The Project: CLS INRA

Computational Literary Studies Infrastructure

- Aimed at integrating the fragmented landscape of literary data
- Geared toward standardising access and reuse of data in digital libraries
- Dedicated to advancing infrastructure as a diverse network of people, data, tools and methods within transdisciplinary theoretical frameworks

14 institutions across Europe

Over 35 researchers and contributors

17 Deliverables to date
28 by project end

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101004984

The People: Authors

Julie M. Birkholz, Ghent University, Royal Library of Belgium; Ingo Börner, University of Potsdam; Floor Buschenhenke, Huygens Institute for History and Culture of the Netherlands; Joanna Byszuk, Instytut Języka Polskiego Polskiej Akademii Nauk; Sally Chambers, Ghent University, Royal Library of Belgium; Vera Maria Charvat, ACDH-CH; Silvie Cinková, Charles University; Tess Dejaeghere, Ghent University; Anna Dijkstra, Huygens Institute for History and Culture of the Netherlands; Julia Dudar, University of Trier; Matej Ďurčo, ACDH-CH; Maciej Eder, Instytut Języka Polskiego Polskiej Akademii Nauk; Jennifer Edmond, DARIAH ERIC, University of Dublin Trinity College; Evgeniia Fileva, University of Trier; Frank Fischer, Freie Universität Berlin, DARIAH-EU; Vicky Garnett , Trinity College Dublin; Françoise Gouzi, DARIAH-EU; Serge Heiden,École normale supérieure de Lyon; Sarah Hoover, University of Galway; Michal Křen, Charles University; Els Lefever, Ghent University; Michał Mrugalski, Humboldt-Universität zu Berlin; Ciara L. Murphy, Technological University Dublin; Carolin Odebrecht, Humboldt-Universität zu Berlin; Eliza Papaki, DARIAH-EU; Marco Raciti, DARIAH-EU; Emily Ridge, University of Galway; Salvador Ros, UNED; Christof Schöch, University of Trier; Artjoms Šeļa, Instytut Języka Polskiego Polskiej Akademii Nauk; Toma Tasovac, DARIAH-EU; Justin Tonra, University of Galway; Erzsébet Tóth-Czifra, DARIAH-EU; Peer Trilcke, University of Potsdam; Karina van Dalen-Oskam, Huygens Institute for History and Culture of the Netherlands; Lisanne M. van Rossum, ; Vera Yakupova, Trinity College Dublin

Literary Methods for All: CLS INFRA

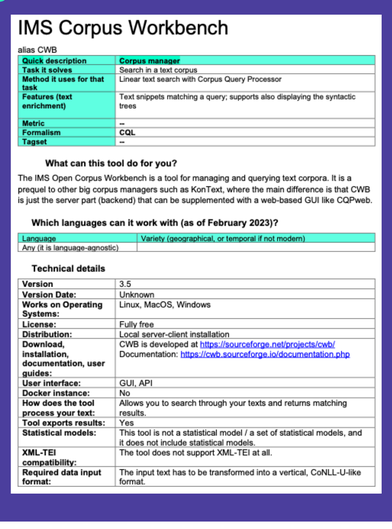
Discover Multilingual Text-Mining Tools

A Resource Guide

Reviews, lists and describes 27 NLP tools that form a Corpus-Enrichment and NLP toolchain

Text mining (or text analytics) tools assist in extracting patterns and non-obvious relationships from text. These are employed in chatbots, automatic summarization of online reviews, customer profiling, and many other applications. The D4.1 survey completed earlier by CLS INFRA showed that those with beginning to intermediate skills in text mining and the techniques of NLP (Natural Language Processing) often feel intimidated by the vastness of these disciplines.

This output, D8.1 (https://zenodo.org/records/7951060), summarises a series of important text-mining tools, their uses, the languages in which they work, and the integrations between these tools in an accessible format. These are divided into categories including Annotation, Corpus Management, NLP, and Poetry Processing.



IMS Corpus Workbench

alias CWB

Quick description	**Corpus manager**
Task it solves	Search in a text corpus
Method it uses for that task	Linear text search with Corpus Query Processor
Features (text enrichment)	Text snippets matching a query; supports also displaying the syntactic trees
Metric	--
Formalism	CQL
Tagset	--

What can this tool do for you?

The IMS Open Corpus Workbench is a tool for managing and querying text corpora. It is a prequel to other big corpus managers such as KonText, where the main difference is that CWB is just the server part (backend) that can be supplemented with a web-based GUI like CQPweb.

Which languages can it work with (as of February 2023)?

