OSCARS

Open Science Clusters' Action for Research & Society

Funded Project

Opravidlo 2.0

Presenter: Hana Zizkova, Masaryk University, iD 0000-0002-6483-6603

Implemented by

MASARYK UNIVERSITY

OSCARS

What problem(s) did you plan to solve?

- Opravidlo is a proofreading service for Czech with explanations
- Currently, it is rule-based with high precision, low recall, and speed issues
- Currently, its UI exists only in Czech, although it is used by language learners

- We will solve the above issues by
  - Using AI models to increase the recall, while keeping high precision
  - Providing explanations and links to Internet Language Reference Book
  - Internationalizing the UI to ease the use by language learners

What have you done to solve the problem?

- We translated the UI into four additional languages based on the size of minorities in the Czech Republic
- We analyzed the types of user texts
- We published a dataset of grammatical rules
- Based on the existing rules, we created a set of datasets for training the language model
- We are training the language model

OSCARS

What are the key results achieved to date and how have you made them available to the broader community?

- Social impact: free public proofreading service
  - Already existing, with thousands of users (>200,000 requests per month)
- Scientific impact
  - **Czech Proofreading Rules**: A collection of proofreading rules, minimum Python code how to used them, and ontology of proofreading phenomena (grammar errors, style, typos) http://hdl.handle.net/11234/1-6001
  - **Three published articles** describing the development of Opravidlo
  - **Errant extended vocabulary**: ERRANT + ERRANT-CZ + Opravidlo extension error-type vocabulary http://hdl.handle.net/11234/1-6065

How will make your results sustainable over time - How will the scientific community/-ies further exploit them?

- Opravidlo 2.0 is freely available proofreading tool, under the licence of NLP Centre of Masaryk University, https://nlp.fi.muni.cz/en/LicenceWebService
- Other outputs such as datasets are licensed under CC-BY-SA within the LINDAT infrastructure.
- Sustainability is ensured by the operation of the Opravidlo 2.0, which we plan to do within the NLP Centre of Masaryk University.

OSCARS

Who is doing it?

- Aleš Horák   0000-0001-6348-109X

- Emma Bednaříková
- Dana Hlaváčková   0000-0002-9918-0958
- Jakub Machura   0000-0002-6623-3064
- Zuzana Nevěřilová   0000-0002-7133-9269
- Pavel Rychlý   0000-0001-5097-4610
- Patrik Stano
- Tereza Vrabcová   0009-0009-5674-3827
- Hana Žižková   0000-0002-6483-6603