Distant Reading for European Literary History. A COST Action

Distant Reading

Christof Schöch (Trier, Germany)
Vilnius Lectures, Sept. 2020

website: http://www.distant-reading.net
slides: https://dh-trier.github.io/talks

cost
EUROPEAN COOPERATION
IN SCIENCE & TECHNOLOGY

Funded by the Horizon 2020 Framework Programme
of the European Union

Overview

1. What are COST Actions?
2. What is DISTANT READING about?
3. Text Collection Building
4. Methods and Tools
5. Literary History and Theory
6. Conclusions

1. What are COST Actions?

Networking Grants

- Support network and capacity building around specific research topics
- Networks of European countries and international partners
- Around 300 Actions are running in parallel (very few of them in the Humanities)

Some Actions relevant to DH

- Interedition, 2008-2012
- European Network of e-Lexicography, 2013-2017
- Reassembling the Republic of Letters, 2015-2018
- NEP4DISSENT - Cultures of Dissent, 2017-2021
- Nexus Linguarum, 2019-2023

Some key features

- Action duration is usually 4 years
- Structured into working groups
- No funding for staff, only for networking activities
- Various forms of "networking activities"

COST networking activities

- Working Group Meetings
- Training Schools
- Short Term Scientific Missions
- Conference Grants

2. What is *Distant Reading for European Literary History* about?

The term ”Distant Reading”

- Term first used by Franco Moretti ("second-hand" reading)
- Has then broadened in meaning to mean any computational analysis of literary texts
- Narrower term: stylometry, in the sense of quantitative methods for authorship attribution
- New term: Computational Literary Studies

Action: Research objectives

- Resources: Build a multilingual reference collection of European novels ("ELTeC")
- Methods: Explore, evaluate, adapt and share computational methods of text analysis for ELTeC
- Theory: Think through consequences of digital data and methods for literary history and theory

Action: Networking objectives

- Bring together corpus linguists, computational linguists, digital literary scholars, literary historians and theorists
- Spread and share competencies in the three areas above among these groups
- Support relevant collaborative grant proposals on the national and European levels

Current Network

- 32 countries are involved
- 200+ scholars are participating
- 4 Working Groups
- several spin-off projects

3. Text Collection Building

European Literary Text Collection (ELTeC)

- Comparable sets of novels for at least 10 European languages
- Each set: 100 novels published between 1840 and 1920
- Extensions (chronologically or simply additional texts)
- WG leads: Carolin Odebrecht (DE), Lou Burnard (UK), Borja Navarro Colorado (SP), Martina Scholger (AT)
- Currently: more than 1000 novels published
- More information: https://distantreading.github.io/ELTeC/

