EUROPEAN
LANGUAGE
DATA SPACE

Common European Language Data Space
The LDS Workshop in Czechia

Jan Hajič (Charles University, Czechia)
hajic@ufal.mff.cuni.cz

02-12-2024 LDS Country Workshop, Prague, Czechia
https://language-data-space.ec.europa.eu

Automatic transcription and translation (ELITR.EU results)

Speech to text translation into 40 languages

A Common European Language Data Space – funded under contract LC-01936389 with the European Union.

EUROPEAN
LANGUAGE
DATA SPACE

Unleashing the potential of data – for EU businesses and citizens

Context: Language Technologies, Artificial Intelligence and (Large) Language Models

- Language Technologies (LT) and Artificial Intelligence (AI) need language data
- Large Language Models (LLMs) lead now the AI race, needing vast amounts of language data to train them
- Europe’s languages are largely under-resourced, except English
- A concerted effort for the collection of enormous amounts of language data for all European languages is very much needed
  - Plus evaluation data for multilingual, culturally-aware applications and AI Act compliance

EUROPEAN
LANGUAGE
DATA SPACE

LDS and the Digital Europe Programme

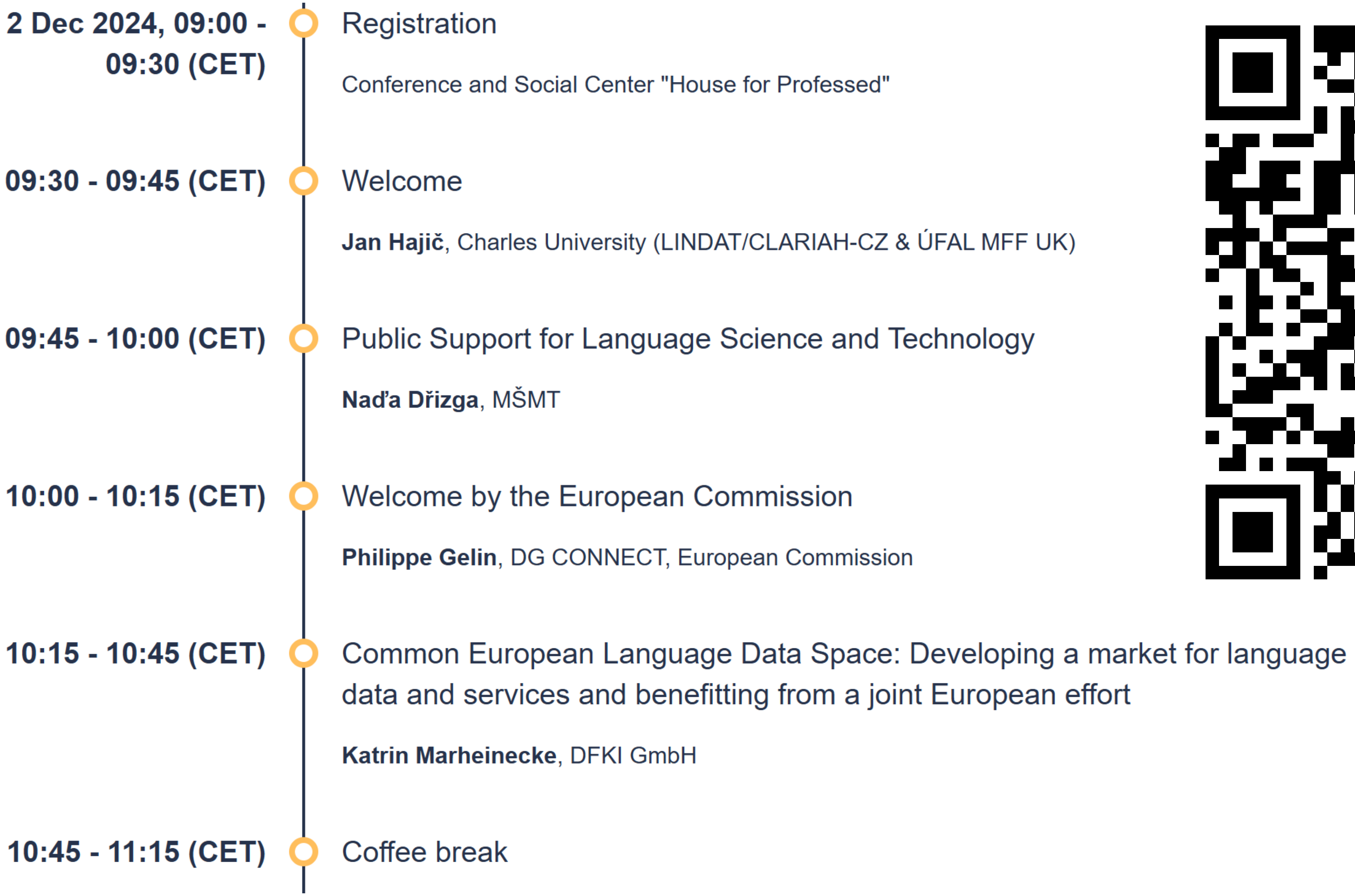
LDS and Digital Europe Programme

- The LDS builds upon initiatives like the European Language Resource Coordination (ELRC), the European Language Grid (ELG) and European Language Equality (ELE) projects
- LDS’ long-term sustainability: the ALT-EDIC
  - ALT-EDIC established 2024, Czechia is member since May 2024
- Funding possibilities
  - Mainly through the application- and industry-oriented Digital Europ Programme
  - First calls closed in May 2024
    - CSA already signed into a project: ALT-EDIC4EU (coordinated by ALT-EDIC)
    - Two technical projects shall follow soon (~40 mil. EUR each, 50-75% EU contribution):
      - One for building multilingual LLMs (followup to HPLT, TrustLLM and other current projects) – under negotiation
      - One for building applications over multilingual LLMs (under review)

