Using large language models to enrich corpus metadata

The case of novels in COHA

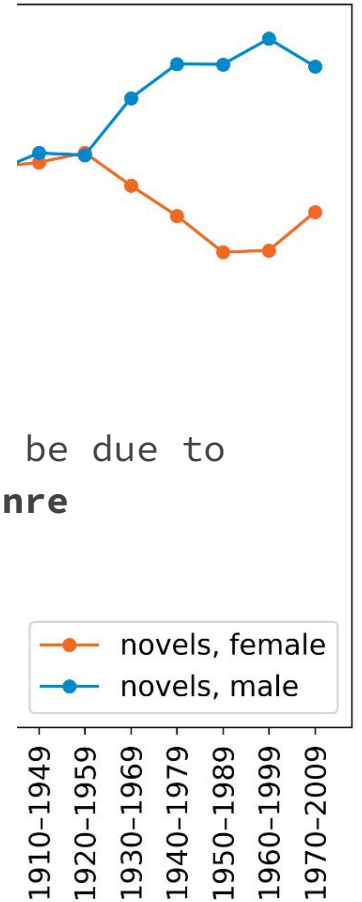
Tanja Säily, Jukka Suomela, Florent Perek, Jimena Jiménez Real & Turo Vartiainen

Introduction

- Rise of big data in (historical) corpus linguistics
  - Pros: new kinds of questions, less frequent phenomena
  - Cons: messier; less detailed metadata (e.g. Vartiainen & Säily 2024)
- New trend: metadata enrichment using machine learning
  - Öhman et al. (2019): gender metadata for fiction in the *Corpus of Historical American English* (COHA)
  - Menzel et al. (2021): discourse fields for the *Royal Society Corpus* with topic modelling
- This talk: metadata enrichment with large language models (LLMs)

Background

- Ongoing work on COHA: gender variation in the productivity of constructions
  - Säily et al. (in press); Säily & Vartiainen (in press)
  - Perek et al. (2024), *way*-cx: Could the gender differences be due to genre imbalance in novels over time? No **metadata on subgenre**
- New project: “Social roots of language change: Investigating change with enriched corpus data”
  - PI: Turo Vartiainen
  - Collaborators: e.g. Tanja Säily, Mark Davies
  - Research Council of Finland, 2024–2028
  - One aim: produce **author metadata for COHA**



Goals

---

1. Quantify how well LLMs can solve metadata annotation tasks
2. Develop best practices for accurate and resource-efficient metadata enrichment
3. Produce a metadata-enriched version of COHA
4. Study if this metadata can shed light on phenomena observed in prior work

Material

- *Corpus of Historical American English* (COHA)
  - 400 MW, 1810–2009
- Fiction section: c. 50% of the data
  - **Gender metadata** for authors developed by Öhman et al. (2019)
  - Promising material for sociolinguistic investigation: a more speech-like genre (dialogue)
  - Types of fiction (e.g. short stories, drama, movie scripts) unevenly distributed over time (Säily & Vartiainen in press)
    → restriction to **novels only, c. 150 MW** (based on cohaTexts.xls)

Metadata enrichment

Piloting metadata enrichment with LLMs

- We want to annotate the novels in COHA for subgenre and author metadata
- Training data of LLMs includes sources like Wikipedia that discuss authors and genres
- How far can we get by giving LLMs just the **author, title and publication year** of novels in COHA?
  - This metadata is freely available from English-Corpora.org
  - Plus some examples annotated manually

Metadata categories to be annotated (1)

- Genre (cf. *Brown Corpus*, *TV Corpus*)
  - General fiction
  - Adventure and Western
  - Fantasy
  - Historical fiction
  - Horror
  - Mystery and detective fiction
  - Romance
  - Science fiction
  - not a novel

Metadata categories to be annotated (2)

- Target audience
  - Adult, young adult, children (cf. *British National Corpus*)
- Publication year

- Author gender
  - Male, female
- Author year of birth

- Uncertain cases labelled as “none”

Manually annotated sample

- 345 novels from COHA
  - Sampled by decade and gender (Öhman et al. 2019)
  - Pre-1950s: (max) 5 per decade and gender, 1950s-: 10
- Split into 3 datasets
  - **Training: 50** – used as examples in the system prompt
  - **Validation: 150** – tried out **95 combinations** of models, prompts, etc. to see what works best
  - **Test: 145** – used to assess the performance of the **best** model/prompt combination in previously-unseen samples

OpenAI LLMs tested in validation phase

- gpt-4.1-2025-04-14
- gpt-4.5-preview-2025-02-27 (expensive)
- o3-mini-2025-01-31
- gpt-4o-2024-08-06
- gpt-4o-mini-2024-07-18 (cheap)

- **“Structured Outputs” API:** ensure well-formed JSON output
  - Category labels listed in output structure specification
- **Temperature:** mostly kept at 0 to minimize surprises

System prompt: many variations

We have got JSON records that describe well-known novels, with fields "author", "title", and "year". This information may be somewhat unreliable; for example, "year" is not necessarily the year when the novel was first published. Sometimes the "author" field might contain the year of birth and the year of death of the author, or some other useful identifying information. There may also be some entries that describe books that are not novels. For each record we need to find this information:

- year = the correct year when the book was first published
- author_gender = the gender of the author
- author_year_of_birth = the author's year of birth
- genre = the genre of the book
- target_audience = the age of the target audience

Set "genre" = "not a novel" if the book is not a novel (e.g. a collection of short stories or poetry). Leave those fields empty that cannot be reliably determined, or if they are ambiguous (e.g. multiple authors with mixed genders).

