Mapping the Inner Lives of Characters in the European Novel 1840-1920

Tamara Radak (University of Vienna), Lou Burnard (Independent Scholar), Pieter Francois (University of Oxford & Alan Turing Institute), Fotis Jannidis (Justus-Maximilians-Universität Würzburg), Diana Santos (Linguateca & University of Oslo)

Distant Reading for European Literary History
Final COST Action Event
April 21st, 2022

Acknowledgements

Our sincere thanks to everybody who contributed data, ideas, questions, answers, and/or discussion points at several stages of the process:

Rosario Arias, Ronan Crowley, Marko Juvan, Cvetana Krstev, Jessie Labov, Luiza Marinescu, Stefano Ondelli, Gabór Palkó, Benedikt Perak, Michael Preminger, Antonija Primorac, Jan Rybicki, Christof Schöch, Ranka Stankovic, Janicke Stensvaag Kaasa, Aase Kristine Tveit, Andrejka Žejn

Structure

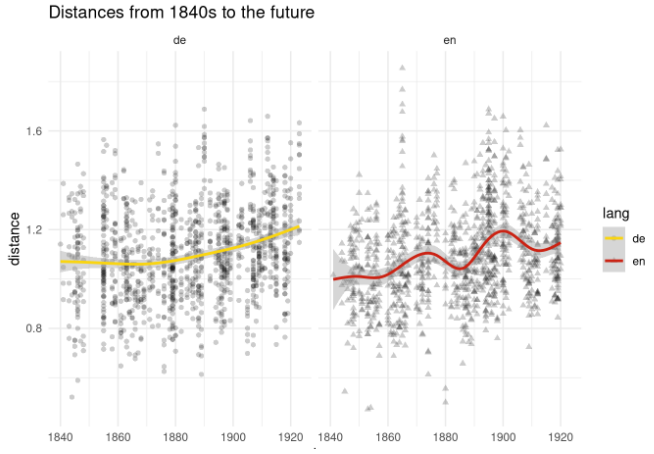
1. Aims and Hypothesis
2. Methodologies
3. Practical Procedure & First Results
4. Challenges and Next Steps

Aims and Hypothesis, Starting Point

- broad starting question: **How could questions regarding periodisation be operationalised for ELTeC (and potentially, other corpora)?**
- starting point: common narrative that the inner life of characters became a central preoccupation of the European novel in the mid- to late-nineteenth and early twentieth centuries → temporal frame commonly associated with (traditional accounts of) modernist novel and with psychological realism
- stylometry: general trend of increasing difference in style between earlier and later novels in ELTeC; graph below: perspective anchored in 1840; steadily increasing distance over time/two upticks: 1870s, around 1900