EUROPEAN
LANGUAGE
DATA SPACE

Workshop Programme

https://language-data-space.ec.europa.eu/events/lds-country-workshop-czechia-2024-12-02_en



2 Dec 2024, 09:00 - 09:30 (CET)
Registration

Conference and Social Center "House for Professed"

09:30 - 09:45 (CET)
Welcome

Jan Hajič, Charles University (LINDAT/CLARIAH-CZ & ÚFAL MFF UK)

09:45 - 10:00 (CET)
Public Support for Language Science and Technology

Naďa Dřizga, MŠMT

10:00 - 10:15 (CET)
Welcome by the European Commission

Philippe Gelin, DG CONNECT, European Commission

10:15 - 10:45 (CET)
Common European Language Data Space: Developing a market for language data and services and benefitting from a joint European effort

Katrin Marheinecke, DFKI GmbH

10:45 - 11:15 (CET)
Coffee break

11:15 - 11:30 (CET) The role of data in the Conversational AI

Jan Cuřín, MAMA AI

11:30 - 12:30 (CET) Panel session: Language Data and Language Technologies in Czechia and for Czech

Jan Hajič, Charles University (moderator)

Jan Cuřín, MAMA AI
Aleš Tamchyna, Phrase
Naďa Dřizga, MŠMT
Veronika Krejčířová, Seznam.cz
Jirka Hana, Geneea
Petr Schwarz, Phonexia

12:30 - 13:30 (CET) Lunch break

13:30 - 14:00 (CET) Keynote: Utilising Language Data at an AI-driven Localisation Technology Company

Aleš Tamchyna, Phrase

14:00 - 14:45 (CET) Panel session: Language Data production, management and market development

Jan Hajič, Charles University (moderator)

Jan Cuřín, MAMA AI
Aleš Tamchyna, Phrase
Veronika Krejčířová, Seznam.cz
Jirka Hana, Geneea
Petr Schwarz, Phonexia

14:45 - 15:00 (CET) Conclusions

Jan Hajič, Charles University

EUROPEAN
LANGUAGE
DATA SPACE

Practical Information:

Please share your feedback with us: *Evaluation form*

Do you need a *Certificate of attendance*?

Common European Language Data Space

https://ec.europa.eu/eusurvey/runner/LDS_WS1-CZ_Feedback

Thank you!

Enjoy the workshop!

A Common European Language Data Space – funded under contract LC-01936389 with the European Union.