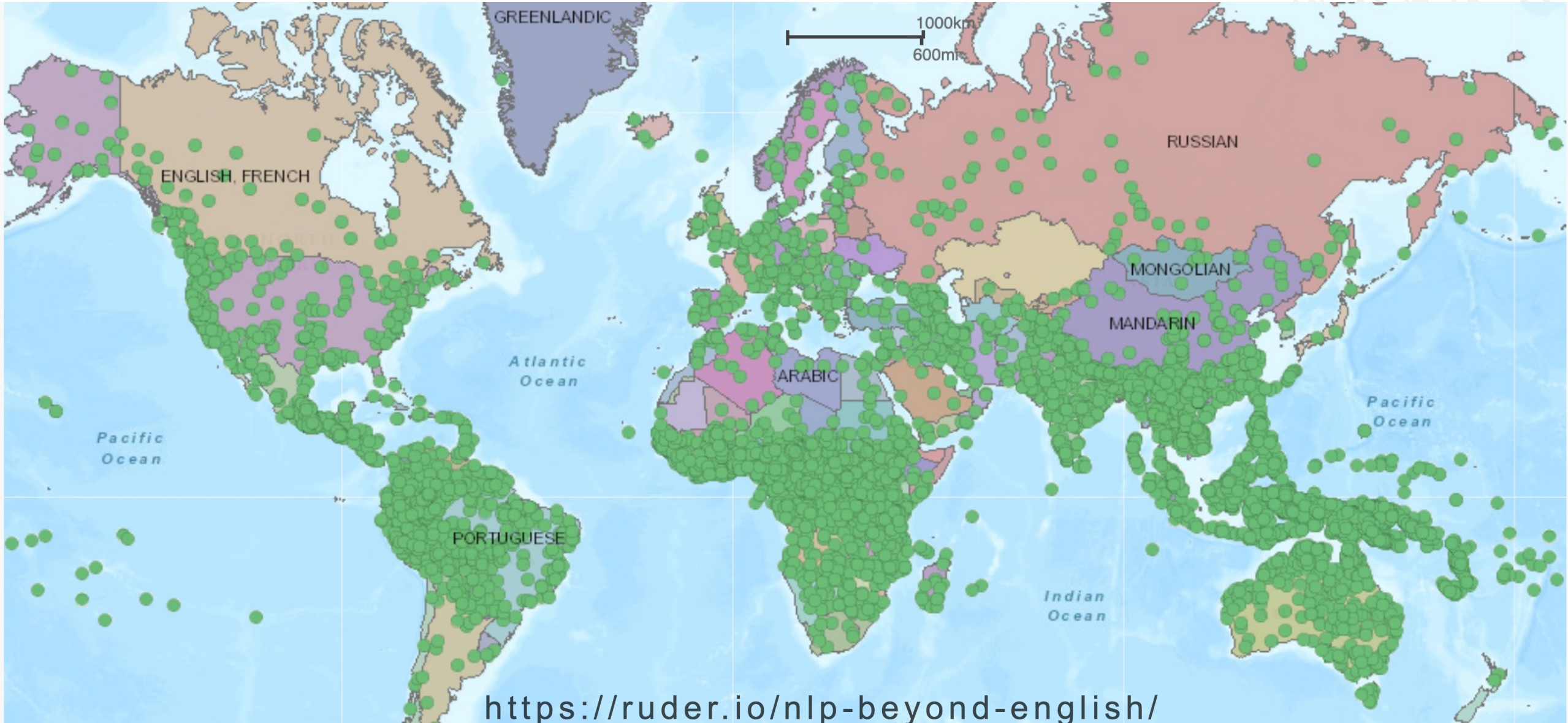
Low Resource Languages

Meriem Beloucif

UPPSALA
UNIVERSITET

Low Resource Languages



https://ruder.io/nlp-beyond-english/

What are low resource languages?

