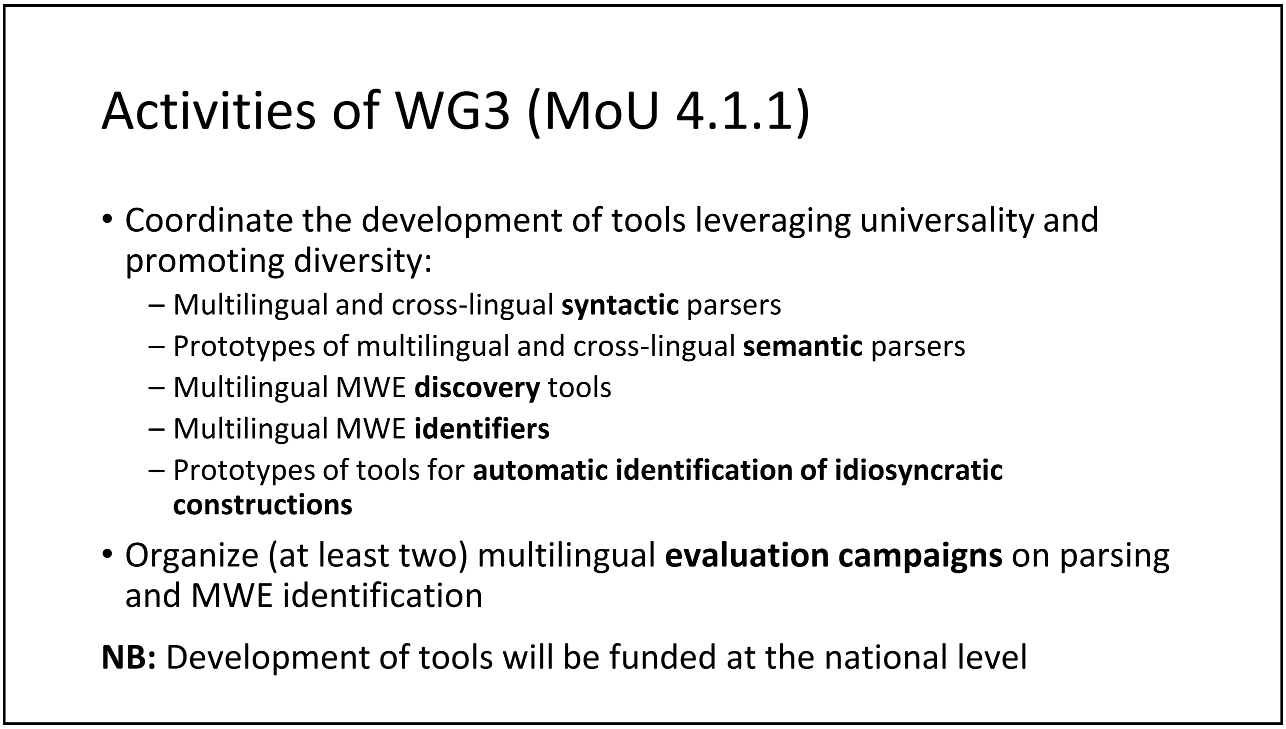
# UniDive WG3

Multilingual and Cross-Lingual Language Technology

## Welcome

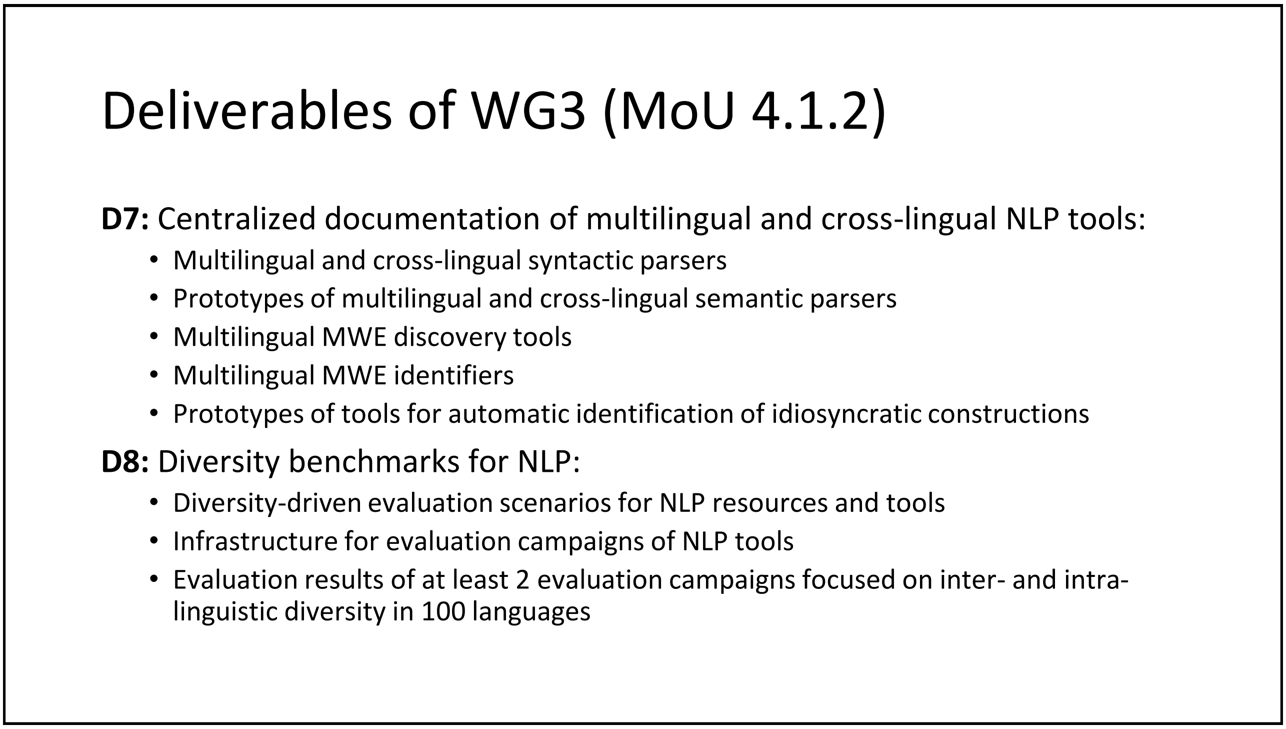
- UniDive is about universality, diversity and idiosyncrasy
- WG3 is about multilingual and cross-lingual processing
- It is up to us to define what this means



## Activities of WG3 (MoU 4.1.1)

- Coordinate the development of tools leveraging universality and promoting diversity:
  - Multilingual and cross-lingual **syntactic** parsers
  - Prototypes of multilingual and cross-lingual **semantic** parsers
  - Multilingual MWE **discovery** tools
  - Multilingual MWE **identifiers**
  - Prototypes of tools for **automatic identification of idiosyncratic constructions**
- Organize (at least two) multilingual **evaluation campaigns** on parsing and MWE identification

**NB:** Development of tools will be funded at the national level



## Deliverables of WG3 (MoU 4.1.2)

**D7:** Centralized documentation of multilingual and cross-lingual NLP tools:

- Multilingual and cross-lingual syntactic parsers
- Prototypes of multilingual and cross-lingual semantic parsers
- Multilingual MWE discovery tools
- Multilingual MWE identifiers
- Prototypes of tools for automatic identification of idiosyncratic constructions

**D8:** Diversity benchmarks for NLP:

- Diversity-driven evaluation scenarios for NLP resources and tools
- Infrastructure for evaluation campaigns of NLP tools
- Evaluation results of at least 2 evaluation campaigns focused on inter- and intra-linguistic diversity in 100 languages

## What do we want to do together?

The MoU defines the goals and scope of UniDive

However:

- The original proposal was written several years ago
- The field of NLP is changing rapidly
- UniDive is essentially a network for collaboration
- We are free to propose new activities consistent with the overall theme

Also:

- An important goal is to create synergy between existing initiatives
- How do we organize ourselves to achieve this?
- Volunteering for different tasks is crucial

