14:40 - 12:30
Demonstrations of the UPSKILLS Learning Content Blocks

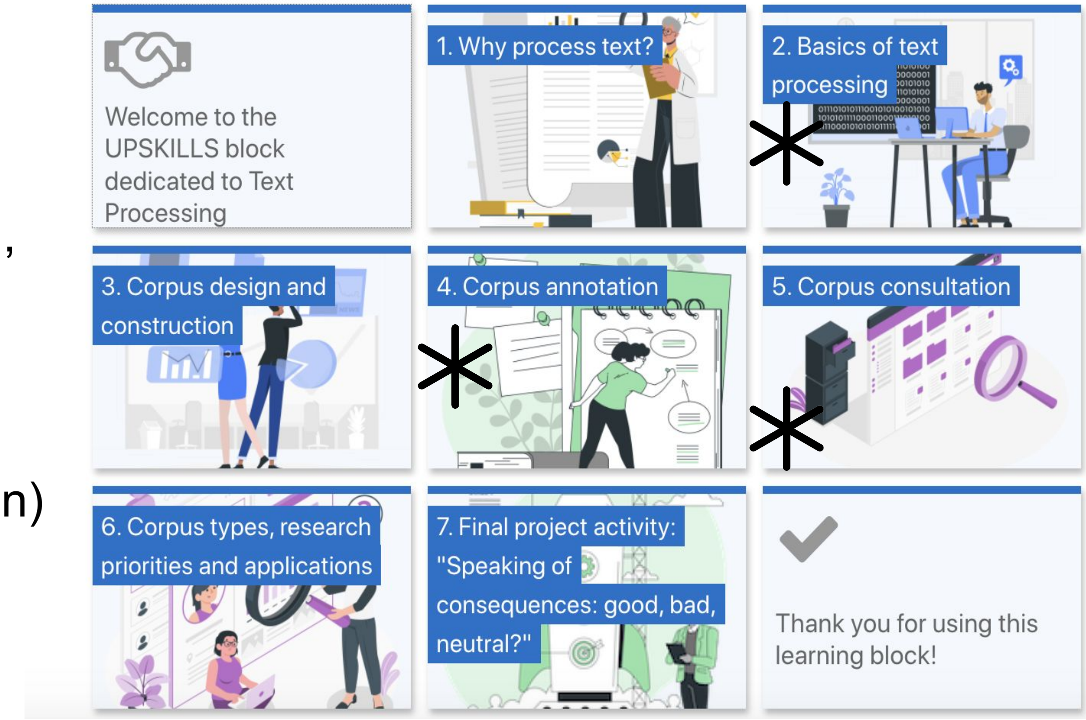
Text Processing

Novella Tedesco
Alma Mater Studiorum - University of Bologna

Learning outcomes

- Necessary **theoretical notions** about corpus linguistics and text processing
- **Technical** skills, including creation, management and use of language corpora for linguistic analysis, basics of text processing (such as regular expressions, text annotation)
- **Research** skills
- **Communication**, **interpersonal** and **organizational** skills

Course structure



Welcome to the UPSKILLS block dedicated to Text Processing

1. Why process text?

2. Basics of text processing ✱

3. Corpus design and construction

4. Corpus annotation ✱

5. Corpus consultation ✱

6. Corpus types, research priorities and applications

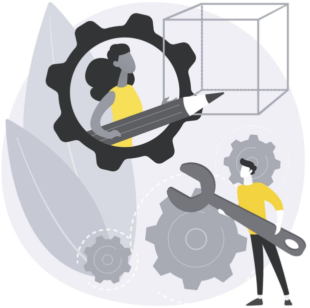
7. Final project activity: "Speaking of consequences: good, bad, neutral?"

Thank you for using this learning block!

✱ *adaptation of materials by DigiLing*

MAIN CHARACTERISTICS OF THE COURSE

- **Research-based** teaching and learning
- **Interactive** content slides and activities
- Final **tests**
- Resources for **self-study**
- Options for **gamification**

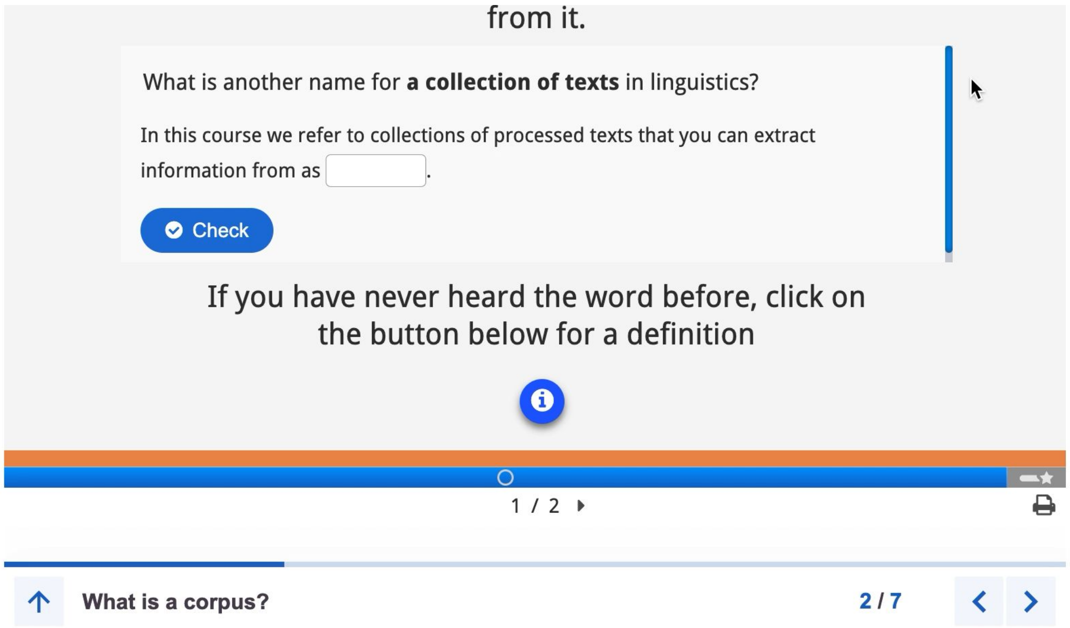


TECHNICAL ASPECTS → Creating learning contents with **H5P**

Some examples from the actual course…

An interactive presentation from Unit 1

Why is text processing Important for you?



from it.

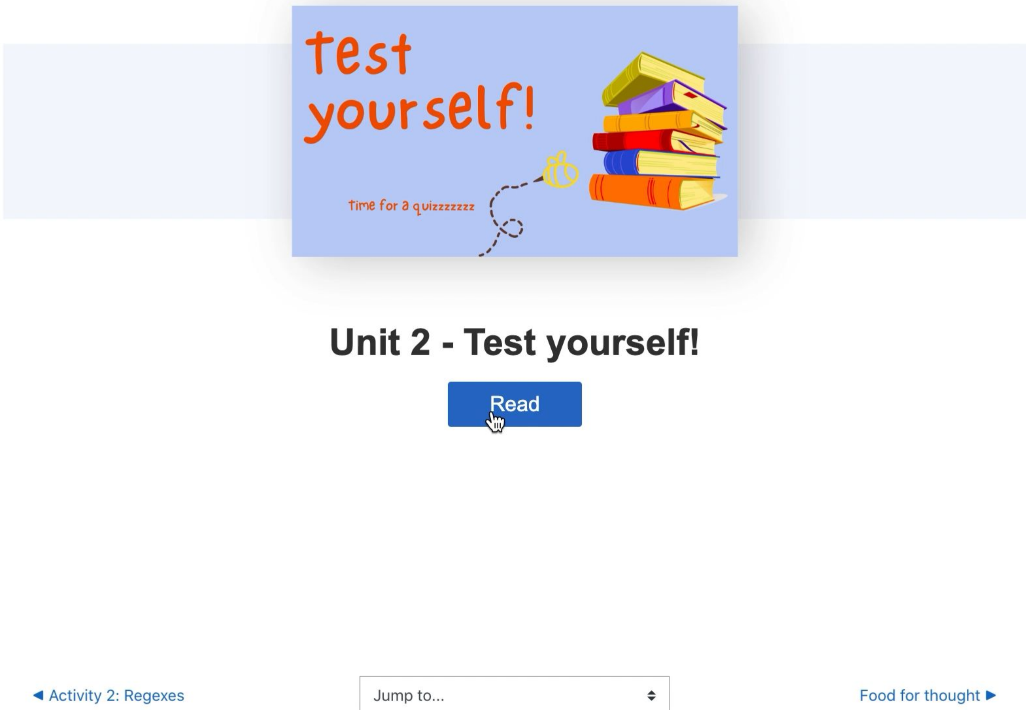
What is another name for **a collection of texts** in linguistics?

