UNIVERSITÄT TRIER

Kompetenzzentrum
Trier Center for Digital Humanities

Towards Computational Comparative Literary Studies

Adressing the Challenges of Multilingualism

Christof Schöch
(Trier University, Germany)

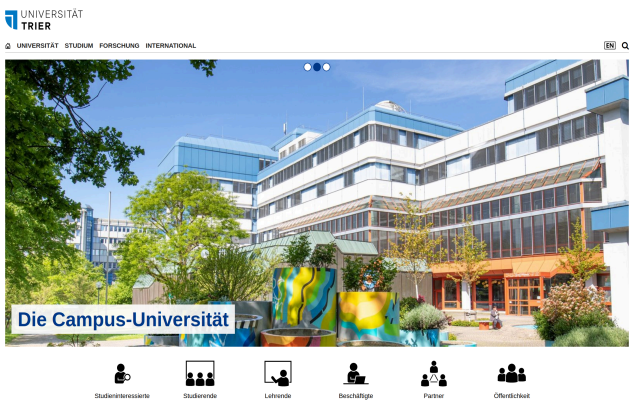
KEASTWEST Conference 2024
Dongguk University, Seoul, South Korea

25 May 2024

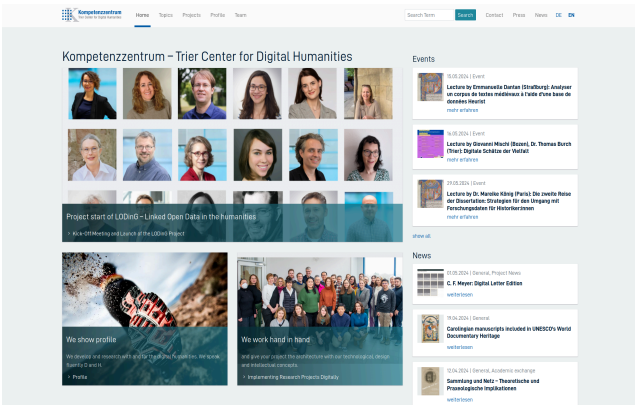
Kompetenzzentrum
Trier Center for Digital Humanities

Introduction

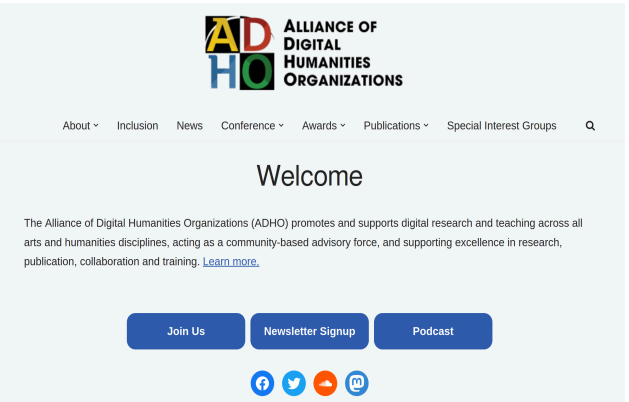
Multilingualism and me



Trier University



Trier Center for Digital Humanities



Alliance of Digital Humanities Organizations

Computational Comparative Literary Studies?

- What could CCLS be?
  - Literary Studies
  - but: Computational (using digital data and methods)
  - and: Comparative (transnational, transmedial)
- Many challenges for conversion
  - requires multiple areas of expertise
  - significant challenges of multilingualism

Three attempts at CCLS

1. Corpus Building: The Diversity Paradox
2. Data Modeling: Linked Open Data
3. Text Analysis: Multilingual Stylometry

(1) Corpus Building: The Diversity Paradox

See: distant-reading.net

The COST Action ‘Distant Reading for European Literary History’

Home > Browse Actions > Distant Reading for European Literary History (DISTANT-READING)

Description | Management Committee | Main Contacts and Leadership | Working Groups and Membership

Description

This Action's challenge is to create a vibrant and diverse network of researchers jointly developing the resources and methods necessary to change the way European literary history is written. Grounded in the Distant Reading paradigm (i.e. using computational methods of analysis for large collections of literary texts), the Action will create a shared theoretical and practical framework to enable innovative, sophisticated, data-driven, computational methods of literary text analysis across at least 10 European languages. Fostering insight into cross-national, large-scale patterns and evolutions across European literary traditions, the Action will facilitate the creation of a broader, more inclusive and better-grounded account of European literary history and cultural identity. To accomplish this, the Action will:

1. build a multilingual European Literary Text Collection (ELTeC), ultimately containing around 2,500 full-text novels in at least 10 different languages, permitting to test methods and compare results across national traditions;
2. establish and share best practices and develop innovative computational methods of text analysis adapted to Europe's multilingual literary traditions;
3. consider the consequences of such resources and methods for rethinking fundamental concepts in literary theory and history.

