DigiTS
Center for Digital
Text Scholarship

I Annual Progress Meeting
17.02.2026

Jakobi 2, Tartu | Zoom

UNIVERSITY OF TARTU
Centre for Digital Humanities

Funded by
the European Union

Agenda

14.00 Prof. Liina Lindström – Welcome
14.05 Prof. Maciej Eder – Overview of DigiTS Year 1 + Upcoming Activities
14.25 Thiago Dumont Oliveira – „The French Drama Revolution: Political Economy and Literary Production, 1700-1900"
14.45 Botond Szemes„Characters and Narratives: Different Approaches in Computational Drama Analysis"
15.05 Kristiina Vaik – „(Web) Corpora Without Fixed Category Labels: An Alternative Approach"
15.25 Coffee break
15.45 Prof. Maciej Eder – „Exploring Text Similarity Measures: New Approaches"
16.05 Sofia Kriuchkova – „Corpora as a Tool for Evaluating Gender Stereotypes in Language Use“
16.25 Bhumika Bhattacharyya – „Applying Computational Linguistics Perspectives on Disinformation: State of the Art, Limitations and Emerging Research Gaps"
16.45 Prof. Liina Lindström – Wrap-up

Thiago Dumont Oliveira

Siena, Italy	Basel, Switzerland	Turin, Italy	ASU, United States	Cambridge, UK
2016-2019	2019-2021	2021-2023	2023-2024	2025-
PhD	Postdoc	Postdoc	Postdoc	Research Associate

UNIVERSITY OF CAMBRIDGE
Centre of Latin-American Studies

DigiTS
Center for Digital Text Scholarship

thiago@ut.ee
@thiagodoliveira.bsky.social
https://sites.google.com/view/tdoliveira

The French Drama Revolution: Political Economy and Literary Production, 1700-1900

Cultural History

Political Economy

Ideology

Narratives

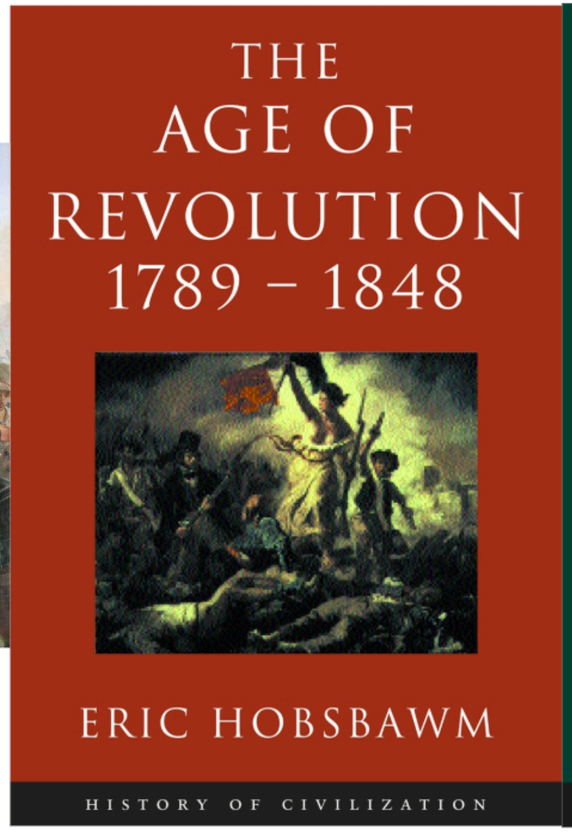
Class

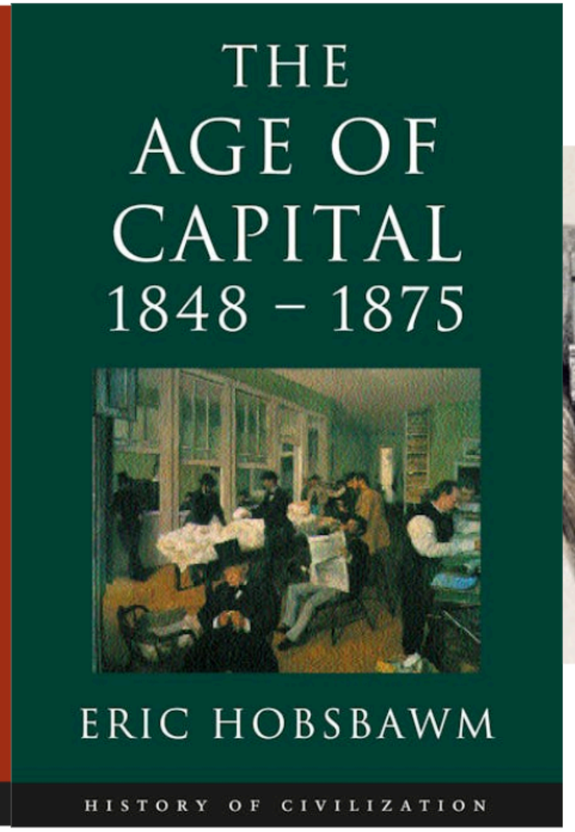
Power

Thiago Dumont Oliveira

Question, Data & Methods

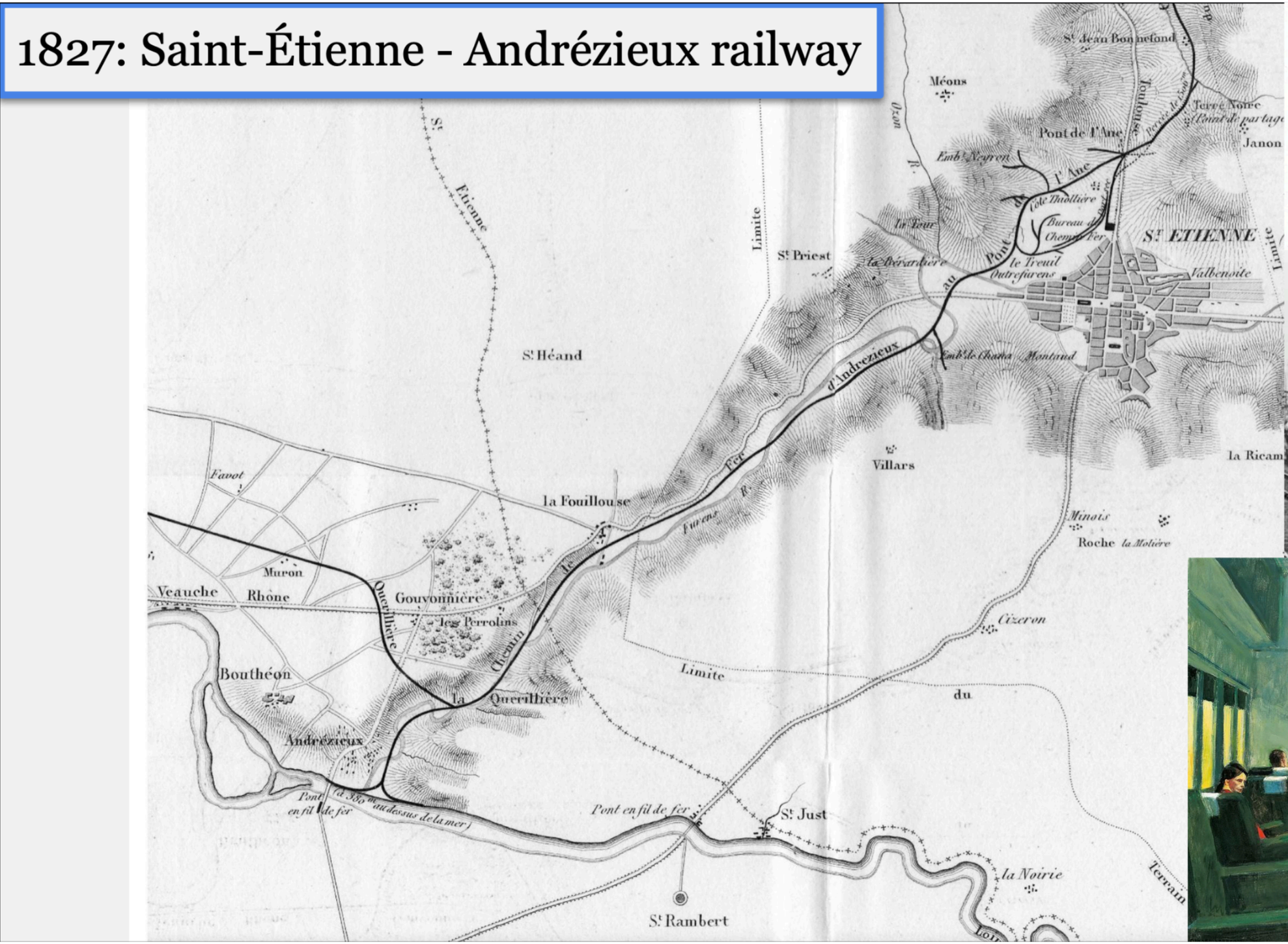
- **Question:** How did French drama change between 1700-1900 vis-à-vis shifting political-economic circumstances?

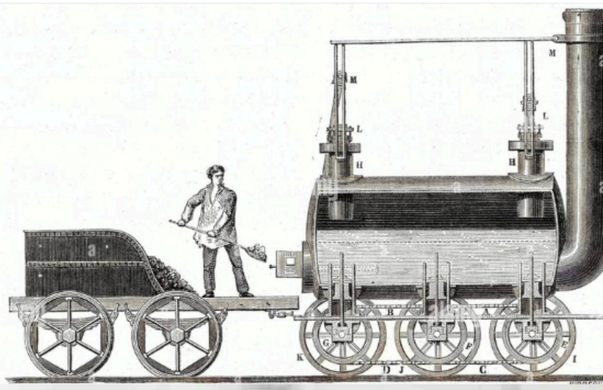


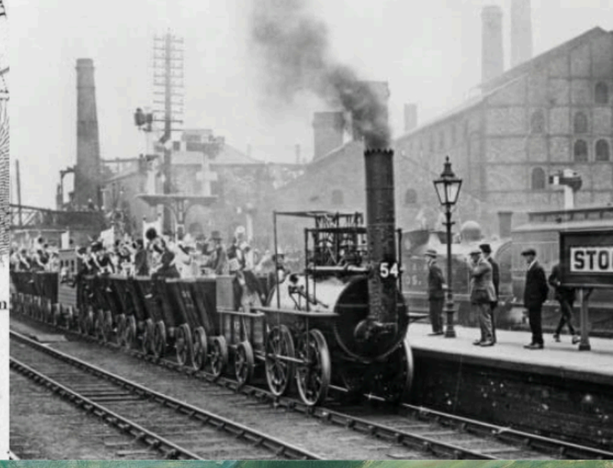


1700 1900

1827: Saint-Étienne - Andrézieux railway







Question, Data & Methods

- **Question:** How did French drama change between 1700-1900 vis-à-vis shifting political-economic circumstances?
- **Data:** 1215 French plays published between 1700 and 1900.
- **Source:** DraCor's French-language corpora (https://dracor.org/)
- **Methods:**

1. Latent Dirichlet Allocation (LDA)
2. Non-negative Matrix Factorization (NMF)
3. Jensen-Shannon Divergence (JSD)
4. Multidimensional Scaling (MDS)

Results - LDA:

Topic Prevalence, Total Variation (1700–1900), 15 Top Words, Translations and 5 Representative Works.