In this course we refer to collections of processed texts that you can extract information from as ________.

If you have never heard the word before, click on the button below for a definition

Some examples from the actual course…

Final test - Unit 2

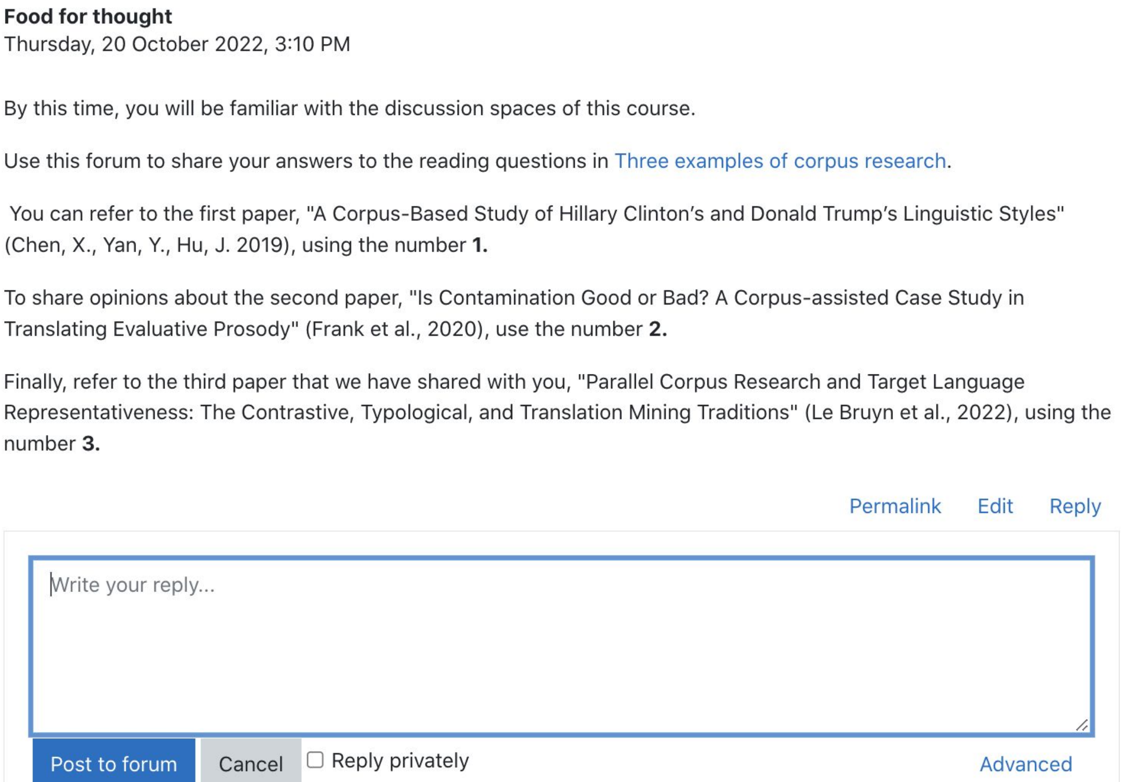


Unit 2 - Test yourself!

Some examples from the actual course…

A space for discussion

Food for Thought - Unit 6



Food for thought
Thursday, 20 October 2022, 3:10 PM

By this time, you will be familiar with the discussion spaces of this course.

Use this forum to share your answers to the reading questions in Three examples of corpus research.

You can refer to the first paper, "A Corpus-Based Study of Hillary Clinton's and Donald Trump's Linguistic Styles" (Chen, X., Yan, Y., Hu, J. 2019), using the number **1.**

To share opinions about the second paper, "Is Contamination Good or Bad? A Corpus-assisted Case Study in Translating Evaluative Prosody" (Frank et al., 2020), use the number **2.**

Finally, refer to the third paper that we have shared with you, "Parallel Corpus Research and Target Language Representativeness: The Contrastive, Typological, and Translation Mining Traditions" (Le Bruyn et al., 2022), using the number **3.**

14:40 - 12:30 Demonstrations of the UPSKILLS Learning Content Blocks

Introduction to Language Data: Standards and Repositories

Iulianna van der Lek

Darja Fišer

With contributions from:

Francesca Frontini

Alexander Konig

Willem Elbers

Learning outcomes

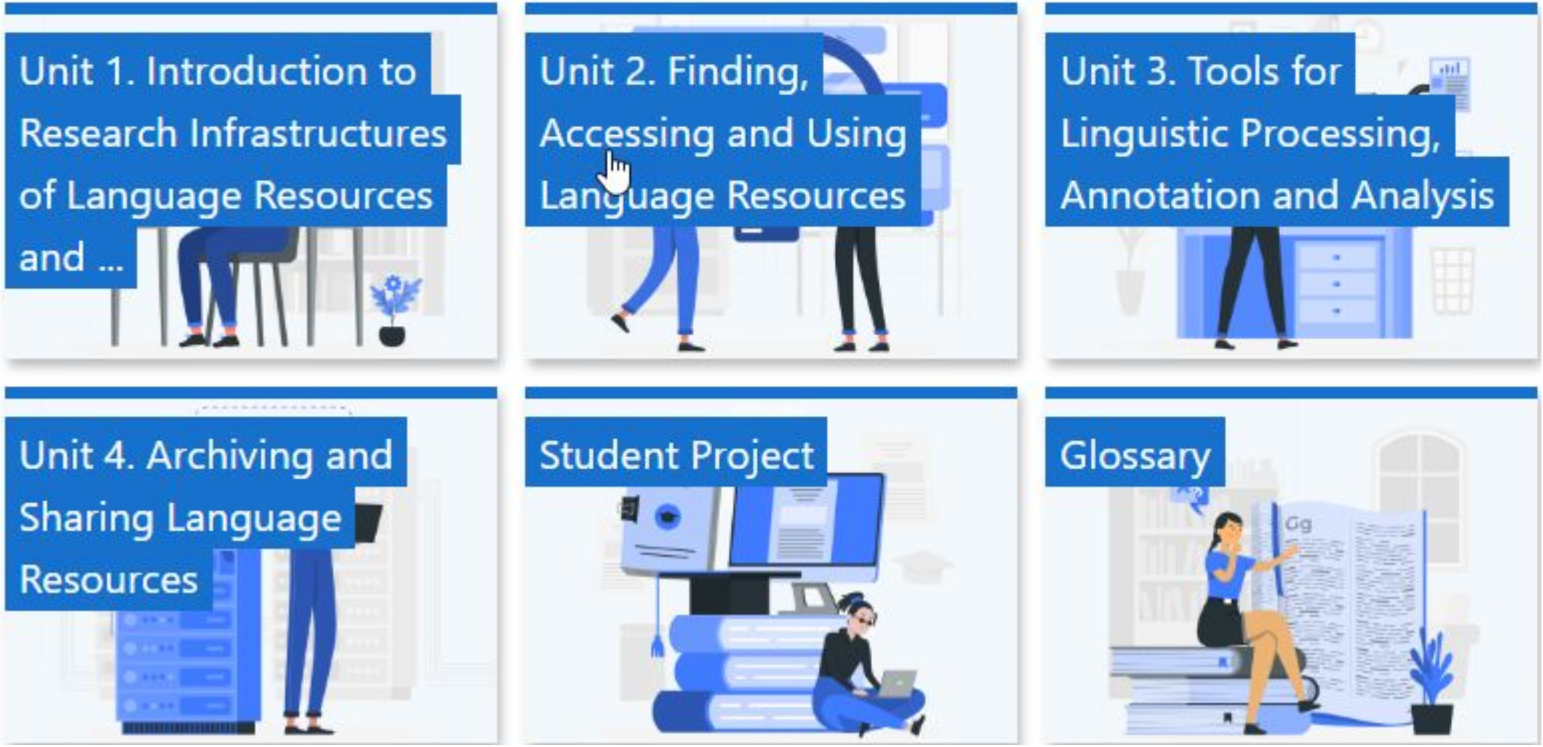
4-5 ECTS

By the end of this unit block, students will be able to:

- Explain what a language resource is and the role that research infrastructures play in the research data lifecycle in the context of Open Science and FAIR
- Use certified research data repositories to search, find and access language resources and datasets
- Process, annotate, and analyse different types of corpora in online environments according to standards and formats used by the community
- Archive and share language resources.