Action Details

- MoU - 015/17
- CSO Approval date - 23/06/2017
- Start date - 03/11/2017
- End date - 30/04/2022
- Former end date - 02/11/2021

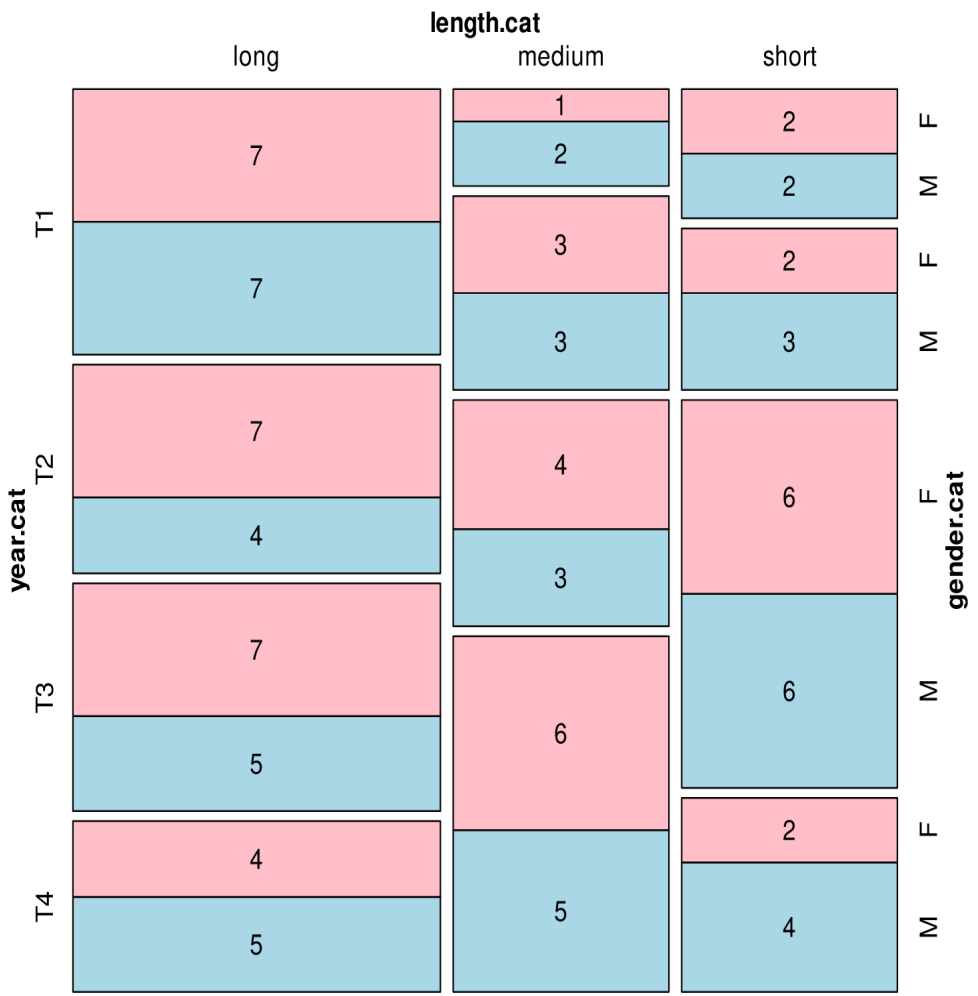
https://www.distant-reading.net/

The ‘European Literary Text Collection’ (ELTeC)