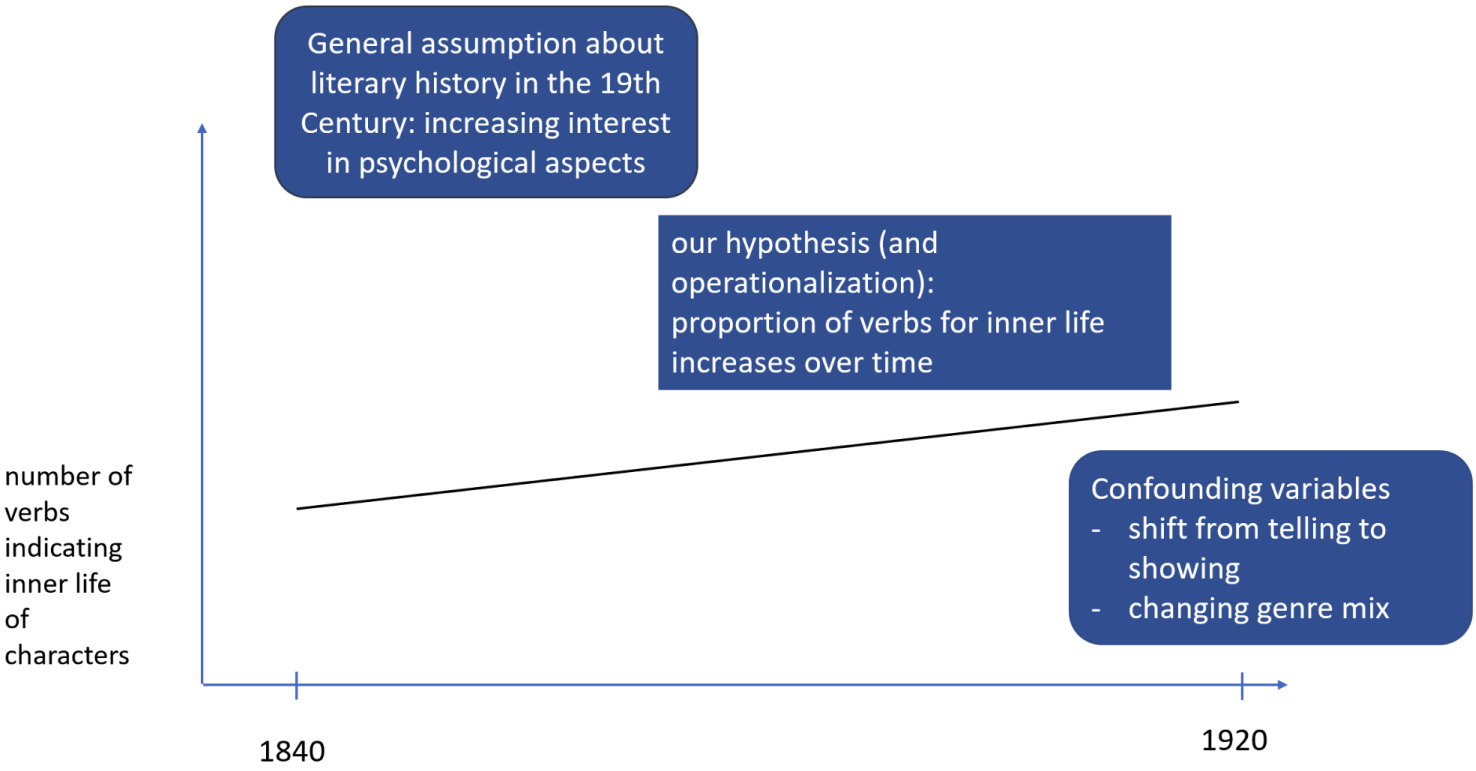
  → **Could the upward trend be connected to a stronger focus on the inner lives of characters in the later novels?**

  → **How can such a question be operationalised for ELTeC?**



Stylometric experiment conducted by Artjoms Šeļa (presented at the Belgrade Training School, March 2022)

https://github.com/distantreading/WG2/blob/master/Exploring_ELTeC_TS/1.2_Stylometry/Advanced%20hands-on/01_distance_and_time.html



Aims

- **long-term/overall aim**: operationalising questions of periodisation for the ELTeC corpus
- **short-term aim:** developing two methodologies and comparing them → is one method quicker than the other or less of a technical challenge? What are the benefits and disadvantages of each?
- explore possibilities as well as limitations of ELTeC for addressing questions relating to distant reading, literary history and periodisation
- make **explicit and visible** the process: choices made and challenges encountered during such a project often remain invisible but identifying and reflecting on them on a meta level could prove useful for future projects and researchers working with similar data or questions

Languages:

collections with **level-2 encodings**:

English, French, German, Hungarian, Norwegian, Portuguese, Romanian, Serbian, Slovenian

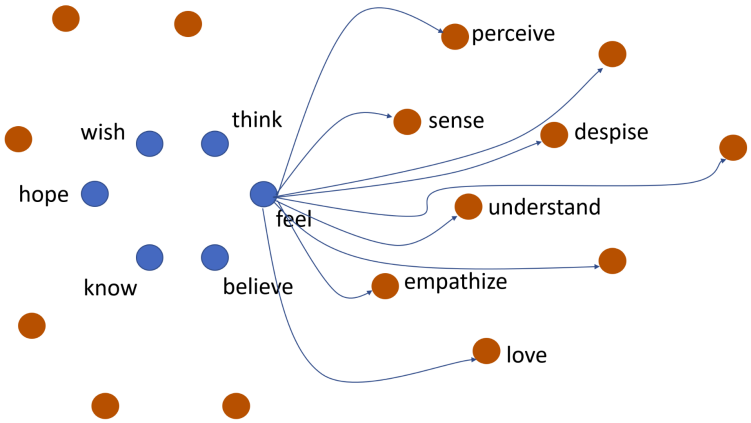
One Approach: Seed Words + Word Embeddings

Starting from six seed word in each language we collect ~ 100 words covering a wide range of different mental states by retrieving the nearest neighbours for each word using a language-specific word embedding.