Prerequisites: Introduction to Text Processing (UniBo)

Course structure in Moodle



Unit 1. Introduction to Research Infrastructures of Language Resources and ...

Unit 2. Finding, Accessing and Using Language Resources

Unit 3. Tools for Linguistic Processing, Annotation and Analysis

Unit 4. Archiving and Sharing Language Resources

Student Project

Glossary

Highlights

- Learn by doing
- Interactive content slides in H5P and learning activities
- Take-home assignments and resources for self-study
- Modular: lessons can be picked and combined

Example of assignment

Search for 5 corpora in the CLARIN Resource Families on a topic that interest you and assess their FAIRness by answering the questions below:

- Findability: Are the corpora findable via Google/Bing, VLO and OLAC?
- Accessibility: Is the data accessible?
- Interoperability: In which format is the data available?
- Reusability: Is there documentation available on formats, methods and licensing?
- Other: Is the data openly available, is there a corpus paper or a dedicated website available?

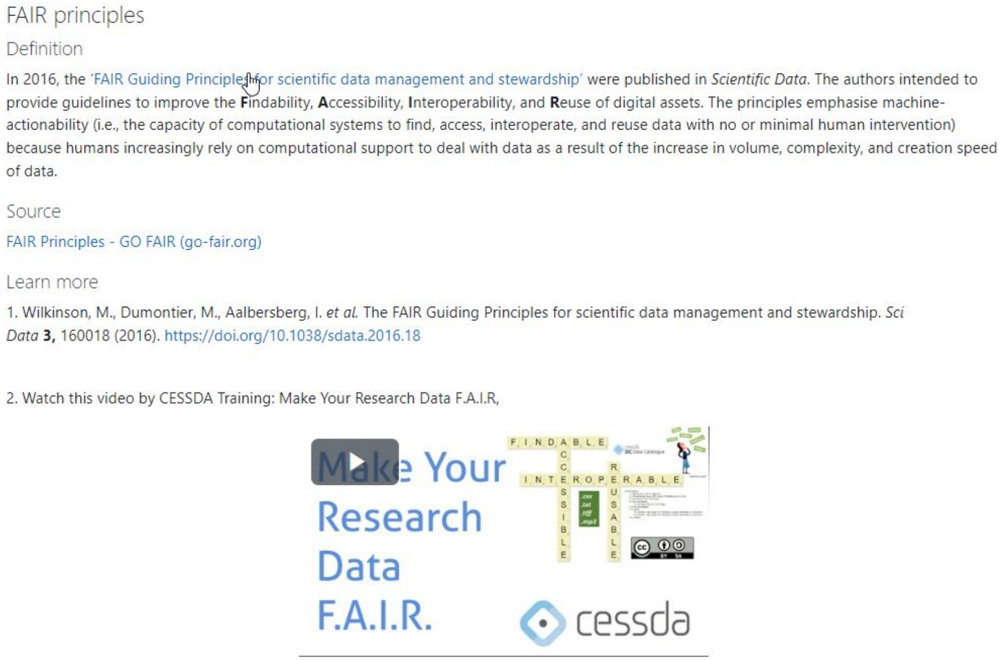
Delivery format: Blog post (800 words max).

Learning activity based on:

Frey, J.-C., König, A., & Stemle, E. W. (2019). How FAIR are CMC Corpora? Proceedings of the 7th Conference on CMC and Social Media Corpora for the Humanities (CMC-Corpora2019), 25–30. https://cmccorpora19.sciencesconf.org/resource/page/id/15.

Glossary

Key concepts related to repositories, standards and research infrastructures



FAIR principles

Definition

In 2016, the 'FAIR Guiding Principles for scientific data management and stewardship' were published in *Scientific Data*. The authors intended to provide guidelines to improve the **F**indability, **A**ccessibility, **I**nteroperability, and **R**euse of digital assets. The principles emphasise machine-actionability (i.e., the capacity of computational systems to find, access, interoperate, and reuse data with no or minimal human intervention) because humans increasingly rely on computational support to deal with data as a result of the increase in volume, complexity, and creation speed of data.

Source

FAIR Principles - GO FAIR (go-fair.org)

Learn more

1. Wilkinson, M., Dumontier, M., Aalbersberg, I. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* **3,** 160018 (2016). https://doi.org/10.1038/sdata.2016.18

2. Watch this video by CESSDA Training: Make Your Research Data F.A.I.R,

Upskilling your “Introduction to language variation class”

Margherita Pallottino
&
Genoveva Puskas

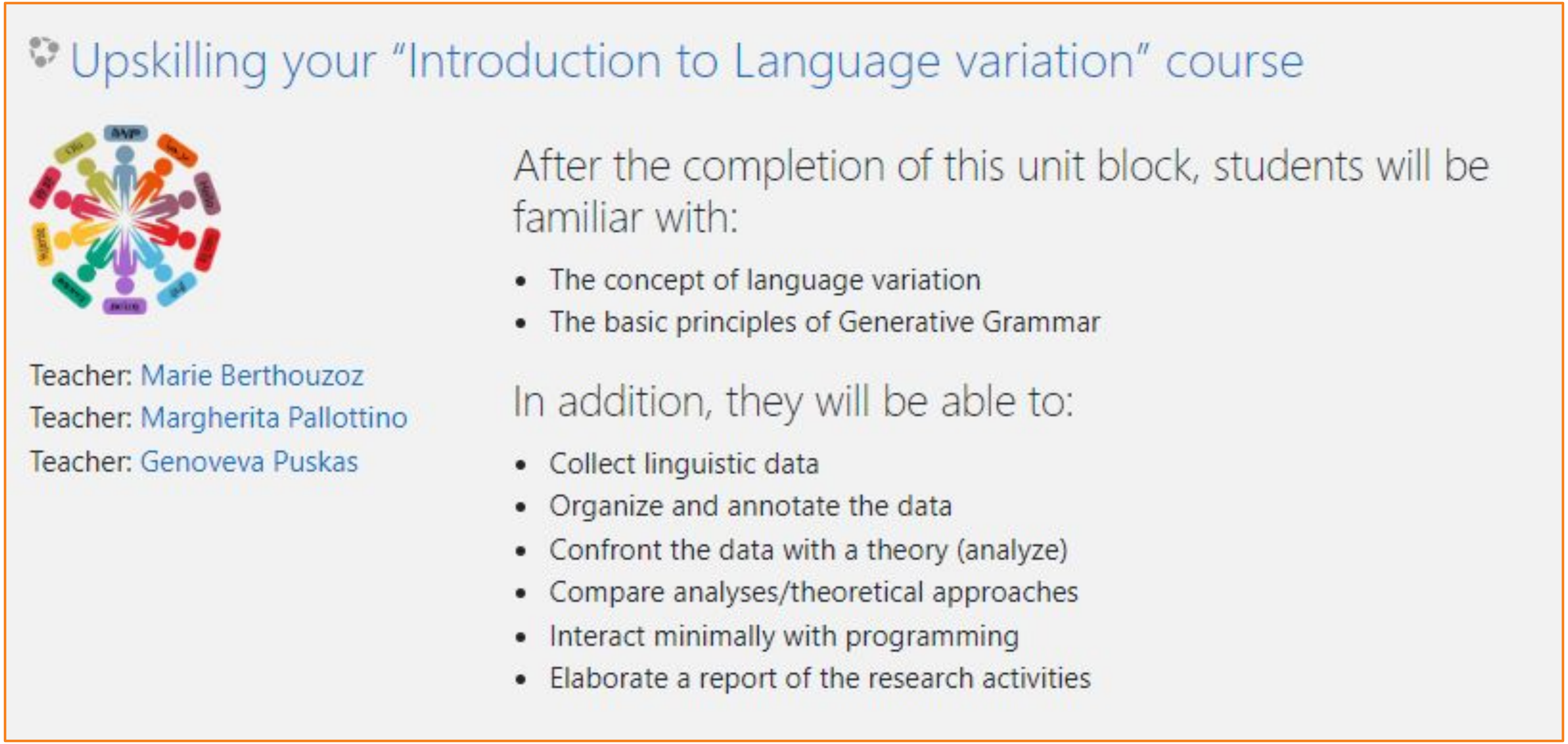
Topic 2, Class 3

The bloc in THEORY

You institution has rigid curricula which can hardly be changed from one year to the other, but you still find the UPSKILLS proposal interesting…

Upskills is not just a repository of content blocs, but it also is an approach to teaching, which aims at promoting the development of additional skills in an interactive class environment.

