TermiKnowledge

Insights into data acquisition and data preparation for an online resource on COVID-19 terminology

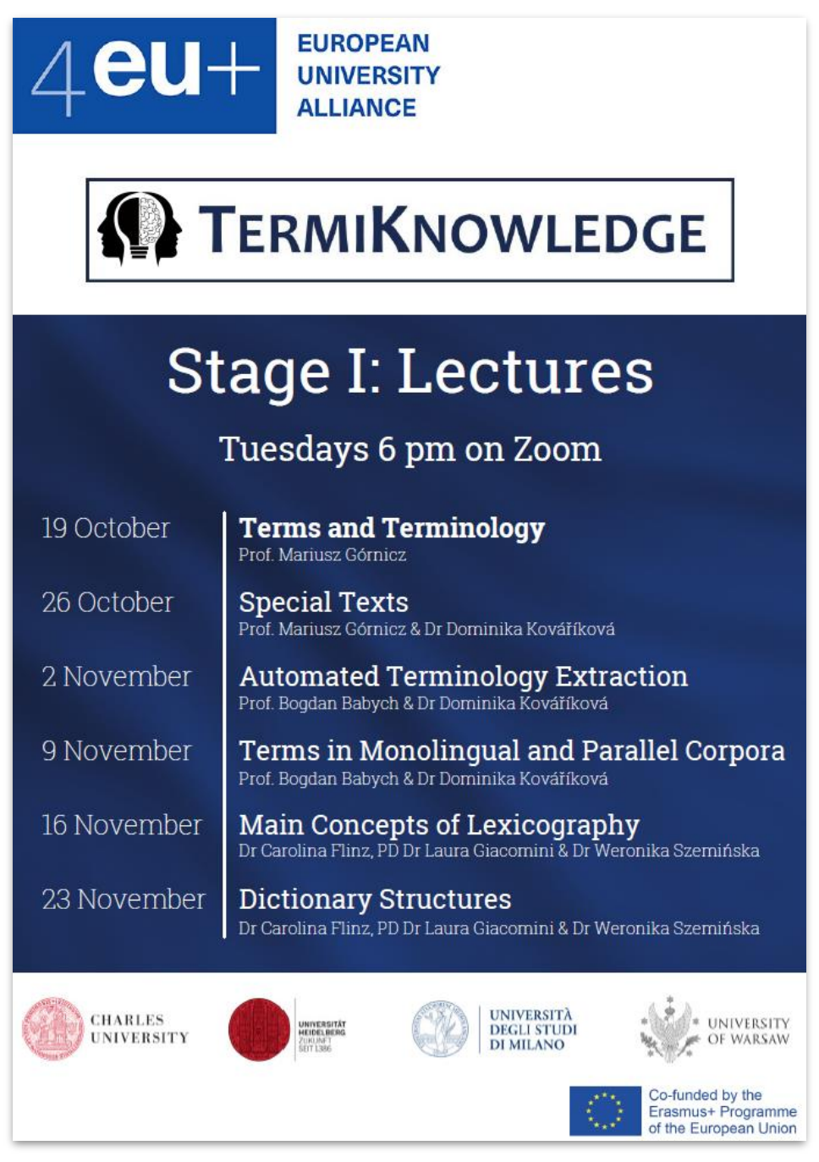
Carolina Flinz (carolina.flinz@unimi.it) – Laura Giacomini (laura.giacomini@iued.uni-heidelberg.de) – Weronika Szemińska (w.szeminska@uw.edu.pl)

OBJECTIVES

- TermiKnowledge is an **international course in creating terminological resources** which took place between October 2021 and June 2022.
- Participants were **BA and MA students in translation, (corpus) linguistics, language and literature**
- The **goal** of TermiKnowledge was to teach students how to create a corpus, extract terms manually and using software, identify their equivalents in other languages, design a multilingual knowledge base and create terminological entries.
- The **4EU+ shared competencies** acquired during the course were:
  - data literacy – corpus work teaches students how to cope with specific terminological problems, use corpus tools and verify data;
  - multilingualism – the project offered an opportunity to compare concept systems and terminology across languages;
  - critical thinking – the students were involved in taking decisions during the entire lexicographic process (extent of the domain, headwords, access structure, entry structure etc.).
- The **multilingual knowledge base has been published on-line** for public use: https://terminology.mimuw.edu.pl/.

