

# Digital Humanities

Continuing our project simulation

Matteo Di Cristofaro

UniMoRe

15/04/2024

**In the news**

# An update on AI in EU

**NOTE: the paper is a pre-print!**

As powerful LLMs like GPT-4 and Gemini, image and video generators rise, their very momentum throws into stark relief the question of the adequacy of existing and forthcoming EU legislation. In this article, we discuss some key legal and regulatory concerns brought up by Generative AI and LLMs regarding liability, privacy, intellectual property, and cybersecurity. The EU's response to these concerns should be contextualised within the guidelines of the Artificial Intelligence Act (AIA), which comprehensively addresses the design, development, and deployment of AI models, including Generative AI within its scope. Where we identify gaps or flaws in the EU legislation, we offer some recommendations to ensure that Generative AI models evolve lawfully.

Generative AI in EU Law: Liability, Privacy, Intellectual Property, and Cybersecurity

# Our project

# Simulating a project

We are going to develop a project from start to finish, using Youtube videos - in the form of subtitle tracks - as source data.

To do so, we are going to follow these steps:

1. ~~Define a topic and a purpose~~
2. ~~Select the Youtube subtitles data to be used as source~~
3. ~~Collect the data and metadata~~
4. ~~Select the metadata (included in the Youtube data) that will support our purpose~~
5. ~~Create the corpus files~~
6. Compile the corpus in SketchEngine
7. Explore the corpus through SketchEngine

# 1. Aims/purposes/research questions

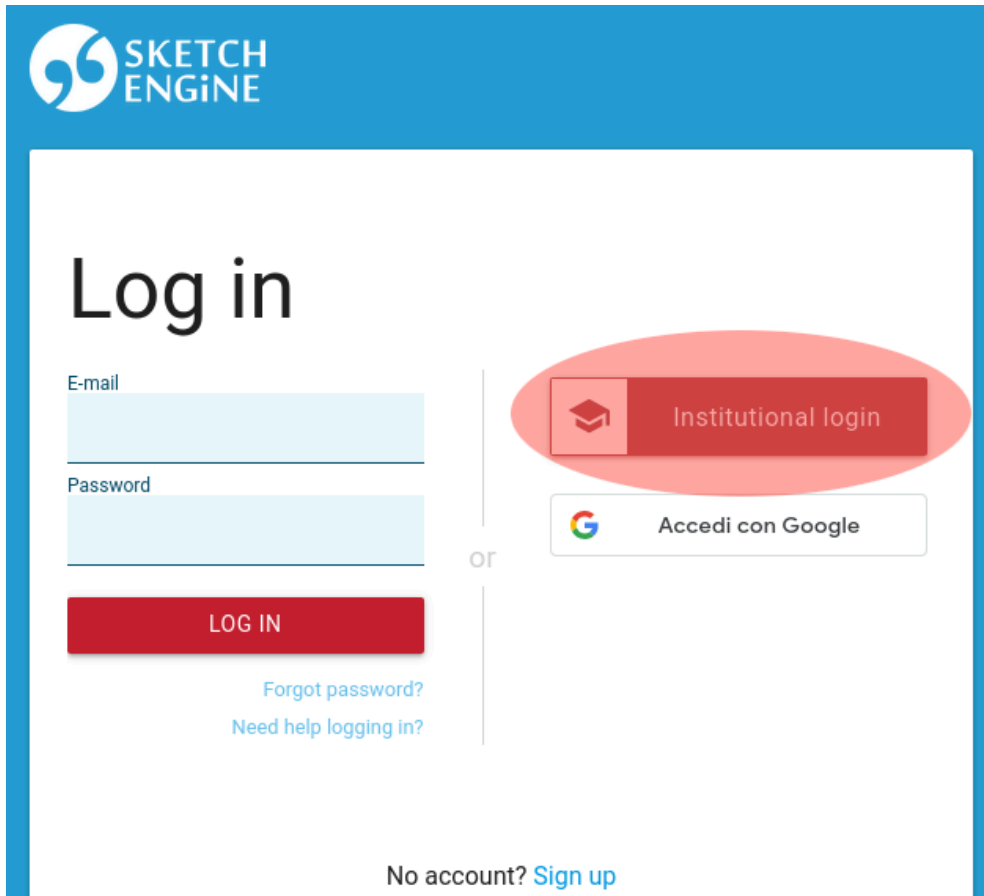
- do they only review food, or also location/bill?
- adjectives used to describe waiters, to understand problems related with rest. service -politeness
- strategies used to promote restaurants that sponsored the video
- allergies information/details
- how do staff/manager react to unfamiliar situations?
- vegan/vegetarian information/details
- use of specialised or general language/terms/etc...
- use of evaluative language

