Computational Linguistics and Digital Humanities

John Nerbonne

Rijksuniversiteit Groningen &

Albert-Ludwigs-Universität, Freiburg

Tübingen

WS 2023-24

Hauptseminar Tübingen

› Winter semester 2023-24

› Idea today (2023-24): sketch issues in course, motivation

› Get an idea of student interest, experience – perhaps to aid in selection of material

Plan

› What's DIGITAL HUMANITIES?

› Two new aspects: scale and replicable methods

› But also:

- Opportunities
- Need to engage, not just provide tools

› Where's it going? How successful will it be?

› Some skepticism from other DHers (Thaller)

› Wrap up, encouragement!

What are the humanities?

> › Humanities aims to understand
>> – Expressions of the human mind (Hegel)
>> – For example, art, architecture, history, language, literature, media, music, material culture, …
>> – But humanities also includes anthropology, archaeology, philosophy, religious studies, …
> › DH provides new tools, opportunities

What's DH?

› … for the purpose of this course

› Humanities studied digitally

› Digital humanities
  – Humanities scholarship where the computer plays an essential role, e.g. culturomics (Google books) study
  – Beyond tasks such as word processing, literature search, email, …

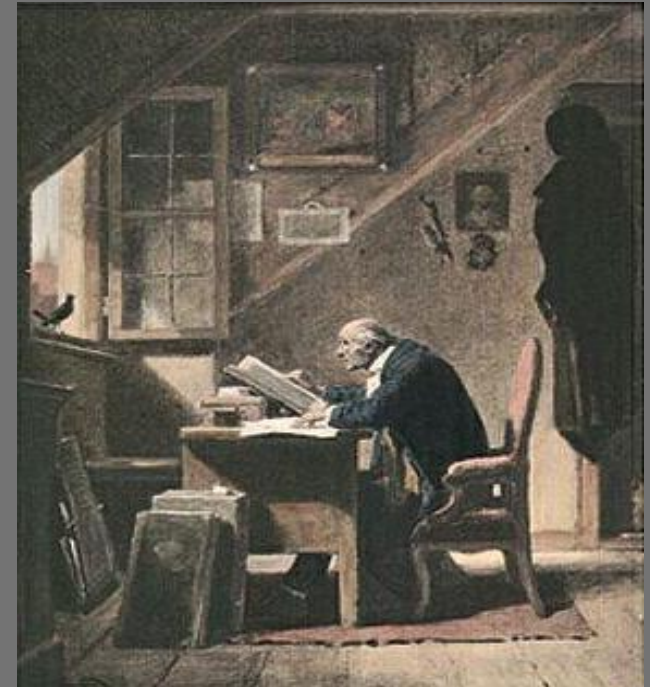› Focus here on language and text-related topics (closest to CL)
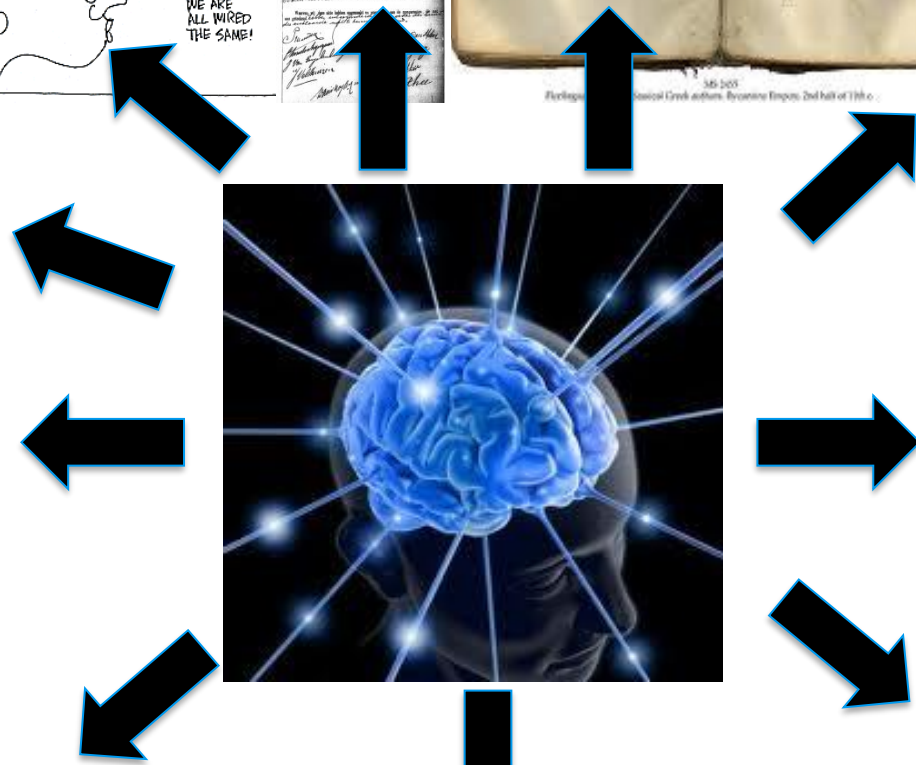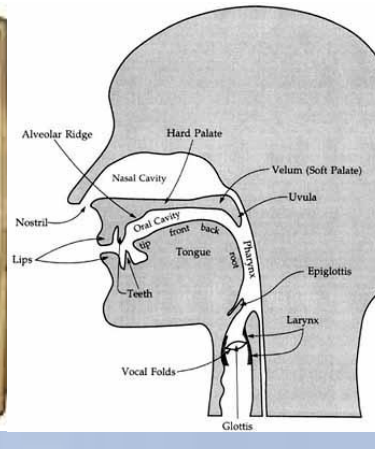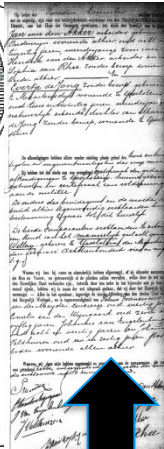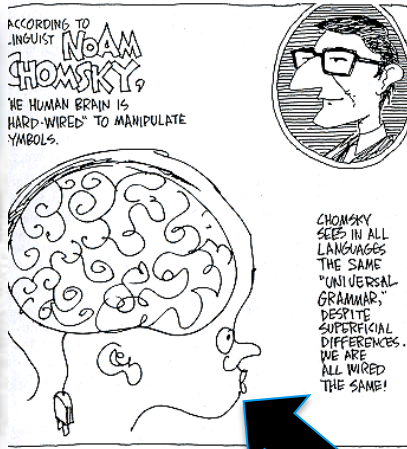
Plan