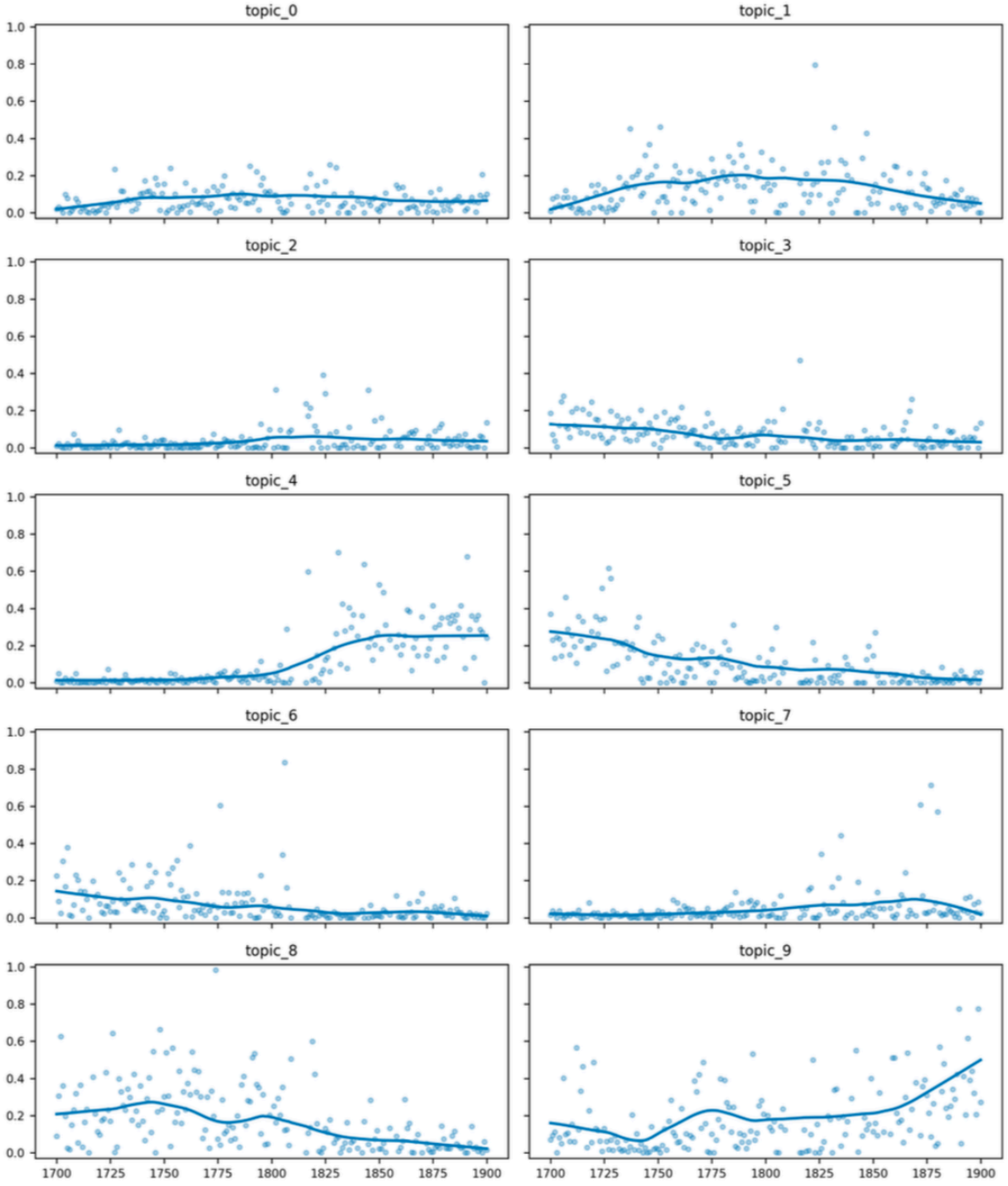
Topic	Prevalence	$\Delta_{1700–1900}$	Top words (15)	Translation	Representative Works
0	0.07	≈ 0	français, auteur, france, art, talent, goût, ouvrage, paris, public, génie, nature, opéra, sujet, vérité, état	French, author, France, art, talent, taste, work, Paris, public, genius, nature, opera, subject, truth, state	Diderot: *Entretien entre d'Alembert et Diderot* Anonymous: *Ouverture de la Séance* Sacy: *L'Ile Déserte ou Le Naufrage* Diderot: *Rêve de d'Alembert* Sacy: *Le Testament*
1	0.13	≈ 0	mère, secret, sentiment, frère, fils, oncle, soeur, lettre, amitié, soin, âme, espérer, tromper, tendre, tendresse	mother, secret, sentiment, brother, son, uncle, sister, letter, friendship, care, soul, hope, deceive, tender, tenderness	Genlis: *La Mère Rivale, Comédie* Colleville: *Sophie et Derville* Carmontelle: *Les Faux Indifférents* Carmontelle: *Le Faux Empoisonnement* Graffigny: *Cénie*
2	0.04	+0.05	molière, pièce, théâtre, jouer, comédie, scène, rôle, acteur, comédien, auteur, poète, roi, don, talent, rire	Molière, play, theatre, play, comedy, stage, role, actor, comedian, author, poet, king, gift, talent, laugh	Jouy: *Les Bancs de la Promenade* Diderot: *Paradoxe sur le Comédien* Sacy: *La Modestie* Sacy: *La Répétition* Chalmeton: *A Jean Racine*
3	0.07	−0.09	encor, bien, époux, affaire, soin, sot, sage, hymen, tour, discours, fou, diable, rire, loi, vieux	again, good, husband, affair, care, fool, wise, marriage, turn, speech, madman, devil, laugh, law, old	Voltaire: *Le Dépostaire* Voltaire: *L'Enfant Prodigue* Regnard: *Démocrite* Regnard: *Les Ménechmes, ou Les Jumeaux* Barbier: *Le Faucon*
4	0.12	+0.37	pauvre, vieux, nuit, soir, roi, mère, fleur, mort, rêve, soleil, voix, peur, âme, bras, souvenir	poor, old, night, evening, king, mother, flower, death, dream, sun, voice, fear, soul, arm, memory	Beissier: *L'Oiseau Bleu, Saynète* Adenis: *Le Nouveau Né* Monselet: *Par la Poste* Crossonnois: *Un Monsieur qui ne Veut Plus Fumer* Beissier: *La Nuit de Noël, Comédie*
5	0.10	−0.24	affaire, épouser, mariage, lisette, mari, chevalier, marier, maîtresse, mère, adieu, fâcher, soeur, honnête, marquis, vérité	affair, marry, marriage, Lisette, husband, knight, wed, mistress, mother, farewell, anger, sister, honest, marquis, truth	Dancourt: *Colin-Maillard* Marivaux: *Le Dénouement Imprévu* Marivaux: *La Commère* Audiffret: *L'Epreuve* Marivaux: *La Double Inconstance*
6	0.07	−0.13	amant, tendre, doux, feu, aimable, ardeur, charmant, charme, objet, flamme, beauté, léandre, dou, désir, âme	lover, tender, sweet, fire, kind, ardor, charming, charm, object, flame, beauty, Leander, gentle, desire, soul	Voisenon: *Hilas et Zélis* Saint-Gilles Lenfant: *La Feinte Heureuse* La Motte: *Le Carnaval et la Folie* Favart: *L'Amant Déguisé* Hénault: *Le Temple des Chimères*
7	0.05	+0.09	mort, fils, terre, mère, ville, parole, demeure, roi, étranger, tuer, frère, sang, vieillard, ennemi, mer	death, son, land, mother, city, word, dwelling, king, foreigner, kill, brother, blood, old man, enemy, sea	Anonymous: *Oedipe à Colone* Crébillon: *Electre* Anonymous: *Oedipe Roi* La Harpe: *Philoctète* Beissier: *Antigone*
8	0.18	−0.27	roi, sang, fils, mort, sort, vertu, loi, peuple, crime, gloire, cruel, soin, fureur, encor, horreur	king, blood, son, death, fate, virtue, law, people, crime, glory, cruel, care, fury, again, horror	Chamfort: *Mustapha et Zéangir* Crébillon: *Pyrrhus* Voltaire: *Saül* Chabanon: *Eudoxie* Crébillon: *Atrée et Thyeste*
9	0.17	+0.23	argent, diable, affaire, pauvre, payer, mari, maison, franc, voyon, manger, dame, lettre, papa, garçon, boire	money, devil, affair, poor, pay, husband, house, franc, rogue, eat, lady, letter, father, boy, drink	Archambault: *Janot* Archambault: *Janot chez le Dégraisseur* Renard: *La Demande* Gueullette: *Le Chapeau de Fortunatus* Carmontelle: *Les Voyageurs*

poor, old, night, evening, king, mother, flower, death, dream, sun, voice, fear, soul, arm, memory

Bourgeois Life

king, blood, son, death, fate, virtue, law, people, crime, glory, cruel, care, fury, again, horror

Aristocratic Life

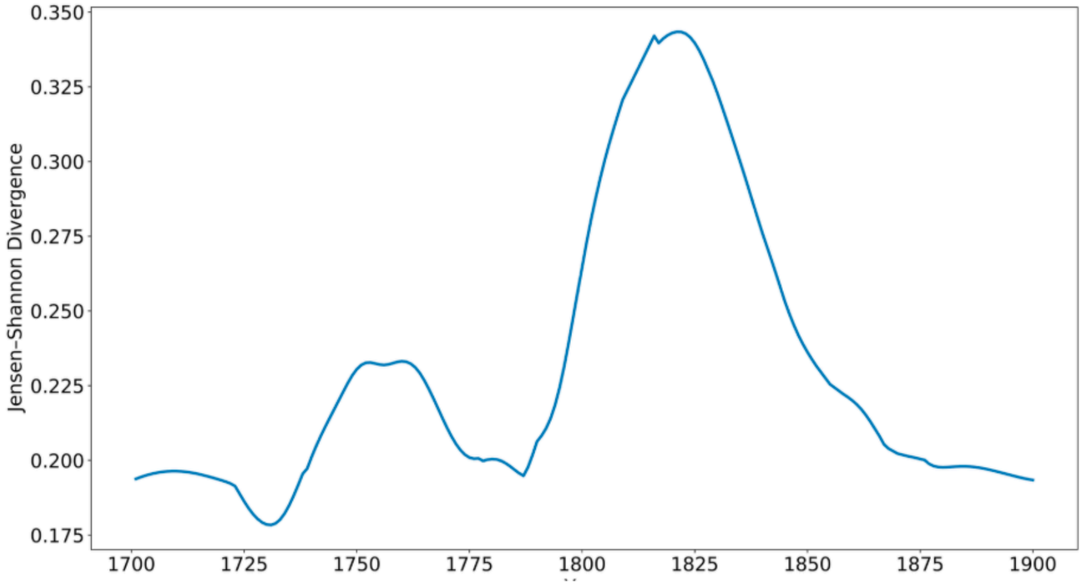


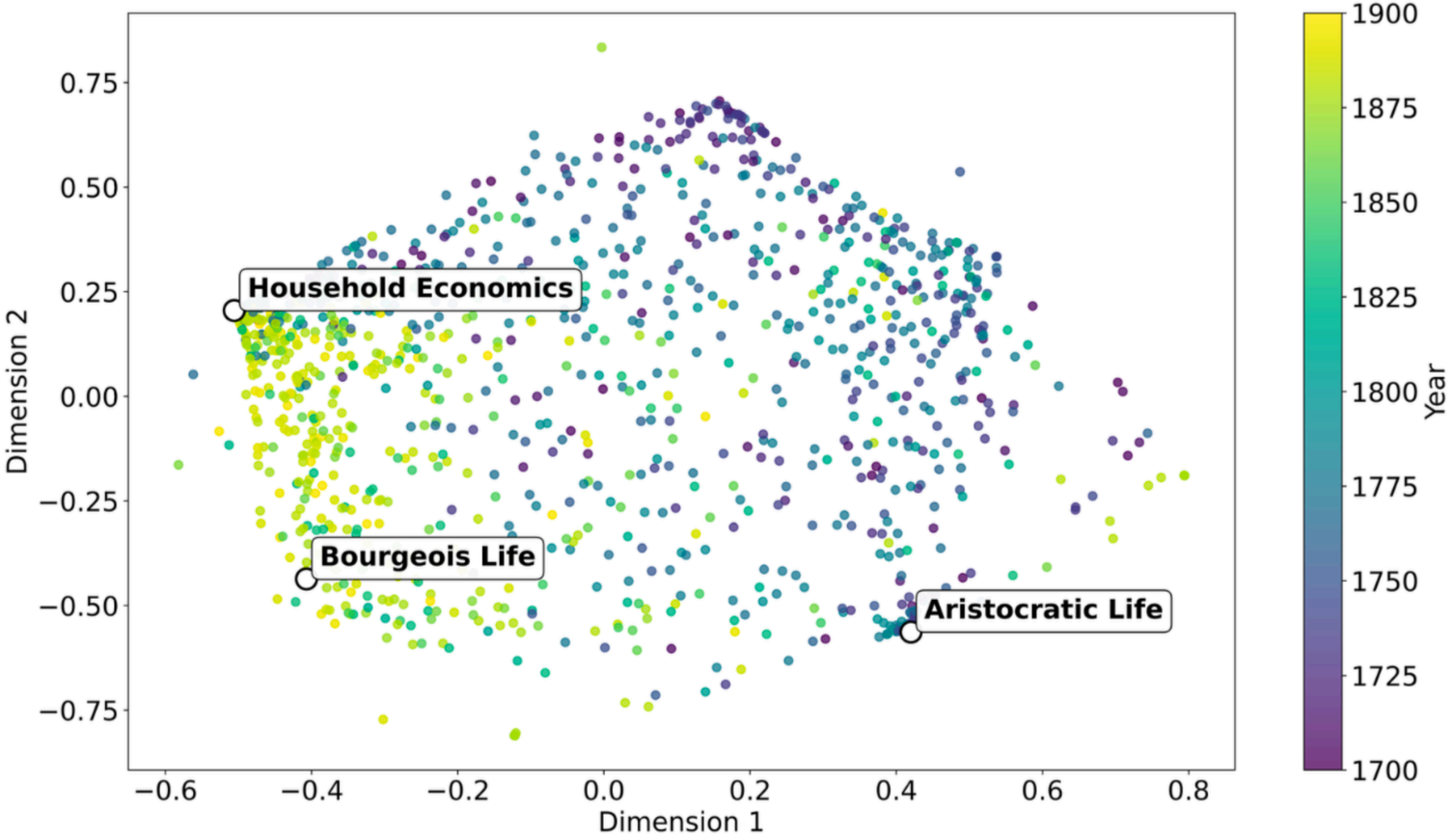
money, devil, affair, poor, pay, husband, house, franc, rogue, eat, lady, letter, father, boy, drink

Household Economics

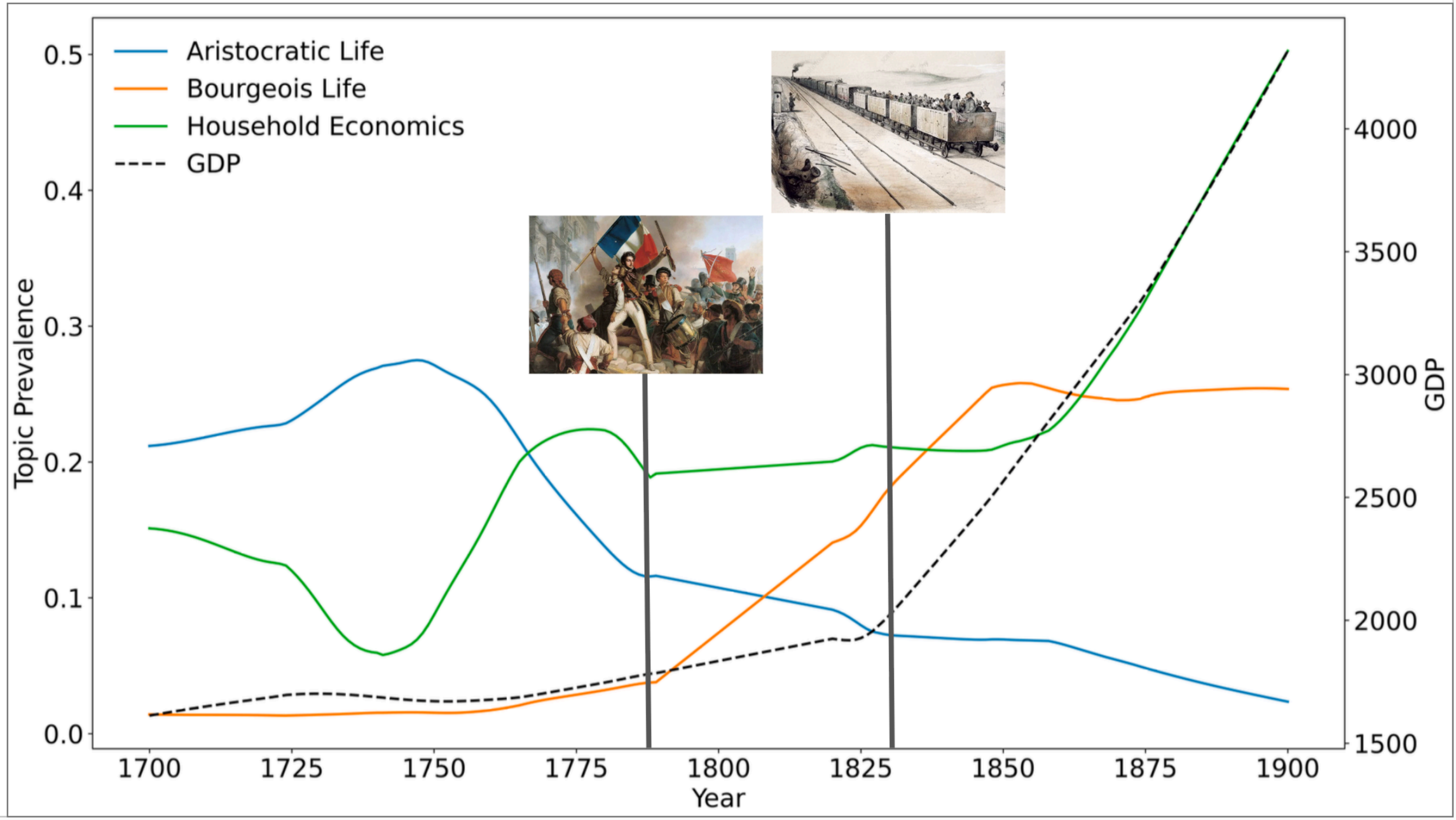
Results - JSD and Semantic Map:

Labels are at the centroid of the 5 representative works.