# 5. Create the corpus files

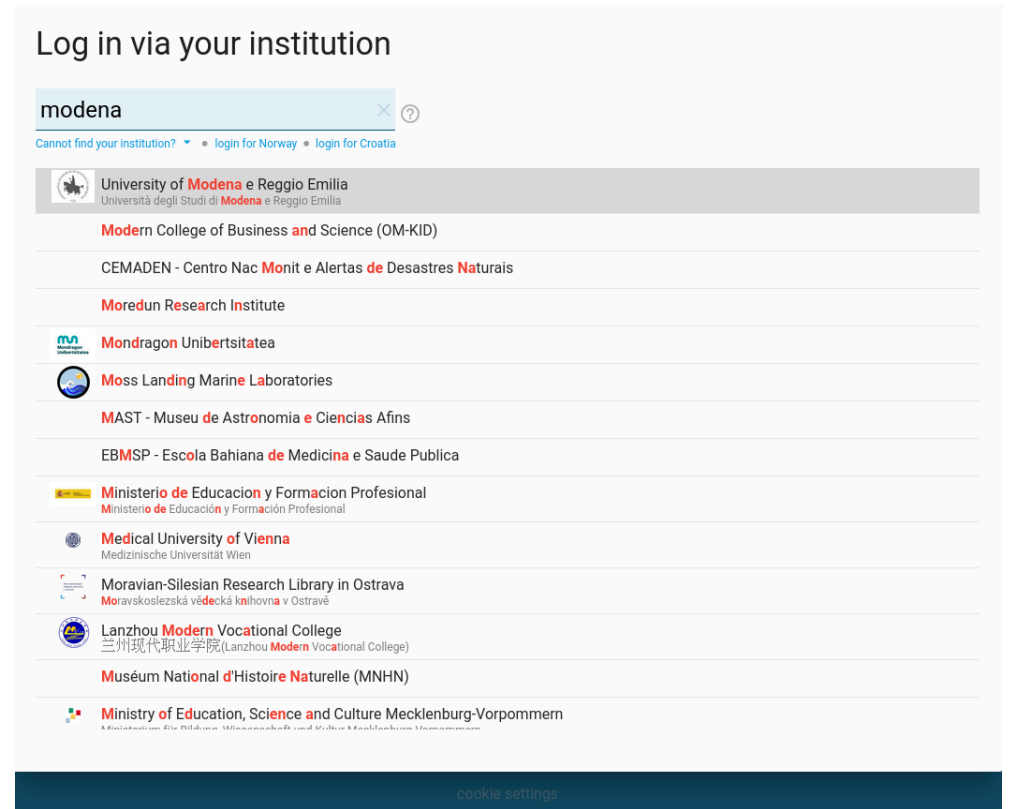
Let's see some more details about how the corpus was created...

# 6. Compile the corpus in SketchEngine

Select *Institutional login* from <https://app.sketchengine.eu>, then choose *University of Modena e Reggio Emilia*, then login using your UniMoRe credentials.



The image shows the SketchEngine login interface. At the top left is the SketchEngine logo. The main heading is "Log in". Below it are two input fields: "E-mail" and "Password". To the right of these fields is a red button labeled "Institutional login", which is highlighted with a red oval. Below the "Institutional login" button is a Google login button labeled "Accedi con Google". At the bottom of the login section is a red button labeled "LOG IN". Below the "LOG IN" button are two links: "Forgot password?" and "Need help logging in?". At the very bottom of the page is a link: "No account? Sign up".





