DIKU Bits
TUESDAY LECTURES

13 SEPTEMBER 2022

What do multilingual language models learn about language?

Karolina Ewa Stanczak
PhD, Natural Language Processing, DIKU

12.15 - 13.00
diku.dk/diku-bits

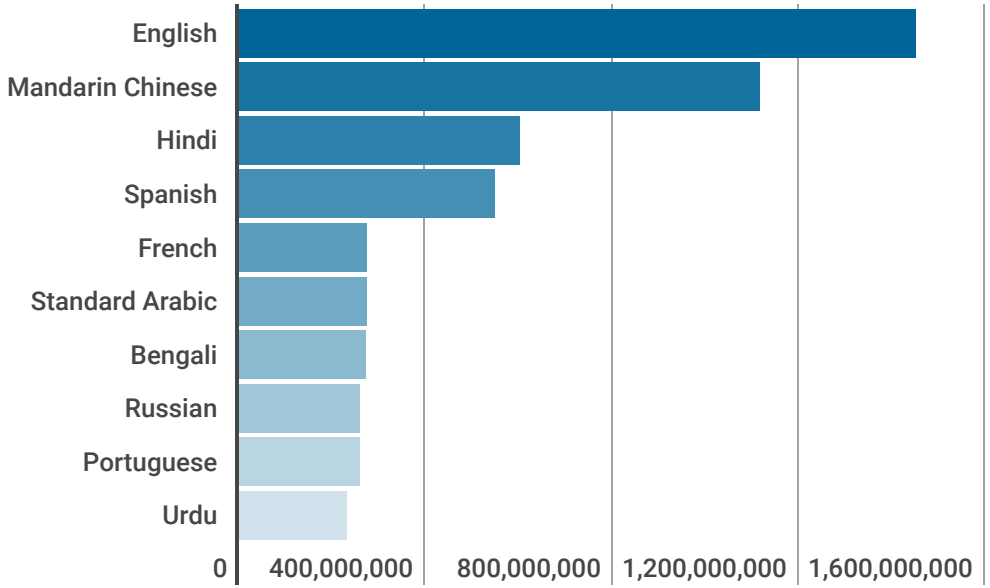
Multilingual World

Ethnologue
@theEthnologue

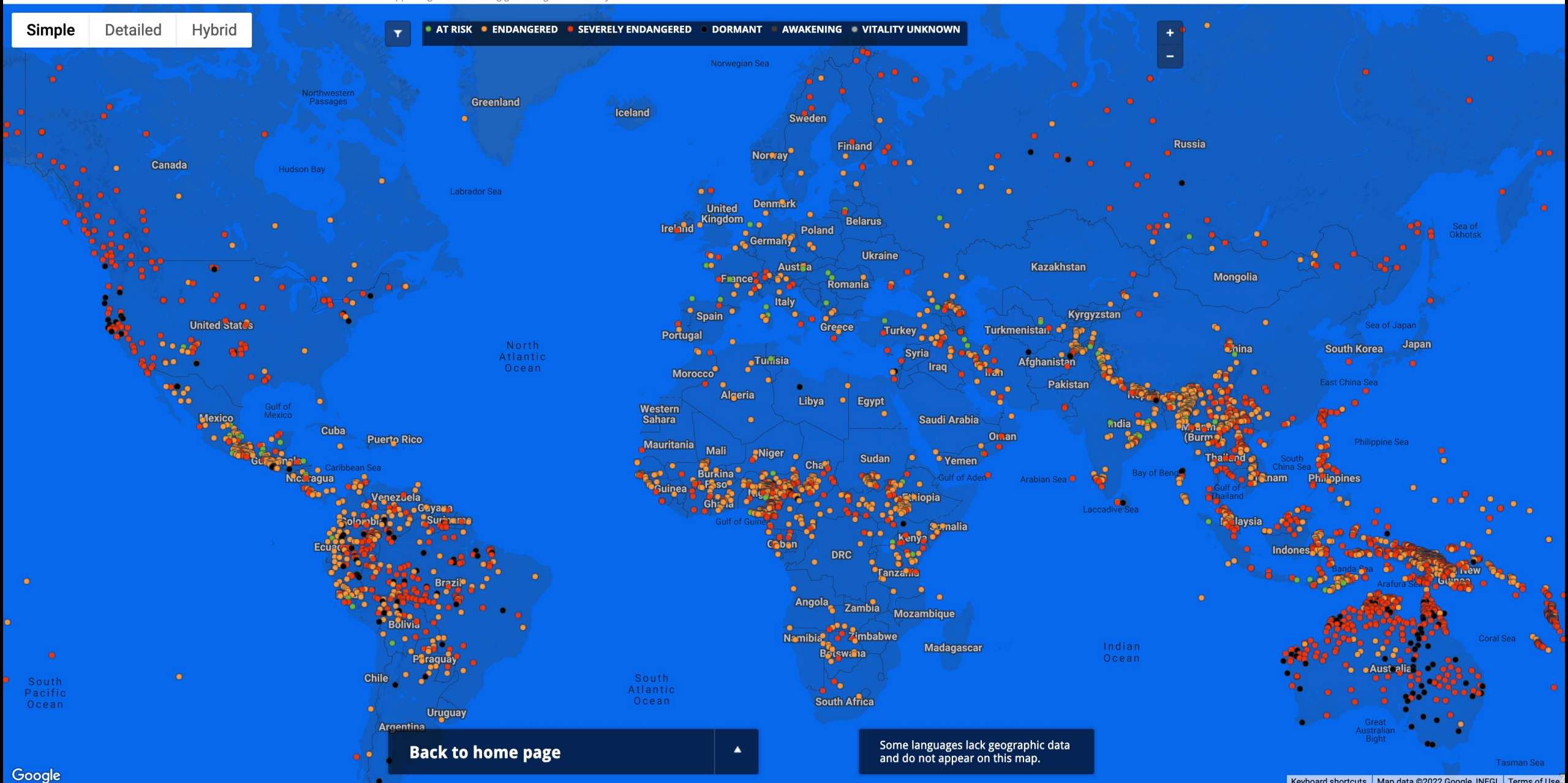
7,151 languages are spoken in the world, as of 2022. 23 languages account for more than half the world's population.