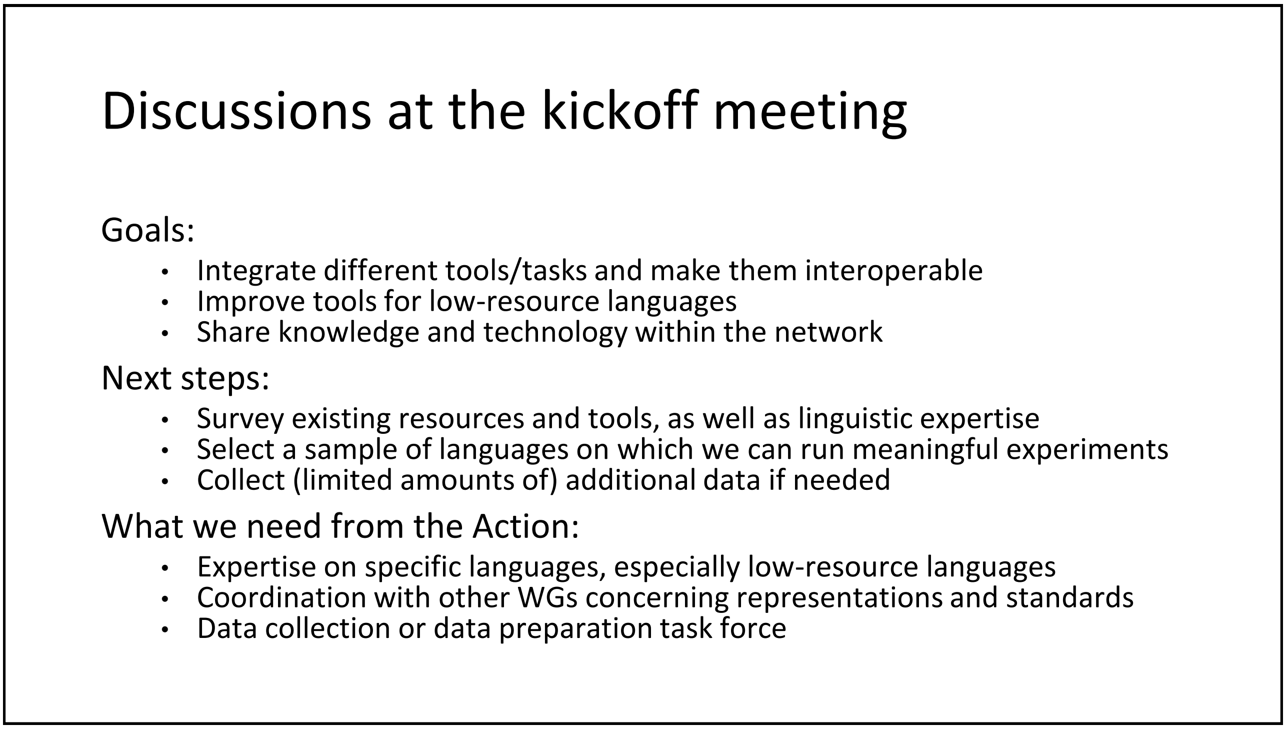
## Plan for this meeting

Session 1:

- Introduction to WG3 (15 min)
- Ideas and expectations – brainstorming (30 min)
- Documentation of tools – initial discussion (30 min)

Session 2:

- Recap of Session 1 (5 min)
- Evaluation campaigns – initial discussion (30 min)
- Next steps – brainstorming (30 min)
- Wrapping up (10 min)



# Discussions at the kickoff meeting

Goals:

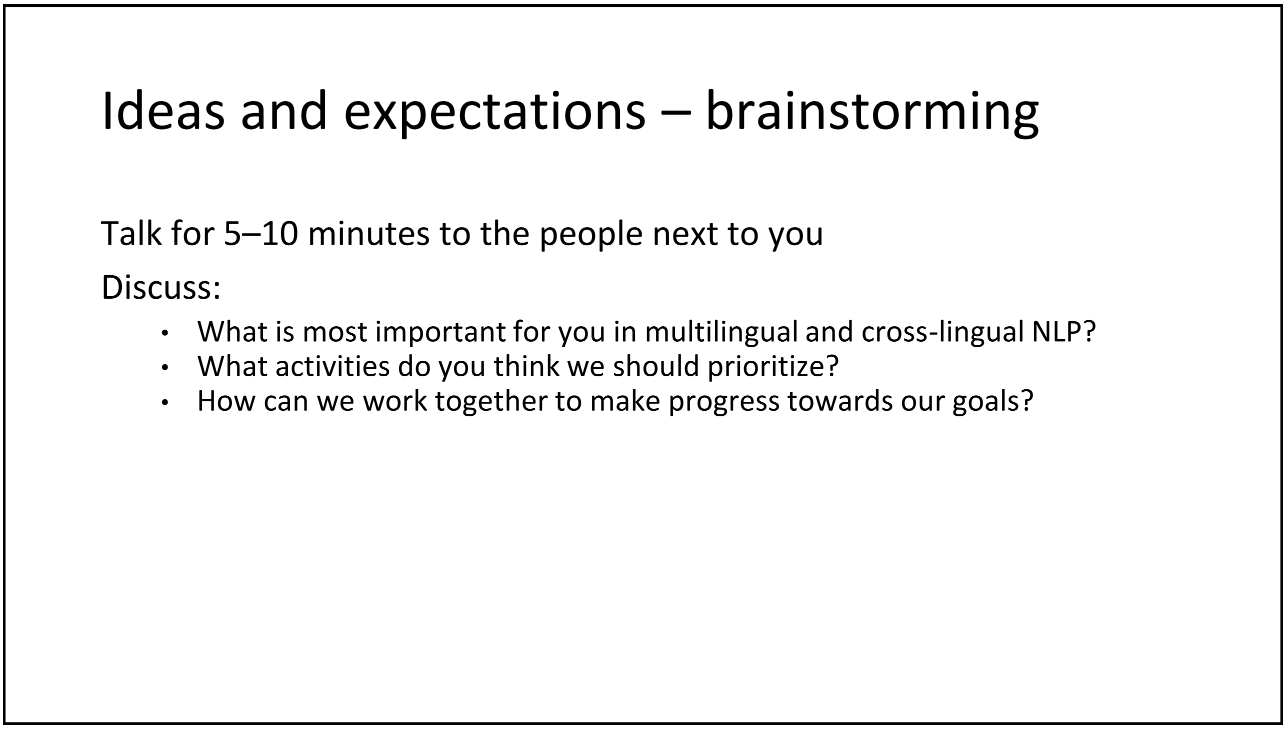
- Integrate different tools/tasks and make them interoperable
- Improve tools for low-resource languages
- Share knowledge and technology within the network

Next steps:

- Survey existing resources and tools, as well as linguistic expertise
- Select a sample of languages on which we can run meaningful experiments
- Collect (limited amounts of) additional data if needed

What we need from the Action:

- Expertise on specific languages, especially low-resource languages
- Coordination with other WGs concerning representations and standards
- Data collection or data preparation task force



# Ideas and expectations – brainstorming

Talk for 5–10 minutes to the people next to you

Discuss:

- 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?

# Initial discussion
# Centralized documentation of NLP tools

**What & Where & How**

- What type of tools do we want to include?
- Where do we want to keep this documentation?
- How are we going to create this documentation/repository?

# What type of tools do we want to include?

**MOU D7:** Centralized documentation of multilingual and cross-lingual NLP tools:

- Multilingual and cross-lingual syntactic parsers
- Prototypes of multilingual and cross-lingual semantic parsers
- Multilingual MWE discovery tools
- Multilingual MWE identifiers
- Prototypes of tools for automatic identification of idiosyncratic constructions

Any tool type to include or exclude from this deliverable?

## Where do we want to keep this documentation?

- What do we currently have as the NLP community?

