



SSHOC and LINDAT/CLARIAH-CZ

Jan Hajič
ÚFAL MFF UK & LINDAT
Prague, CZ



The SSHOC project



- SSHOC
 - Social Sciences and Humanities Open Cloud

The banner features the 'cessda eric' logo at the top left, which includes the text 'Consortium of European Social Science Data Archives' and 'European Research Infrastructure Consortium'. In the center, the 'SSHOC' logo is displayed in a stylized font, followed by the text 'Social Sciences & Humanities Open Cloud'. At the bottom left, there is a 3D illustration of two yellow construction cranes lifting blue blocks to form a cloud shape. On the right side of the banner, the text 'INFRAEOSC-04-2018' and 'Ensure the connection of research infrastructures' is written in white.



SSHOC project



- Research Infrastructures involved
 - CESSDA (coordinator) – social sciences
 - Consortium of European Social Science Data Archives
 - ESS – European Social Surveys
 - SHARE – The Survey of Health, Ageing and Retirement in Europe ...
 - CLARIN – Language Resources (& technology)
 - DARIAH – Digital Humanities and Arts
 - E-RIHS – History
- Each RI participates also through many LTPs



SSHOC project



- Budget
 - 11.5M EUR
 - 40 months duration, 2019 – IV/2022

- Pillars of SSHOC

- | | |
|--|------------|
| 1. User-centric Services | WP 3,4,5,6 |
| 2. Tools for FAIR data management | WP 3,4 |
| 3. Access mechanisms | WP 5,6,7 |
| 4. Rules of participation and compliance for actors | WP 7 |
| 5. Architecture to federated e-infrastructures | WP 7 |
| 6. Federative yet efficient Governance structure | WP 8 |



LINDAT (CU) in SSHOC



- LINDAT (J. Hajič)
 - Research Infrastructure (MSMT CZ), part of Clarin
 - LINDAT/CLARIN: 2016-2019, P. Straňák, J. Mišutka
 - LINDAT/CLARIAH-CZ: 2019-2022, B. Vidová
 - Combines membership in EU CLARIN and DARIAH networks
 - 11 institutions in the Czech combined network
 - UK, MU, Academy of Sciences, National and Moravian Libraries, NG, NFA
 - <http://lindat.cz> (<http://clariah.lindat.cz>)
- SSHOC involvement
 - ESS: (automatic) translation of surveys
 - Min. 3 language pairs (English to and from Czech, Russian, German, French)
 - CLARIN: Task 3.3 – provide basic linguistic processing
 - UDPipe, 70+ languages – particular languages for SSHOC tbd



Text and Data Mining (CLARIN)



- WP3 – T3.3
 - Task leader: Maria Eskevich (CLARIN ERIC)
 - TDM for social sciences
 - 4 topics
 - CUNI in at least one of them
 - Use of basic linguistic analysis applies to all four topics
 - Topic 2 explicitly focuses on linguistic processing for TDM, into existing tools
 - Results:
 - Easy to use pipeline for languages in question
 - Based on UDPipe tool (Straka et al.)
 - Evaluated for „threshold“ (UDPipe: 70 languages, but not all up to the challenge, and some not of interest)
- WP3 – T3.1 Multilingual terminology
 - Small CUNI involvement (ling. Tools)



Translation for ESS

- WP4 – T4.2
 - Main contact: Diana Zavala-Rojas (UPF)
 - Adapt SoA MT models for the domain of social surveys
 - Deep NN (Neural Machine Translation)
 - Transformer architecture (M. Popel et al., WMT 2018 and WMT 2019 winner for en-cs both ways)
 - People: Dušan Variš, Jakub Arnold (both PhD students)
 - Results:
 - Trained MT systems in three (actually, four) language pairs
 - Next:
 - Integration in social survey preparation/translation workflow
 - Best practices for use of MT in social surveys



Machine Translation

LINDAT

Repository

Corpus Search

TreeQuery

TreeX

More Apps

About

CLARIN

LINDAT Translation

Success

Translate

Docs

You'll be shown the status (started/queued) of your translation task above this panel.

Source

Target

Czech

English

☐ advanced

Input sentences

Translation

Meč datovaný do mladší doby bronzové je zdobený jednoduchou rytou linií obíhající kolem ostří, které je stále ostré jako břitva, konstatovala archeoložka rychnovského muzea Martina Beková.

„Bronzový meč s jazykovitou rukojetí je datovaný do období kolem roku 1200 před naším letopočtem, náleží kultuře lužické. Nálezy této kultury jsou ve východních Čechách četné, ale neplatí to o mečích,“ řekla Beková.

Podle ní v posledních desetiletích nálezů pravěkých mečů není v celé ČR více než pět.

Přesné místo objevu archeologové tají z důvodu ochrany lokality.

The sword, dated to a younger Bronze Age, is decorated with a simple engraved line orbiting a blade that is still razor-sharp, Rychnov Museum archaeologist Martina Bek noted.

"The bronze sword with the tongue-like handle is dated to around 1200 B.C., belonging to a culture of debt. Findings of this culture are numerous in Eastern Bohemia, but this is not true of swords," Bekova said.

According to her, in recent decades there are no more than five finds of ancient swords in the entire Czech Republic.

The exact location of the discovery is being kept secret by archaeologists to protect the site.

Translate

Choose file

Credits: The service runs systems trained by Martin Popel (en-cs; cs-en; en-fr; fr-en), Shantipriya Parida (en-hi) and Dušan Variš (en-ru, ru-en, en-de, de-en).



Relation to LINDAT (1)



- LINDAT will have to
 - Service the tools once deployed to the SSHOC portal / cloud
 - Perhaps through the European Language Grid?
 - <http://europeanlanguagegrid.eu>
 - Continue development of MT technology
 - To stay abreast with quality
 - To provide domain-specific translation models



Relation to LINDAT (2)



- LINDAT will have to
 - Continue to provide basic language analysis tools
 - UD languages – or at least the 24 EU languages
 - Business partner languages (BRIC etc.)
 - Service the linguistic analysis implementation
 - Domain specific tools?
 - Develop new tools
 - e.g., request for gender-specific job titles



Future needs

- LINDAT/CLARIAH-CZ (in SSHOC/EOSC)
 - Repository service for data (also for collecting data from SSH domains)
 - Part of basic LINDAT RI mission
 - But depends on volume, frequency etc.
 - Services
 - Permanent service needs staff – “legacy lock”
 - Hardware and space
 - Even if green computing-oriented, at least
 - 20-40M CZK for equipment every \$ years
 - 1-2M CZK for power (energy)
 - 4-5M for IT support (HR)



LINDAT in SSHOC



Thank you!