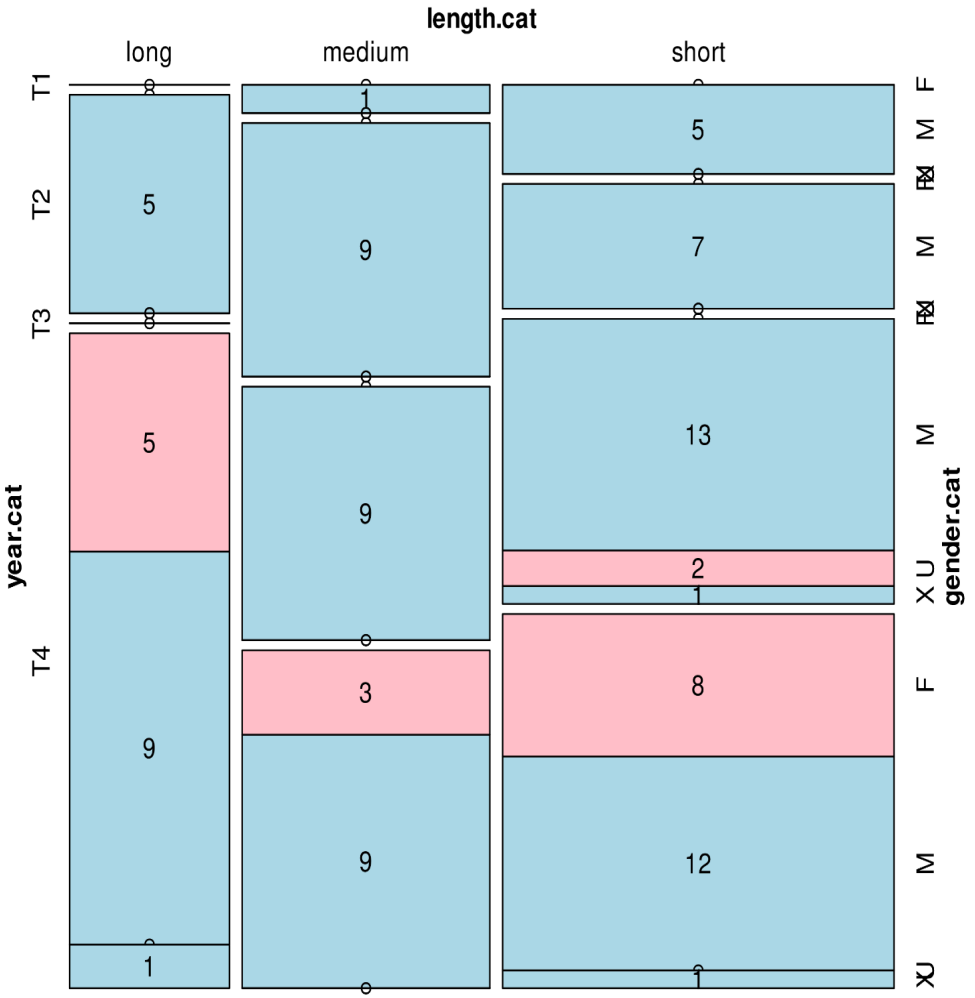
ELTeC-core				AUTHORSHIP				LENGTH			TIME SLOT					REPRINT COUNT		
Language	Last update	Texts	Words	Male	Female	1-title	3-title	Short	Medium	Long	1840-59	1860-79	1880-99	1900-20	range	Frequent	Rare	E5C
cze	2021-04-09	100	5621667	88	12	62	6	43	49	8	12	21	39	28	27	1	19	80.00
deu	2022-04-19	100	12738842	67	33	35	9	20	37	43	25	25	25	25	0	48	46	96.92
eng	2022-11-17	100	12227703	49	51	70	10	27	27	46	21	22	31	26	10	32	68	100.00
fra	2022-01-24	100	8712219	66	34	58	10	32	38	30	25	25	25	25	0	44	56	101.54
gsw	2023-03-30	100	6408326	73	27	32	9	45	40	15	6	16	19	59	53	0	0	66.15
hun	2022-01-24	100	6948590	79	21	71	9	47	31	22	22	21	27	30	9	32	67	100.00
pol	2022-06-01	100	8500172	58	42	1	33	33	35	32	8	11	35	46	38	39	61	80.00
por	2022-03-15	100	6799385	83	17	73	9	40	41	19	13	37	19	31	24	26	60	94.62
rom	2022-05-31	100	5951910	79	16	59	9	49	31	20	6	21	25	48	42	24	76	83.08
slv	2022-02-02	100	5682120	89	11	26	5	53	39	8	2	13	36	49	47	48	52	78.46
spa	2022-05-16	100	8737928	78	22	46	10	34	35	31	23	22	29	26	7	46	54	100.00
srp	2022-03-17	100	4931503	92	8	48	11	55	39	6	2	18	40	40	38	38	62	80.77

ELTeC-plus				AUTHORSHIP				LENGTH			TIME SLOT					REPRINT COUNT		
Language	Last update	Texts	Words	Male	Female	1-title	3-title	Short	Medium	Long	1840-59	1860-79	1880-99	1900-20	range	Frequent	Rare	E5C
gle	2022-04-08	1	24471	1	0	1	0	1	0	0	0	0	0	1	1	0	1	1.54
gre	2022-01-24	17	98607	13	4	14	1	17	0	0	0	2	8	7	8	4	7	52.31
hrv	2022-01-26	21	1440018	21	0	4	0	6	12	3	6	12	2	1	11	1	0	23.08
ita	2022-05-05	70	5535905	59	11	29	5	26	30	14	8	18	21	23	15	39	4	70.77
lit	2022-05-25	32	947634	25	7	18	1	24	3	5	6	4	6	16	12	9	23	60.00
lav	2022-04-28	31	2553907	27	4	14	1	10	14	7	0	2	6	23	23	4	26	52.31
nor	2022-11-12	58	3686837	40	18	22	12	28	19	11	5	3	32	18	29	32	26	70.77
swe	2021-04-11	58	4960085	29	28	18	8	16	24	18	15	3	20	20	17	17	41	76.92
ukr	2021-04-09	50	1840062	37	13	23	7	34	13	3	5	10	11	24	19	30	20	70.77