Six seed words: *feel*, *think*, *believe*, *know*, *hope*, *wish*

Problems:

- requires filtering the noise from the list of nearest neighbours
- seed words have a cognitive bias
- wide variety of WE approaches



Another Approach: Most Frequent Verbs - Top Ten

Based on the sorted frequency counts of verb lemmas of the ELTeC collection, from level 2, human annotators selected the ten most frequent verbs related to inner life in each collection – both unambiguous ones or those that may also have other meanings, not relating to inner life

→ two lists of 10 verbs each, one unambiguous, one ambiguous (with obvious overlap)

Verbs that have one meaning relating to inner life and others which are not related to inner life

- Hungarian *bír* means own and like
- Portuguese *esperar* means wait and hope
- Romanian *aduce* means bring and remember
- English *touch*

Practical procedure

Both approaches resulted in a list of verb forms, one containing ca. 40-90 items and the other two lists containing ten verbs.

```
<list n='manual' xml:lang="por">
 <lemma form="ver" freq="20206" inner="m" trans="see/understand"/>
 <lemma form="saber" freq="17307" inner="m" trans="know/taste"/>
 <lemma form="querer" freq="13595" inner="y" trans="want"/>
 <lemma form="sentir" freq="5221" inner="y" trans="feel"/>
 <lemma form="conhecer" freq="5099" inner="y" trans="know"/>
 <lemma form="esperar" freq="4237" inner="m" trans="hope/wait"/>
 <lemma form="pensar" freq="3982" inner="y" trans="think"/>
 <lemma form="amar" freq="3202" inner="y" trans="love"/>
```

In the next step, we counted the occurrences of each of these verbs in each work, normalizing the count by the number of all verbs in that text.

```
textId year verbs innerVerbs aimer connaître croire entendre regarder
savoir sembler trouver voir vouloir
FRA00101 1860 3889 310  17 9 28 22 18 52 5 47 83 29
FRA00102 1883 5499 465  112 21 38 16 17 55 32 30 77 67
FRA00201 1910 7577 682  26 20 41 75 96 63 49 93 128 91
FRA00301 1858 16808 929  35 52 69 62 46 153 77 122 143 170
FRA00302 1894 25832 1613  101 120 134 105 79 262 97 212 269 234
```

First results

Most languages display a tilde / wave which was not really predicted. Only Romanian and English show a slight uptick in the end of the period.

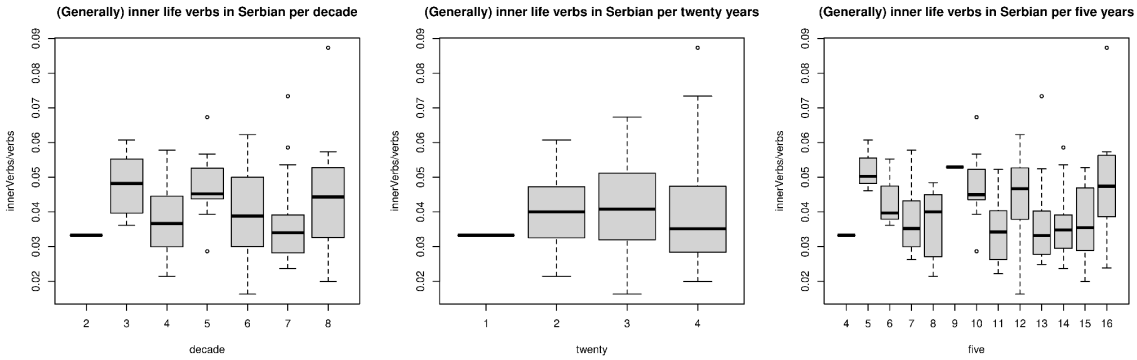
Full disclosure: https://foxglove.hypotheses.org/669 about the data gathering