5:00 PM · Feb 28, 2022 · Twitter Web App

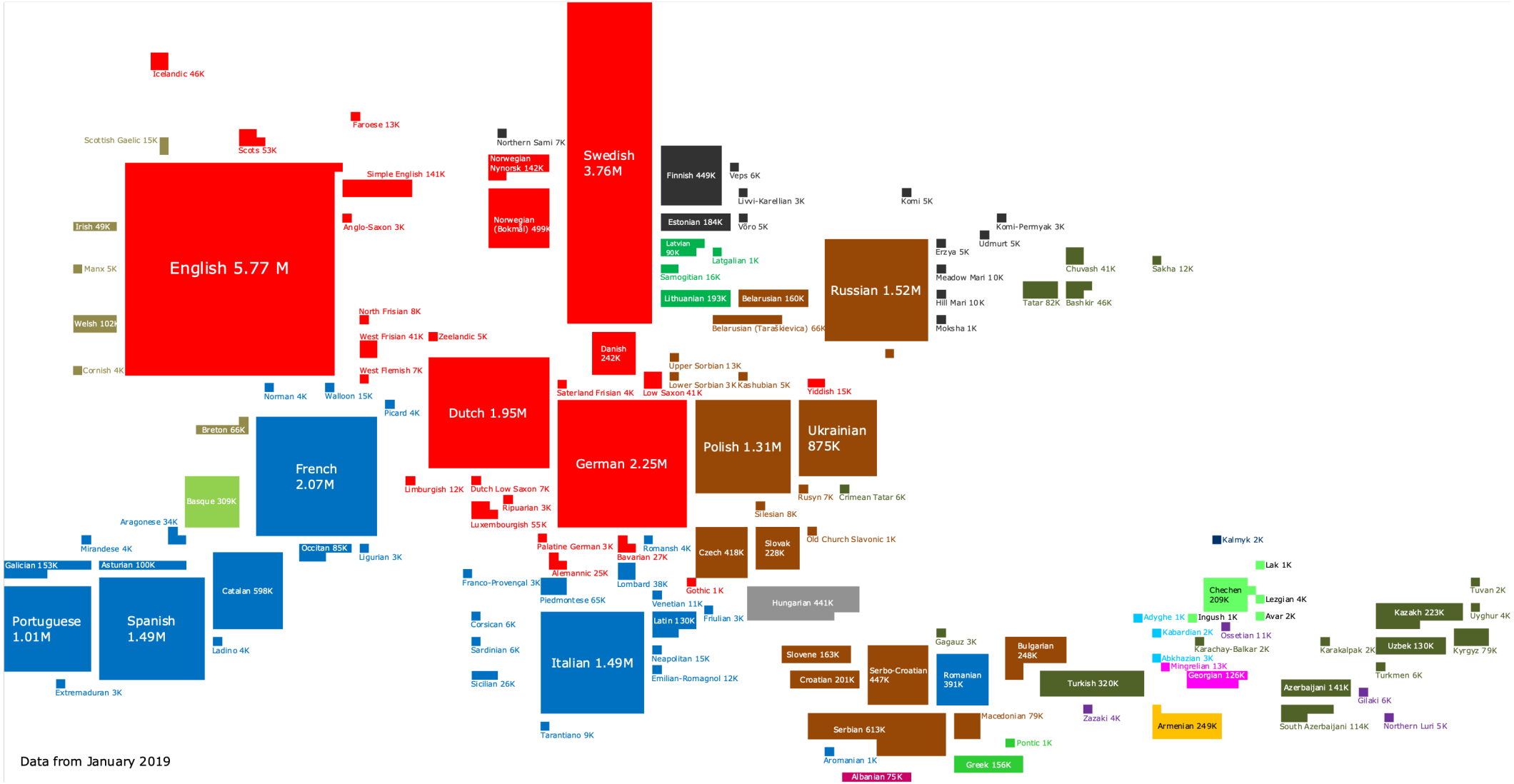
Top 10 most spoken languages, 2022



Ethnologue


Simple
Detailed
Hybrid
AT RISK
ENDANGERED
SEVERELY ENDANGERED
DORMANT
AWAKENING
VITALITY UNKNOWN
Norwegian Sea
Greenland
Iceland
Northwestern Passages
Canada
Hudson Bay
Labrador Sea
Sweden
Finland
Norway
Russia
United Kingdom
Denmark
Ireland
Belarus
Poland
Germany
Ukraine
Austria
France
Romania
Italy
Spain
Portugal
Greece
Kazakhstan
Mongolia
Kyrgyzstan
Turkmenistan
Turkey
Syria
Iraq
Iran
Afghanistan
Pakistan
China
Sea of Okhotsk
Sea of Japan
South Korea
Japan
East China Sea
Philippine Sea
North Atlantic Ocean
United States
Mexico
Gulf of Mexico
Cuba
Puerto Rico
Caribbean Sea
Nicaragua
Venezuela
Guyana
Suriname
Colombia
Ecuador
Brazil
Bolivia
Paraguay
Chile
Uruguay
Argentina
South Pacific Ocean
South Atlantic Ocean
Morocco
Tunisia
Algeria
Libya
Egypt
Western Sahara
Mauritania
Mali
Niger
Chad
Sudan
Burkina Faso
Guinea
Ghana
Gulf of Guinea
Nigeria
Cameroon
Ethiopia
Somalia
Kenya
DRC
Tanzania
Angola
Zambia
Mozambique
Namibia
Zimbabwe
Botswana
South Africa
Madagascar
Saudi Arabia
Oman
Yemen
Gulf of Aden
Arabian Sea
Laccadive Sea
India
Bay of Bengal
Thailand
Gulf of Thailand
South China Sea
Philippines
Malaysia
Indonesia
Banda Sea
Arafura Sea
New Guinea
Coral Sea
Australia
Great Australian Bight
Tasman Sea
Indian Ocean
Back to home page
Some languages lack geographic data and do not appear on this map.
Google
Keyboard shortcuts
Map data ©2022 Google, INEGI
Terms of Use

