Sentence length across ELTeC collections and Gutenberg Fiction

Distant Reading

Christof Schöch (Trier, Germany)

Distant Reading Closing Conference, April 21-22, 2022
https://christofs.github.io/krakow22/

Kompetenzzentrum Trier Center For Digital Humanities (TCDH)
Universität Trier
cost EUROPEAN COOPERATION IN SCIENCE & TECHNOLOGY
Funded by the Horizon 2020 Framework Programme of the European Union

Overview

1. The issue with sentence length
2. Methods used
3. Findings from Gutenberg Fiction
4. Findings from ELTeC collections
5. Influence of direct speech
6. Conclusion

The issue with sentence length

Why care about sentence length?

Why care about sentence length?

- It is a proxy for syntactic complexity and one aspect of readability

Why care about sentence length?

- It is a proxy for syntactic complexity and one aspect of readability
- It probably interacts with other features of texts, like narrator / character speech

Why care about sentence length?

- It is a proxy for syntactic complexity and one aspect of readability
- It probably interacts with other features of texts, like narrator / character speech
- It might vary also with first-person vs. third-person narration

Why care about sentence length?

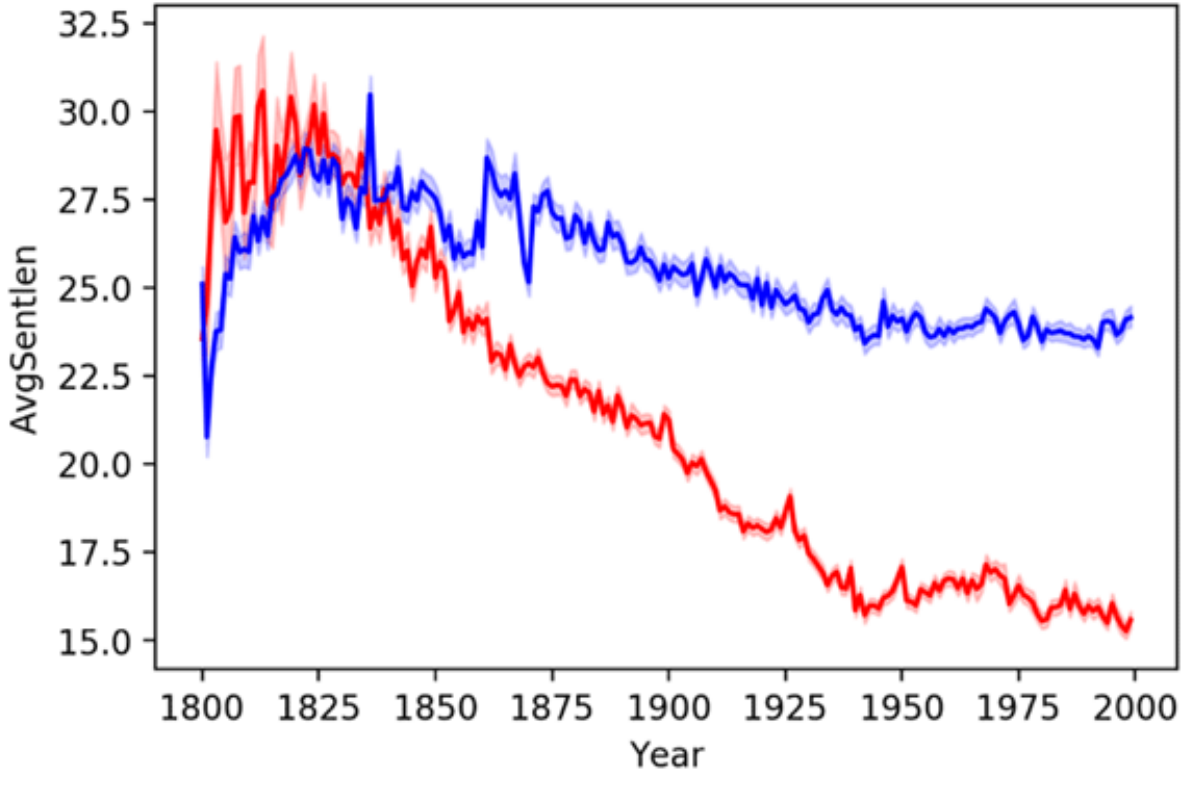
- It is a proxy for syntactic complexity and one aspect of readability
- It probably interacts with other features of texts, like narrator / character speech
- It might vary also with first-person vs. third-person narration
- People have assumed a decline of sentence length for a long time
  - but is it true, generally?
  - and specifically for different European literatures?

Biber and Conrad: English fiction

Date	Author	Novel	Average sentence length
1720	Daniel Defoe	*Life and Adventures of Duncan Campbell*	144
1720	William Pitts	*The Jamaica Lady*	44
1736	Eliza Haywood	*Adventures of Eovaai*	74
1751	Henry Fielding	*Amelia*	42
1764	Horace Walpole	*The Castle of Otranto*	27
1778	Clara Reeve	*The Old English Baron*	40
1818	Jane Austen	*Persuasion*	28
1828	David Moir	*The Life of Mansie Wauch*	24
1850	Herman Melville	*White-Jacket*	27
1880	Edward Bellamy	*Dr. Heidenhoff's Process*	26
1897	Stephen Crane	*The Third Violet*	18
1923	P. G. Wodehouse	*The Inimitable Jeeves*	25
1969	Kurt Vonnegut	*Slaughterhouse-Five*	15
1970	Saul Bellow	*Mr. Sammler's Planet*	13
1977	P. D. James	*Death of an Expert Witness*	16
1988	Toni Morrison	*Beloved*	20
1989	Robert Ludlum	*The Icarus Agenda*	18

- 500 words from each of 17 novels, 1720-1989
- Pretty clear decline, but very small sample
- Reference: Biber and Conrad 1989

Hathi Trust



- Fiction (red) vs. non-fiction (blue)
- Clear decline for fiction between ca. 1820 and 1940
- Source: Hathi 1M dataset; Bagga and Piper 2022

Method(s)

Sampling

- For ELTeC corpora, the full corpus is used
- For Gutenberg Fiction, random sampling is performed, partly with stratification by decade

How to establish average sentence length?

