

Multilingual Language Technology in the age of Artificial Intelligence and Deep Neural Networks

Jan Hajič

Institute of Formal and Applied Linguistics
Computer Science School
Faculty of Mathematics and Physics
Charles University, Prague, Czech Republic

Using pictures created by Francesca Frontini, Fnaciska de Jong, Dieter van Uytvanck, Georg Rehm, Andy Way, Pavel Straňák

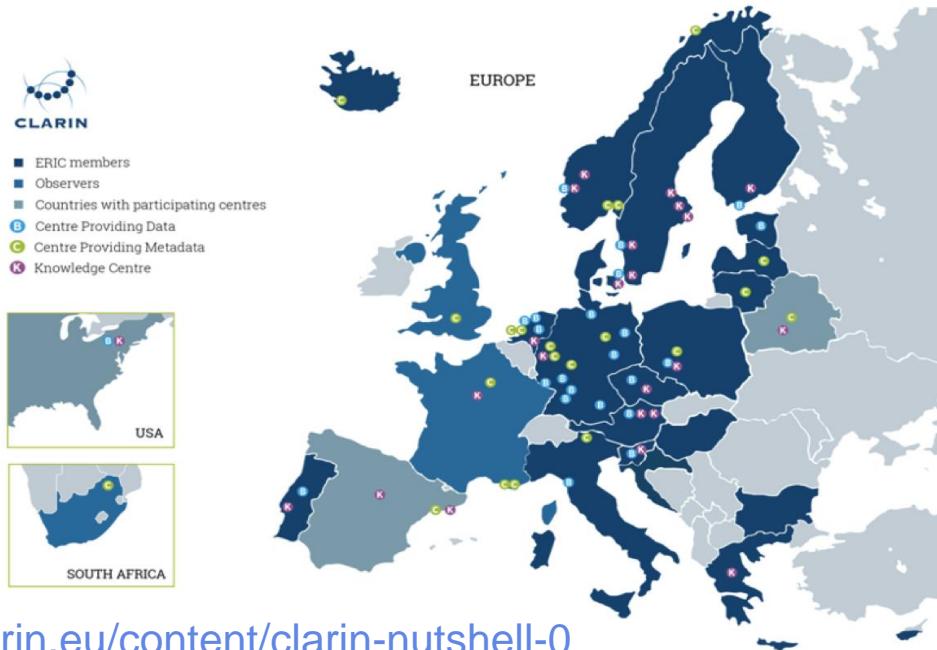


Outline

- European Networks
 - Language Resources & Technology
 - Fostering Language Equality
- Example of a complex Research Infrastructure
 - LINDAT/CLARIAH-CZ: resources, tools, services
- Future of Language Technology
 - AI methods / Large Language and Translation Models
 - Natural Language Understanding
- Conclusions

European Language Technology Networks: CLARIN

- CLARIN: since 2008,
CLARIN ERIC since 2012
 - >50 centres
 - 21 members + 3 observers
- Provides
 - Resources
 - Tools
 - Services
- Standardized, open, virtual,
distributed research
infrastructure

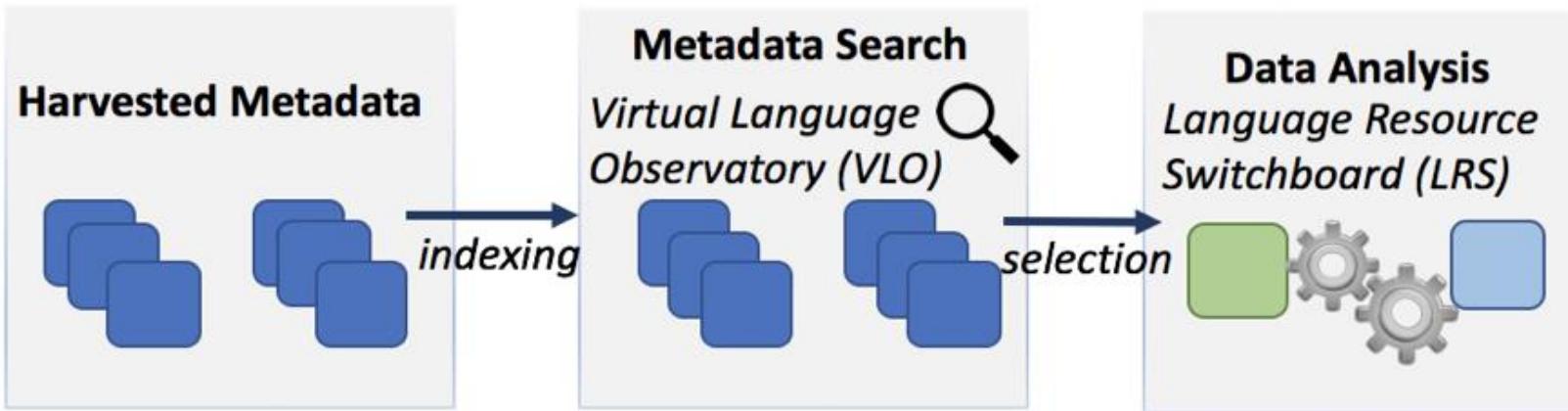


<https://www.clarin.eu/content/clarin-nutshell-0>



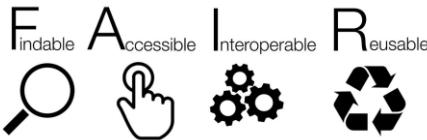
European Language Technology Networks: CLARIN