Boxplots per

- decade
- 20 years
- 5 years

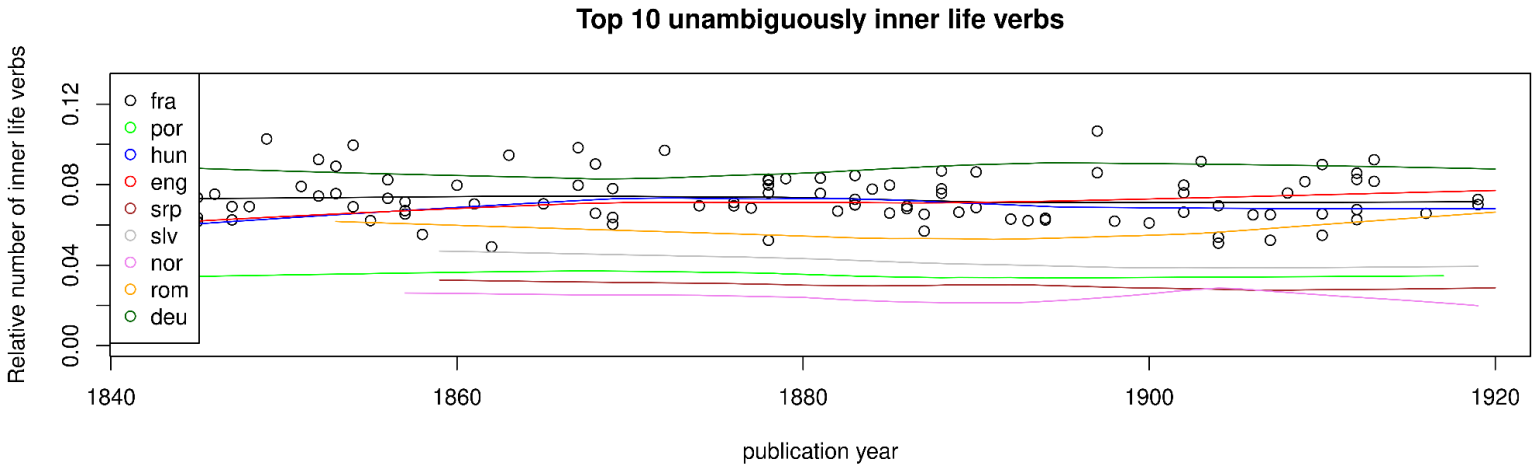
Serbian here:

