

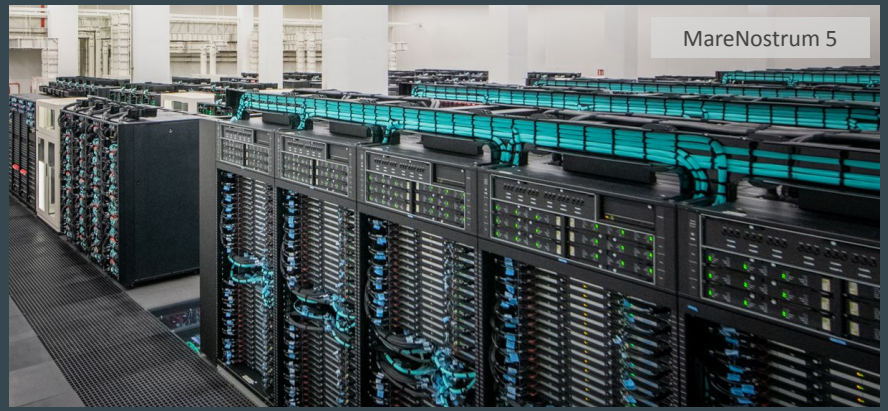
Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Computational Text Analysis for research in the Humanities and Social Sciences

Overview, Limits and Opportunities

Mariona Coll Ardanuy
Computational Social Sciences and Humanities Lab

Barcelona Supercomputing Center
Centro Nacional de Supercomputación





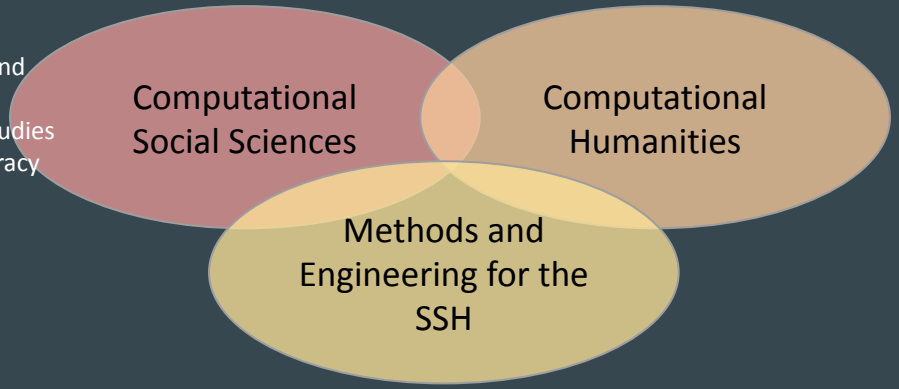
- Research in Computer Sciences, Life Sciences, Earth Sciences, Engineering
- Since 2024: Computational Social Sciences and Humanities (CSSH)

Computational Social Sciences and Humanities Lab (CSSH Lab)

Director: Mercè Crosas; Group Leaders: Caterina Calsamiglia, Hèctor Orengo, Clara Guasch

Understand current and past large-scale, complex questions in human societies, with the assistance of large datasets, Artificial Intelligence and High-Performing Computing.

- Public Policy Intelligence and Impact
- Science and Technology Studies
- Digital Society and Democracy



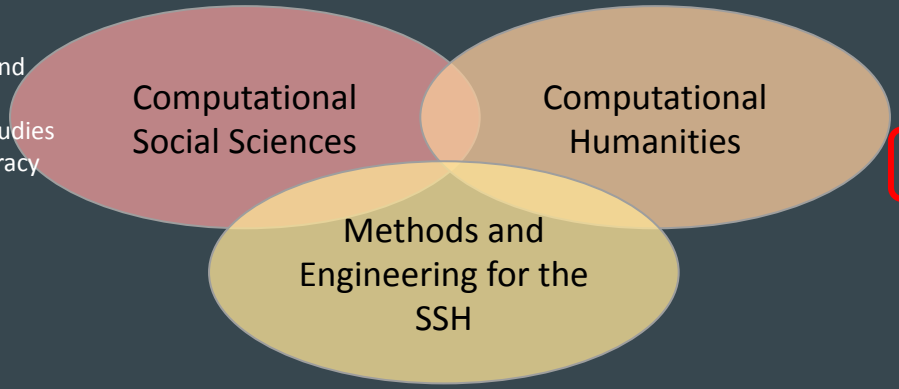
- Computational Archaeology
- Cultural Heritage: archives and museums
- Natural Language Processing for the Humanities

- FAIR data stewardship
- Metadata standards and sensitive data
- AI models, statistics & computational methods, workflows & tools

Computational Social Sciences and Humanities Lab (CSSH Lab)

Director: Mercè Crosas; Group Leaders: Caterina Calsamiglia, Hèctor Orengo, Clara Guasch

Understand current and past large-scale, complex questions in human societies, with the assistance of large datasets, Artificial Intelligence and High-Performing Computing.