- The technical infrastructure

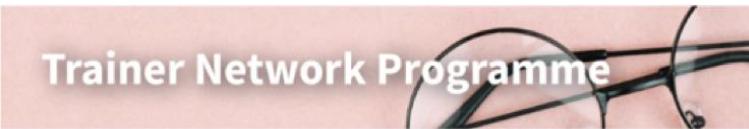


<https://vlo.clarin.eu/>

<https://switchboard.clarin.eu/>



European Language Technology Networks: CLARIN



<https://www.clarin.eu/content/clarin-for-researchers>

<https://www.clarin.eu/content/knowledge-sharing>

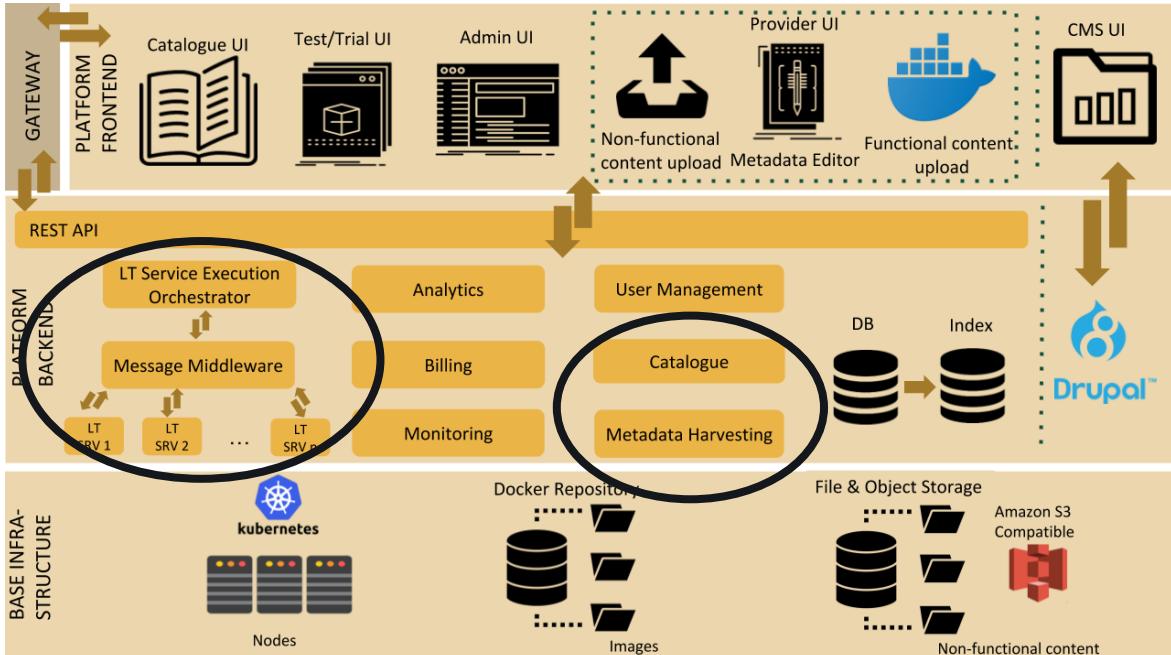
European Language Technology Networks: ELG

- European Language Grid
 - Centralized, all-in-one:
 - Language Resources (data, metadata, links to existing repositories)
 - Software tools, packages
 - (Running) services
 - Demos, catalogues (resources, institutions, stakeholders)
 - Marketplace for the European LT business space
 - Contributed from 15 external projects (FSTP funding)
 - And previous and concurrently running H2020 and infra projects
 - <https://www.european-language-grid.eu>



European Language Technology Networks: ELG

- Architecture
 - Services
 - Dockers
 - Data
 - Copied
 - Harvested
 - (CLARIN nodes)
 - Created
 - Curated



European Language Technology Networks: ELG



TOOLS & SERVICES

ELG provides access to LT services that can be tested directly through the user interface or called through the REST API.

FEATURED MOST POPULAR MOST RECENT

Cogito Discover Sentiment Analysis

270 VIEWS UPDATED 01.02.2022

Provides a sentiment score (positive or negative) for the entities recognized in the text, and an overall score for the whole set of entities in the...

HENSOLDT ANALYTICS Named Entity...

115 VIEWS UPDATED 21.12.2021

HENSOLDT ANALYTICS MediaMiningIndexer NED - named entity detection engine that provides classification of named entities of following types...

Elhuyar Basque ASR



LANGUAGE RESOURCES

ELG provides access to datasets, corpora, models and other resources from all over Europe, for all European languages.