Where “additional” refers to the fact that an instructor must not replace the disciplinary content, for instance, theoretical linguistic, with practical notions, but integrate the two aspects.



Upskilling your "Introduction to Language variation" course

Teacher: Marie Berthouzoz

Teacher: Margherita Pallottino

Teacher: Genoveva Puskas

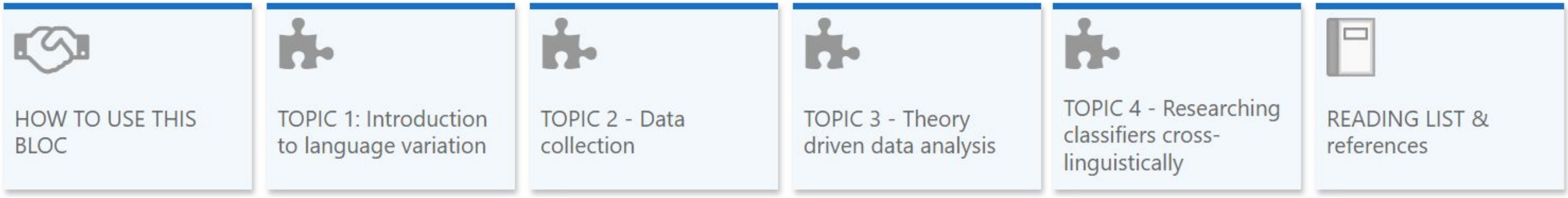
After the completion of this unit block, students will be familiar with:

- The concept of language variation
- The basic principles of Generative Grammar

In addition, they will be able to:

- Collect linguistic data
- Organize and annotate the data
- Confront the data with a theory (analyze)
- Compare analyses/theoretical approaches
- Interact minimally with programming
- Elaborate a report of the research activities

All of it with the support of a game!



HOW TO USE THIS BLOC

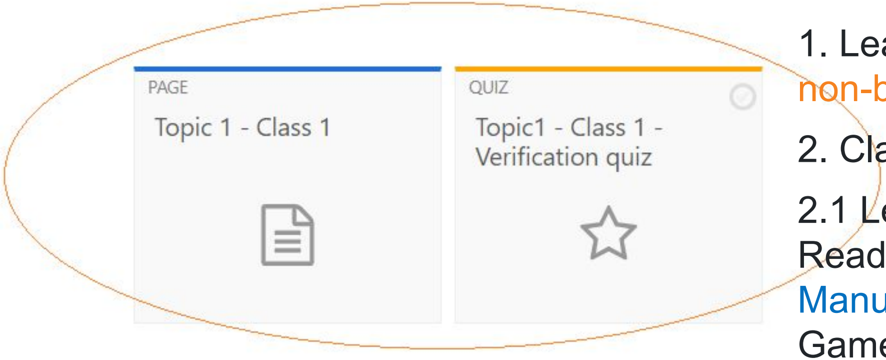
TOPIC 1: Introduction to language variation

TOPIC 2 - Data collection

TOPIC 3 - Theory driven data analysis

TOPIC 4 - Researching classifiers cross-linguistically

READING LIST & references



PAGE

Topic 1 - Class 1

QUIZ

Topic1 - Class 1 - Verification quiz

1. Learning Outcomes: Provide examples of a non-binary/binary linguistic feature/attribute

2. Class layout Organization of the time in class

2.1 Learning material
Readings: Radford (2009, pp. 1-19); “Guess the language” Manual of instructions
Games: *Guess the Language!©”* Hasbro *“Guess who? ®“*

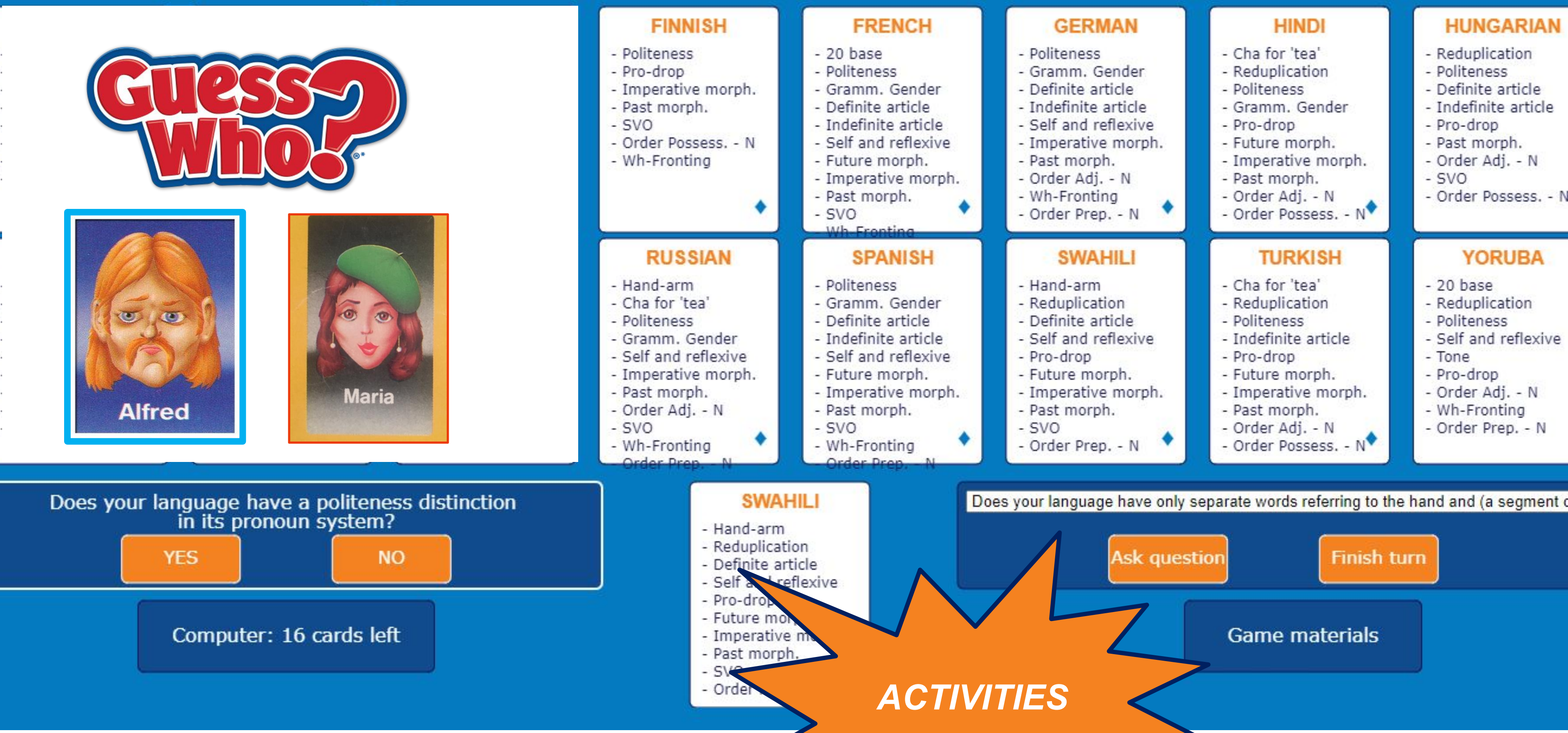
2.2 In-class activity: "Break the ice with a guessing game“ material needed, goal of the activity, carrying out the activity, estimated time

2.3 Debrief and introduction to Generative Grammar

3. Proposed homework

4. Assessment

5. Detailed workload (for the teacher, for the students)



Guess Who?

