WG3: Multilingual and Cross-Lingual Language Technology

Task 3.3: Conceptions of Multilinguality
preliminary analysis of survey responses

Adriana Pagano (Universidade Federal de Minas Gerais)
Ilan Kernerman (Lexicala by K Dictionaries)

WG3 Meeting – June 10, 2024

UniDive

AIM

- Define the concepts of
  - *multilingual*
  - *cross-lingual*
  - *translingual*

  in the context of Language Technology

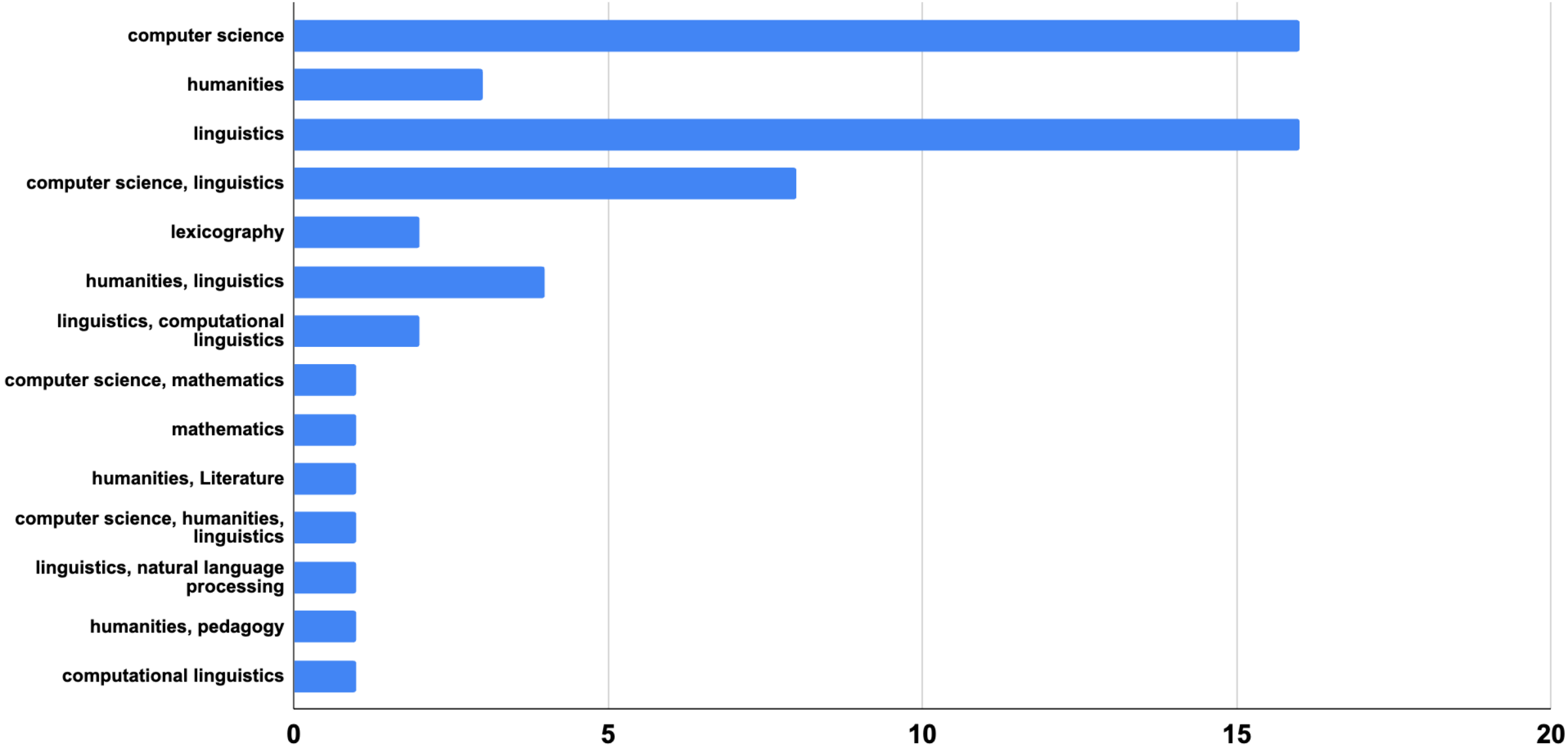
UniDive

SURVEY

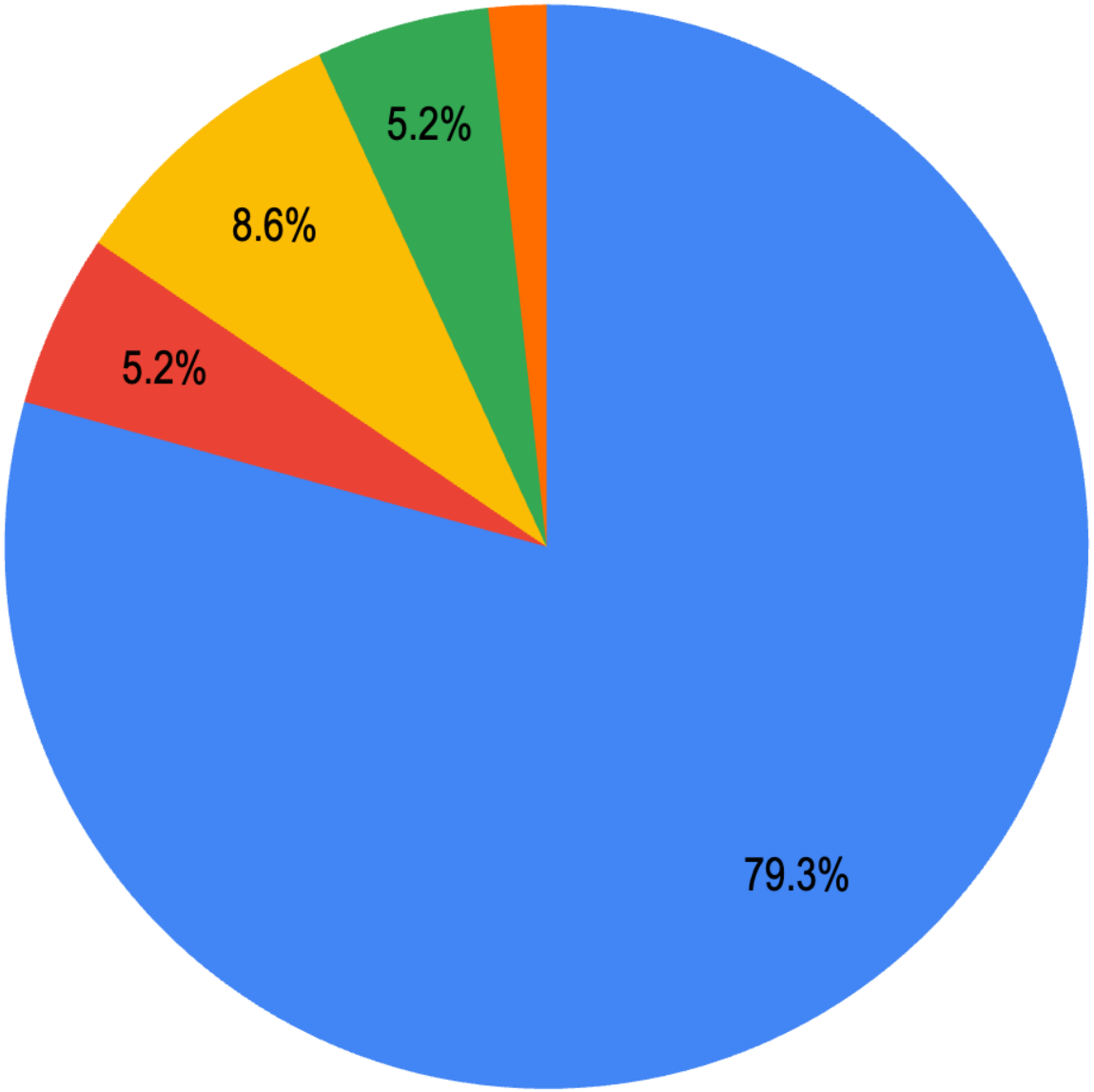
- launched April 6 – closed April 25
- open to all UniDive members
- 58 respondents

UniDive

BACKGROUND


computer science
humanities
linguistics
computer science, linguistics
lexicography
humanities, linguistics
linguistics, computational linguistics
computer science, mathematics
mathematics
humanities, Literature
computer science, humanities, linguistics
linguistics, natural language processing
humanities, pedagogy
computational linguistics
0
5
10
15
20

PRESENT OCCUPATION



- Academic faculty or researcher
- Academic faculty or researcher, Industry researcher
- Student (including PhD student)
- Academic faculty or researcher, Student (including PhD student)
- Other

How many languages are implicated in the concept *multilingual*?

- more than one language
- more than two languages
- 3 or more languages (two is *bilingual*)
- 4 or more languages
- 5 and more would be OK

UniDive

What can be *multilingual*?

- Resource / Data / Corpus
  - contains data for several languages
  - possibly in separate sub-resources / sub-corpora
  - might be parallel
  - often comparable in genre, size, etc.

UniDive

What can be *multilingual*?

- Text