Using all verbs

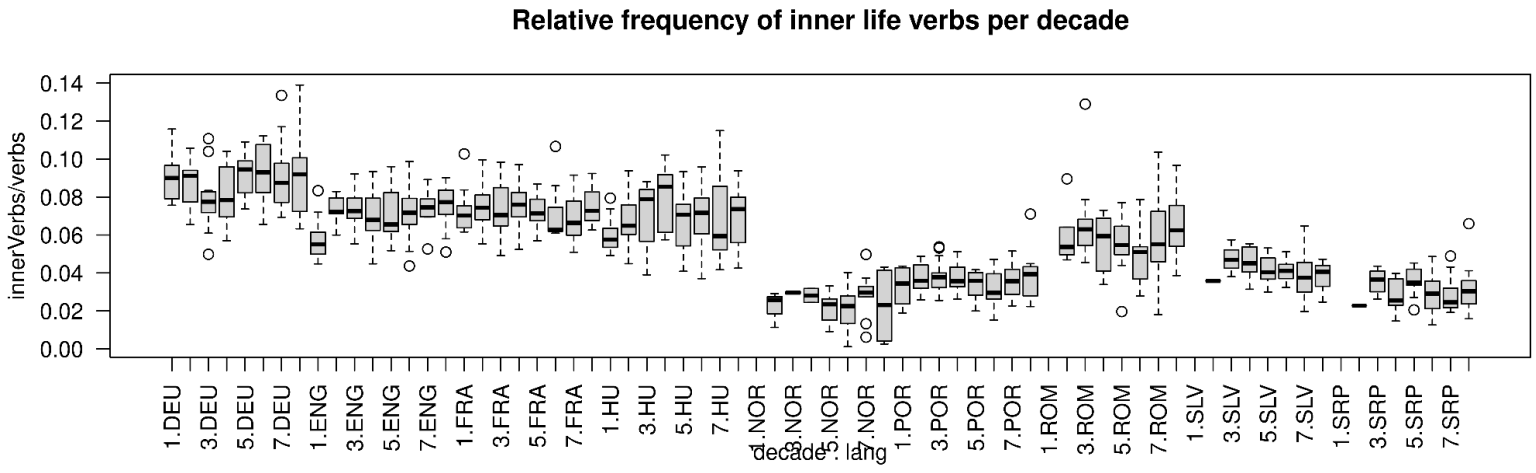


A bird’s eye view of all languages

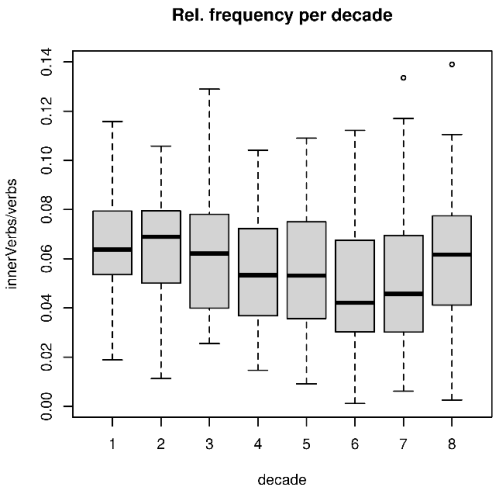
Using all data points and drawing a line:

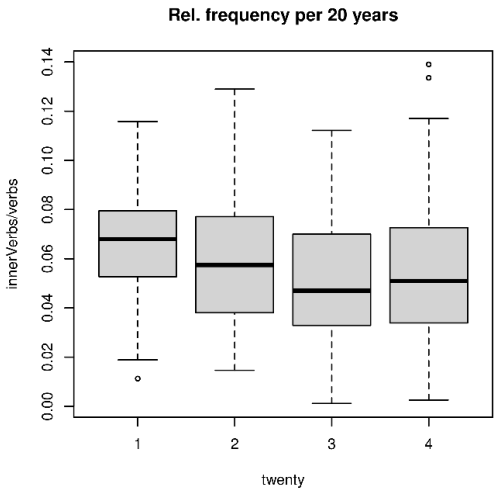


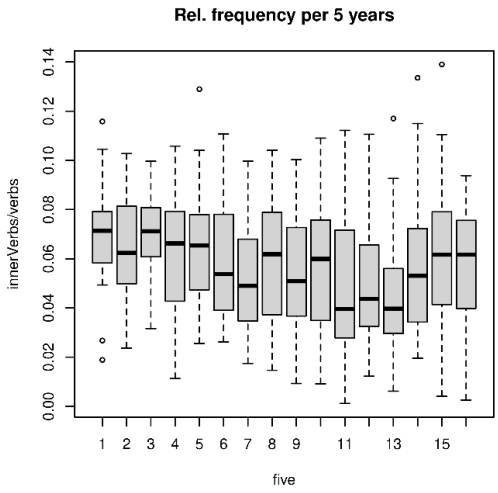
Looking at the progression per language, per decade



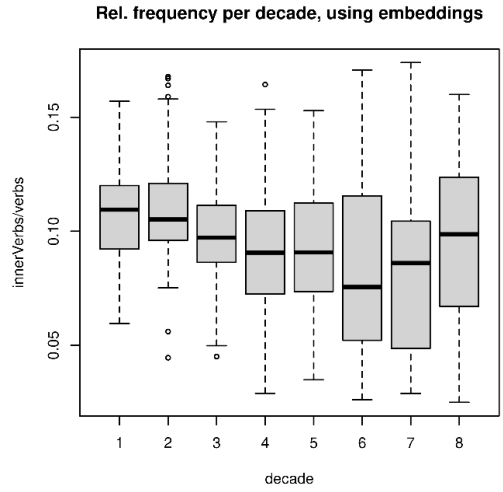
The whole collection of 9 languages all together