Alfred

Maria

FINNISH
- Politeness
- Pro-drop
- Imperative morph.
- Past morph.
- SVO
- Order Possess. - N
- Wh-Fronting

FRENCH
- 20 base
- Politeness
- Gramm. Gender
- Definite article
- Indefinite article
- Self and reflexive
- Future morph.
- Imperative morph.
- Past morph.
- SVO
- Wh-Fronting

GERMAN
- Politeness
- Gramm. Gender
- Definite article
- Indefinite article
- Self and reflexive
- Imperative morph.
- Past morph.
- Order Adj. - N
- Wh-Fronting
- Order Prep. - N

HINDI
- Cha for 'tea'
- Reduplication
- Politeness
- Gramm. Gender
- Pro-drop
- Future morph.
- Imperative morph.
- Past morph.
- Order Adj. - N
- Order Possess. - N

HUNGARIAN
- Reduplication
- Politeness
- Definite article
- Indefinite article
- Pro-drop
- Past morph.
- Order Adj. - N
- SVO
- Order Possess. - N

RUSSIAN
- Hand-arm
- Cha for 'tea'
- Politeness
- Gramm. Gender
- Self and reflexive
- Imperative morph.
- Past morph.
- Order Adj. - N
- SVO
- Wh-Fronting
- Order Prep. - N

SPANISH
- Politeness
- Gramm. Gender
- Definite article
- Indefinite article
- Self and reflexive
- Future morph.
- Imperative morph.
- Past morph.
- SVO
- Wh-Fronting
- Order Prep. - N

SWAHILI
- Hand-arm
- Reduplication
- Definite article
- Self and reflexive
- Pro-drop
- Future morph.
- Imperative morph.
- Past morph.
- SVO
- Order Prep. - N

TURKISH
- Cha for 'tea'
- Reduplication
- Politeness
- Indefinite article
- Pro-drop
- Future morph.
- Imperative morph.
- Past morph.
- Order Adj. - N
- Order Possess. - N

YORUBA
- 20 base
- Reduplication
- Politeness
- Self and reflexive
- Tone
- Pro-drop
- Order Adj. - N
- Wh-Fronting
- Order Prep. - N

Does your language have a politeness distinction in its pronoun system?

YES NO

Computer: 16 cards left

SWAHILI
- Hand-arm
- Reduplication
- Definite article
- Self a reflexive
- Pro-drop
- Future mo
- Imperative m
- Past morph.
- SV
- Order

Does your language have only separate words referring to the hand and (a segment o

Ask question Finish turn

Game materials

ACTIVITIES

ACTIVITIES

Break the ICE in class

Teach how to interact with a JavaScript file

Teach different data collection techniques

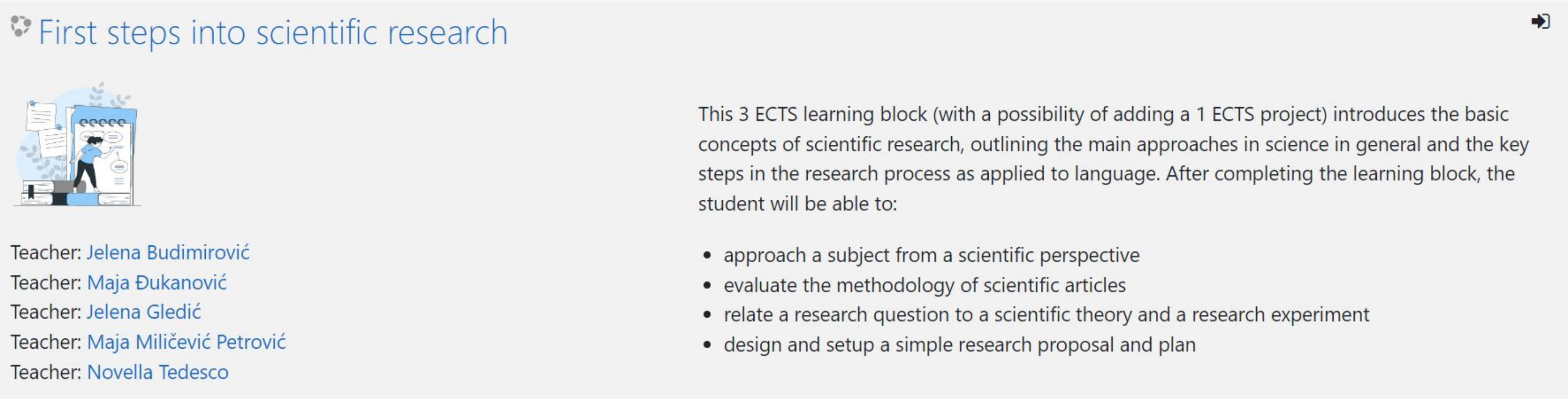
Teach what language diversity comes down to be

Teach how to organize data in a meaningful way

THANK YOU & HAVE FUN PLAYING!

Download *Guess the Language!* ©
from the chat on Zoom: GamePack.zip

Open index.html to play on your browser



First steps into scientific research

Teacher: Jelena Budimirović
Teacher: Maja Đukanović
Teacher: Jelena Gledić
Teacher: Maja Miličević Petrović
Teacher: Novella Tedesco

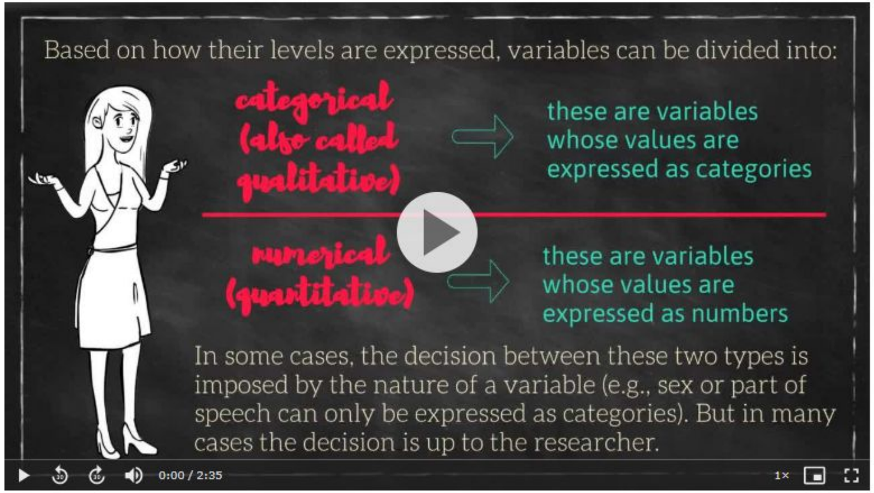
This 3 ECTS learning block (with a possibility of adding a 1 ECTS project) introduces the basic concepts of scientific research, outlining the main approaches in science in general and the key steps in the research process as applied to language. After completing the learning block, the student will be able to:

- approach a subject from a scientific perspective
- evaluate the methodology of scientific articles
- relate a research question to a scientific theory and a research experiment
- design and setup a simple research proposal and plan

First steps into scientific research

Benefit from existing materials - Movetia online course *Revisiting research training in linguistics: theory, logic, method*

Definitions with examples that can be used individually as needed



Based on how their levels are expressed, variables can be divided into:

categorical (also called qualitative) → these are variables whose values are expressed as categories

numerical (quantitative) → these are variables whose values are expressed as numbers

