Challenges and Strategies in Cross-Cultural NLP

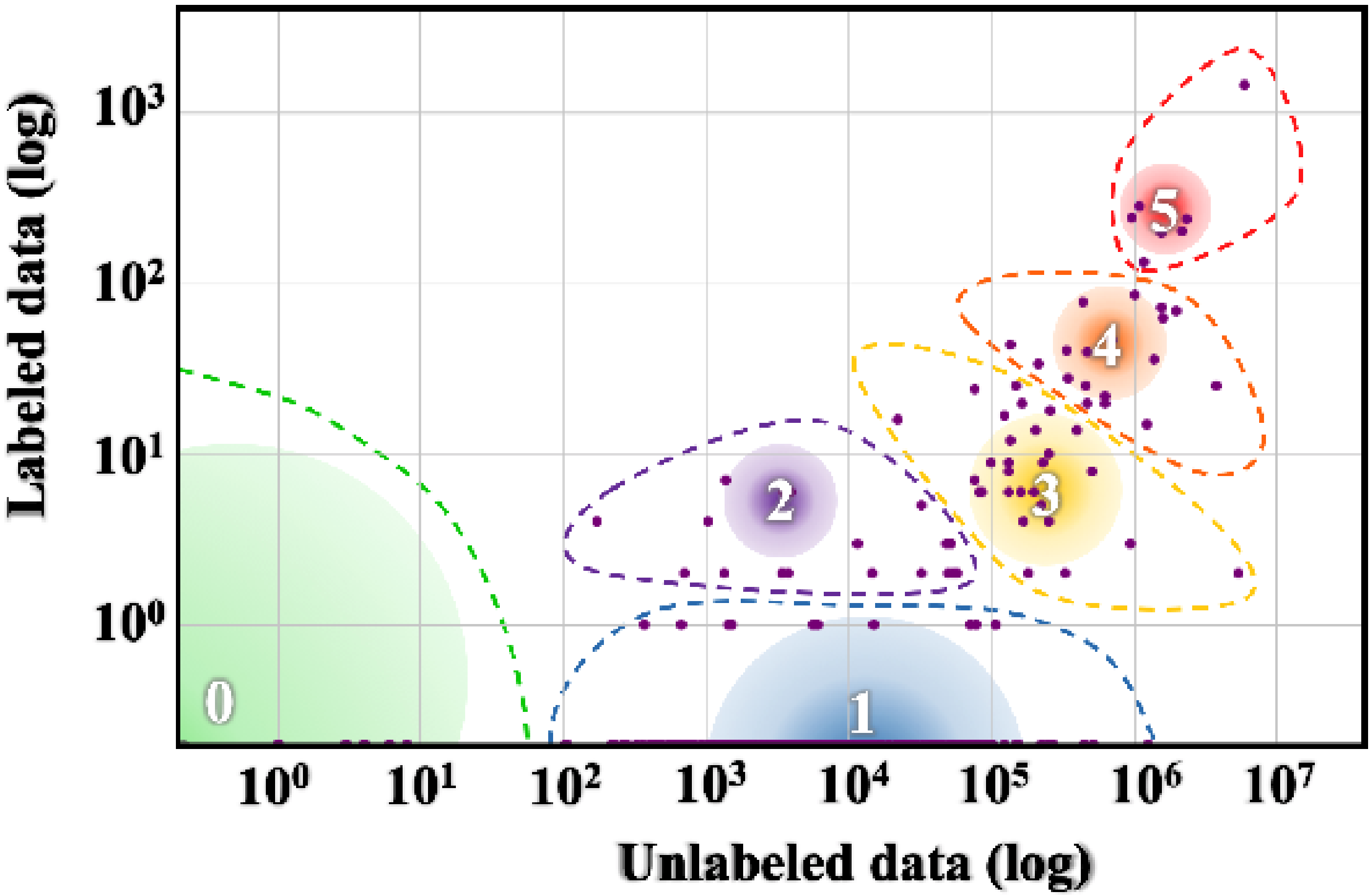
Daniel Hershcovich, Stella Frank,
Heather Lent, Miryam de Lhoneux,
Mostafa Abdou, Stephanie Brandl,
Emanuele Bugliarello, Laura Cabello Piqueras,
Ilias Chalkidis, Ruixiang Cui, Constanza Fierro,
Katerina Margatina, Phillip Rust
and Anders Søgaard