FEATURED MOST POPULAR MOST RECENT

SardiStance Dataset

83 VIEWS UPDATED 12.05.2021

The SardiStance dataset collects 3,242 tweets written in Italian about the "Movimento delle Sardine", retrieved by means of the keywords...

Bangor University's Corpus of Welsh...

63 VIEWS UPDATED 25.08.2021

This is a collection of Welsh language sentences released under a CC0 license and collected by members of the Language Technologies...

Motion Capture 3D Sign Language Data...



ORGANISATIONS

ELG provides access to information about all European LT organisations, including universities and research centres.

FEATURED MOST POPULAR MOST RECENT

Institute for Bulgarian Language

65 VIEWS UPDATED 29.09.2021

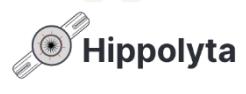
The Institute for Bulgarian Language at the Bulgarian Academy of Sciences is a leading research organisation focused on theoretical and...

VoiceLab AI

46 VIEWS UPDATED 11.03.2021

VoiceLab is a technology firm from Gdańsk, working on automatic speech recognition (ASR) and natural language understanding (NLU). Our...

Centre for Translation Studies



cubbitt

Search



- Examples in the catalog
- Machine translation services
 - Cubbitt

Language resources & technologies

Service functions

Languages

Media types

Licences

Conditions of use

Source

5 search results for cubbitt



CUBBITT Translation Models (en-cs) (v1.0)

8 views

version: unspecified

CUBBITT En-Cs translation models, exported via TensorFlow Serving, available in the Lindat translation service (<https://lindat.mff.cuni.cz/services/translation/>). Models are compatible with Tensor2tensor version 1.6.6. F

Keywords: machine · neural machine · transformer · cubbitt · translation · translation

Languages: English · Czech

Creative Commons Attribution Non Commercial Share Alike 4.0 International
Licence: Licence: Creative Commons Attribution Non Commercial Share Alike 4.0 International



CUBBITT Translation Models (en-fr) (v1.0)

51 views

version: unspecified

CUBBITT En-Fr translation models, exported via TensorFlow Serving, available in the Lindat translation service (<https://lindat.mff.cuni.cz/services/translation/>). Models are compatible with Tensor2tensor version 1.6.6. F



CLARIAH-CZ

European Language Technology Networks: ELE/ELE2

- European Language Equality
 - <https://european-language-equality.eu/>
 - Foster and promote multilinguality and equal (digital technology) treatment of European languages
 - Draw up **strategic research agenda and roadmap**
 - From current state of Language Technology to **Digital Language Equality**
 - Provide path for implementing the agenda by 2030
- Now in second phase (2022-2023)
 - Open Call for contributions – everybody welcome



European Language Technology Networks: ELE/ELE2

- Open Call



About ▾ Strategic Agenda Open Call Deliverables Events ▾ News ▾ Contact



Open Call for SRIA Contribution Projects

European Language Equality initiative opens a call for SRIA Contribution Projects!

The SRIA Contribution Projects, are meant to provide meaningful, so far missing, convincing, compelling input for the strategic agenda and roadmap. Become a part of the European Language Equality initiative and help us by contributing to the overall success and uptake of the strategic agenda by submitting a proposal that addresses one of the ten key topics.

You can submit your project proposals until **29th November 2022**.

Research organisations, NGOs, incorporated associations and companies legally established in the EU Member States are eligible for funding. Projects should last



[Call Documentation](#)
[Guide for Applicants](#)
[Third Party Agreement](#)
[Evaluation Criteria](#)

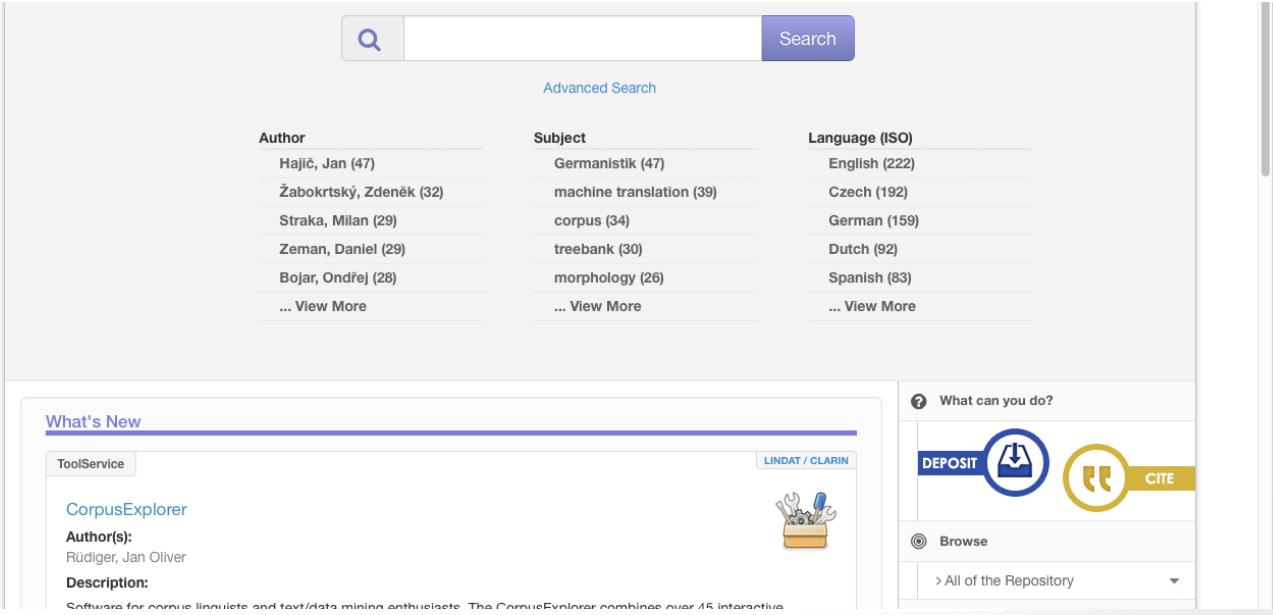
- <https://european-language-equality.eu/open-call/>