› What's DIGITAL HUMANITIES?

› Two new aspects: scale and replicable methods

› Where's it going? How successful will it be?

› Part II, later

What's DH add to understanding?

› Replicable techniques
  · …computing is replicable!

› More precise definitions

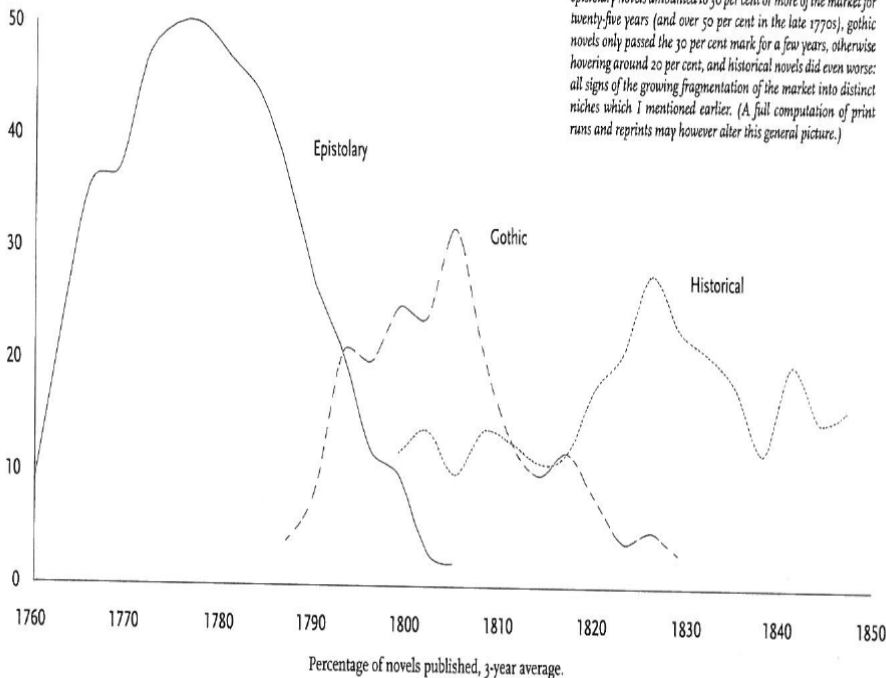› Scale!

› Research questions evolve more slowly

Spitzweg's "lonely scholar" no longer suffices

Larger scale views

FIGURE 8: *Market quotas of British hegemonic forms, 1760–1850*

All works of art, and not only parodies, are created either as a parallel or an antithesis to some model. The new form makes its appearance not in order to express a new content, but rather to replace an old form that has already outlived its artistic usefulness.