High- and low-resource languages



European Wikipedias article count 2019 map

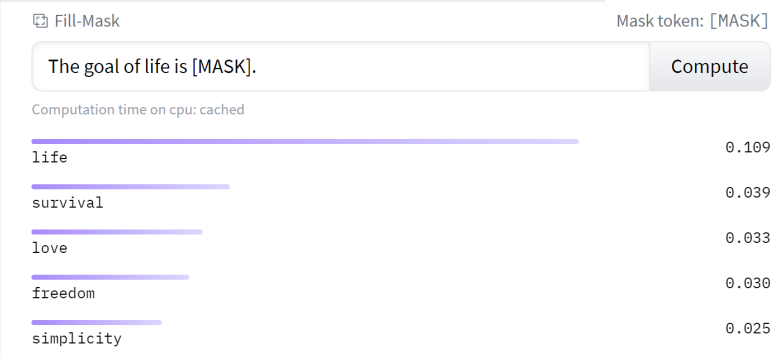
Language Models

Language model is a probability distribution over sequences of words with language modelling being the task of predicting what word comes next.

- Formal definition:

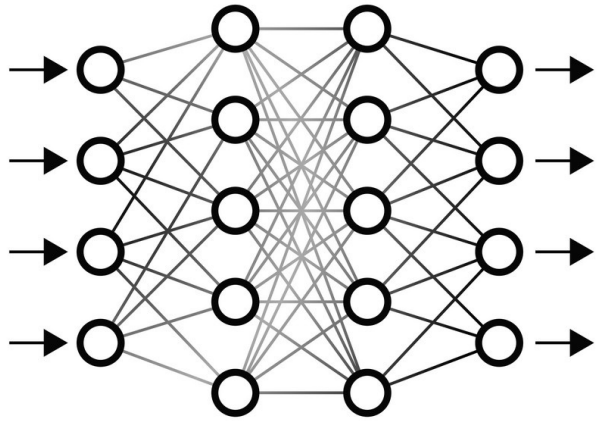
> Given a sequence of words $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(n)}$, compute the probability distribution of the next word $\mathbf{x}^{(n+1)}$:
>
> $$P(\mathbf{x}^{(n+1)} | \mathbf{x}^{(n)}, \ldots, \mathbf{x}^{(1)})$$
>
> where $\mathbf{x}^{(n+1)}$ can be any word from a given vocabulary.

- Areas of application: language generation, error correction, translation



Multilingual Language Models

Danish
English
Hindi
Kinyarwanda
Polish
…



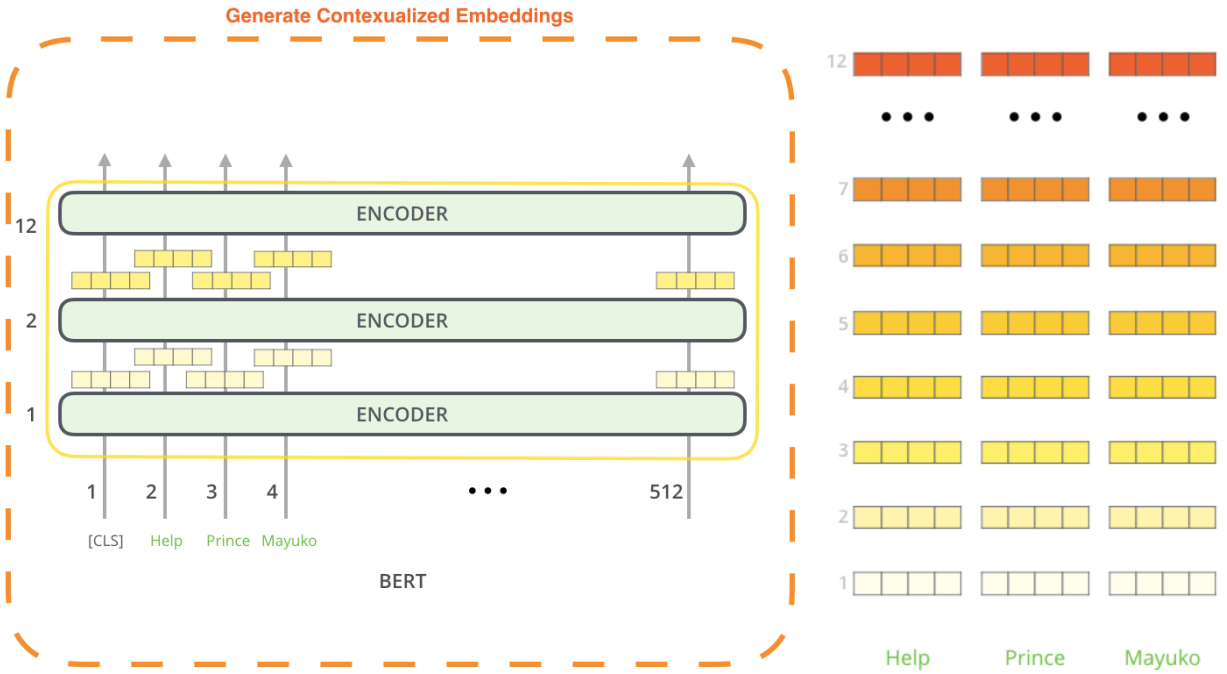
Danish
English
Hindi
Kinyarwanda
Polish
…

- A single model is trained on a mixed dataset from multiple languages.
- Current multilingual language models are trained on as many as 100 or 200 languages!