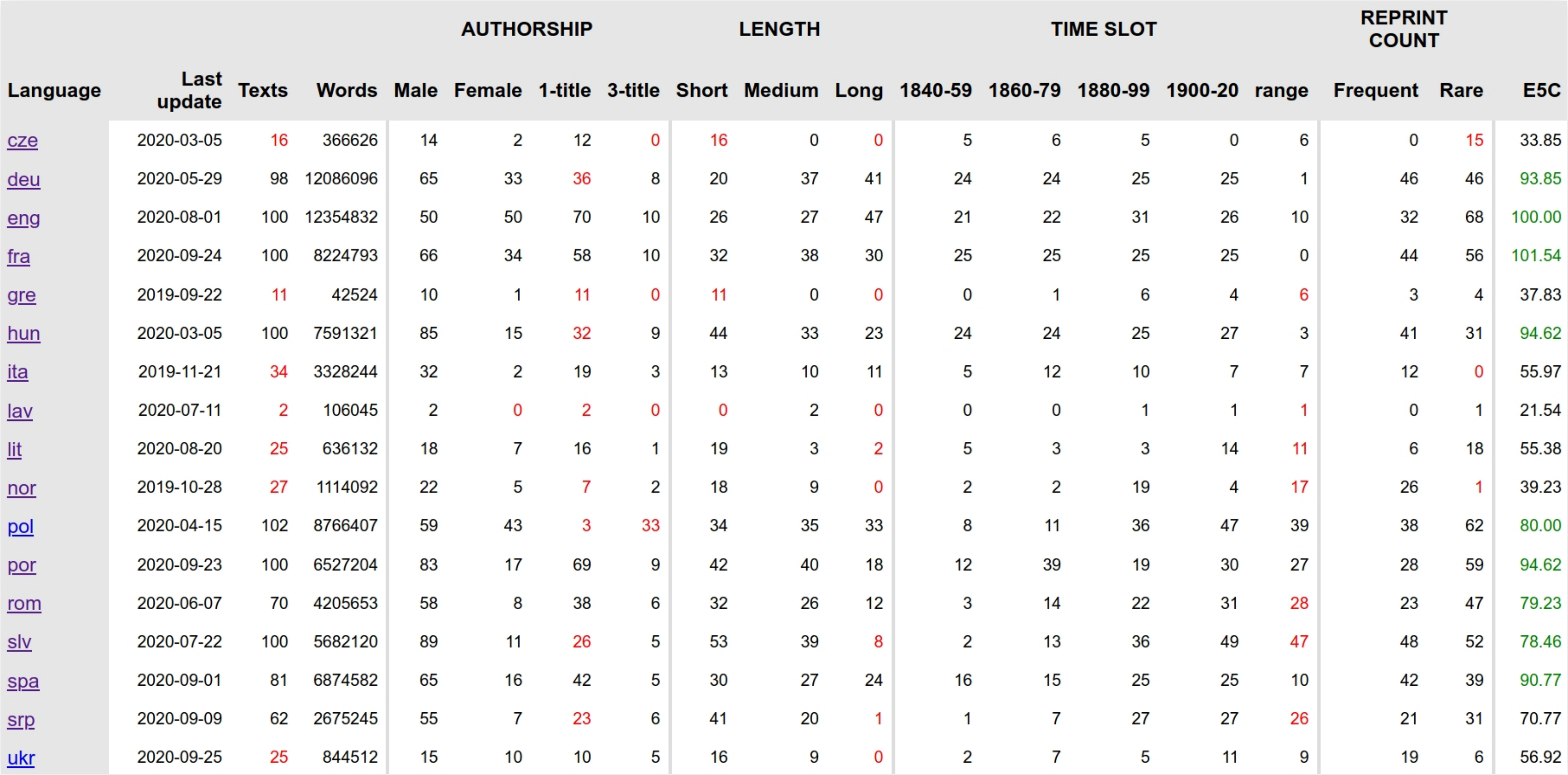
Collection building: text selection

- (1) Eligibility: In order to be included, a text must…
  - have been first published as a book between 1850 and 1919
  - have been published in a European country within a decade from their first publication
  - be a novel, i.e. a fictional prose narrative of a minimum length of 10,000 words
  - have originally been written in the language of the given subcollection

Collection building: text selection

- (2) Composition: Among the novels in each language subcollection…
  - at least 10% (ideally more) have been written by female authors
  - at least 30% are rarely reprinted novels
  - at least 20% are short (10-50k words), at least 20% are long (>100k words) novels
  - 9-11 authors are represented with three novels
  - the novels should be spread out evenly across the entire period