A closer look: corpus composition in ELTeC



English ELTeC corpus



Romanian ELTeC corpus

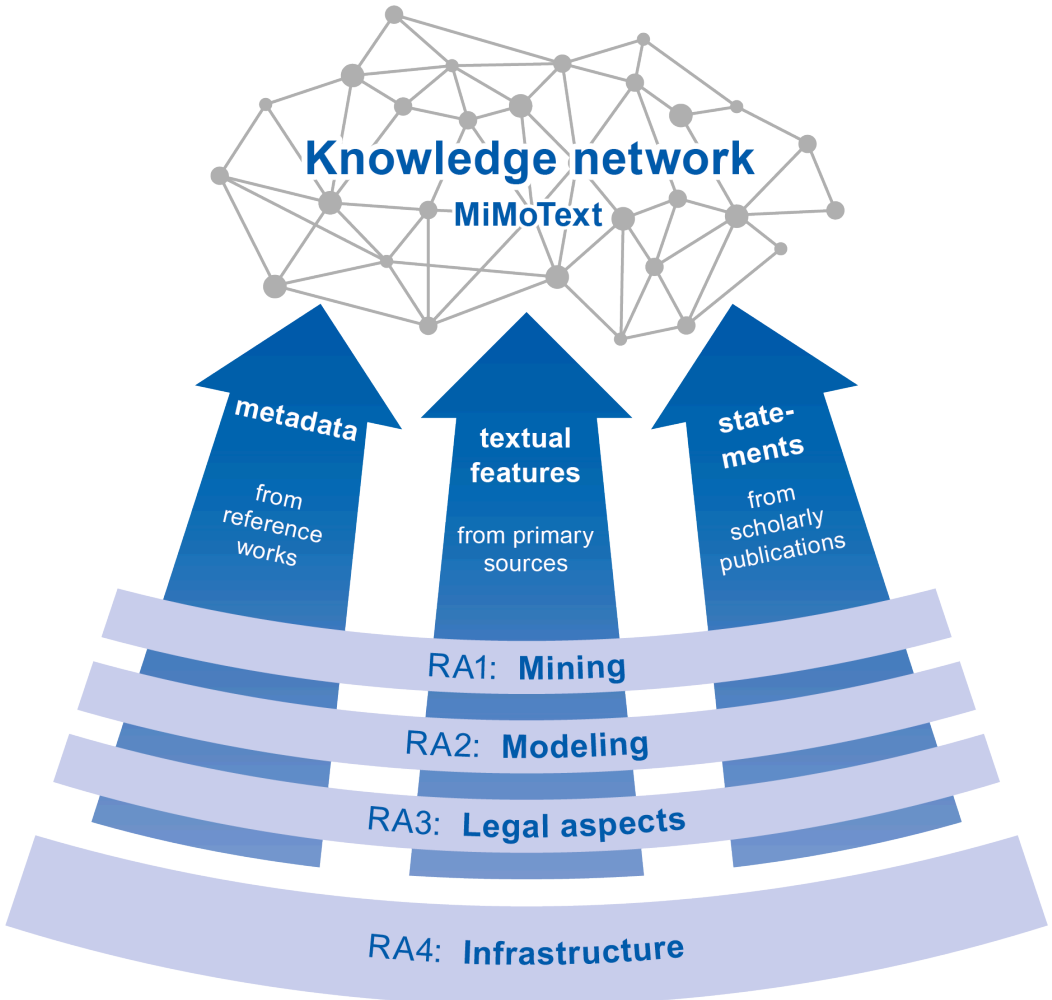
The Diversity Paradox

- ELTeC design goals: enable meaningful cross-language investigations
  - Balance with respect to key text characteristics
    (text length, author gender, prestige)
  - Inclusivity with respect to language-based literary traditions
- Consequence: the ‘diversity paradox’
  - If the criteria are too loose, balance is compromimsed
    (many, but invalid, corpora)
  - If the criteria are too strict, inclusivity is compromised
    (valid, but few, corpora)
  - In both cases, meaningful cross-language investigations are impossible