Contextualised Embeddings

Pre-trained language models can be used to create contextualised word embeddings.

- Word embeddings: vector representation of a word in a vocabulary
- Encoding: mapping an input sequence to a sequence of continuous representations



The output of each encoder layer along each token's path can be used as a feature representing that token.

Success of Multilingual Language Models

- Multilingual pre-trained models display an impressive ability to transfer knowledge between languages as well as to perform zero-shot learning!

Model	D	#vocab	en	fr	de	ru	zh	sw	ur	Avg
Monolingual baselines										
BERT	Wiki	40k	84.5	78.6	80.0	75.5	77.7	60.1	57.3	73.4
	CC	40k	86.7	81.2	81.2	78.2	79.5	70.8	65.1	77.5
Multilingual models (cross-lingual transfer)										
XLM-7	Wiki	150k	82.3	76.8	74.7	72.5	73.1	60.8	62.3	71.8
	CC	150k	85.7	78.6	79.5	76.4	74.8	71.2	66.9	76.2
Multilingual models (translate-train-all)										
XLM-7	Wiki	150k	84.6	80.1	80.2	75.7	78	68.7	66.7	76.3
	CC	150k	**87.2**	**82.5**	**82.9**	**79.7**	**80.4**	**75.7**	**71.5**	**80.0**

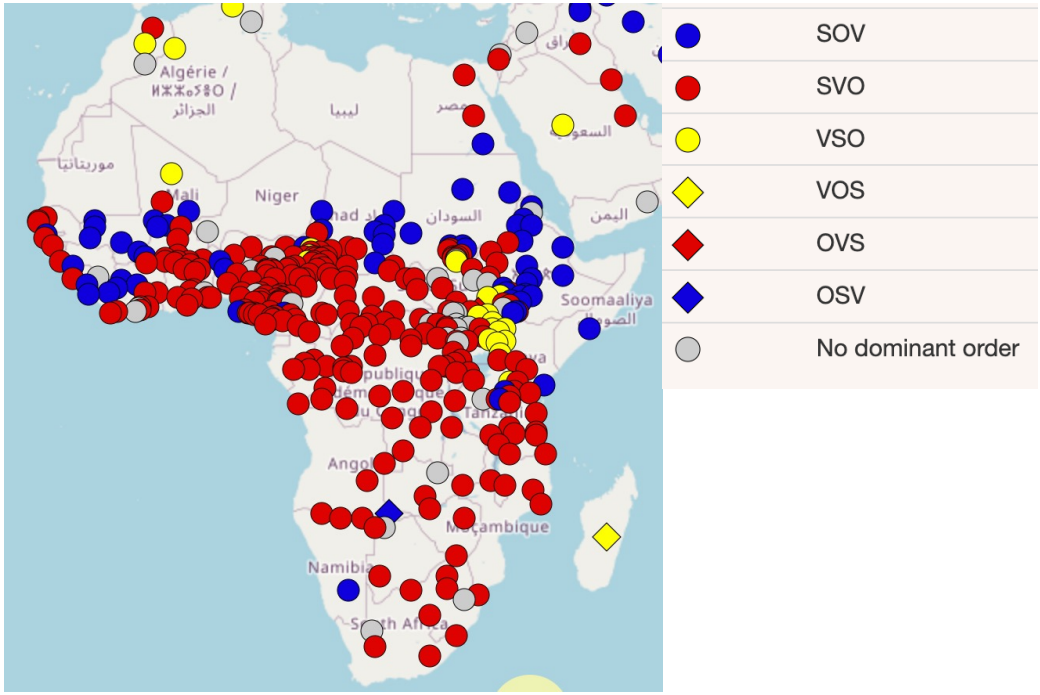
What is it about the multilingual language models that enables these results?