ORGANISATION

- TermiKnowledge was run under the **4EU+ Alliance** by an international team of **experts in terminology, corpus linguistics, translation and knowledge representation** from **four European universities** (University of Warsaw – UW, Heidelberg University – HU, Charles University in Prague – CU, and Milan University – MU):
  - Mariusz Górnicz (PI, UW), Weronika Szemińska (UW), Laura Giacomini (HU), Bogdan Babych (HU), Dominika Kováříková (CU), Carolina Flinz (MU), Jerzy Tyszkiewicz (UW)
  - student assistants: Weronika Stefańska (UW), Pia Kruse (HU), Petr Louda (CU), Rita Luppi (MU)
- The project included **two courses**:
  - Semester I from mid-October 2021 to March 2022
  - Semester II from March 2022 to June 2022
- Each course consisted in
  - on-line lectures and research-based classes,
  - collaborative practical tasks on the creation of terminological resources as well as
  - short mobilities involving final works towards publication.
- Assessment was based on portfolios.



4EU+ EUROPEAN UNIVERSITY ALLIANCE

TermiKnowledge

Stage I: Lectures

Tuesdays 6 pm on Zoom

19 October – **Terms and Terminology**
Prof. Mariusz Górnicz

26 October – **Special Texts**
Prof. Mariusz Górnicz & Dr Dominika Kováříková

2 November – **Automated Terminology Extraction**
Prof. Bogdan Babych & Dr Dominika Kováříková

9 November – **Terms in Monolingual and Parallel Corpora**
Prof. Bogdan Babych & Dr Dominika Kováříková

16 November – **Main Concepts of Lexicography**
Dr Carolina Flinz, PD Dr Laura Giacomini & Dr Weronika Szemińska

23 November – **Dictionary Structures**
Dr Carolina Flinz, PD Dr Laura Giacomini & Dr Weronika Szemińska

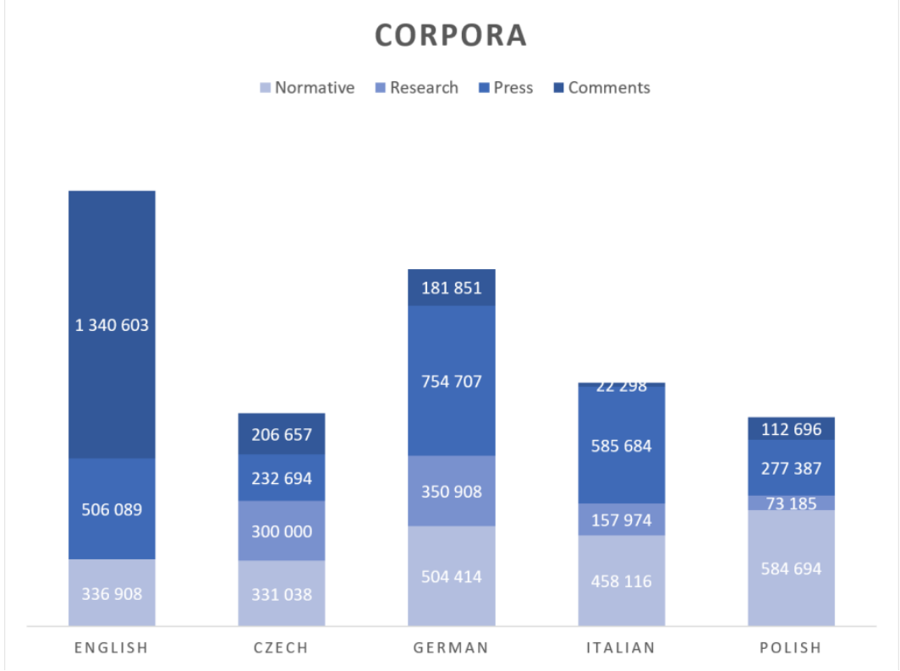
Charles University – Universität Heidelberg Zukunft seit 1386 – Università degli Studi di Milano – University of Warsaw