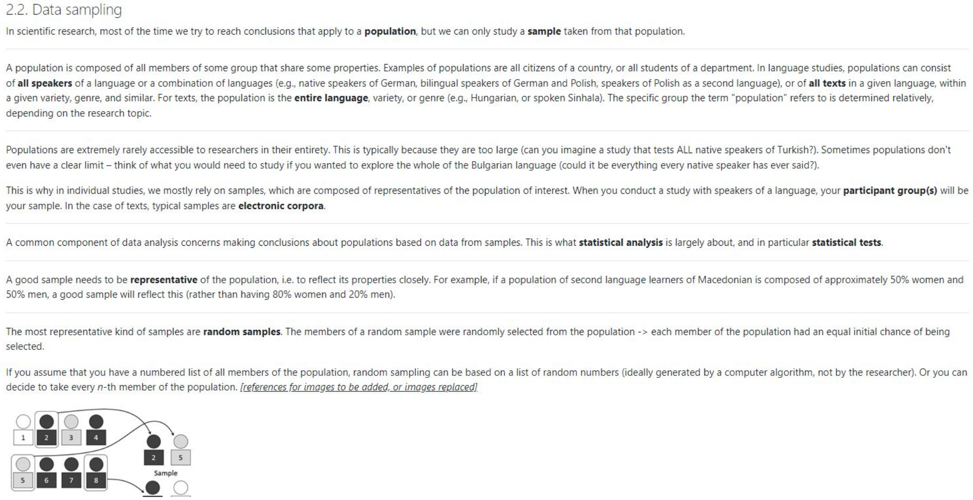
In some cases, the decision between these two types is imposed by the nature of a variable (e.g., sex or part of speech can only be expressed as categories). But in many cases the decision is up to the researcher.

MC2019 Unit 06_03

Revisiting research training in linguistics: theory, logic, method

Share Embed

Universität Zürich^UZH



2.2. Data sampling

In scientific research, most of the time we try to reach conclusions that apply to a **population**, but we can only study a **sample** taken from that population.

A population is composed of all members of some group that share some properties. Examples of populations are all citizens of a country, or all students of a department. In language studies, populations can consist of **all speakers** of a language or a combination of languages (e.g., native speakers of German, bilingual speakers of German and Polish, speakers of Polish as a second language), or of **all texts** in a given language, within a given variety, genre, and similar. For texts, the population is the **entire language**, variety, or genre (e.g., Hungarian, or spoken Sinhala). The specific group the term "population" refers to is determined relatively, depending on the research topic.

Populations are extremely rarely accessible to researchers in their entirety. This is typically because they are too large (can you imagine a study that tests ALL native speakers of Turkish?). Sometimes populations don't even have a clear limit – think of what you would need to study if you wanted to explore the whole of the Bulgarian language (could it be everything every native speaker has ever said?).

This is why in individual studies, we mostly rely on samples, which are composed of representatives of the population of interest. When you conduct a study with speakers of a language, your **participant group(s)** will be your sample. In the case of texts, typical samples are **electronic corpora**.

A common component of data analysis concerns making conclusions about populations based on data from samples. This is what **statistical analysis** is largely about, and in particular **statistical tests**.

A good sample needs to be **representative** of the population, i.e. to reflect its properties closely. For example, if a population of second language learners of Macedonian is composed of approximately 50% women and 50% men, a good sample will reflect this (rather than having 80% women and 20% men).

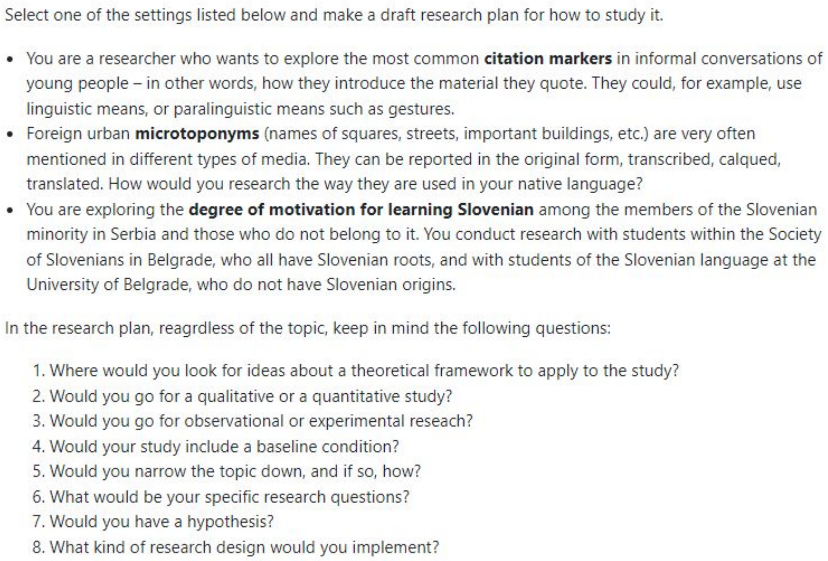
The most representative kind of samples are **random samples**. The members of a random sample were randomly selected from the population -> each member of the population had an equal initial chance of being selected.

If you assume that you have a numbered list of all members of the population, random sampling can be based on a list of random numbers (ideally generated by a computer algorithm, not by the researcher). Or you can decide to take every *n*-th member of the population. *[references for images to be added, or images replaced]*

First steps into scientific research

Practical exercises, for example:

Making a research plan



Select one of the settings listed below and make a draft research plan for how to study it.

- You are a researcher who wants to explore the most common **citation markers** in informal conversations of young people – in other words, how they introduce the material they quote. They could, for example, use linguistic means, or paralinguistic means such as gestures.
- Foreign urban **microtoponyms** (names of squares, streets, important buildings, etc.) are very often mentioned in different types of media. They can be reported in the original form, transcribed, calqued, translated. How would you research the way they are used in your native language?
- You are exploring the **degree of motivation for learning Slovenian** among the members of the Slovenian minority in Serbia and those who do not belong to it. You conduct research with students within the Society of Slovenians in Belgrade, who all have Slovenian roots, and with students of the Slovenian language at the University of Belgrade, who do not have Slovenian origins.

In the research plan, reagrdless of the topic, keep in mind the following questions:

1. Where would you look for ideas about a theoretical framework to apply to the study?
2. Would you go for a qualitative or a quantitative study?
3. Would you go for observational or experimental reseach?
4. Would your study include a baseline condition?
5. Would you narrow the topic down, and if so, how?
6. What would be your specific research questions?
7. Would you have a hypothesis?
8. What kind of research design would you implement?

Critically thinking about the relevance of cultural context in scientific research

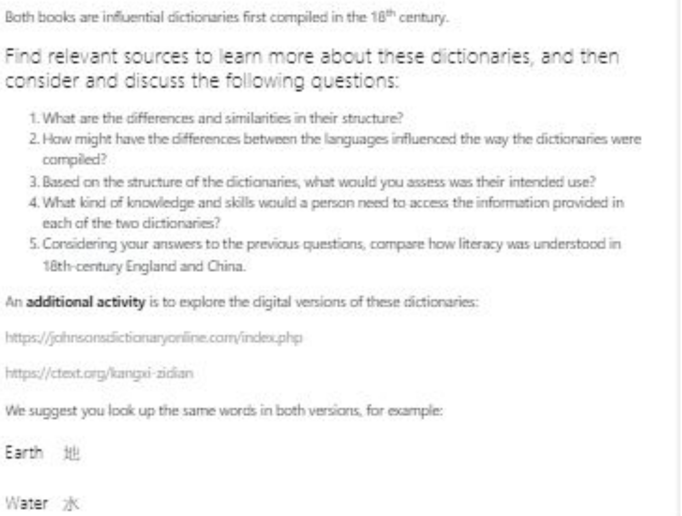
Consider two books:

A Dictionary of the English Language, compiled by Samuel Johnson

Image: Samuel Johnson, Public domain, via Wikimedia Commons

The *Kangxi Dictionary* of the Chinese language, commissioned by Emperor Kangxi and compiled by tens of scholars

Image: Malcolm I'Anson, Public domain, via Wikimedia Commons



Both books are influential dictionaries first compiled in the 18th century.

Find relevant sources to learn more about these dictionaries, and then consider and discuss the following questions:

1. What are the differences and similarities in their structure?
2. How might have the differences between the languages influenced the way the dictionaries were compiled?
3. Based on the structure of the dictionaries, what would you assess was their intended use?
4. What kind of knowledge and skills would a person need to access the information provided in each of the two dictionaries?
5. Considering your answers to the previous questions, compare how literacy was understood in 18th-century England and China.

