

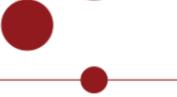
Finding Meaning in Data across Languages

Sprogteknologisk Konference
30 November 2022

Daniel Hershcovich

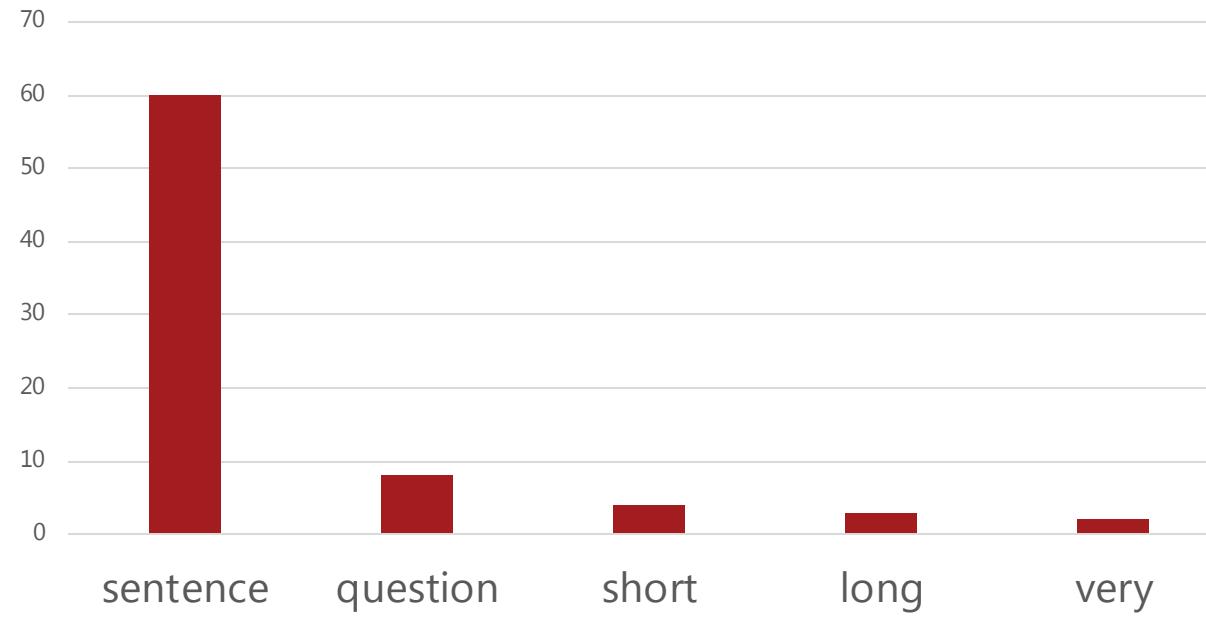
Department of Computer Science
(DIKU)

UNIVERSITY OF COPENHAGEN



Language models

What is the next word in this



Language modeling: given text, estimate the probability distribution of the next word
(usually based on huge text corpora)

Language models

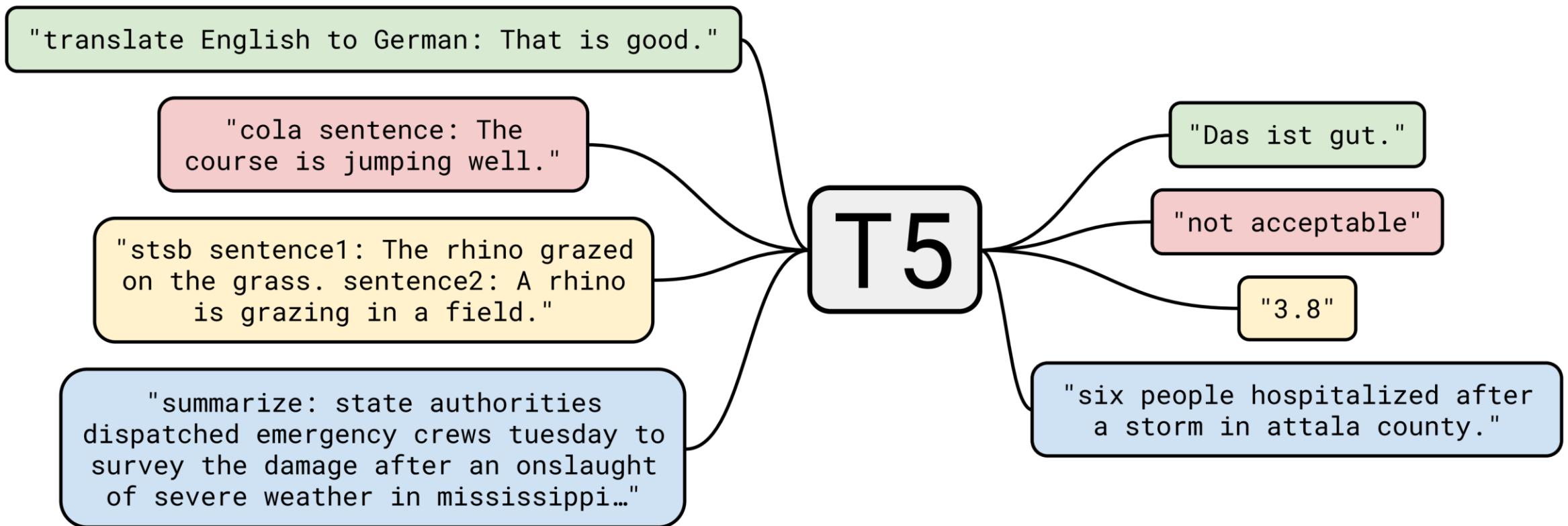
Paradigm shift in NLP since ~2020:
"Any" task can be cast as language modeling



Language Models are Few-Shot Learners (Brown et al., NeurIPS 2020)



Language models



Language models

Instruction finetuning

Please answer the following question.
What is the boiling point of Nitrogen?

Chain-of-thought finetuning

Answer the following question by reasoning step-by-step.
The cafeteria had 23 apples. If they used 20 for lunch and bought 6 more, how many apples do they have?

Multi-task instruction finetuning (1.8K tasks)

Inference: generalization to unseen tasks

Q: Can Geoffrey Hinton have a conversation with George Washington?
Give the rationale before answering.

Language model

-320.4F

The cafeteria had 23 apples originally. They used 20 to make lunch. So they had $23 - 20 = 3$. They bought 6 more apples, so they have $3 + 6 = 9$.

Text-to-image

*Medieval painting of
a monk eating a poster
session on language
technology*

DALL·E 2



Text-to-code with language models

Exemplars

Q: Who edited a film that M1 and M2 produced
A: <Exemplar Answer> ...

Subproblems

Q: What was directed by M3
A: <Predicted Answer> ...

Input

Q: What was produced by an art director that
M1 and M2 employed and was directed by M3

LM

Output

A: SELECT DISTINCT WHERE {
?x0 produced_by ?x1 . ?x1 a art_director .
M1 employed ?x1 . M2 employed ?x1 .
?x0 directed_by M3 }

Knowledge base

answer

Language models

Masked/bidirectional/LMs:

- ELMo ([Peters et al., 2018](#))
- BERT ([Devlin et al., 2019](#))
- RoBERTa ([Liu et al., 2019](#))
- ...

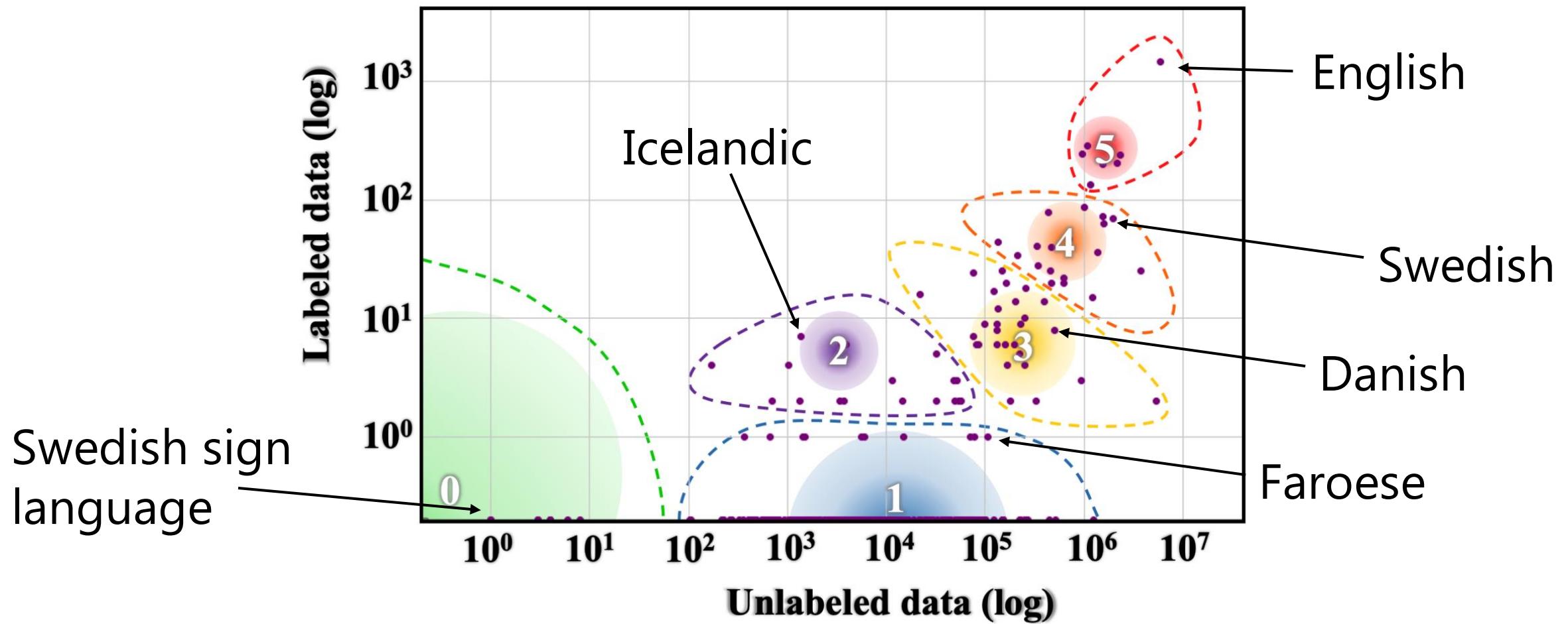


Causal/generative/autoregressive LMs:

- GPT-2 ([Radford et al., 2018](#))
- GPT-3 ([Brown et al., 2020](#))
- T5 ([Raffel et al., 2019](#))
- T0 ([Sanh et al., 2021](#))
- BART ([Lewis et al., 2020](#))
- FLAN ([Wei et al., 2021](#))
- ...