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

THE LEXICOGRAPHIC PROCESS (SEMESTER I)

a) CORPUS COMPILATION

- five languages: English, Czech, Italian, German, and Polish
- four corpora per language:
  - **Normative** – legal regulations such as EU directives, national regulations, and medical guidelines
  - **Research** – published research papers
  - **Press** – general press articles
  - **Comments** – texts in online comments sections under press articles



* The already existing *Covid-19* corpus available on SketchEngine served as the English Research corpus. With its 224,061,570 words, it was by far the largest corpus in the project – and had to be omitted in the chart.

b) TERMINOLOGY WORK

- Keyword extraction
- Keyword analysis and validation
- Contrastive analysis of keywords
- Headword selection (similar across corpora)
- Data preparation:
  - decision on the form of the headword
  - search for synonyms and variants
  - collocation extraction
  - search for definitions and encyclopaedic information
  - search for examples
  - preparation of notes

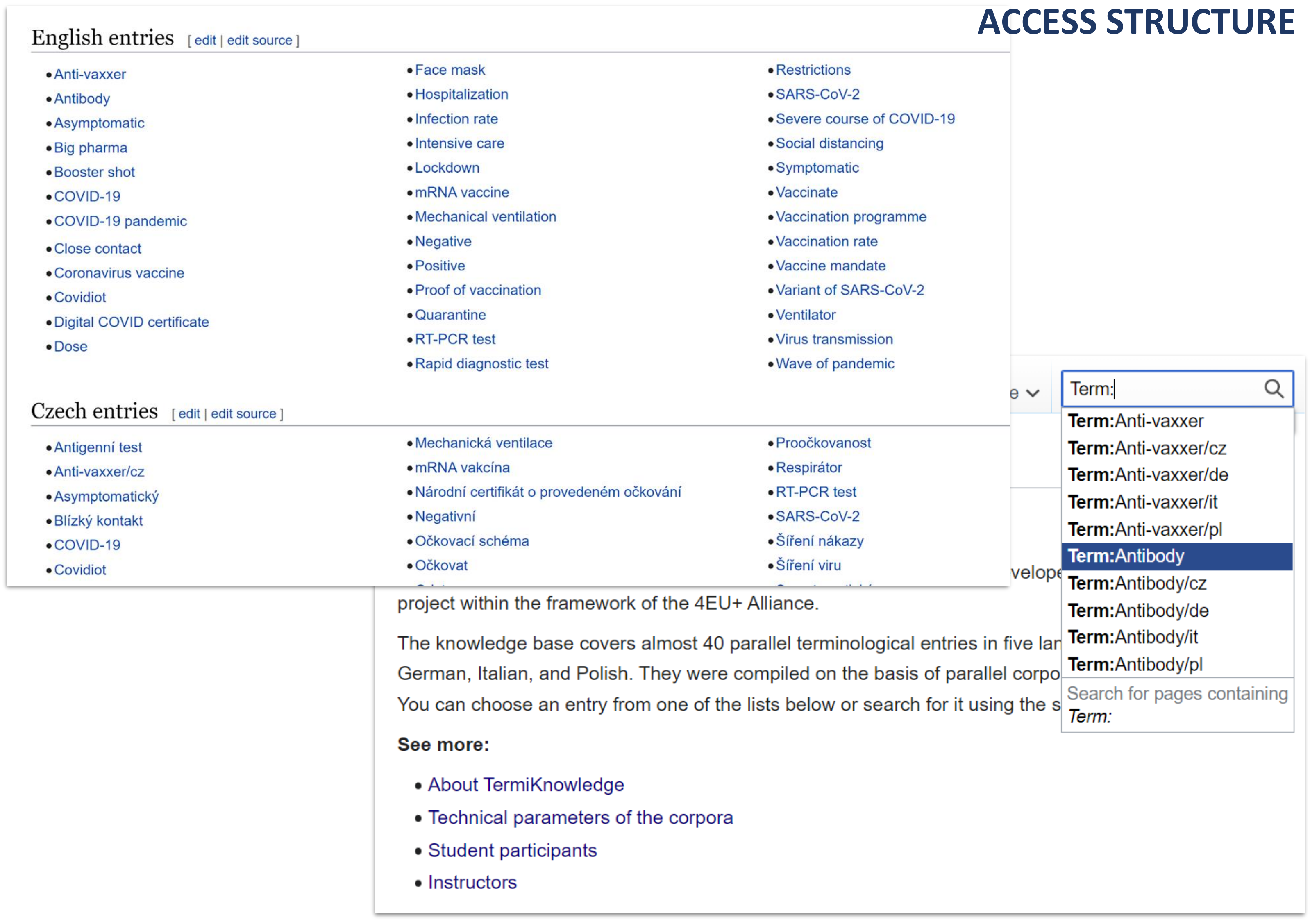
c) ABSTRACT MICROSTRUCTURE

TERM	
Contents	
Related terms	
NORMATIVE:	PRESS:
- Synonyms and variants	- Synonyms and variants
- Definition	- Description
- Encyclopaedic information	- Examples
- Examples	- Collocations
- Collocations	- Keyness
- Keyness	
- Note	
RESEARCH:	COMMENTS:
- Synonyms and variants	- Synonyms and variants
- Definition/description	- Description
- Examples	- Examples
- Collocations	- Collocations
- Keyness	- Keyness

d) ENTRY COMPILATION

- The entries were compiled and stored in a dedicated application designed by a team of IT students from UW.