An **additional activity** is to explore the digital versions of these dictionaries:

https://johnsonsdictionaryonline.com/index.php

https://ctext.org/kangxi-zidian

We suggest you look up the same words in both versions, for example:

Earth 地

Water 水

Co-funded by the
Erasmus+ Programme
of the European Union

Collecting data from human participants

Tihana Kraš (University of Rijeka)
Marko Simonović (University of Graz)

The Third UPSKILLS Multiplier Event, Utrecht, 4 November 2022

Learning content unit block basic information

Workload: 6 ECTS

Designers: Tihana Kraš*, Martina Podboj*, Marko Simonović †

Designers' affiliation: *University of Rijeka, † University of Graz

Thematically, the unit block is structured in four parts:

1. General
2. Morphophonology and morphosyntax
3. Second language acquisition (SLA)
4. Sociolinguistics

+ Two student projects (3 ECTS)

Topics covered (subunits)

GENERAL (Part 1)

- Relevant terms and distinctions
- Ethics in linguistic research with human participants
- Population sampling in linguistic research

MORPHOPHONOLOGY AND MORPHOSYNTAX (Part 2)

- Judgement data
- Judgement tasks
- Creating an experiment/survey
- Data elicitation

Topics covered (subunits)

SECOND LANGUAGE ACQUISITION (Part 3)

- Comprehension tasks in SLA
- Elicited production tasks in SLA
- Acceptability judgement tasks in SLA

SOCOLINGUISTICS (Part 4)

- Ethnographic fieldwork in sociolinguistics
- The sociolinguistic interview
- Surveys and questionnaires in socolinguistic research

Student projects

- “Irregular” phonological alternations
- Second language acquisition of English morphosyntax

General subunit structure

Preparatory reading assignment or a warm-up activity

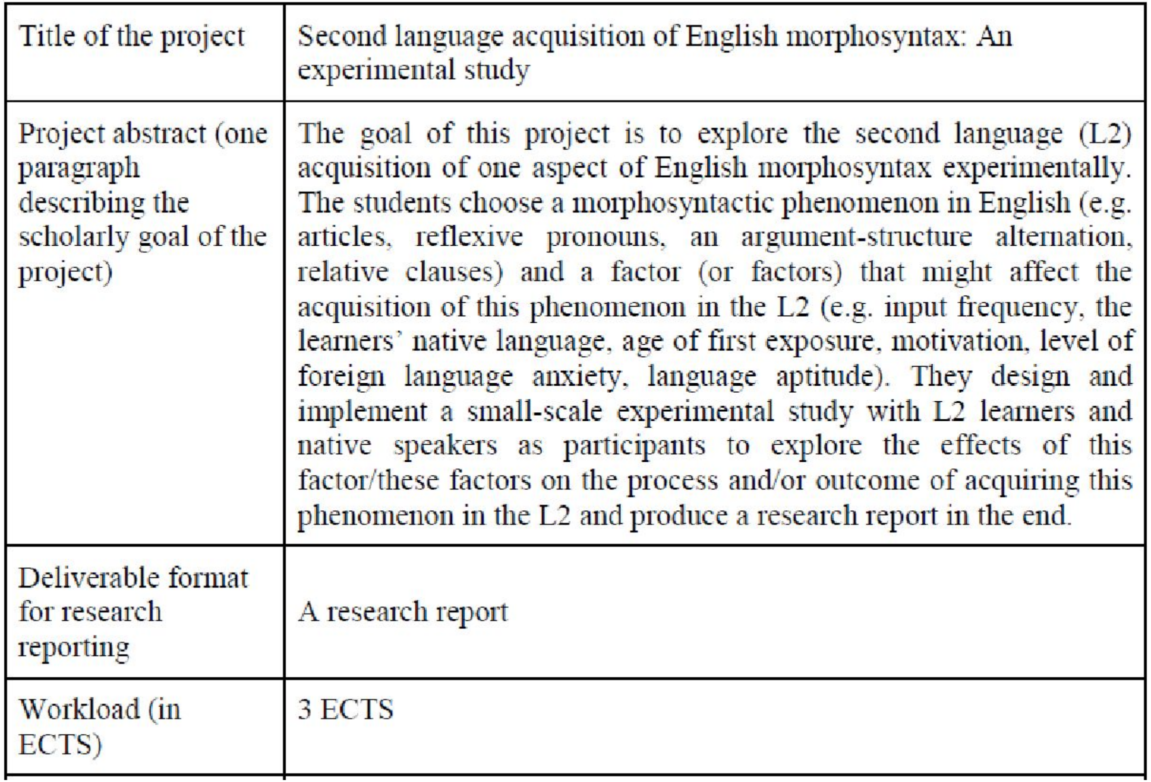
PPT presentation with notes for teachers accompanied by handouts for students

Optional reading assignment

Practical assignment

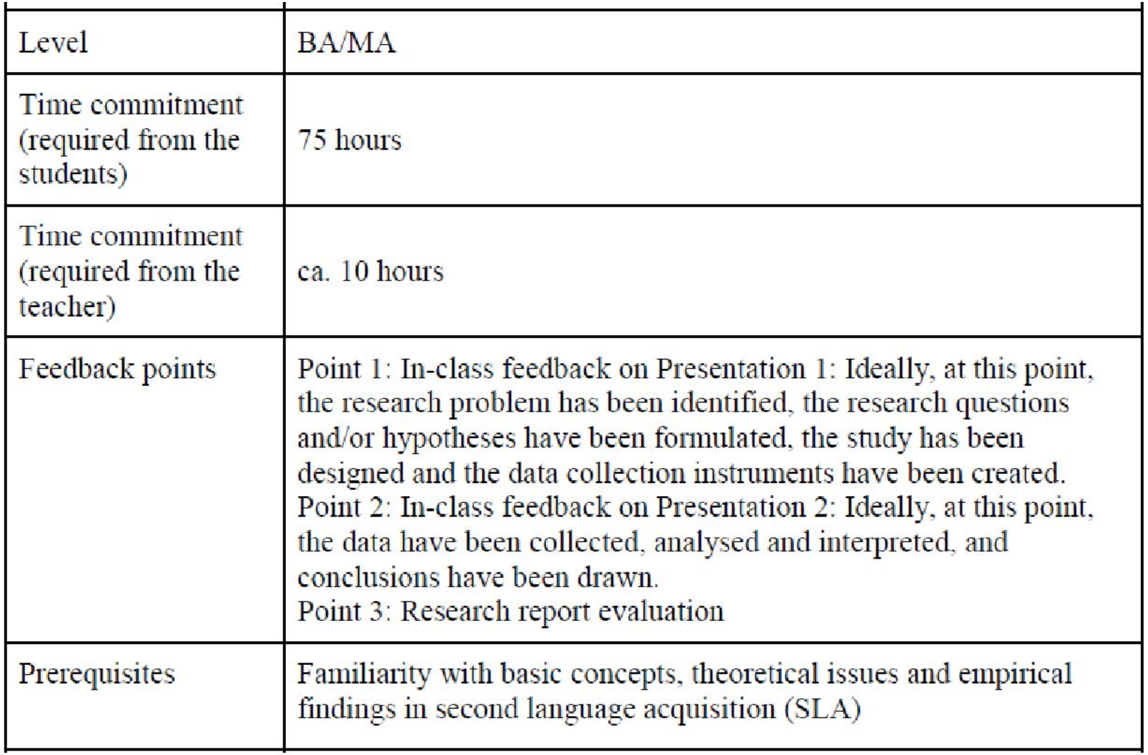
Quizz

Example of a student project



Title of the project	Second language acquisition of English morphosyntax: An experimental study
Project abstract (one paragraph describing the scholarly goal of the project)	The goal of this project is to explore the second language (L2) acquisition of one aspect of English morphosyntax experimentally. The students choose a morphosyntactic phenomenon in English (e.g. articles, reflexive pronouns, an argument-structure alternation, relative clauses) and a factor (or factors) that might affect the acquisition of this phenomenon in the L2 (e.g. input frequency, the learners' native language, age of first exposure, motivation, level of foreign language anxiety, language aptitude). They design and implement a small-scale experimental study with L2 learners and native speakers as participants to explore the effects of this factor/these factors on the process and/or outcome of acquiring this phenomenon in the L2 and produce a research report in the end.
Deliverable format for research reporting	A research report
Workload (in ECTS)	3 ECTS