Results - Topic Prevalence vs Real GDP per capita in 2011 US$



Results: Max-Share Shock Model: Structural Vector Autoregressions (SVAR)

Cultural History

Political Economy

gdp shock and responses of variables

1820-1900

MODEL: gdp, human capital, life expectancy, household, bourgeois, aristocratic

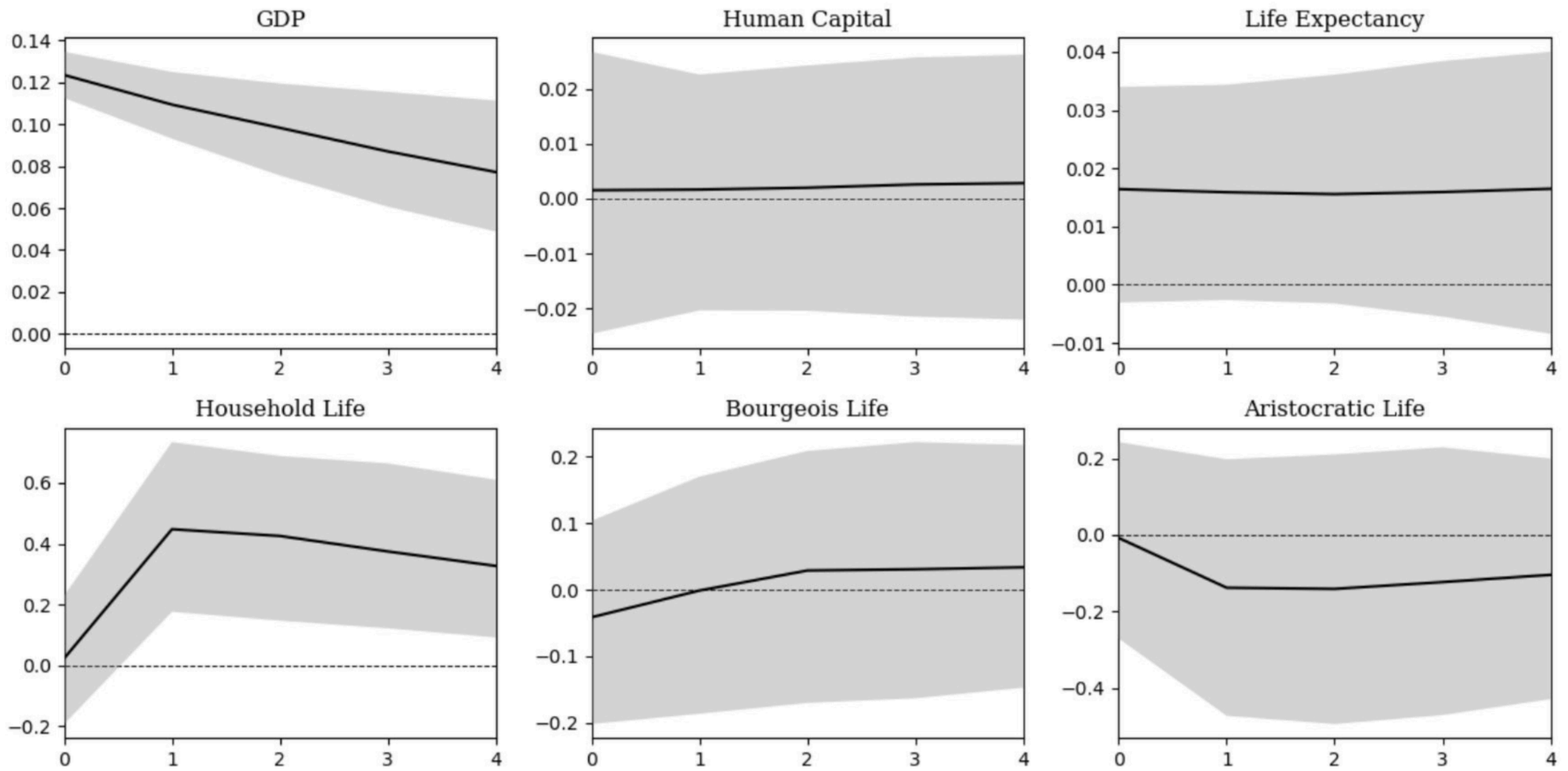
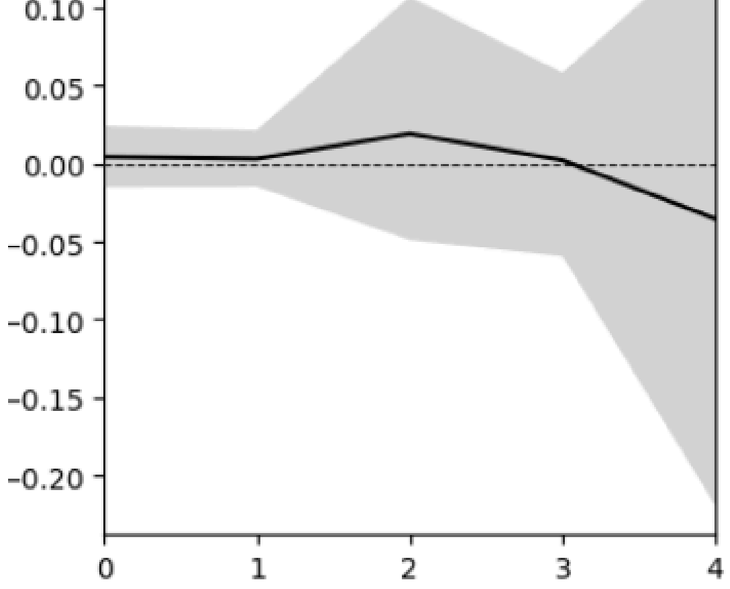
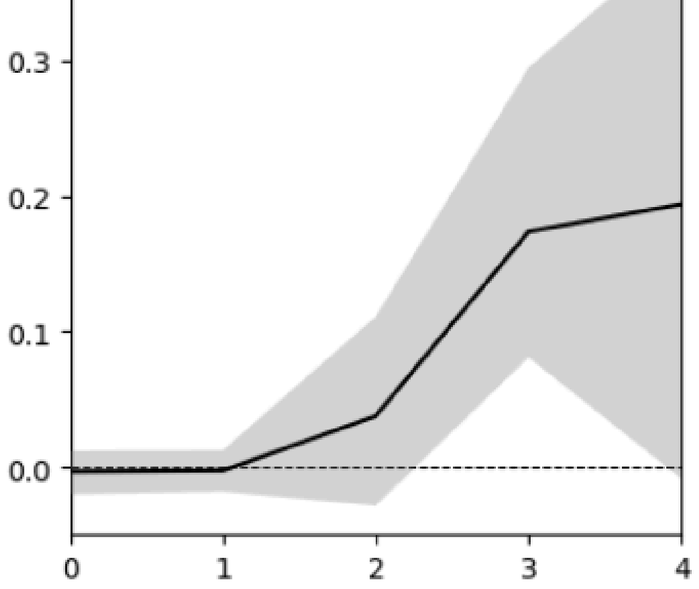


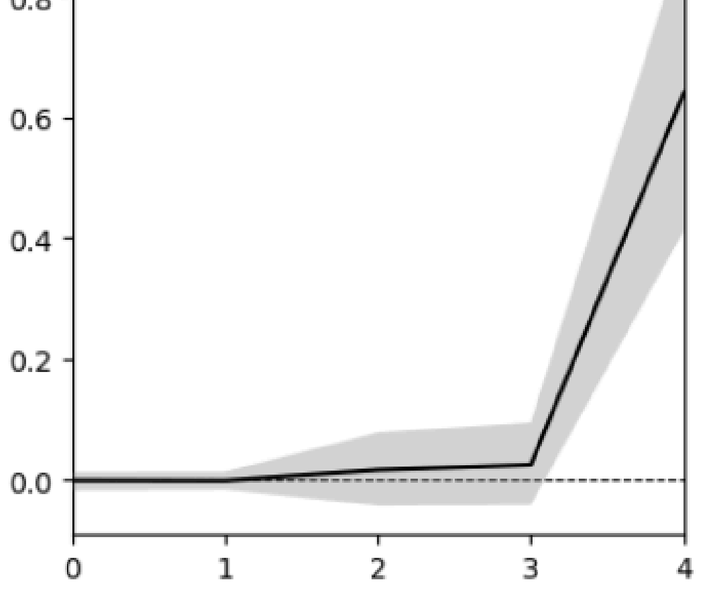
fig1 fig2 tab1 fig3 fig4 fig5 tab2 **fig6** tab3 fig7 fig8 tab4

Results: Max-Share Shock Model: Structural Vector Autoregressions (SVAR)

Cultural History

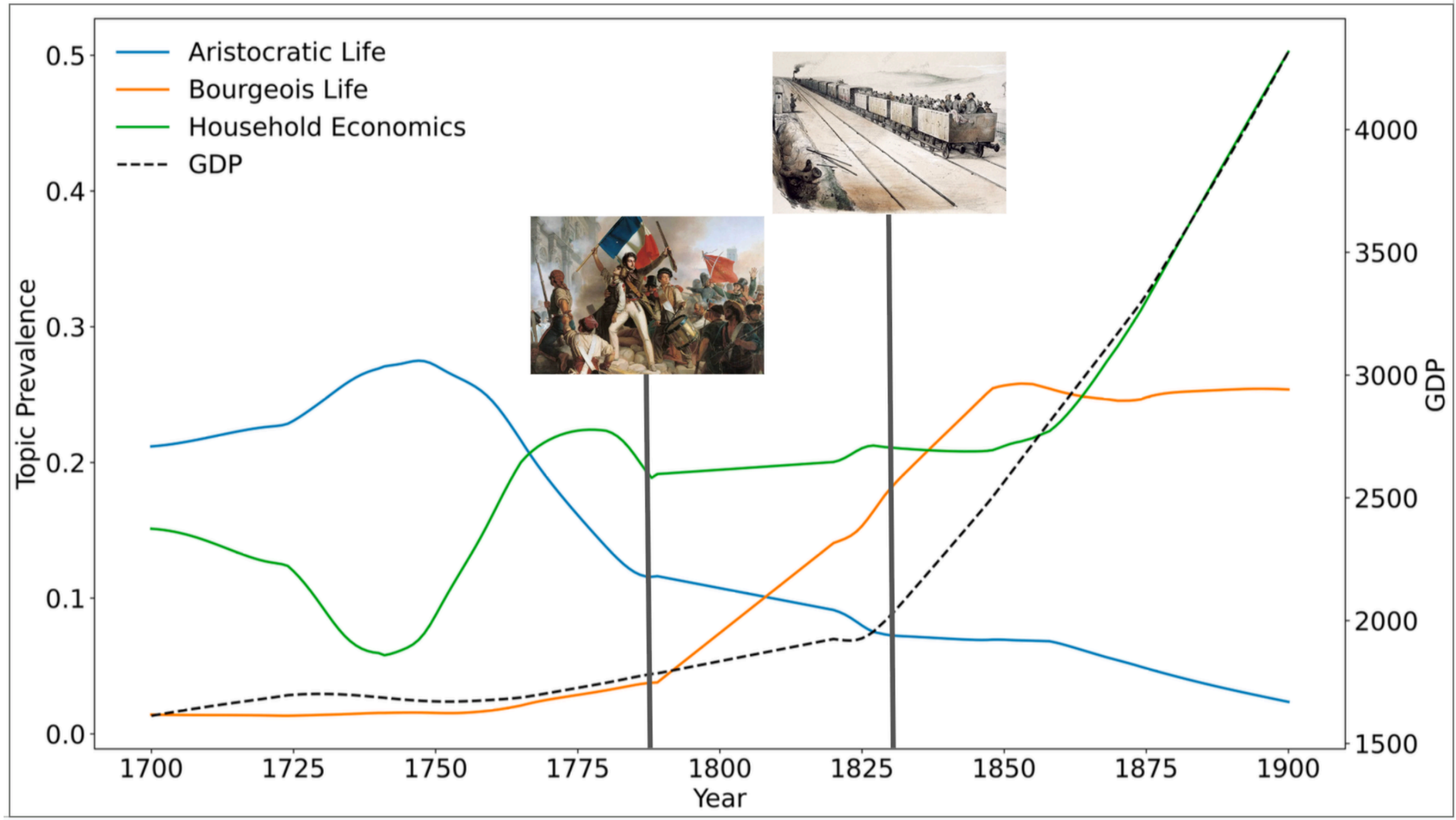






Political Economy


MERCI
BEAUCOUP


Aristocratic Life
Bourgeois Life
Household Economics
GDP
Topic Prevalence
GDP
Year
0.0
0.1
0.2
0.3
0.4
0.5
1500
2000
2500
3000
3500
4000
1700
1725
1750
1775
1800
1825
1850
1875
1900

Characters and Narratives

Different Approaches in Computational Drama Analysis

Botond Szemes, szemes.botond@ut.ee

Funded by
the European Union

DigiTS
Center for Digital
Text Scholarship

Computational Drama Analysis

- **Structure**
  - Netwrok-based metrics
  - Count-based metrics
- **Language**
  - Style
  - Topic

How they relate?