How to establish average sentence length?

- Basic approach: establish number of tokens and number of sentences

How to establish average sentence length?

- Basic approach: establish number of tokens and number of sentences
- Either: Using level2 encoding, use @type="SENT" as marker of sentence boundaries

How to establish average sentence length?

- Basic approach: establish number of tokens and number of sentences
- Either: Using level2 encoding, use @type="SENT" as marker of sentence boundaries
- Or: Using level1 encoding or plain text, use language-specific spacy tokenizer + sentencizer

How to establish average sentence length?

- Basic approach: establish number of tokens and number of sentences
- Either: Using level2 encoding, use @type="SENT" as marker of sentence boundaries
- Or: Using level1 encoding or plain text, use language-specific spacy tokenizer + sentencizer
- Strong correlation between these two approaches

How to establish average sentence length?

- Basic approach: establish number of tokens and number of sentences
- Either: Using level2 encoding, use @type="SENT" as marker of sentence boundaries
- Or: Using level1 encoding or plain text, use language-specific spacy tokenizer + sentencizer
- Strong correlation between these two approaches
- See Viera, Picoli and Mendes 2018 for a comparison of approaches

How to establish average sentence length?

- Basic approach: establish number of tokens and number of sentences
- Either: Using level2 encoding, use @type="SENT" as marker of sentence boundaries
- Or: Using level1 encoding or plain text, use language-specific spacy tokenizer + sentencizer
- Strong correlation between these two approaches
- See Viera, Picoli and Mendes 2018 for a comparison of approaches
- Scatterplot: Novels by publication year and average sentence length

Test for significant difference

Test for significant difference

- Equal-sized samples from early and later time slice

Test for significant difference

- Equal-sized samples from early and later time slice
- Density plot: for visual check of overlap and range

Test for significant difference

- Equal-sized samples from early and later time slice
- Density plot: for visual check of overlap and range
- Significance test: Mann-Whitney-U-Test

Findings based on Gutenberg Fiction

What do I mean by 'Gutenberg Fiction'?

What do I mean by 'Gutenberg Fiction'?

- Sample from the Gutenberg Project Corpus

What do I mean by 'Gutenberg Fiction'?

- Sample from the Gutenberg Project Corpus
- Downloaded using tool by Gerlach and Font-Clos 2018: 63.208 items

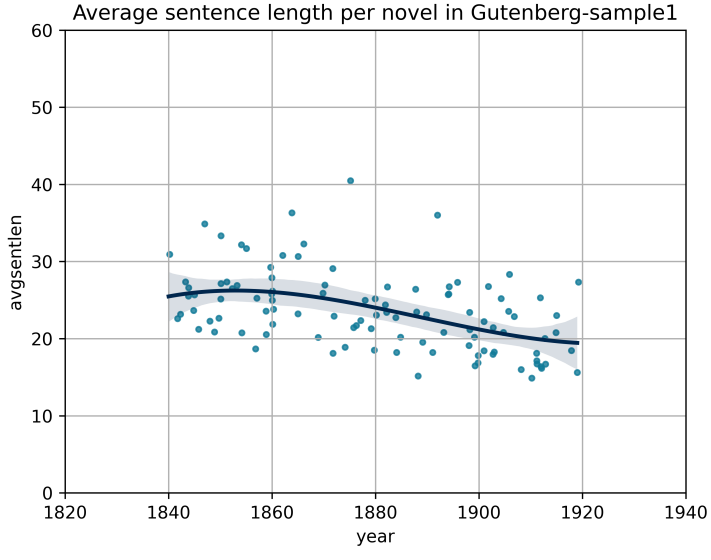
What do I mean by 'Gutenberg Fiction'?

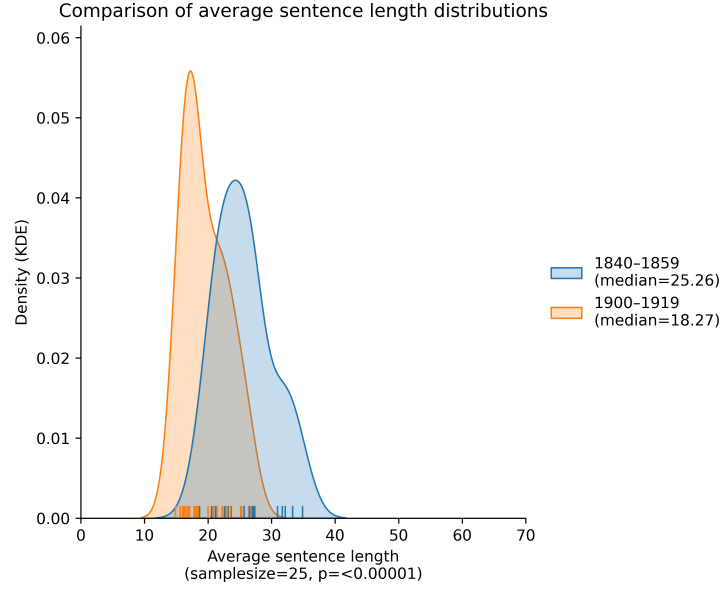
- Sample from the Gutenberg Project Corpus
- Downloaded using tool by Gerlach and Font-Clos 2018: 63.208 items
- Filtered out everything except English-language narrative fiction: 18.738 texts

What do I mean by 'Gutenberg Fiction'?