Example of a student project



Level	BA/MA
Time commitment (required from the students)	75 hours
Time commitment (required from the teacher)	ca. 10 hours
Feedback points	Point 1: In-class feedback on Presentation 1: Ideally, at this point, the research problem has been identified, the research questions and/or hypotheses have been formulated, the study has been designed and the data collection instruments have been created. Point 2: In-class feedback on Presentation 2: Ideally, at this point, the data have been collected, analysed and interpreted, and conclusions have been drawn. Point 3: Research report evaluation
Prerequisites	Familiarity with basic concepts, theoretical issues and empirical findings in second language acquisition (SLA)

UP SKILLS

Co-funded by the Erasmus+ Programme of the European Union

A glimpse into language data science

Maja Miličević Petrović,
Adriano Ferraresi, Paul Marty

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

L-Università ta' Malta

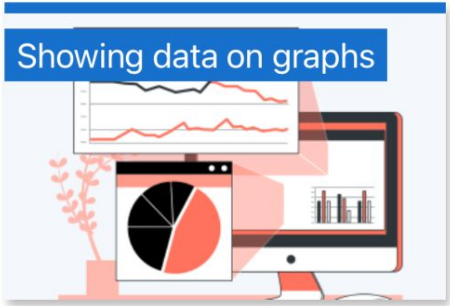
Guidelines and Best Practices for Research-Based Teaching
Utrecht, 4 November 2022

Outline

Statistics 101

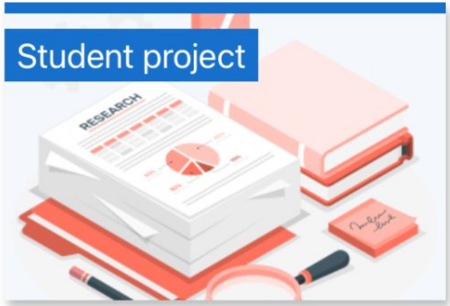
Working with R

Calculating summary numbers


Showing data on graphs

The logic behind inferential statistics

Some simple statistical tests


Student project

🧑‍🏫 **Key concepts**

🧑‍🎓 **Activities**

💾 **Data and code**

Formats

- **Theoretical and methodological contents**
  ⇒ Moodle books

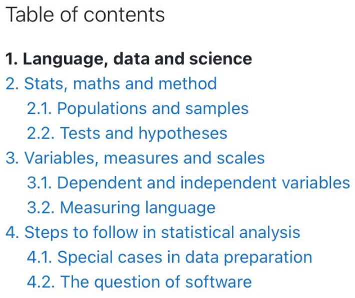
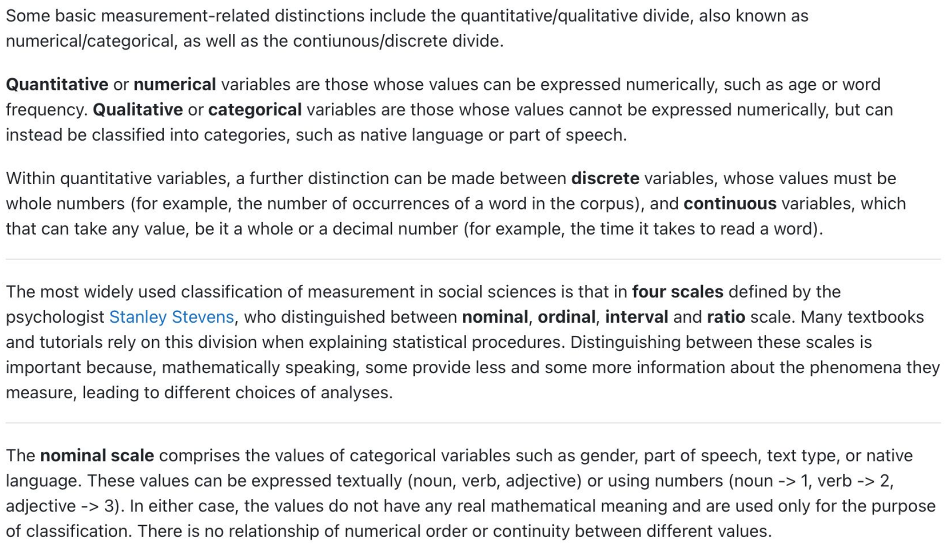


Table of contents



Some basic measurement-related distinctions include the quantitative/qualitative divide, also known as numerical/categorical, as well as the contiunous/discrete divide.

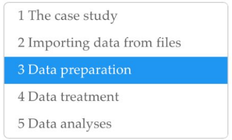
Quantitative or **numerical** variables are those whose values can be expressed numerically, such as age or word frequency. **Qualitative** or **categorical** variables are those whose values cannot be expressed numerically, but can instead be classified into categories, such as native language or part of speech.

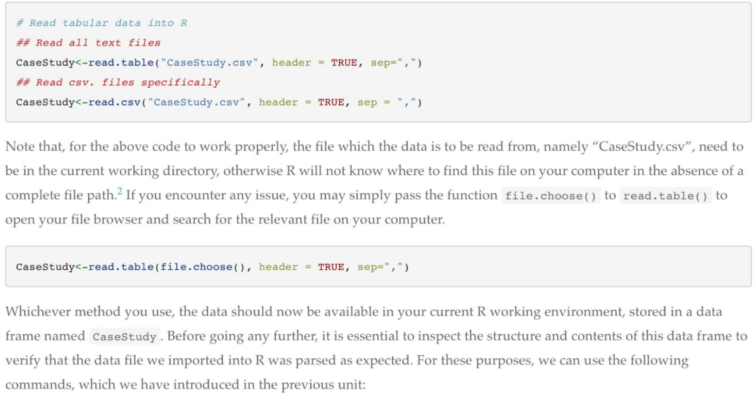
Within quantitative variables, a further distinction can be made between **discrete** variables, whose values must be whole numbers (for example, the number of occurrences of a word in the corpus), and **continuous** variables, which that can take any value, be it a whole or a decimal number (for example, the time it takes to read a word).

The most widely used classification of measurement in social sciences is that in **four scales** defined by the psychologist Stanley Stevens, who distinguished between **nominal**, **ordinal**, **interval** and **ratio** scale. Many textbooks and tutorials rely on this division when explaining statistical procedures. Distinguishing between these scales is important because, mathematically speaking, some provide less and some more information about the phenomena they measure, leading to different choices of analyses.

The **nominal scale** comprises the values of categorical variables such as gender, part of speech, text type, or native language. These values can be expressed textually (noun, verb, adjective) or using numbers (noun -> 1, verb -> 2, adjective -> 3). In either case, the values do not have any real mathematical meaning and are used only for the purpose of classification. There is no relationship of numerical order or continuity between different values.

- **Exercises**
  ⇒ R scripts / R markdown





```
# Read tabular data into R
## Read all text files
CaseStudy<-read.table("CaseStudy.csv", header = TRUE, sep=",")
## Read csv. files specifically
CaseStudy<-read.csv("CaseStudy.csv", header = TRUE, sep = ",")
```

Note that, for the above code to work properly, the file which the data is to be read from, namely "CaseStudy.csv", need to be in the current working directory, otherwise R will not know where to find this file on your computer in the absence of a complete file path.[2] If you encounter any issue, you may simply pass the function `file.choose()` to `read.table()` to open your file browser and search for the relevant file on your computer.

```
CaseStudy<-read.table(file.choose(), header = TRUE, sep=",")
```

Whichever method you use, the data should now be available in your current R working environment, stored in a data frame named `CaseStudy`. Before going any further, it is essential to inspect the structure and contents of this data frame to verify that the data file we imported into R was parsed as expected. For these purposes, we can use the following commands, which we have introduced in the previous unit:

(Re)use

The materials are...

- Downloadable
- Modifiable
- Modular

Different levels of (re)use possible:

- Full course
- Course units
- Books
  - Book chapters
  - Individual paragraphs
- Exercises