The image shows the "Log in via your institution" page. At the top is the heading "Log in via your institution". Below it is a search bar containing the text "modena". To the right of the search bar are two icons: a close icon (X) and a help icon (?). Below the search bar are two links: "Cannot find your institution?" and "login for Norway" and "login for Croatia". Below these links is a list of institutions. The first institution is "University of Modena e Reggio Emilia", which is highlighted with a grey background. Below it are several other institutions, including "Modern College of Business and Science (OM-KID)", "CEMADEN - Centro Nazionale di Monitoraggio e Allerta per i Disastri Naturali", "Moredun Research Institute", "Mondragon Unibertsitatea", "Moss Landing Marine Laboratories", "MAST - Museo de Astronomia e Ciencias Afins", "EBMSP - Escola Bahiana de Medicina e Saude Publica", "Ministerio de Educacion y Formacion Profesional", "Ministerio de Educacion y Formacion Profesional", "Medical University of Vienna", "Moravian-Silesian Research Library in Ostrava", "Lanzhou Modern Vocational College", "Muséum National d'Histoire Naturelle (MNHN)", and "Ministry of Education, Science and Culture Mecklenburg-Vorpommern". At the bottom of the page is a link: "cookie settings".





# 6. Compile the corpus in SketchEngine /2


Now, from the dashboard, select *NEW CORPUS* in the top-right section, and then follow the step-by-step creation process until...


**DASHBOARD**   


**TESTYT\_DT** **CORPUS INFO** **MANAGE CORPUS**


 **Word Sketch**  
Collocations and word combinations


 **Thesaurus**  
Synonyms and similar words


 **Parallel Concordance**  
Translation search


 **N-grams**  
Multiword expressions (MWEs)


 **Trends**  
Diachronic analysis, neologisms


 **OneClick Dictionary**  
Automatic dictionary drafting


 **Word Sketch Difference**  
Compare collocations of two words

 **Concordance**  
Examples of use in context





 **Wordlist**  
Frequency list

 **Keywords**  
Terminology extraction

 **Text type analysis**  
Statistics of the whole corpus


 **Bilingual terms**  
Bilingual terminology extraction

**RECENTLY USED CORPORA** **NEW CORPUS**

testyt_dt	English	3,415	
testyt	English	3,415	
British Academic Written English Corpus (BAWE)	English	6,968,089	
testeer	English	130	