ELTeC Overview



				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	2020-03-05	16	366626	14	2	12	0	16	0	0	5	6	5	0	6	0	15	33.85
deu	2020-05-29	98	12086096	65	33	36	8	20	37	41	24	24	25	25	1	46	46	93.85
eng	2020-08-01	100	12354832	50	50	70	10	26	27	47	21	22	31	26	10	32	68	100.00
fra	2020-09-24	100	8224793	66	34	58	10	32	38	30	25	25	25	25	0	44	56	101.54
gre	2019-09-22	11	42524	10	1	11	0	11	0	0	0	1	6	4	6	3	4	37.83
hun	2020-03-05	100	7591321	85	15	32	9	44	33	23	24	24	25	27	3	41	31	94.62
ita	2019-11-21	34	3328244	32	2	19	3	13	10	11	5	12	10	7	7	12	0	55.97
lav	2020-07-11	2	106045	2	0	2	0	0	2	0	0	0	1	1	1	0	1	21.54
lit	2020-08-20	25	636132	18	7	16	1	19	3	2	5	3	3	14	11	6	18	55.38
nor	2019-10-28	27	1114092	22	5	7	2	18	9	0	2	2	19	4	17	26	1	39.23
pol	2020-04-15	102	8766407	59	43	3	33	34	35	33	8	11	36	47	39	38	62	80.00
por	2020-09-23	100	6527204	83	17	69	9	42	40	18	12	39	19	30	27	28	59	94.62
rom	2020-06-07	70	4205653	58	8	38	6	32	26	12	3	14	22	31	28	23	47	79.23
slv	2020-07-22	100	5682120	89	11	26	5	53	39	8	2	13	36	49	47	48	52	78.46
spa	2020-09-01	81	6874582	65	16	42	5	30	27	24	16	15	25	25	10	42	39	90.77
srp	2020-09-09	62	2675245	55	7	23	6	41	20	1	1	7	27	27	26	21	31	70.77
ukr	2020-09-25	25	844512	15	10	10	5	16	9	0	2	7	5	11	9	19	6	56.92

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

Collection building: Text encoding

- All texts are encoded in XML-TEI (Text Encoding Initiative)
- There are three levels of encoding, with increasingly detailed markup
  - level 0: very simple / minimal markup
  - level 1: richer, more 'semantic' markup
  - level 2: text with linguistic annotation

Some challenges

- Different states of digitization in various formats (e.g.: French vs. Romanian)
- Most metadata relevant to composition is not included in catalogs; e.g. novel type, author gender
- Varying writing systems used (e.g.: Romanian 'transition alphabet')
- Varying traditions of novel length (e.g.: few Slovenian 'long' novels)
- corpus composition criteria as a double-edged sword (under-represented categories in under-represented literary traditions)
- need for stability vs. adjustment of composition and encoding guidelines