Typology of Languages

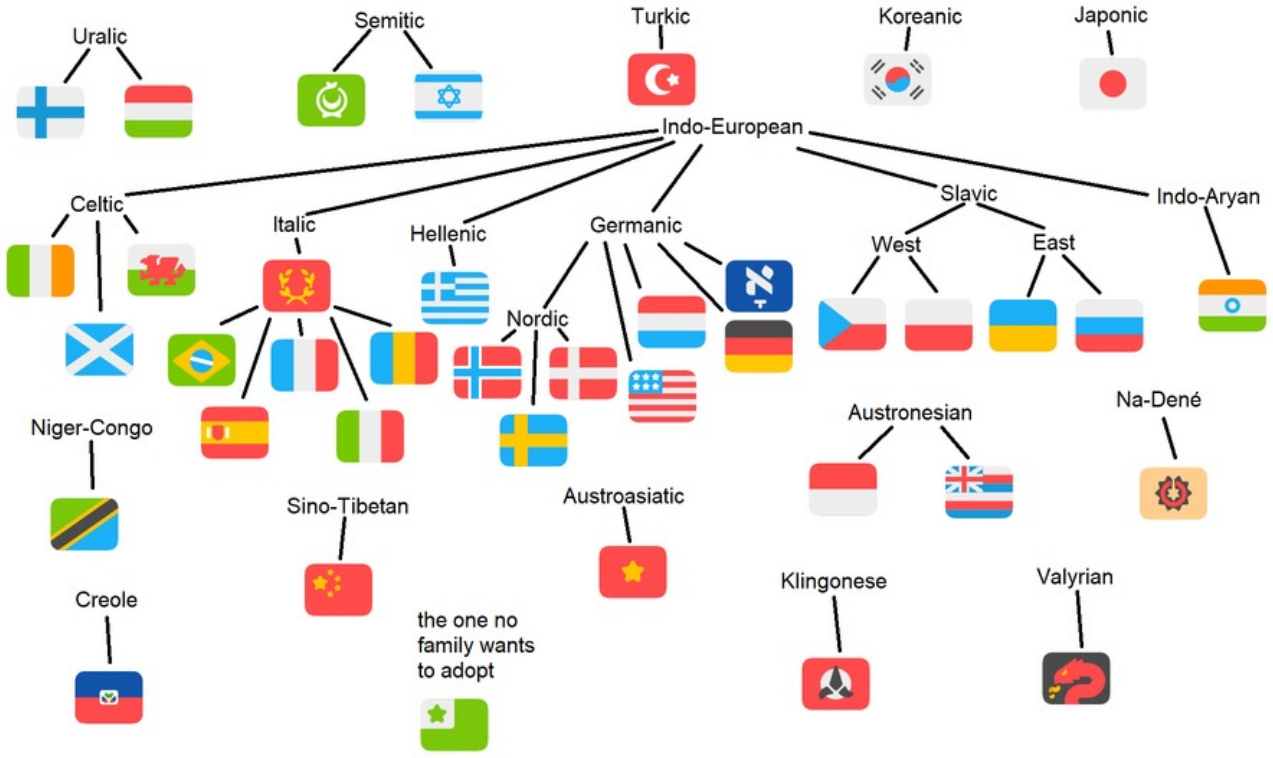
Linguistic Typology: is a subfield of linguistics that studies and classifies languages according to their structural features.

Subfields of linguistic typology:

- **Morphological**: study of the formation and internal structure of words
- **Morphosyntactic:** relationship between syntax and morphology
- **Word Order**



Typology of Languages



- Languages build families and genera based on their relatedness
- The world's languages share universal features at a deep level
- Linguistic universal: Nearly all languages of the world make a distinction between nouns and verbs

Typological Databases

Tag	Description	Example	Tag	Description	Example
JJ	adjective	*small*	UH	interjection	*oops, oh*
JJR	adject., comparative	*smaller*	VB	verb, base form	*fly*
JJS	adject., superlative	*smallest*	VBD	verb, past tense	*flew*
LS	list item marker	*1, one*	VBG	verb, gerund	*flying*
MD	modal	*can, could*	VBN	verb, past participle	*flown*
NN	noun, singular or mass	*dog*	VBP	verb, non-3sg pres	*fly*
NNS	noun, plural	*dogs*	VBZ	verb, 3sg pres	*flies*
NNP	proper noun, sing.	*London*	WDT	wh-determiner	*which, that*

```
# sent_id = dev-0
# text = Hvor kommer julemanden fra?
1   Hvor    hvor    ADV _   ADV 2   advmod  _   _
2   kommer  komme   VERB    _   ACT;PRS;V;IND   0   root    _   _
3   julemanden  julemand    NOUN    _   N;MASC+FEM;DEF;SG   2   nsubj   _   _
4   fra fra ADP _   ADP 1   case    _   SpaceAfter=No
5   ?   ?   PUNCT   _   _   2   punct   _   _
```

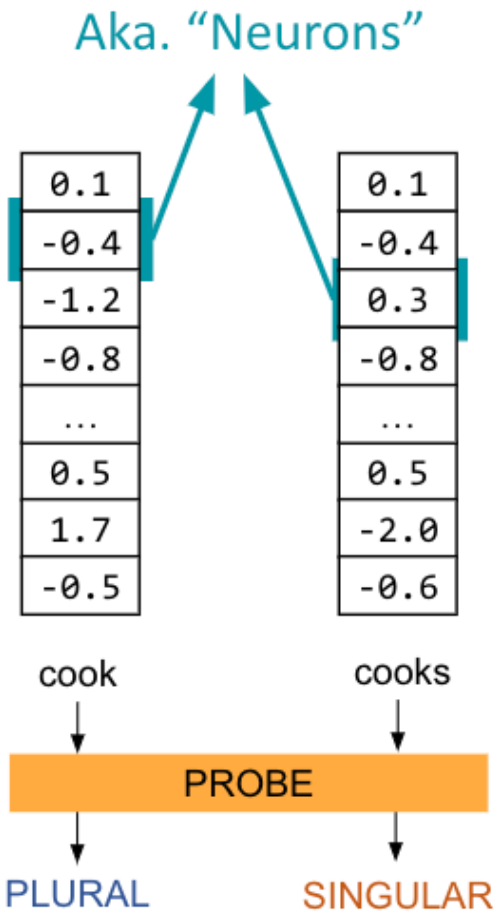
Penn Treebanks: project which has produced approximately 7 million words of part-of-speech tagged text, 3 million words of skeletally parsed text!

Universal Dependencies: a framework for consistent annotation of grammar across languages. It includes dependencies, part-of-speech tags, and the morphosyntactic tagsets.

Probing for Typology

Probing: An analysis field that aims to decipher what the representations know

- Train a ***supervised classifier*** to recover a linguistic property from a neural net
- High performance suggests that a property is encoded in the representation



Extrinsic vs Intrinsic Probing

Extrinsic probes

Layer-level resolution

How good is layer X at encoding Y?

Does reprsentation X encode Y?