- This version performed the best

System prompt: add examples

(few-shot learning)

```
Here are some examples of valid input records and corresponding output records:

Input: {"author": "Andrews, V. C. (Virginia C.)", "title": "Dawn /", "year": 1990}

Output:
{"year":1990,"author_gender":"female","author_year_of_birth":1923,"genre":"Horror","ta
rget_audience":"young adult"}


Input: {"author": "James T. Farrell", "title": "Yet Other Waters", "year": 1952}

Output:
{"year":1952,"author_gender":"male","author_year_of_birth":1904,"genre":"General
fiction","target_audience":"adult"}
```

- Number of examples given: 0–50 (training dataset)

User prompt: just the input data

{"author": "Clark, Mary Higgins. ", "title": "Pretend you don't see her /", "year": 1998}

User prompt: add Wikipedia results (retrieval-augmented generation)

```
Wikipedia results: {"id": 479225, "key":
"Mary_Higgins_Clark", "title": "Mary Higgins Clark",
"excerpt": "<span class="searchmatch">Mary <span
class="searchmatch">Higgins <span class="searchmatch">Clark
(born <span class="searchmatch">Mary Theresa Eleanor <span
class="searchmatch">Higgins; December 24, 1927 - January 31,
2020) was an American author of suspense novels. Each of
<span class="searchmatch">her 51", "matched_title": null,
"description": "American novelist and writer (1927–2020)"}
...
```

Output formats

(chain-of-thought reasoning)

---

1. Ask for a minimal JSON record
2. Ask for a JSON record that starts with a “notes” field for free-text explanations

“Emilie Baker Loring was a prolific American author known for her romance novels. Her works often feature themes of love, honor, and adventure, appealing to a broad audience. 'To Love and to Honor' is one of her many romance novels, published posthumously.”

Observations from validation

- **gpt-4o** works well, no reason to pay for gpt-4.5
- **Many examples** (up to 50) help a lot
- With many examples:
  - Precise prompt not that important
  - “Reasoning” not necessary → faster and cheaper
  - Wikipedia results not necessary → simplifies the process a lot
- Our final choices:
  - model: OpenAI gpt-4o-2024-08-06
  - system prompt: with 50 examples
  - user prompt: minimal JSON, no Wikipedia results
  - output: minimal JSON

Validation vs. test

Best performing model + prompt on **validation** dataset:
gpt-4o examples-50 fix1
(overall accuracy: **88.3%**)

- **Genre: 71.3%**
- Target audience: 94.0%
- Publication year: 91.3%
- **Author gender: 95.3%**
- Author year of birth: 89.3%

→ Moved on to test dataset

Performance on **test** dataset:

gpt-4o examples-50 fix1
(overall accuracy: **85.7%**)

- **Genre: 68.3%**
- Target audience: 88.3%
- Publication year: 89.0%
- **Author gender: 95.2%**
- Author year of birth: 87.6%

Lower as expected but same ballpark
→ moved on to entire COHA fiction

Common mistakes: genre (test dataset)

- **Historical fiction** (precision 75%, recall 55%)
  → 7 × General fiction, 2 × Romance, 1 × Adventure and Western
- **Adventure and Western** (precision 73%, recall 44%)
  → 4 × General fiction, 3 × not a novel, 2 × Historical fiction,
  1 × Mystery and detective fiction
- **General fiction** (precision 60%, recall 77%)
  → 4 × not a novel, 3 × Romance, 1 × Adventure and Western,
  1 × Historical fiction
- **Romance** (precision 58%, recall 64%)
  → 4 × General fiction

Most common erroneous annotations: 20 × General fiction,
7 × not a novel, 5 × Romance, 4 × Historical fiction
- sometimes a matter of interpretation!

Common mistakes: author gender (test dataset)

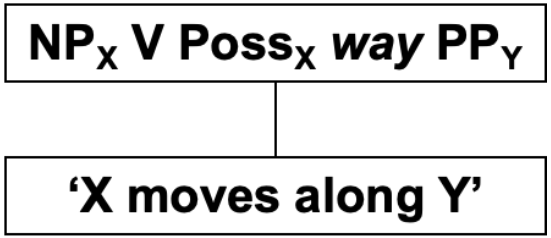
- **None** (precision 60%, recall 67%)
  → 3 × Male
  - E.g. multiple authors, house name used by publishing company
- **Female** (precision 100%, recall 97%)
  → 2 × none
  - COHA metadata has author:None, human fetched information manually
- **Male** (precision 96%, recall 97%)
  → 2 × none
  - Vermilye Taylor: human incorrect, none/female ok → recall 100%!

Full classification of all COHA fiction

- Total ≈ 11,000 queries
  - ≈ 41.3 million input tokens, 98% cached
  - ≈ 0.3 million output tokens
- Costs ≈ **56 USD**