All trained (almost) only on
English text

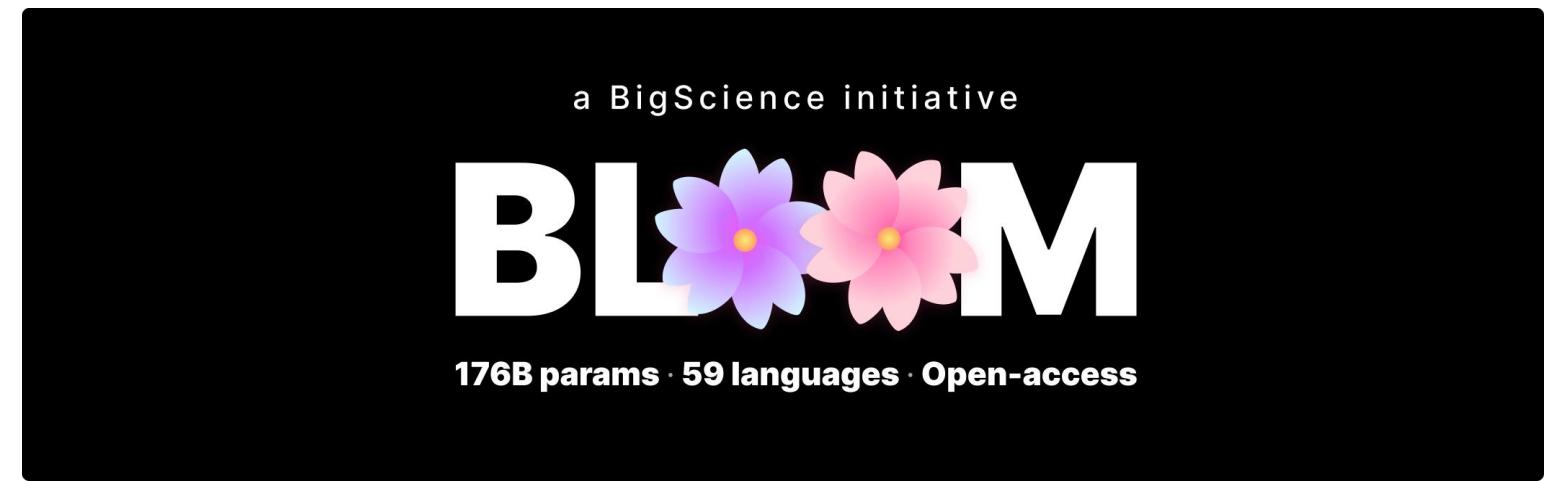
Resource disparity for languages



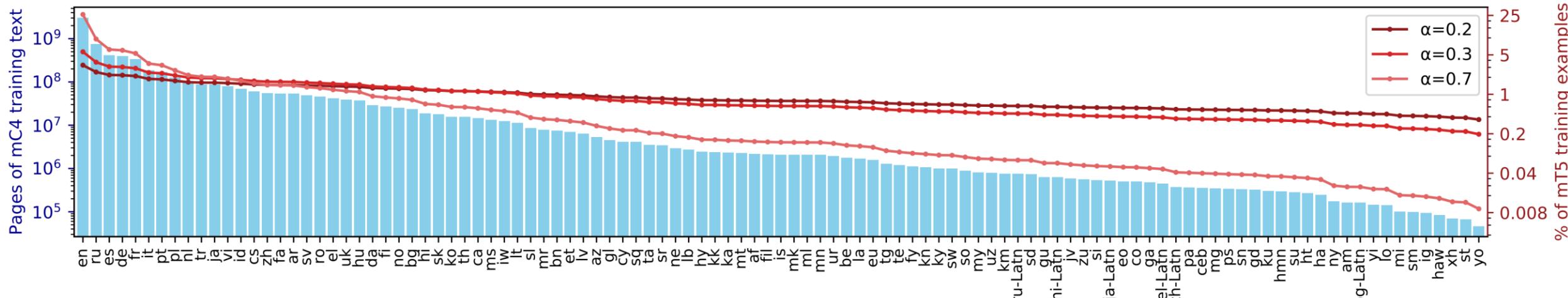
The State and Fate of Linguistic Diversity and Inclusion in the NLP World
(Joshi et al., ACL 2020)

Multilingual language models

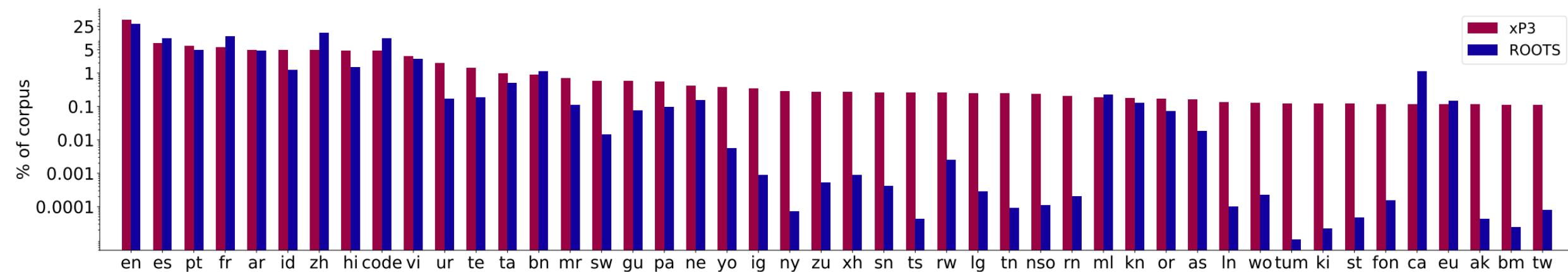
- mBERT ([Devlin et al., 2019](#))
- XLM, XLM-R ([Conneau et al., 2020](#))
- mBART ([Liu et al., 2020](#))
- mT5 ([Xue et al., 2021](#))
- XGLM ([Lin et al., 2021](#))
- BLOOM ([Le Scao et al., 2022](#))
- ...



Language distribution in multilingual language models



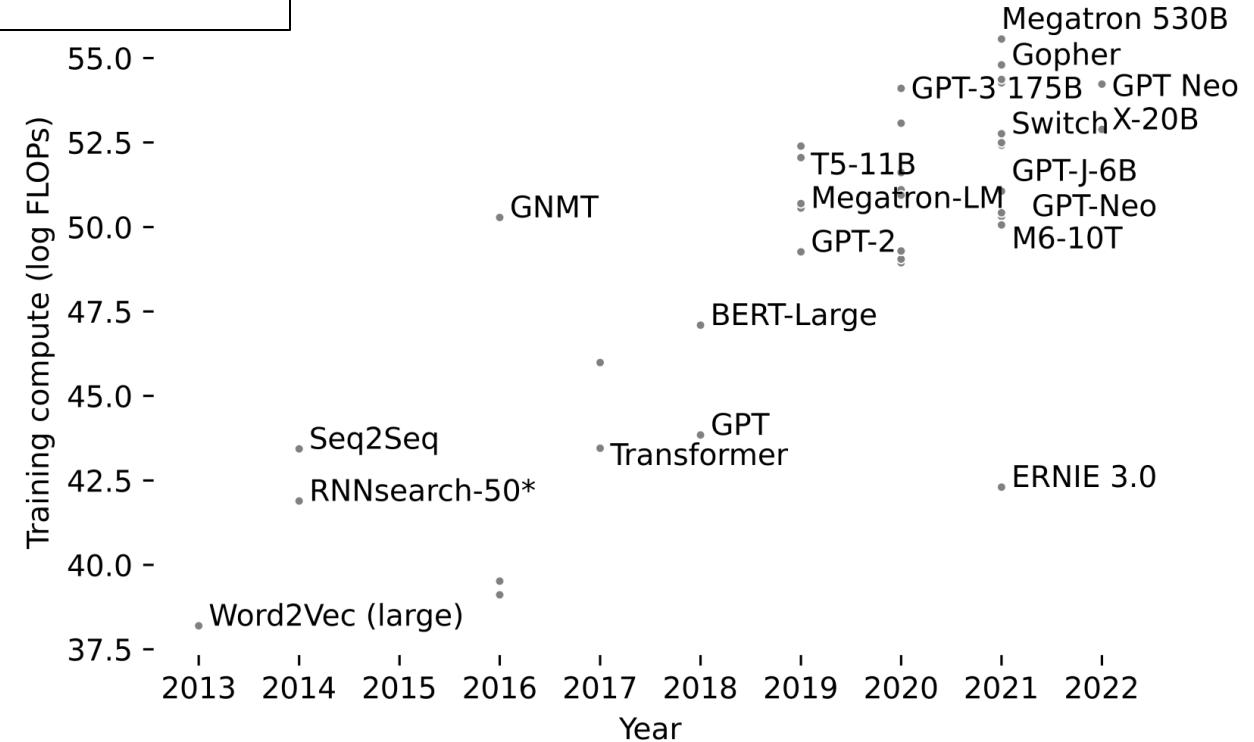
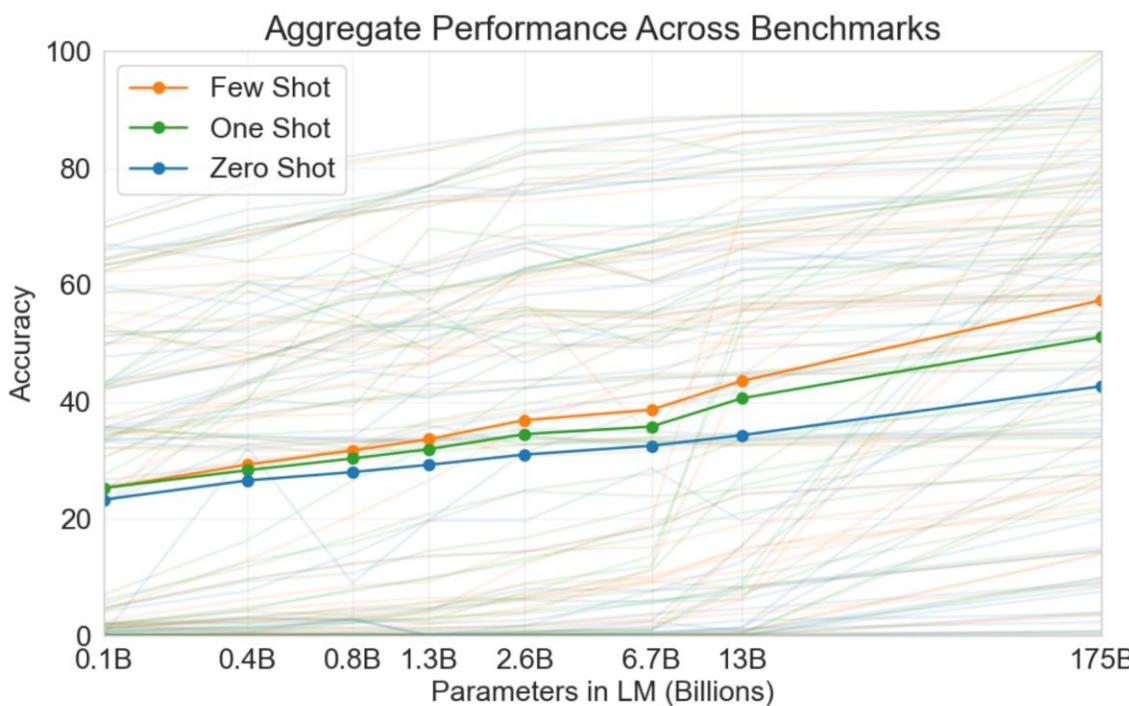
mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer (Xue et al., NAACL 2021)



