WG3: Multilingual and cross-lingual language technology

Task 3.3: Conceptions of Multilinguality

UniDive 3rd general meeting
Budapest, 30 January 2025

Adriana Pagano (Universidade Federal de Minas Gerais)
Ilan Kernerman (Lexicala by K Dictionaries)
Nesl̈ihan Önder Özdemïr (Bursa Uludag University)
Natasha Ringblom (Umeå universitet)

AIM

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

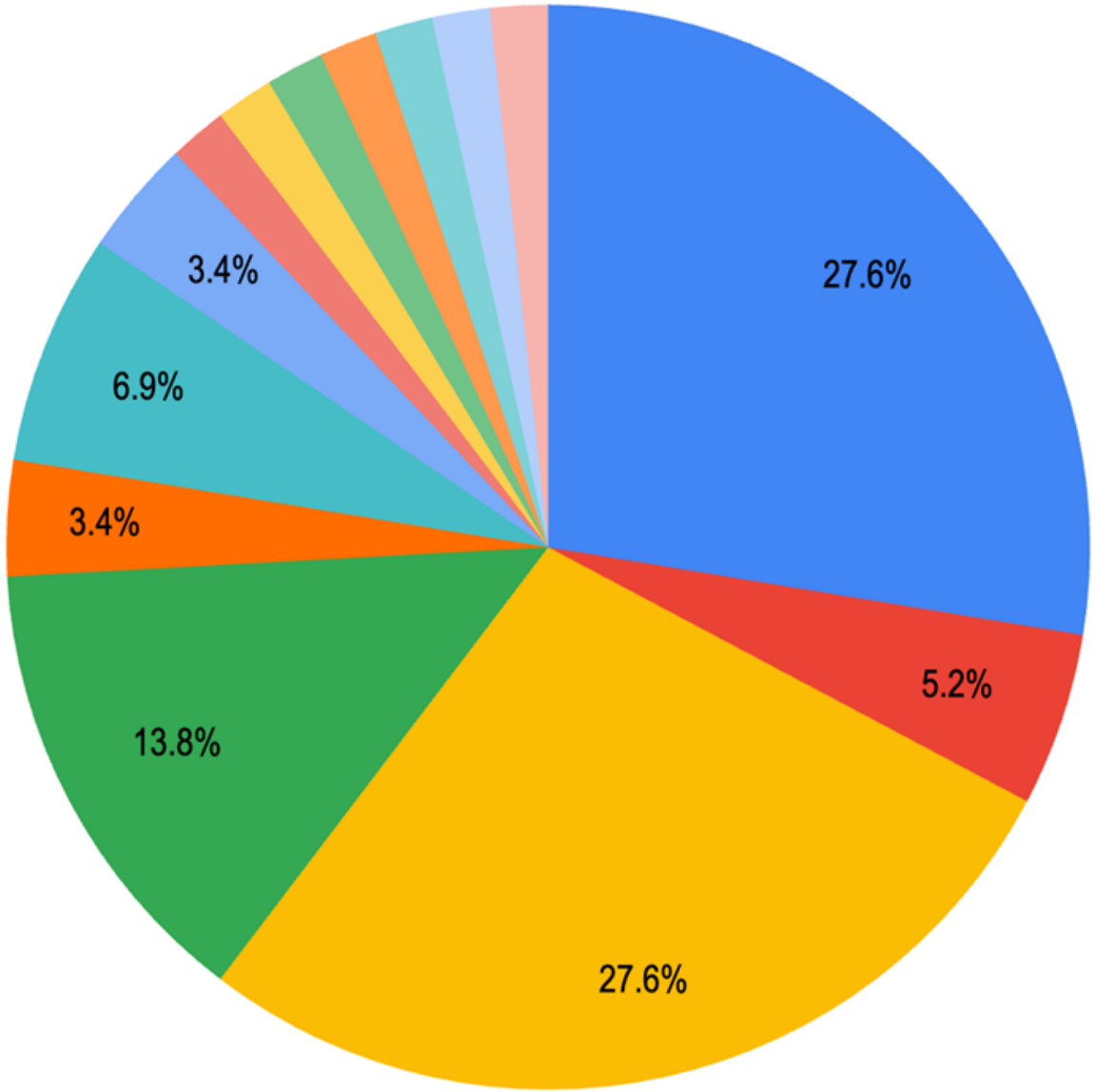