Case: *way*-construction

The *way*-construction

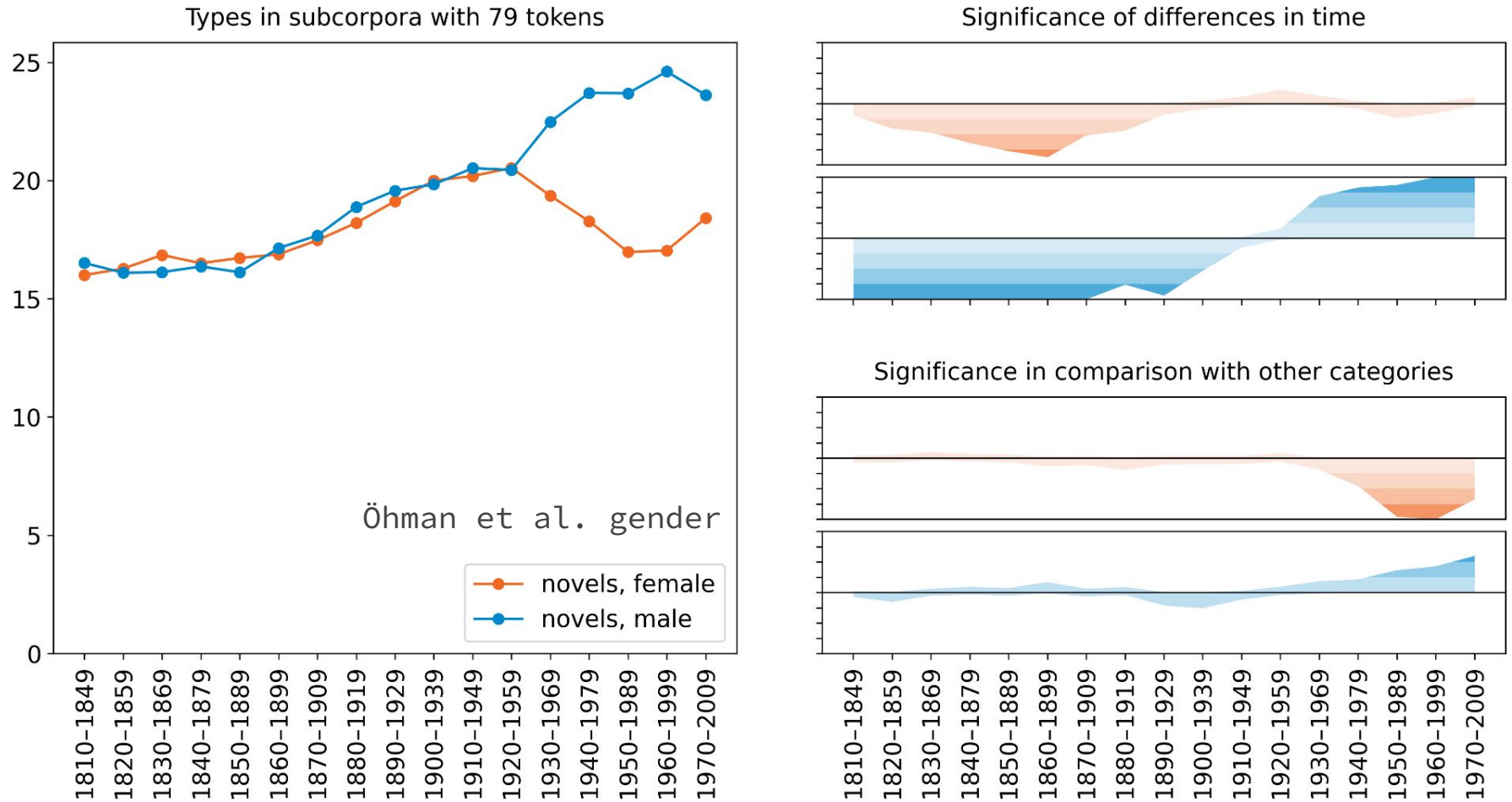


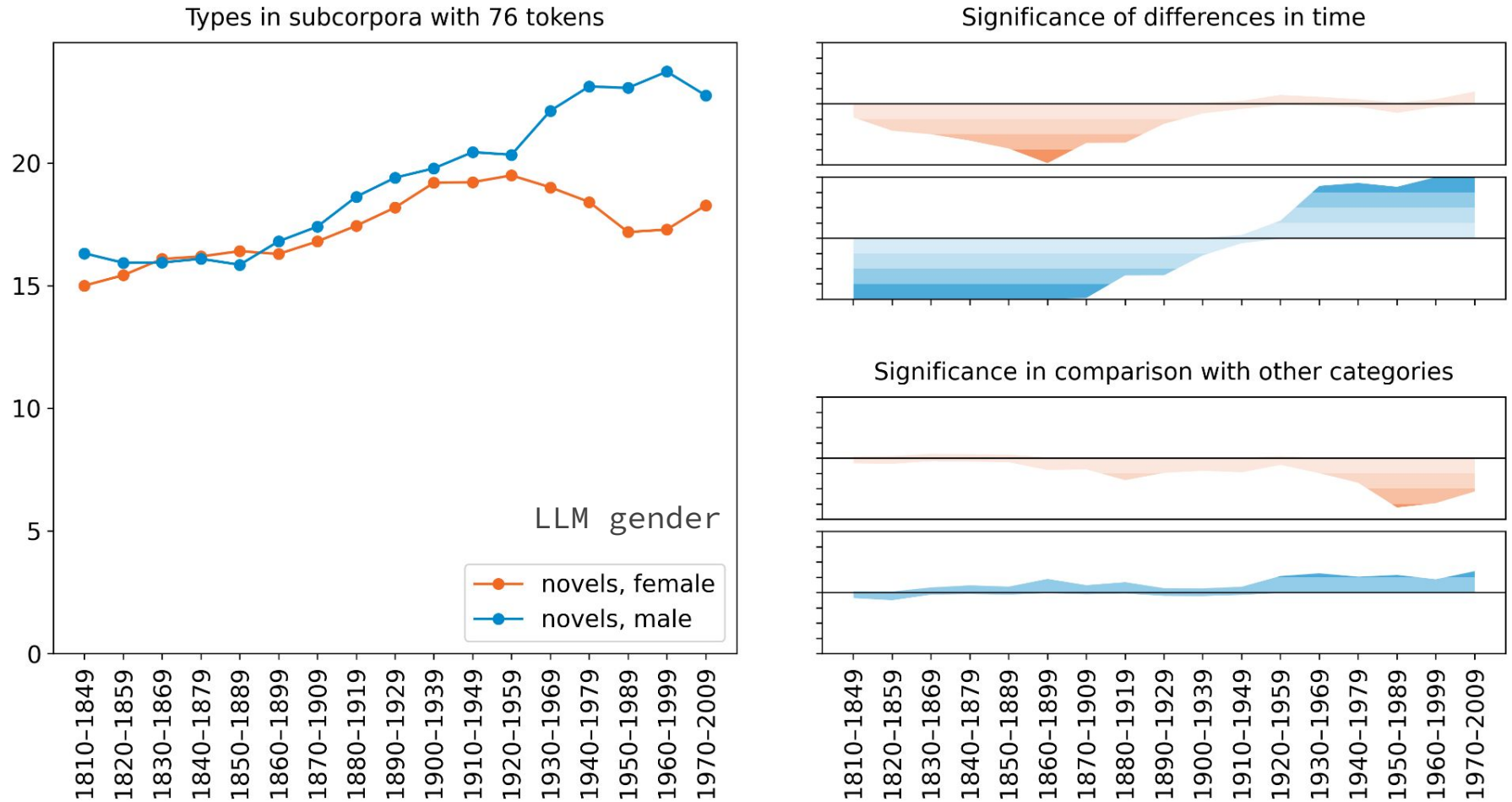