[BLOOM: A 176B-Parameter Open-Access Multilingual Language Model](#) (Le Scao et al., 2022)

Diminishing returns?

Newer and larger models perform better but require more and more resources and energy



Language Models are Few-Shot Learners (Brown et al., NeurIPS 2020)

Towards Climate Awareness in NLP Research (Hershcovich et al., EMNLP 2022)

Can we do more with less data?

Besides large text corpora, explicit **meaning representation** can help...



Performance

Inductive bias
Access to structured data
Reasoning ability



Understanding

Interpretability
Theoretical analysis
Fine-grained control



Generalization

Languages
Domains
Tasks

Finding meaning by decomposition

[Meaning], [Representation] and [Parsing]

1. What we mean, 2. How to represent (something), 3. How to parse (something)

[Meaning Representation] and [Parsing]

1. How to represent what we mean, 2. How to parse (something)

[Meaning [Representation and Parsing]]

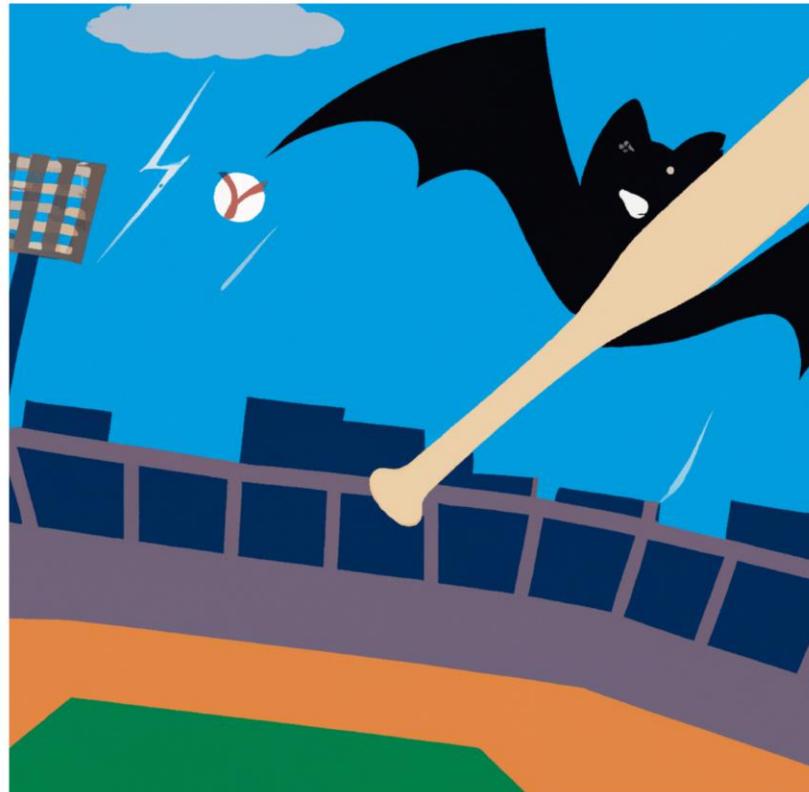
1. How to represent what we mean, 2. How to parse what we mean

[Meaning Representation] and [Parsing (to Meaning Representation)]

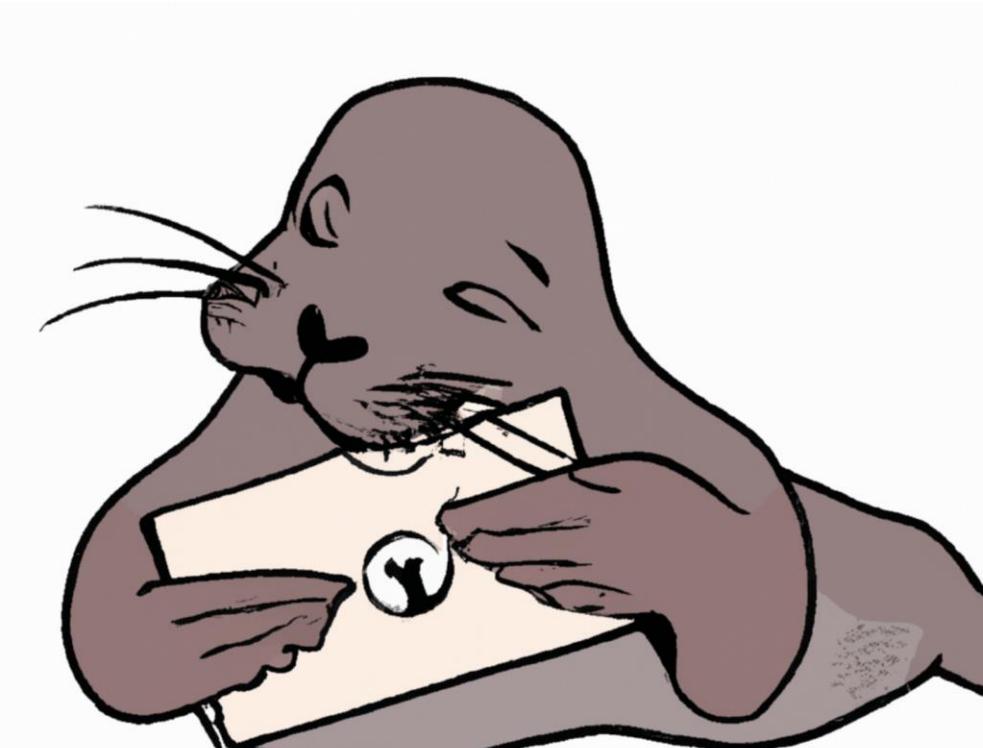
1. How to represent what we mean, 2. How to parse (1)

Meaning in text-to-image

A bat is flying over a baseball stadium



a seal is opening a letter



Meaning in text-to-code

What (was produced by ((a art director) that (M1 and M2 employed)) and (was directed by M3))

What (was produced by ((a art director) that (M1 and M2 employed)))

What (was directed by M3)

What (was produced by (a art director))

What (was produced by ((a art director) that (M1 employed)))

Meaning representation for analysis of language models

Context: A piece of paper was later found on which he had written his last statements in **two** languages, Latin and German. Only **one** statement was in Latin and the rest in German.

Question: In what language were **most** statements written?