- Sample from the Gutenberg Project Corpus
- Downloaded using tool by Gerlach and Font-Clos 2018: 63.208 items
- Filtered out everything except English-language narrative fiction: 18.738 texts
- Established year of publication for many of them using several heuristics
  - Information from Wikidata
  - Information from Worldcat
  - Years of author's birth and death
  - Sanity checks

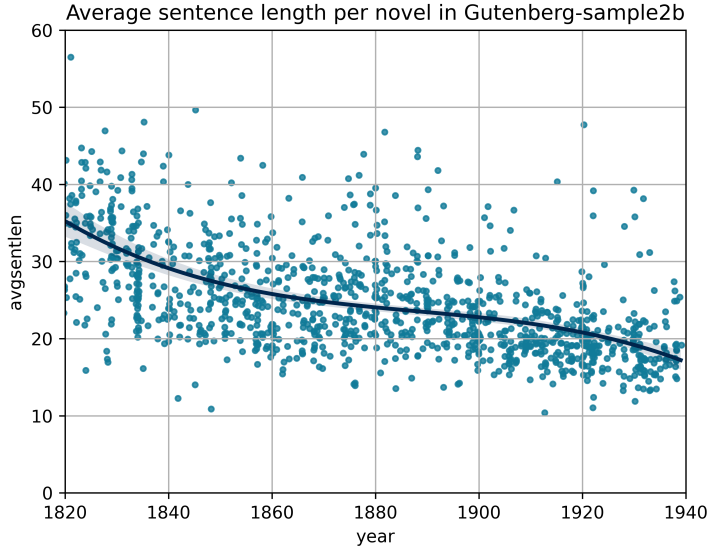
Gutenberg Fiction (100, 1840-1920)

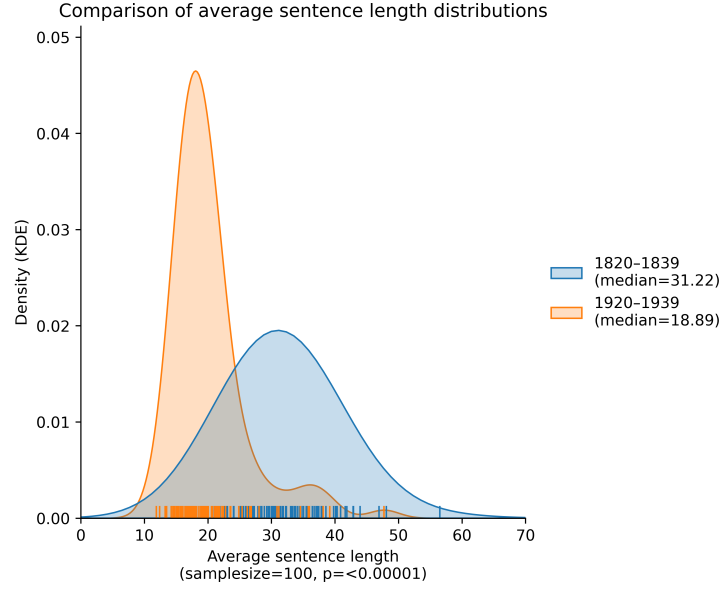




- Sample size: 100 novels
- Suitable for comparison with ELTeC-eng
- Significant difference in average sentence length

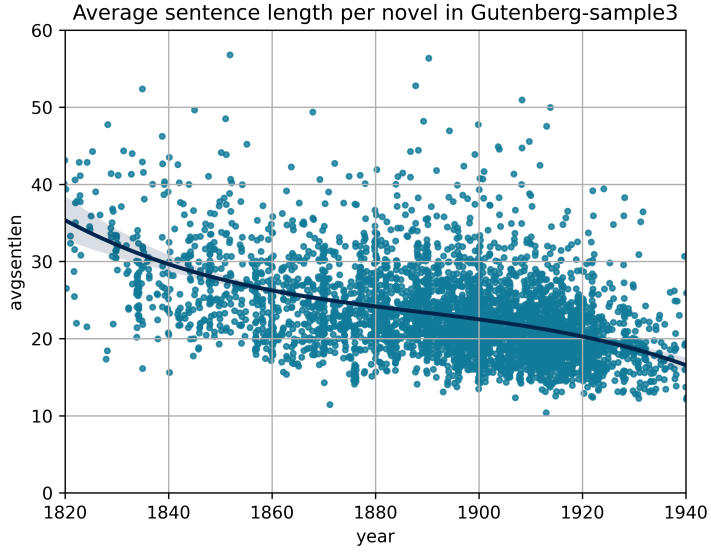
Gutenberg Fiction (1150, 1820-1920)

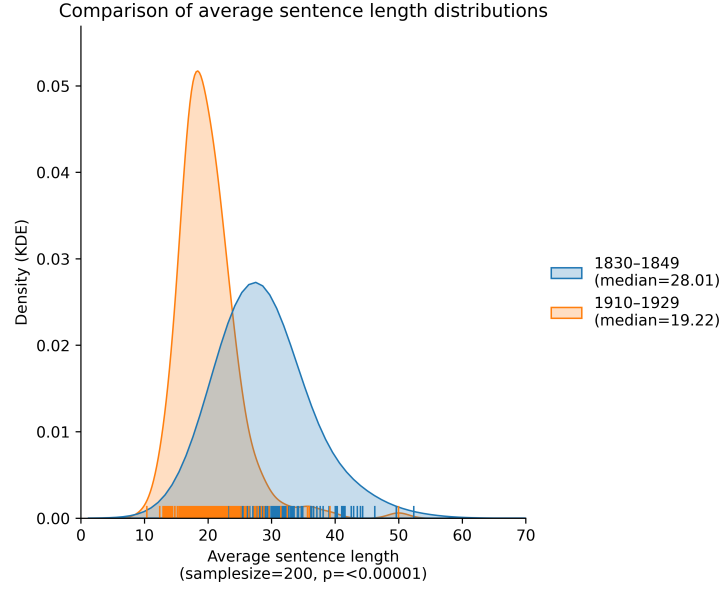




- Sample size: 1150 novels
- Longer period, chronologically-stratified sample
- Significant difference in average sentence length