Language	Variety (geographical, or temporal if not modern)
Any (it is language-agnostic)	

Technical details

Version	3.5
Version Date:	Unknown
Works on Operating Systems:	Linux, MacOS, Windows
License:	Fully free
Distribution:	Local server-client installation
Download, installation, documentation, user guides:	CWB is developed at https://sourceforge.net/projects/cwb/ Documentation: https://cwb.sourceforge.io/documentation.php
User interface:	GUI, API
Docker instance:	No
How does the tool process your text:	Allows you to search through your texts and returns matching results.
Tool exports results:	Yes
Statistical models:	This tool is not a statistical model / a set of statistical models, and it does not include statistical models.
XML-TEI compatibility:	The tool does not support XML-TEI at all.
Required data input format:	The input text has to be transformed into a vertical, CoNLL-U-like format.

Data for the People

Toolkit Report for Data Sharing

Reproducibility of research is at the core of open science. The relationship between measurable reproducibility and shared insights takes center stage in philosophy of science and methodology of literary studies.
Based on the review of literature on, and real-life challenges to data sharing (the latter informed by a series of conversations with outstanding researchers in CLS), this deliverable provides readily comprehensible and easily implementable recommendations and templates organized along the lines of Research Data Life Cycle.

It covers research good practices in the context of CLS corpora and data with regard to:

planning and designing data

creating and collecting data

preparing and enriching data

preserving and publishing data

re-using data

What's the Use?

Non-academic Applications of (Computational) Literary Studies

This research explores the potential of CLS beyond academia. By profiling potential users from diverse fields such as policy, consultancy, journalism, publishing, medicine/psychology (bibliotherapy), and artistic practice, the project envisions CLS as a versatile tool.

User scenarios, drawn from interviews with non-literary research users, outline tasks in fields ranging from history research to narrative-driven futures institutes.

The research identifies four models, showcasing how fictional narratives contribute to diverse areas, including cultural identity claims, prototyping future scenarios, building predictive models, and fostering empathy or narrative medicine.

Helping non-academics take advantage of the opportunities offered by CLS tools functions as a key driver for the sustainability and impact of the infrastructure.

Deliverable 3.5, report expected June 2024. Follow ClsInfra.io for details!

Unlocking CLS Research

Navigating Methods and Issues with the Survey Grid

The Survey of Methods, D3.2 (https://zenodo.org/records/7892112) presented on an interactive grid (https://clsinfra.io/resources/d3-2-methods) provides an accessible introduction to practices, methods and issues that are prominent within the current landscape in CLS (Schöch, Dudar, Fileva, eds. 2023).

Though it is not intended as a primer, the Survey Grid provides targeted information in a focused way that is suitable for a quick overview of a few CLS subfields like authorship attribution, genre analysis, literary history, gender analysis, and canonicity.In this sense, the survey, while far from exhaustive, can also serve as an annotated bibliography and as a guide to further reading. If you are looking for a resource to introduce your colleague or a student to computational study of literature -- this can easily be the first link to send them.

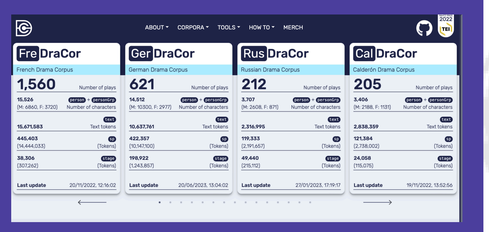
Growing a Corpus:

Versioning Living and Programmable Corpora

Digital corpora, which are proving more and more to be the most important epistemic objects of Computational Literary Studies (CLS), are by no means always static objects. On the contrary, it is becoming increasingly clear that the digitisation of our cultural heritage needs to be understood as an ongoing process, which also implies that a number of the epistemic objects of CLS must be conceptualized as genuinely dynamic.

Deliverable D7.3 includes a report as well as a technical prototype for Programmable Corpora: the Drama Corpora Platform "DraCor".
The main access points to the different components of the DraCor system are:

- Code and Data on GitHub: https://github.com/dracor-org
- DraCor Front-end: https://dracor.org/
- DraCor API: https://dracor.org/doc/api



In a broader sense, this report is also an exploration of the traces left by a living corpus in the technical space of a Git-based version control system. The traces are recovered using a method that we call "algorithmic corpus archaeology" – a method which we recommend to all those who embark on the epistemological adventure of working with living and programmable corpora.

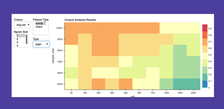
Please Pick up Your Spanners:

Showcases for the Application of CLS Methods and Tools

The four showcases brought together in Deliverable 3.3 illustrate, in a concrete, visual and interactive way, how some of the key methods in CLS work when they are applied to collections of literary texts.

Multilingual Stylometry Showcase

- Designed for: scholars and students, linguists
- Topics: the intersection of language, corpus composition, and stylometric methods of authorship attribution.



Averell

- Designed for: scholars and students of poetry
- Function: to help scholars create their own corpora of poetry by merging different corpora together

Detecting Small Worlds in a Corpus of Thousands of Theater Plays

- Designed for: scholars and students with basic knowledge of network theory
- Topics: typology of theatre plays from a network perspective

Poetrylab + rantanplan

- Designed for: readers, scholars, researchers in linguistics, literature, and culture
- Topics: Spanish poetic forms