Character types

- Most research has focused on predefined character groups or types within a supervised classification framework (e.g., protagonist, schemer, gender, age, social class).
- The aim of this work is to complement these approaches with a bottom-up, **unsupervised** method.
- Instead of one multidimensional model, the method contrasts netwrok- and count-based metrics in the frist phase,
- and then examine whether the groups formed on the basis of structural characteristics also display distinct linguistic patterns through **keyword** analysis

Character types

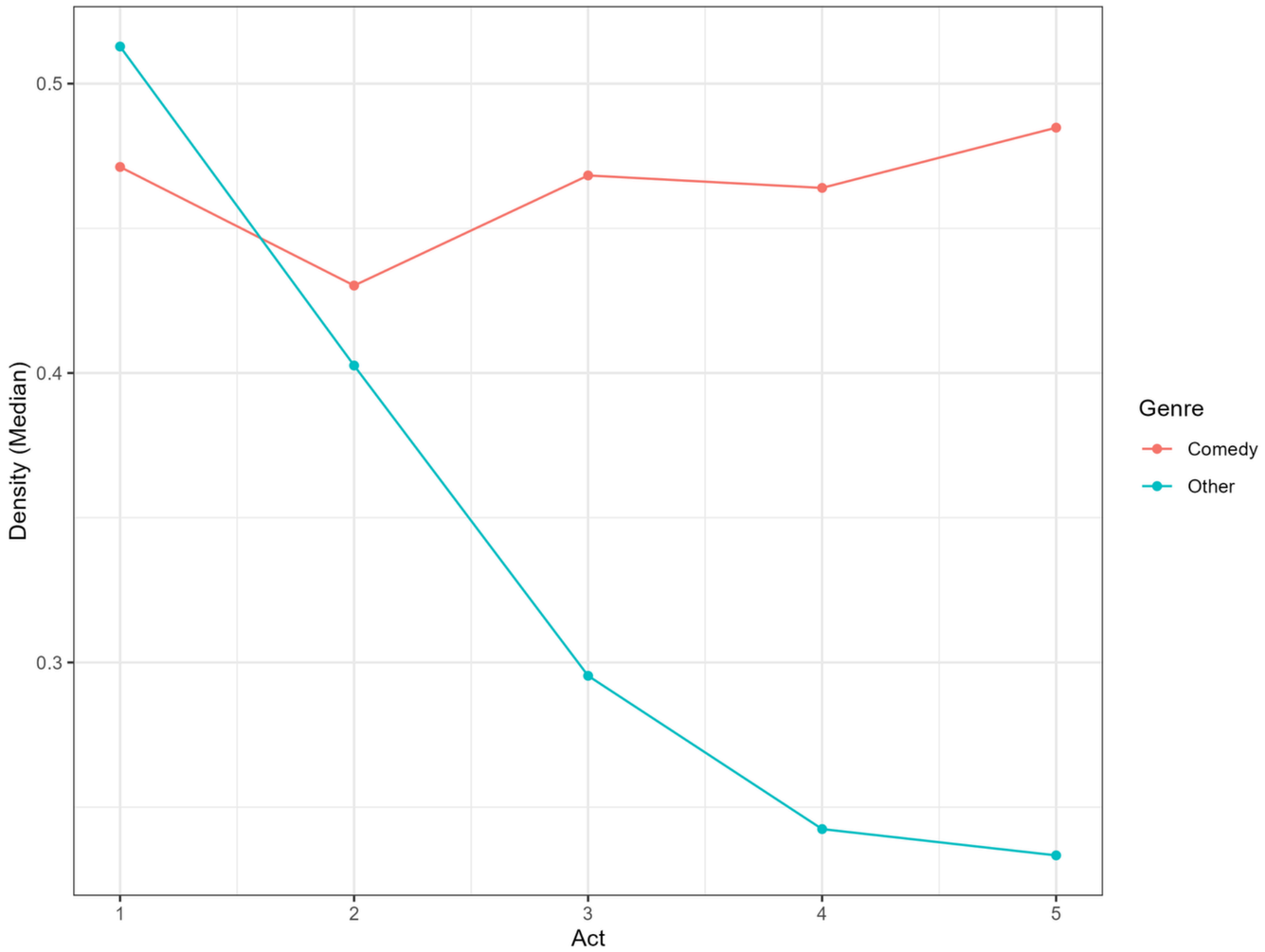
- Number of words, Number of Speech Acts and Betweenness Centrality
- Dominant characters: who speak the most and are also central to the social world of the drama,
- Connectors: who are central in the social network but do not speak the most,
- and Speakers: who have many lines but are less central in the network.

Character types

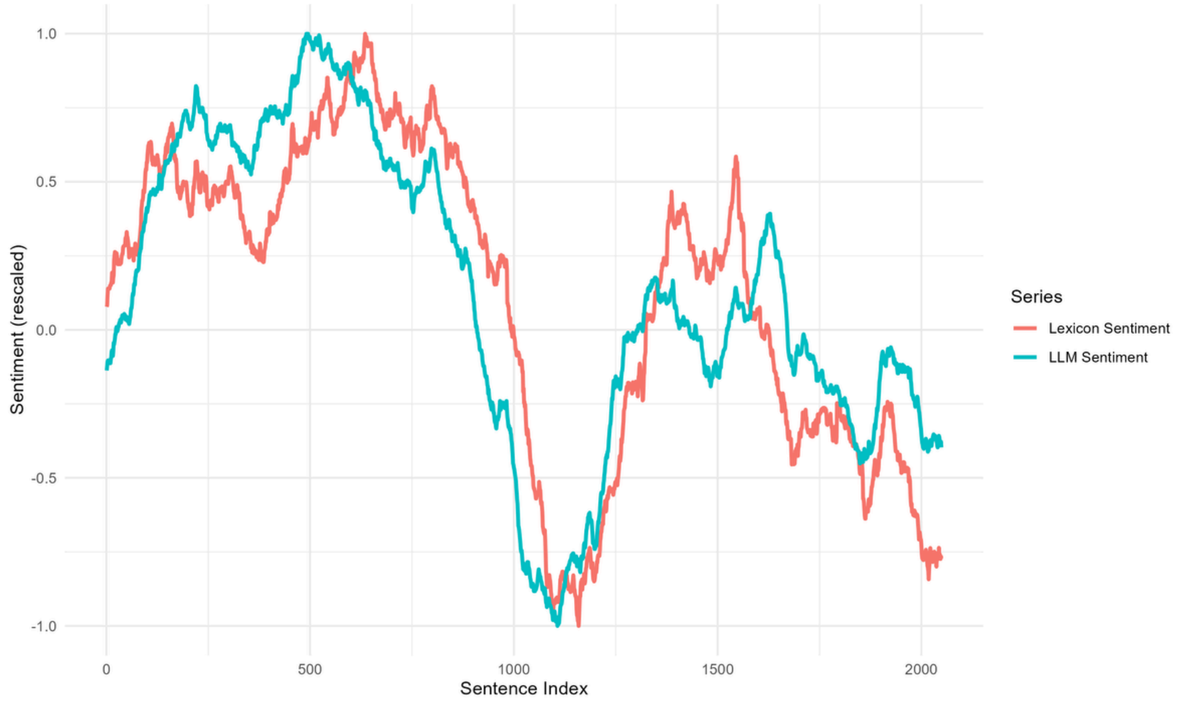
- Case study on Shakespeare: meaningful categories useful for interpretation and comparison
- The groups established differ not just based on structural features but linguistically based on keyword analysis
- **Speakers** can be characterized as characters who articulate themselves and mediate social bonds primarily through speech. Their language tends to foreground emotional and social expression, frequently giving voice to inner thoughts and relational concerns, and is oriented toward address and dialogue.
- **Connectors**, on the other hand, emerge as characters whose language is more closely aligned with managing, observing, and coordinating social interactions. Their speech is typically oriented toward socially functional tasks—such as mediating the external world, enforcing rules, or advancing the plot—rather than toward sustained self-expression.

Narrative arcs

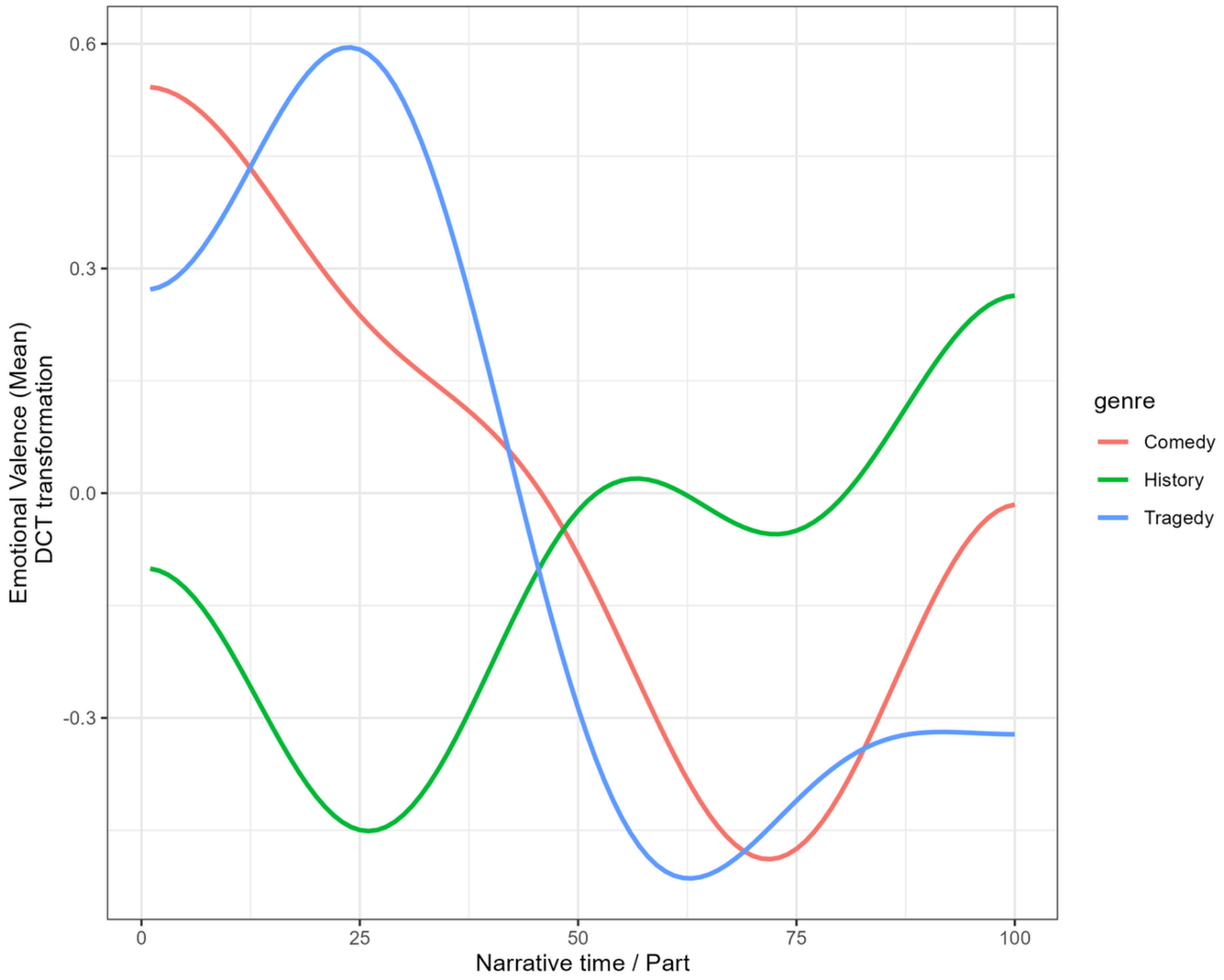
Compare the comulative changes of network metrics with arcs based on sentiment analysis



Narrative arcs



Narrative arcs



Thank you for your attention!

(Web) Corpora Without Fixed Category Labels: An Alternative Approach

Kristiina Vaik
Research Fellow in Digital Humanities

Funded by the European Union

DigiTS
Center for Digital Text Scholarship

Who am I?

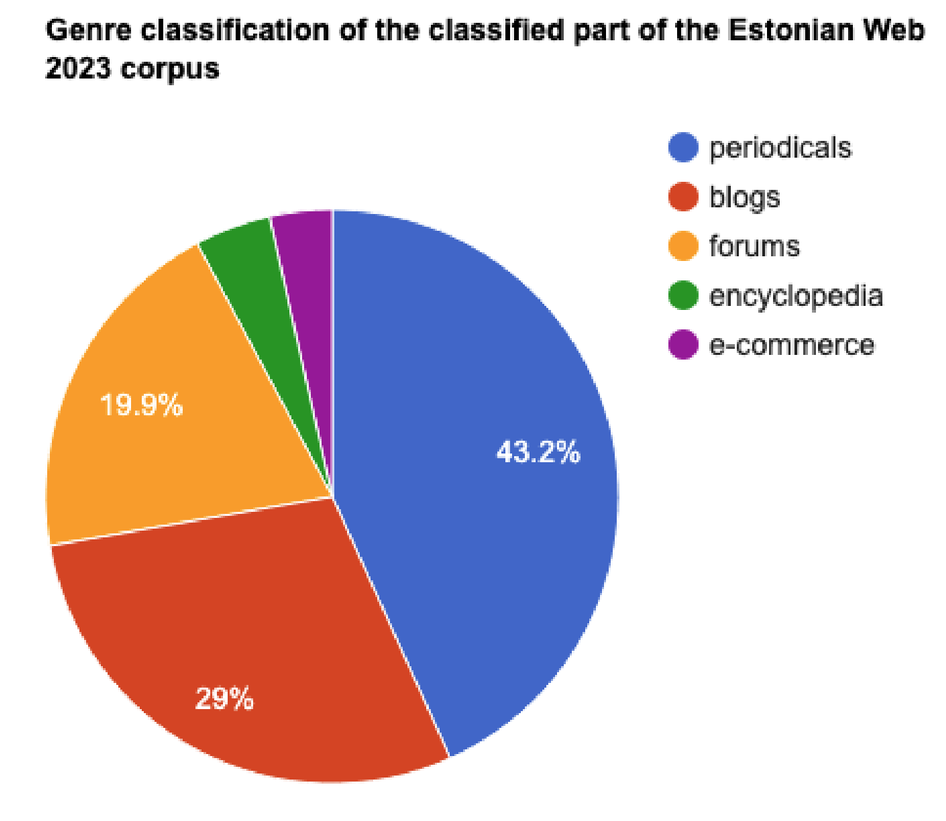
- currently a Research Fellow in Digital Humanities
- education:
  - BA and MA in General Linguistics, University of Tartu
  - PhD (2024), University of Tartu
- research interests:
  - automatic processing of natural language (morphology, syntax, semantics)
  - large language corpora and text mining (as a broad term)
  - sense of mission: working with the Estonian language