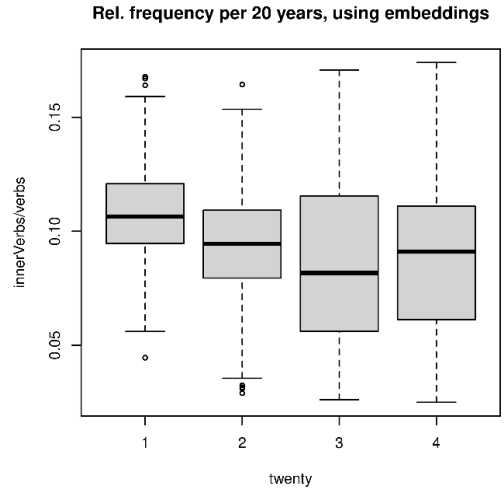


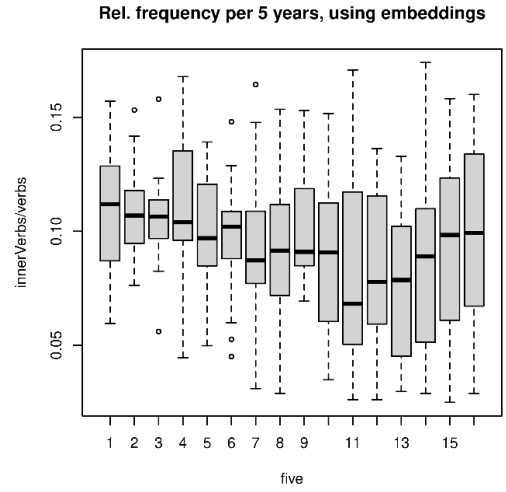




Using the embeddings results for four languages

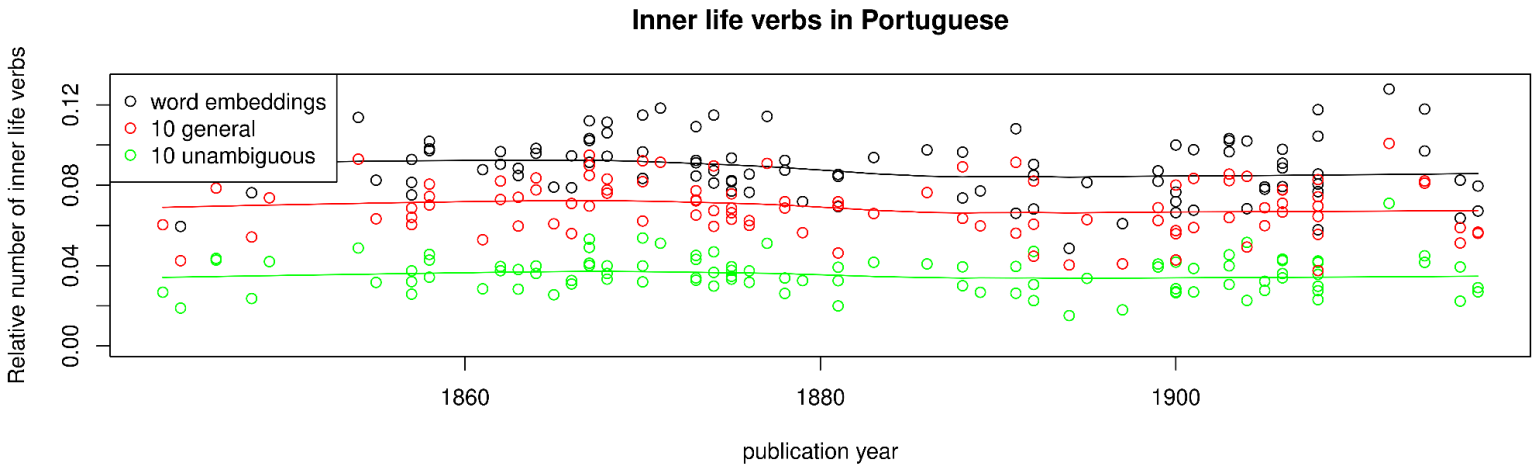




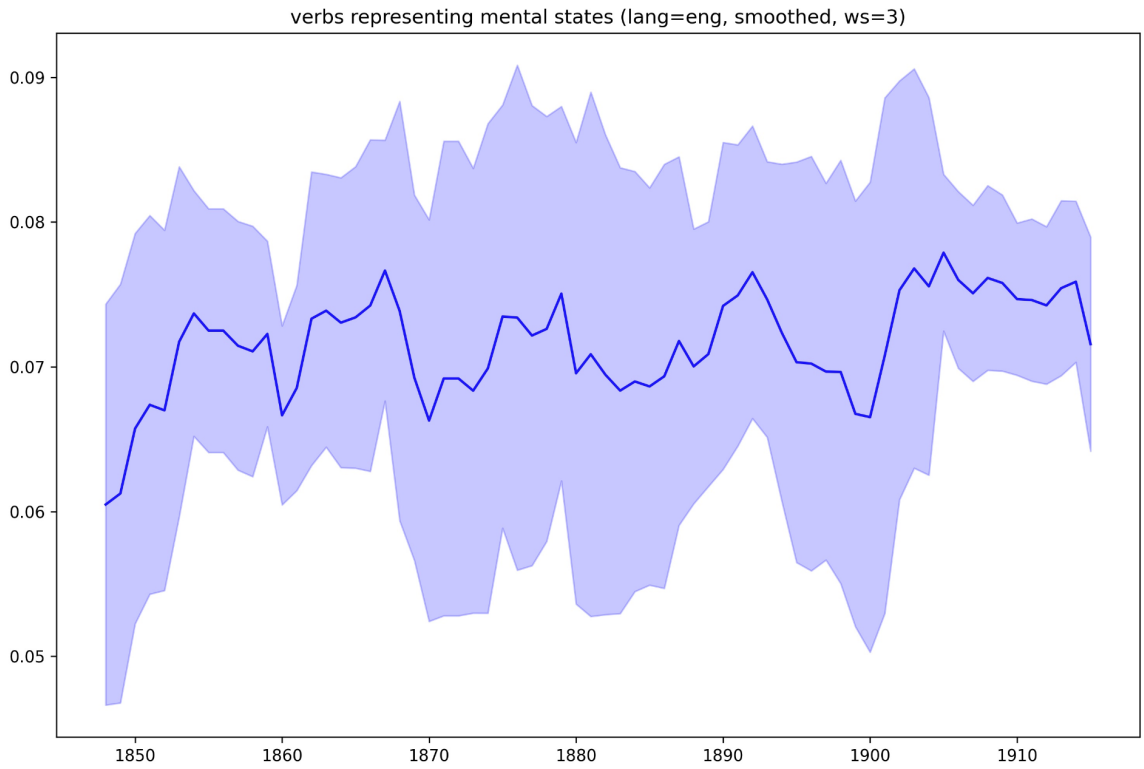


The two methodologies (top ten and word embeddings)

→ both methodologies pick up essentially the same signal



Smoothing with a rolling window of 3 years



Results and Challenges

https://github.com/COST-ELTeC/ELTeC-data/tree/main/Outputs