  Examples:

  1 - Universal Dependencies Project WebPage

    - Over 200 treebanks in over 100 languages & a list of tools working on UD resources
    - Issues are raised on GitHub and resolved through discussion

  2 - CLARIN/LINDAT

    - Language resources and tools
- For Unidive, what do we understand from documentation? (e.g., a repository, a web page, a Github webpage)
- What type of platforms and processes will be used to gather information?

## How are we going to create this documentation?

- How do we organize the work?
- Who are our members experienced on such platforms?
- Volunteers from the Unidive community?

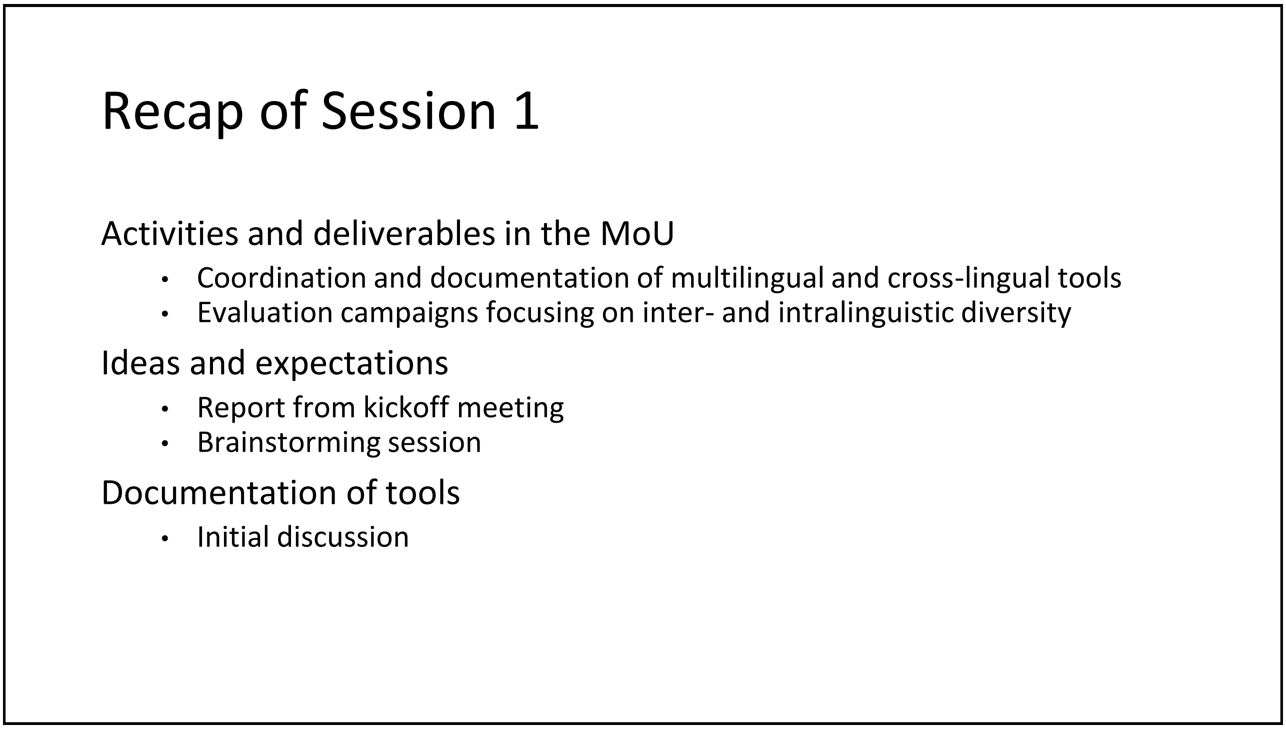
# Plan for this meeting

Session 1:

- Introduction to WG3 (15 min)
- Ideas and expectations – brainstorming (30 min)
- Documentation of tools – initial discussion (30 min)

Session 2:

- Recap of Session 1 (5 min)
- Evaluation campaigns – initial discussion (30 min)
- Next steps – tasks and volunteers (25 min)
- Presentation of ELE (10 min)
- Wrapping up (5 min)



