

Ineo: Dutch CLARIN resources and tools enter the classroom

Third Hybrid UPSKILLS Multiplier Event:
Guidelines and Best Practices for Research-Based Teaching

Antal van den Bosch
Utrecht University / CLARIN ERIC BoD

4 November 2022






1



2

CLARIAH-NL's CLARIN heritage 


Flemish-Dutch joint infrastructure programmes:

• Spoken Dutch Corpus (CGN)	1998-2004	10 mNFL
• STEVIN	2004-2009	11.4 mEUR
• CLARIN-NL & CLARIN Flanders	2009-2014	9 mEUR

National funding for Large-scale research infrastructure CLARIAH (with DARIAH):

• CLARIAH-SEED	2015-2014	4 mEUR
• CLARIAH-CORE	2014-2019	12.6 mEUR
• CLARIAH-PLUS	2019-2025	13.8 mEUR

3

General information 

Name: CLARIAH-PLUS, www.clariah.nl

Period: 1 Jan 2019 – 31 Dec 2023


Funder: NWO, Netherlands (Grant 184.034.023)

Funding Instrument: National Roadmap for Large-Scale Research Facilities

Budget: 27.163 m euro


- 13.879 m contribution by NWO
- 12.684 m contribution in-kind
- 600 k euro cash contribution by KNAW

4

Context 

- Extends the CLARIAH research infrastructure
- Preceded by earlier national roadmap projects CLARIN-NL, CLARIAH-SEED and CLARIAH-CORE
- Integrates part of the European [CLARIN](#) and [DARIAH](#) research infrastructures
- Close collaboration with related infrastructure projects: [Nederlab](#), [Goiden Agents](#), [Athens](#), [Amsterdam Time Machine](#) and others
- New Roadmap call 2021: Joint effort with SSH Large-scale Infrastructure ODISSEI and others

5

Work packages & centers 

Data and Compute centres:

- WP2: [KNAW Humanities Cluster](#)
- WP3: [Meertens, Instituut voor de Nederlandse Taal, TLA-MPI](#)
- WP4: [International Institute of Social History \(IISH\)](#)
- WP5: [Netherlands Institute for Sound & Vision \(NISV\)](#)
- WP6: [National Library of the Netherlands \(KB\)](#)

6

Clariah

Current services

Selection of services created in CLARIAH-CORE:

- WPS: [LaMachine](#) ([Service Overview](#) & [Collection Overview](#))
- WP4: [DataLegend](#)
- WPS: [MediaSuite](#)
- New resource infrastructure Ineo

7



8

Clariah

General information

Name: Ineo (Latin: I start with...), www.ineo.tools

Goals:

- Offer users **one starting point** to find digital resources, instead of different portals for each humanities domain or resource type;
- with a **coherent** design, user interface and user experience;
- **based on user** requirements and preferences

9

Clariah

Present situation

A collage of three overlapping screenshots of existing CLARIAH services: 'CLARIAH Tools and Services' (a table of tools), 'DataLegend' (a map-based interface), and 'CLARIAH' (a general overview page with logos for COW, Cetic, and Droid).

10

Clariah

Resource page (concept)

Tab Overview
Fast and easy resource scanning, orientation and checking of usefulness.

Tab Learn
Description of available instruction and educational material.

Tab Mentions
Publications, presentations, media appearances, etc.

Tab Metadata
Machine readable, standardized metadata, for harvesting and archiving.


A screenshot of the proposed Ineo resource page concept, showing a clean layout with a search bar, navigation tabs, and a main content area displaying a resource card.

11



12

Corpus Research for Dutch Language and Literature / NE3V18001

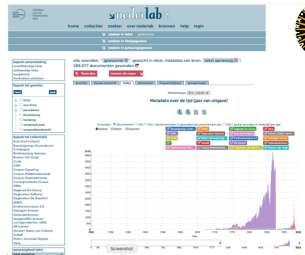



Course goals:

- Students are familiarized with a number of important corpora and tools for the humanities, among which **Nederlab**, **DBNL**, **Sonar** and the **Spoken Dutch Corpus**;
- They understand the way these corpora and tools have been built; which data can be found in these corpora; what the features of these data are and what the possibilities and limitations of these tools;
- Are able to ask meaningful questions to these corpora, accounting for their possibilities and limitations;
- Are able to do research on the corpora (using the tools);
- Are able to order, interpret and analyze the data.


13

Nederlab

14

Spoken Dutch Corpus



/Instituut voor de Nederlandse taal / taalmaterialen

Alle taalmaterialen Over deze website

Corpus Gesproken Nederlands (CGN)

Het Corpus Gesproken Nederlands (CGN) is een verzameling van 900 uur (bijna 9 miljoen woorden) hedendaagse Nederlands gesproken, afkomstig van Vlaanderen en Nederlanders. De spraakgegevens bestaan uit voerbestandrijke digitale en digitale transcripties (i.e. audiofiles, transcripties en annotaties (syntactisch, POS-tag), Metadata, lexica, frequentielijsten en de corpusplaatsoverzichten. Deze materialen worden ook wel het CGN genoemd.