Gutenberg Fiction (4080, 1820-1940)

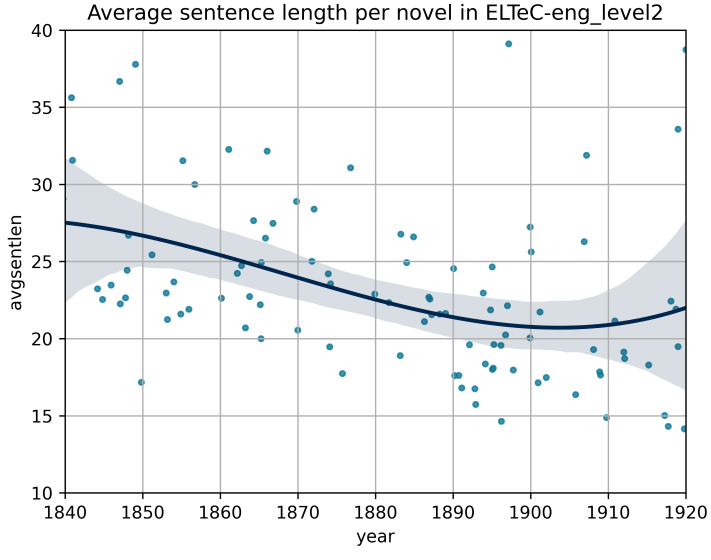


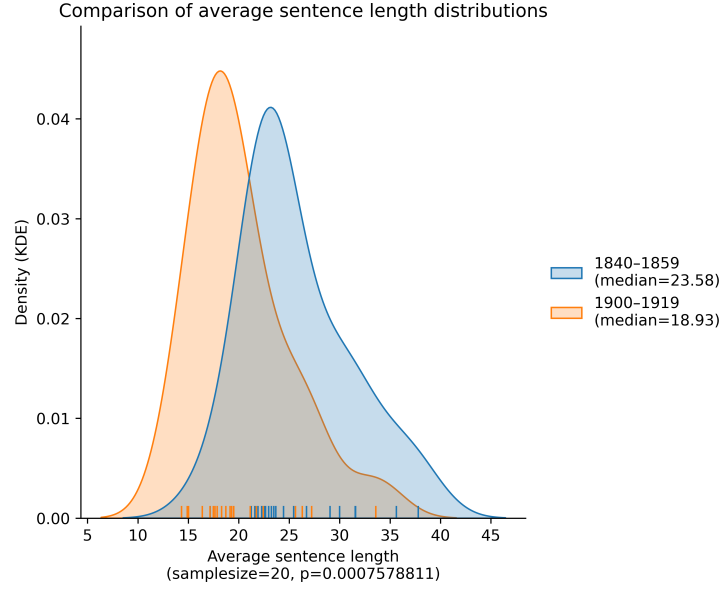


- Sample size: 4080 novels
- Longer period, unbalanced sample
- Significant difference in average sentence length

Findings based on ELTeC

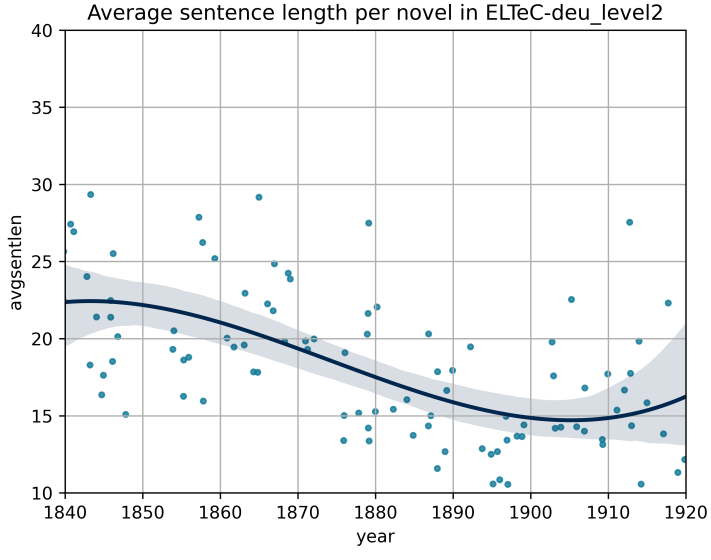
ELTeC-eng (1840-1920)

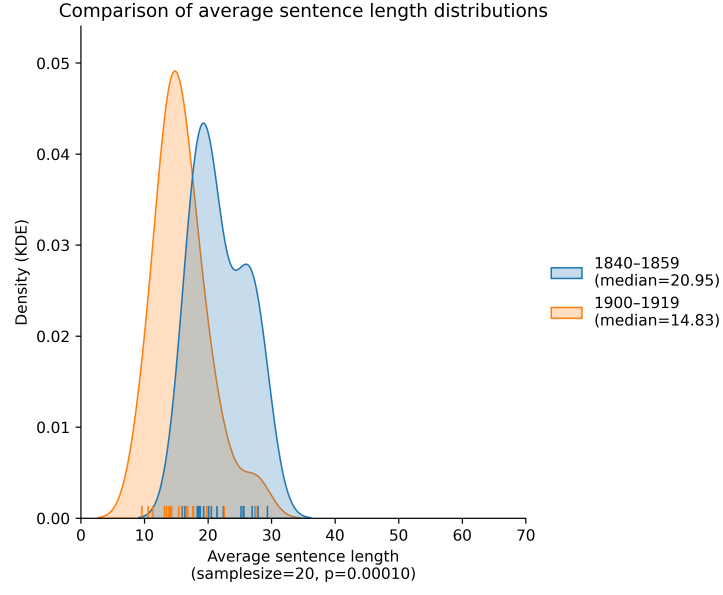




- Corpus size: 100 novels
- Standard period, quite even spread
- Significant decline in average sentence length
- Overall very similar to Gutenberg Fiction results

