unsere Methoden

- Vorverarbeitung:
  - Problem: Verschiedene Corpora - verschiedener Textaufbau
  - Lösung → Überführung in einheitliches XML Format
  - Für spezifische Fragestellungen jedoch zeitaufwändig jedes XML einzeln zu parsen und zu prüfen, ob es für eine Fragestellung relevant ist
  - Lösung → Metadatendatenbank mit Verweisen
  - Zusätzlich sind mehrere Wortstemmingverfahren implementiert wurden (Porter2, Lovins, Paicehusk)

Unsere Methoden

- Vorverarbeitung:
  - Problem: Verschiedene Corpora - verschiedener Textaufbau
  - Lösung → Überführung in einheitliches XML Format
  - Für spezifische Fragestellungen jedoch zeitaufwändig jedes XML einzeln zu parsen und zu prüfen, ob es für eine Fragestellung relevant ist
  - Lösung → Metadatendatenbank mit Verweisen
  - Zusätzlich sind mehrere Wortstemmingverfahren implementiert wurden (Porter2, Lovings, Paicehusk)

XML Format

<xml>
    <author> ... </author>
    <title> ... </title>
    <year> … </year>
    <text>
        (<w>...</w>)*
    </text>
</xml>

Erstes Schema der Datenbank

- Text = (id,year,author_id,title,file)
- Author = (id,name,is_from)

- Erlaubt es schnell über die Texte zu suchen
- Mit einfachen Mitteln ist es möglich diese zu erweitern.
- Zurzeit benutzte DB-Engine: SQLite3 (automatisch bei Python dabei)

Wie analysieren wir zurzeit

- Zunächst statistische Messungen geplant.
- Beispiel: Zu wieviel Prozent kommen Shakespeares Wörter in bestimmten Texten/Jahren/Werken von Autoren vor
- Algorithmen die zurzeit benutzt werden:
  - Tf-idf
  - Prozentuale Rechnungen auf den Mengen der Wörter

Ergebnisse

- Mehrere Parser für verschiedene Texttypen für die vorhandenen Corpora und Überführung in gleiches XML Format
- Datenbank mit verschiedenen Metainformationen

→ Es ist ein System entstanden, welches einfach erweiterbar ist und für verschiedene Fragestellungen genutzt werden kann

- Es folgen bald die ersten Zahlen zu den Fragestellungen

The Outlook

- Next Tasks
  - Create Database
  - Run analysis

The Outlook

- Expected state at the end of the Vorlesungszeit
  - Diagrams like googles n-gram to display the occurrence of shakespeare words over time

The Outlook

- Expected state at the end of the Semester
  - Visualization of geographical dissemination (Ausbreitung) of the words

The Outlook

- Optional goals
  - Generalized software (independent of Corpus, words)
  - Extend our software to work with phrases (n-grams)
  - Webinterface to improve usability

The Outlook

- Possible truncation of the project
  - Reduce quality standards:

    unfixed bugs

    low quality source code

    inaccurate visualizations

Why you should fund our project?

- methodological advances
  - Open Source alternative to Google's Ngram Viewer
  - supports arbitrary corpora
  - uses more information than occurrence per year
    - More use cases

Why you should fund our project?

- What does it tell us about human history etc.?
  - how authors influence language in general
  - how knowledge distributes regarding geographical information about authors and place of publishing
  - Even more thinks depending on which information and corpora are included.