Published at ACL 2022

LIIR Journal Club, KU Leuven
21 April 2022

UNIVERSITY OF COPENHAGEN

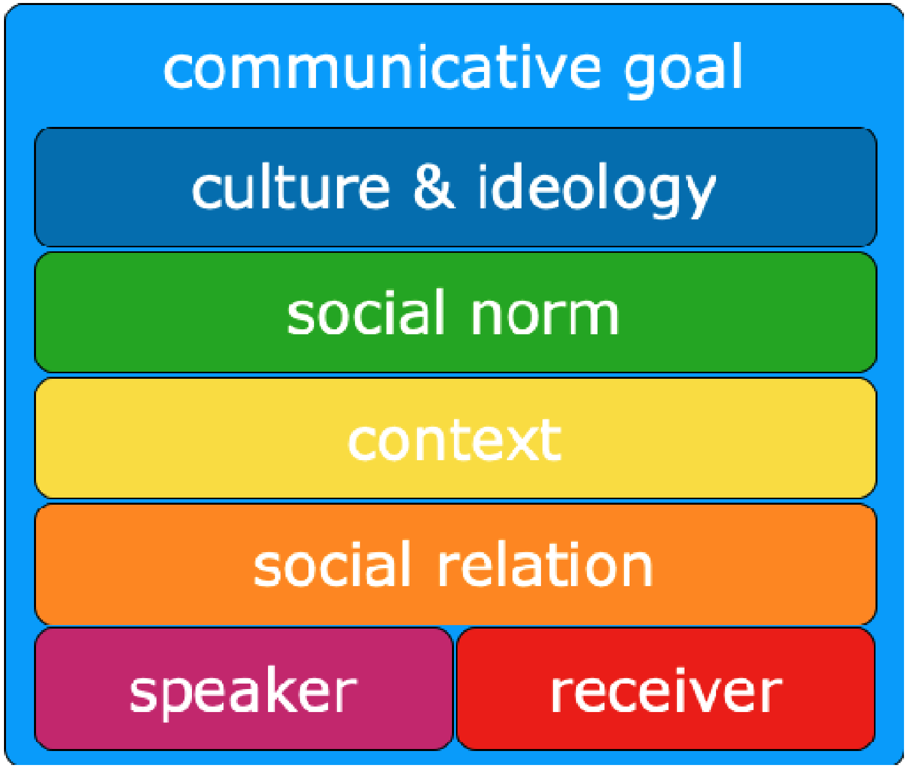
Resource disparity for languages



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

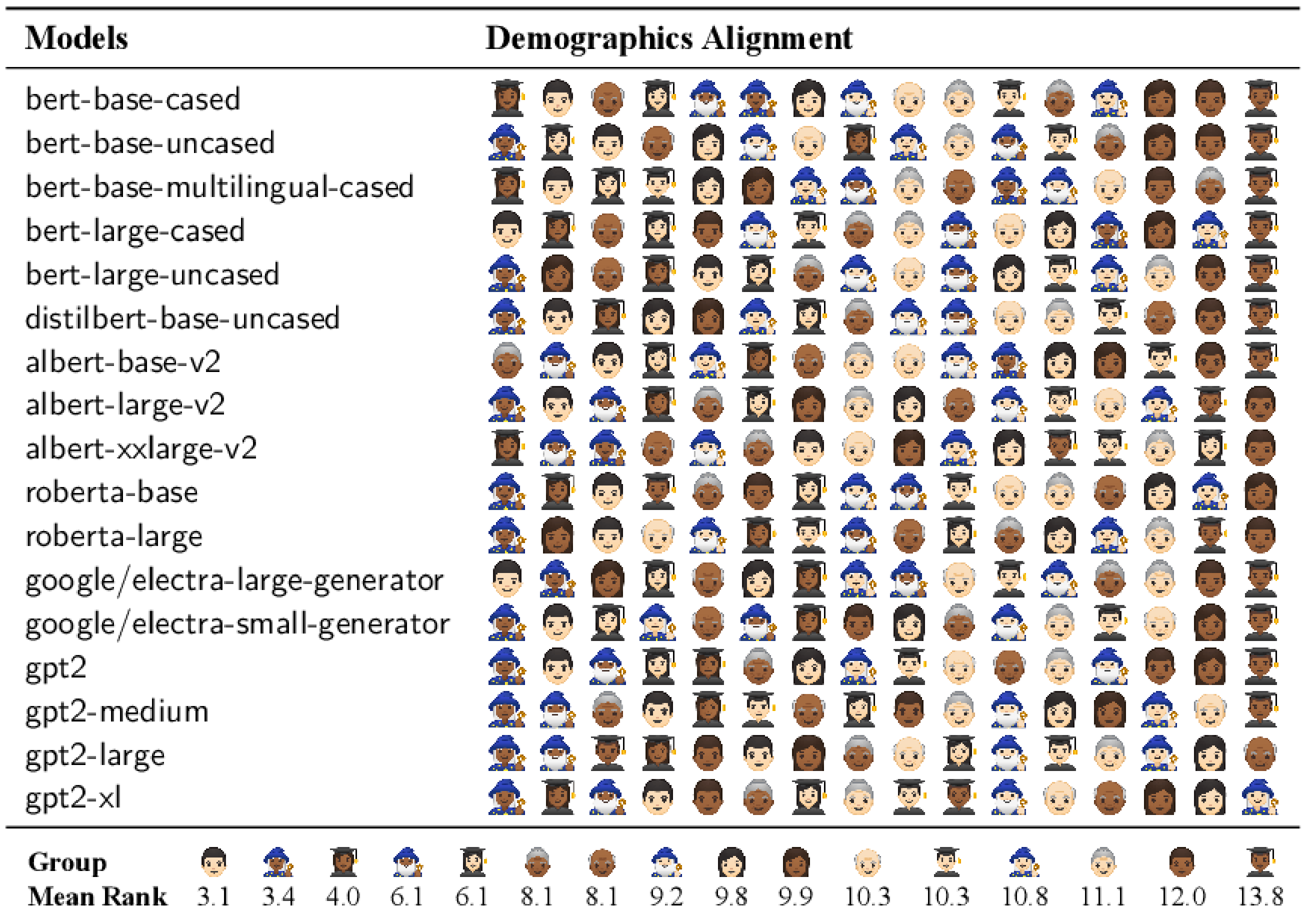
Social factors

NLP is for people (not just languages)