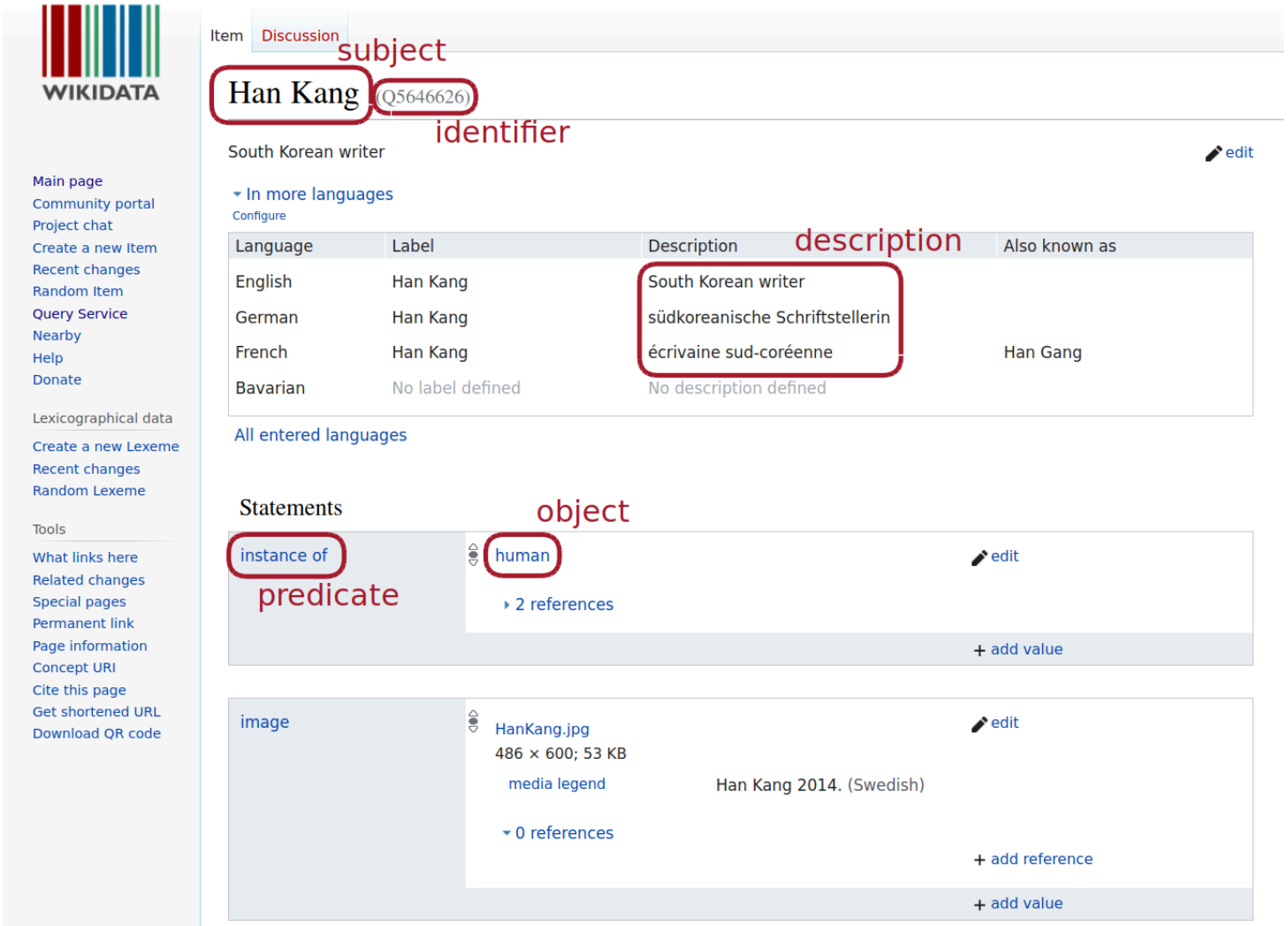
(2) Data Modeling: Linked Open Data

See: mimotext.uni-trier.de/en

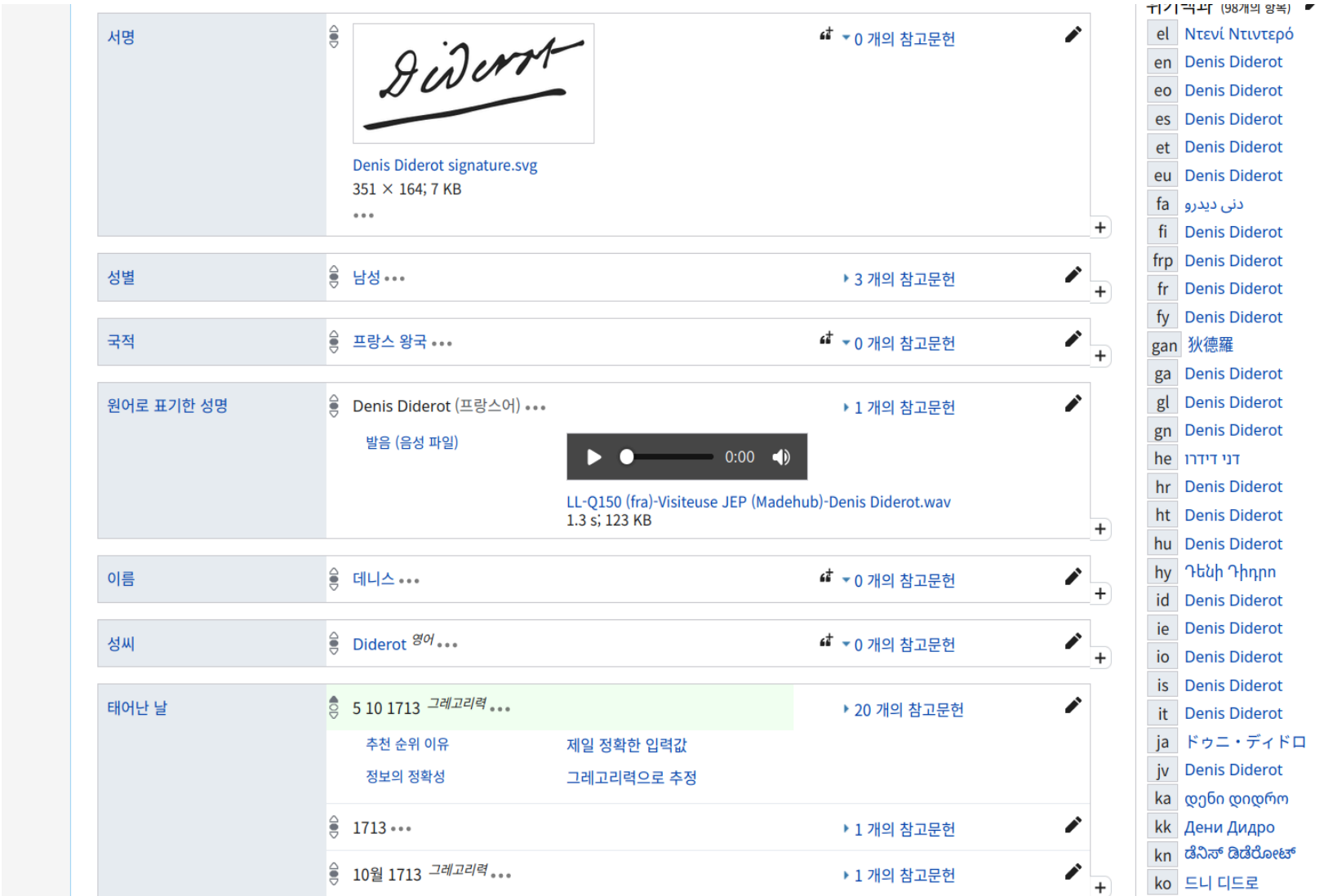
The project ‘Mining and Modeling Text’