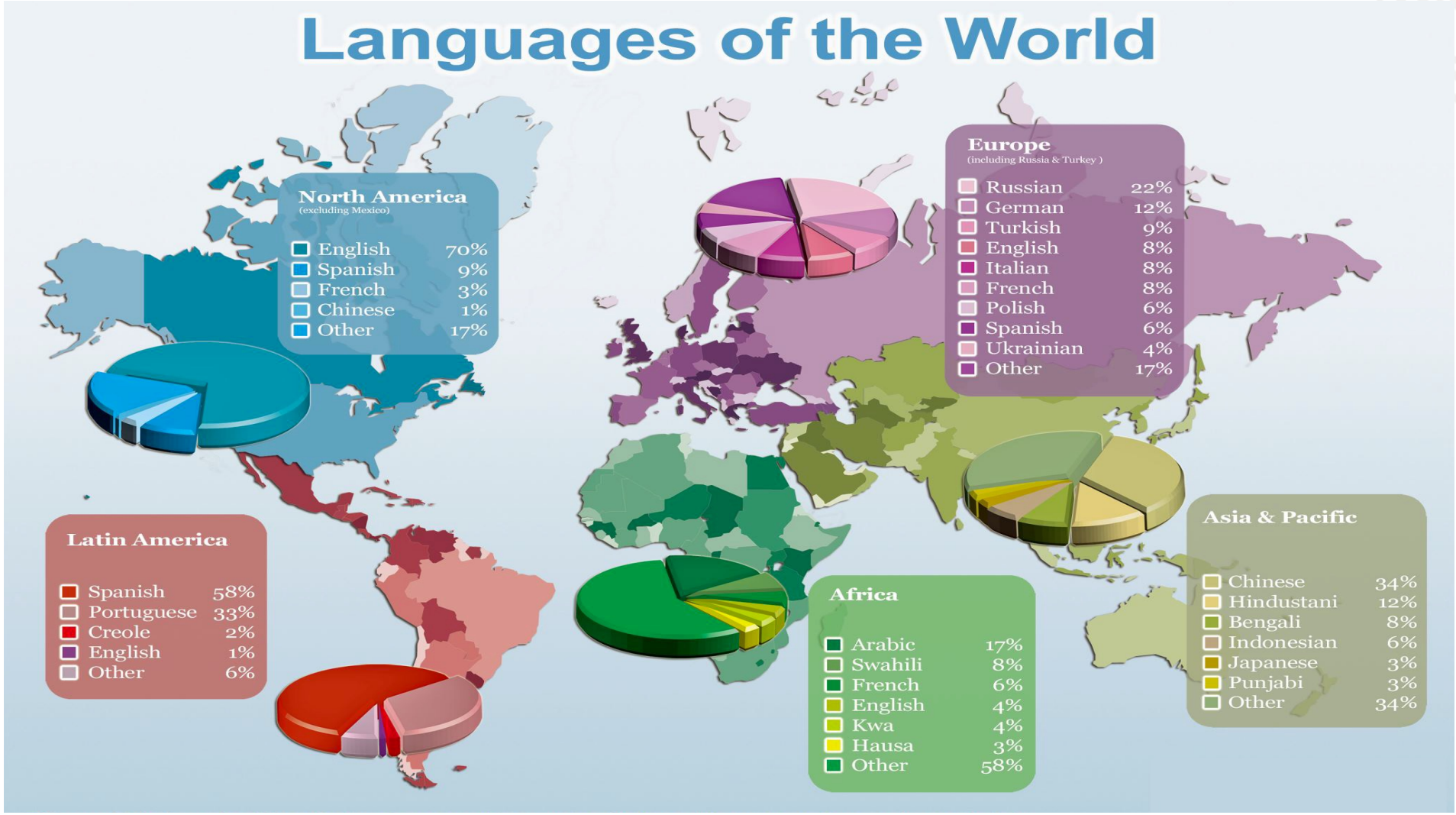
- **a minority language**
  - Xhosa which is spoken by 7 million people in South Africa
- **a less-studied language**
  - Turkish which is spoken by around 88 million people
- **a resource poor language**
  - Bengali although it is spoken by around 200 million people

UPPSALA UNIVERSITET

How many languages are there in the world?

UPPSALA
UNIVERSITET

There are around 7100 languages worldwide but 40% are threatened



https://www.ethnologue.com/

UPPSALA UNIVERSITET

But...

you can only Google search in just over 130 different languages!

UPPSALA
UNIVERSITET

Why is NLP for LR languages important?

There are several reasons why NLP for LRL is crucial (Ruder 2020):

- **The societal perspective**
- **The linguistic perspective**
- **The ML perspective**
- **The cultural and normative perspective**
- **The cognitive perspective**