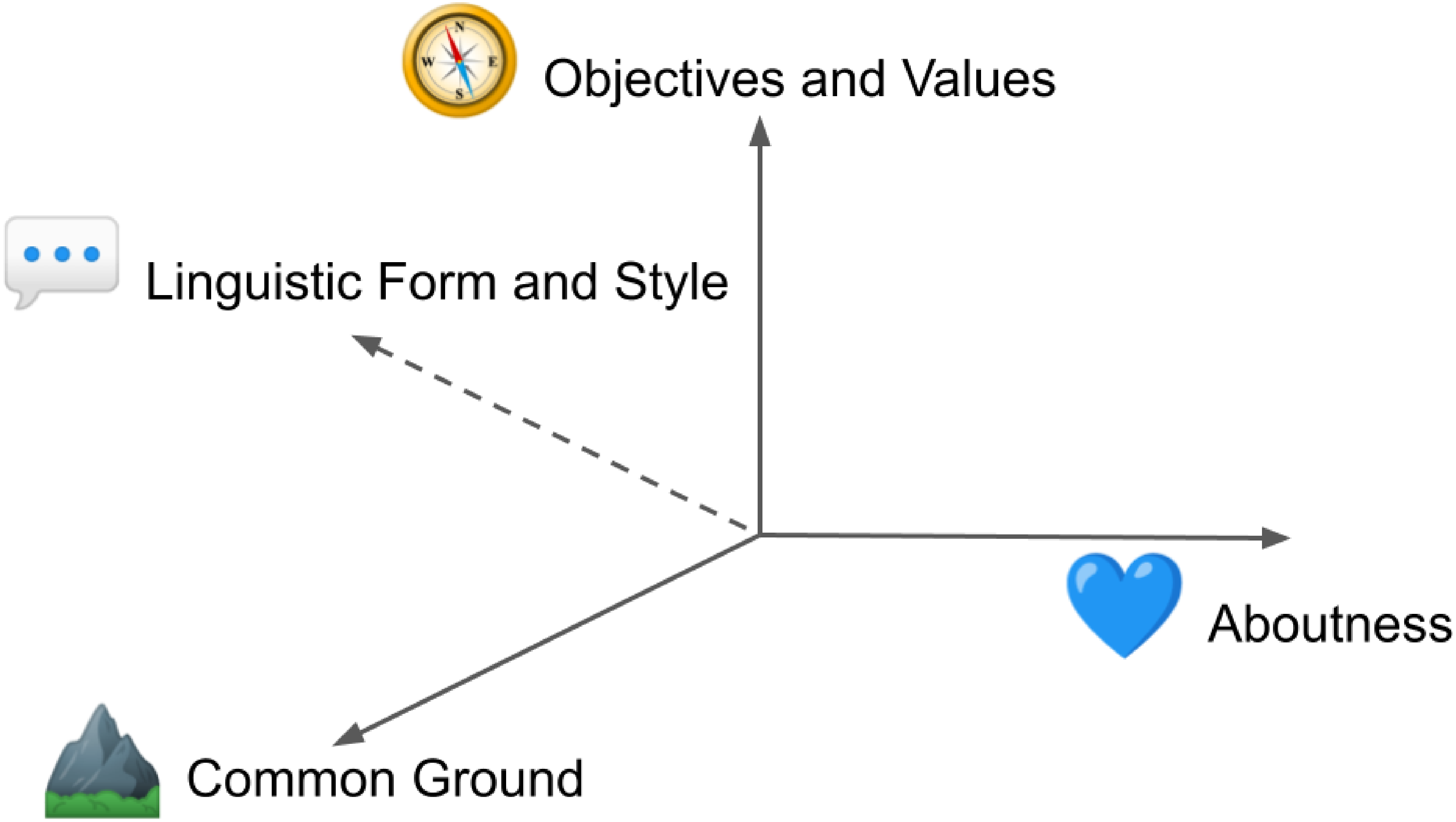
The Importance of Modeling Social Factors of Language: Theory and Practice
(Hovy & Yang, NAACL 2021)

Social bias in language models



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

Dimensions of culture



Form

How we express ourselves in language

- Morphosyntax
- Word choice
- Style

Style

Stylistic aspects of linguistic form:

- Directness
- Formality
- Politeness
- Emotional expression

Levels of granularity

Linguistic and cultural variation within groups

Idiolect
Individual, personality

Sociolect, dialect
Social group or region, sub-culture

Standardised language
Country, national culture

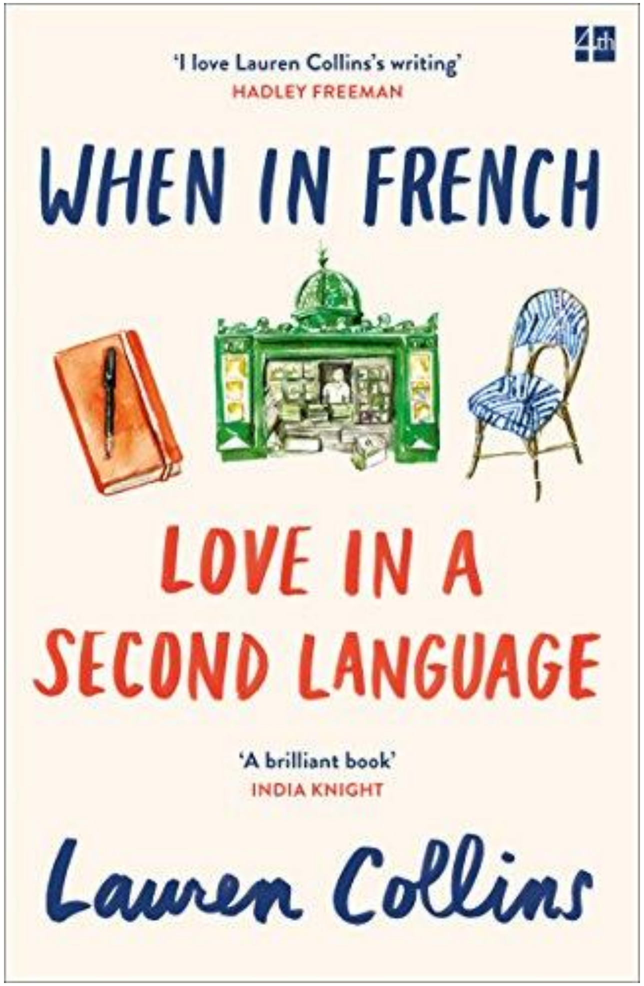
Language, language family
International cultures

