

Data and Corpusmanagement with Python: Theory and Applications (A005869)

Course size (nominal values; actual values may depend on programme)

Credits 5.0 **Study time 150 h**

Course offerings in academic year 2026-2027

A (semester 2) Dutch Gent

Lecturers in academic year 2026-2027

Van Hee, Cynthia	LW22	lecturer-in-charge
De Langhe, Loic	LW22	co-lecturer
Swaelens, Colin	LW06	co-lecturer

Offered in the following programmes in 2026-2027

	crdts	offering
Bachelor of Arts in Applied Language Studies: a combination of at least two languages(main subject Dutch, English, Language Technology)	5	A
Bachelor of Arts in Applied Language Studies: a combination of at least two languages(main subject Dutch, French, Language Technology)	5	A
Bachelor of Arts in Applied Language Studies: a combination of at least two languages(main subject Dutch, German, Language Technology)	5	A
Preparatory Course Master of Arts in Multilingual Communication: a combination of at least two languages	5	A
Preparatory Course Master of Arts in Translation: a combination of at least two languages	5	A

Teaching languages

Dutch

Keywords

Python, programming, corpus construction and analysis, data processing, data visualization, language technology.

Position of the course

This course builds on the course *Inleiding Programmeren voor de Humanities* (A005868) and focuses specifically on **Python-supported programming techniques for collecting, managing (in various formats), analyzing, and visualizing data**. This knowledge provides students with the necessary background and key skills for the courses *NLP and Linguistic Analysis* (A005870) and *Machine Learning and Feature Engineering* (A005871) in the second year (Ba2).

Contents

Python project management

- Managing Python projects using tools like PyCharm, virtual environments, and Google Colab.
- Handling file structures and file path references.

Data collection

- Importing existing corpora in various formats such as csv, tsv, json, and xml.
- Processing and analyzing different types of Python objects, including dictionaries, nested XML structures, and tabular data.
- Automatically generating new corpora using APIs and web crawling techniques.

Data management and analysis

- Searching data using regular expressions and parsing techniques.
- Manipulating, analyzing, and translating data into corpus statistics using Python libraries such as Pandas.
- Saving data in various formats such as csv, tsv, json and XML, and in data

frames.

Data visualization

- Visualizing corpus data and analyses with tools like Matplotlib.
- Using Python to export data from objects into Excel spreadsheets.

Initial competences

A solid foundational knowledge of Python programming is required (as covered in *Inleiding Programmeren voor de Humanities*).

In addition, students are expected to be familiar with basic concepts in digital text analysis. Therefore, they are required to enroll simultaneously in the course *Inleiding tot de Digitale Tekstanalyse*.

Final competences

- 1 Students can import, clean, and prepare corpora in various file formats for analysis, including tasks such as tokenization and handling encoding issues.
- 2 Students are able to independently construct corpora by applying Python-based web scraping techniques.
- 3 Students are capable of analyzing corpora and extracting relevant statistics using analytical techniques and tools such as regular expressions and the Pandas Python library.
- 4 Students can export texts in multiple formats and visualize basic corpus statistics using appropriate Python libraries.

Conditions for credit contract

Access to this course unit via a credit contract is unrestricted: the student takes into consideration the conditions mentioned in 'Starting Competences'

Conditions for exam contract

This course unit cannot be taken via an exam contract

Teaching methods

Seminar

Extra information on the teaching methods

Interactive lectures: the theoretical course material is made available in advance via Ufora, allowing class time to be used as much as possible for practice. Students are expected to actively participate during the sessions.

Study material

Type: Slides

Name: slides

Indicative price: Free or paid by faculty

Optional: no

References

Course content-related study coaching

During lectures (plenary discussion and remediation) and practice sessions.

Lecture recordings may be made available upon request.

Assessment moments

end-of-term assessment

Examination methods in case of periodic assessment during the first examination period

Written assessment

Examination methods in case of periodic assessment during the second examination period

Written assessment

Examination methods in case of permanent assessment

Possibilities of retake in case of permanent assessment

not applicable

Calculation of the examination mark

100% written exam.