4. Methods and Tools

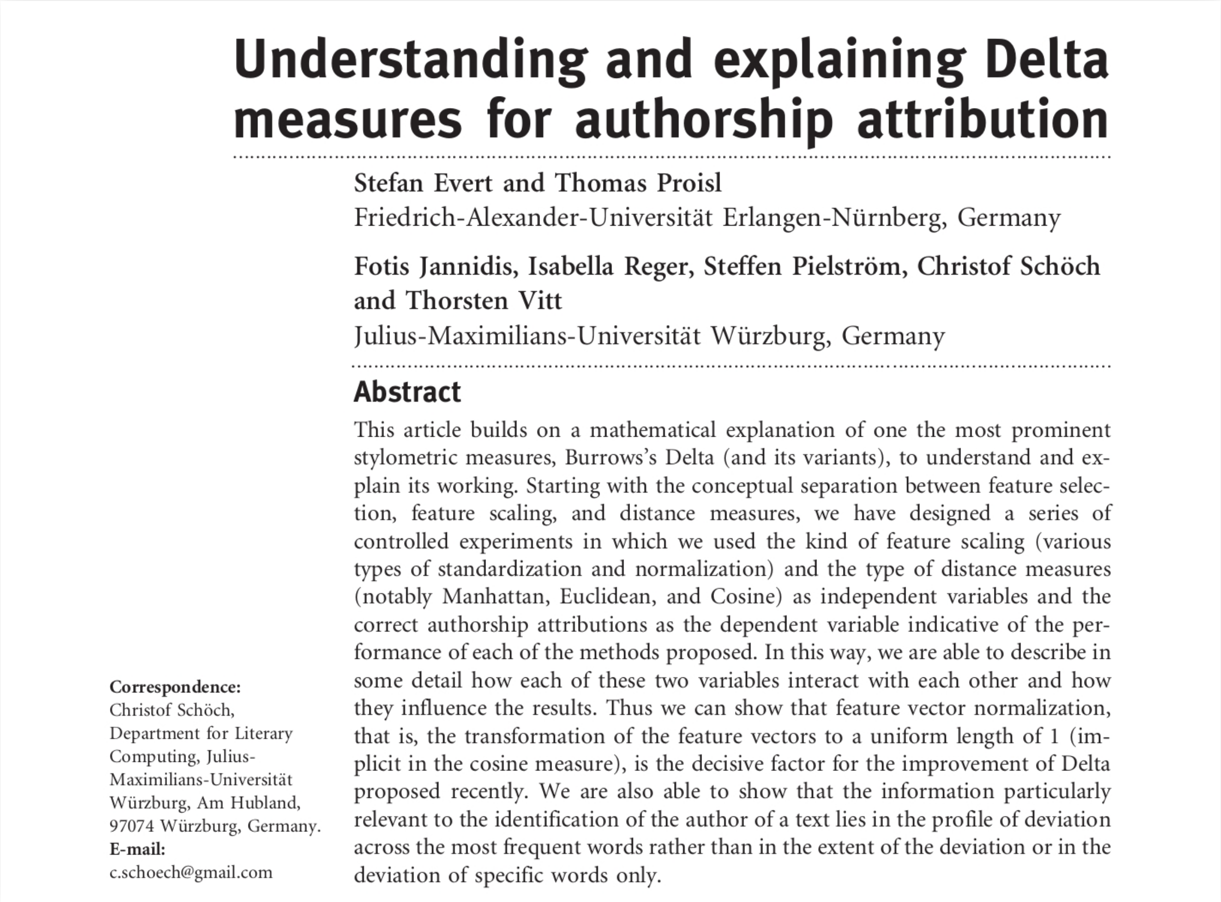
Objectives

- Adapt Distant Reading methods to multiple European languages
- Develop cross-linguistic use of Distant Reading methods
- Spread Distant Reading competencies across Europe (ECI / ITC)
- WG leads: Joanna Byszuk (PL), George Mikros (GR), Fotis Jannidis (DE), Yaakov HaCohen-Kerner (ISR)

Example for Adaptation: Stylometry

- Computational Authorship Attribution
- Based on the word frequency profiles of texts
- Multiple small differences are decisive (Burrows)
- Applications
  - Elena Ferrante / Domenico Starnone + Anita Raja
  - Robert Galbraith (Joan K. Rowlings): The Cuckoo's Calling
  - Corneille and Molière
  - many more

Example for Adaptation: Stylometry



Understanding and explaining Delta measures for authorship attribution

Stefan Evert and Thomas Proisl
Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany

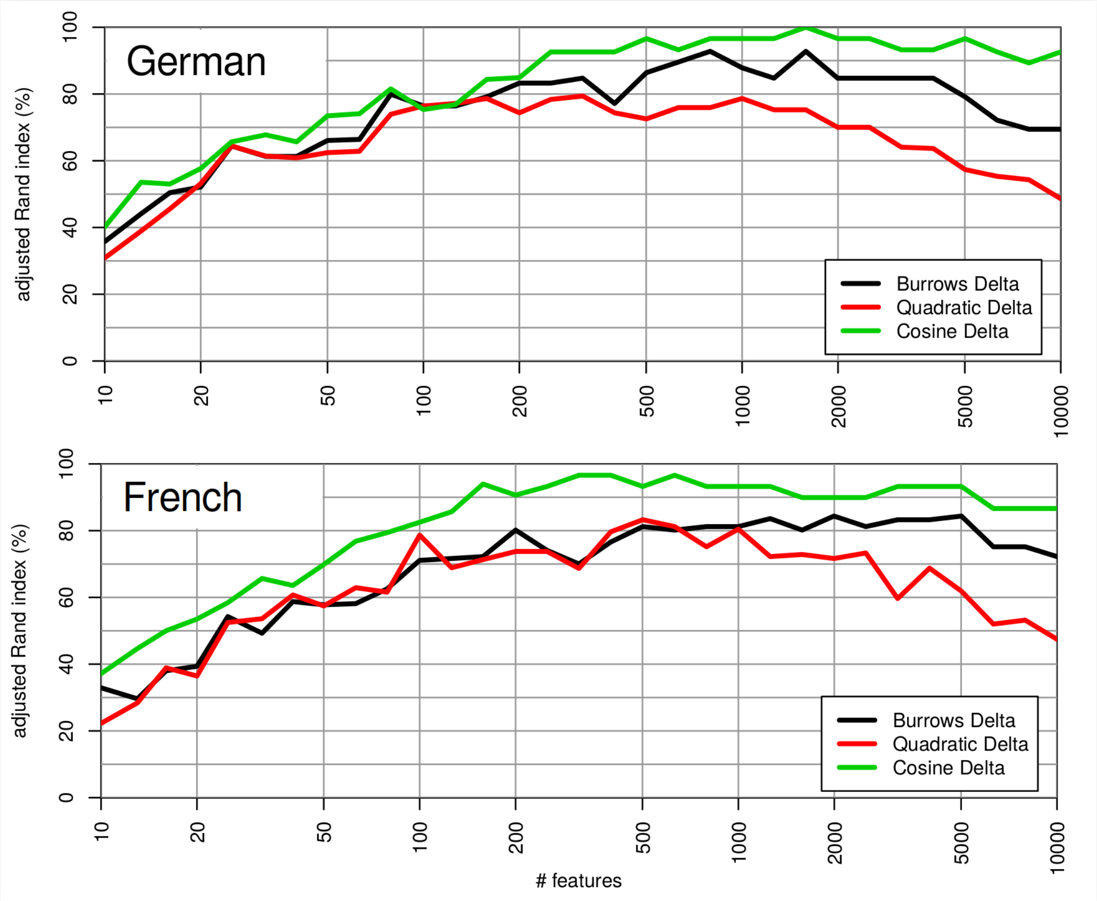
Fotis Jannidis, Isabella Reger, Steffen Pielström, Christof Schöch and Thorsten Vitt
Julius-Maximilians-Universität Würzburg, Germany

Abstract

This article builds on a mathematical explanation of one the most prominent stylometric measures, Burrows's Delta (and its variants), to understand and explain its working. Starting with the conceptual separation between feature selection, feature scaling, and distance measures, we have designed a series of controlled experiments in which we used the kind of feature scaling (various types of standardization and normalization) and the type of distance measures (notably Manhattan, Euclidean, and Cosine) as independent variables and the correct authorship attributions as the dependent variable indicative of the performance of each of the methods proposed. In this way, we are able to describe in some detail how each of these two variables interact with each other and how they influence the results. Thus we can show that feature vector normalization, that is, the transformation of the feature vectors to a uniform length of 1 (implicit in the cosine measure), is the decisive factor for the improvement of Delta proposed recently. We are also able to show that the information particularly relevant to the identification of the author of a text lies in the profile of deviation across the most frequent words rather than in the extent of the deviation or in the deviation of specific words only.

Correspondence:
Christof Schöch,
Department for Literary Computing, Julius-Maximilians-Universität Würzburg, Am Hubland, 97074 Würzburg, Germany.
E-mail:
c.schoech@gmail.com

Example for Adaptation: Stylometry



Cross-Language Distant Reading: Direct Speech Recognition

Proceedings of 1st Workshop on Language Technologies for Historical and Ancient Languages, pages 100–104
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