Common ground

Shared knowledge based on which people reason and communicate

- Conceptualisation
- Commonsense

Common ground



- Conceptualisation
- Commonsense
- Stories
- Metaphors
- Clichés
- …

Conceptualisation

Objects	Colours	Kinship
Space	Time	Events

Objects

Swahili
leso

Visually Grounded Reasoning across Languages and Cultures

(Liu et al., EMNLP 2021)

Events and rituals

Tamil
ஜல்லிக்கட்டு
jallikattu

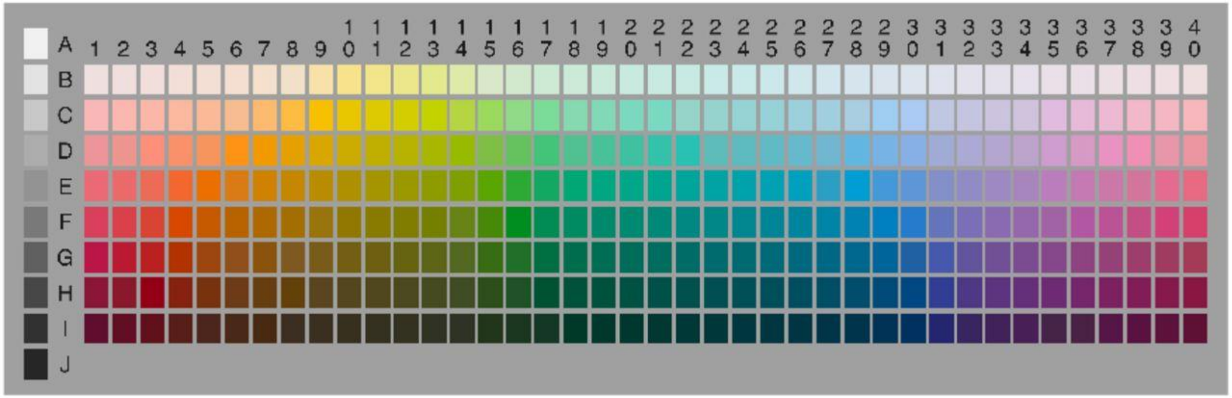
Visually Grounded Reasoning across Languages and Cultures

(Liu et al., EMNLP 2021)

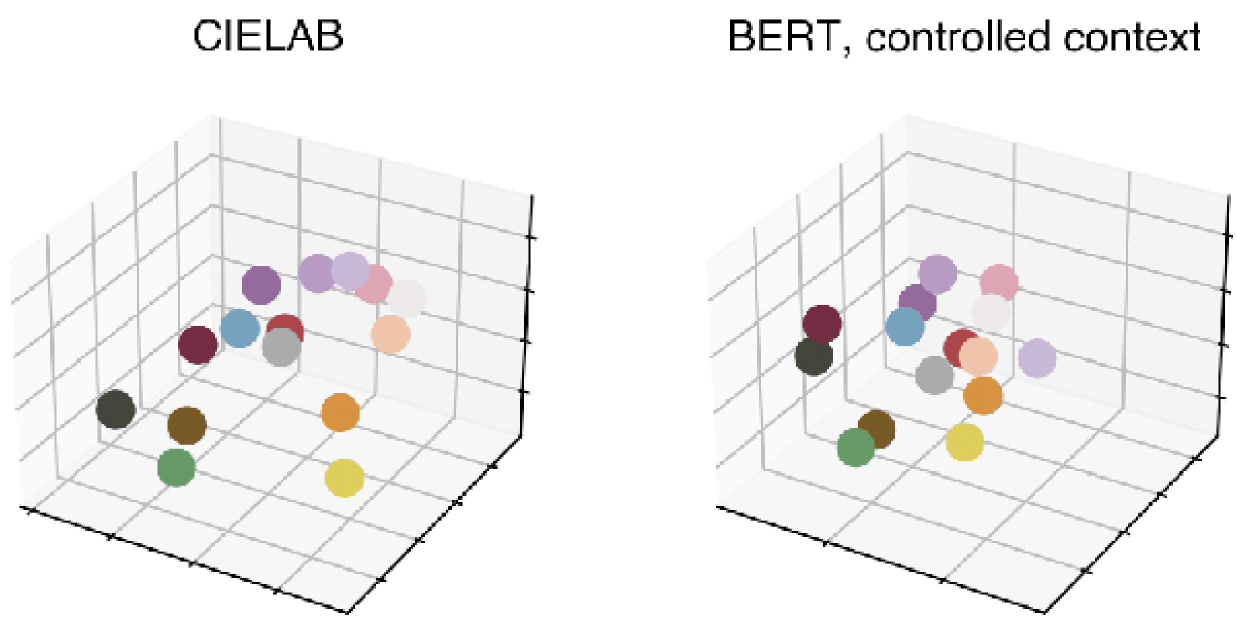
Visual concepts include culture-specific activities that cannot be mapped across cultures.