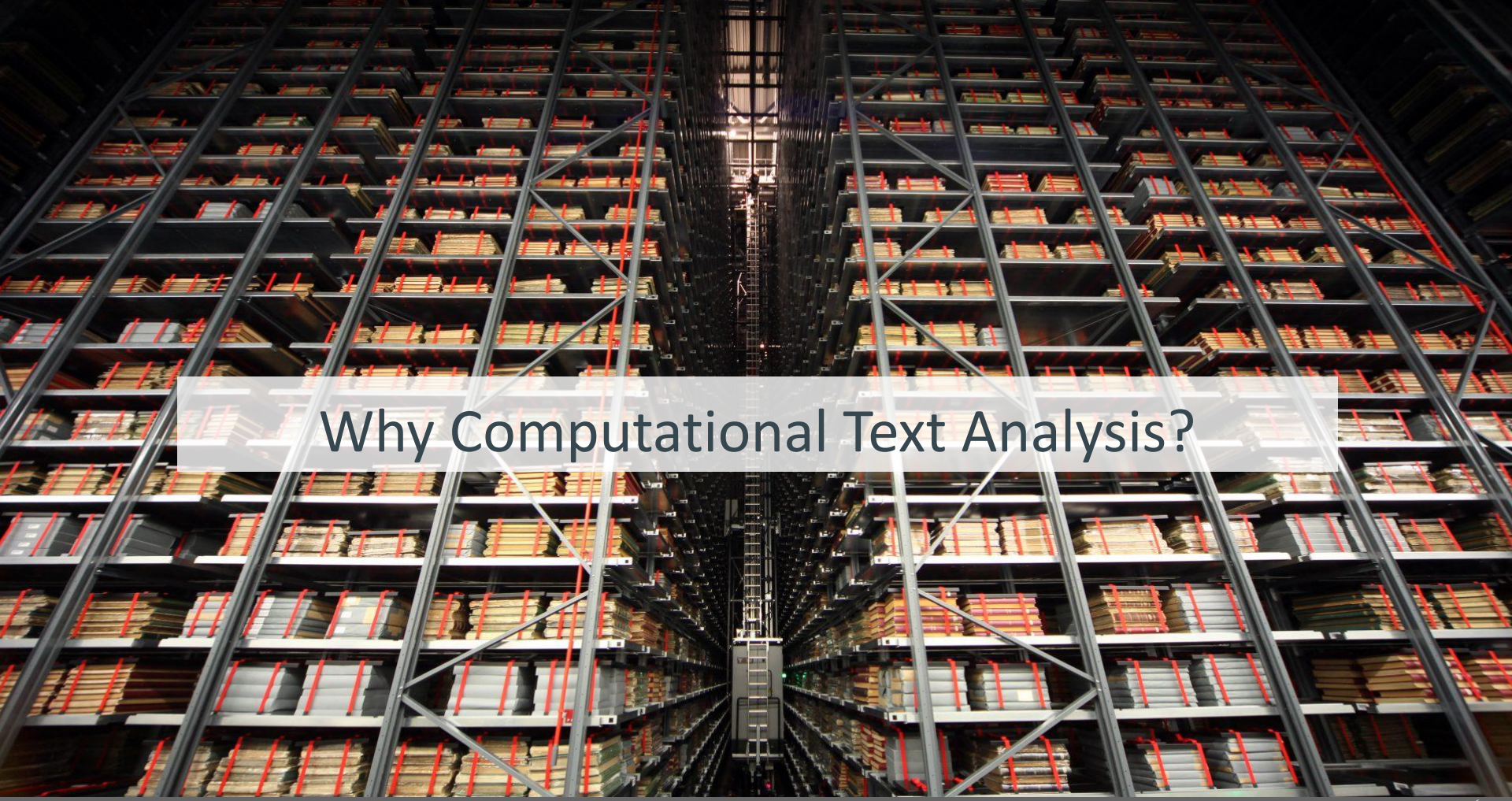
- Public Policy Intelligence and Impact
- Science and Technology Studies
- Digital Society and Democracy

- Computational Archaeology
- Cultural Heritage: archives and museums
- Natural Language Processing for the Humanities

- FAIR data stewardship
- Metadata standards and sensitive data
- AI models, statistics & computational methods, workflows & tools

Why Computational Text Analysis?

Why Computational Text Analysis?



British Newspaper Archive (45 km of shelves). Source: The British Library

Automating Text Analysis

- To be able to process large volumes of text data.
- To highlight patterns that are too subtle or wide-ranging to comprehend.
- To measure and study complex ideas or abstract concepts systematically.

Automating Text Analysis

- To be able to process large volumes of text data.
- To highlight patterns that are too subtle or wide-ranging to comprehend.
- To measure and study complex ideas or abstract concepts systematically.

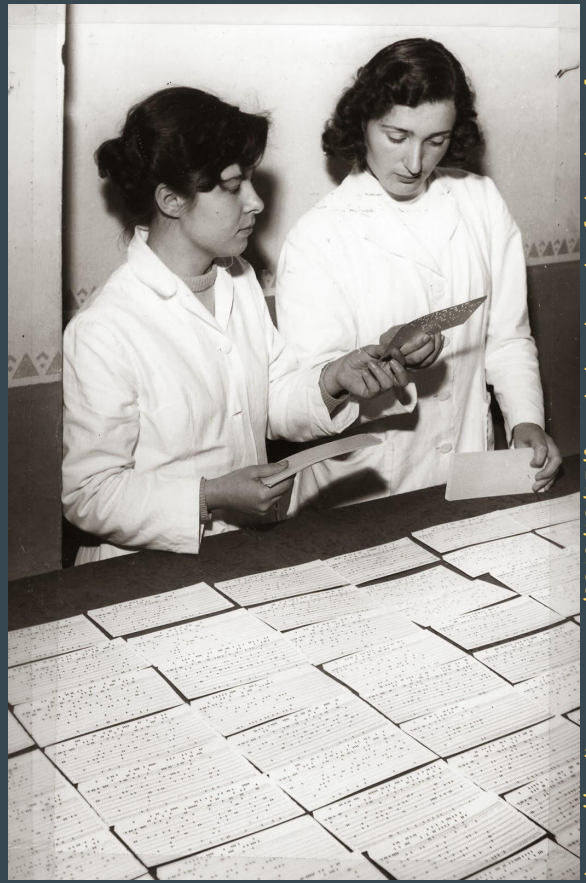
Text as Data (Grimmer & Stewart, 2013): Text as a convenient proxy for understanding human thought, behavior, and broader social, economic, and cultural trends or dynamics.