- Verb + Possessive + *way* + PP

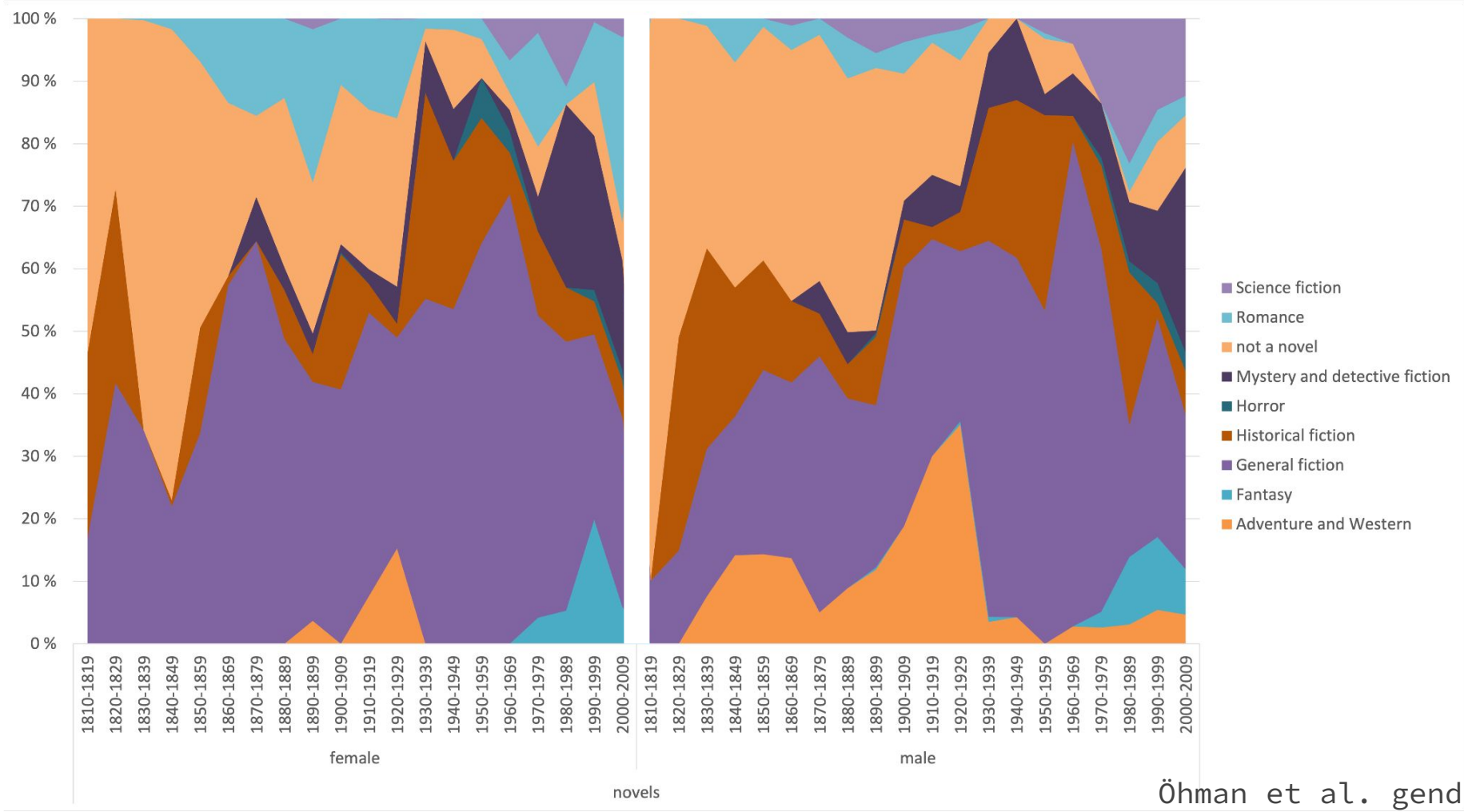
  They hacked their way through the jungle.

  We pushed our way into the bar.