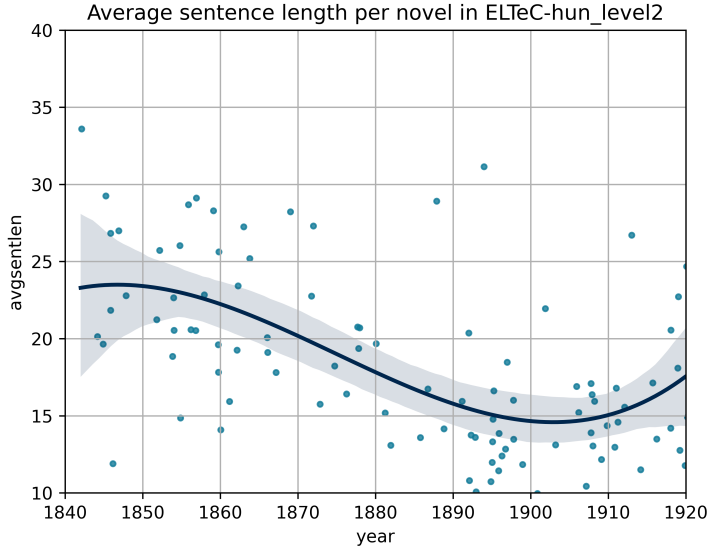
ELTeC-deu (1840-1920)

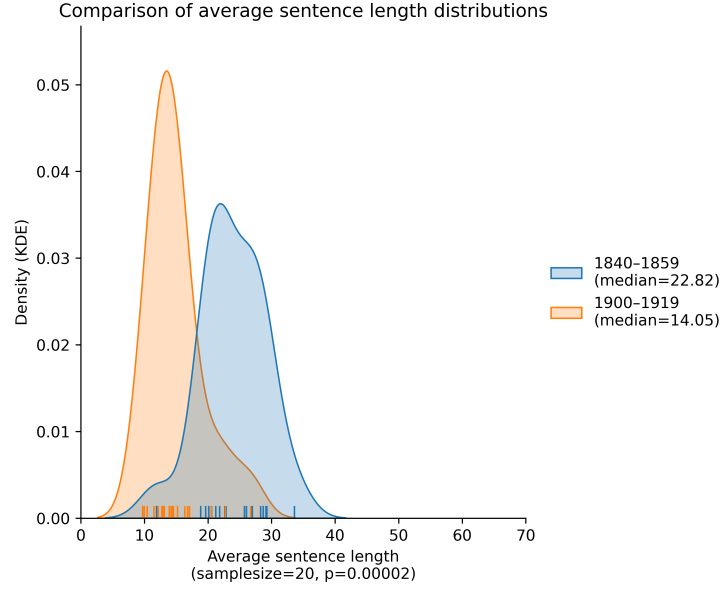




- Corpus size: 100 novels
- Standard period, quite even spread
- Significant decline in average sentence length

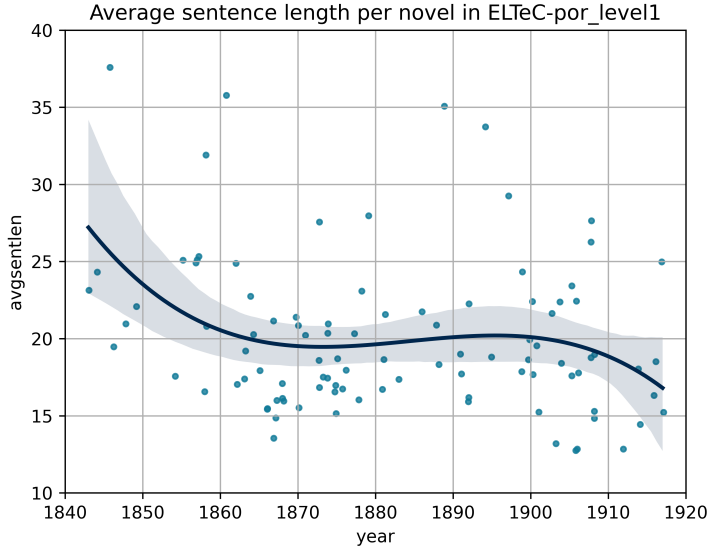
ELTeC-hun (1840-1920)

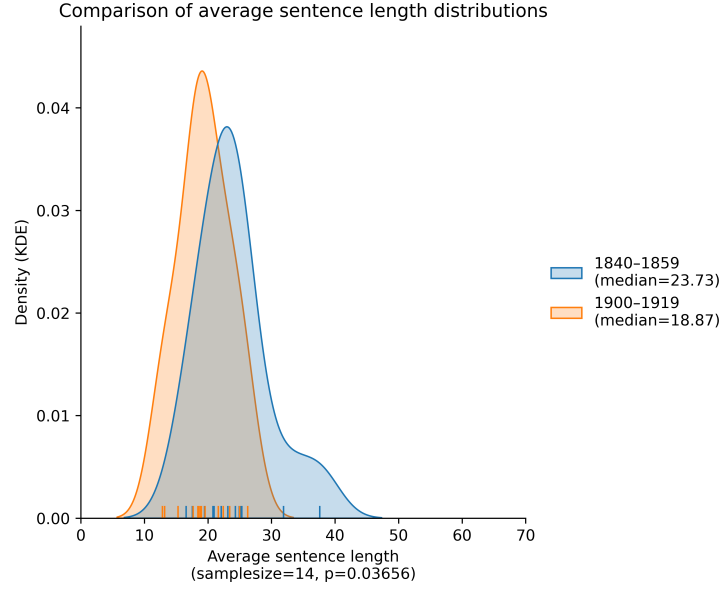




- Corpus size: 100 novels
- Standard period, quite even spread
- Significant decline in average sentence length

