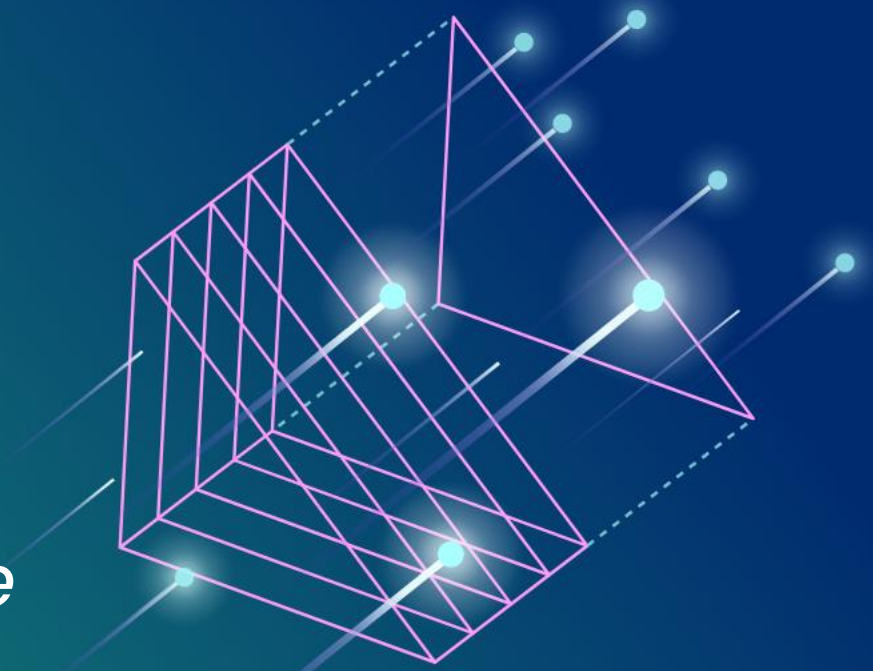
EUROPEAN BIG DATA VALUE FORUM

Getting value out of language data: technologies, platforms and solutions

Language Technology Research in Prague

Jan Hajič, Charles University, Prague



ORGANISED BY:

BDV BIG DATA VALUE ASSOCIATION

IN COLLABORATION WITH:

European Commission

Plan4all

VSB TECHNICAL UNIVERSITY OF OSTRAVA | IT4INNOVATIONS NATIONAL SUPERCOMPUTING CENTER

UNDER THE AUSPICES OF:

EU2022.CZ

Research in Language Technology @ Charles University & LINDAT/CLARIAH-CZ

METHODS

- Deep learning
  - Use less data (& be green)
    - Learn to use synthetic data, esp. in MT
  - Multimodal systems
  - Large Language Models (HPLT)
- Neurosymbolic learning
  - Human in the loop

DATA (LANGUAGE RESOURCES)

- Collecting data
  - All modalities an settings
    - Speech, text, multilingual
    - Dialog systems
- Annotated, symbolic data
  - Semantics, language understanding
  - Universal Dependencies (120+ languages)
  - Cooperation on Universal Meaning Representation

Ecosystem in Language Technology

DATA REPOSITORIES AND ENVIRONMENTS

- Open, accessible („FAIR“)
  - European Open Sceince Cloud
- Part of pan-european efforts
  - European Language Resource Consortium
  - CLARIN
  - European Language Grid
  - (Upcoming) Language Data Space