- The app was based on the familiar MediaWiki engine.
- All entries are subject to the CC-BY licence.
- The students worked on specific entry sections corresponding to the working group they belonged to / the type of corpus they worked on, i.e. Normative, Research, Press, or Comments.
- All students could propose related terms, which were then verified and added by the instructors or student assistants.

ACCESS STRUCTURE



English entries [edit | edit source]

- Anti-vaxxer
- Antibody
- Asymptomatic
- Big pharma
- Booster shot
- COVID-19
- COVID-19 pandemic
- Close contact
- Coronavirus vaccine
- Covidiot
- Digital COVID certificate
- Dose
- Face mask
- Hospitalization
- Infection rate
- Intensive care
- Lockdown
- mRNA vaccine
- Mechanical ventilation
- Negative
- Positive
- Proof of vaccination
- Quarantine
- RT-PCR test
- Rapid diagnostic test
- Restrictions
- SARS-CoV-2
- Severe course of COVID-19
- Social distancing
- Symptomatic
- Vaccinate
- Vaccination programme
- Vaccination rate
- Vaccine mandate
- Variant of SARS-CoV-2
- Ventilator
- Virus transmission
- Wave of pandemic

Czech entries [edit | edit source]

- Antigenní test
- Anti-vaxxer/cz
- Asymptomatický
- Blízký kontakt
- COVID-19
- Covidiot
- Mechanická ventilace
- mRNA vakcína
- Národní certifikát o provedeném očkování
- Negativní
- Očkovací schéma
- Očkovat
- Proočkovanost
- Respirátor
- RT-PCR test
- SARS-CoV-2
- Šíření nákazy
- Šíření viru

Term:
Term:Anti-vaxxer
Term:Anti-vaxxer/cz
Term:Anti-vaxxer/de
Term:Anti-vaxxer/it
Term:Anti-vaxxer/pl
Term:Antibody
Term:Antibody/cz
Term:Antibody/de
Term:Antibody/it
Term:Antibody/pl
Search for pages containing *Term:*