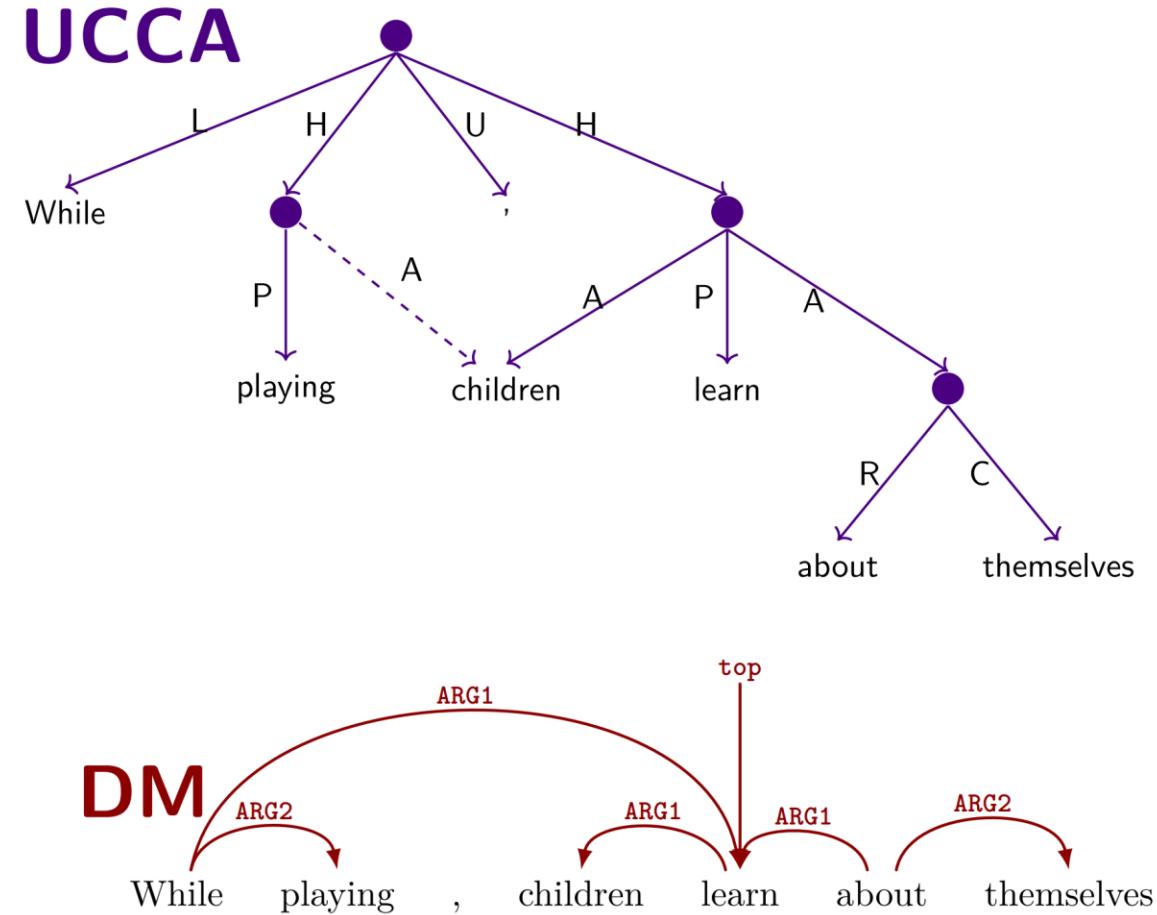
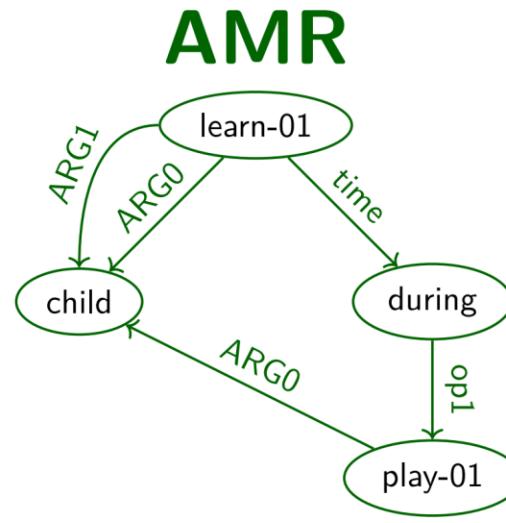
Answer: German

Predicted answer

(RoBERTa): Latin and German

Generalized Quantifiers	Logical Denotation	RoBERTa avg. acc.
some(A)(B) = 1	$A \cap B \neq \emptyset$	83.7
all(A)(B) = 1	$A \subseteq B$	85.3
more than k the(A)(B) = 1	$ A \cap B > k$	68.2
less than k the(A)(B) = 1	$ A \cap B < k$	91.7
k (A)(B) = 1	$ A \cap B = k$	87.8
between p and k the(A)(B) = 1	$p < A \cap B < k$	70
the p/k (A)(B) = 1	$ A \cap B = p \cdot (A /k)$	77.8
the k% (A)(B) = 1	$ A \cap B = k \cdot (A /100)$	72.2
most (A)(B) = 1	$ A \cap B > A \setminus B $	80.9
few (A)(B) = 1	$ A \cap B < A \setminus B $	78.3
each other (A)(B) = 1	$\forall a \in (A \cap B) \exists b \in (A \cap B) (a \neq b)$	84.1

Meaning representation frameworks



Universal Conceptual Cognitive Annotation (UCCA)

Design principles

- Cross-linguistic portability and stability
- Accessibility to non-expert annotators
- Modularity of semantic components