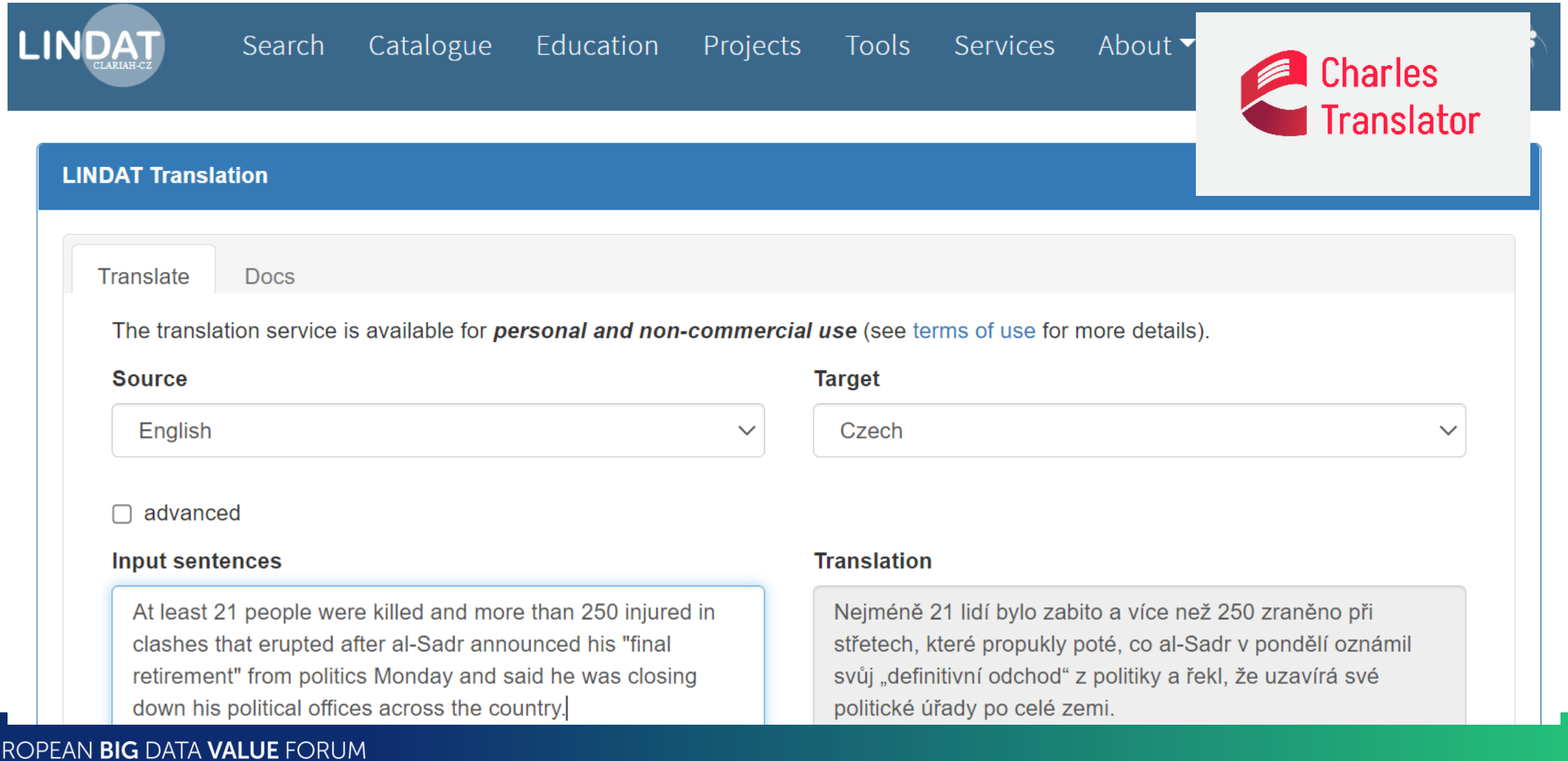
CORE TECHNOLOGIES

- Large Language / Translation models
  - EU and national projects
    - HPLT, OpenGPT-X, OpusMT
  - National projects
- HPC support
  - LUMI, EuroHPC, national HPCs (IT4I)

Services through Research Infrastructure

- LINDAT/CLARIAH-CZ – since 2010, part of CLAIRN ERIC network with 25 countries
  - **Machine Translation service**
    - Similar to Google Translate, Bing, DeepL, …
  - **Charles Translator** en-cs system (CUBBITT)
    - Better than professionals (Popel et al., 2020, Nat. Comm.) on translating news texts
    - Technology: Deep Neural Networks (DNNs), Transformer with specific data feed mixture
    - Trained on very large data - parallel cs/en, cs/uk, … texts
    - *Ukrainian – Czech* available as a special model + app with speech support

Services through Research Infrastructure



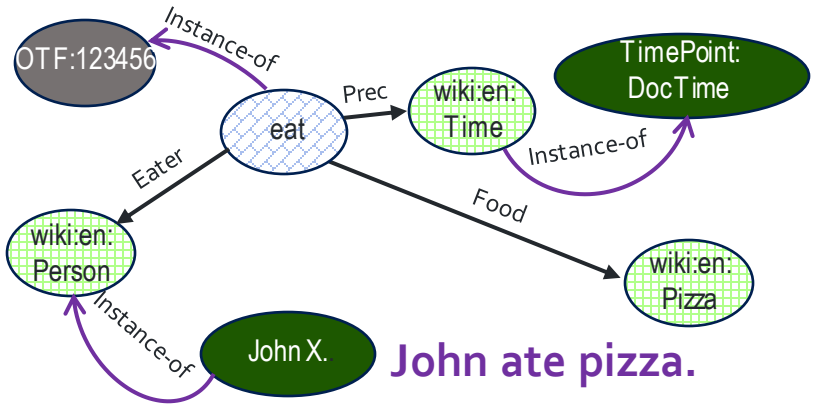
Future Research: Natural Language Understanding

- 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…
    - Graphs (with known properties and algorithms)
    - Ontologies to include contents (LUSyD project)
    - Logic: inferencing, contradiction detection, deduction, …
  - Neurosymbolic systems (Machine learning using [some] symbolic knowledge)

Natural Language Understanding

- The Principles
  - Knowledge Representation Graph
    - Nodes representing ***Events & Entities***
    - Edges: ***relations*** between them
  - All nodes ***grounded*** in
    - Human-comprehensible ***Ontologies*** (Wikidata, dbpedia, Wikipedia, SynSemClass…)
  - Nodes *linked* to original text/speech/…
    - For Machine Learning in general, applications



EUROPEAN BIG DATA VALUE FORUM

Thank you for attention
Questions?

ORGANISED BY:

BDV BIG DATA VALUE ASSOCIATION

IN COLLABORATION WITH:

European Commission

Plan4all

VSB TECHNICAL UNIVERSITY OF OSTRAVA | IT4INNOVATIONS NATIONAL SUPERCOMPUTING CENTER

UNDER THE AUSPICES OF:

EU2022.CZ