Colour

World Colour Survey



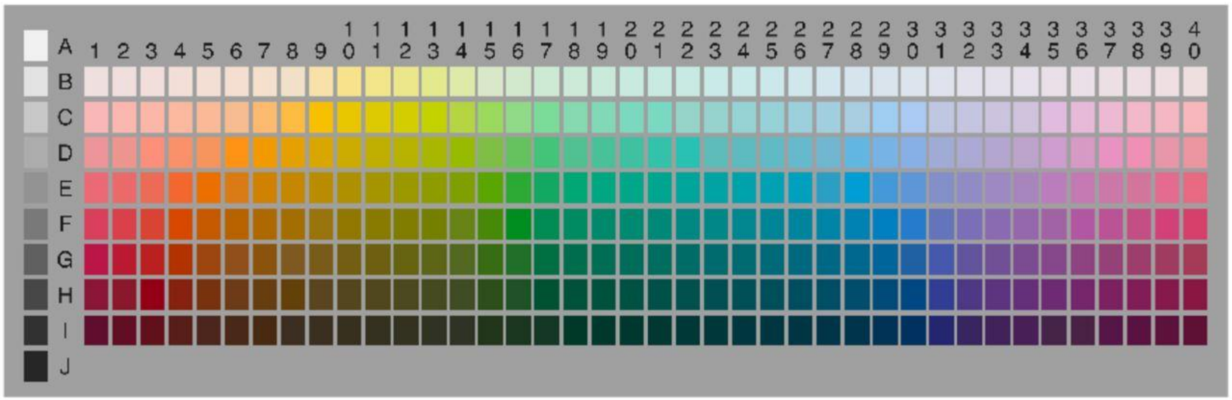
Probing colour



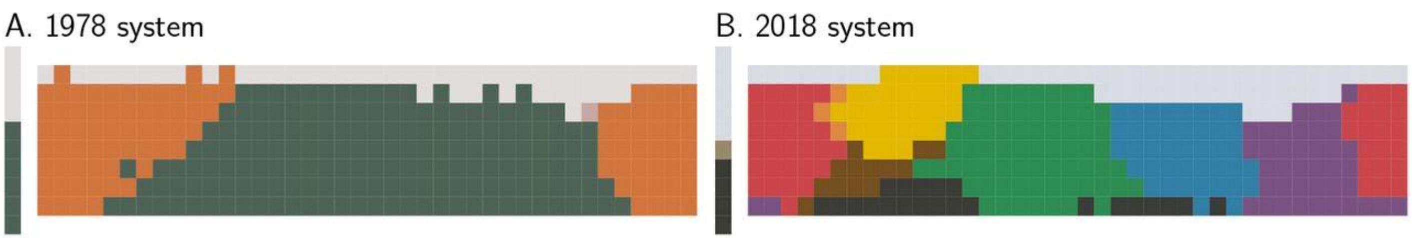
English BERT aligns with English-speaking Americans.
(What about others?)

Can Language Models Encode Perceptual Structure Without Grounding? A Case Study in Color
(Abdou et al., CoNLL 2021)

Differences in colour grounding



Nafaanra, a language of Ghana and Côte d'Ivoire



[The evolution of color naming reflects pressure for efficiency: Evidence from the recent past]

(Zaslavsky et al., Journal of Language Evolution 2022)

Commonsense

- Physical
- Social
- Taxonomic
- Temporal

"Commonsense is the basic level of practical knowledge and reasoning concerning everyday situations and events that are commonly shared among most people."

Commonsense Reasoning for Natural Language Processing
(Sap et al., ACL 2020 Tutorial)