# Recap of Session 1

Activities and deliverables in the MoU

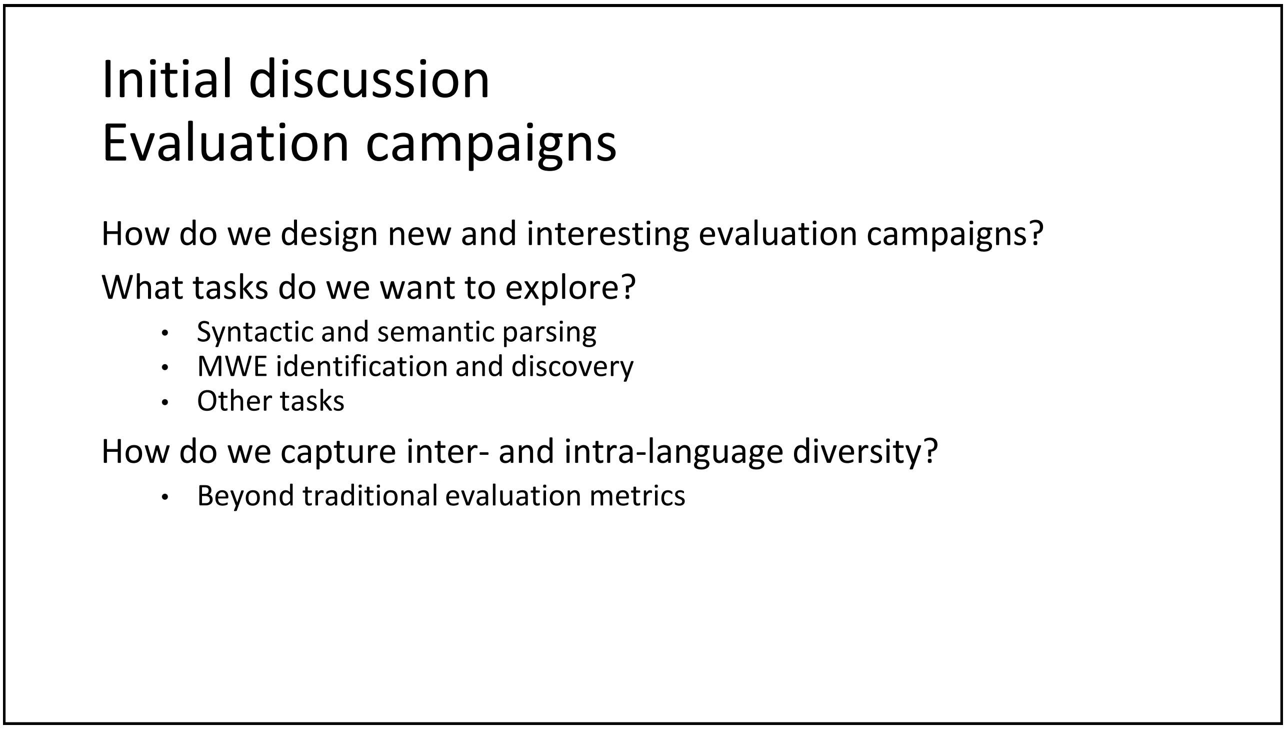
- Coordination and documentation of multilingual and cross-lingual tools
- Evaluation campaigns focusing on inter- and intralinguistic diversity

Ideas and expectations

- Report from kickoff meeting
- Brainstorming session

Documentation of tools

- Initial discussion



# Initial discussion
# Evaluation campaigns

How do we design new and interesting evaluation campaigns?

What tasks do we want to explore?

- Syntactic and semantic parsing
- MWE identification and discovery
- Other tasks

How do we capture inter- and intra-language diversity?