LINDAT/CLARIAH-CZ

- CLARIN “node” at <https://lindat.cz>
- Certified technical centre
 - Repository
 - Data, tools, metadata, licensing information
 - Single sign-on, persistent IDs
 - Services
 - Distributed network for Digital Humanities and Arts
 - Libraries, gallery, film archive, historical and other academy partners
 - Also part of Dariah ERIC and (soon) of EHRI ERIC



- Core point: data repository – standardized, certified
 - Preserve and find language data and NLP tools



The screenshot shows the LINDAT/CLARIAH-CZ search interface. At the top is a search bar with a magnifying glass icon and a "Search" button. Below it is a link to "Advanced Search". The main area displays search results categorized by Author, Subject, and Language (ISO). The "Author" section lists Hajič, Jan (47), Žabokrtský, Zdeněk (32), Straka, Milan (29), Zeman, Daniel (29), Bojar, Ondřej (28), and a "... View More" link. The "Subject" section lists Germanistik (47), machine translation (39), corpus (34), treebank (30), morphology (26), and a "... View More" link. The "Language (ISO)" section lists English (222), Czech (192), German (159), Dutch (92), Spanish (83), and a "... View More" link. At the bottom left is a "What's New" section featuring "ToolService" and "CorpusExplorer". The "CorpusExplorer" section includes fields for "Author(s)" (Rüdiger, Jan Oliver) and "Description". On the right side, there is a sidebar titled "What can you do?" with buttons for "DEPOSIT" (blue circle with a drop icon), "CITE" (yellow circle with a speech bubble icon), "Browse" (radio button icon), and a dropdown menu for "All of the Repository".





LINDAT/CLARIN

- OPEN  ACCESS (as much as it can)
- > 500 registered users
 - submitters & users signing license
- 200+ Data Records
 - > 1000 Metadata Records
- 1001 languages
 - 200 TB+ Data in Repository
 - + 1PB of UCS Shoah Found
- 40,000 accesses per month



EUROPEAN
LANGUAGE
GRID



LINDAT/CLARIN Repository Home / Search

Search

[Advanced Search](#)

Limit your search

Author

Subject

Rights

Language (ISO)

Type

Contain Files

Community

Showing 1 through 10 out of 1038 results

1
2
3
>
104

Corpus

[AKCES 2 ver. 2](#)

(Charles University in Prague, ÚČJTK / 2013-12-18)

Author(s): Šebesta, Karel ; Holáňová, Hana

This item contains 1 file (3.85 MB).

Publicly Available

LexicalConceptualResource

[A Gold Standard Word Alignment for English-Swedish \(2015-10-12\)](#)

(Linköping University / 2015-10-12)

Author(s): Ahrenberg, Lars ; Holmqvist, Maria

This item contains 1 file (590 KB).

Publicly Available

ToolService

[MorphoDiTa: Morphological Dictionary and Tagger](#)

(Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics (UFAL) / 2014-02-14)

Author(s): Straka, Milan ; Straková, Jana

This item contains no files.

LINDAT / CLARIN

LINDAT/CLARIAH-CZ

- Safe preservation
 - upload and don't worry
- Discovery & Reuse
- Direct data citation
 - Google Scholar, Datasets
- Licensing
 - Open Access
 - ... more options
- Versioning
- Worldwide (for everyone)

How to Deposit

Only authenticated users can deposit items. If you cannot find your home organisation in the Login dialog list of organisations then register at clarin.eu and authenticate using "clarin.eu website account". In case you cannot use any authentication method above or if you encounter a problem, do not hesitate to contact our [Help Desk](#) and we can create a local account for you.

Step 1: Login

To start a new submission you have to login first. Click Login under My Account in the right menu panel.



Fig1. Menu Login

Step 2: Starting a new submission

Now you have a new menu item 'Submissions' under My Account. Click on Submissions to go to the Submissions screen.