Automating Text Analysis

Traditional approaches (mostly until 2010s)

1. Preprocess: split into words: filter, normalize.
2. Count/compare: word frequencies, collocations, co-occurrences, saliency.
3. Model/visualize: concordances, topic modelling, clustering, classification, word clouds, networks.

"Language is complex. Not all of [its] complexity is necessary to effectively analyze texts." (Grimmer & Stewart, 2013)



https://melissaterras.org/2013/10/15/for-ada-lovelace-day-father-busas-female-punch-card-operatives/ Source: CIRCSE Research Centre.

Benefits

1. Goes beyond anecdotal evidence, enables patterns in data to surface
2. Precise estimates of typicality ("how common"), spots outliers
3. Provides a framework for qualitative analysis, clearly defines limits of research
4. Emphasises methodology, validation, reproducibility

Risks

1. Loss of context.
2. Society and human behaviour reduced to word frequencies.
3. Risk of appearing neutral and objective. Illusion of objective truth, hiding biases and subjective decisions.
4. Errors in design cascade to errors in interpretation.

From Word Counts to Representing Word Meaning

Representing language computationally:
The distributional hypothesis

"You shall know a word by the company it keeps."
(Firth, 1957)
Words that share similar contexts tend to have similar meanings.
(Harris, 1954)

What is '***tesgüino***'?

1. A bottle of *tesgüino* is on the table.
2. Everyone likes *tesgüino*.
3. *Tesgüino* makes you drunk.
4. We make *tesgüino* out of corn.

Example from: Lin, Dekang. "Automatic retrieval and clustering of similar words." In ACL-COLING, 1998

Representing language computationally:

The distributional hypothesis

	bottle	*food*	*like*	*drunk*	*corn*
tesgüino	*y*	*n*	*y*	*y*	*y*
wine	*y*	*n*	*y*	*y*	*n*
tortilla	*n*	*y*	*y*	*n*	*y*
cotton	*n*	*n*	*y*	*n*	*n*

tesgüino	[1, 0, 1, 1, 1, …]
wine	[1, 0, 1, 1, 0, …]
tortilla	[0, 1, 1, 0, 1, …]
cotton	[0, 0, 1, 0, 0, …]



Representing language computationally:
Towards Language Models (1/2)

2013: Static word embeddings (Word2Vec): we can “predict” the meanings of words from large amounts of text. Every word form is represented as a vector, regardless of its meaning in context. (E.g. Apple/apple: one vector).

Representing language computationally:
Towards Language Models (2/2)

2017/18: Contextual word embeddings (Transformers: BERT/GPT): the meaning of words changes depending on the context, so each token gets a different representation depending on its context.

Representing language computationally

Masked language models (e.g., BERT)

Predicting the future is	[MASK]	but not impossible.

Fine-tune for specific tasks (such as text classification or named entity recognition) via examples.

Representing language computationally

Masked language models (e.g., BERT)

Predicting the future is	[MASK]	but not impossible.

Fine-tune for specific tasks (such as text classification or named entity recognition) via examples.

Generative models (e.g., GPT)

Predicting the future is hard, but not	???

Fine-tuned for instruction-tuning (w/wo reasoning).

A Language Model reflects the language of the corpus it has been trained from:

"They were told that the [MASK] stopped working."

Top predictions of BERT language models trained on English books from…

1760-1850	1850-1875	1875-1890
1. man	1. men	1. men
2. prisoners	2. people	2. miners
3. men	3. miners	3. machines
4. people	4. engine	4. engine
5. slaves	5. women	5. people

Hosseini, K., Beelen, K., Colavizza, G., & Coll Ardanuy, M. (2021). *Neural Language Models for Nineteenth-Century English*. Journal of Open Humanities Data, 7: 22, pp. 1–6.

Language Modelling Meets Text Analysis

Data preprocessing

. . . preparing text data for analysis

Most text data (especially non digital-born) exist in formats that are not machine-readable or usable for text analysis.