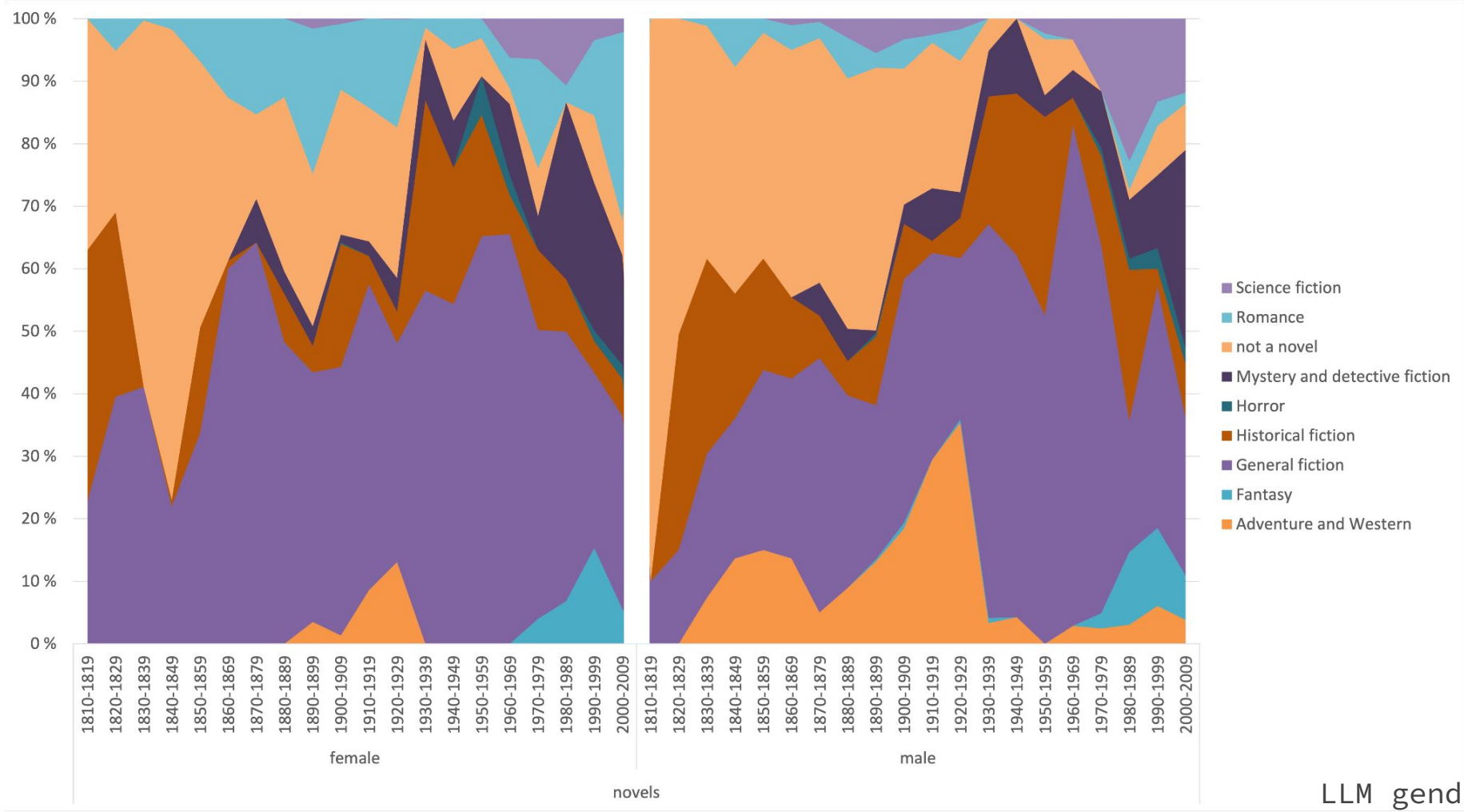
- We study the productivity of the *way*-construction by measuring **type frequencies**
  - I.e. how many different items in the verb slot in different time periods


Types in subcorpora with 79 tokens
25
20
15
10
5
0
Öhman et al. gender
novels, female
novels, male
1810–1849
1820–1859
1830–1869
1840–1879
1850–1889
1860–1899
1870–1909
1880–1919
1890–1929
1900–1939
1910–1949
1920–1959
1930–1969
1940–1979
1950–1989
1960–1999
1970–2009
Significance of differences in time
Significance in comparison with other categories
1810–1849
1820–1859
1830–1869
1840–1879
1850–1889
1860–1899
1870–1909
1880–1919
1890–1929
1900–1939
1910–1949
1920–1959
1930–1969
1940–1979
1950–1989
1960–1999
1970–2009