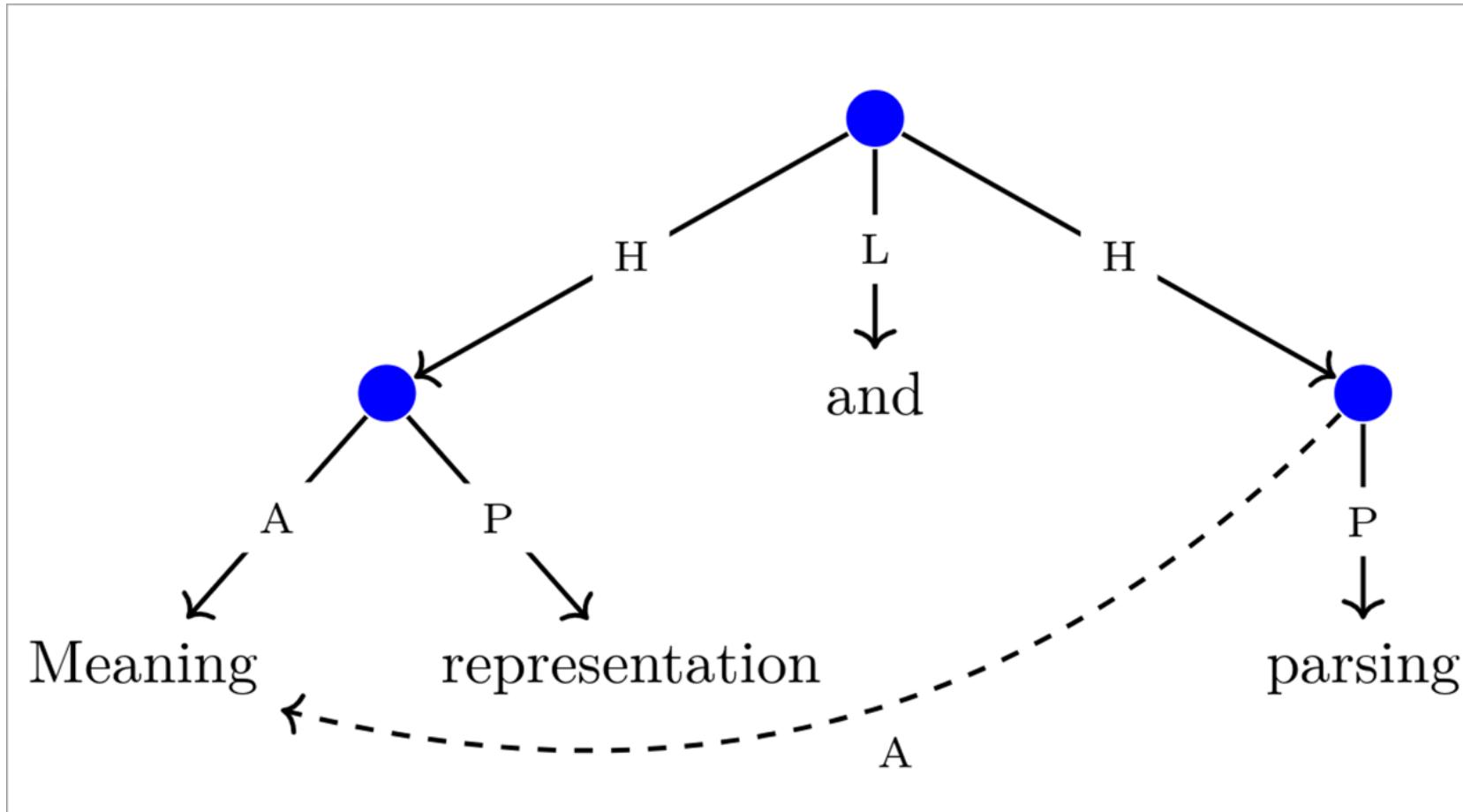
Corpora

- English, German, French, Russian, Hebrew & Turkish

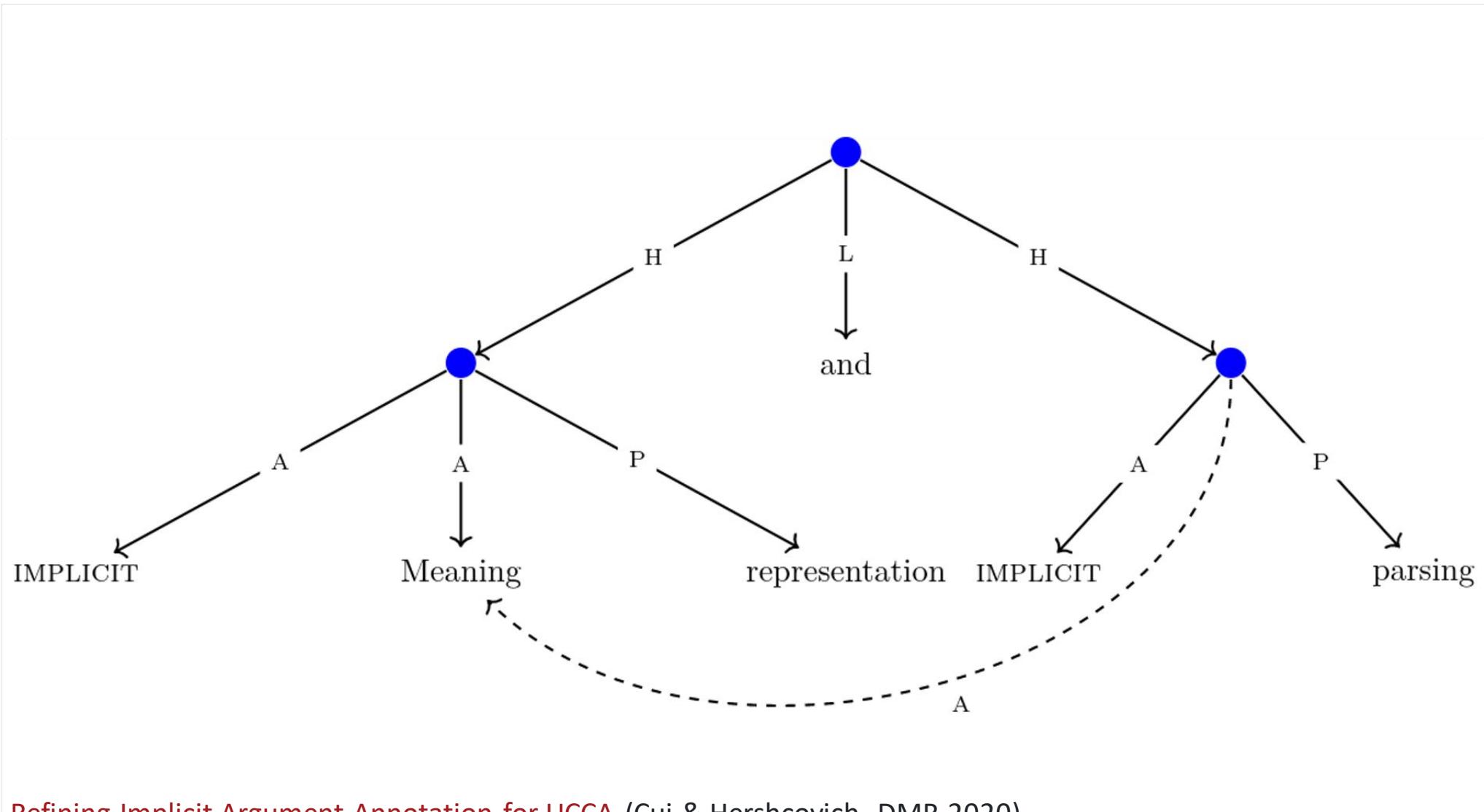
Applications

- Text simplification
- Machine translation
- Relation extraction
- Textual process description

UCCA example



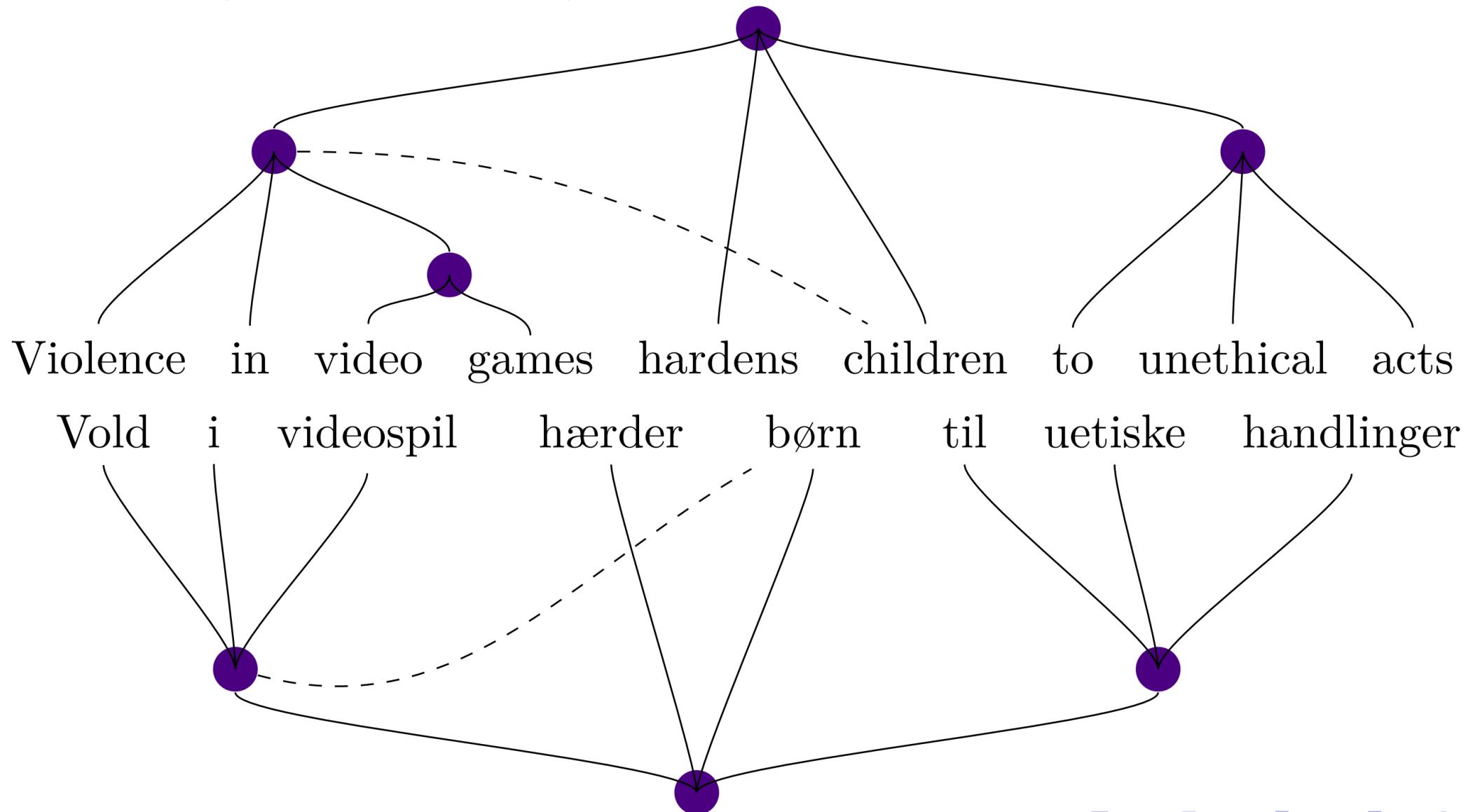
Implicit relations in UCCA



[Refining Implicit Argument Annotation for UCCA](#) (Cui & Hershcovitch, DMR 2020)

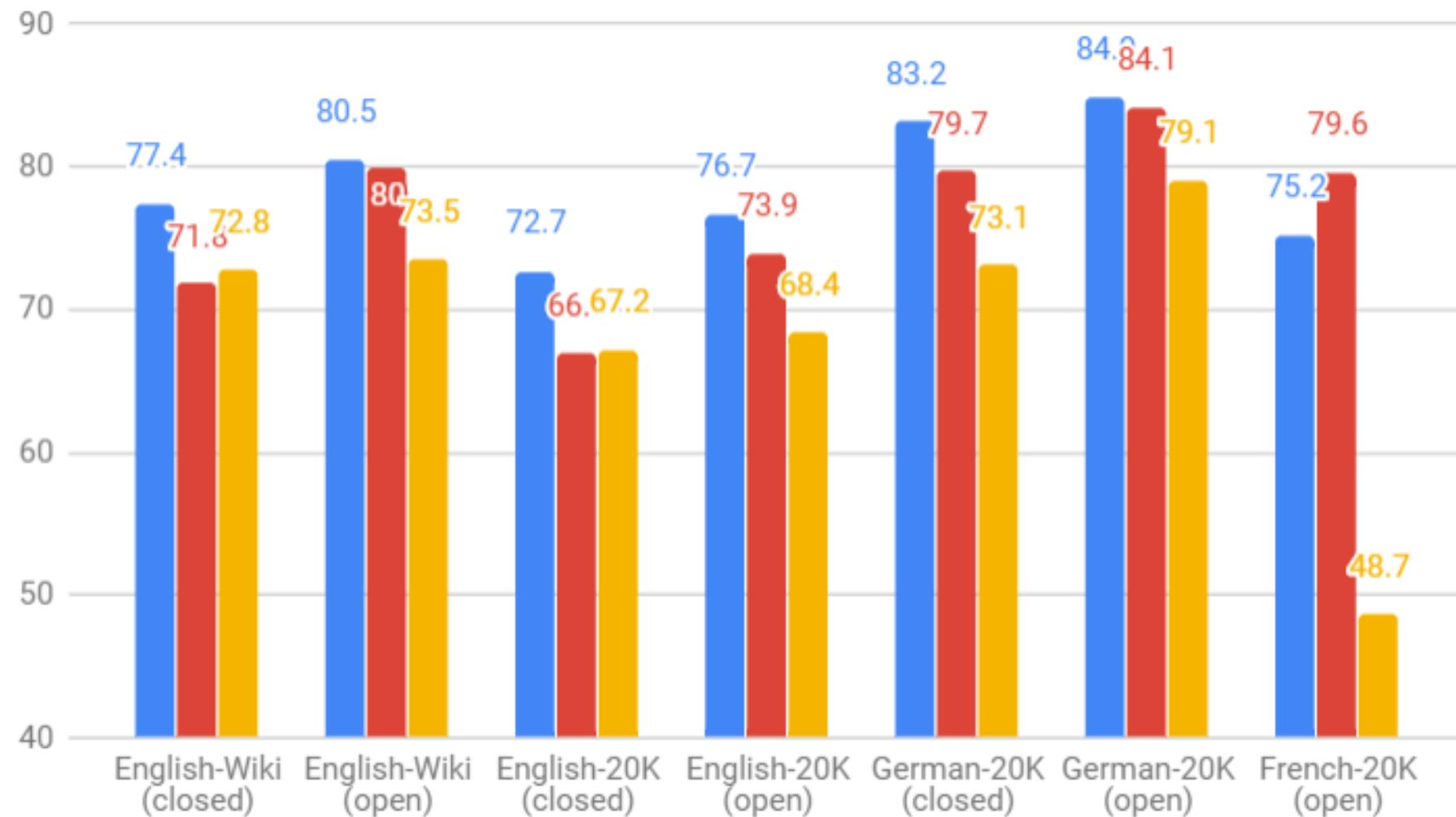
[Great Service! Fine-grained Parsing of Implicit Arguments](#) (Cui & Hershcovitch, IWPT 2021)