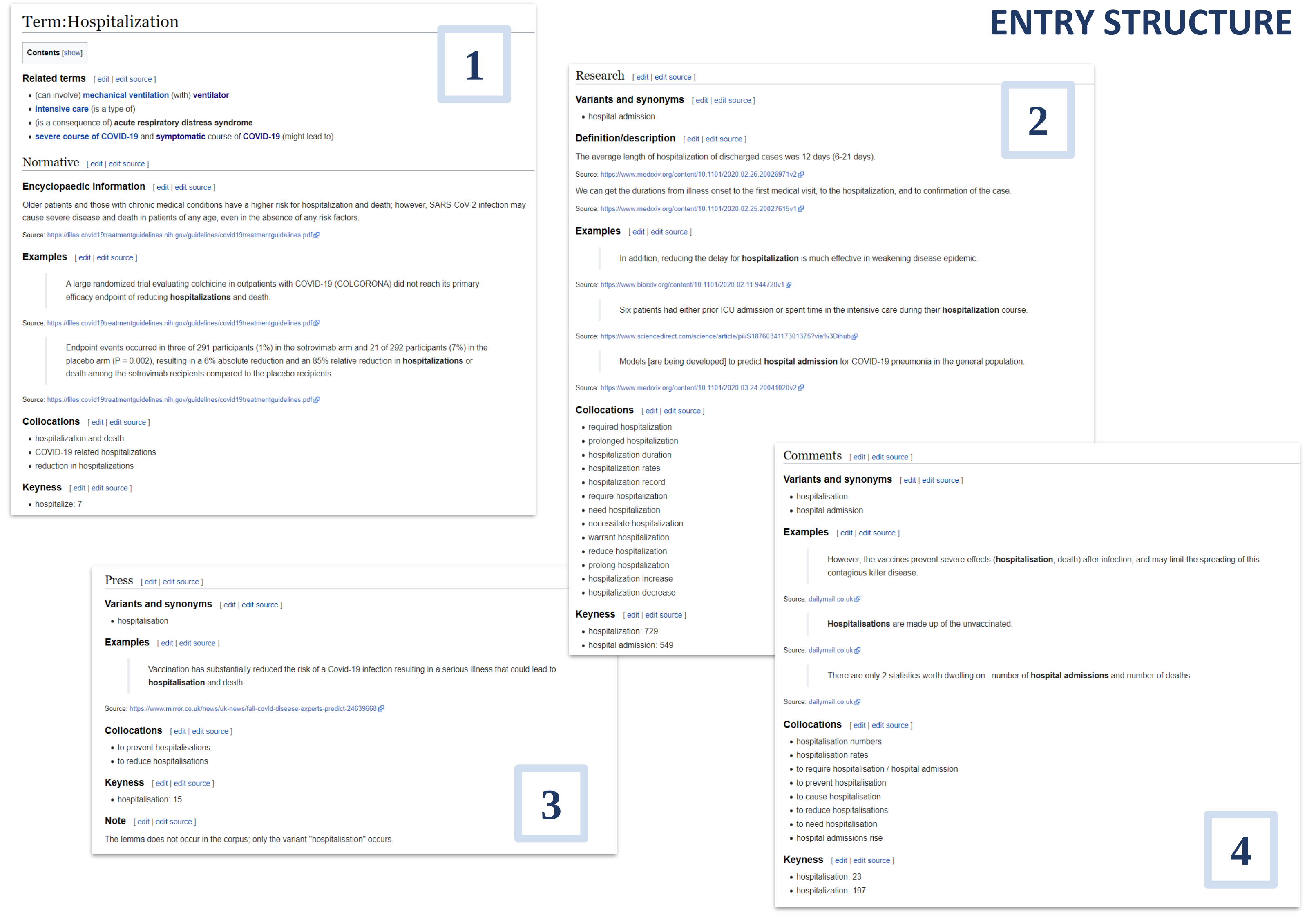
...velope project within the framework of the 4EU+ Alliance.

The knowledge base covers almost 40 parallel terminological entries in five lan German, Italian, and Polish. They were compiled on the basis of parallel corpo You can choose an entry from one of the lists below or search for it using the s

See more:

- About TermiKnowledge
- Technical parameters of the corpora
- Student participants
- Instructors

ENTRY STRUCTURE



1

Term:Hospitalization

Contents [show]

Related terms [edit | edit source]

- (can involve) **mechanical ventilation** (with) **ventilator**
- **intensive care** (is a type of)
- (is a consequence of) **acute respiratory distress syndrome**
- **severe course of COVID-19** and **symptomatic** course of **COVID-19** (might lead to)

Normative [edit | edit source]

Encyclopaedic information [edit | edit source]

Older patients and those with chronic medical conditions have a higher risk for hospitalization and death; however, SARS-CoV-2 infection may cause severe disease and death in patients of any age, even in the absence of any risk factors.

Source: https://files.covid19treatmentguidelines.nih.gov/guidelines/covid19treatmentguidelines.pdf

Examples [edit | edit source]

> A large randomized trial evaluating colchicine in outpatients with COVID-19 (COLCORONA) did not reach its primary efficacy endpoint of reducing **hospitalizations** and death.

Source: https://files.covid19treatmentguidelines.nih.gov/guidelines/covid19treatmentguidelines.pdf

> Endpoint events occurred in three of 291 participants (1%) in the sotrovimab arm and 21 of 292 participants (7%) in the placebo arm (P = 0.002), resulting in a 6% absolute reduction and an 85% relative reduction in **hospitalizations** or death among the sotrovimab recipients compared to the placebo recipients.