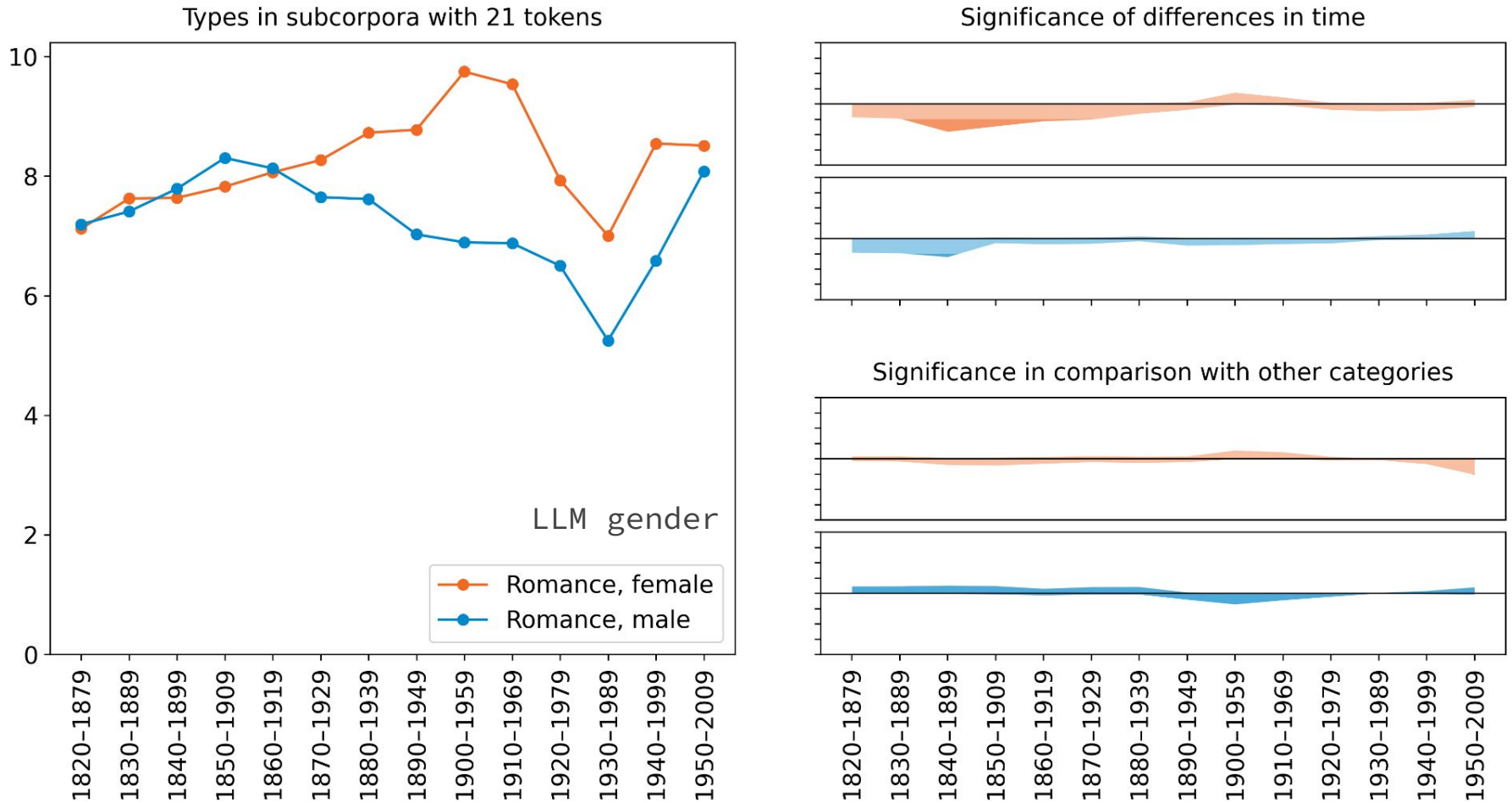

Types in subcorpora with 76 tokens
0
5
10
15
20
LLM gender
novels, female
novels, male
1810–1849
1820–1859
1830–1869
1840–1879
1850–1889
1860–1899
1870–1909
1880–1919
1890–1929
1900–1939
1910–1949
1920–1959
1930–1969
1940–1979
1950–1989
1960–1999
1970–2009
Significance of differences in time
Significance in comparison with other categories
1810–1849
1820–1859
1830–1869
1840–1879
1850–1889
1860–1899
1870–1909
1880–1919
1890–1929
1900–1939
1910–1949
1920–1959
1930–1969
1940–1979
1950–1989
1960–1999
1970–2009

