MUNI
FI

# Fine-Grained Language Relatedness for Zero-Shot Silesian-English Translation

**Edoardo Signoroni**

MUNI
FI

# Outline

- Introduction
- Methodology
  - Languages
  - Data
  - Models
- Experiments
- Results
- Conclusions

# Introduction – Low Resource languages

>7000 living languages

plus:
varieties;
dialects;
slangs;
code-switching;
code-mixing;
... and more

but most of these are “Left-Behinds” or **Low-resource languages**, since the biggest MT system online supports a grand total of 133

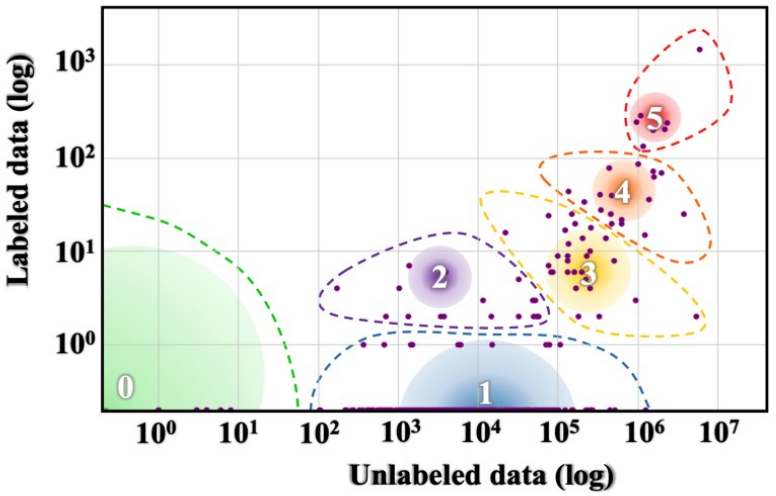



Figure 2: Language Resource Distribution: The size of the gradient circle represents the number of languages in the class. The color spectrum VIBGYOR, represents the total speaker population size from low to high. Bounding curves used to demonstrate covered points by that language class.

Blasi et al. (2022), Joshi et al. (2020)

MUNI
FI

# Introduction – Low Resource languages

Decreasing the digital divide

Dealing with inequalities of information access and production

Mitigating  cross-cultural biases

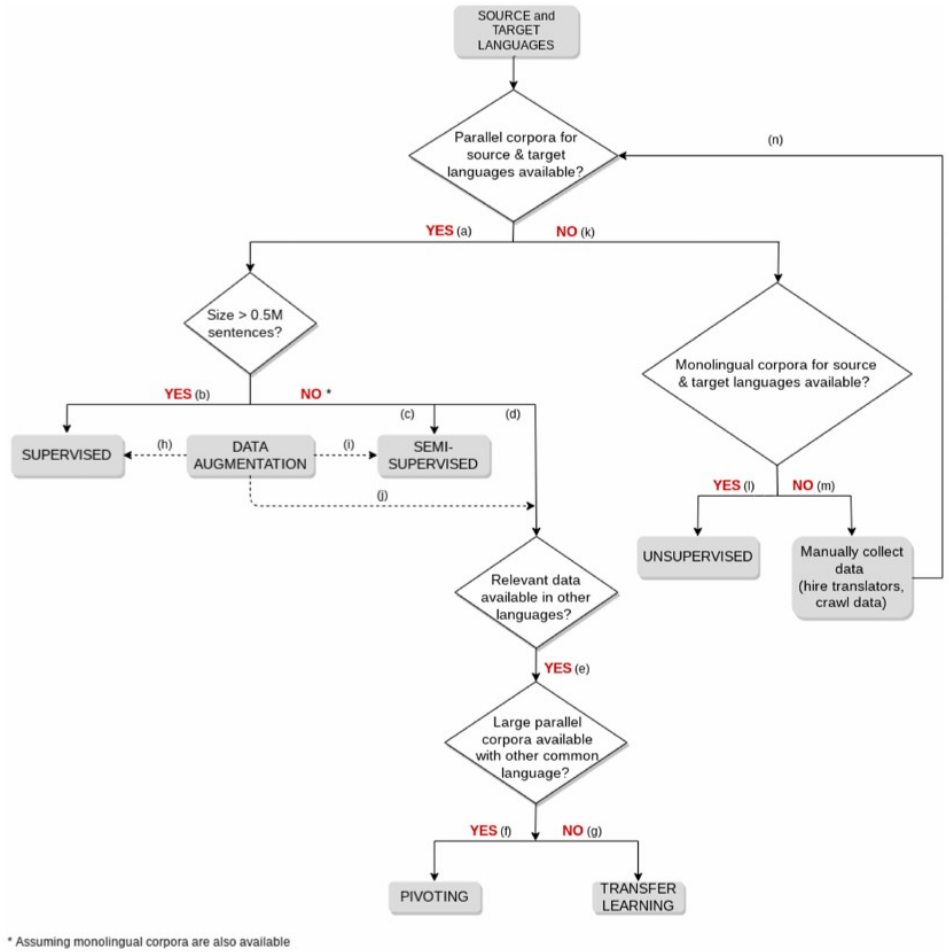
Deploying NLP technologies for underrepresented languages