Cross-linguistic stability in UCCA



UCCA parsing

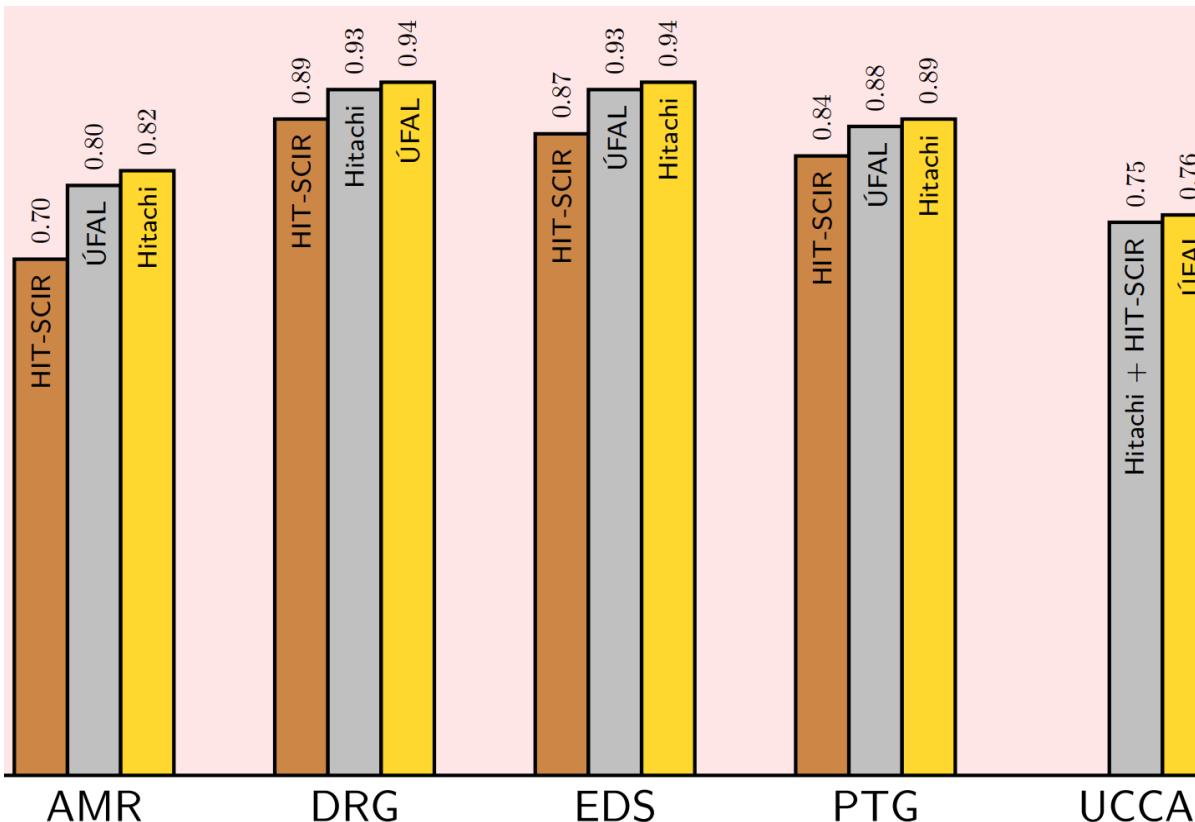
Successful cross-lingual transfer



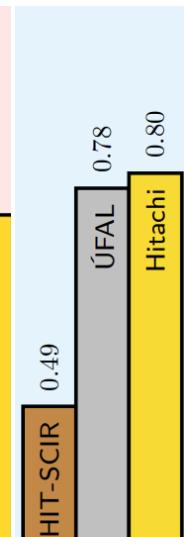
Meaning representation parsing

Successful monolingual parsing
in different languages

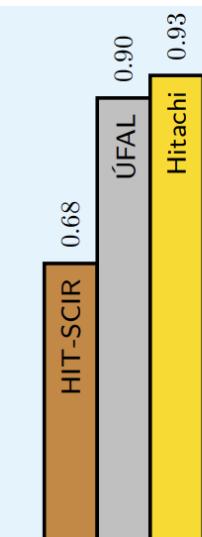
English



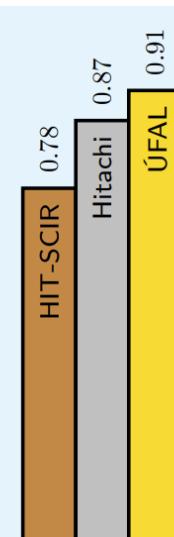
Chinese



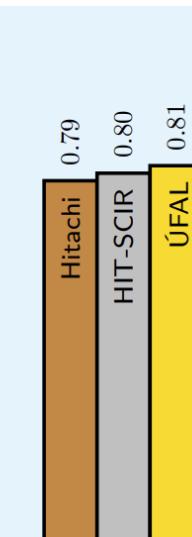
German



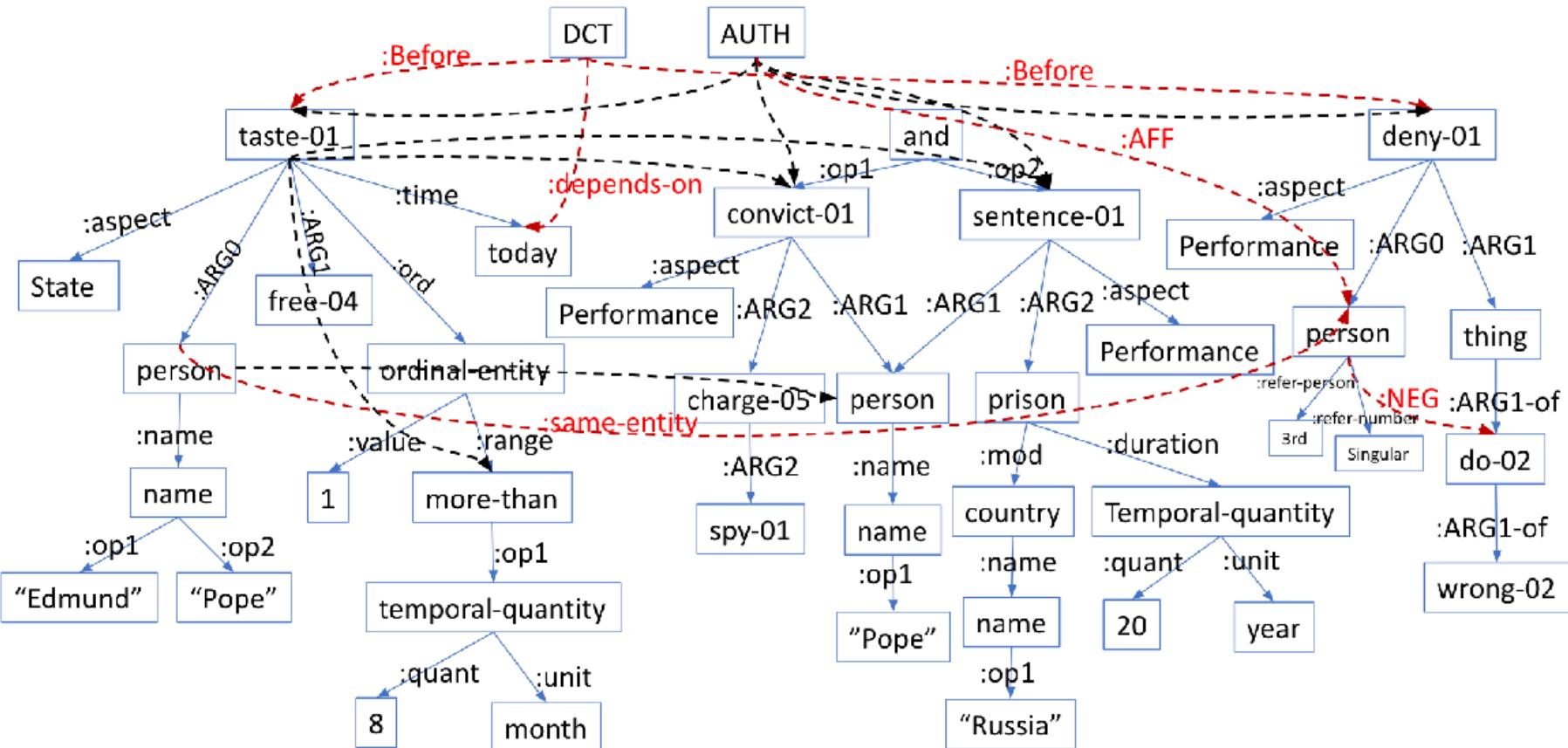
Czech



German



Uniform Meaning Representation



"Edmund Pope tasted freedom today for the first time in eight months."

"Pope was convicted on spying charges and sentenced to 20 years in a Russian prison."

"He denied any wrong-doing."

Compositional generalization

“

**"THE ABILITY TO SYSTEMATICALLY GENERALIZE TO
COMPOSED TEST EXAMPLES OF A CERTAIN
DISTRIBUTION AFTER BEING EXPOSED TO THE
NECESSARY COMPONENTS DURING
TRAINING ON A DIFFERENT DISTRIBUTION"**

Train set

Who directed inception?

Did Greta Gerwig produce Goldfinger?

Test set

Did Greta Gerwig direct Goldfinger?