Culture-dependent

Visual commonsense

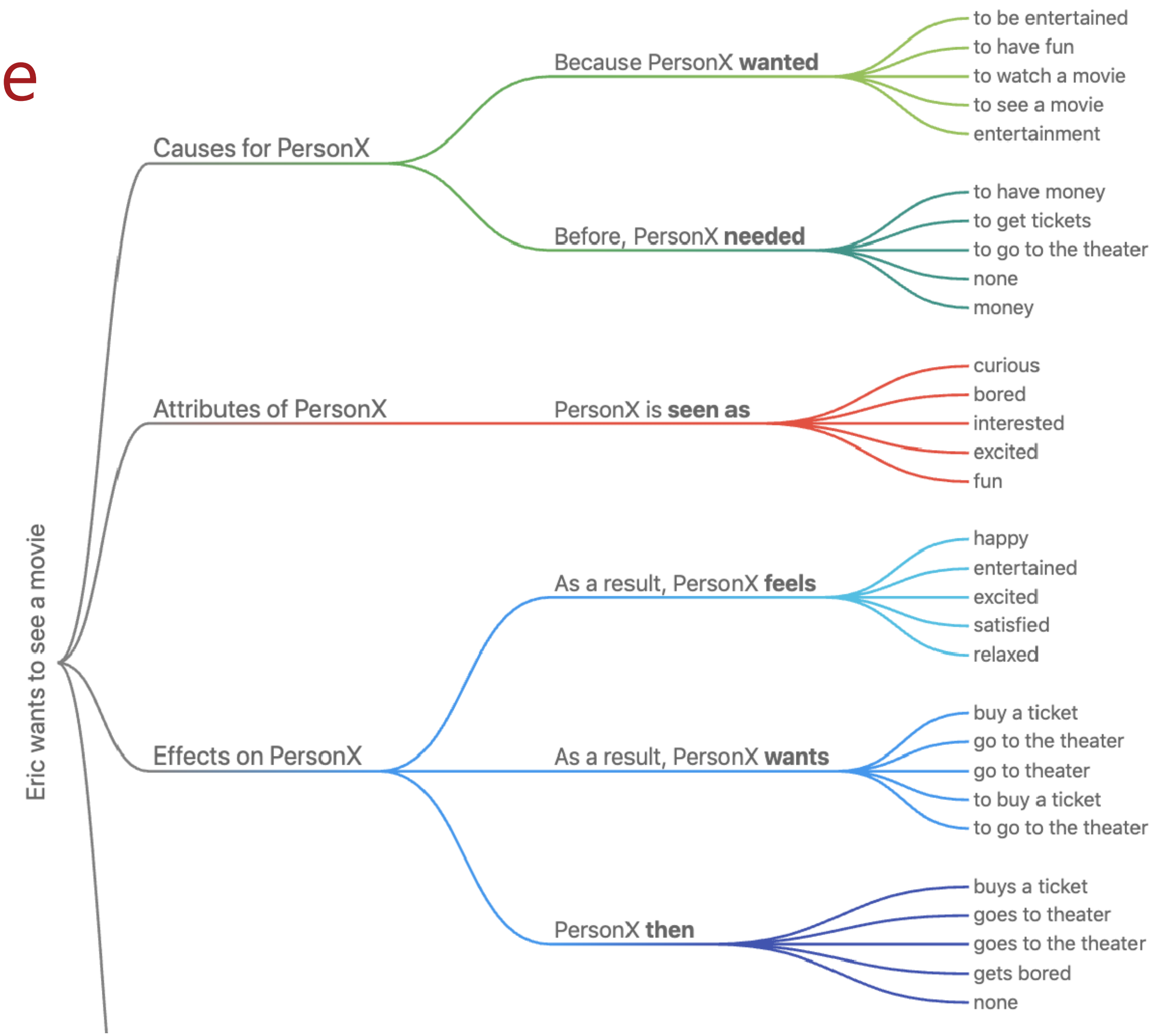
Bola basket (Indonesian) *Mpira wa kikapu* (Swahili) 篮球 (Chinese) *Basketbol* (Turkish) கூடைப்பந்தாட்டம் (Tamil)

[Visually Grounded Reasoning across Languages and Cultures]

(Liu et al., EMNLP 2021)

Commonsense

Some knowledge is "universal", other culture-specific



COMET: Commonsense Transformers for Automatic Knowledge Graph Construction
(Bosselut et al., ACL 2019)

Commonsense

Towards an Atlas of Cultural Commonsense for Machine Reasoning
(Acharya et al., CSKGs 2021)

Knowledge bias in language models

"[X] was created in [Y]"

en
Japan (170), Italy (56)

de
Deutschland (217), Japan (70)

nl
Nederland (172), Italië (50)

it
Italia (167), Giappone (92)

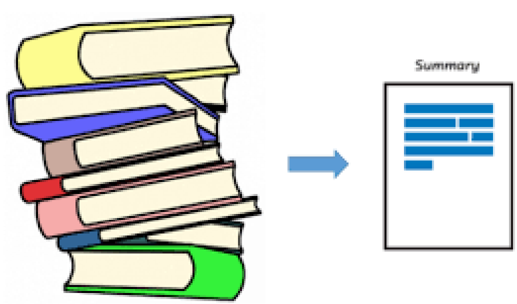
The language of prompting affects the model's answer to prompts

Multilingual LAMA: Investigating Knowledge in Multilingual Pretrained Language Models (Kassner et al., EACL 2021)

Aboutness

What content do people *care about*?