Understanding cross-linguistic differences

Preserving linguistic diversity

MUNI FI

# Introduction – LR Machine Translation




Parallel source-target data is needed

If these are not available, another model can be used in a **zero-shot** manner

Previous work shows that using data from **related languages** improves the performance of LR MT

However, this work focused on training and fine tuning and on high-level "**horizontal**" relatedness

Instead, we look at relatedness in a fine-grained and "**vertical**" way, using zero-shot Silesian-English translation as our study case

FACULTAS ARTIS INFORMATICÆ • UNIVERSITAS MASARYKIANA •

MUNI
FI

# Languages

| Language | ISO Code | Group | Script | Classification |
|---|---|---|---|---|
| Silesian | szl | West Slavic, Lechitic, Polish-Silesian | Latin | - |
| Polish | pol | West Slavic, Lechitic, Polish-Silesian | Latin | 4 |
| Czech | ces | West Slavic, Lechitic, Czech-Slovak | Latin | 3 |
| Croatian | hrv | South Slavic, Western South Slavic | Latin | 2 |
| Serbian | srp | South Slavic, Western South Slavic | Cyrillic | 1 |
| Ukrainian | ukr | East Slavic, Ukrainian-Rusyn | Cyrillic | 1 |
| Maltese | mlt | Afro-Asiatic, Semitic, Arabic, ... | Latin | 0 |

MUNI
FI

# Data

250k sentence pairs for each language

Croatian, Serbian, Ukrainian, Maltese from MaCoCu (Banon et al. EAMT2022)

Czech from CzEng 2.0 (Kocmi et al. arXiv2020)

Polish from WikiMatrix (Schwenk et al. EACL2021)

Silesian – English from Flores-200 (Goyal et al. 2021)

MUNI
FI

# Models

Multilingual T5 variants:

byT5-small (byte/character level) (Xue et al. TACL2022)

mT5-small (subword level) (Xue et al. NAACL2021)

Both pretrained on the mC4 corpus (15.2% is Slavic text)