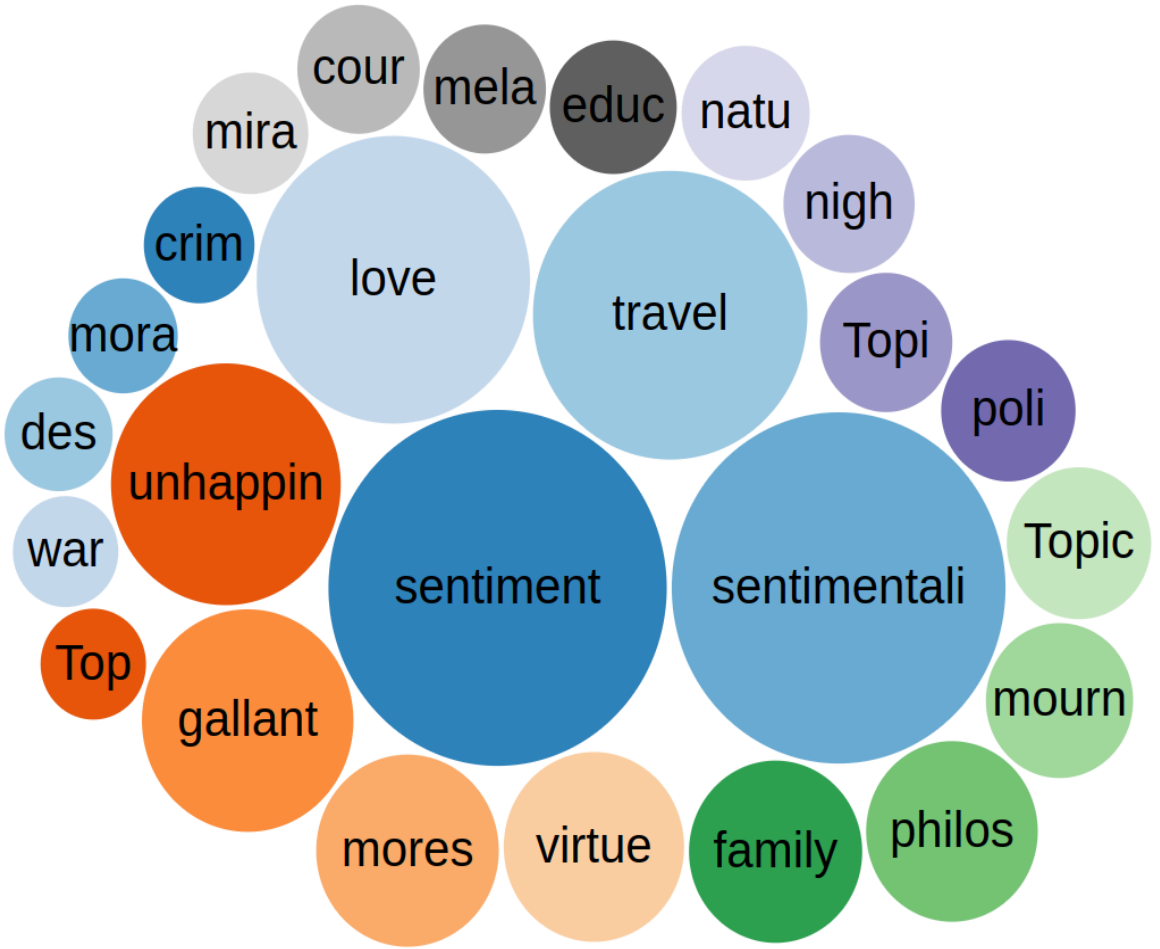
Linked Open Data: Simple Statements



Linked Open Data: Multilingualism



MiMoText Base: Query for themes in novels

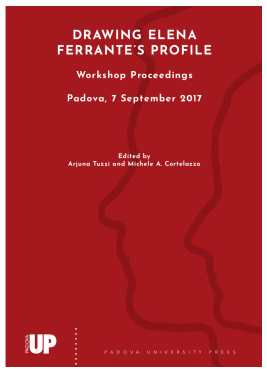


(3) Text Analysis: Multilingual Stylometry

See: showcases.clsinfra.io

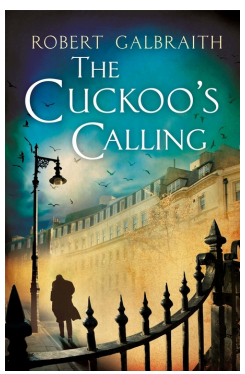
High-profile cases of stylometric authorship attribution