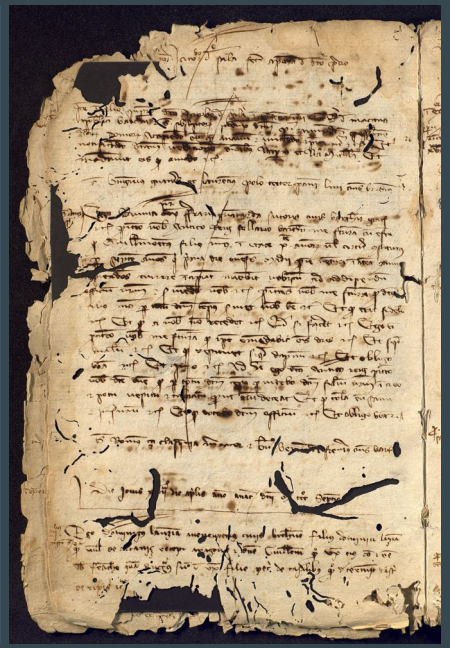
Advances in Language Modelling have been key for:

- Text extraction
- Textual metadata enrichment

Data preprocessing

Example of Text Extraction

Source: Arxiu Històric de Protocols de Barcelona, 28/1, f. 7v



<CONTINUE>
[Item, in alio instrumento dict]us Iacobus de Sala fecit apocham de dicto precio.
Item, in alio instrumento dictus Iacobus de Sala, in presencia mei notarii Ri de Maçana, publici Barchinone, et testium subscriptorum, in dominio dictum Pm Raimundi et possessionem dictarum domorum, accipiendo eum per manum et mittendo ipsius inter, et cetera, et notificando dictam vendicionem dominabus na Perua et na Michaela. Et mandavit eis quod amodo, et cetera.
Testes: Berengarius Guaneris et Laurencius Polo, textor panni lini, cives Barchinone.

Ego Rumia, uxor Ferrarii Guitardi, sutoris, civis Barchinone, gratis, et cetera, promitto vobis Antico Reig, sellario Barchinone, me facturam cum efectu quod Guillemonetus, filius meus, in etate Xcem annorum vel circiter, constitutus per septem annos qui prima die mensis madii proxime venturi in antea continu[e] [com]putandos currere incipiant manebit vobiscum ad addiscendum officium vestrum et serviendo vobis, et cetera. Promittens vobis me facturam quod dictus filius meus per totum dictum tempus serviet vobis bene, et cetera, et quod erit fidelis, et cetera, et quod a vobis non recedet, et cetera. Quod si faceret, et cetera. Ego tamen promitto vobis me facturam quod ipse emendabit omes diez, et cetera. Et si quid malum, et cetera. Et quod restituet si quid dampnum, et cetera. Et obligo bona, et cetera. Et iuro, et cetera. Ad hec ego dictus Anticus Reig promitto vobis dicte domine quod per totum dictum tempus providebo dictum filium vestrum in cibo et potu, vestitu et calciatu, prout eum deceat; et quod colam eum sanum et infirmum, et cetera. Et quod docebo dictum officium, et cetera. Et obligo bona, et cetera.
Testes: Romeus ça Clasquerin, mercator, et Bnus Be[…], frenerius, cives Barchinone.

Die iovis X[X]IIª die aprilis, anno a Nativitate Domini M CCCº (LX) septimo.
Ego Dominicus Lanuzia, aventurerius, civis Barchinone, filius Dominici Lanuzia, quondam, ville de [Al]caniç, recepi Aragon[i]e, et domine Guilleme, quondam, uxoris eius, confiteor et recognosco vobis Franche, quam ducturus sum in uxorem, filie Petri de Casalibus, quondam, de termino castri de Ripis, et cetera.
<CONTINUE>

Automatic Transcription: https://gitlab.bsc.es/cssh/releases/scalable-htr (Carrascosa, Coll Ardanuy, et al, 2026)

Data preprocessing

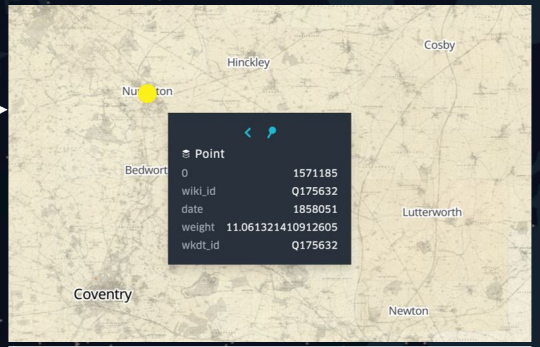
Example of Textual Metadata Enrichment

Toponym resolution: The task of finding and geolocating mentions of places, e.g. in news articles.

E.g.: Geolocation of 4,500 articles between 1880 and 1888 from *The Stourbridge Observer* newspaper (Newspaper image: *FindMyPast*). →