MUNI
FI

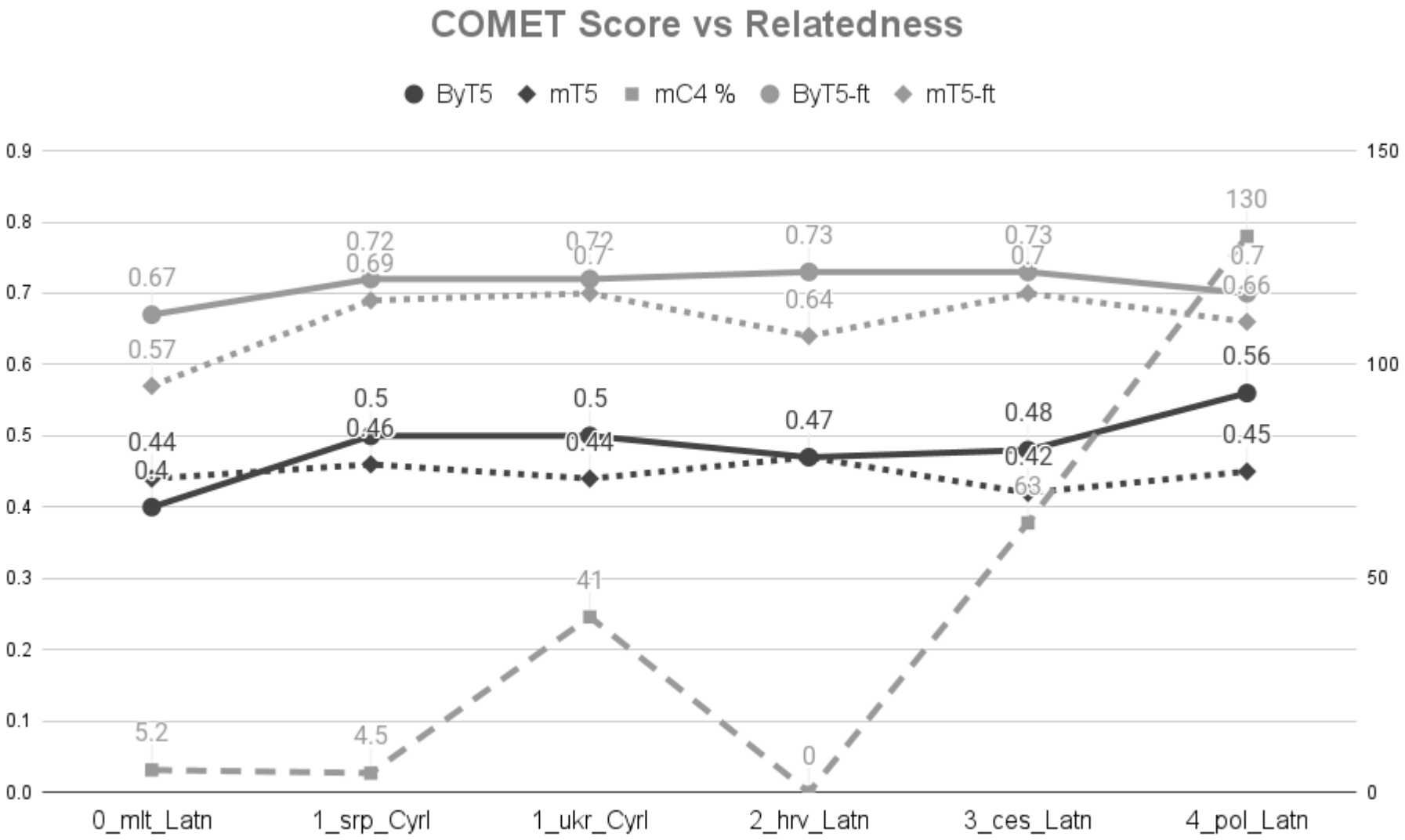
# Experiments

Fine tune the models on the related data

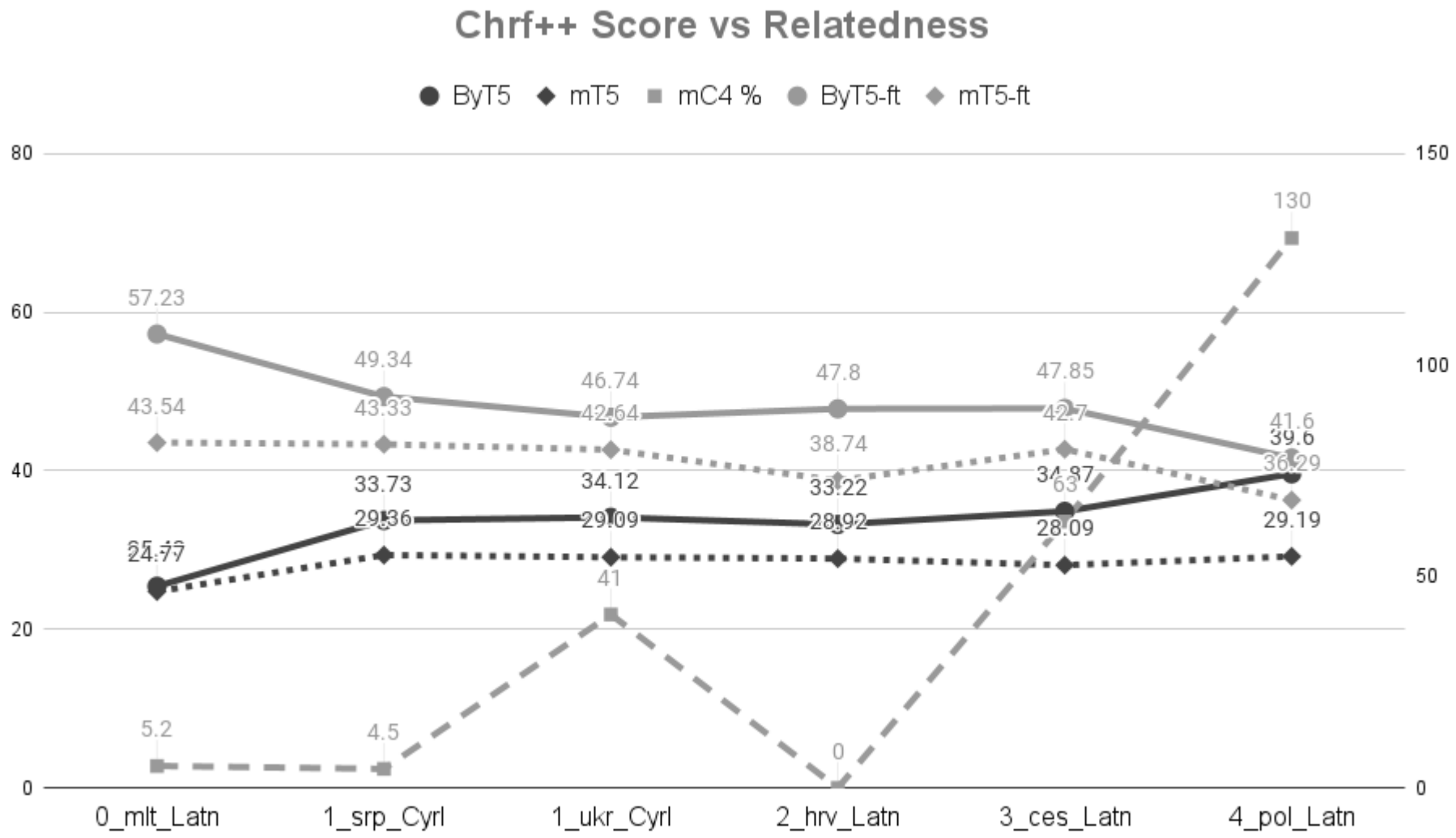
Generate the translations from Silesian

Evaluate the output with ChrF++ (Popovic WMT2017) and COMET (Rei et al. EMNLP2020)

# Results - COMET



# Results - ChrF++



MUNI
FI

# Results - Summary

ByT5 is generally better than mT5

The quality of the system on the fine-tune language does not seem to matter

The amount of pre-training data in the LLM does not seem to lead to big changes

The varies according to the model and to the metric

MUNI
FI

# Conclusions

Confirmed the assumption that using a related language helps

In this specific case, the closest language led to the best results, but the impact of relatedness at a finer scale it is not clear

Future work could involve more language families, and model of a different size (since preliminary results suggest that *large* models may have a different behavior)

MUNI
FI

NLP Centre

@edo_signoroni

edoardosignoroni.github.io

FACULTAS ARTIS INFORMATICÆ • UNIVERSITAS MASARYKIANA •