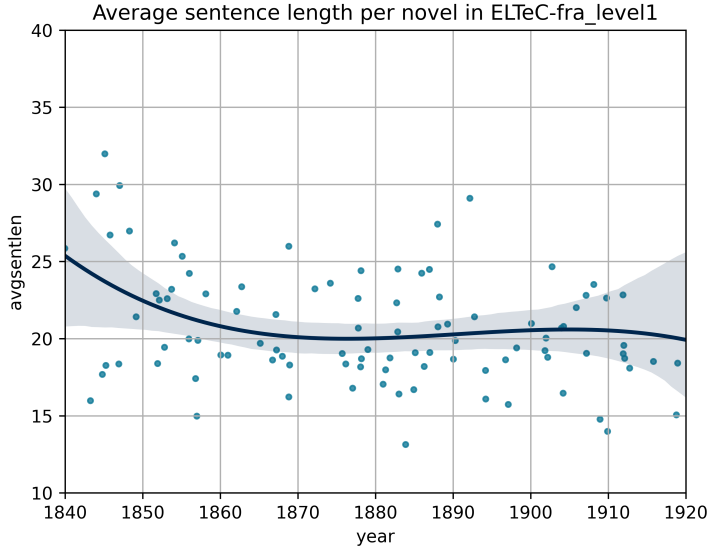
ELTeC-por (1840-1920)

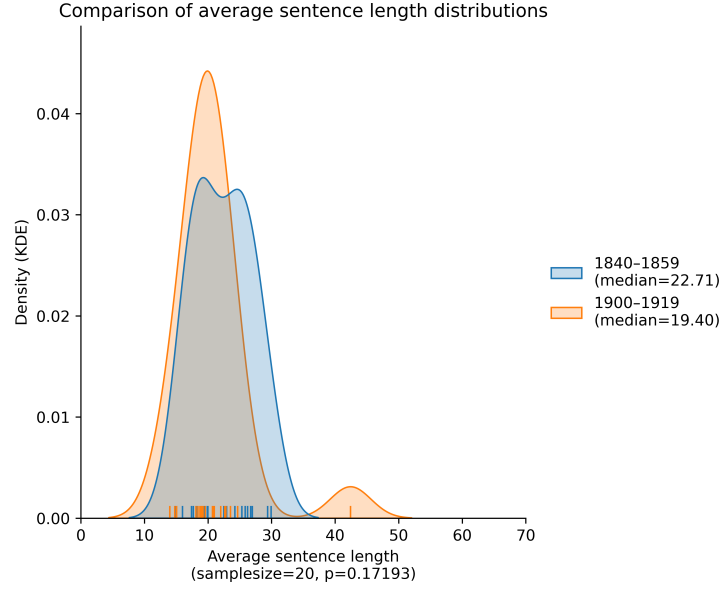




- Corpus size: 100 novels
- Standard period, quite even spread
- Significant decline in average sentence length

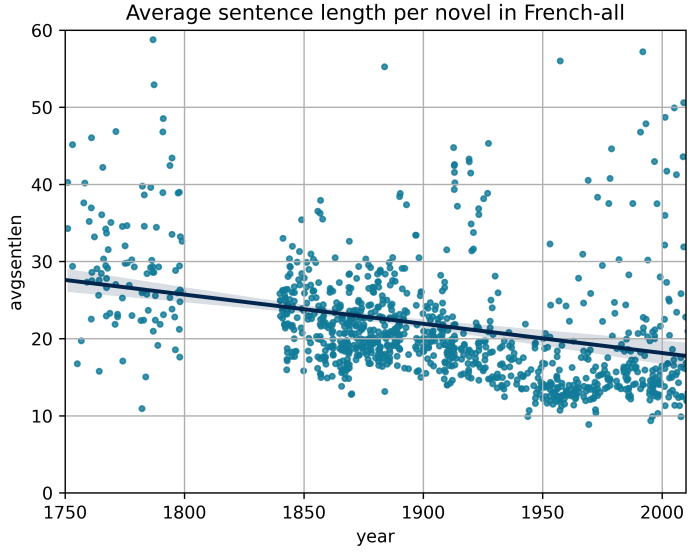
ELTeC-fra (1840-1920)

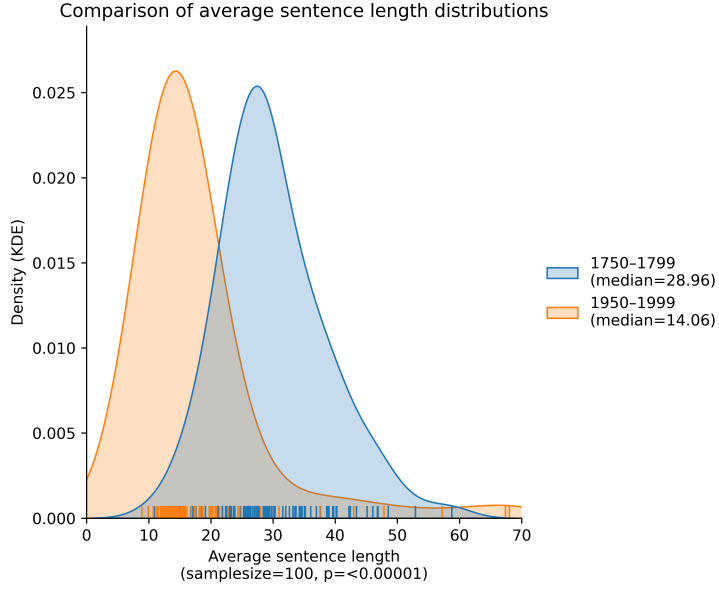




- Corpus size: 100 novels
- Standard period, quite even spread
- No significant decline in average sentence length (!)