Viktor Shklovsky, *A Theory of Prose*

As more and more novels are published every year, the hegemony of a single genre tends to become less and less absolute: whereas epistolary novels amounted to 30 per cent or more of the market for twenty-five years (and over 50 per cent in the late 1770s), gothic novels only passed the 30 per cent mark for a few years, otherwise hovering around 20 per cent, and historical novels did even worse: all signs of the growing fragmentation of the market into distinct niches which I mentioned earlier. (A full computation of print runs and reprints may however alter this general picture.)

Epistolary

Gothic

Historical

Percentage of novels published, 3-year average.

> Harvard-Google culturomics (<u>TED Talks</u>)

> Moretti's "distant reading"

> Jocker's work on 19th cent. novels (below)

O quam cito transit gloria mundi!

› Thomas à Kempis, 15th cent.

› Warhol's 15 min. fame for all – old idea!

  · So what's new?

› Culturomics measures!

  · ½-life cut 50% in 50 yr
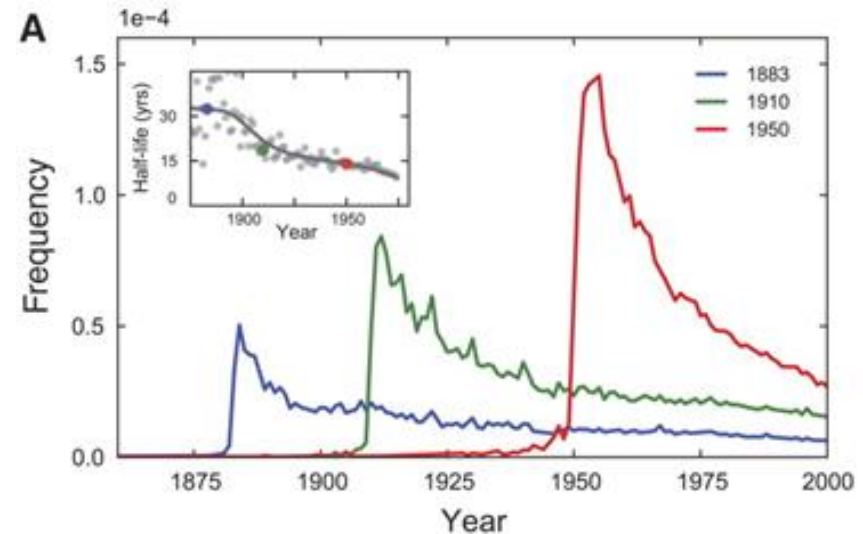
› In general, new techniques lead to new ideas

SIN_AND_SALVATION
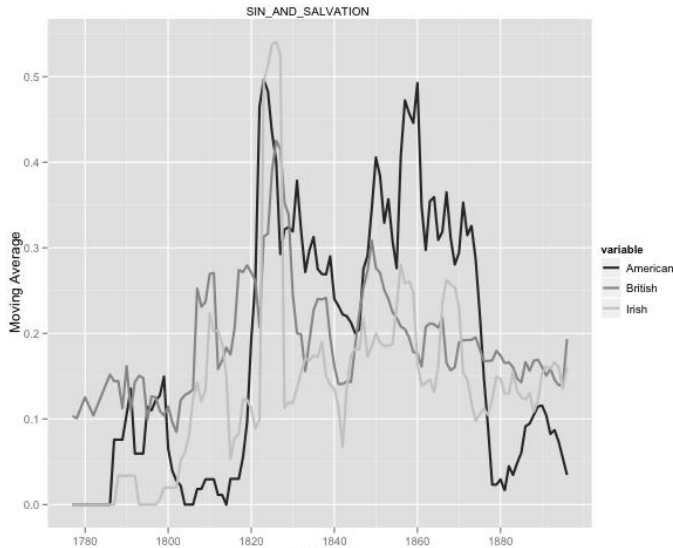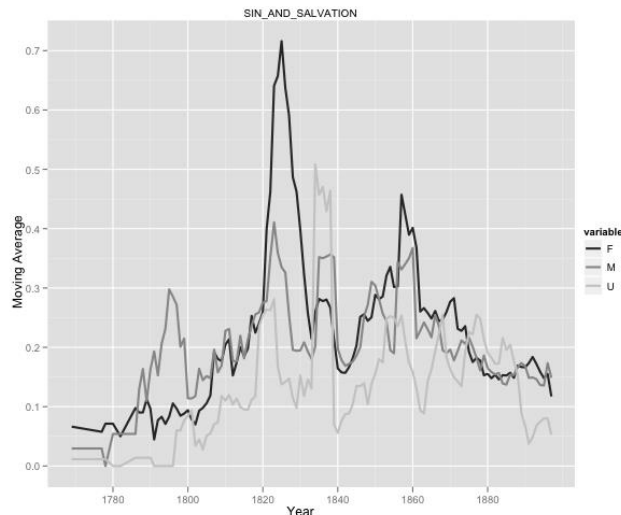