Source: https://files.covid19treatmentguidelines.nih.gov/guidelines/covid19treatmentguidelines.pdf

Collocations [edit | edit source]

- hospitalization and death
- COVID-19 related hospitalizations
- reduction in hospitalizations

Keyness [edit | edit source]

- hospitalize: 7

2

Research [edit | edit source]

Variants and synonyms [edit | edit source]

- hospital admission

Definition/description [edit | edit source]

The average length of hospitalization of discharged cases was 12 days (6-21 days).

Source: https://www.medrxiv.org/content/10.1101/2020.02.26.20026971v2

We can get the durations from illness onset to the first medical visit, to the hospitalization, and to confirmation of the case.

Source: https://www.medrxiv.org/content/10.1101/2020.02.25.20027615v1

Examples [edit | edit source]

> In addition, reducing the delay for **hospitalization** is much effective in weakening disease epidemic.

Source: https://www.biorxiv.org/content/10.1101/2020.02.11.944728v1

> Six patients had either prior ICU admission or spent time in the intensive care during their **hospitalization** course.

Source: https://www.sciencedirect.com/science/article/pii/S1876034117301375?via%3Dihub

> Models [are being developed] to predict **hospital admission** for COVID-19 pneumonia in the general population.

Source: https://www.medrxiv.org/content/10.1101/2020.03.24.20041020v2

Collocations [edit | edit source]

- required hospitalization
- prolonged hospitalization
- hospitalization duration
- hospitalization rates
- hospitalization record
- require hospitalization
- need hospitalization
- necessitate hospitalization
- warrant hospitalization
- reduce hospitalization
- prolong hospitalization
- hospitalization increase
- hospitalization decrease

Keyness [edit | edit source]

- hospitalization: 729
- hospital admission: 549

3

Press [edit | edit source]

Variants and synonyms [edit | edit source]

- hospitalisation

Examples [edit | edit source]

> Vaccination has substantially reduced the risk of a Covid-19 infection resulting in a serious illness that could lead to **hospitalisation** and death.

Source: https://www.mirror.co.uk/news/uk-news/fall-covid-disease-experts-predict-24639668

Collocations [edit | edit source]

- to prevent hospitalisations
- to reduce hospitalisations

Keyness [edit | edit source]

- hospitalisation: 15

Note [edit | edit source]

The lemma does not occur in the corpus; only the variant "hospitalisation" occurs.

4

Comments [edit | edit source]

Variants and synonyms [edit | edit source]

- hospitalisation
- hospital admission

Examples [edit | edit source]

> However, the vaccines prevent severe effects (**hospitalisation**, death) after infection, and may limit the spreading of this contagious killer disease.

Source: dailymail.co.uk

> **Hospitalisations** are made up of the unvaccinated.

Source: dailymail.co.uk

> There are only 2 statistics worth dwelling on...number of **hospital admissions** and number of deaths