**boot camp** A **face-to-face** course in using Sketch Engine.  
Brno, CZ, 26–27 April 2023  
**REGISTRATION**

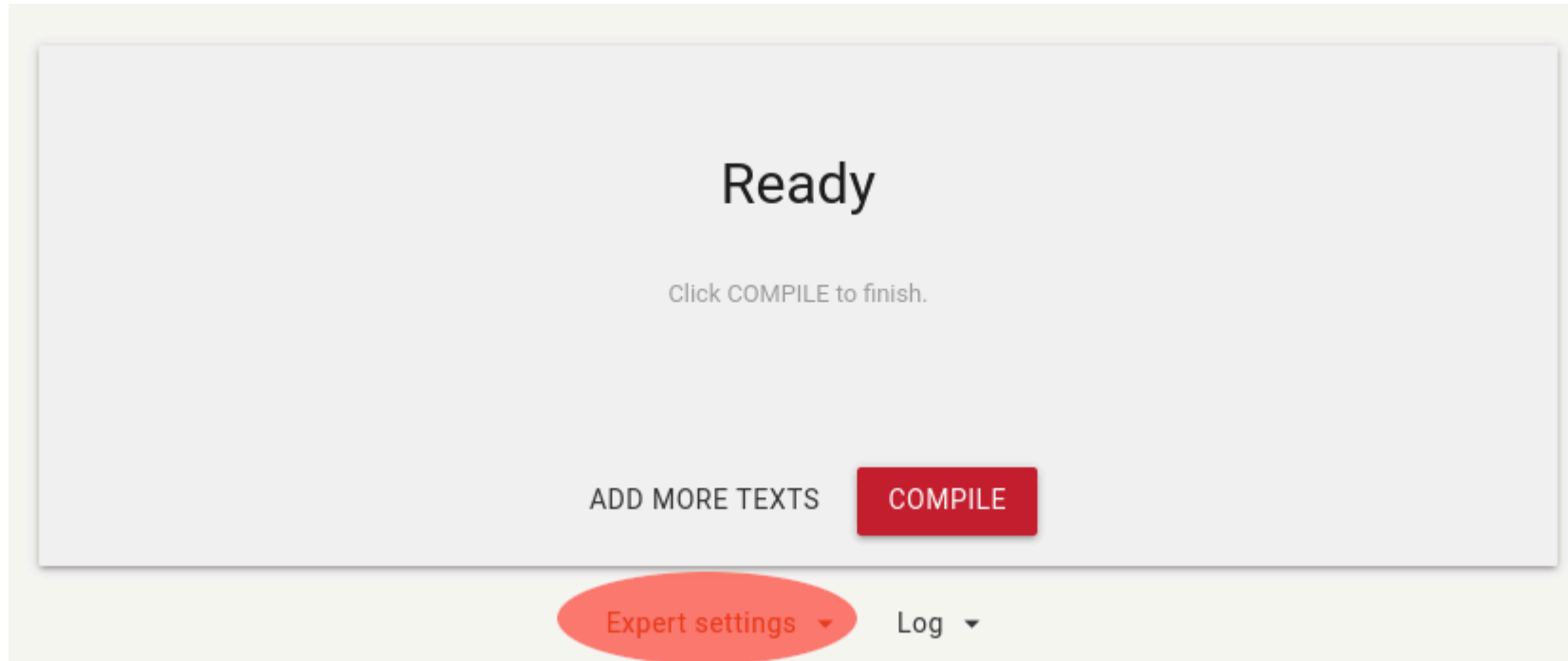
**MY SEARCH HISTORY** **ANNOTATIONS**



Only favourites ☐ Only history ☐

## 6. Compile the corpus in SketchEngine /3

When you reach this step, **click on Expert settings**



## 6. Compile the corpus in SketchEngine /3

When you reach this step, **click on Expert settings** and then **select all the metadata attributes you want to use in your corpus**.

### EXPERT SETTINGS

Duplicated content ?

☐ Remove duplicated content

Structures and attributes to keep ?

all | none

✓ s (504)

✓ p (500) —

✓ is\_ac

✓ time

✓ g (382)

✓ text (3) —

✓ date\_ts

✓ date\_y

✓ format

✓ likes

✓ title

## 6. Compile the corpus in SketchEngine /4

Once finished, leave all the other options untouched, and click on SAVE AND COMPILE.

Term grammar ?

- ☒ English terms 3.1 ?
- ☐ English terms 3.0 ?
- ☐ English (TreeTagger - PennTB) for term extraction 2.3 ?
- ☐ None (no term extraction)

+

Structure name for documents ? doc

Structure for document counts ? Same as name for documents ▼

CANCEL SAVE AND COMPILE

# Corpus and Corpus Linguistics

# Defining Corpus Linguistics (CL)

"a group of methods that use specialist computer programmes to study language in large bodies of machine-readable text" (Brookes and McEnery 2020:378)

# Defining Corpus Linguistics (CL)

"a group of methods that use specialist computer programmes to study language in large bodies of machine-readable text" (Brookes and McEnery 2020:378)

CL is commonly associated with a *quantitative*-only perspective on language

but is it really so?

# More than numbers

“What started as a methodological enhancement but included a quantitative explosion (I am referring here to the quantity of data processed thanks to the aid of the computer) has turned out to be a theoretical and qualitative revolution in that it has offered insights into the language that have shaken the underlying assumptions behind many well-established theoretical positions in the field.” (Tognini Bonelli 2012:17)

## How so?



# More than numbers /2

CL is based on the assumption – corroborated by more than 60 years of research – that a link exists between **frequency of repetition** of words and the **meaning/connotation** they have in a language.