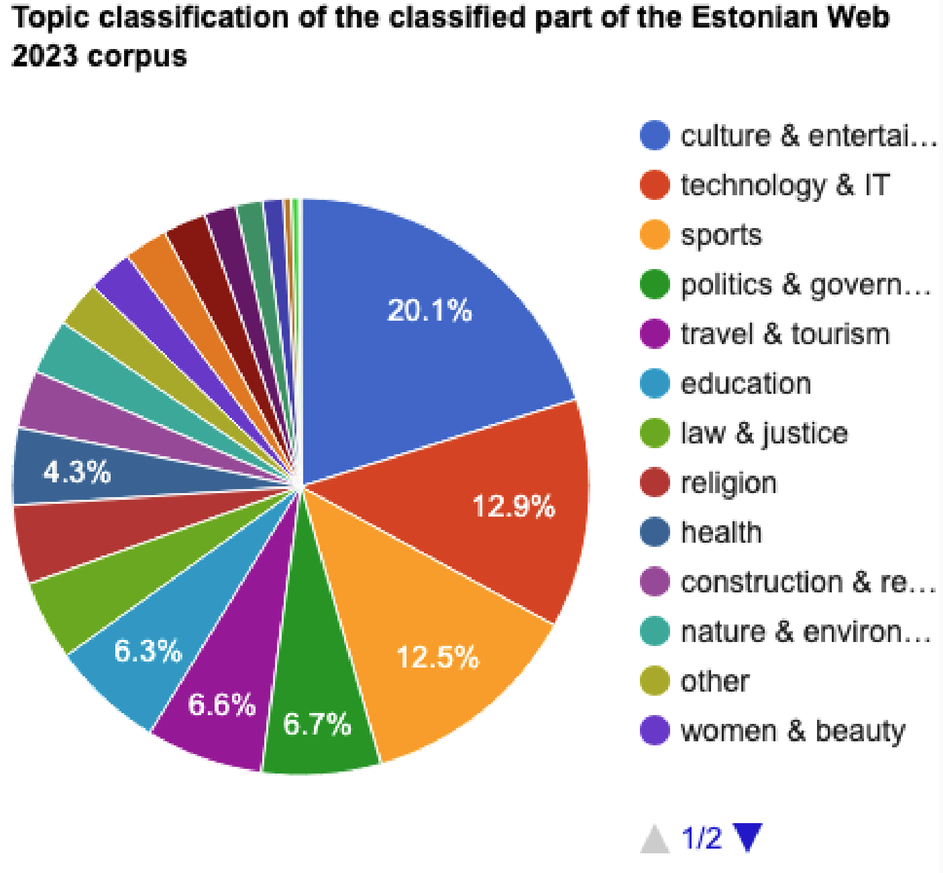
The Core Problem

Web corpora have lots of data, but...

- the composition is unknown
- texts do not fit into the existing genre taxonomies because
    - user- and computer-generated content, e.g. blogs, posts on social networks, content marketing, etc. (Santini et al. 2010),
    - the language of LLMs: instruction-tuned models generate text that does not align with traditional genre conventions (Reinhart et al. 2025).
- automatic classification requires reliable predefined labels
    - but category boundaries are fuzzy which means that reliably annotated datasets are a challenge (Suchomel 2020).

etTenTen 2023 content





* The TenTen corpus family is a set of comparable web text corpora, i.e. collections of texts that have been crawled from the Web and processed to match the same standards.

Rethinking the Starting Point

- traditional approaches: classify texts into fixed categories

- the proposed approach → from discrete categories to continuous functional dimensions

- instead of “Which category does this text belong to?” → “What is the communicative function of this text?”
  - to convey information? express emotion?

- assume the function manifests through a set of co-occurring linguistic features,

- develop a language- and corpus-independent framework and model continuous functional dimensions in multidimensional space

The 12 Dimensional Space

D1 abstractness	D7 impersonality
D2 affectivity	D8 temporality
D3 instructability	D9 interactivity
D4 information density	D10 subjectivity
D5 spontaneity	D11 complexity
D6 formality	D12 argumentativity

These are not categories.

A text is not “formal” or “informal”.

Instead, texts are viewed on a continuum - a text is more or less X.

Research Questions

(RQ1) do these dimensions exist in actual data?

- annotation study:
  - a subset from the etTenTen corpus (Koppel & Kallas 2022), - a continuous scale Likert approach: dimension is present in a text strongly, moderately, weakly, or not present at all.

- for consistency: inter-annotator agreement
- for collinearity and latent factors: correlation and exploratory factor analysis.

(RQ2) if yes, how do they vary linguistically?

- automatic feature extraction
  - lexical/textual (e.g., type/token ratio, avg. sentence length, core verbs) and grammatical (verb categories, parts of speech, subject-object relations) features etc.
- quantitative analysis to determine the linguistic fingerprints

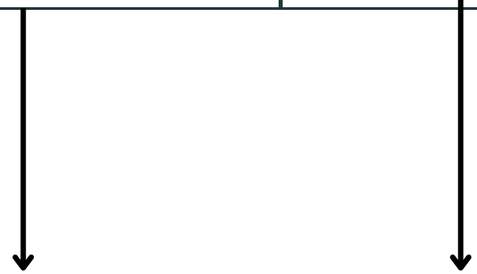
Do dimensions actually exist?

substantial and moderate IAA

Dimension	α	Dimension	α
subjectivity	.76	instructability	.47
affectivity	.74	interactivity	.46
formality	.63	impersonality	.42
spontaneity	.6	temporality	.4

fair IAA

Dimension	α
complexity	.38
argumentativity	.38
abstractness	.33
information density	.25

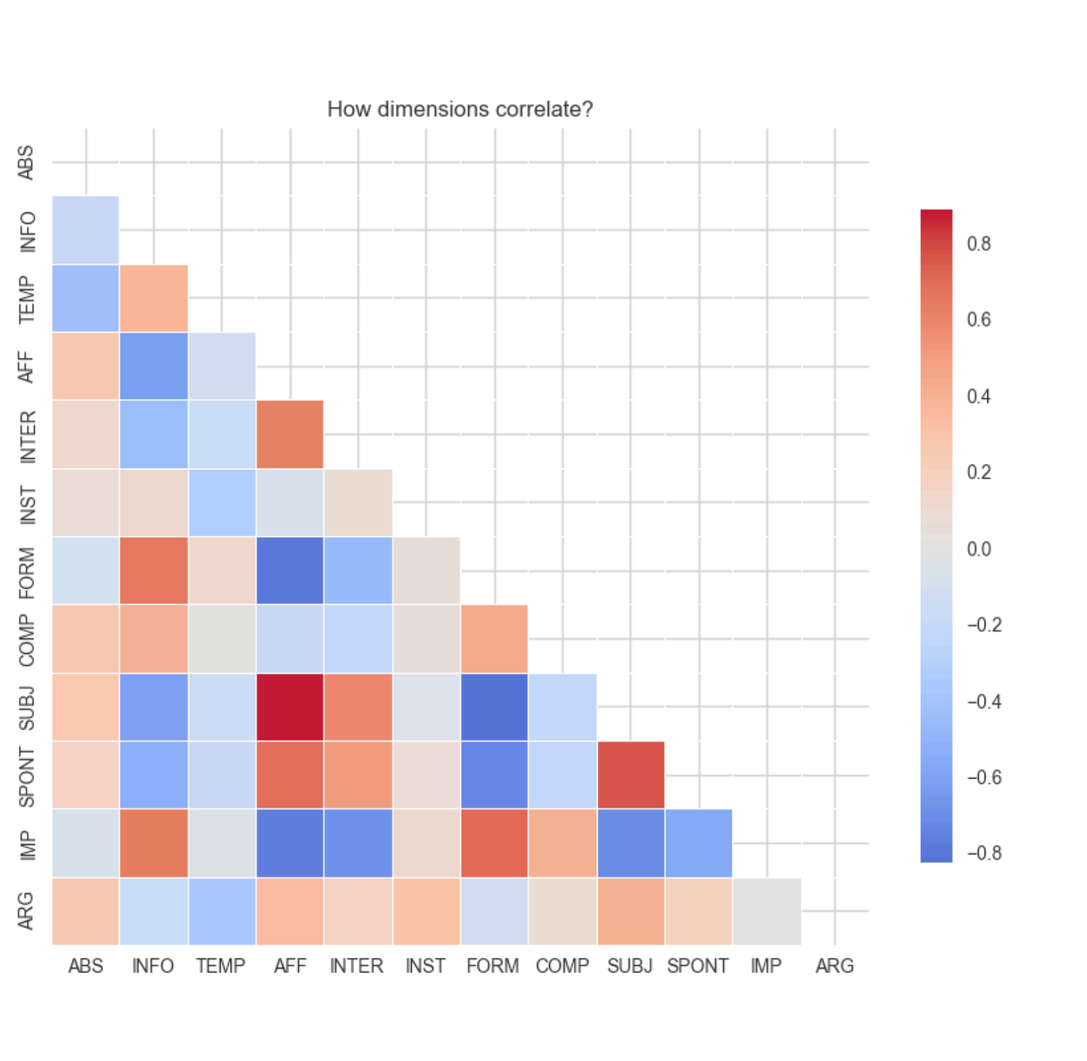


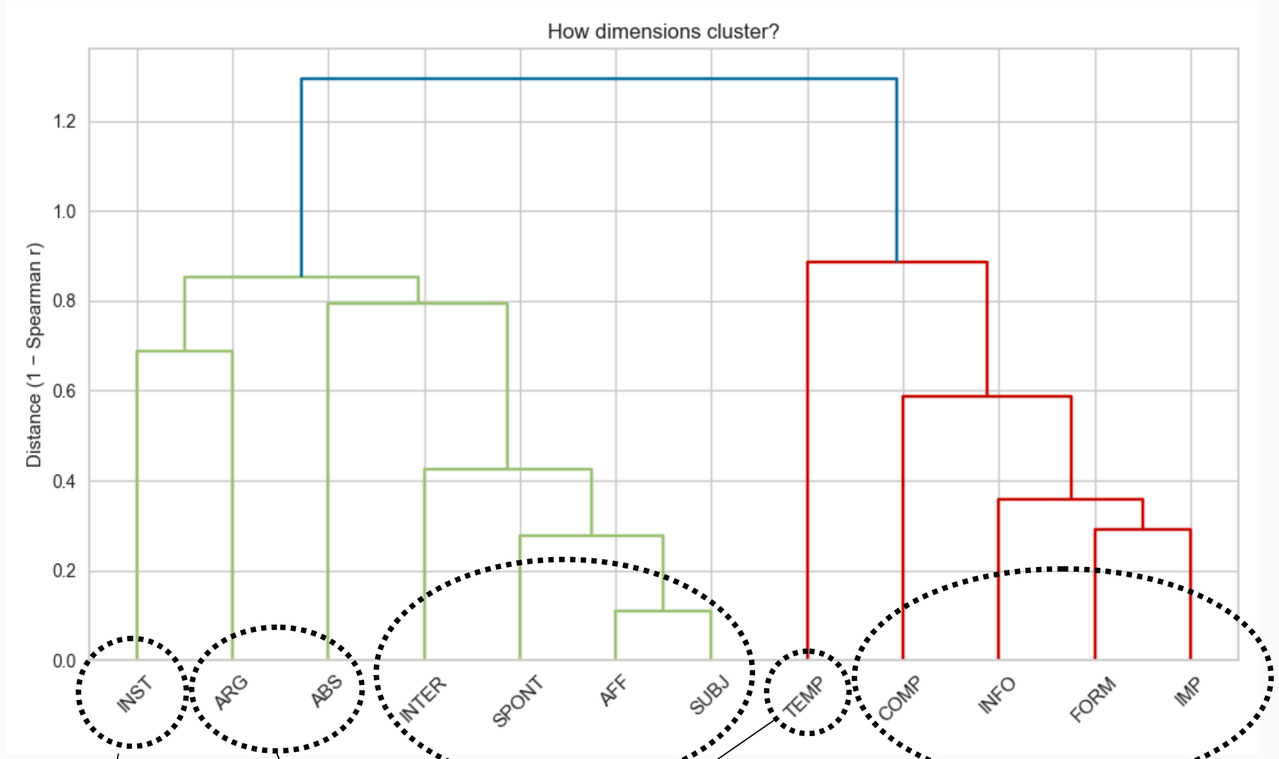
satisfactory

not great, not terrible

??

Overall, the results suggest that these functional dimensions are not arbitrary. They can be reliably identified in actual data.





Key takeaway: this dimensional space has structure and these dimensions are not random: they form somewhat meaningful groupings.

Factor 3

Factor 2

Factor 1

verbal vs nominal

Dimensional Variation?

- **all dimensions exhibited a distinct linguistic profile**, for example

  - interactivity can be characterized by having a smaller vocabulary, using more vocatives and other features associated with verbal style of writing (e.g., interjections, modal/finite verbs, 2nd prs plural pronouns (such as ‘teie’ in Estonian));

  - **temporal** texts use less infinite verb forms, more past tense and numerical modifiers (which shows that sequential order of event happening in the past is important);

  - more **formal** texts use longer words and sentences, and use grammatical features which are more common for nominal style of writing (e.g., impersonal voice, nouns, nominal modifiers);

  - more **subjective** texts use features which are related to simpler sentence structures, ‘see’ (it) pronoun and other grammatical features which are more common for verbal style of writing.

  - etc.