William Shakespeare:
Craig and Kinney (2009)



Elena Ferrante:
Tuzzi and Cortelazzo (2018)

Molière and Corneille:
Cafiero and Camps (2019)



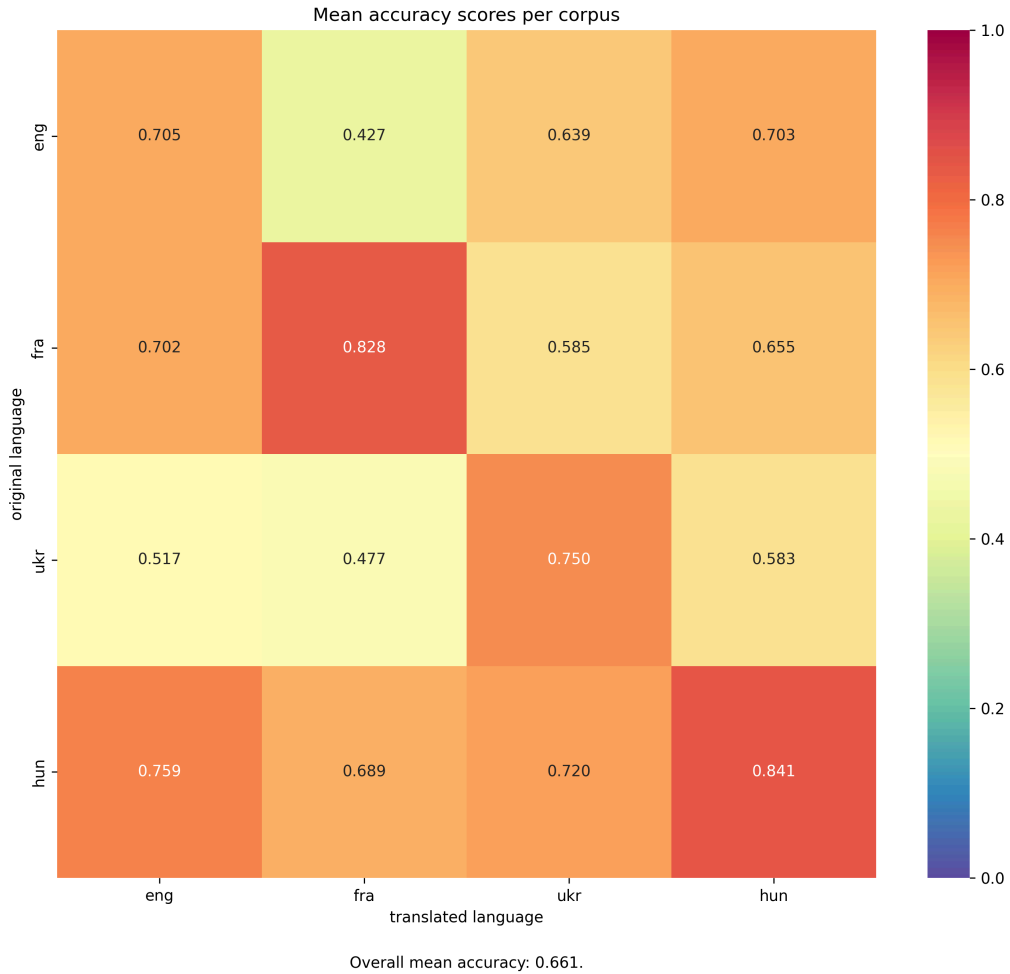
Galbraith / Rowling:
Juola (2015)

Multilingual stylometry?

translation ↦ ↧ original	fra	eng	hun	ukr
fra	fra-fra	fra-eng	fra-hun	fra-ukr
eng	eng-fra	eng-eng	eng-hun	eng-ukr
hun	hun-fra	hun-eng	hun-hun	hun-ukr
ukr	ukr-fra	ukr-eng	ukr-hun	ukr-ukr

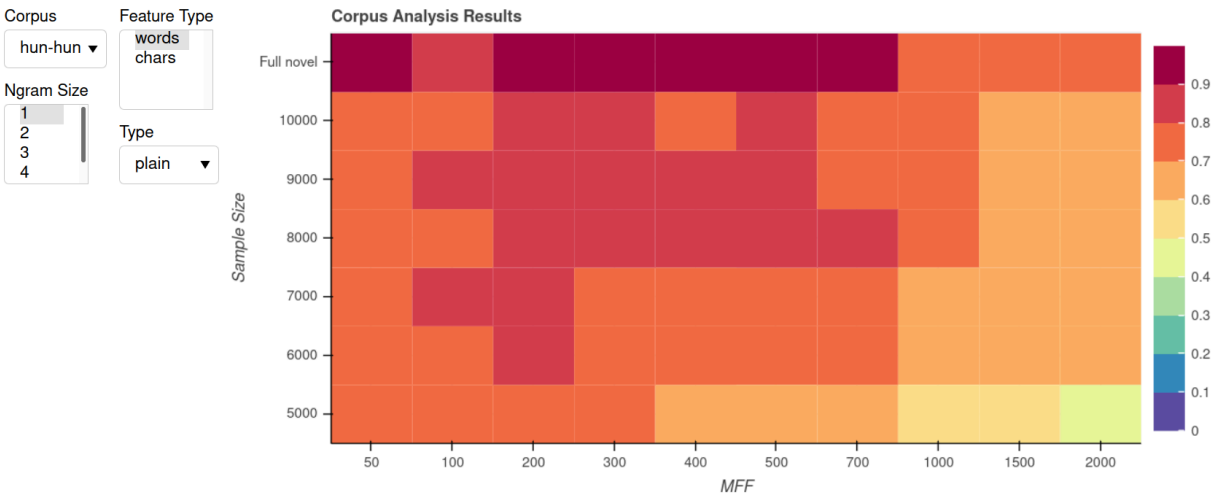
- Using corpora from the European Literary Text Collection (ELTeC)
- Translated entirely into the other languages using DeepL