Detecting Direct Speech in Multilingual Collection of 19th-century Novels

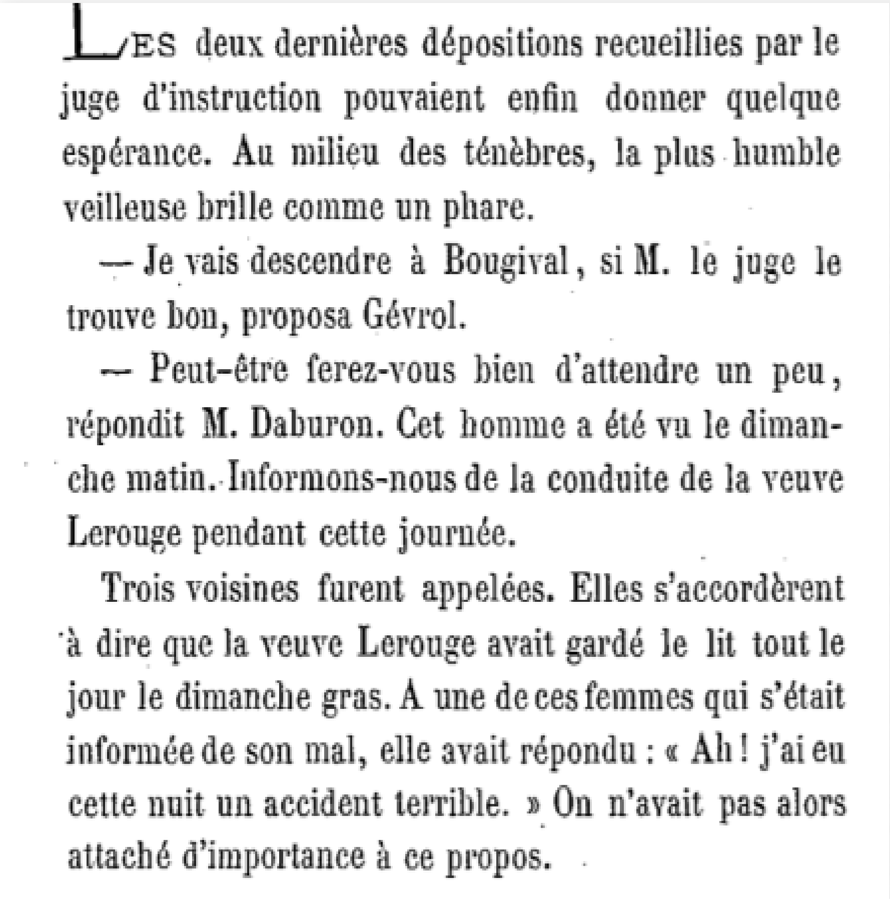
**Joanna Byszuk[1], Michał Woźniak[1], Mike Kestemont[2], Albert Leśniak[1],
Wojciech Łukasik[1], Artjoms Šeļa[1,3], Maciej Eder[1]**
[1]Institute of Polish Language, Polish Academy of Sciences; [2]University of Antwerp; [3]University of Tartu
Mickiewicza 31, 31120 Kraków, Poland; Prinsstraat 13, 2000 Antwerpen, Belgium; Ülikooli 18, 50090 Tartu, Estonia
{joanna.byszuk, michal.wozniak, albert.lesniak, wojciech.lukasik, artjoms.sela, maciej.eder}@ijp.pan.pl
mike.kestemont@uantwerp.be

Abstract

Fictional prose can be broadly divided into narrative and discursive forms with direct speech being central to any discourse representation (alongside indirect reported speech and free indirect discourse). This distinction is crucial in digital literary studies and enables interesting forms of narratological or stylistic analysis. The difficulty of automatically detecting direct speech, however, is currently under-estimated. Rule-based systems that work reasonably well for modern languages struggle with (the lack of) typographical conventions in 19th-century literature. While machine learning approaches to sequence modeling can be applied to solve the task, they typically face a severed skewness in the availability of training material, especially for lesser resourced languages. In this paper, we report the result of a multilingual approach to direct speech detection in a diverse corpus of 19th-century fiction in 9 European languages. The proposed method fine-tunes a transformer architecture with multilingual sentence embedder on a minimal amount of annotated training in each language, and improves performance across languages with ambiguous direct speech marking, in comparison to a carefully constructed regular expression baseline.