Who produced inception?

Multilingual Compositional Wikidata Questions (MCWQ)

Lang. Question

En	Did Lohengrin's male actor marry Margarete Joswig
He	האם ה שחקן ה גברי של לוהנגרין עם מרגרט יוסוויג
Kn	ಯೋಹೆಂಗ್ರಿನ್ ಅವರ ಪುರುಷ ನಟ ವಿವಾಹವಾದರು ಮಾರ್ಗರೇಟ್ ಜೋಸ್ವಿಗ್
Zh	Lohengrin 的 男 演员 嫁给 了 Margarete Joswig 吗

SPARQL Query:

```
ASK WHERE { ?x0 wdt:P453 wd:Q50807639 . ?x0  
wdt:P21 wd:Q6581097 . ?x0 wdt:P26 wd:Q1560129 .  
FILTER ( ?x0 != wd:Q1560129 )}
```

answer



MCWQ



Multilingual compositional generalization benchmark



mT5 achieves similar within-language generalization across languages



Zero-shot cross-lingual generalization fails

Limitations of compositional generalization benchmarks

Synthetic & unnatural data

Mostly automatic translation

No cultural adaptation

Social factors

NLP is for people (not just languages)



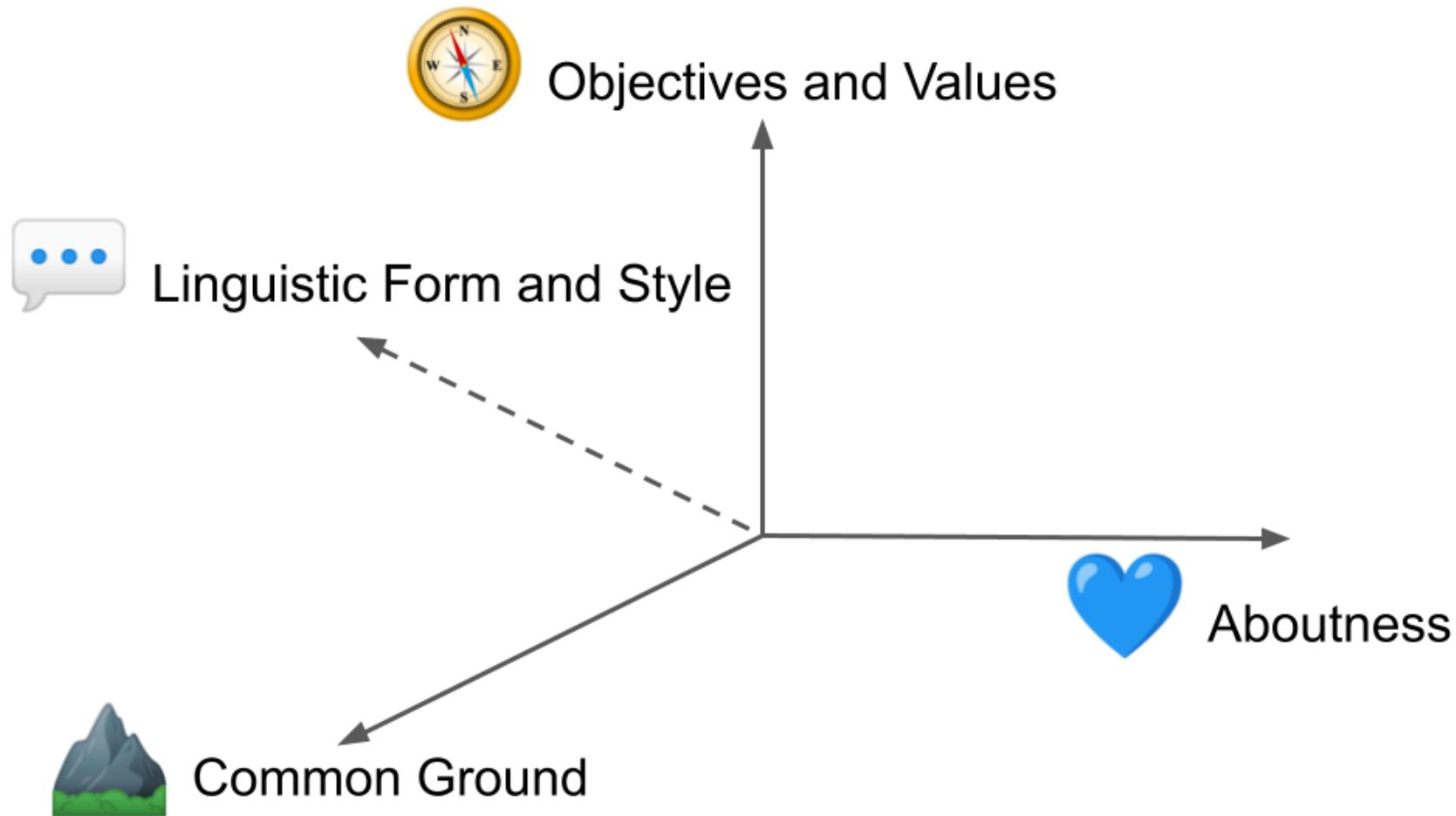
[The Importance of Modeling Social Factors of Language: Theory and Practice](#)
(Hovy & Yang, NAACL 2021)

Social bias in language models

Models	Demographics Alignment															
bert-base-cased																
bert-base-uncased																
bert-base-multilingual-cased																
bert-large-cased																
bert-large-uncased																
distilbert-base-uncased																
albert-base-v2																
albert-large-v2																
albert-xxlarge-v2																
roberta-base																
roberta-large																
google/electra-large-generator																
google/electra-small-generator																
gpt2																
gpt2-medium																
gpt2-large																
gpt2-xl																
Group																
Mean Rank	3.1	3.4	4.0	6.1	6.1	8.1	8.1	9.2	9.8	9.9	10.3	10.3	10.8	11.1	12.0	13.8

Sociolectal Analysis of Pretrained Language Models
(Zhang et al., EMNLP 2021)

Cultural awareness in NLP



Challenges and Strategies in Cross-Cultural NLP
(Hershcovich et al., ACL 2022)

Form 

*How we express
ourselves in
language*

Morphosyntax

Word choice

Style

Levels of granularity

Linguistic and cultural variation within groups



Idiolect	Sociolect, dialect	Standardised language	Language, language family
Individual, personality	Social group or region, sub-culture	Country, national culture	International cultures

Common ground

Shared
knowledge based
on which people
reason and
communicate

Conceptualisation

Commonsense

Commonsense

Some knowledge is "universal", other culture-specific

Color of wedding dress

In traditional [X] weddings, the color of wedding dress is usually [MASK].

EN

पारंपरिक [X] शादियों में दुल्हन की पोशाक का रंग आमतौर पर [MASK] होता है।

HI

...

Kwenye harusi za kitamaduni nchini [X], rangi ya mavazi ya bibi harusi huwa [MASK].

SW

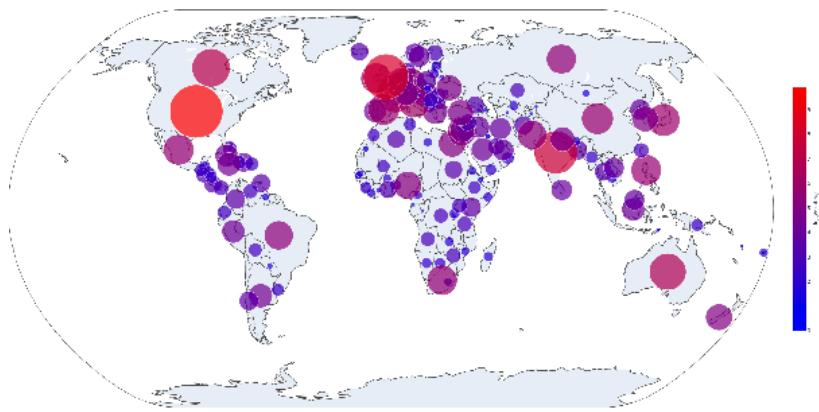
[X] (Country name)	[MASK]
American	
Chinese	
Indian	
Iranian	
Kenyan	

[X] (Country name)	[MASK]
अमेरिका	
चीनी	
भारतीय	
फारसी	
केन्या	

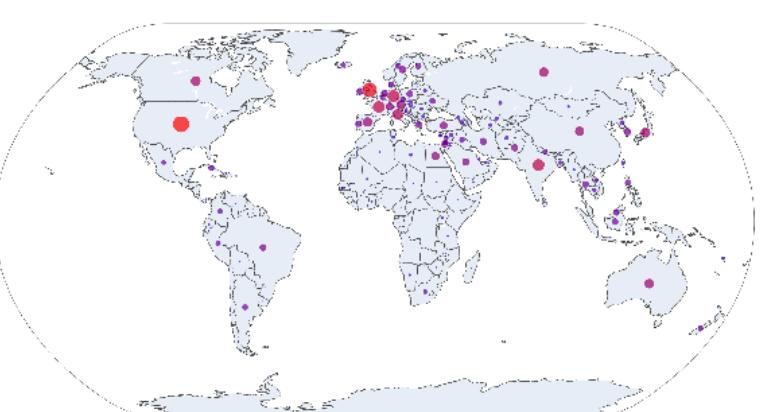
Aboutness ❤️

What content do people *care about*?