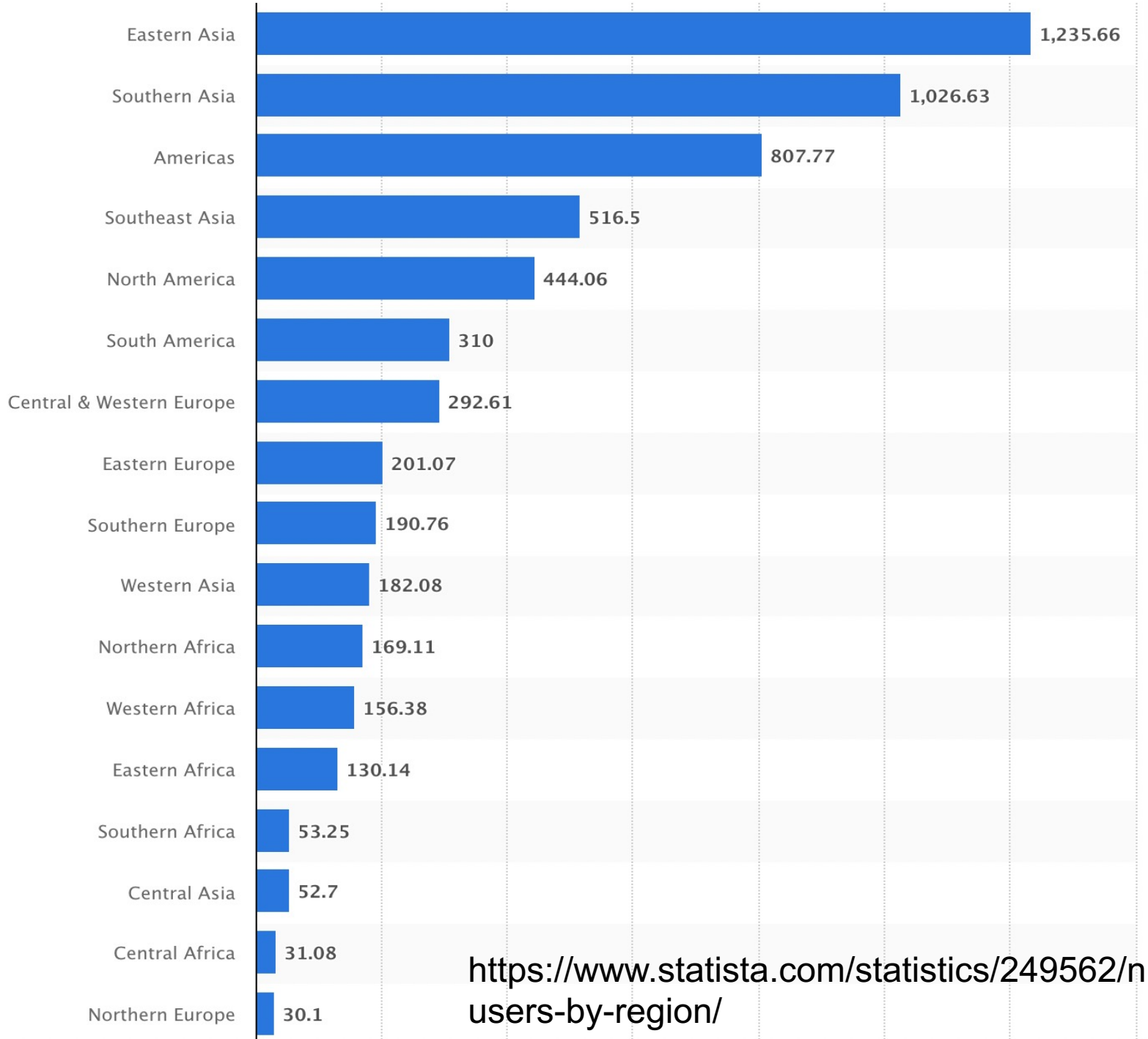
UPPSALA
UNIVERSITET

The societal perspective

UPPSALA
UNIVERSITET

Inequality of information

https://www.ethnologue.com/



The Web does not just connect machines, it connects people - Tim Berners-Lee

UPPSALA UNIVERSITET

https://www.statista.com/statistics/249562/number-of-worldwide-internet-users-by-region/

The linguistic perspective

- High resource language such as English, German and Chinese are not linguistically representative of the world's other languages
  - mostly Indo-European languages, morphologically poor
- By working on automatizing English, and ignoring other languages, we are missing potential insights that could help understand the correlation between different languages (Artetxe et al., 2020)

UPPSALA UNIVERSITET

The linguistic perspective

- There are 192 typological features or structural and semantic properties of a language
  - such as order of subject, object, and verb in a language
- 48% of all feature categories exist only in the low-resource languages (Joshi et al., 2020)
- How can we build efficient models that generalize properly if we ignored such a large subset of typological?

The ML perspective

- Current models are stochastic parrots (Bender et al. 2021)
- Neural models often overlook the complexities of morphologically rich languages (Tsarfaty et al., 2020)
- Focusing on high-resource languages, we are building methods that work well only when large amounts of labelled and unlabelled data are available

UPPSALA
UNIVERSITET

The cultural and normative perspective

- By building models around English, we are building agents and AI systems that normalizes concepts as the source English does
- Taboo topics vary from a society to another
- When travelling, we always ask about societal norms for the societies we are visiting, our AI systems should do the same

UPPSALA UNIVERSITET

The cognitive perspective

How do children learn languages?