Keywords: direct speech recognition, multilingual, 19th century novels, deep learning, transformer, BERT, ELTeC

Cross-Language Distant Reading: Direct Speech Recognition

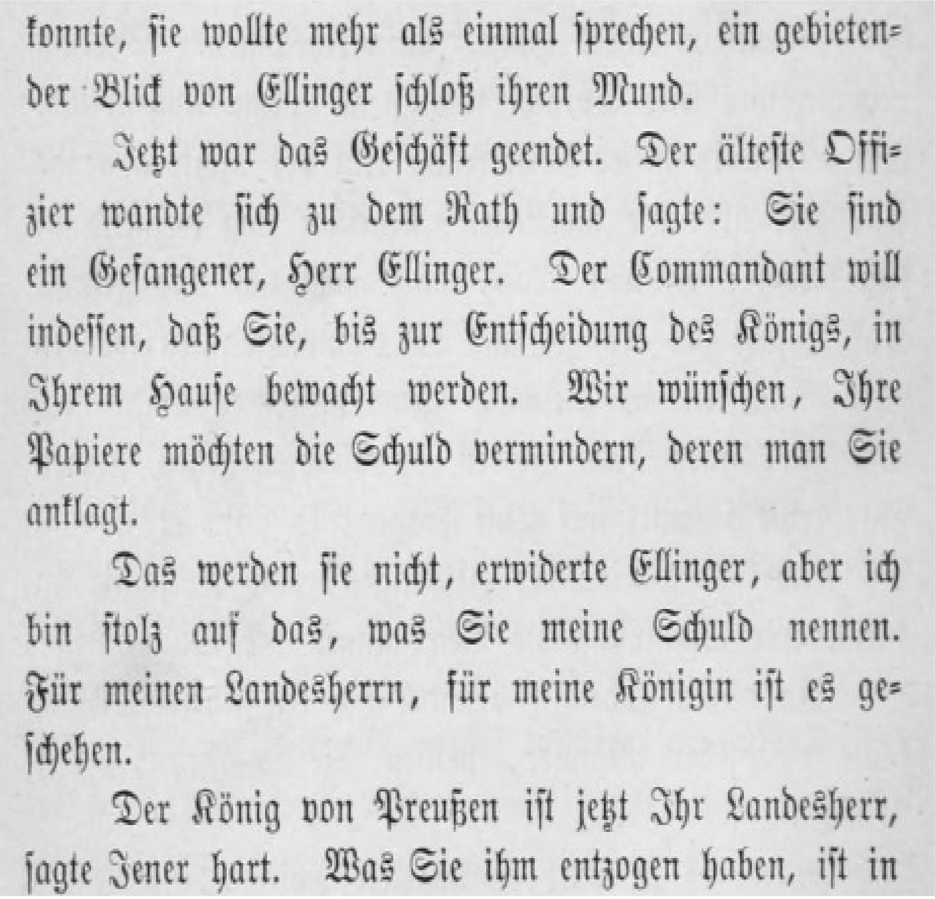


LES deux dernières dépositions recueillies par le juge d’instruction pouvaient enfin donner quelque espérance. Au milieu des ténèbres, la plus humble veilleuse brille comme un phare.

— Je vais descendre à Bougival, si M. le juge le trouve bon, proposa Gévrol.

— Peut-être ferez-vous bien d’attendre un peu, répondit M. Daburon. Cet homme a été vu le dimanche matin. Informons-nous de la conduite de la veuve Lerouge pendant cette journée.

Trois voisines furent appelées. Elles s’accordèrent à dire que la veuve Lerouge avait gardé le lit tout le jour le dimanche gras. A une de ces femmes qui s’était informée de son mal, elle avait répondu : « Ah ! j’ai eu cette nuit un accident terrible. » On n’avait pas alors attaché d’importance à ce propos.



konnte, ſie wollte mehr als einmal ſprechen, ein gebietender Blick von Ellinger ſchloß ihren Mund.