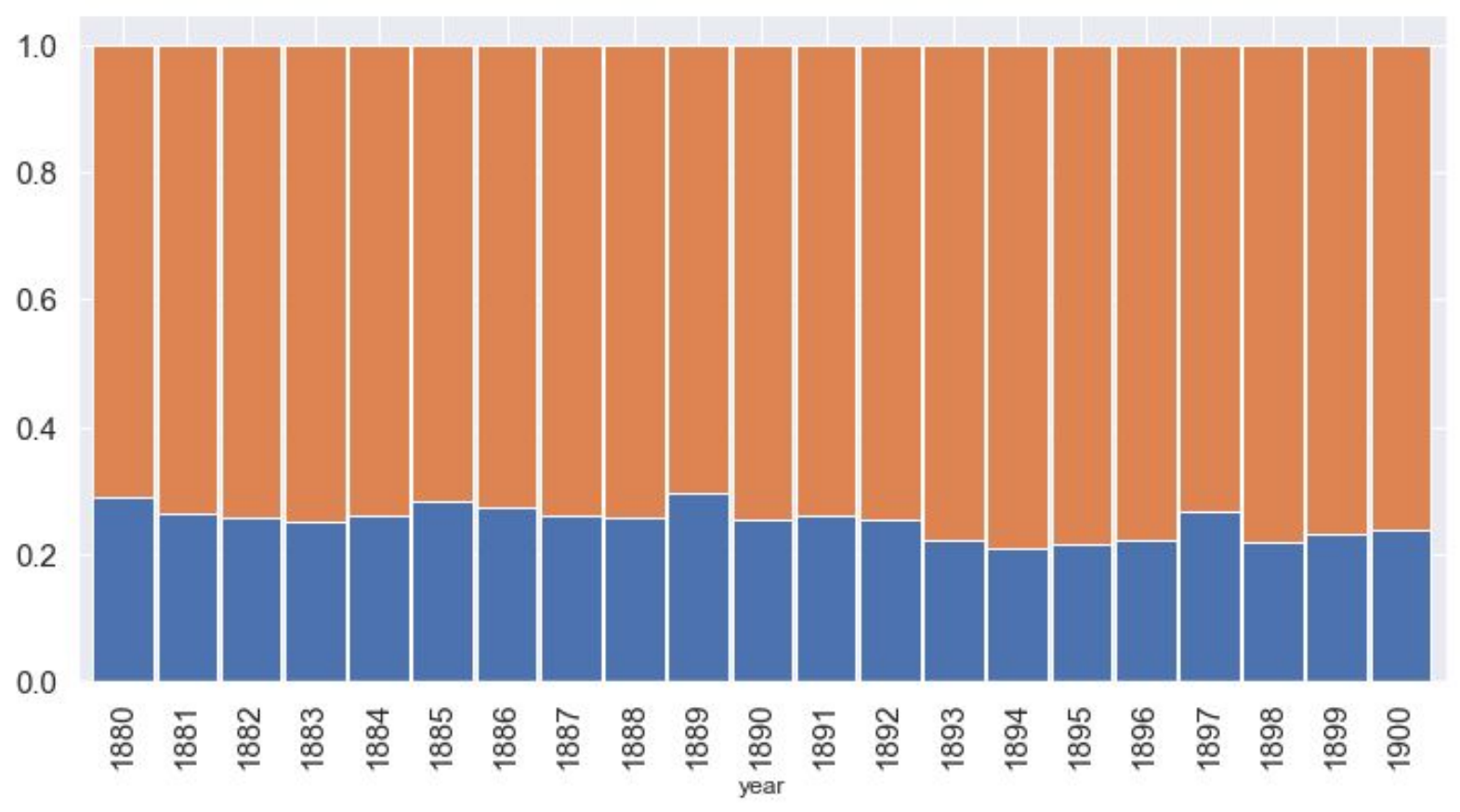

THE STOURBRIDGE OBSERVER

Coll Ardanuy, M., et al (CHR 2023).



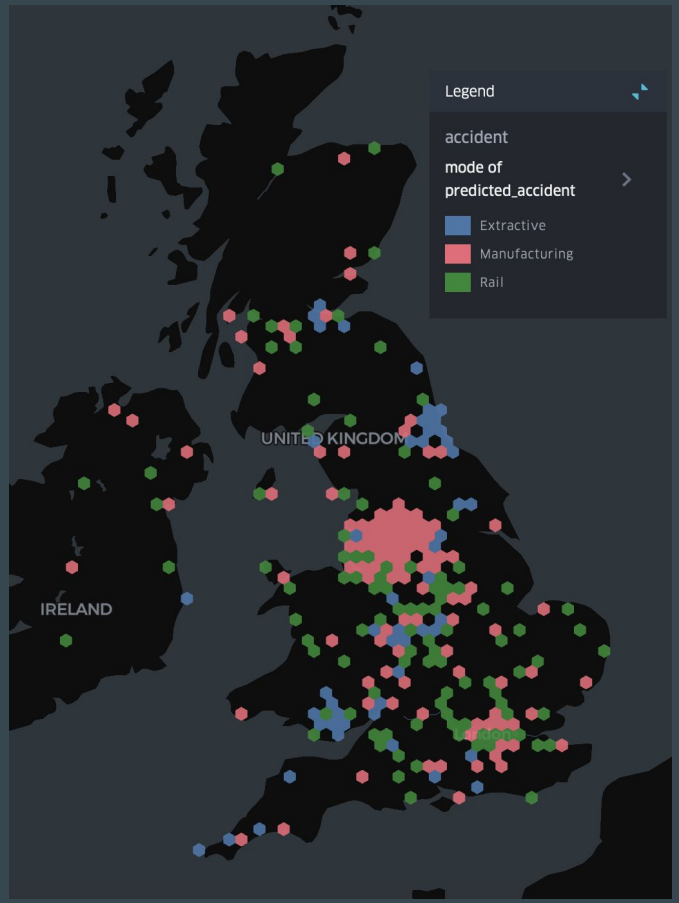
"After the accident had happened , and I saw the damage done, I ran my engine to Nuneaton to obtain medical and other assistance, and stop all coming trains ." *The Sun*, 13/05/1858

Places mentioned in 2,500 newspaper issues published between 1880 and 1900 across 56 regional working-class newspapers (2.7M toponyms)



In the British Isles / Outside the British Isles

Coll Ardanuy, M., et al (2023). "The Past is a Foreign Place: Improving Toponym Linking for Historical Newspapers." Computational Humanities Research (2023)



Data preprocessing

Example of Textual Metadata Enrichment

Toponym resolution: The task of finding and geolocating mentions of places, e.g., in news articles.