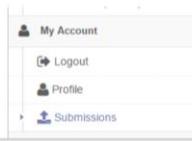


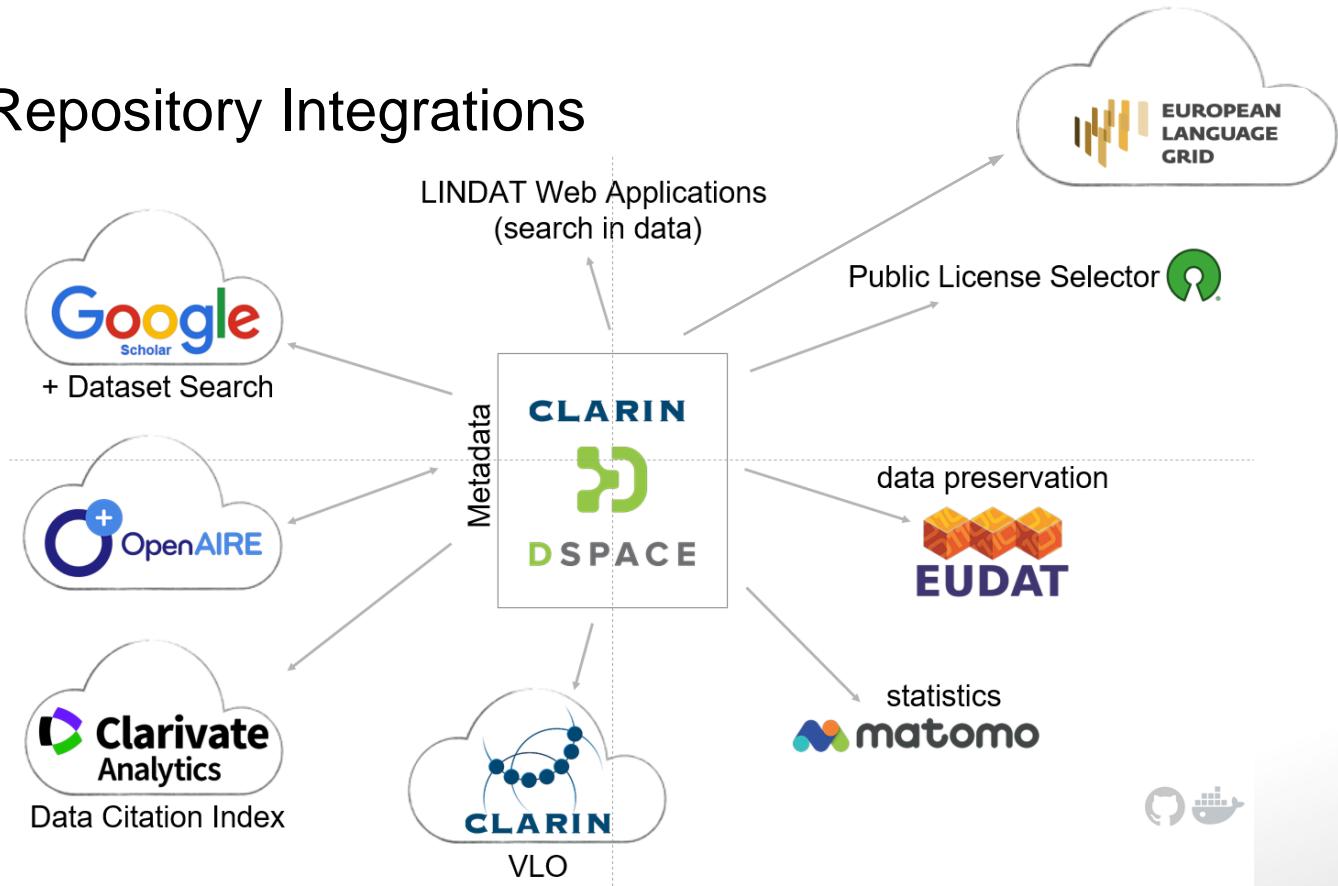
Fig2. Menu Submissions

Now you should be on the main Submission and Workflow tasks page where you can view your incomplete/archive submissions. Click on the 'Start another submission' link to start a new Submission.