  in the context of Language Technology

ACTIVITIES 2024

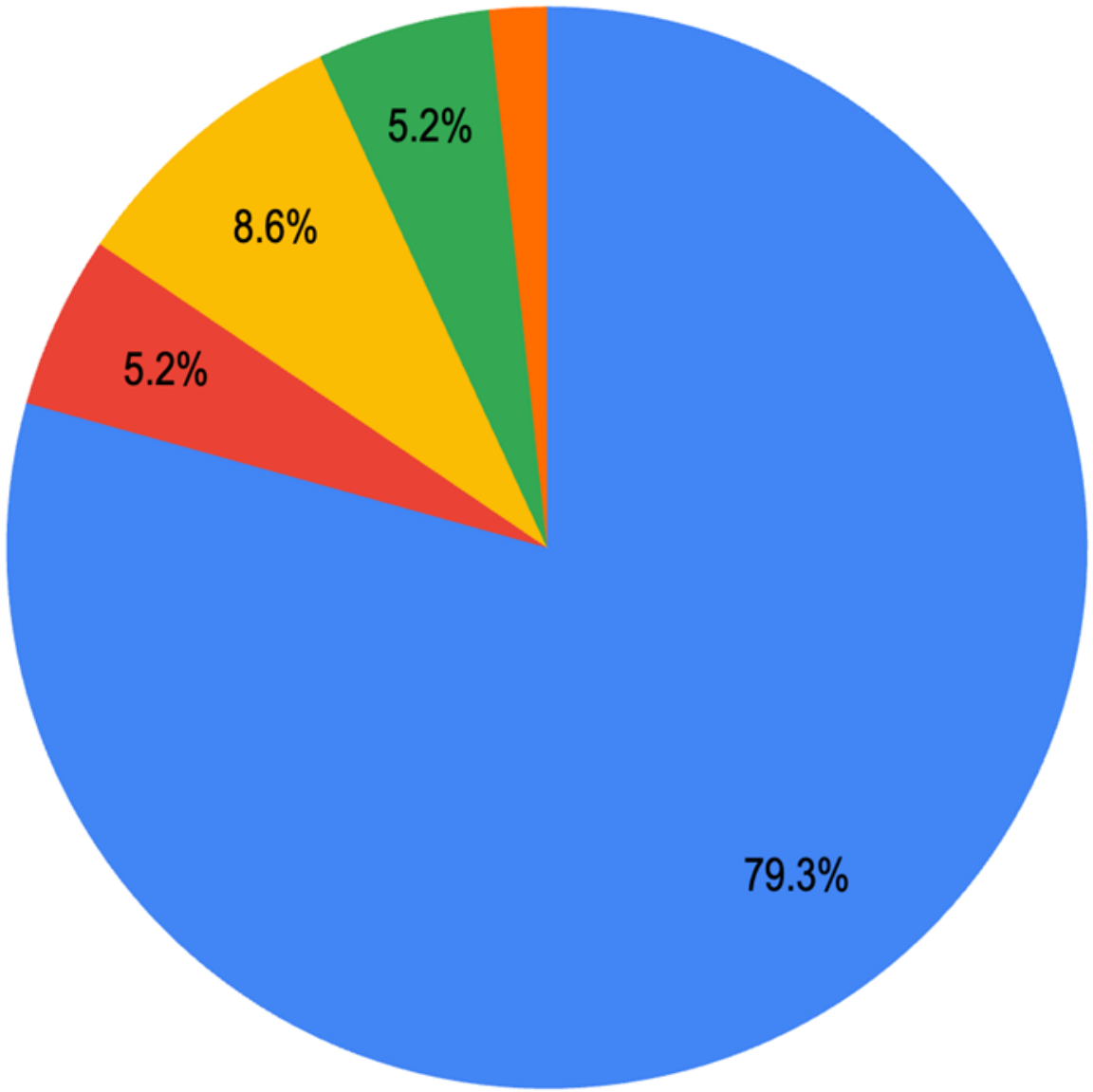
- **April**. Survey design and application
- **May-Oct**. Analysis and report of survey results
- **Nov-Dec**. Exploration of NLP online glossaries and textbooks
- Review of publications from ACL and other major journals
- Proposal currently under review for the Special Issue of *Language and Intercultural Communication*