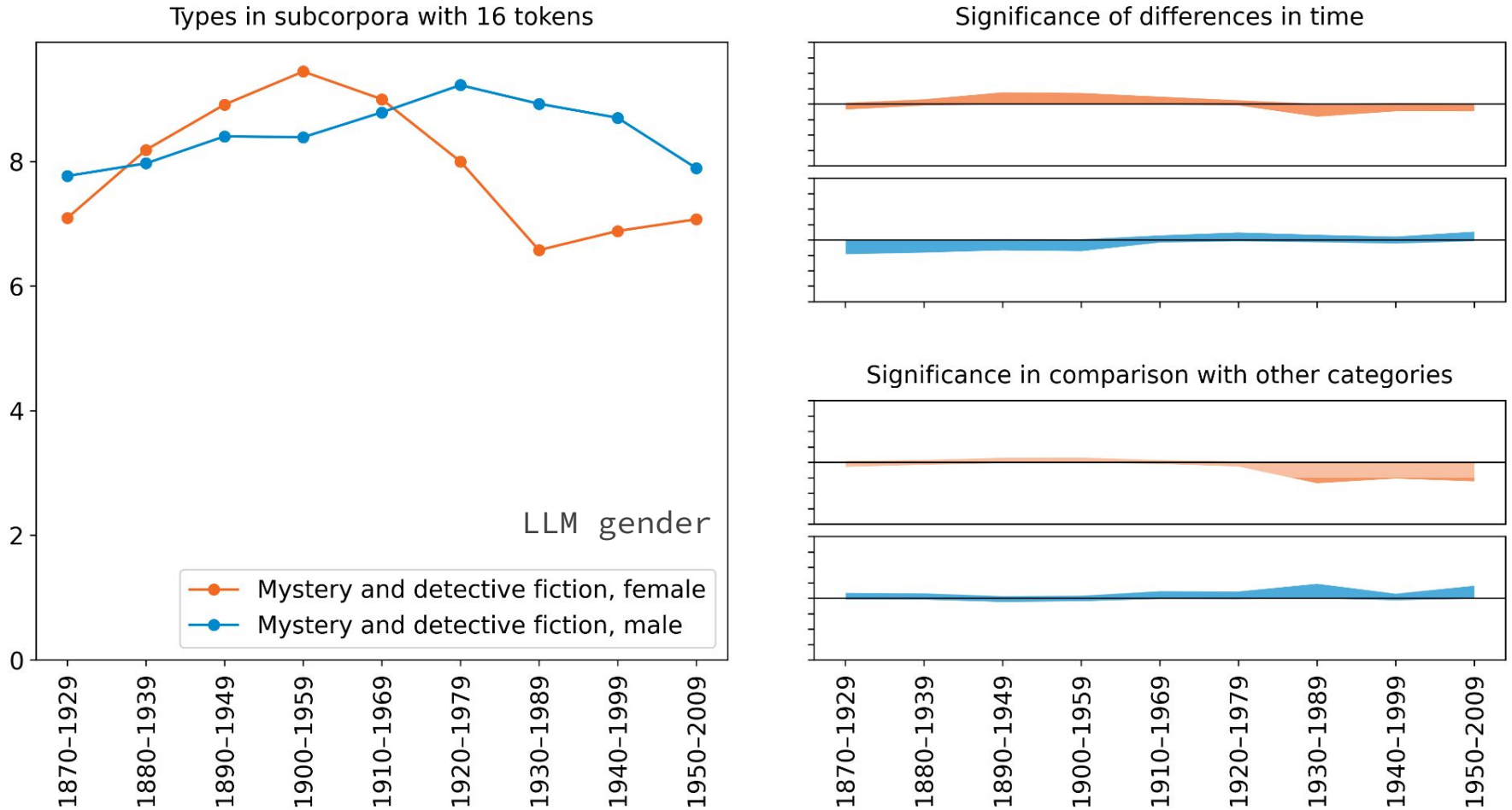


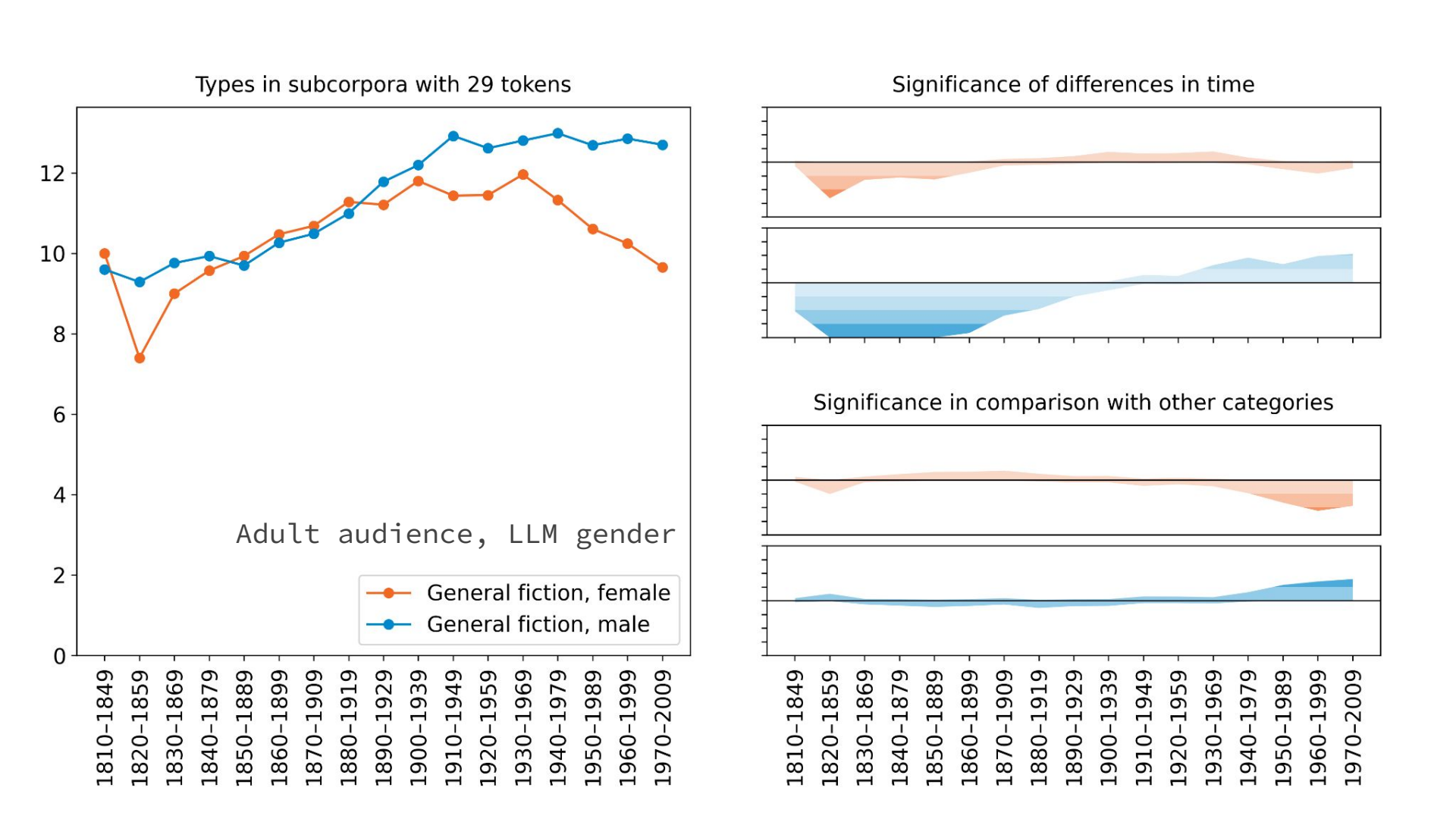
Öhman et al. gender


100 %
90 %
80 %
70 %
60 %
50 %
40 %
30 %
20 %
10 %
0 %
1810-1819
1820-1829
1830-1839
1840-1849
1850-1859
1860-1869
1870-1879
1880-1889
1890-1899
1900-1909
1910-1919
1920-1929
1930-1939
1940-1949
1950-1959
1960-1969
1970-1979
1980-1989
1990-1999
2000-2009
female
male
novels
Science fiction
Romance
not a novel
Mystery and detective fiction
Horror
Historical fiction
General fiction
Fantasy
Adventure and Western