Intrinsic probes

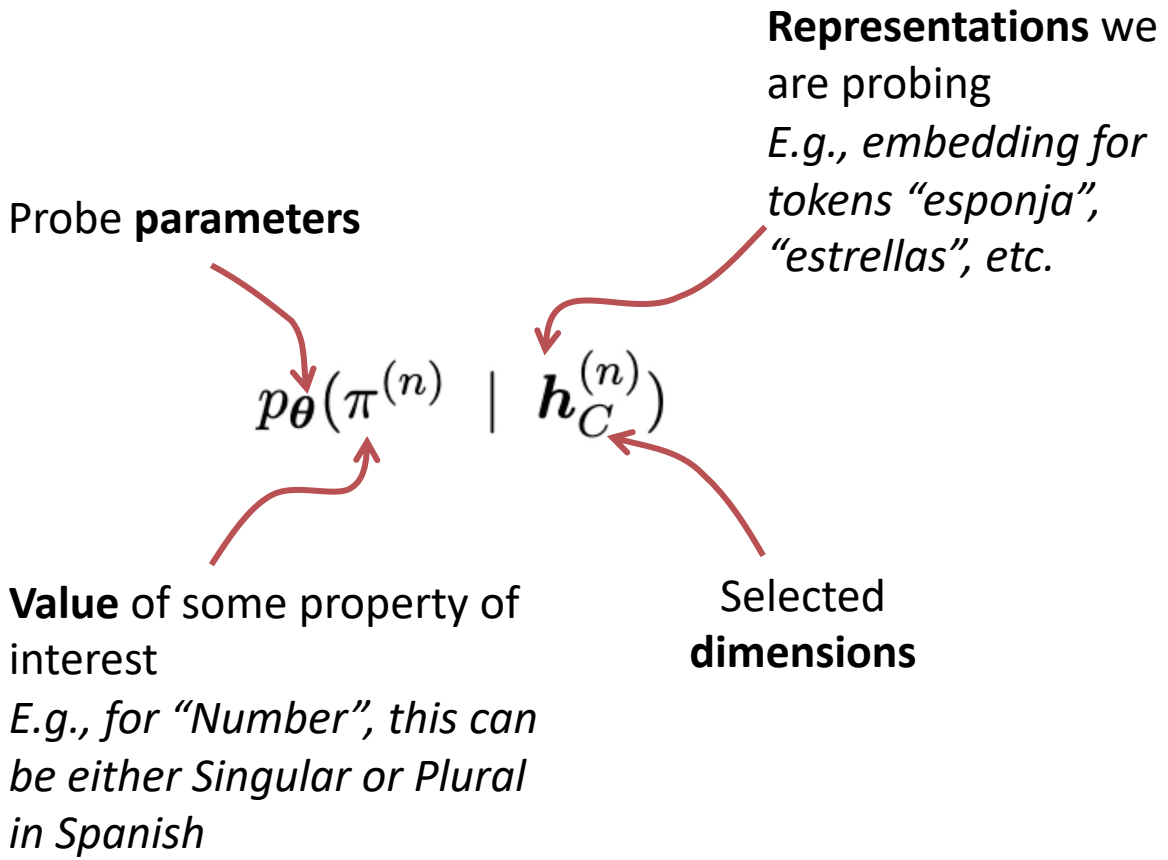
Neuron-level resolution

How dispersed is information of Y in representation X?

Intrinsic Probing

In intrinsic probing, our goal is to find the size k subset of neurons $C \subseteq D$, which are most informative about the property π of interest given a representation vector $\boldsymbol{h}$.

$$C^\star = \underset{\substack{C \subseteq D, \\ |C|=k}}{\operatorname{argmax}} \sum_{n=1}^{N} \log p_{\boldsymbol{\theta}}\left(\pi^{(n)} \mid \boldsymbol{h}_C^{(n)}\right)$$



Application Areas of Intrinsic Probing

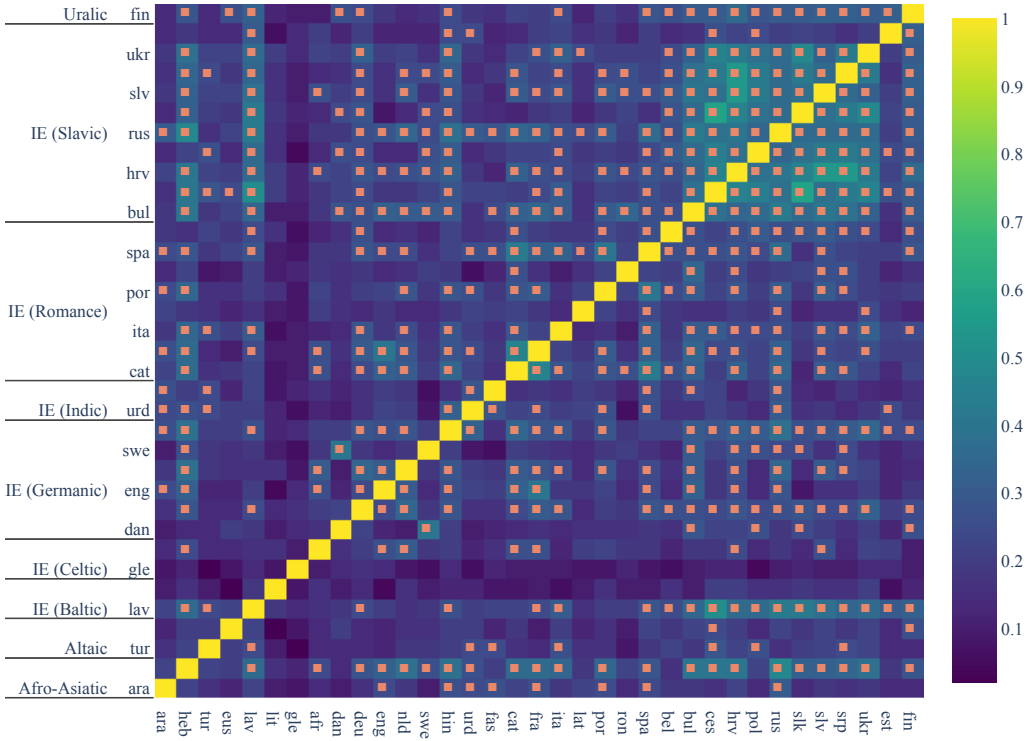
Identifying individual neurons responsible for encoding linguistic features of interest has previously been shown:

- to increase **model transparency** (Bau et al., 2019)
- to mitigate **potential biases** (Vig et al., 2020)
- in controllable **text generation** (Bau et al., 2019)
- to analyse **linguistic capabilities** of language models (Lakretz et al., 2019)

Most Informative Neurons wrt Typology

- *Our hypothesis*: multilingual language models implicitly align morphosyntactic information that fulfils a similar grammatical function across languages (such as gender for nouns and mood for verbs)

- Overlap of sub-words between cognates in related languages plays a role in the process of multilingual generalisation



Overlap in Most Informative Neurons

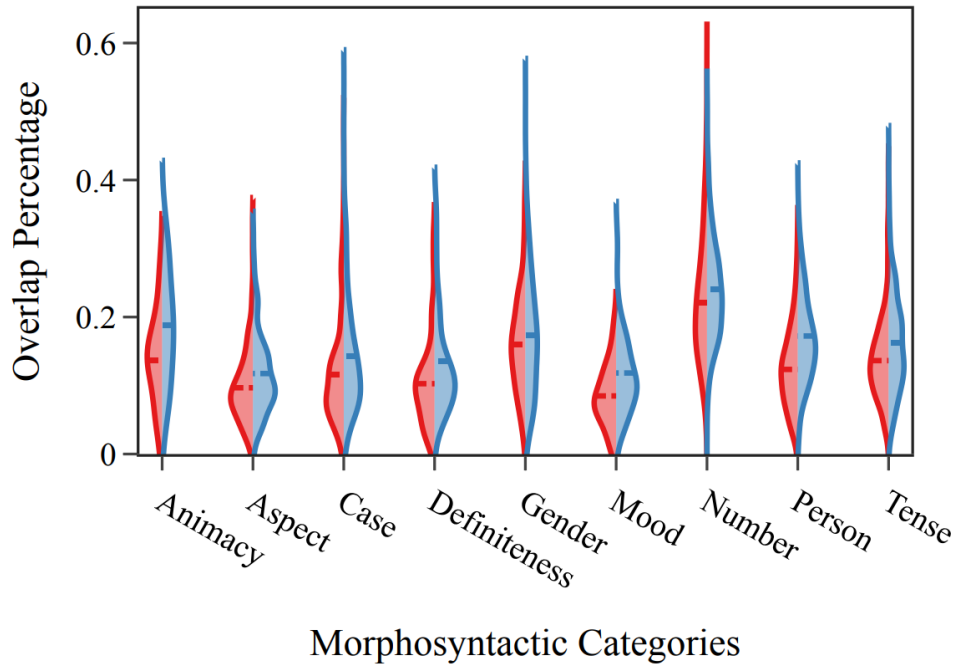


Figure 1: Percentages of neurons most associated with a particular morphosyntactic category that overlap between pairs of languages. Colours in the plot refer to 2 models: m-BERT (red) and XLM-R-base (blue).

≈ 20% of neurons among the top-50 most informative ones overlap on average, but this number may vary dramatically across categories!

Morphosyntactic Categories I

While significant overlap is accentuated in categories such as comparison, polarity, and number, neurons for categories such as mood, aspect, and case are shared by only a handful of language pairs.

This finding may be partially explained by the different number of values each category can take.

	m-BERT	XLM-R-base	XLM-R-large	Total
Definiteness	0.11	0.22	0.13	45
Comparison	0.20	0.90	0.50	10
Possession	0.00	0.00	0.00	1
Aspect	0.03	0.10	0.09	153
Polarity	0.33	0.67	0.33	3
Number	0.40	0.51	0.74	666
Animacy	0.14	0.57	0.32	28
Mood	0.00	0.07	0.05	105
Gender	0.15	0.32	0.19	378
Person	0.08	0.25	0.13	276
POS	0.04	0.27	0.70	861
Case	0.10	0.18	0.17	300
Tense	0.08	0.23	0.12	325
Finiteness	0.09	0.18	0.09	45

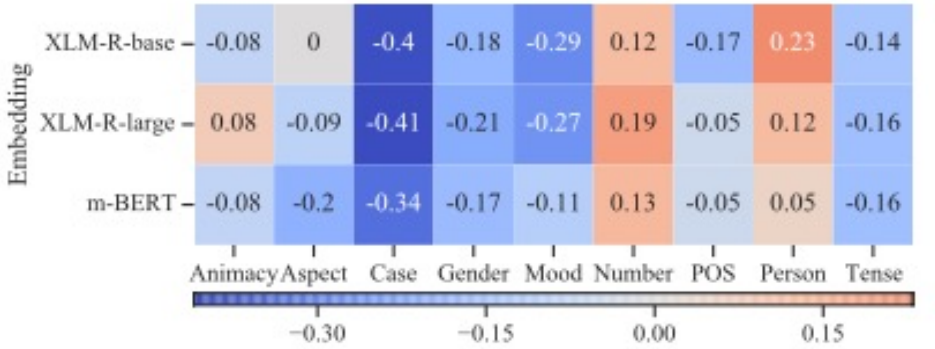
Table 1: Proportion of language pairs with statistically significant overlap in the top-50 neurons for an attribute (after Holm–Bonferroni (Holm, 1979) correction). We compute these ratios for each model. The final column reports the total number of pairwise comparisons.