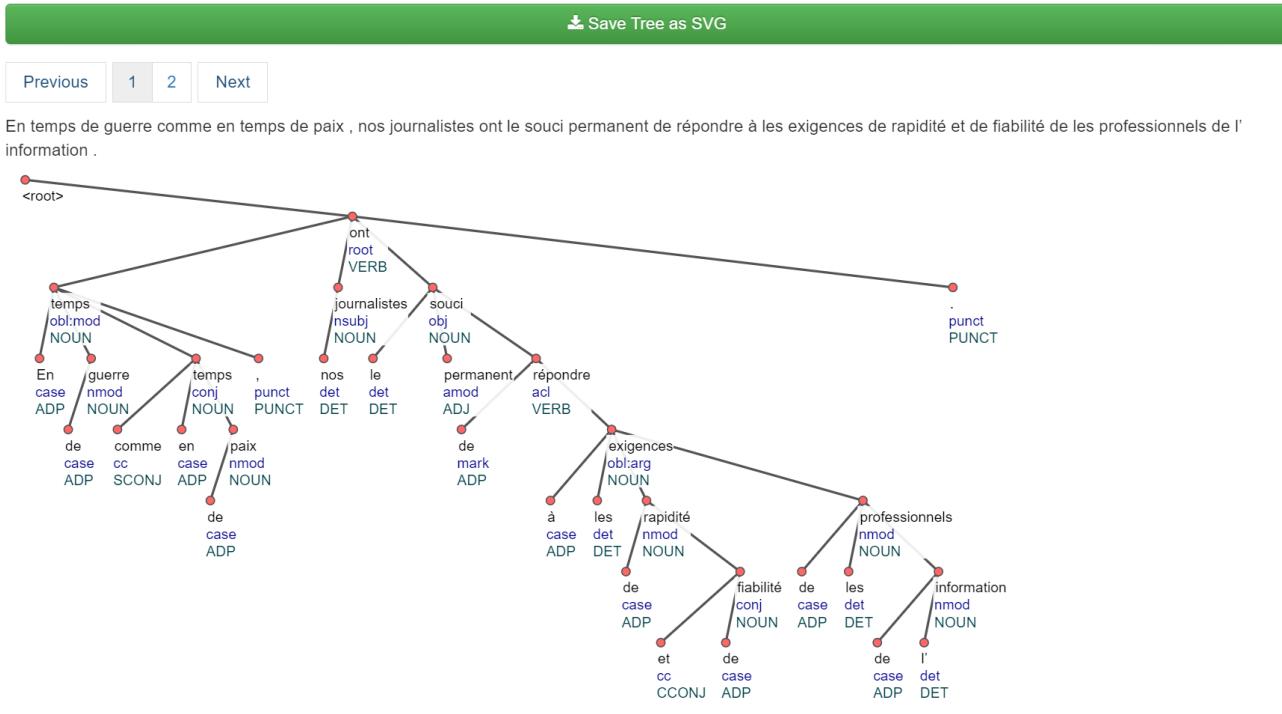



LINDAT/CLARIAH-CZ

- Repository Integrations



- Text analysis services – UDPipe 2.0 (and other)



- Machine Translation service
 - Similar to Google Translate, Bing, DeepL



Search Catalogue Education Projects Tools Services About ▾



LINDAT Translation

Translate Docs

The translation service is available for **personal and non-commercial use** (see [terms of use](#) for more details).

Source Target

English Czech

advanced

Input sentences

At least 21 people were killed and more than 250 injured in clashes that erupted after al-Sadr announced his "final retirement" from politics Monday and said he was closing down his political offices across the country.

Translation

Nejméně 21 lidí bylo zabito a více než 250 zraněno při střetech, které propukly poté, co al-Sadr v pondělí oznámil svůj „definitivní odchod“ z politiky a řekl, že uzavírá své politické úřady po celé zemi.



Language Technology Future

- In past few years...
 - Huge success in applying Deep Learning
 - End2end systems, large language models, Multilanguage models
 - Need for large data (and computing resources)
 - Advances in applications
 - Speech
 - Language
 - Multimodal (vision, image, video, ...)
 - Integrated structured and unstructured data

Language Technology Future

- In past few years...
 - Huge success in applying Deep Learning
 - End2end systems, **large language models, Multilanguage models**
 - **Need for large data (and computing resources)**
 - Advances in applications
 - Speech
 - Language
 - Multimodal (vision, image, video, ...)
 - Integrated structured and unstructured data

The HPLT (Hippolyta) project

- High-Performance Language Technology
- Horizon Europe DATA call, 2022-2025
- Goals
 - Collect large data from Internet Archive (SF, CA, USA)
 - Approx. 12 PB
 - Extract text, clean, identify, deduplicate, pseudonymize, describe, ...
 - Train language and translation models: 24 EU + min. 16 other
 - xBERTy, GPT-x, Transformer, future SoTA
 - make them openly available (OpusMT, Huggingface, possibly other repos)
 - Evaluate models – keep a dashboard
 - Demonstrate use of EU HPC Centres in a distributed manner
 - Huge compute demands: just for cleaning, 20 mil. CPU hours

The HPLT (Hippolyta) project

- Project partners
 - University of Edinburgh (Ken Heafield): scientific coordinator,
 - Charles University (UFAL/LINDAT, Jan Hajic, Dusan Varis, Jindrich Helcl, Martin Popel, Pavel Stranak, Barbor Hladka)
 - coordinator
 - University of Helsinki (Finland, Jorg Tiedemann, OpusMT)
 - University of Turku (Finland, Sampo Pyysalo, Filip Ginter)
 - University in Oslo (Norway, Stephan Oepen)
 - Prompsit (Spain, Gema Ramirez)
 - HPCs:
 - CESNET (Czechia, Ludek Matyska, David Antos)
 - Sigma2 (Norway, Hans Eide)
 - Cooperation with LUMI, EuroHPC, Karolina (IT4Innovations), possibly others

Natural Language Understanding

- Common theme of many roadmaps, agendas, etc.
- Are we there yet?
 - Large language models (esp. GPT-x) can generate fluent text (incl. theatre plays! See e.g. the TheAltre project)
 - Translation beats humans (some domains)
 - Dialog systems can pass Turing test (in the order of minutes or low tens of minutes)
- Questions remain:
 - Explainability, interpretability, post-output error detection, (de)bias(ing), ethical issues, ...