- Beyond traditional evaluation metrics

## CoNLL 2017/2018:
## From Raw Text to Universal Dependencies

Massively multilingual shared tasks:

- Training data for 45/57 languages
- Test data for 49/57 languages (including 14 parallel test sets)

Focus on end-to-end parsing:

- Word segmentation
- Part-of-speech tagging
- Lemmatization
- Morphological analysis
- Labeled dependency parsing

Coincided with the launch of UD 2.0

## IWPT 2020/2021:
## Parsing into Enhanced Universal Dependencies

Similar setup:

- End-to-end parsing including word segmentation

Different output:

- Enhanced UD representations
- Adds implicit dependencies not encoded in the basic representation
- Representation is a general graph (not tree)
- Representation may contain empty nodes

Smaller set of languages

- Training and test data from 17 languages

# PARSEME 2017/2018/2020: Identification of Verbal MWEs

Smaller set of languages:

- 18/20/14 languages

Focus on:

- reflexive verbs
- light verb constructions
- verbal idioms
- verb-particle constructions
- multi-verb constructions
- adpositional verbs

# Other shared tasks

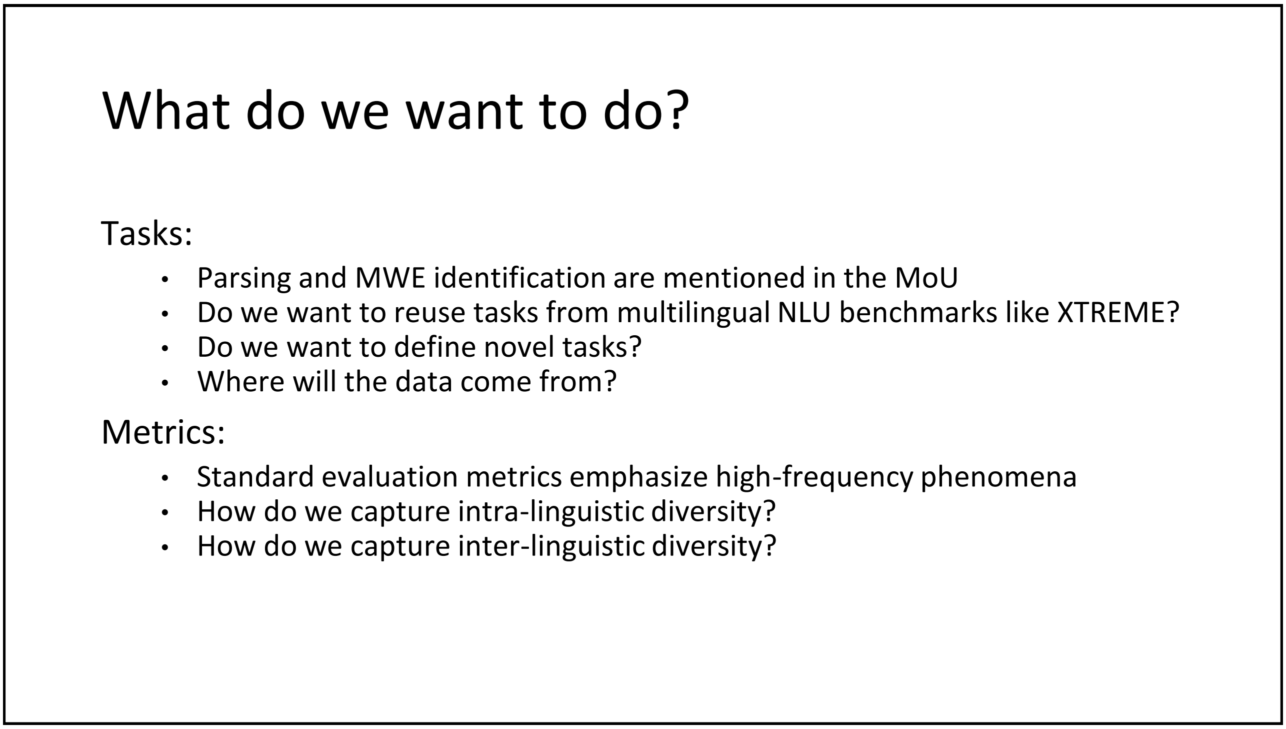
Semantic parsing tasks:

- SemEval 2016: Meaning Representation Parsing (AMR)
- SemEval 2019: Cross-Lingual Semantic Parsing (UCCA)
- CoNLL 2019/2020: Cross-Framework Meaning Representation Parsing

Measuring language complexity (2018)

Multilingual coreference resolution (CRAC 2022)