- **most features are shared across dimensions** (expected result because linguistic features serve overlapping communicative functions)

  - especially the AFF-INTER-SPONT-SUBJ vs FORM-IMP-INFO

Literature

- Koppel, K. and Kallas, J. (2022). Estonian National Corpus 2013–2021: The largest collection of Estonian language data [‘Eesti keele Ühendkorpuste sari 2013–2021: mahukaim eestikeelsete digitekstide kogu’]. Eesti Rakenduslingvistika Ühingu aastaraamat = Estonian papers in applied linguistics, 18:207–228.
- Reinhart, A., Markey, B., Laudenbach, M., Pantusen, K., Yurko, R., Weinberg, G., and Brown, D. W. (2025). Do LLMs write like humans? Variation in grammatical and rhetorical styles. Proceedings of the National Academy of Sciences, 122(8).
- Santini, M., Mehler, A., and Sharoff, S. (2010). Riding the Rough Waves of Genre on the Web, pages 3–30. Springer Netherlands.
- Suchomel, V. (2020). Better Web Corpora For Corpus Linguistics And NLP. Ph.D. disseration, Masaryk University, Faculty of Informatics, Brno.

Exploring Text Similarity Measures: New Approaches

Maciej Eder
University of Tartu | Polish Academy of Sciences

2026-02-16

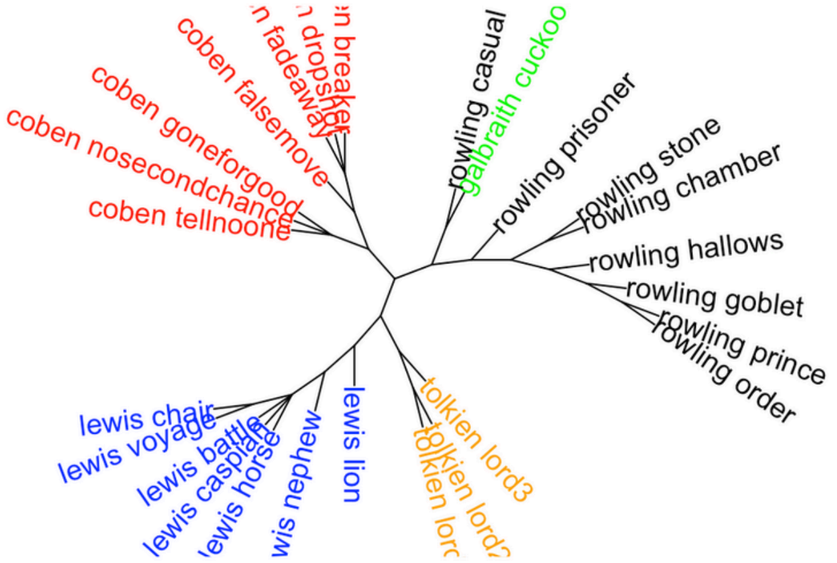
DigiTS

From features to similarities

	the	and	to	of	a	was	I	in
coben_breaker	3.592	1.175	2.163	1.376	2.519	1.502	1.445	1.176
coben_dropshot	3.588	1.179	2.122	1.269	2.375	1.567	1.497	1.040
coben_fadeaway	3.931	1.445	2.200	1.213	2.306	1.323	1.330	1.198
coben_falsemove	3.625	1.613	2.134	1.237	2.401	1.375	1.346	1.109
coben_goneforgood	3.834	1.817	2.153	1.176	1.962	1.733	3.814	1.131
coben_nosecondchance	4.098	1.589	2.271	1.206	1.992	1.758	3.855	1.151
coben_tellnoone	4.102	1.790	2.031	1.246	2.176	1.418	3.499	1.162
galbraith_cuckoos	4.523	2.267	2.494	2.179	2.141	1.656	1.127	1.380
lewis_battle	5.051	3.405	2.138	2.138	1.960	1.511	0.902	1.284
lewis_caspian	4.865	3.592	2.153	2.144	2.168	1.353	1.115	1.212
lewis_chair	4.973	3.221	1.997	2.103	2.354	1.405	1.073	1.214
lewis_horse	4.885	3.487	2.306	2.224	2.322	1.403	1.195	1.298
lewis_lion	5.141	3.699	2.295	2.185	2.100	1.346	0.813	1.162
lewis_nephew	4.482	2.856	2.070	2.231	2.311	1.571	1.179	1.355
lewis_voyage	5.222	3.279	2.261	2.114	2.244	1.583	1.048	1.153
rowling_casual	4.749	2.639	2.625	2.108	1.763	1.646	0.561	1.443
rowling_chamber	4.415	2.344	2.352	1.877	2.001	1.481	0.882	1.168
rowling_goblet	4.483	2.426	2.486	2.022	1.791	1.423	0.849	1.117

DigiTS

What we hope to get



DigiTS

What is a distance?

take any two texts:

```
                the   and    to    of     a   was     I    in    he  said   you
lewis_lion    5.141 3.699 2.295 2.185 2.100 1.346 0.813 1.162 1.087 1.426 1.141
tolkien_lord1 5.624 3.782 2.074 2.597 1.916 1.313 1.492 1.419 1.221 0.825 0.872
```

subtract the values vertically:

```
          the    and    to     of     a   was      I     in     he  said   you
       -0.483 -0.083 0.221 -0.412 0.184 0.033 -0.679 -0.257 -0.134 0.601 0.269
```

then drop the minuses:

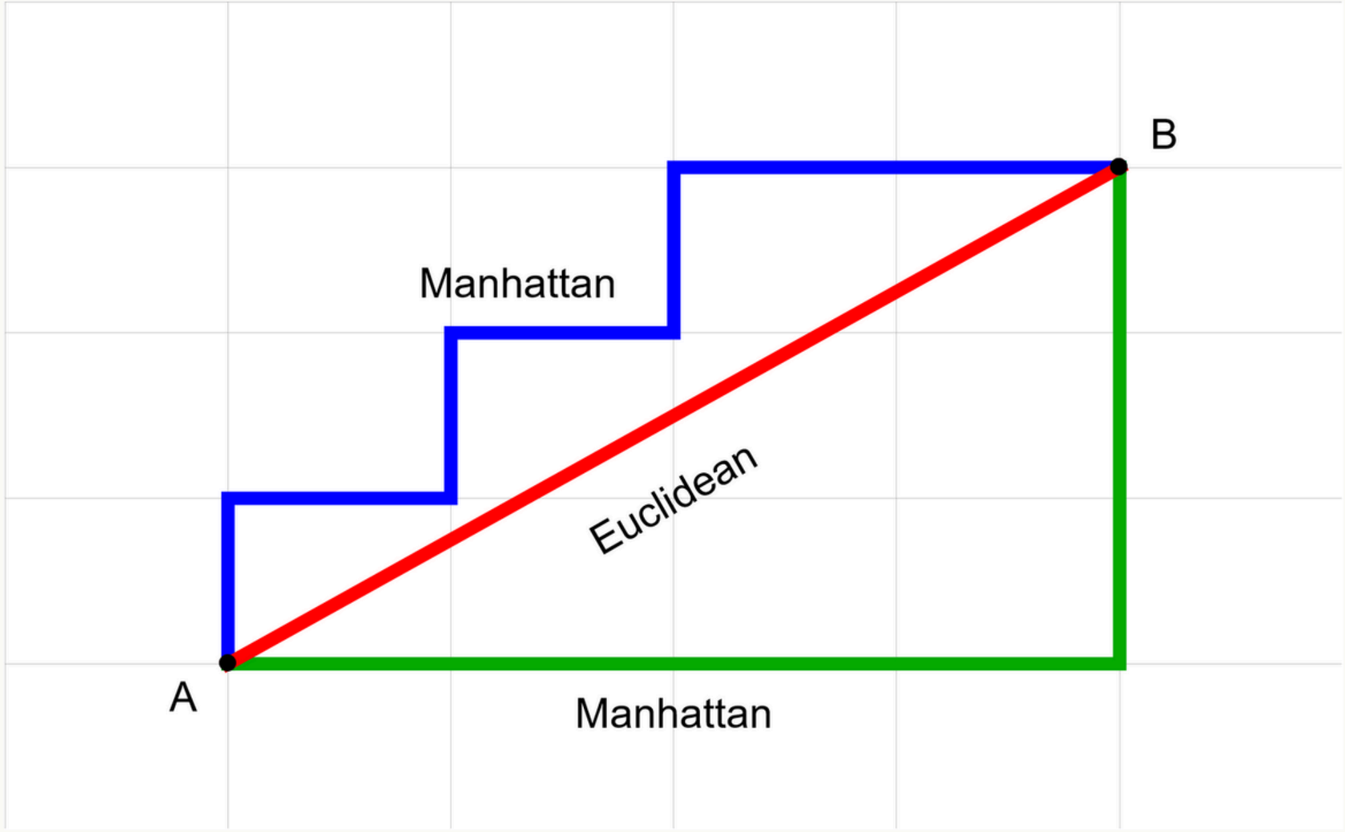
```
              the   and    to    of     a   was     I    in    he  said   you
            0.483 0.083 0.221 0.412 0.184 0.033 0.679 0.257 0.134 0.601 0.269
```

sum up the obtained values:

```
[1] 3.356
```

DigiTS

Manhattan vs. Euclidean



DigiTS

Euclidean distance

between any two texts represented by two points A and B in an n-dimensional space can be defined as:

$$\delta_{AB} = \sqrt{\sum_{i=1}^{n}(A_i - B_i)^2}$$

where A and B are the two documents to be compared, and A_i, B_i are the scaled (z-scored) frequencies of the i-th word in the range of n most frequent words.

DigiTS

Manhattan distance

can be formalized as follows:

$$\delta_{AB} = \sum_{i=1}^{n} |A_i - B_i|$$

which is equivalent to

$$\delta_{AB} = \sqrt[1]{\sum_{i=1}^{n} |A_i - B_i|^1}$$

(the above weird notation will soon become useful)

DigiTS

Euclidean and Manhattan are siblings!

$$\delta_{AB} = \sqrt[2]{\sum_{i=1}^{n}(A_i - B_i)^2}$$

vs.

$$\delta_{AB} = \sqrt[1]{\sum_{i=1}^{n}|A_i - B_i|^1}$$

For that reason, Manhattan and Euclidean are named L1 and L2, respectively.

DigiTS

An (infinite) family of distances

- The above observations can be further generalized
- Both Manhattan and Euclidean belong to a family of (possible) distances:

$$\delta_{AB} = \sqrt[p]{\sum_{i=1}^{n} |A_i - B_i|^p}$$

where p is both the power and the degree of the root.

DigiTS

The norms L1, L2, L3, … (and beyond)

- The power p doesn't need to be a natural number
- We can easily imagine norms such as L1.01, L3.14159, L1¾, L$\sqrt{2}$ etc.
- Mathematically, $p < 1$ doesn't satisfy the formal definition of a norm…
- … yet still, one can easily imagine a dissimilarity L0.5 or L0.0001.
- (plus, the so-called Cosine Distance doesn't satisfy the definition either).

DigiTS

🧐 How do the norms from a wide range beyond L1 and L2 affect text classification?

DigiTS

Data

Four full-text datasets used:

- 99 English novels by 33 authors,
- 99 Polish novels by 33 authors,
- 28 books by 8 American Southern authors:
  - Harper Lee, Truman Capote, William Faulkner, Ellen Glasgow, Carson McCullers, Flannery O'Connor, William Styron and Eudora Welty,
- 26 books by 5 fantasy authors:
  - J.K. Rowling, Harlan Coben, C.S. Lewis, and J.R.R. Tolkien.

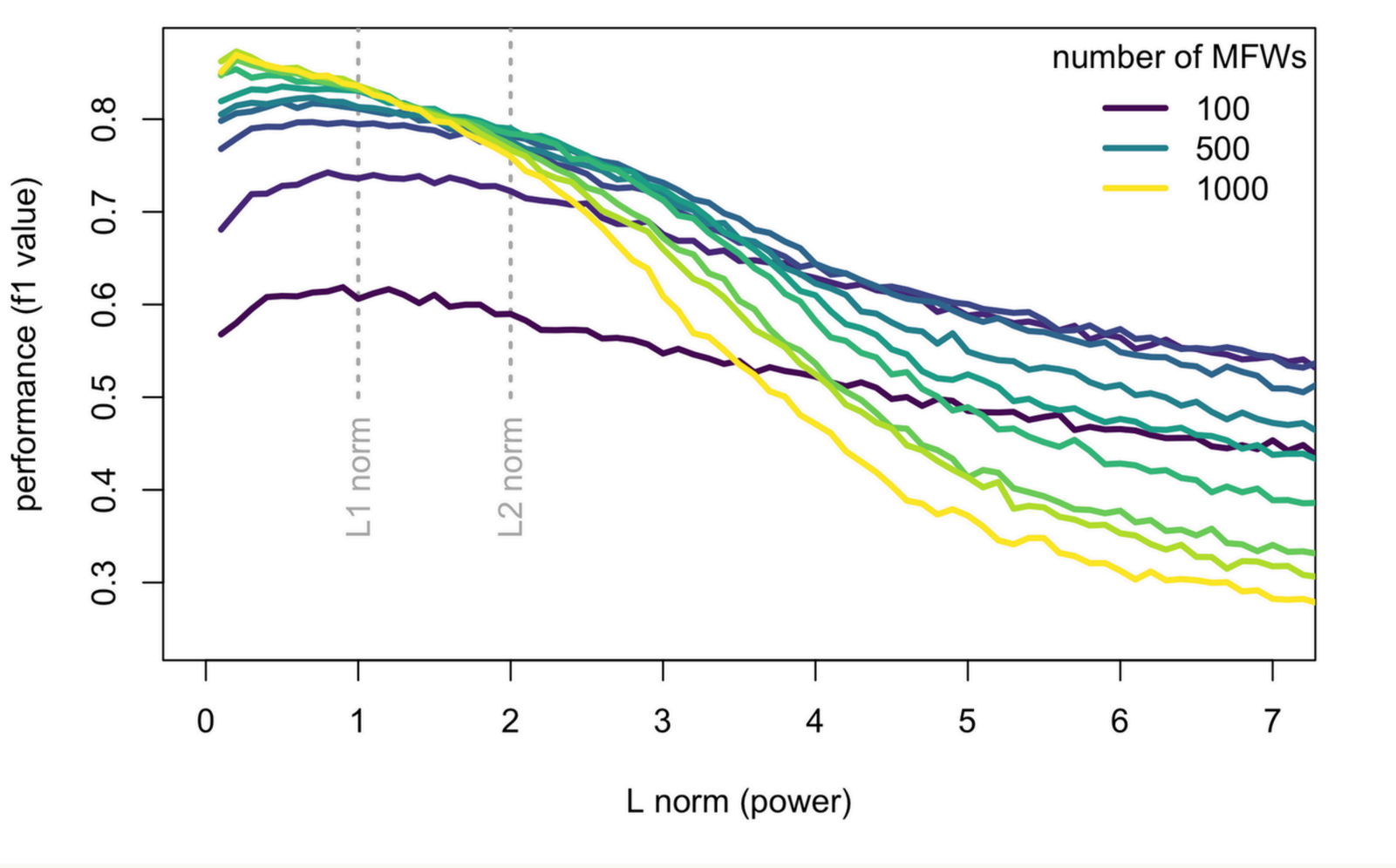
DigiTS

Method

- A supervised classification experiment was designed
- Aimed at authorship attribution
- leave-one-out cross-validation scenario
  - 100 independent bootstrap iterations…
  - … each of them involving 50% randomly selected input features (most frequent words)
  - The procedure repeated for the ranges of 100, 200, 300, …, 1000 most frequent words.
- The whole experiment repeated iteratively for L0.1, L0.2, …, L10.
- The performance in each iteration evaluated using accuracy, recall, precision, and the F1 scores.