Text classification: The task of classifying texts (e.g., news articles) into classes, such as “type of accident mentioned”.

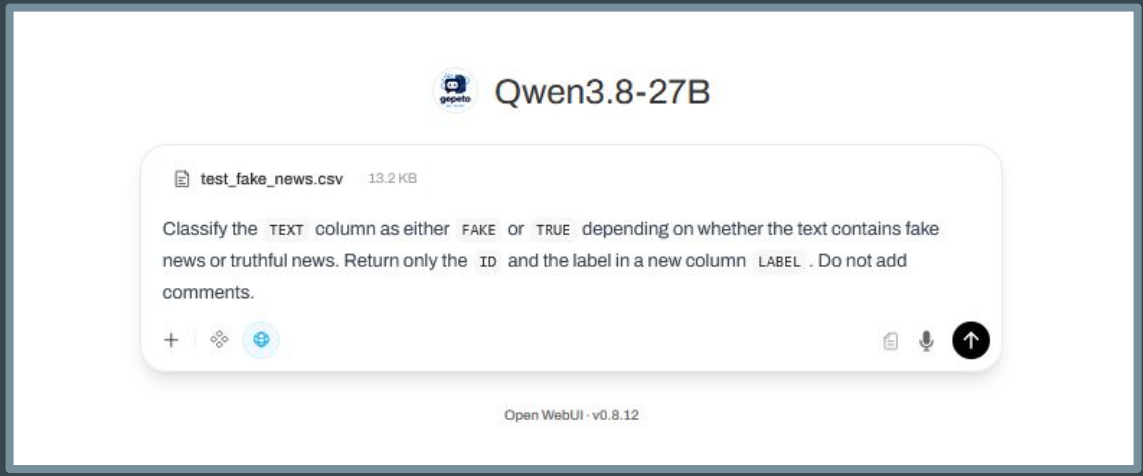
Data preprocessing

. . . preparing text data for large-scale analysis

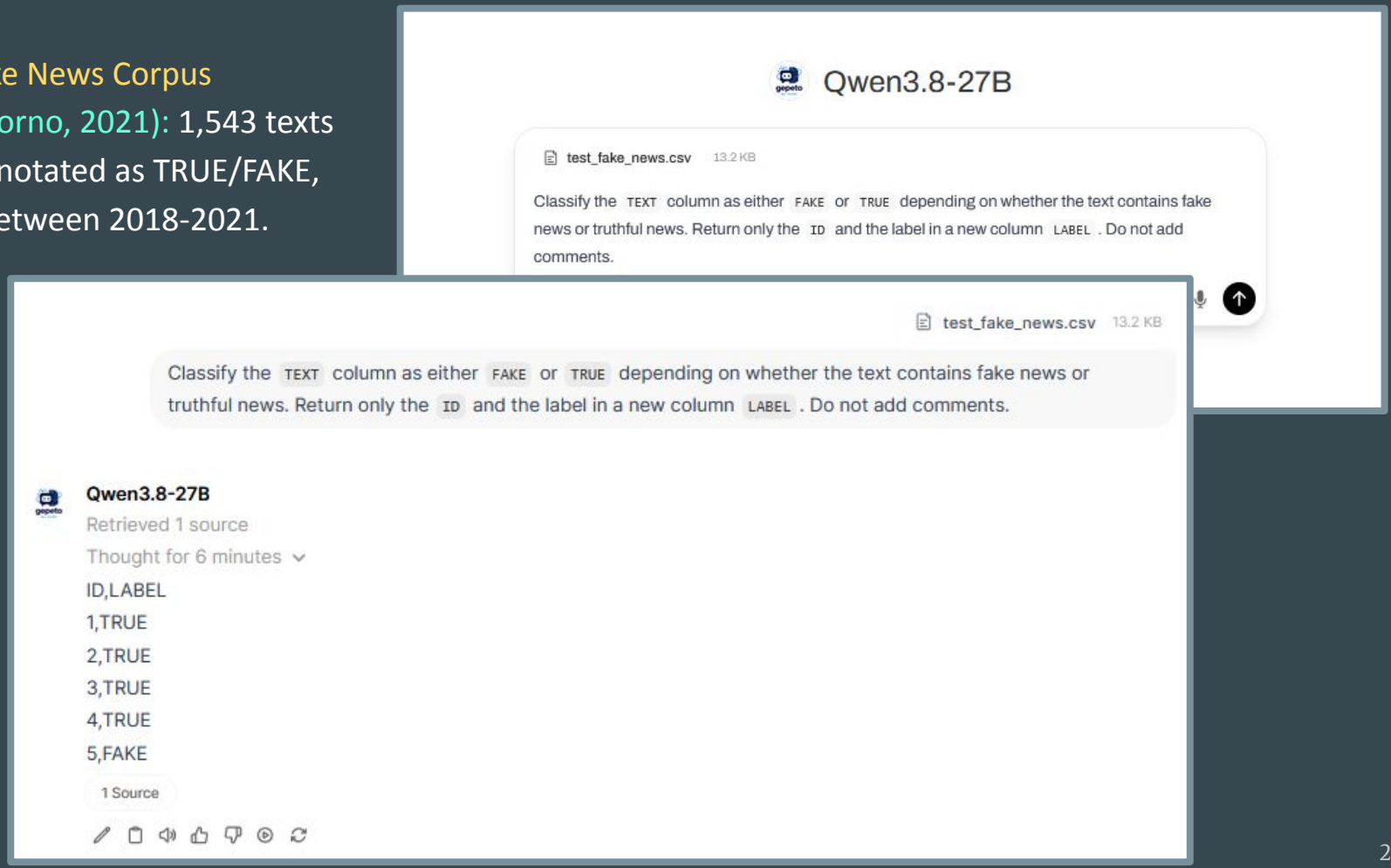
- LLMs for annotation:
  - Tasks: Classification, entity/relation extraction, biography parsing, Q&A (and more!)
  - Returns structured datasets that can directly be used for analysis/visualization.
- Current state:
  - For most languages and domains: Off-the-shelf LLMs can directly be used (via zero-/few-shot prompting w/wo RAG) to reach strong accuracy.
  - For very low-resource languages or specialised domains: Training may be needed.
- Validation (!!!):
  - Domain expertise + Quantitative evaluation

