High-performance Language Technologies aka *Hippolyta*

Jan Hajič

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

The HPLT (*Hippolyta*) project

- High-Performance Language Technology
- Horizon Europe DATA call, 2022-2025
- Goals
  - Collect large data from Internet Archive (San Francisco, 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, UK (**Ken Heafield**, *scientific coordinator;* Barry Haddow)
  - Charles University, Czechia (ÚFAL/LINDAT: **Jan Hajič**, Dušan Variš, Jindřich Helcl, Martin Popel, Pavel Straňák, Barbora Vidová Hladká)
    - Coordinator organization
  - University of Helsinki, Finland (Jörg Tiedemann, OpusMT)
  - University of Turku, Finland (Sampo Pyysalo, Filip Ginter)
  - University in Oslo, Norway (Stephan Oepen, Andrey Kutuzov)
  - Prompsit, Spain (Gema Ramirez)
  - HPCs:
    - CESNET, Czechia (Luděk Matyska, David Antoš)
    - Sigma2, Norway (Hans Eide)
    - Cooperation with LUMI, EuroHPC, Karolina (IT4Innovations), possibly others

Thank you!

https://ufal.mff.cuni.cz
https://lindat.cz
https://lindat.cz/services
Twitter: @LindatClariahCZ
Twitter: @ufal_cuni