**Frequency** can therefore indicate the level of **conventionalisation** a pair of form and meaning has in a language or a discourse.

# The "building blocks" of CL: a gentle introduction / a quick 'refresher'

Type, token, lemma

Frequencies and frequencies lists

Dispersion

Concordances and Key-Word-In-Context (KWIC)

Collocations and N-grams

Keywords

# Type, token, lemma

A **type** is the single wordform of each word in a corpus, while **token** refers to an occurrence of a type in a corpus: hence in the text

Computer science is the study of computation, automation, and information. Computer science spans theoretical disciplines (such as algorithms, theory of computation, and information theory) to practical disciplines (including the design and implementation of hardware and software). Computer science is generally considered an area of academic research and distinct from computer programming.

the word ‘computer’ counts as one type occurring four times - hence four tokens (even though...).

**Lemma** refers instead to a categorisation that groups “wordforms that are related by being inflectional forms of the same base word” (McEnery and Hardie 2012:245), and may also be called the ‘root’ form of a word: ‘computers’ is the plural inflectional form of the lemma ‘computer’.

# Frequencies and frequencies lists

Frequencies, as the name suggests, indicate how many times a set of words appears in a corpus. They can be **absolute** (also called raw) or **normalised**: the former expresses the total number of occurrences each element has in the corpus, while the latter are adjusted to reflect the weight each element has in the whole using e.g. a basis of 1 million words as reference – a common baseline in corpus linguistics (cf. Brezina 2018:43).

Frequencies normalised per million words (PMW) are obtained by dividing the number of occurrences of an element by the total number of elements contained in the corpus – in the case of words, this is the total number of tokens – and multiplying this value by 1 million. Normalised frequencies are particularly relevant when **comparisons** across different (sub-)corpora need to be made as they allow to evaluate the weight each word has in its respective corpus.

# Dispersion

Dispersion allows the researcher to understand in how many parts of a corpus (e.g. documents) an element appears.

“Imagine you have a corpus which contains a substantial sample from a book on whelks (small sea creatures somewhat similar to snails). Naturally, the word whelk will appear many times in this book because it is about whelks. In general English, however, whelk is not a particularly common word because it is specific to a single genre/register – books and articles on sea life. However, here comes the problem: when we construct a frequency list based on our corpus that includes the book on whelks the word whelk will appear among fairly frequent items by virtue of being repeated many times in a single text. If we base our investigation on word frequency alone, our results will be extremely misleading.” (Brezina 2018:47)

Therefore frequency answers the question “how often does x happen?” whereas dispersion asks “in how many contexts will you encounter x at all?” (Gries and Ellis 2015:232)

So, **frequency should always be paired with dispersion!**

# Concordances and Key-Word-In-Context (KWIC)

A concordance is a snippet of text in which the searched character string (a word, a pattern of words, a part-of-speech tag, or any other element a corpus contains) is shown in its original context, with some of the original textual context on its left and on its right. Concordances are obtained through Key-Word-In-Context (KWIC) concordancing – a term that is oftentimes used in corpus tools as label to access this functionality. Concordances are the linguists’ “weapon of first and last resort [...], the place where quantification and interpretation meet” (Hardie 2017).

# Collocations and N-grams

**Collocations** are “co-occurrence patterns observed in corpus data” (McEnery and Hardie 2012:123), i.e. combinations of two (or more) elements that exhibit a certain degree of attraction (e.g. conventionalisation) in the corpus under scrutiny.

# Collocations and N-grams

**Collocations** are “co-occurrence patterns observed in corpus data” (McEnery and Hardie 2012:123), i.e. combinations of two (or more) elements that exhibit a certain degree of attraction (e.g. conventionalisation) in the corpus under scrutiny.

For collocations, we can specify a **span** indicating the width of the collocation window: this is the maximum number of elements to the left and to the right of the **node** (i.e. the searched element) the software should consider when scouting for potential candidates. This information is commonly written through the notation  $-/+5$  or  $5L, 5R$ .