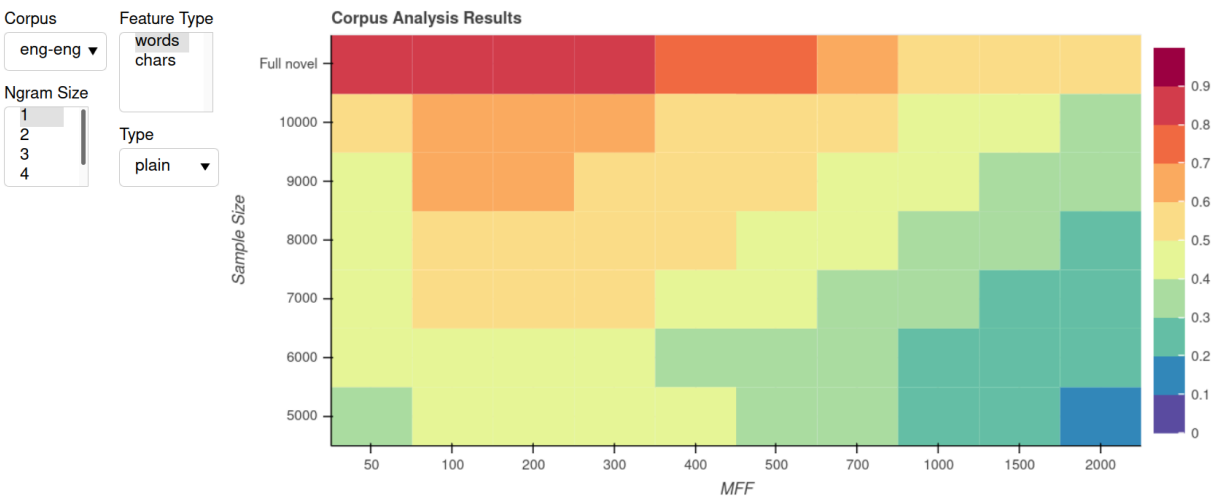
Some first results



More information: Dudar et al. (in progress).

Full interactive showcase





Conclusion

Take-home message

- Good, multilingual corpora are rare and hard to build
- Linked Open Data is a huge opportunity for multilingual data modeling
- Text analysis is still primarily multi-lingual rather than cross-lingual (but multilingual LLMs are in the process of changing that)

Lessons learned

- Multilingual research is multicultural research
- ‘Computational Comparative Literary Studies’ requires multiple competencies
- Nobody can learn everything: we need interdisciplinary collaboration
- Let’s learn from each other: Computational and Comparative Literary Studies

Thank you for your kind attention!

References

Cafiero, Florian, and Jean-Baptiste Camps. 2019. “Why Molière Most Likely Did Write His Plays.” *Science Advances* 5 (11): eaax5489. https://doi.org/10.1126/sciadv.aax5489.

Craig, Hugh, and Arthur F. Kinney. 2009. *Shakespeare, Computers, and the Mystery of Authorship*. Cambridge University Press.

Dudar, Julia, Evgeniia Fileva, Artjoms Šeļa, and Christof Schöch. in progress. “Multilingual Stylometry: The Influence of Corpus Composition and Language on the Performance of Authorship Attribution Using Corpora from the European Literary Text Collection (ELTeC).” *Tbc*, in progress.

Juola, Patrick. 2015. “The Rowling Case: A Proposed Standard Protocol for Authorship Attribution.” *Digital Scholarship in the Humanities* 30 (suppl. 1): 100–113. https://doi.org/10.1093/llc/fqv040.

Schöch, Christof, Maria Hinzmann, Julia Röttgermann, Katharina Dietz, and Anne Klee. 2022. “Smart Modelling for Literary History.” *International Journal of Humanities and Arts Computing* 16 (1): 78–93. https://doi.org/10.3366/ijhac.2022.0278.

Schöch, Christof, Roxana Patras, Tomaž Erjavec, and Diana Santos. 2021. “Creating the European Literary Text Collection (ELTeC): Challenges and Perspectives.” *Modern Languages Open*, no. 1: 25. https://doi.org/10.3828/mlo.v0i0.364.

Tuzzi, Arjuna, and Michele A. Cortelazzo, eds. 2018. *Drawing Elena Ferrante’s Profile: Workshop Proceedings*. Padova: Padova UP.