Spanish Fake News Corpus (Gómez-Adorno, 2021): 1,543 texts expertly annotated as TRUE/FAKE, compiled between 2018-2021.

Spanish Fake News Corpus (Gómez-Adorno, 2021): 1,543 texts expertly annotated as TRUE/FAKE, compiled between 2018-2021.



Spanish Fake News Corpus (Gómez-Adorno, 2021): 1,543 texts expertly annotated as TRUE/FAKE, compiled between 2018-2021.



"All models are entangled in discourse; they have histories, and they act politically." (Wharton, 2021)

Risks and Limitations of LLMs for Text Analysis

- LLMs as "stochastic parrots" (Bender et al., 2021): good at generating convincing language, optimized for plausibility.
- Factuality: what is correct, what are models confidently making up?
- Bias and fairness: reproduction and amplification of biases and stereotypes learned from the training data. Whose perspective is being channeled?
- Interpretability: black boxes (some, particularly opaque).
- Data contamination: has the model already seen your data?
- Privacy/cost/environment: esp. with proprietary models (Claude, Gemini, ChatGPT)

Leveraging the limitations of language models

A promising avenue for research in SSH

Investigate the (inner) workings and behaviour of language models:

- Where and how do different models fail? What can we learn from their mistakes? What can we learn from their predictions?
- What are their limits for tasks they are not optimized for? In which ways are they biased?
- Study the gap between expectation (what the model predicts) and observation (what the text actually says).

Use Case: Study atypical uses of word “machine”

Living with Machines project (2019-2023); Text Machine project (2026-2027)

- 19thC Britain: a society being transformed by industrialization.
- Identify contexts where machines are being attributed human-like characteristics, features, or behaviour:
  - To explore how machines have been imagined in the nineteenth century from lifeless mechanical objects to human-like agents that feel and think.
  - To study changing attitudes of society towards technological change.
  - To investigate the notion of agency and responsibility in humans vs machines.
- Precedent for today’s discussion on AI (Turing, 1950: “Can machines think?”)

Coll Ardanuy, et al (2020). "Living Machines: A Study of Atypical Animacy." COLING

Observation vs prediction: detecting biases and shifts

Language models to capture cultural or social bias or changing trends.

"Our sewing [MASK] stood near the wall where grated windows admitted sunshine, and their hymn to Labour was the only sound that broke the brooding silence."

Observation vs prediction: detecting biases and shifts

Language models to capture cultural or social bias or changing trends.

"Our sewing [MASK] stood near the wall where grated windows admitted sunshine, and their hymn to Labour was the only sound that broke the brooding silence."