Natural Language Understanding

- Possible direction of further research
 - Both large(r) language models and symbolic methods
 - Combination?
- Can knowledge representation be useful?
 - What “knowledge representation”?
- Goal of many research projects
 - Building ontologies: Cyc (and successors), Babelnet, but also Wikipedia, wikidata, databases, ...
 - Building semantic representations
 - From the language point of view
 - From the logical/reasoning/inference side

Natural Language Understanding

- Human-like perception of read/heard language (text, voice)



- Key point: **representation** (of the knowledge/content acquired)
- Unclear how people remember and structure knowledge and facts...
 - ... Mathematics and Logic to the rescue
 - Graphs (with known properties and algorithms)
 - Graph search, graph transformations, graph matching
 - Ontologies to include contents
 - Logic: inferencing, contradiction detection, deduction, ...

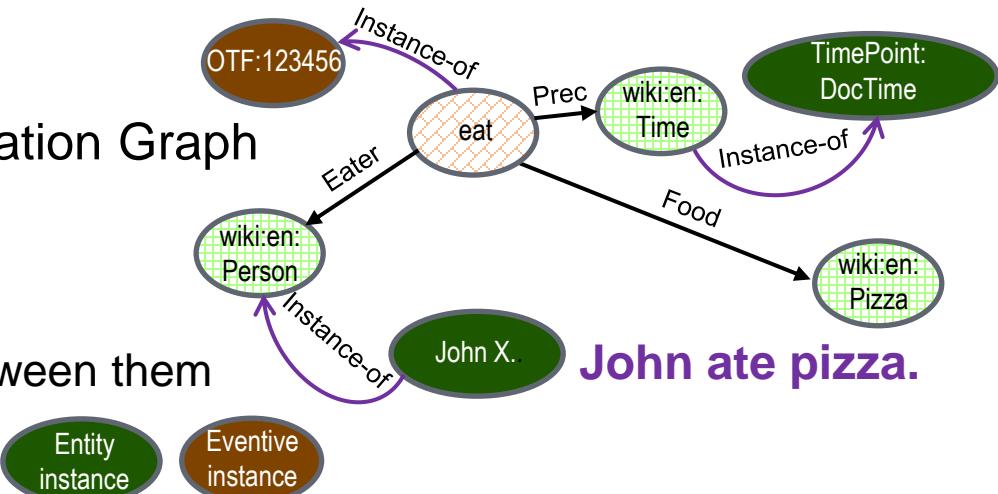


Natural Language Understanding

- LUSyD: Language Understanding: from Syntax to Discourse
 - Project 2020-2024, Czech Science Foundation
 - Linguistic & Computational (DNN, Language Models)
 - Relation between language form and knowledge representation
 - Morphology, syntax, semantics, discourse
 - Come up with **Knowledge Representation**
 - Link to existing language (annotation/lexical) resources
 - Demonstrate on **multiple** (as many as possible) **languages**
 - Plan: Czech, English, German, Spanish, Chinese, Korean (Arabic, ...)
 - Create **eventive grounding ontology**, populated with multiple languages (SynSemClass), linked to FrameNet, WordNet etc.