Source: dailymail.co.uk

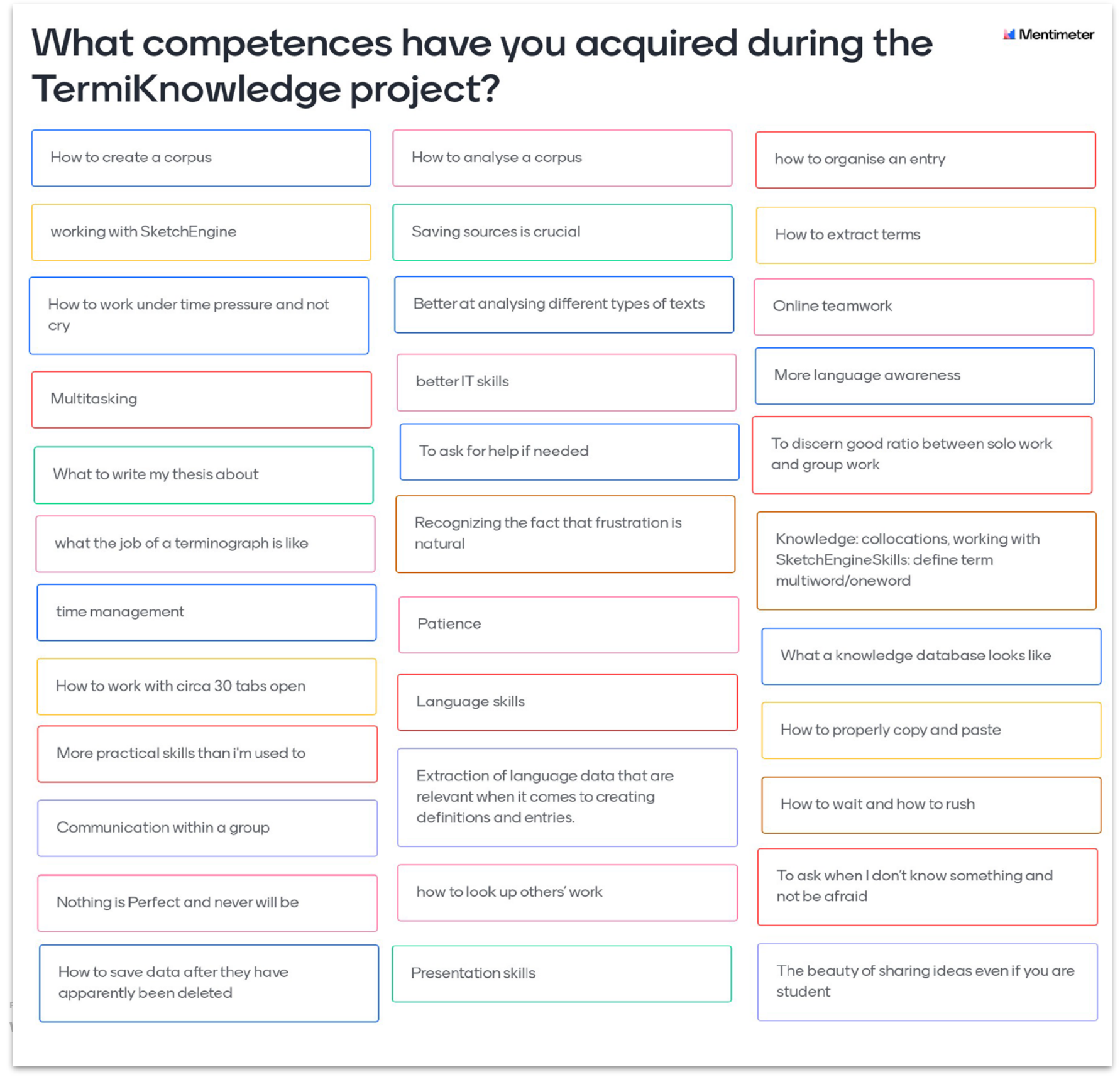
Collocations [edit | edit source]

- hospitalisation numbers
- hospitalisation rates
- to require hospitalisation / hospital admission
- to prevent hospitalisation
- to cause hospitalisation
- to reduce hospitalisations
- to need hospitalisation
- hospital admissions rise

Keyness [edit | edit source]

- hospitalisation: 23
- hospitalization: 197

CHALLENGES AND DISCOVERIES



What competences have you acquired during the TermiKnowledge project?