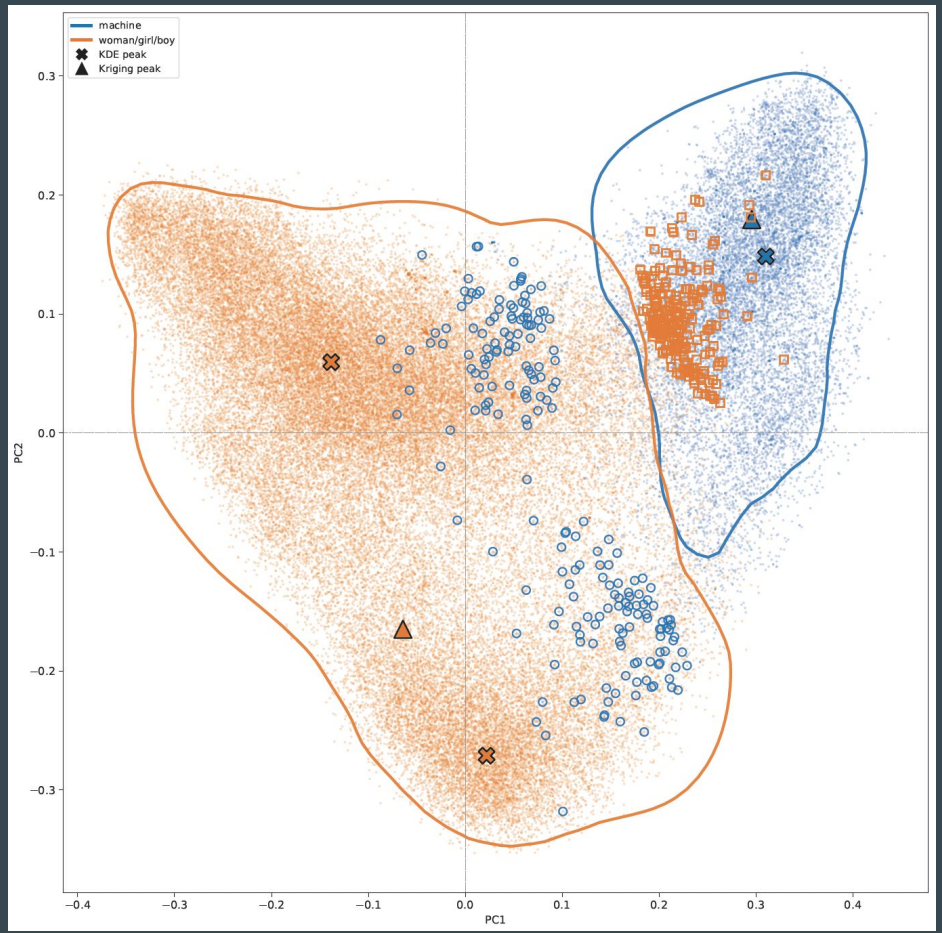
Predictions:

- 1760-1850 model: **'girls' 'women' 'sisters'**
- 2000-2020 model: **'machines' 'girls' 'women'**

Why Study Atypical Uses of Words?

- Rare by definition, and hard to retrieve from corpora.
- Interesting from an interpretative point of view:
  - Often deliberate: conveying emotion, beauty, or insight;
  - A window onto the social, cultural, and psychological experience/world of the writer.
- Examine how biases, prejudices and social stereotypes are embedded in the language models, how these change through time and corpora.

Coll Ardanuy, et al (2020). "Living Machines: A Study of Atypical Animacy." COLING



Pedrazzini, N., et al 2026 (in progress)

«Our servants, like mere machines, move on their mercenary track without feeling.» (1850)

«The misfortune was that she had been too much reduced to the state of a machine.» (1870)

«The Analytical Engine has no pretensions whatever to originate anything. It can do whatever we know how to order it to perform.» (Ada Lovelace, 1843)

«A soldier is, of all descriptions of men, the most completely a machine.» (1888)

«This hateful world had caught him and transformed him to a machine.» (1897)

«The private adventurer and the locomotive were the untechnical missionaries.» (1888)

Conclusions

- Advances in data science, open science, AI, and compute have opened new research possibilities for the Social Sciences and Humanities, when language models are used critically.
- Domain expertise in SSH is essential; interdisciplinary collaboration (computational+domain expertise) enables new forms of research.
- Bidirectionality: SSH to shape AI; AI to advance SSH.
- Creatively leveraging the limitations of language models is a promising (and exciting) research direction.

Thank you!

Credits

These slides present research conducted as part of the Text Machine project (with Ruth Ahnert, Kaspar Beelen, Daniel Wilson, Nilo Pedrazzini, Paolo Marangio), the Living with Machines project (with Text Machine team plus Federico Nanni and Mia Ridge) and the BSC Computational Archival History research line (with Adrián Carrascosa, Coral Cuadrada, Ramon Sarobe, Alexandra Málaga, Laura Servat and Mercè Crosas).

References (1/2)

Kasra Hosseini, Kaspar Beelen, Giovanni Colavizza, and Mariona Coll Ardanuy (2021). Neural Language Models for Nineteenth-Century English. Journal of Open Humanities Data, 7: 22, pp. 1–6.

Adrián Carrascosa, Mariona Coll Ardanuy, Laura Servat, Alexandra Málaga, Ramon Sarobe, Coral Cuadrada, and Mercè Crosas. 2026. Scalable-htr: A Framework for Scalable and Reproducible Handwritten Text Recognition on Historical Documents. In Proceedings of the 26th ACM/IEEE Joint Conference on Digital Libraries (JCDL '26). Association for Computing Machinery, New York, NY, USA, Article 104, 1–6. https://doi.org/10.1145/3805696.3846511

References (2/2)

Mariona Coll Ardanuy, Federico Nanni, Kaspar Beelen, and Luke Hare. "The Past is a Foreign Place: Improving Toponym Linking for Historical Newspapers." In Proceedings of Computational Humanities Research, pp. 368-390. 2023.

Mariona Coll Ardanuy, Federico Nanni, Kaspar Beelen, Kasra Hosseini, Ruth Ahnert, Jon Lawrence, Katherine McDonough, Giorgia Tolfo, Daniel CS Wilson, and Barbara McGillivray. "Living machines: A study of atypical animacy." In Proceedings of the 28th international conference on computational linguistics, pp. 4534-4545. 2020.

Daniel CS Wilson, Mariona Coll Ardanuy, Kaspar Beelen, Barbara McGillivray, and Ruth Ahnert. "The living machine: a computational approach to the nineteenth-century language of technology." Technology and Culture 64, no. 3 (2023): 875-902.