UPPSALA
UNIVERSITET

In the LR group, we will focus on 3 themes

- Progress of Low Resource in NLP
  - State of the art
  - Overview of LRL NLP
- Transfer Learning for different NLP Tasks
  - Study transfer learning methods
  - Transfer learning in three different NLP applications
- Low Resource in the era of Large Models
- …

UPPSALA
UNIVERSITET

Potential Projects (Not limited to)

- Add a new language to Multilingual BERT
- Sentiment classification using crosslingual transfer for you choice language
- Exploring data augmentation for fine-tuning multilingual BERT towards a specific task (NER, POS)
- Crosslingual projection of labels, a case study between english Semantic role labeling and low resource semantic role labeling

UPPSALA
UNIVERSITET

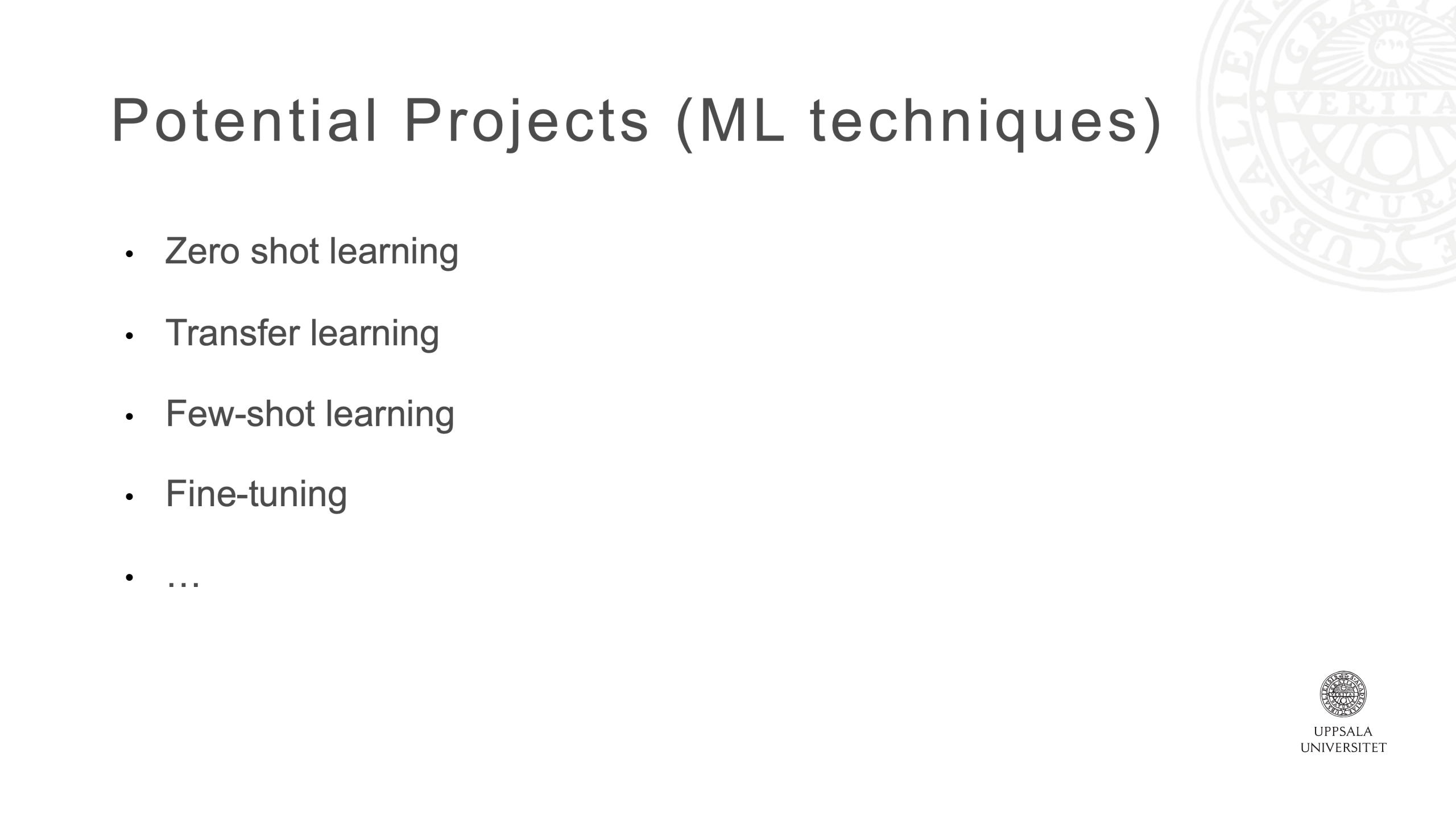
Potential Projects (NLP fields)

- Part of speech tagging
- Semantic parsing
- Question Answering
- Machine Translation
- Preprocessing.
- Sentiment Analysis
- ….

UPPSALA
UNIVERSITET

Potential Projects (ML techniques)

- Zero shot learning
- Transfer learning
- Few-shot learning
- Fine-tuning
- …



UPPSALA
UNIVERSITET