Multilingual idiomaticity detection (SemEval 2022)



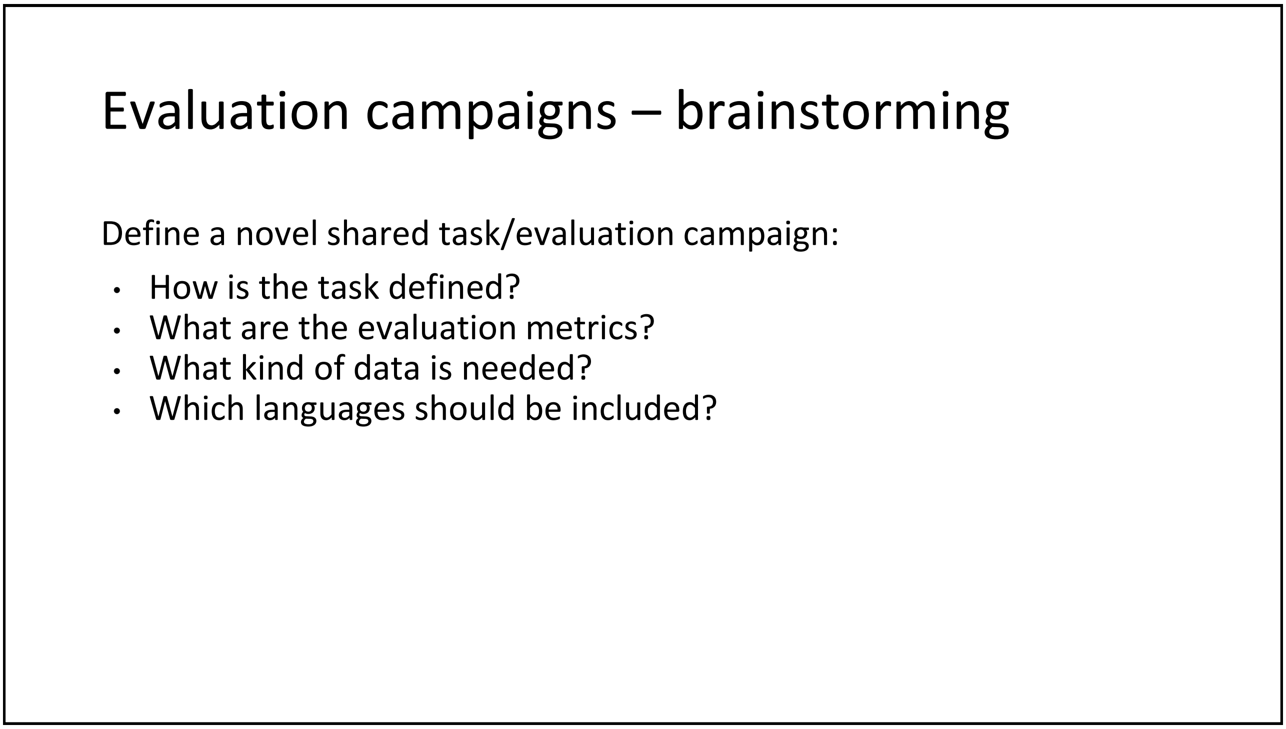
## What do we want to do?

Tasks:

- Parsing and MWE identification are mentioned in the MoU
- Do we want to reuse tasks from multilingual NLU benchmarks like XTREME?
- Do we want to define novel tasks?
- Where will the data come from?

Metrics:

- Standard evaluation metrics emphasize high-frequency phenomena
- How do we capture intra-linguistic diversity?
- How do we capture inter-linguistic diversity?



## Evaluation campaigns – 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?

# Next steps

Next WG3 meeting will be held in İstanbul, September 8, 2023

To discuss:

- What should be the focus of the meeting?
- What can we do to prepare the ground before the meeting?
- How do we organize the work?
- Who will contribute?

Tentative proposal:

- Task 1: A taxonomy for multilingual and cross-lingual LT
- Task 2: An infrastructure for multilingual and cross-lingual LT

Other ideas?