A SNAPSHOT OF THE SURVEY RESPONDENTS' PROFILE

BACKGROUND


27.6%
5.2%
27.6%
13.8%
3.4%
6.9%
3.4%
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

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

SURVEY INSIGHTS

- All the respondents provided definitions for
  multilingual
  cross-lingual
- Most respondents not familiar with – or never heard of –
  translingual

MULTILINGUAL vs CROSS-LINGUAL

- **Multilingual** applies to
  resource
  data
  corpus
  text
  app
  model
  use
  tool

- **Cross-lingual** applies to
  model
  tool

MULTILINGUAL vs CROSS-LINGUAL

- **Multilingual model**
  - encodes knowledge on several languages
  - can process data in more than one language
  - works on many languages, all at once (same model) or with similar approach (separate models)
  - has been trained using data from several languages
  - used to solve tasks for more than one language

- **Cross-lingual model**
  - built on one language, and then adapted (tuned) to another
  - involves knowledge transfer or linking
  - uses resources in one language to process another language
  - enables interaction and transfer between languages

MULTILINGUAL – OTHER USES

- Systems that can deal with multiple languages
  - no necessary interaction between languages during learning
- *Language-independent* whereas the learning method can be applied to any natural language
  - i.e., having no language-specific requirements
- *Polyglot* as a term to refer to models trained multilingually
  - i.e., a single model for multiple languages, e.g., by parameter sharing between languages in networks during training

CROSS-LINGUAL LEARNING

- A case of *transfer-learning* in which both source and target domains can be sets of languages
  - used to enhance low-resource languages
- *Multilingual learning* considered to be a type of *cross-lingual* learning

TRANSLINGUAL

- Used in other fields, e.g., Applied Linguistics
  - mixing multiple languages e.g.,
    translanguaging
    codeswitching
    codemixing
  - containing features from the different languages in contact
  - For example: language used by the Warao refugees in Brazil (Warao language + Spanish + Brazilian Portuguese)

NLP PAPERS on TERM USAGE

- **Pan, S. J. and Yang, Q. 2009.** A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10), 1345-1359.
- **Panigrahi, S., Nanda, A. and Swarnkar, T. 2021.** A Survey on Transfer Learning. In Mishra, D., Buyya, R., Mohapatra, P. and Patnaik, S. (eds.) *Intelligent and Cloud Computing. Smart Innovation, Systems and Technologies*, vol. 194. Singapore: Springer. https://doi.org/10.1007/978-981-15-5971-6_83
- **Wang, D. and Zheng, T. F. 2015.** Transfer learning for speech and language processing. In *2015 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA),* 1225-1237. IEEE.

POPULAR NLP TEXTBOOKS

- Brief mention of
  - *cross-linguistic* typology
  - *cross-linguistic standards*, e.g., **Universal Dependencies**
  - *multilingual* texts, e.g., *Wikipedia*
  - *(massively) multilingual* language model, e.g. XLM-RoBERTa trained on 100 languages
  - cross-lingual information retrieval
  - multilingual information retrieval

Jurafsky, D. and Martin, J. 2025. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models, 3rd edition. Online manuscript published Jan 2025. https://web.stanford.edu/~jurafsky/slp3.
Mitkov, R. (Ed.). 2022. The Oxford Handbook of Computational Linguistics. Oxford University Press.

NLP GLOSSARIES

- No reference or definition found
  - https://nlpworld.co.uk/nlp-glossary/
  - https://nlp-techniques.org/nlp-glossary-of-terms/
  - https://infosysbpm.com/glossary/natural-language-processing.html
  - https://argilla.io/blog/nlp_glossary/
  - https://nlppod.com/glossary-nlp-terms/

SUGGESTION

- How can technologies deal with **translingual** data,
  - i.e., data involving features of several language systems
- Can single language models be leveraged to approach cases of **translinguality**

UniDive

Thank you for your attention!

Questions?

Comments?