# Collocations and N-grams /2

**N-grams** are distinct from collocations because in the latter the collocates have no fixed order/structure, while in the former the elements always appear in the same order.

# Keywords

A keyword is an element (e.g. a word or a part-of-speech tag) derived from a quantitative comparison between two datasets, linguistically expressing the characteristics of a collection of texts (the **analysis** or **target** corpus) when compared with another collection of texts used as baseline for comparisons (the **reference** corpus) using *keyness metrics*.

Keywords are defined as **positive keywords** when they are key in the analysis corpus - i.e. when they are more *exclusive* of the language represented by the analysis corpus -, and **negative keywords** when they are key in the reference corpus. Between these two extremes, **lockwords** indicate those keywords whose keyness is (quasi-)equal in the two corpora.

## 7. Explore the corpus through SketchEngine: now what?

- Rainbow washing: do fashion companies communicate differently about LGBTQI+ topics during Pride Month?

# 7. Explore the corpus through SketchEngine: now what?

- Rainbow washing: do fashion companies communicate differently about LGBTQI+ topics during Pride Month?

Find all adjectives in videos posted in June (any year)

```
[tag="J.*"] within <text date_m="6" />
```

Find all adjectives in videos posted in any other month (but not June)

```
[tag="J.*"] within <text date_m!="6" />
```

# 7. Explore the corpus through SketchEngine: now what?

- How do Youtubers communicate in the first minute of their videos?

# 7. Explore the corpus through SketchEngine: now what?

- How do Youtubers communicate in the first minute of their videos?

Find all verbs used in subs appearing in the first minute of video:

```
[tag="V.*"] within <p time<="60000" />
```

# 7. Explore the corpus through SketchEngine: now what?

Other queries possibilities:

Find all words in subtitles written by humans:

```
[tag=".*"] within <text autocaption="false" />
```

Find all nouns in videos with at least 2.000 likes:

```
[tag="V.*"] within <text likes>="2000" />
```

Find all adjectives appearing in the first minute of videos posted in June (any year)

```
[tag="J.*"] within (<p time<="60000" /> within <text date_m="6" />)
```

# Unix time for corpus queries

Going back to our previous question about time periods, here are two examples using Unix time in SketchEngine CQL queries.

Find all adjectives in videos created between 22nd May 2022 and 7th June 2022

```
[tag="J.*"] within (<text date_ts>="1653170400" & date_ts<="1654552800" />)
```

Find all verbs in videos created NOT before Fri Aug 21 2020 23:58:20 UTC+0200 (Central European Summer Time)

```
[tag="V.*"] within <text date_ts!<="1598047100" />
```

Find all adjectives in videos created before 22nd May 2022 and after 7th June 2022

```
[tag="J.*"] within (<text date_ts<="1653170400" & date_ts>="1654552800" />)
```



## 8. Why are we exploring? What are we investigating?

Now that our corpus is ready and we know a bit more about how it can be explored and analysed, let us recap what we want to investigate - and maybe find new aims...

Take notes of the research question(s)/aim(s) you wish to pursue through the just-created corpus in [this shared document](#).

At the same time, explore the corpus to find elements (words, constructions, patterns, etc...) that you deem useful to answer your research question(s).

## 9. What have others investigated?

Now let's use **Google Scholar** to find papers/researches that may provide us with new ideas/hints/tools/methods/practices to improve/support our research questions.

A starting point?

| linguistics restaurant reviews

When you find something interesting or useful, add it to the research question you included in the shared document, explaining why it may be useful.

# Other possible explorations

Beside the functionalities offered by SketchEngine, we may explore our corpus using a different "perspective" such as that offered by **semantic domains**. We are going to employ the **USAS (UCREL Semantic Analysis System)** tagging system, which can be applied to our corpus through **LancsBox X**.

# Enough for today!

## Questions? Doubts?

Contact me: [mdicristofaro@unimore.it](mailto:mdicristofaro@unimore.it)

Office hours: send me an email and we will setup an online "ricevimento".

See you next week!