- short-term aim and comparative angle: both methodologies pick up similar signals and show similar trends → method using top ten frequent verbs yields comparable results and is easier to produce than the one involving word embeddings
- explaining/interpreting the visualised data is not straightforward

  → rather great deal of variety in the data

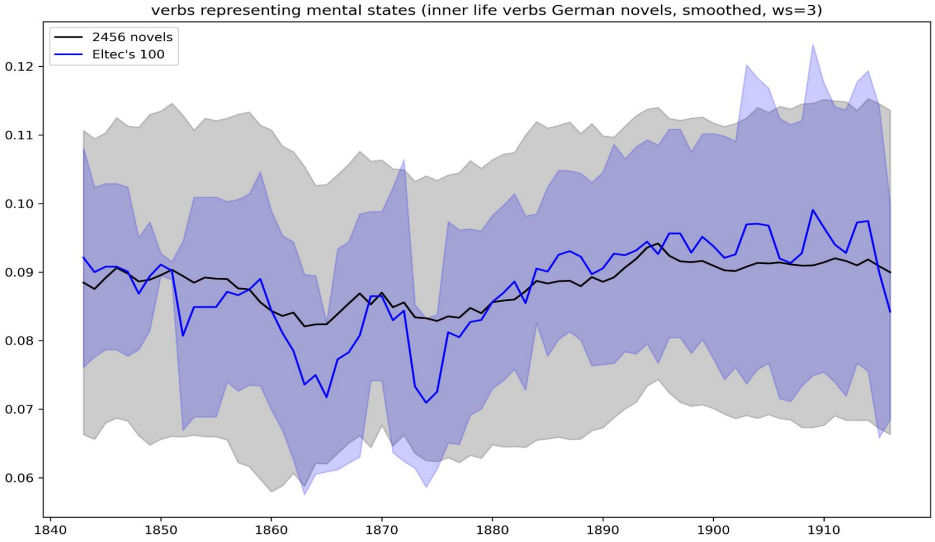
  → confounding variables that cannot entirely be eliminated → experiments planned