SIN_AND_SALVATION

Jockers' 19 cent. novels

› 500 themes analyzed in 3,346 American, British & Irish novels
- Themes "clustered" automatically
- Word based

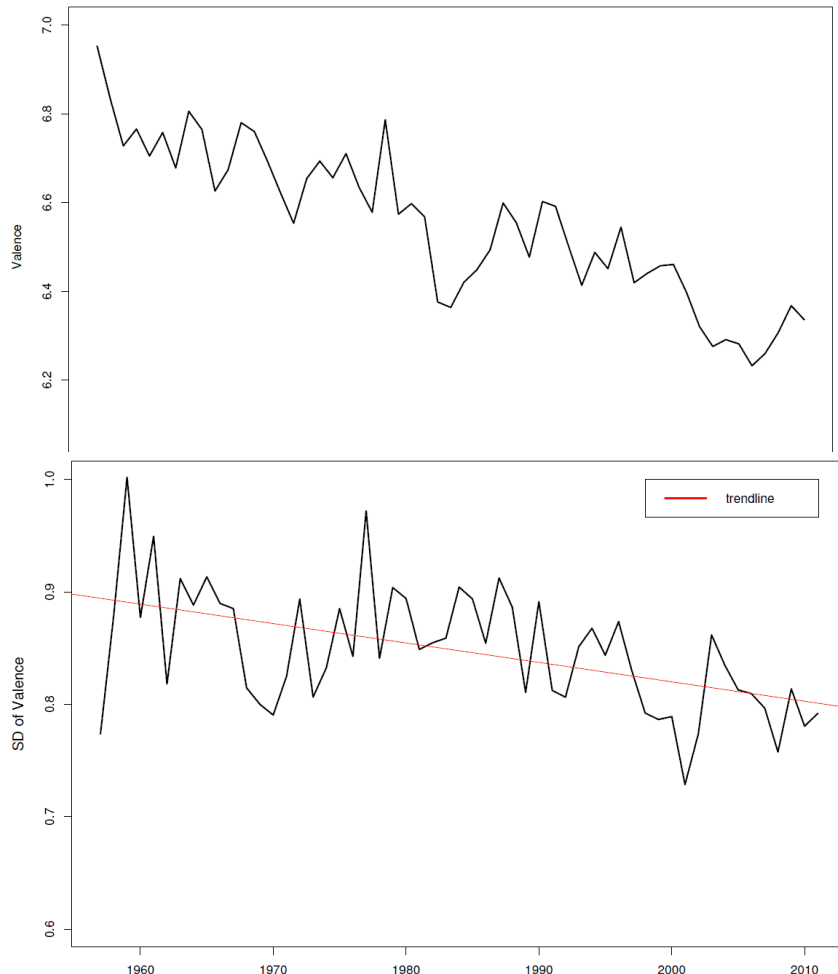› LATENT DIRICHLET ALLOCATION

› Exmpl: Sin and salvation
- Nationalities (above), M/F (below)

› www.matthewjockers.net/macroanalysisbook/

Pop music

> Peter Meindertsma (Gron.)
> Lyrics of 21,000 top-100 pop songs analyzed (1956-2011)
>   - ~ 80% of top-100 songs
> Measured average VALENCE (pleasantness) per song
>   - 'love' ~ 8.72
>   - 'suicide' ~ 1.25
> Valence also falls in variability (lower graph)
> Linked to massive concentration in music industry?
>   - 6 firms control 85%

Plan

> What's DIGITAL HUMANITIES?

> Two new aspects: scale and replicable methods

> Where's it going? How successful will it be?

> Part II, later

How is it going?

› DH is on the rise

› Centers at Harvard, Oxford, Stanford, Chicago, Amsterdam, …

› 14 chairs in Germany (DUZ 12/2013)

› 600+ attendance at DH 2015 (Lausanne)

› Profiting from increased availability of data

› Attractive educationally because of the computer skills it provides

› But what does it add to understanding?

This course

› Open to students in CL and in humanities fields

- So technical topics will not dominate
- If you're interested, you can develop the technical skills (later)

› One 15-20-min. presentation from each student

- Choose subject early (from syllabus)
- Outline (min. ½ page) 2 wk before presentation
- Draft of slides 1 wk before presentation

More introductory remarks

> … but later!

> I want to hear from you about your interests

> … then jump into a cautionary presentation next class

> … in order to motivate some practical work next week

> So I'll continue with the general remarks later

Interlude

> Poll of participants

- Name, major study, CL experience,
- Any special interest in a DH topic?
- Any special interest in other humanities subjects?
  - Literature, philology, history, …