LLM gender


Types in subcorpora with 21 tokens
10
8
6
4
2
0
LLM gender
Romance, female
Romance, male
1820–1879
1830–1889
1840–1899
1850–1909
1860–1919
1870–1929
1880–1939
1890–1949
1900–1959
1910–1969
1920–1979
1930–1989
1940–1999
1950–2009
Significance of differences in time
Significance in comparison with other categories
1820–1879
1830–1889
1840–1899
1850–1909
1860–1919
1870–1929
1880–1939
1890–1949
1900–1959
1910–1969
1920–1979
1930–1989
1940–1999
1950–2009


Types in subcorpora with 16 tokens
LLM gender
Mystery and detective fiction, female
Mystery and detective fiction, male
0
2
4
6
8
1870–1929
1880–1939
1890–1949
1900–1959
1910–1969
1920–1979
1930–1989
1940–1999
1950–2009
Significance of differences in time
Significance in comparison with other categories
1870–1929
1880–1939
1890–1949
1900–1959
1910–1969
1920–1979
1930–1989
1940–1999
1950–2009


Types in subcorpora with 29 tokens
Adult audience, LLM gender
General fiction, female
General fiction, male
0
2
4
6
8
10
12
1810-1849
1820-1859
1830-1869
1840-1879
1850-1889
1860-1899
1870-1909
1880-1919
1890-1929
1900-1939
1910-1949
1920-1959
1930-1969
1940-1979
1950-1989
1960-1999
1970-2009
Significance of differences in time
Significance in comparison with other categories
1810-1849
1820-1859
1830-1869
1840-1879
1850-1889
1860-1899
1870-1909
1880-1919
1890-1929
1900-1939
1910-1949
1920-1959
1930-1969
1940-1979
1950-1989
1960-1999
1970-2009

Discussion

Goals revisited

---

1. Quantify how well LLMs can solve metadata annotation tasks
   - Less ambiguous cases (e.g. gender in COHA): quite well!
2. Develop best practices for accurate and resource-efficient metadata enrichment
   - More examples = better, current generic model ok
3. Produce a metadata-enriched version of COHA
   - Pilot version: github.com/suomela/coha-gpt-enriched-metadata
4. Study if this metadata can shed light on phenomena observed in prior work
   - Genre does not seem to explain gender variation in the *way*-cx

Conclusion

- Remaining challenges
  - Reproducibility? Rerunning the same LLM+prompt produces similar but not identical results (even with temperature=0)
  - Reliability estimates for individual predictions? Human could check less reliable instances
- Future work
  - Compare our gender metadata with Öhman et al. (2019)
    - Discrepancies for human checking
  - Experiment with other methods besides LLMs
  - Use best methods + human checking to generate reliable metadata for current versions of COHA and COCA

References

- Davies, M. 2010–. *The Corpus of Historical American English*: 400 million words, 1810–2009. https://www.english-corpora.org/coha/
- Menzel, K., J. Knappen & E. Teich. 2021. Generating linguistically relevant metadata for the *Royal Society Corpus*. *Research in Corpus Linguistics* 9(1): 1–18.
- Öhman, E., T. Säily & M. Laitinen. 2019. Towards the inevitable demise of *everybody*? A multifactorial analysis of *-one/-body/-man* variation in indefinite pronouns in historical American English. 40th Annual Conference of the International Computer Archive of Modern and Medieval English (ICAME 40), Neuchâtel, Switzerland, June 2019. https://tanjasaily.fi/talks/icame40_ohman_et_al_2019.pdf
- Perek, F., T. Säily & J. Suomela. 2024. Historical sociolinguistics meets constructional change: Gender and the *way*-construction in the *Corpus of Historical American English*. 57th Annual Meeting of the Societas Linguistica Europaea (SLE 2024), Helsinki, Finland, August 2024. https://tanjasaily.fi/talks/sle57_perek_et_al_2024.pdf
- Säily, T., F. Perek & J. Suomela. In press. Variation and change in the productivity of BE *going to* V in the *Corpus of Historical American English*, 1810–2009. *English Language and Linguistics* 29(2).
- Säily, T. & T. Vartiainen. In press. Historical linguistics. M. Mahlberg & G. Brooks (eds.), *Bloomsbury Handbook of Corpus Linguistics*. Bloomsbury Academic.
- Vartiainen, T. & T. Säily. 2024. Engaging with bad (meta)data in historical corpus linguistics. M. Kaunisto & M. Schilk (eds.), *Challenges in Corpus Linguistics: Rethinking Corpus Compilation and Analysis*, 9–34. John Benjamins.

Acknowledgements

- We would like to thank Jaakko Lehtinen and Perttu Hämäläinen for helpful discussions

- This work was supported in part by the Research Council of Finland, grant 363720

Output structure specification (Python, pydantic)

```python
class AuthorGender(str, Enum):
    male = "male"
    female = "female"

class Genre(str, Enum):
    general_fiction = "General fiction"
    adventure_and_western = "Adventure and Western"
    fantasy = "Fantasy"
    historical_fiction = "Historical fiction"
    horror = "Horror"
    mystery_and_detective_fiction = "Mystery and detective fiction"
    romance = "Romance"
    science_fiction = "Science fiction"
    not_a_novel = "not a novel"

class TargetAudience(str, Enum):
    children = "children"
    young_adult = "young adult"
    adult = "adult"

class BookClassification(BaseModel):
    year: Optional[int]
    author_gender: Optional[AuthorGender]
    author_year_of_birth: Optional[int]
    genre: Optional[Genre]
    target_audience: Optional[TargetAudience]
```