Mentimeter

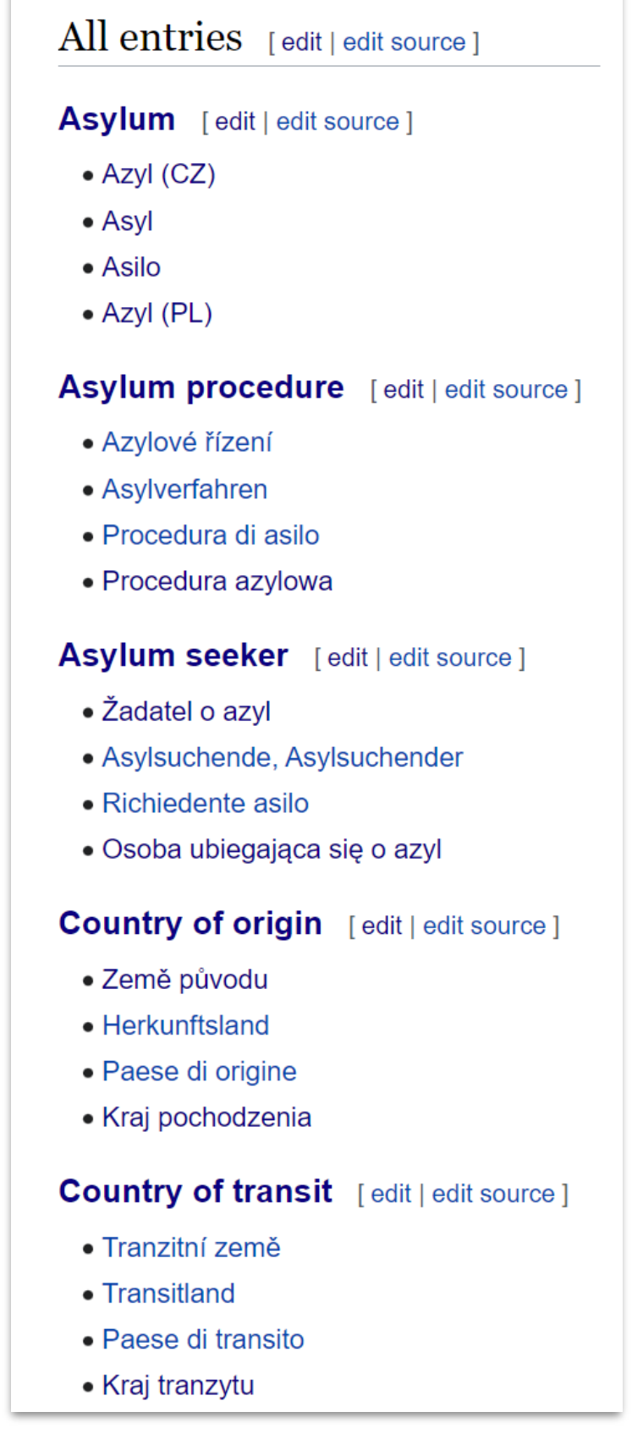
- How to create a corpus
- working with SketchEngine
- How to work under time pressure and not cry
- Multitasking
- What to write my thesis about
- what the job of a terminograph is like
- time management
- How to work with circa 30 tabs open
- More practical skills than i'm used to
- Communication within a group
- Nothing is Perfect and never will be
- How to save data after they have apparently been deleted
- How to analyse a corpus
- Saving sources is crucial
- Better at analysing different types of texts
- better IT skills
- To ask for help if needed
- Recognizing the fact that frustration is natural
- Patience
- Language skills
- Extraction of language data that are relevant when it comes to creating definitions and entries.
- how to look up others' work
- Presentation skills
- how to organise an entry
- How to extract terms
- Online teamwork
- More language awareness
- To discern good ratio between solo work and group work
- Knowledge: collocations, working with SketchEngineSkills: define term multiword/oneword
- What a knowledge database looks like
- How to properly copy and paste
- How to wait and how to rush
- To ask when I don't know something and not be afraid
- The beauty of sharing ideas even if you are student

REFLECTIONS AFTER SEMESTER I

- student portfolios
- MentiMeter survey during mid-term meeting in Heidelberg

CHANGES IN SEMESTER II

- organisation
  - more workshops
  - no working groups
- primary data
  - one corpus per language
  - only normative and press texts
- access structure
  - separate entry lists for each language
  - list of all English entries with equivalent other language entries
  - links to all equivalent entries at the top of each article
- microstructure
  - no entry sections
- implementation
  - more refined term relations



All entries [edit | edit source]

Asylum [edit | edit source]

- Azyl (CZ)
- Asyl
- Asilo
- Azyl (PL)

Asylum procedure [edit | edit source]

- Azylové řízení
- Asylverfahren
- Procedura di asilo
- Procedura azylowa

Asylum seeker [edit | edit source]

- Žadatel o azyl
- Asylsuchende, Asylsuchender
- Richiedente asilo
- Osoba ubiegająca się o azyl

Country of origin [edit | edit source]

- Země původu
- Herkunftsland
- Paese di origine
- Kraj pochodzenia

Country of transit [edit | edit source]

- Tranzitní země
- Transitland
- Paese di transito
- Kraj tranzytu