Natural Questions

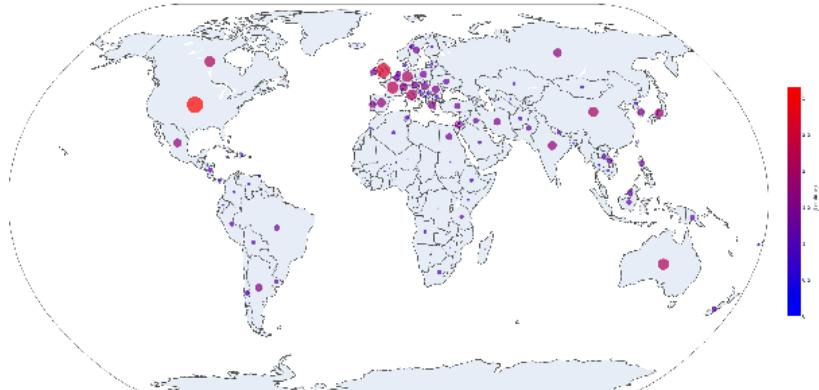


MLQA

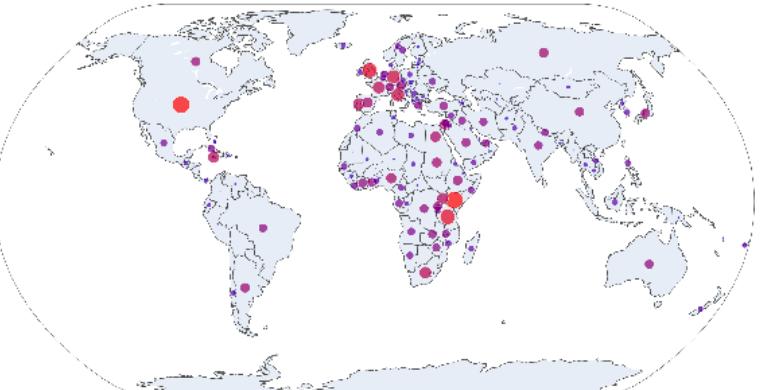


Entities

TyDi-QA (English)



TyDi-QA (Swahili)



Experiences



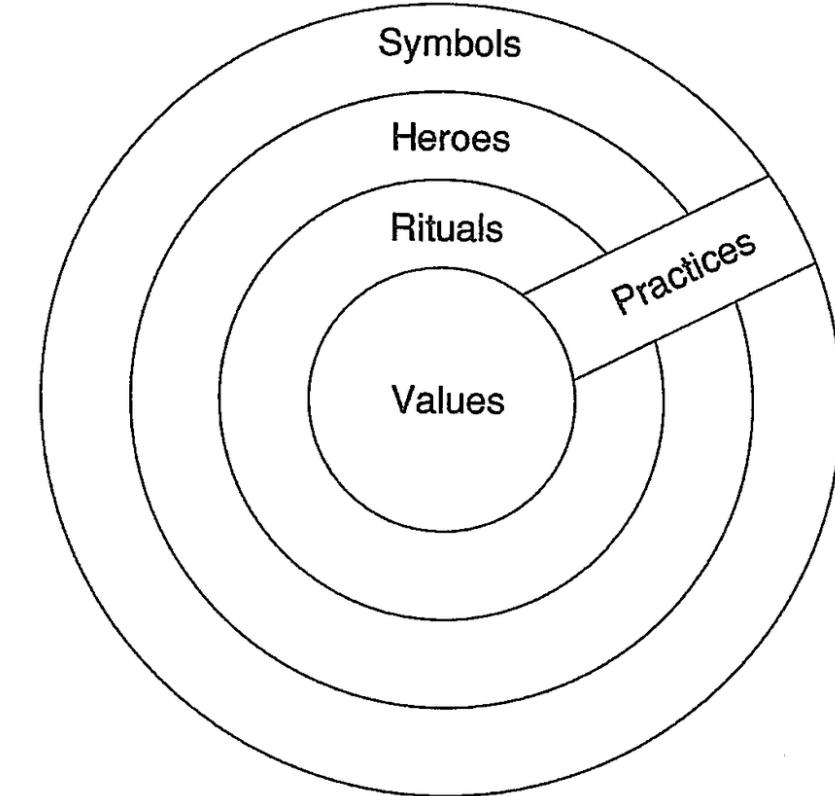
Aspects

Values



Power Distance
Individualism
Masculinity
Uncertainty Avoidance
Long-term Orientation
Indulgence

	Power Distance	Individualism	Masculinity	Uncertainty Avoidance	Long-term Orientation	Indulgence
Turkey	13.600711	18.690817	12.002849	-104.655977	18.402661	-29.212504
Philippines	69.966500	32.454340	-36.896868	68.080674	-29.341779	127.777309
Romania	44.302007	28.049334	1.360547	-44.124610	11.181644	-98.111277
Vietnam	19.073573	36.610564	11.822331	53.483910	5.504491	-167.303567
Malaysia	35.838607	0.000000	0.000000	35.835262	82.649935	45.570108
Korea South	86.411917	-14.096250	9.924329	43.353994	5.085976	-38.421668
Greece	104.289865	-8.447076	-27.989583	58.921055	7.643961	-95.508714
Iran	45.482057	24.832506	-33.998558	-23.384572	-60.234540	-74.847725
Germany	-57.777116	23.726717	35.012510	96.525180	60.957147	-24.038782
Indonesia	39.311610	0.000000	-24.932221	40.816592	24.227209	-50.315727
Pakistan	64.237824	-0.905699	44.611927	154.195160	19.852991	-48.476206
Serbia	-61.397906	-56.702120	-81.248254	-75.697432	-7.394642	-38.726297
Bangladesh	53.278621	70.191660	-31.669899	36.499059	25.463037	-40.400576



Cultures and Organizations: Software of the Mind
(Hofstede, 1991)



World Values Survey

Value bias in language models



Die allermeisten von uns kennen den Zustand völliger Erschöpfung auf der Flucht, verbunden mit Angst um das eigene Leben oder das Leben der Kinder oder der Partner, zum Glück nicht. Menschen, die sich zum Beispiel aus Eritrea, aus Syrien oder dem Nordirak auf den Weg machen, müssen oft Situationen überwinden oder Ängste aushalten, die uns wahrscheinlich schlichtweg zusammenbrechen ließen. Deshalb müssen wir beim Umgang mit Menschen, die jetzt zu uns kommen, einige klare Grundsätze gelten lassen. Diese Grundsätze entstammen nicht mehr und nicht weniger als unserem Grundgesetz, unserer Verfassung.

Values are altered
to reflect US culture



(translation)



- “1. I am in favor of **limiting** immigration.
2. I am in favor of **limiting** immigration for humanitarian reasons.
3. I am in favor of **limiting** immigration for economic reasons.”

The Ghost in the Machine has an American accent: value conflict in GPT-3 (Johnson et al., 2022)

Strategies

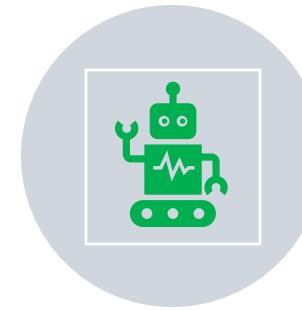


DATA

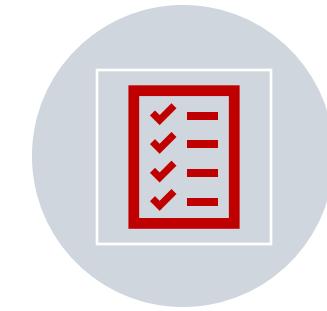
Culture-sensitive curation

Culturally diverse collection

Native data or culturally sensitive translation



MODELS



TASKS

Style transfer

Entity adaptation

Explanation by analogy

Tasks

Entity adaptation



"I saw Merkel eating a Berliner from Dietsch on the ICE"



I saw Biden eating a Boston Cream from Dunkin' Donuts on the Acela

Adapting Entities across Languages and Cultures
(Peskov et al., Findings 2021)

Recipe adaptation

凉拌秋葵

用料

- | | | | |
|------|-------|------------|------|
| • 秋葵 | 20根左右 | • 香油 | 1勺 |
| • 生抽 | 2-3勺 | • 糖 | 1勺 |
| • 醋 | 1勺 | • 蒜 | 3-5瓣 |
| • 蚝油 | 1勺 | • 盐 | 酌量 |
| | | • 绿芥末膏不用也行 | 酌量 |

做法

- 将秋葵洗净放开水中焯2分钟左右。
- 开水中放盐一勺，油一勺，这样秋葵颜色翠绿鲜艳) ...



Chinese Okra Salad

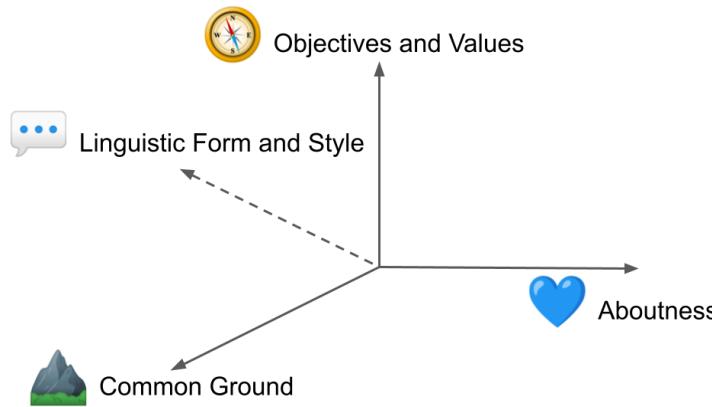
Ingredients

- 8 oz (225 g) okra
- 2 teaspoons light soy sauce (or soy sauce)
- 1/2 teaspoons green Sichuan pepper oil (or more to taste)

Instructions

- Bring a medium pot of water to a boil. Add 1 teaspoon vegetable oil and a pinch of salt...

Summary



(Multilingual) language models are getting better and better

Meaning representations help with efficiency, interpretability, control

We must consider culture in cross-lingual/multilingual NLP

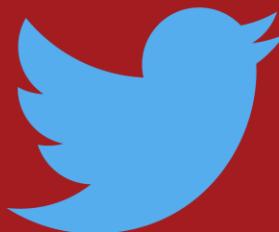
Thanks!



danielhers.github.io



dh@di.ku.dk



@daniel_hers



sigmoid.social/@dh