French (ELTeC+, 1750-2010)

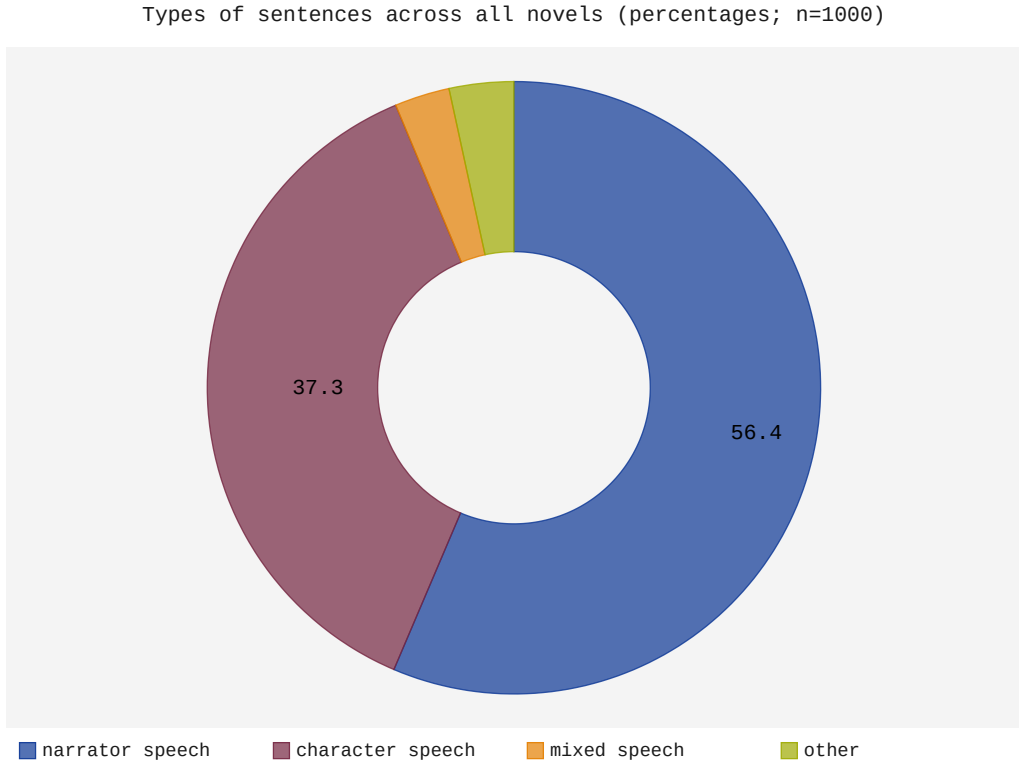




- Corpus size: 1079 novels (fra + ext1 + ext2 + cligs-rv)
- Enlarged period, with gap, uneven spread
- Significant decline in average sentence length