- Related to topic/domain



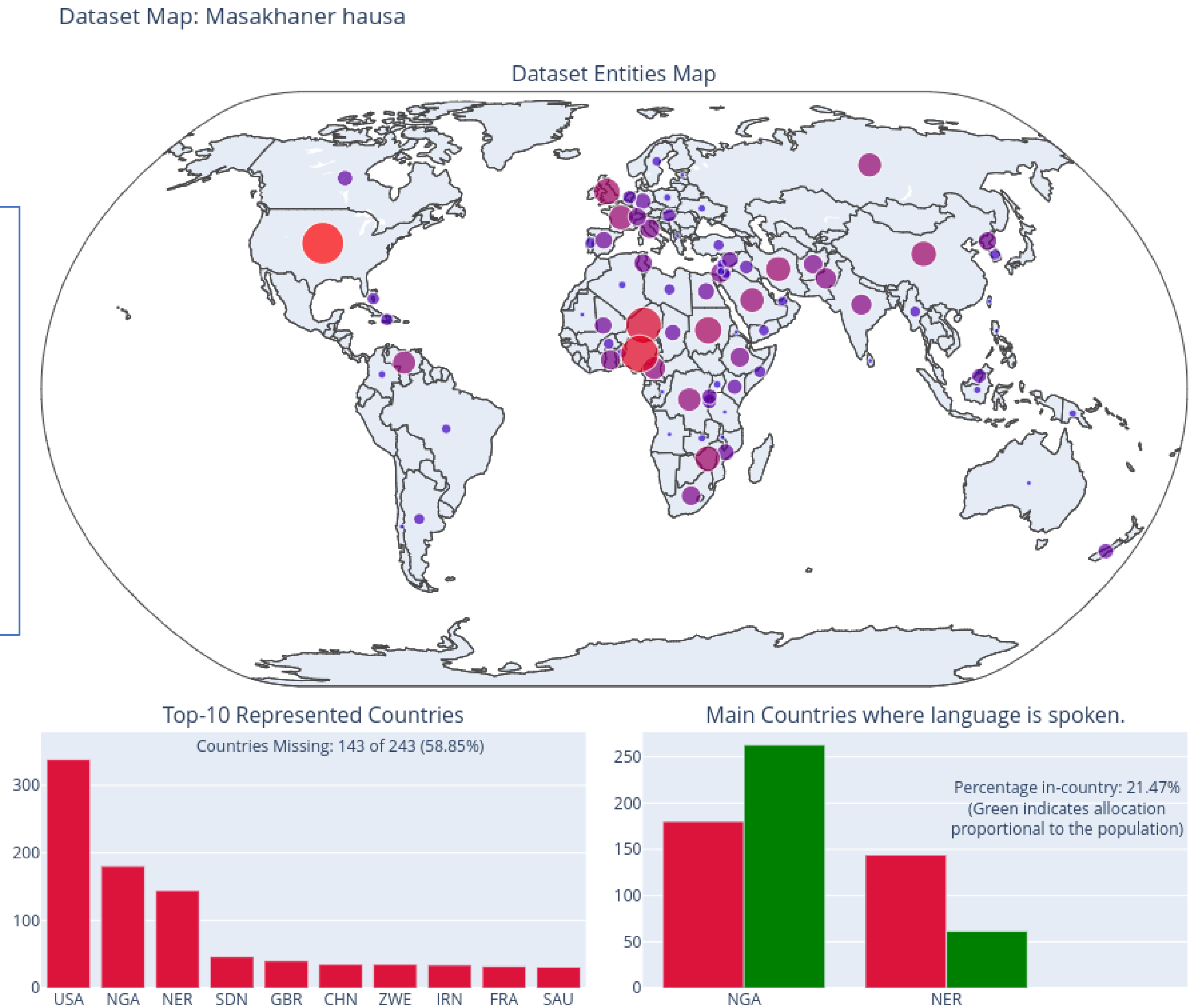
Entities

Experiences

Aspects

Entities

USA & Europe are over-represented in datasets across languages

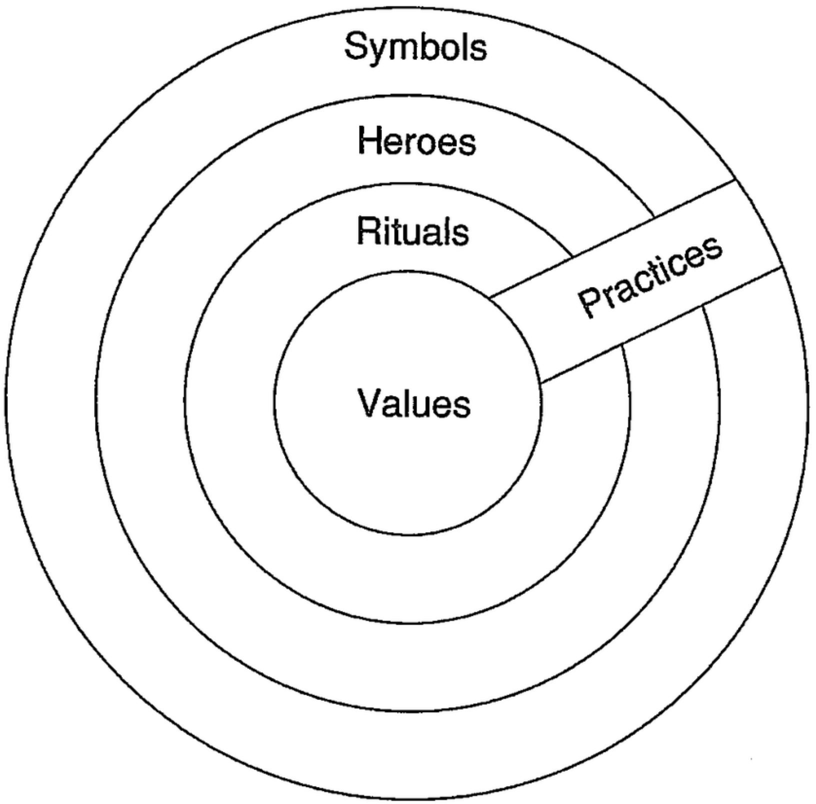


Dataset Geography: Mapping Language Data to Language Users
(Faisal et al., ACL 2022)

