COST Action CA21167: "Universality, diversity and idiosyncrasy in language technology"
WG3 Meeting 2023-03-17 Minutes

Session 1

10.45–11.00 Introduction to WG3 (slides)

11.00–11.30 Brainstorming on ideas and expectations

Discussion questions:

- What is most important for you in multilingual and cross-lingual NLP?
- What activities do you think we should prioritize?
- How can we work together to make progress towards our goals?

Points raised:

- Large language models are most important
- Articulating linguistic theories underlying tools
- Defining idiosyncrasy and diversity
- The user perspective is important
- Supporting low-resource languages through cross-lingual technology
- Supporting low-resource languages through annotation tools
- Supporting low-resource languages through data collection
- Supporting low-resource languages with semantics
- Tools for all languages – start with morphology
- Low-resource language is not a homogeneous concept
- Building resources for specific languages (Serbian)
- Linking corpus resources between languages
- Standardized tools applicable to different languages
- Evaluation of tools – coordinate with other WGs
- Tracking evaluation status for different types of tools
- Improved benchmarking and experimental design
- Organize shared tasks

11.30–12.00 Initial discussion on documentation of tools

Discussion questions:

- Which types of tools do we want to include?
- Where do we want to keep the documentation?
- How do we create this documentation/inventory?

Points raised:

- A huge multidimensional matrix
- A shared repository
- Tools shared between typologically similar languages
- Consider end users
- Too many languages have nothing – document what is missing rather than what exists

- Connect to CLARIN
- Flagship project on MWE
- Include all tools or be selective?
- What about commercial tools?
- What about tools without documentation?

WG tasks emerging from the discussion:

- Define multidimensional taxonomy of tools for documentation
- Define infrastructure and procedure for creating documentation

Session 2

13.30–13.35 Recap of Session 1 (for new participants)

13.35–14.20 Initial discussion on evaluation campaigns

Background on goals and previous shared tasks (slides)

Brainstorming – define a novel shared task/evaluation campaign:

- How is the task defined?
- What are the evaluation metrics?
- What kind of data is needed?
- Which languages should be included?

Ideas:

- Task = provide resources for shared tasks (eval metrics, test sets)
- Instead of a shared task, build a dynamic leaderboard for LMs
- Compare “traditional methods” to LMs on UD and MWE data
- UD parsing with only surprise test languages, minimize training data
- NLP tasks on top of UD data using linguistically defined embeddings
- Distinguish similar languages or dialects (for example, using MWEs)
- Objective: make every language appear at the center of the world
- Collect idiom data using LLMs, evaluate on gold data

14.20–14.30 Next steps

Next WG3 meeting in Istanbul, September 8, 2023
We will focus on documentation of tools
Two tasks in preparation for the meeting:

1. A taxonomy of multi- and cross-lingual language technology
2. An infrastructure for multi- and cross-lingual language technology

Volunteers for these tasks are encouraged to contact WG leaders by email

14.30–14.45 Presentation of the European Language Equality project (slides)