Morphosyntactic Categories II

Generally negative correlation between neuron overlap and number of morphosyntactic categories — prominent exceptions being number and person.

As the inventory of values of a category grows, cross-lingual alignment becomes harder.

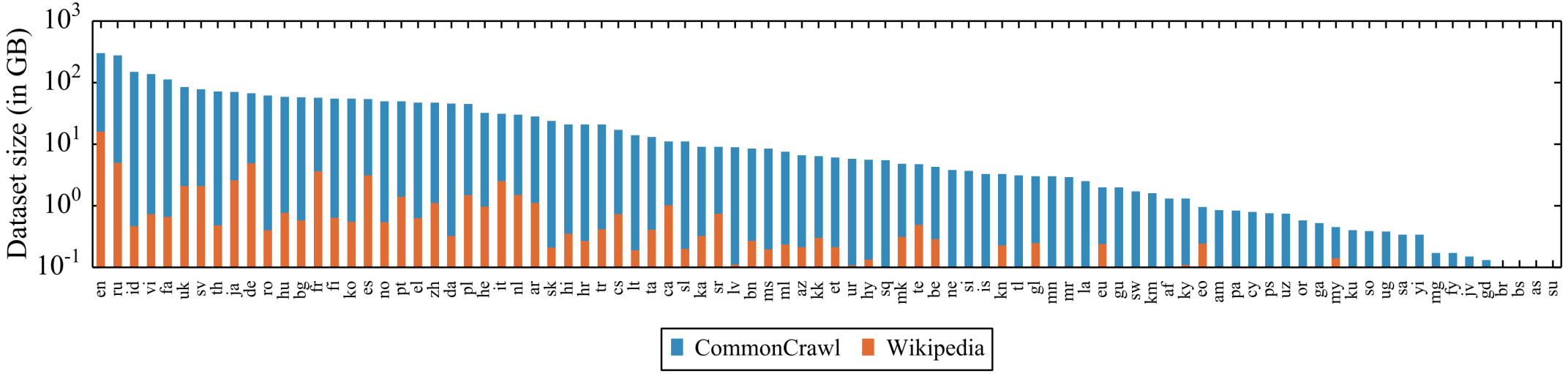
Figure 5: Spearman's correlation, for a given model and morphological category, between the cross-lingual average percentage of overlapping neurons and:



(a) number of values for each morphosyntactic category;

Challenges of Multilinguality

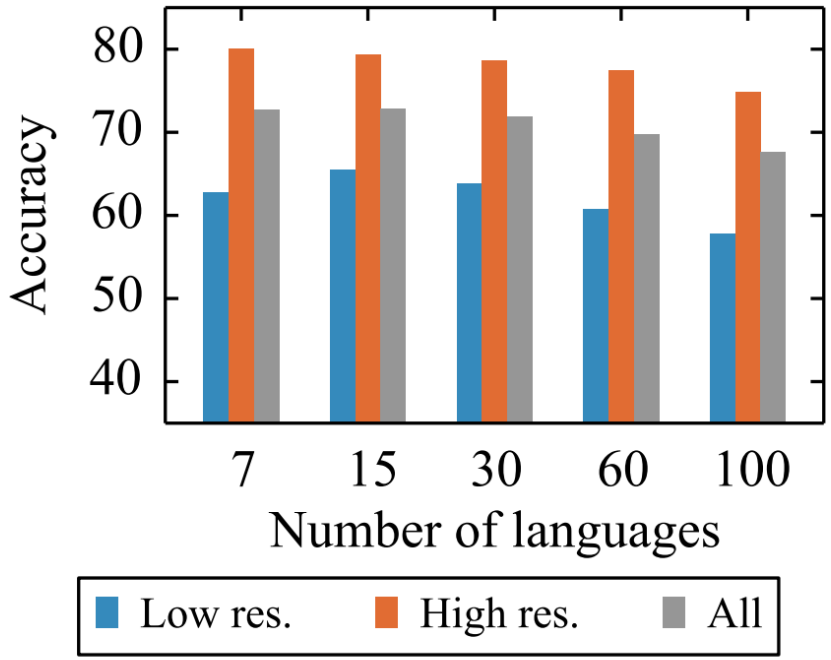
- Data distribution across languages is highly imbalanced!



- Upsampling low resource languages?
- How to evaluate a multilingual language model?

Curse of Multilinguality

- A trade-off between supporting more languages and obtaining better performance on a smaller set of languages.
- Performance drops as more languages are added to the model.
- Recent approaches to mitigate this issue.



Overall Take-Aways I

- Need in NLP to reflect the multilinguality in the world.
- Multilingual language models are able to perform cross-lingual generalisation surprisingly well.
- It remains unclear how pre-trained models actually manage to learn multilingual representations.

Overall Take-Aways II

- Probing offers a way to understand the grammatical knowledge a language model encodes.

- Pre-trained multilingual language models employ the same subset of neurons to encode the same morphosyntactic information.

- NLP is fun:
  - Introduction to NLP: 5100-B1-1E22
  - Fair and Transparent Machine Learning Methods: NDAK22005U

Resources

Papers:

- How Multilingual is Multilingual BERT?
- Investigating Gender Bias in Language Models Using Causal Mediation Analysis
- Modeling Language Variation and Universals: A Survey on Typological Linguistics for Natural Language Processing
- Probing Pretrained Language Models for Lexical Semantics
- Same Neurons, Different Languages: Probing Morphosyntax in Multilingual Pre-trained Models
- The emergence of number and syntax units in LSTM language models
- Understanding the role of individual units in a deep neural network
- Unsupervised Cross-lingual Representation Learning at Scale

CMU CS Class:

- Multilingual Natural Language Processing