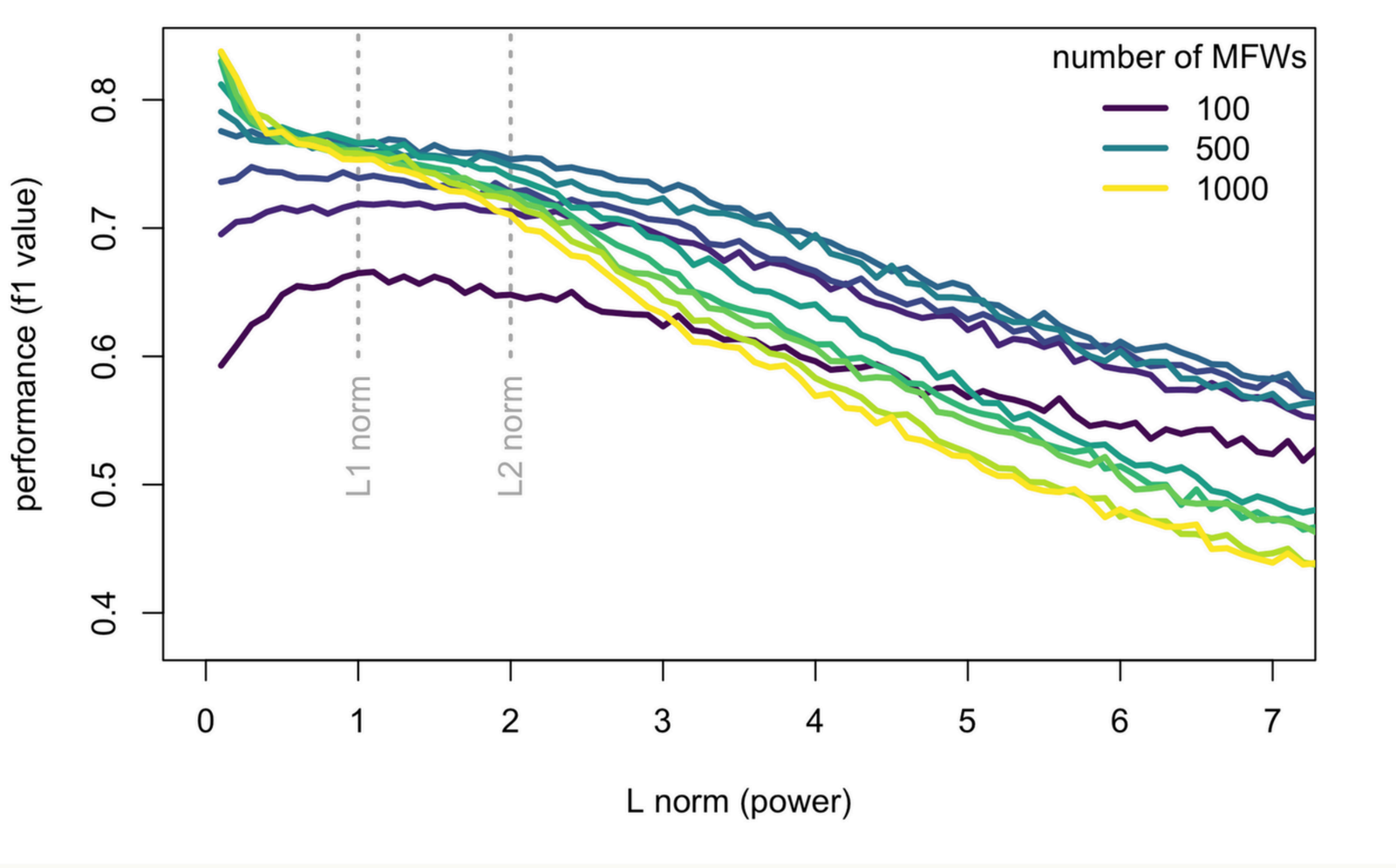
DigiTS

99 English novels by 33 authors



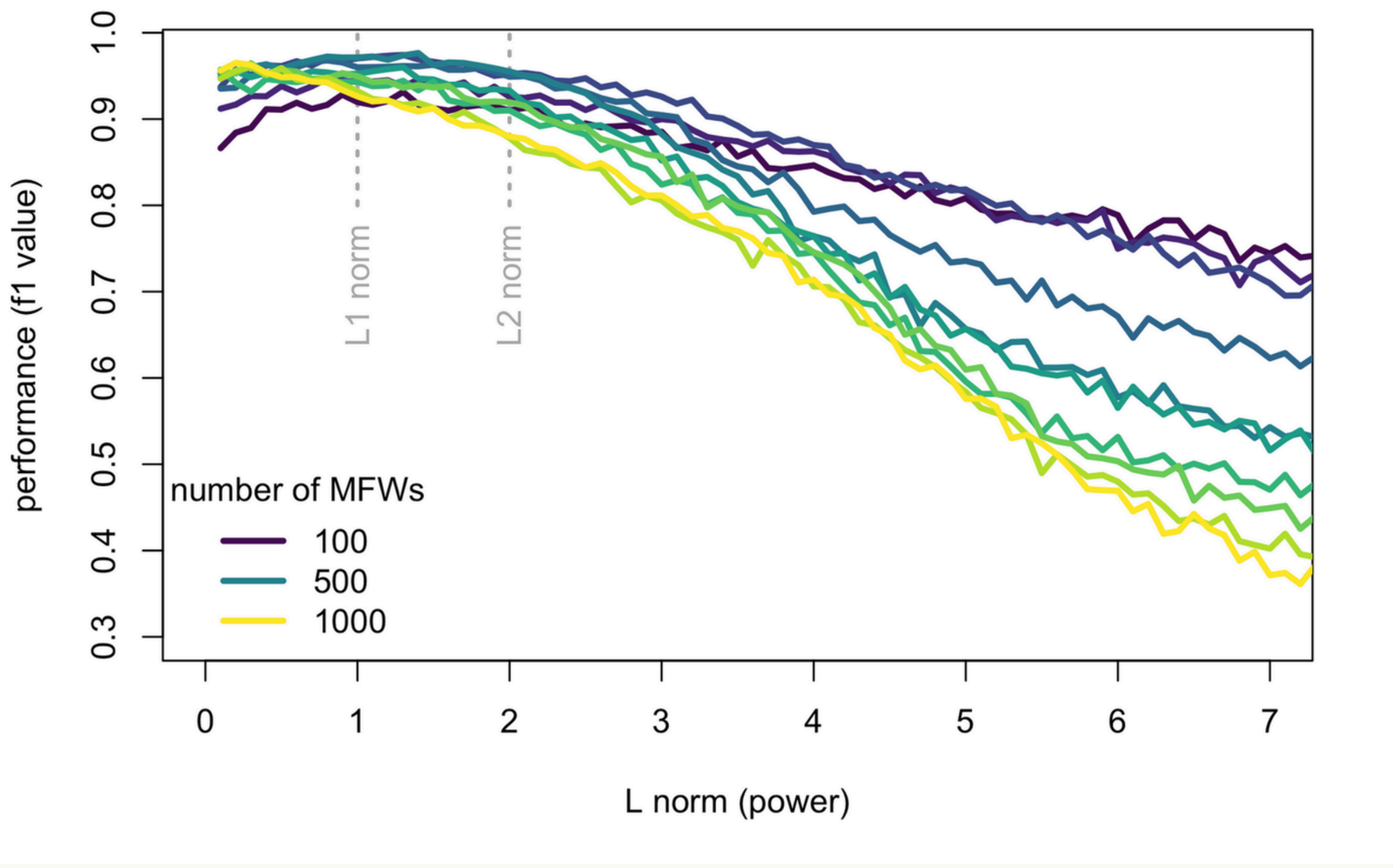
DigiTS

99 Polish novels by 33 authors



DigiTS

28 novels by 8 Southern authors



DigiTS

Plausible explanations

- Small p makes it more important for two feature vectors to have fewer differing features (rather than smaller differences among many features),
- Small p amplifies small differences (important, e.g., for low-frequency features in distinguishing between o difference – for two texts lacking a feature – and a small difference).

Therefore:

- Small p norms might be one way of effectively utilizing long feature vectors.

DigiTS

DigiTS

CORPORA AS A TOOL FOR EVALUATING GENDER STEREOTYPES IN LANGUAGE USE

Sofia Kriuchkova, junior researcher

UNIVERSITY OF TARTU

Funded by
the European Union

DigiTS

Gender Stereotypes and Language

Gender

- Socially constructed
- Learned through socialization
- Culture-dependent

Masculinity → strength, decisiveness, persistence etc.
Femininity → softness, emotionality, politeness etc.
Stereotypes influence:

- behavior
- profession
- language use

Examples:

- “Women are more emotional”
- “Men swear more”

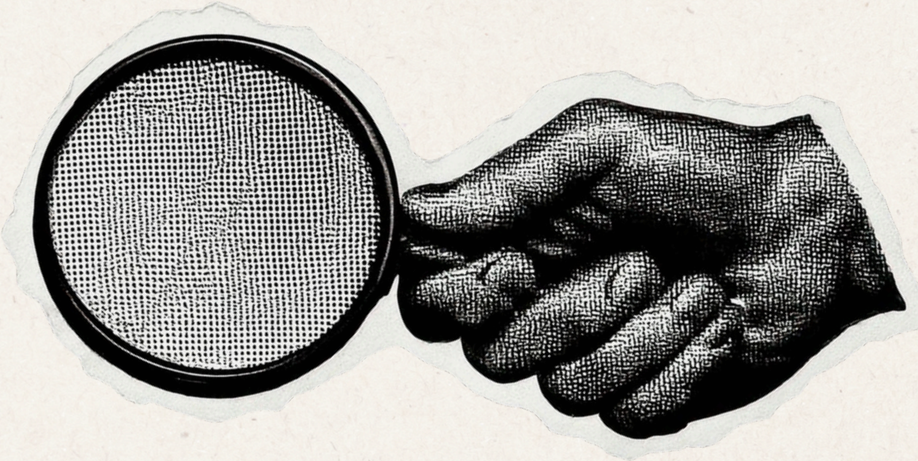
DigiTS

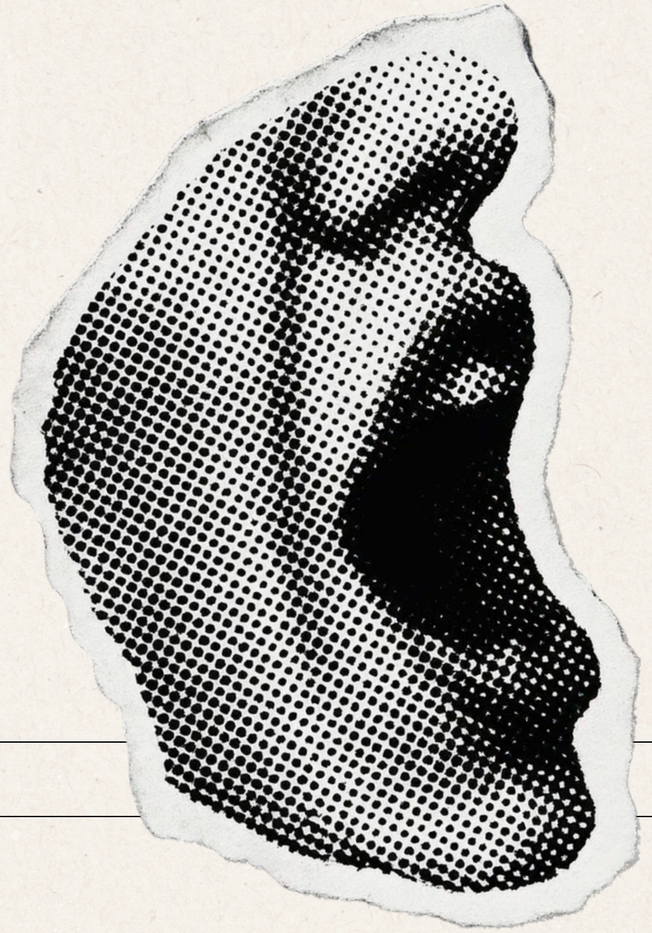
Corpora

- Large collections of authentic texts
- Reflect real usage
- Allow quantitative analysis
- Reduce reliance on intuition

DigiTS

Research Question





- Do Estonian men and women speak differently?
- What kinds of differences can be observed?
- Why might differences exist?