Jetzt war das Geſchäft geendet. Der älteſte Offizier wandte ſich zu dem Rath und ſagte: Sie ſind ein Gefangener, Herr Ellinger. Der Commandant will indeſſen, daß Sie, bis zur Entſcheidung des Königs, in Ihrem Hauſe bewacht werden. Wir wünſchen, Ihre Papiere möchten die Schuld vermindern, deren man Sie anklagt.

Das werden ſie nicht, erwiderte Ellinger, aber ich bin ſtolz auf das, was Sie meine Schuld nennen. Für meinen Landesherrn, für meine Königin iſt es geſchehen.

Der König von Preußen iſt jetzt Ihr Landesherr, ſagte Jener hart. Was Sie ihm entzogen haben, iſt in

5. Literary History and Theory

Objectives

- Think through consequences for literary history and theory
- Key concepts: style, genre, authorship, periodization, canonization, intertextuality, etc.
- WG leads: Antonija Primorac (HR), Rosario Arias (SP)

Example of canonization: concept

- Required reading in school or university
- Enduring interpretability
- Complexity; quality; importance
- Critical and/or commercial success
- Literary novels as non-genre novel

Example of canonization: indicators

- text-external
  - Reprint and sales figures
  - Library holding data
  - Number of entries in academic bibliographies
  - Amount of text in reference literary histories
  - Number of Wikipedias with an article
  - Number of mentions on reading lists
  - Literary prizes received
- text-internal
  - Lexical complexity / difficulty
  - Syntactic complexity / sentence length
  - Combination: Measures of readability
  - Complexity of plot, characters, meaning

Time for questions!

To learn more: websites

- http://distant-reading.net/
- https://github.com/distantreading
- http://www.cost.eu/COST_Actions/ca/CA16204
- https://twitter.com/DistantReading

To learn more: readings

- Burnard, Lou, Schoöch, and Odebrecht. “In Search of Comity: TEI for Distant Reading.” November 1, 2019. https://doi.org/10.5281/zenodo.3552489.
- Byszuk, Joanna, Michał Woźniak, Mike Kestemont, Albert Leśniak, Wojciech Łukasik, Artjoms Šeļa, and Maciej Eder. “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, edited by Rachele Sprungoli and Marco Passarotti, https://lrec2020.lrec-conf.org/media/proceedings/Workshops/Books/LT4HALAbook.pdf:100–104. Paris: European Language Resources Association (ELRA), 2020.
- Cinková, Silvie, et al. “Evaluation of Taggers for 19th-Century Fiction.” DH_Budapest_2019, edited by Gábor Pálko, ELTE, 2019, http://elte-dh.hu/dh_budapest_2019-abstract-booklet/.
- Evert, Stefan, Fotis Jannidis, Thomas Proisl, Steffen Pielström, Thorsten Vitt, Christof Schöch, and Isabella Reger. “Understanding and Explaining Distance Measures for Authorship Attribution.” Digital Scholarship in the Humanities, 2017. https://academic.oup.com/dsh/article-pdf/32/suppl_2/ii4/21298943/fqx023.pdf.
- Patras, Roxana, Ioana Galleron, Camelia GRĂDINARU, Ioana Lionte, and Lucreţia Pascaru. “The Splendors and Mist(Eries) of Romanian Digital Literary Studies: A State-of-the-Art Just before Horizons 2020 Closes Off.” Hermeneia 23 (2019): 207–22.
- Stankovic, Ranka, Francesca Frontini, Tomaž Erjavec, and Carmen Brando. “Named Entity Recognition for Distant Reading in Several European Literatures.” In DH_Budapest_2019, edited by Gábor Pálko. Budapest: ELTE, 2019. http://elte-dh.hu/dh_budapest_2019-abstract-booklet/.
- Schöch, Christof; Roxana Patras; Tomaž Erjavec; Diana Santos: "Creating the European Literary Text Collection (ELTeC): Challenges and Perspectives", submitted.

Danke! · Thank you! · Merci! · Ačiū!