  - having text in several languages

UniDive

What can be *multilingual*?

- App
  - has a user interface in several languages
  - no NLP-specific meaning

UniDive

What can be *multilingual*?

- Model
  - encodes knowledge about several languages
  - can process data in more than one language
  - works on many languages, either all at once (same model) or using a similar approach (separate models)
  - has been trained using data from several languages (e.g. mBERT)
  - used to solve tasks for more than one language (without the need for individual language models)
  - can be applied to several languages, w/out an explicit signal indicating the input language (e.g. a token/input with a language embedding)
  - in training, all languages are treated equally

UniDive

What can be *multilingual*?

- Use
  - using more than two languages in a society or technology

UniDive

What can be *multilingual*?

- Tool
  - applies to many languages by sheer replacement of input data
  - offers the same information in different languages
  - can be applied to several languages, w/out passing the input language as an explicit parameter
  - supports multiple languages specifically, e.g. provided with pre-trained models for multiple languages (if trainable on multiple languages, call it *language-agnostic* NOT *multilingual*)
  - there might be data imbalance, but no formal difference in status between the different languages (as opposed to *cross-lingual* tools that have source and target languages)

UniDive

NEXT STEPS

- Analyze responses for *cross-lingual* and *translingual*
- Present full, detailed results
- Draw conclusions for a paper

UniDive