DigiTS

Research Material

Phonetic Corpus of Estonian Spontaneous Speech (Lippus et al. 2023)

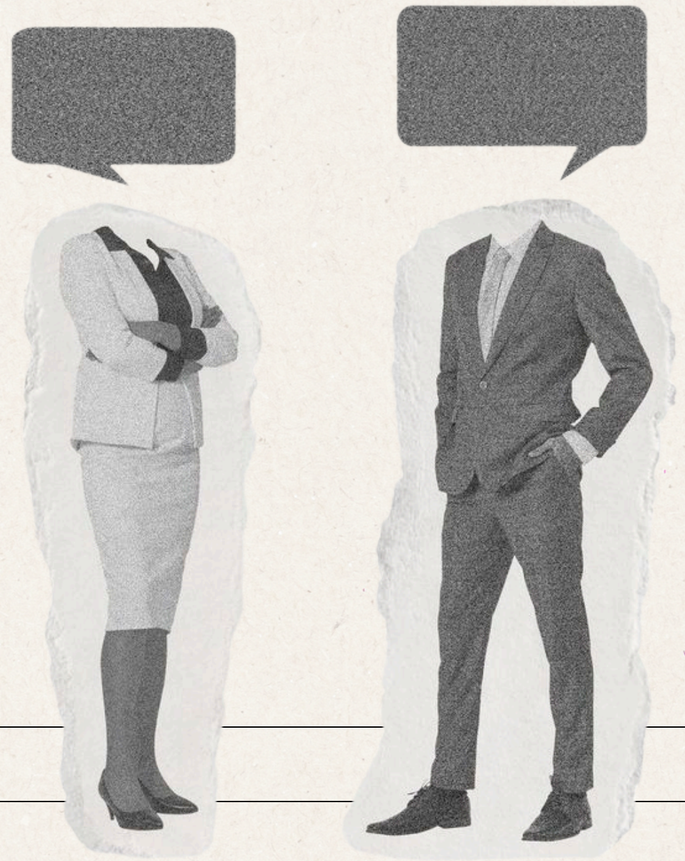
- 90 recorded dialogues
- 150 speakers
  - 73 men
  - 77 women
- Recorded between 2006–2022

Natural, everyday conversations.

Unit of analysis:
→ one speaker in one dialogue

DigiTS

Expectations Based on Stereotypes



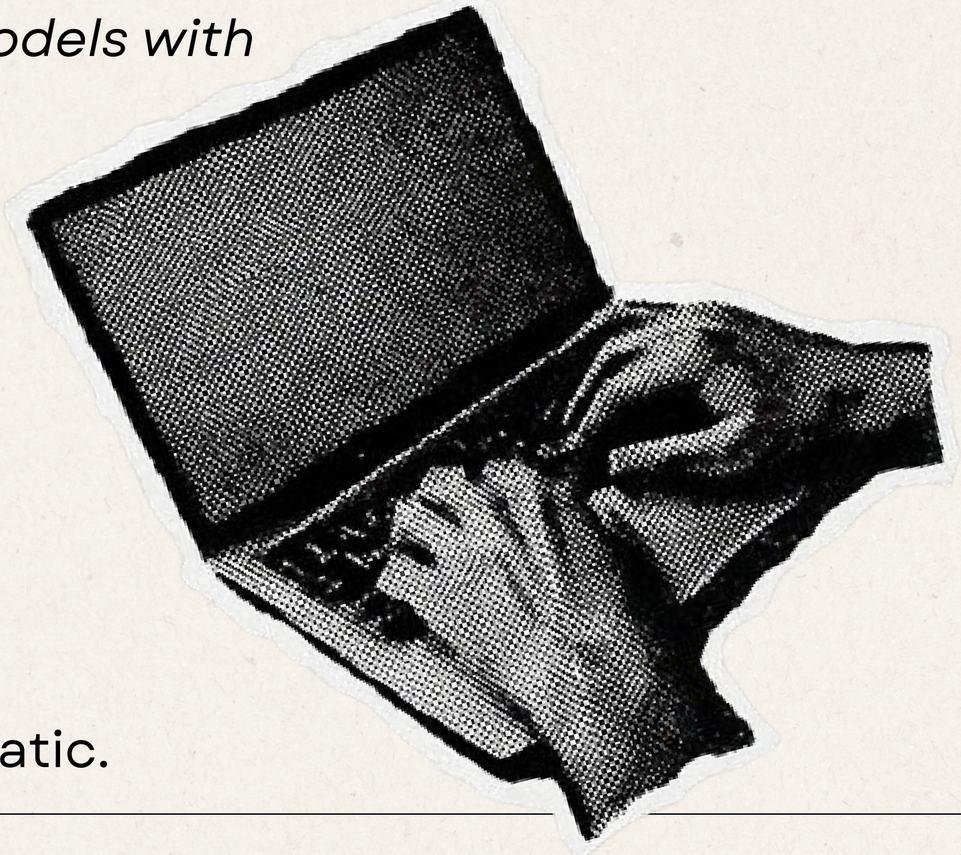
- Women use more probability particles (*vist* ‘probably’, *äkki* ‘maybe’ etc.).
- Men use more generalising and certainty particles (*muidugi* ‘sure’, *tõesti* ‘indeed’ etc.).
- Women use more apologies (*vabandust* ‘sorry’ etc.).
- Women use more intensifiers (*väga* ‘very’, *nii* ‘so’ etc.).
- Women prefer to use more socially acceptable emotional expressions (*jumal!* ‘God!’ etc.), whereas men use more swear words than women do (*sitt!* ‘shit!’ etc.).
- Women use more personal pronouns (*mina* ‘I’, *sina* ‘you’ etc.).

DigiTS

How the Analysis Works

I use statistical modelling (*generalised linear mixed models with a negative binomial distribution*) to:

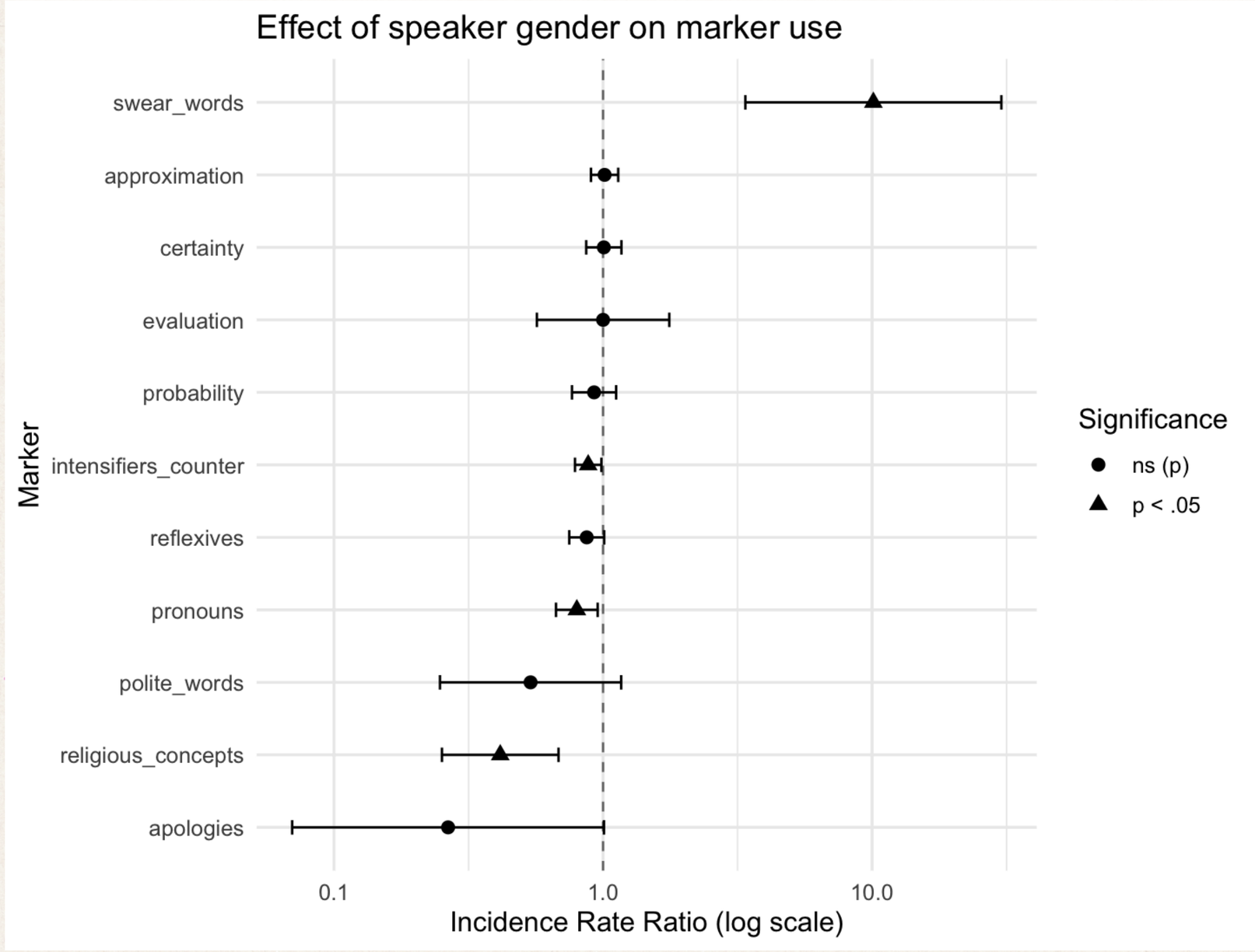
- compare speakers fairly
- control for:
  - how much they speak
  - age
  - recording year
  - communication setting
  - gender of the interlocutor



This allows me to test whether differences are systematic.

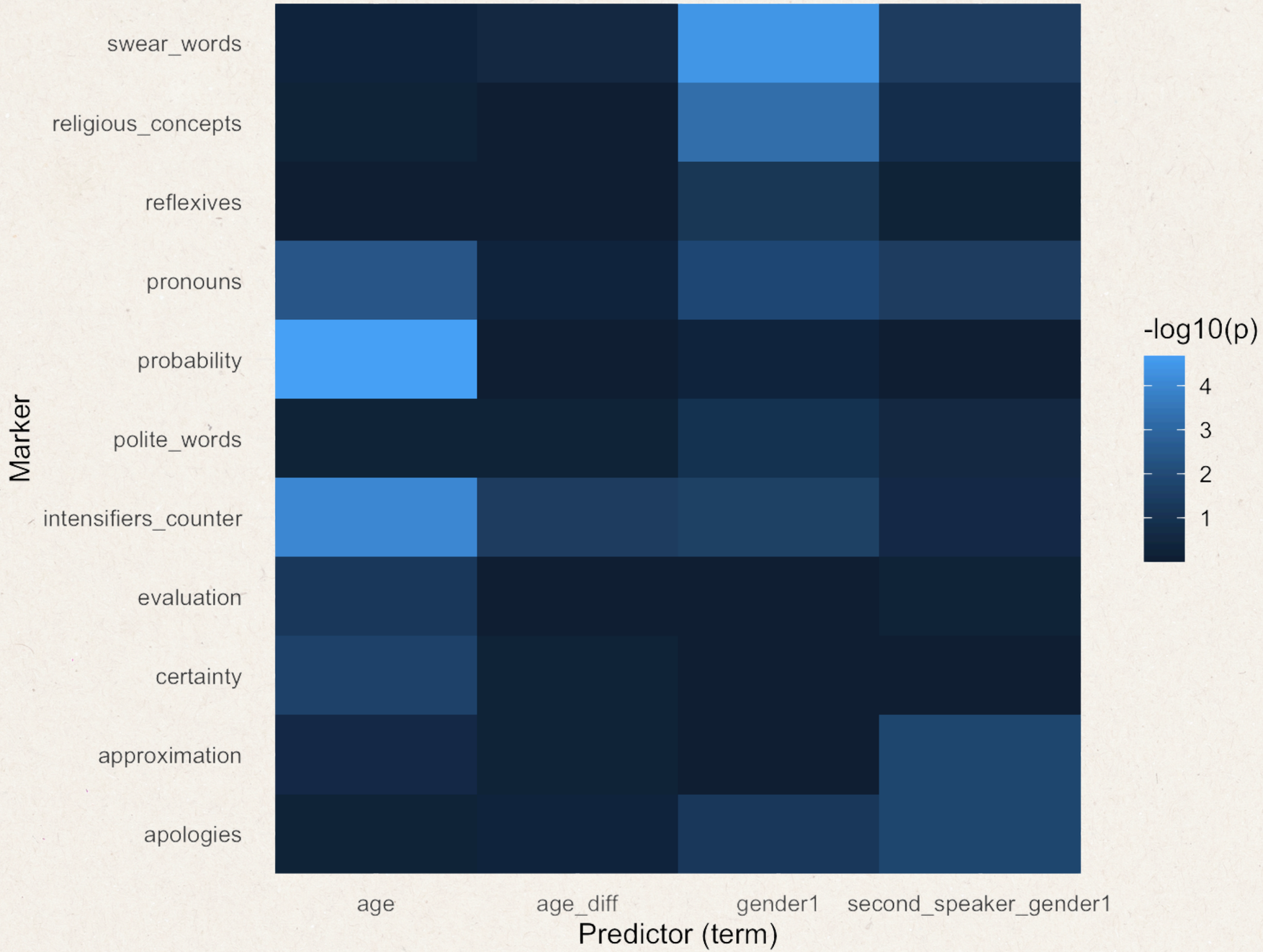
DigiTS

Results



DigiTS

Results



DigiTS

Conclusion

Overall, **gender** does matter, but its influence is selective and marker-specific rather than universal.

Corpora help us:

- evaluate widespread social beliefs
- distinguish real patterns from overgeneralisations
- ground discussions of gender and language in empirical data

They do not explain everything, but they allow us to test assumptions systematically.

Thank you for your attention!

APPLYING COMPUTATIONAL LINGUISTICS PERSPECTIVES ON DISINFORMATION

State of the Art, Limitations and Emerging Research Gaps

By Bhumika Bhattacharyya

17.02.26

WHEN STORIES MISLEAD: WHY THIS RESEARCH

- Fake news spreads faster than facts
- Emotional framing drives engagement
- Algorithms amplify manipulation

WHY DISINFORMATION IS TECHNICALLY & SOCIALLY COMPLEX

- Not Just “True vs False”
- Subtle framing, emotional triggers, cultural references
- AI Challenges : Black-box models, data imbalance (English datasets), low interpretability
- Manipulation is linguistic, contextual, and strategic...not just factual.

TWO LANGUAGES, TWO EXTREMES, ONE INSIGHT

Geopolitical Context :

- India → Large internal misinformation ecosystem
- Estonia → Target of foreign information warfare

Strategic value :

- Tests generalisability, enables transfer learning
- Is manipulation language agnostic?

BUILDING TOOLS THAT UNDERSTANDS HOW LANGUAGE PERSUADES

- High-Quality Data : Building a curated dataset with real+fake news data
- Linguistic modelling : factives, hedges, reporting verbs, modality, evaluative language

Pipeline :

Data → Annotation → Modeling → Interpretation → Evaluation

FROM RESEARCH TO REAL-WORLD IMPACT

- multilingual datasets, framing-aware AI models
- Better disinformation detection, media resilience

Small languages + scalable methods = high impact

THANK YOU