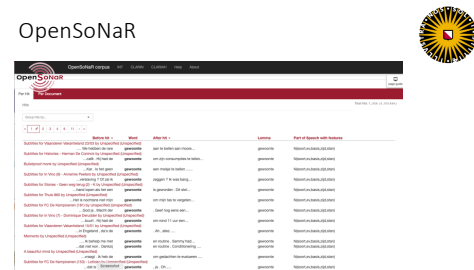

Naast het Corpus Gesproken Nederlands zijn de CGN-annotaties ook open te verkrijgen. Deze annotaties zijn identiek aan het volledige Corpus Gesproken Nederlands, maar dan zonder de geluidsbestanden.

Verwag de beschikbaarheid van dit product ook aangevraagd worden op een extern bestel schijf. Het rekent het ENT €100,00-voorstel- en afhandelingkosten vast.

■ A collection of about 900 hours spoken standard Dutch from Flanders and the Netherlands.



15

OpenSoNaR

16

Humanities Data Analysis: Case Studies with Python

Humanities Data Analysis: Case Studies with Python

Humanities Data Analysis: Case Studies with Python is a practical guide to data-intensive humanities research using the Python programming language. The book, written by Roger Karstner, Mike Kestemont and Adam Rosen, was originally published with Princeton University Press in 2021. It is a revised version of the book, see the publisher's website, and is now available as an Open Access Interactive Jupyter Book.

The book begins with an overview of the place of data science in the humanities, and proceeds to cover data economy, the essential techniques for gathering, cleaning, organizing, and transforming textual and tabular data. Then, drawing from real-world, publicly available data sets that cover a variety of disciplinary contexts, the book opens into detailed case studies. Focusing on textual data analysis, the authors explore such diverse topics as network analysis, genre theory, chronemics, rhetoric, author attribution, mapping, stylometry, topic modeling, and word sense analysis. Exercises and resources for further reading are provided at the end of each chapter.


What is the book about?

- Parsing and Manipulating Data:** Learn to have effectively gather, read, store and parse different data formats, such as CSV, XML, HTML, JSON, and JSONL data.
- Modeling and Data Representation of Texts:** Construct Vector-Space Models for texts and represent data in a tabular format. Learn how to use these and other representations (such as topics) to assess authorship and distance between texts.

Working on Real-World Case Studies

17

MacBERTh and GysBERT models



MacBERTh and GysBERT
Language Models for Historical English and Dutch
PDI-SSH 2020

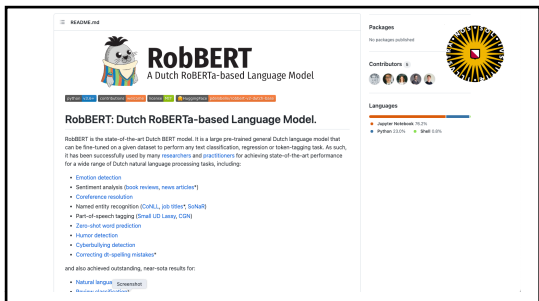
How to cite:

- **MacBERTh (English):** When using the English model (MacBERTh), please cite the following paper (BIBTEX can be found using the 'cite' button in 'Project Publications')
 - Manjaveaux, Enrique & Lauren Fonteyn. 2022. *Adapting vs. Pre-training Language Models for Historical Languages*. *Journal of Data Mining & Digital Humanities* (preprint 1912). <https://doi.org/10.46298/jdmh.1912>
- **GysBERT (Dutch):** We have written up a paper describing the Dutch model and its evaluation, which will (hopefully) be published soon. We will add the citation details as soon as they are known.

MacBERTh and GysBERT are language models (more specifically, BERT models) pre-trained on historical textual material (date range: 1450-1950).

Researchers who interpret and analyse historical textual material are well-aware that languages are subject to change over time, and that the way in which concepts and discourses of class, gender, norms and prestige function in different time periods. As such, it is quite important that the interpretation of textual/linguistic material from the past is not approached from a present-day point-of-view, which is why NLP models pre-

18



The screenshot shows the GitHub repository page for RobBERT. At the top left, there is a profile picture of a cartoon character and the text "RobBERT A Dutch RoBERTa-based Language Model". Below this, there is a description: "RobBERT: Dutch RoBERTa-based Language Model. RobBERT is the state-of-the-art Dutch RoBERTa model. It is a large pre-trained general Dutch language model that can be fine-tuned on a given dataset to perform any text classification, regression or token-tagging task. As such, it has been successfully used for many researches and applications for achieving state-of-the-art performance for a wide range of Dutch natural language processing tasks, including:

- Sentiment analysis
- Conference recognition
- Named entity recognition (CONLL, POS, NER, SNeE)
- Part-of-speech tagging (Dutch UD Lancy, GSN)
- Zero-shot word prediction
- Name detection
- Cyberbullying detection
- Correcting spelling mistakes

and also achieved outstanding, near-sota results for:

- Natural language [questioning](#)
- [Business document classification](#)

19



20