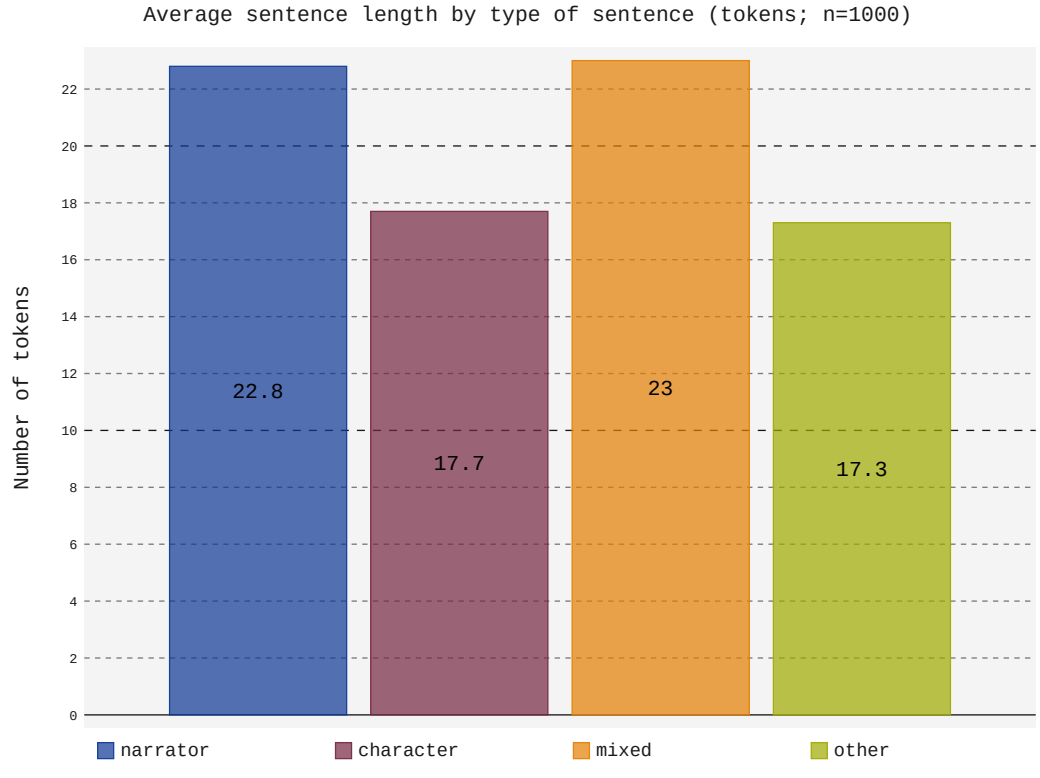
Influence of direct speech

The case of French (1): overall



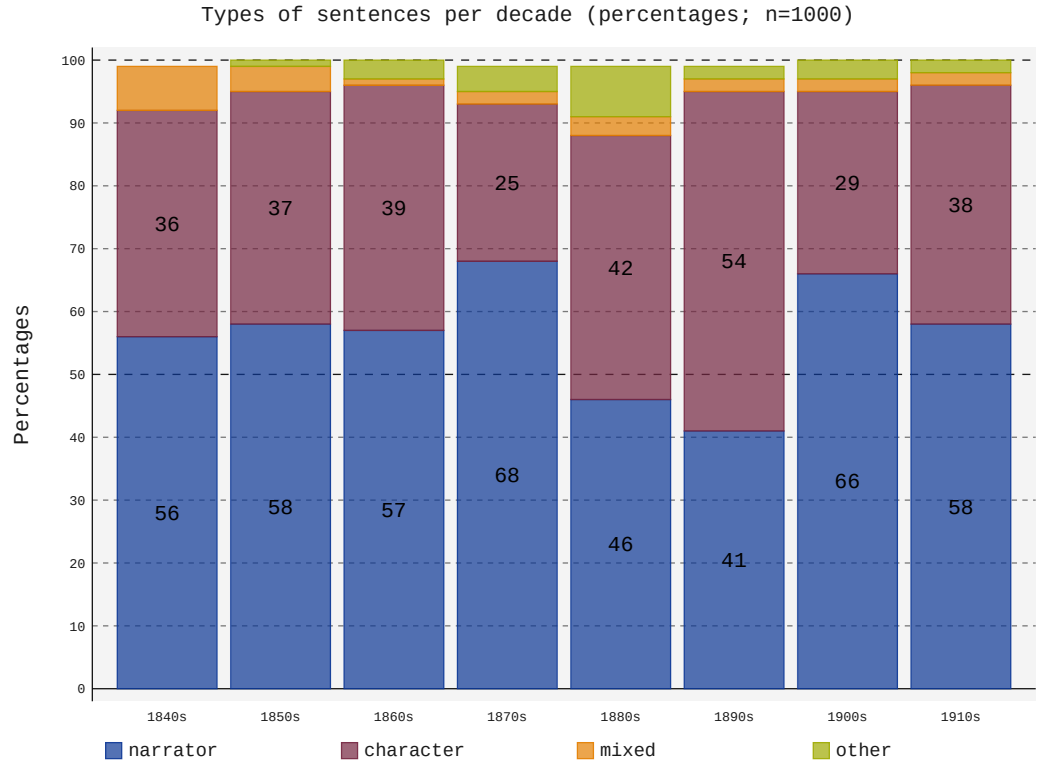
- Overall proportion of character vs. narrator speech
- 56% narrator, 37% character speech

The case of French (3): sentence length



- Average sentence length by speech type
- Character speech does have, typically, shorter sentence length

The case of French (2): per decade



- Proportion of speech type per decade
- Some variation, but no clear trend

Conclusions

Sentence length

Sentence length

- Sentences do get shorter over time, at least:
  - in novels / narrative fiction
  - for several languages
  - between 1840 and 1920

Sentence length

- Sentences do get shorter over time, at least:
  - in novels / narrative fiction
  - for several languages
  - between 1840 and 1920
- Further data needed for link to direct speech
  - French: stable character speech proportion may explain stable sentence length
  - German, English, Hungarian, Portuguese and many more: lack of data

More general issues

More general issues

- We need larger datasets, including for the 20th century, in multiple languages

More general issues

- We need larger datasets, including for the 20th century, in multiple languages
- For existing larger datasets, we need much better metadata (publication data, narrative perspective, subgenre labels)

More general issues

- We need larger datasets, including for the 20th century, in multiple languages
- For existing larger datasets, we need much better metadata (publication data, narrative perspective, subgenre labels)
- For ELTeC, annotation of character speech (or: modes of enunciation) would be important

Thank you!


Danke!
TAKK
TACK
KIITOS
AITÄH
PALDIES
AČIŪ
TAK
DANK U
THANK YOU
DANKE
DZIĘKI
DĚKUJI
ĎAKUJEM
MERCI
OBRIGADO
DANKE
DANKE
KÖSZÖNÖM
MULȚUMESC
GRAZIE
ХВАЛА (HVALA)
БЛАГОДАРЯ (BLAGO-DARYA)
GRACIAS
DANKE
HVALA
ΕΥΧΑΡΙΣΤΩ (EFHAR-ISTÓ)
= SLAWISCHE SPRACHEN
= GERMANISCHE SPRACHEN
= ROMANISCHE SPRACHEN
= FINNO-UGRISCHE SPRACHEN
= BALTISCHE SPRACHEN
= ALBANISCH
= GRIECHISCH
= TÜRKISCH
БЛАГОДАРАМ (BLAGODARAM)
FALEMENDERIT
SPRACHHELD.DE

References

- Bagga, Sunyam, und Andrew Piper. 2022. „HATHI 1M: Introducing a Million Page Historical Prose Dataset in English from the Hathi Trust“. Harvard Dataverse. https://doi.org/10.7910/DVN/HAKKUA.
- Biber, Douglas, und Susan Conrad. 2009. Register, genre, and style. Cambridge textbooks in linguistics. Cambridge, UK; New York: Cambridge University Press.
- Byszuk, Joanna, Micha\l Woźniak, Mike Kestemont, Albert Leśniak, Wojciech \Lukasik, Artjoms Šeļa, und Maciej Eder. 2020. „Detecting Direct Speech in Multilingual Collection of 19th-Century Novels“. In Proceedings of LT4HALA 2020 - 1st Workshop on Language Technologies for Historical and Ancient Languages, 100–104. Marseille, France: European Language Resources Association (ELRA). https://www.aclweb.org/anthology/2020.lt4hala-1.15.
- Gerlach, Martin, und Francesc Font-Clos. 2018. „A standardized Project Gutenberg corpus for statistical analysis of natural language and quantitative linguistics“. arXiv:1812.08092 [physics], Dezember. http://arxiv.org/abs/1812.08092.