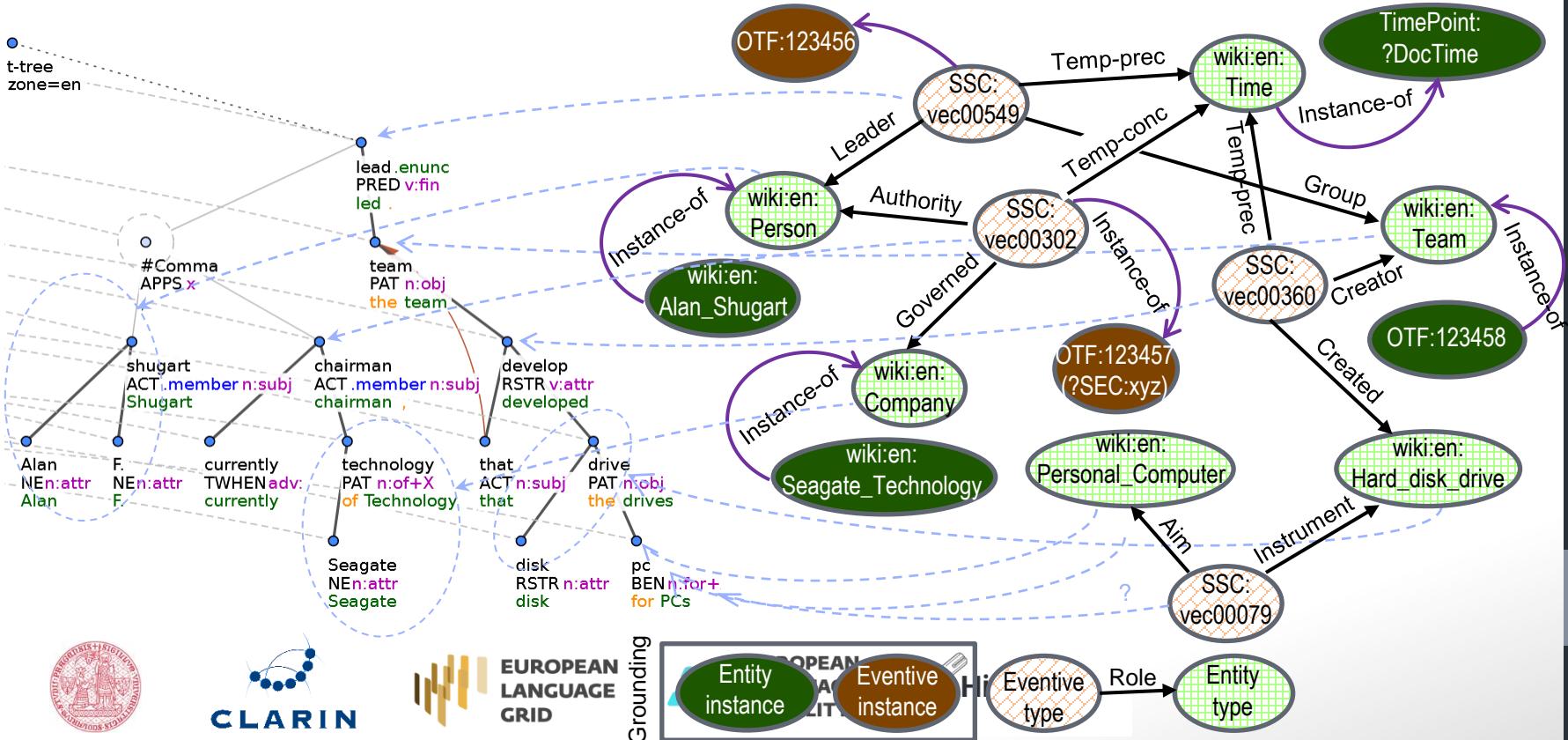
Natural Language Understanding

- The Principles
 - Knowledge Representation Graph
 - Nodes representing
 - Events
 - Entities
 - Edges: relations between them
 - All nodes grounded
 - Ontologies
 - Relations: fixed set
 - Nodes linked to original text/speech/...
 - For Machine Learning in general, applications



NLU: real example (WSJ)

Alan F. Shugart, currently chairman of Seagate Technology, led the team that developed the disk drives for PCs.



Thank you!

<https://ufal.mff.cuni.cz>

<https://lindat.cz>

<https://lindat.cz/services>

Twitter: [@LindatClariahCZ](#)

Twitter: [@ufal_cuni](#)



EUROPEAN
LANGUAGE
GRID



EUROPEAN
LANGUAGE
EQUALITY



Hippolyta

References (NLU)

- Zdeňka Urešová, Karolina Zaczynska, Peter Bourgonje, Eva Fučíková, Georg Rehm, Jan Hajič. 2022. Making a Semnatic Event-type Ontology Multilingual. LREC 2022, Marseille, France (this presentation).
- Petr Sgall, Eva Hajičová, and Jarmila Panevová. 1986. *The meaning of the sentence in its semantic and pragmatic aspects*. D. Reidel, Dordrecht.
- Zdeňka Urešová, Jan Štěpánek, Jan Hajič, Jarmila Panevová, and Marie Mikulová. 2014. *PDT-Vallex*. LINDAT/CLARIN digital library. <http://hdl.handle.net/11858/00-097C-0000-0023-4338-F>.
- Zdeňka Urešová, Eva Fučíková, and Jana Šindlerová. 2016. *CzEngVallex: a bilingual Czech-English valency lexicon*. The Prague Bulletin of Mathematical Linguistics, 105:17–50.
- Zdeňka Urešová, Eva Fučíková, Eva Hajičová, and Jan Hajič. 2018a. Creating a Verb Synonym Lexicon Based on a Parallel Corpus. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC'18)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).
- Zdeňka Urešová, Eva Fučíková, Eva Hajičová, and Jan Hajič. 2018b. Defining verbal synonyms: between syntax and semantics. In Dag Haug, Stephan Oepen, Lilja Ovreliid, Marie Candito, and Jan Hajič, editors, *Proceedings of the 17th International Workshop on Treebanks and Linguistic Theories (TLT 2018)* (Pub. No. 155), pages 75–90, Linköping, Sweden. Universitetet i Oslo, Linköping University Electronic Press.
- Zdeňka Urešová, Eva Fučíková, Eva Hajičová, and Jan Hajič. 2018c. Synonymy in Bilingual Context: The SynSemClass Lexicon. In *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018*, Santa Fe, New Mexico, USA, August 20-26, 2018, pages 2456–2469.