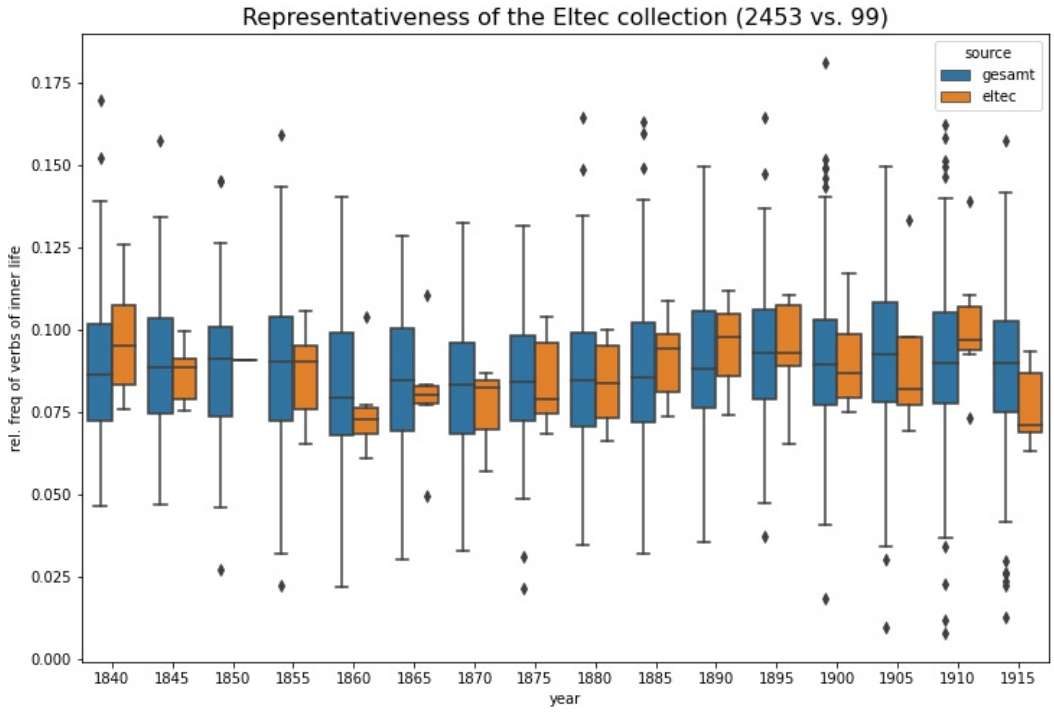
  → complexity of literary trends and techniques that may run counter to each other
- conceptual challenge: delineating inner life
- working on questions of periodisation with a collection that ends in 1920 → not all languages contain works that would be classified as (typically) “modernist” and the heyday of literary modernism occurred later in some literatures → expand approach to corpora that span a longer time span and compare the results
- how representative can a collection of 100 texts be for a time span of 80 years to capture shifts which are much smaller than the individual variations?

Next steps

- further testing and interpretation of data → could different trends be nullifying each other?
- if there is no perceivable trend in the data, this result also needs to be verified as such, using more data
- accounting for the confounding variables – e.g. with regard to the subgenres of the novels
- adjusting parameters:
  - expanding approach to corpora that span a longer/later time span and comparing the results
  - re(de)fining more explicitly and potentially narrowing down the definition of “inner life” employed (e.g. by focussing on different aspects of identification with characters in narrative texts representing perceptions, cognitions, evaluations, emotions and more (van Krieken et al. 2017))

Comparing with more data in the same period (German)





Thank you for your attention, feedback, and ideas!

Bibliography

Bradbury, Malcolm, and James Walter McFarlane. *Modernism: A Guide to Literature 1890-1930*. Penguin Books, 1991, 1976.

ELTeC collections. https://distantreading.github.io/ELTeC.

https://github.com/COST-ELTeC/ELTeC-data/tree/main/Outputs.

https://github.com/distantreading/WG2/blob/master/Exploring_ELTeC_TS/1.2_Stylometry/Advanced%20hands-on/01_distance_and_time.html
.

van Krieken, Kobie, et al. “Evoking and Measuring Identification with Narrative Characters – a Linguistic Cues Framework.” *Frontiers in Psychology*, vol. 8, 2017, p. 1190. doi:10.3389/fpsyg.2017.01190.

Underwood, Ted. *Distant Horizons: Digital Evidence and Literary Change*. University of Chicago Press, 2019.