Values

Objectives and goals people strive for

- What is considered desired or desirable



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

(Meta) values

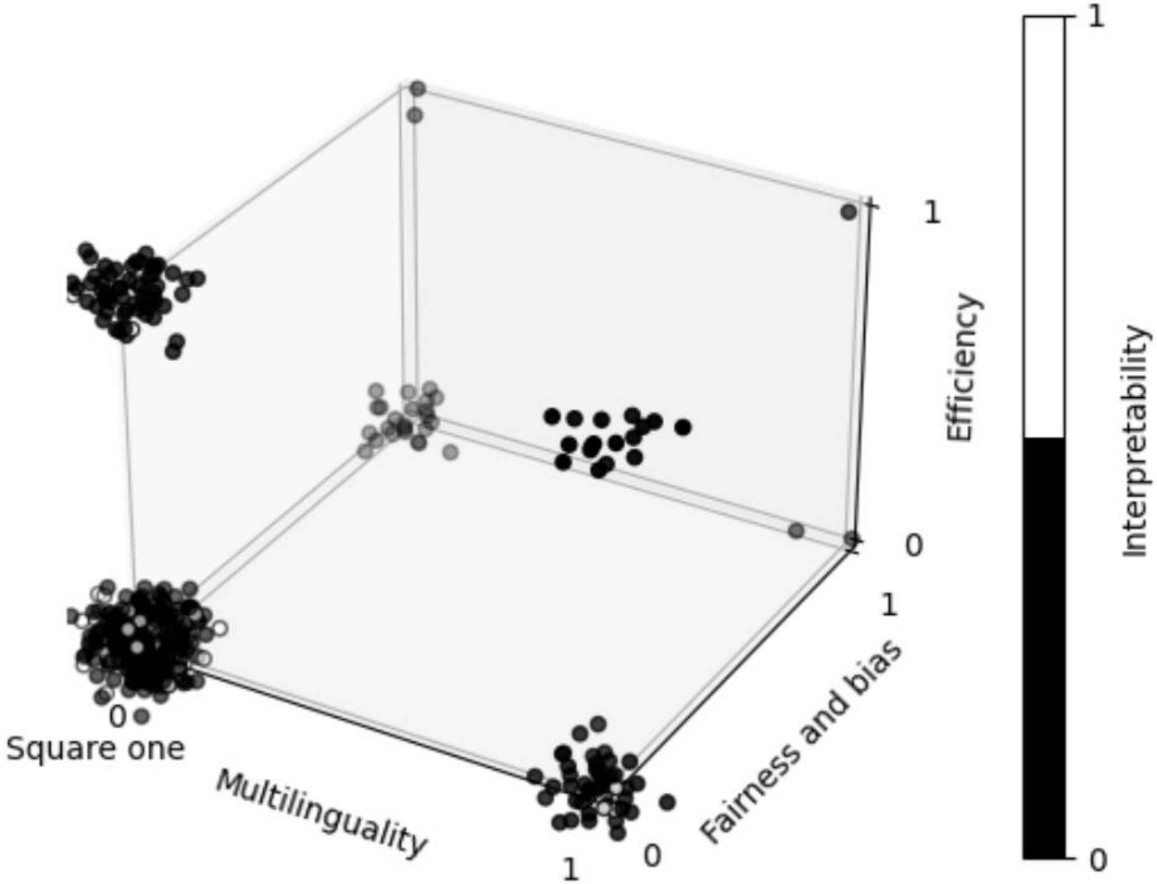
Why are we doing NLP?

- Users may have different goals, often implicit

No single correct answer.

Changing the World by Changing the Data (Rogers, ACL 2021)

Common meta-objectives in NLP



Accuracy, fairness, etc. reflect the values of NLP researchers

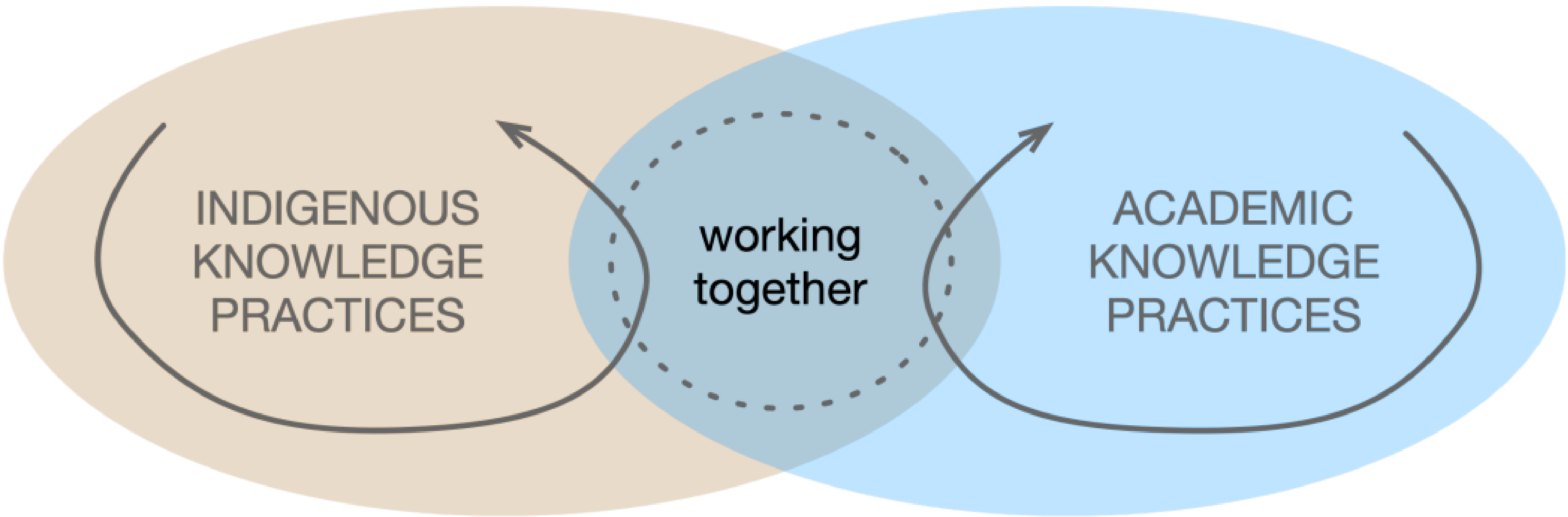
Square One Bias in NLP: Towards a Multi-Dimensional Exploration of the Research Manifold (Ruder et al., ACL 2022)

Conflicting objectives between stakeholders

- Researchers
- Practitioners
- End-users
- Affected communities
- Regulators

Transparent values facilitate adaptation and decision making

Language technology for all (potential) users



Local Languages, Third Spaces, and other High-Resource Scenarios (Bird, ACL 2022)

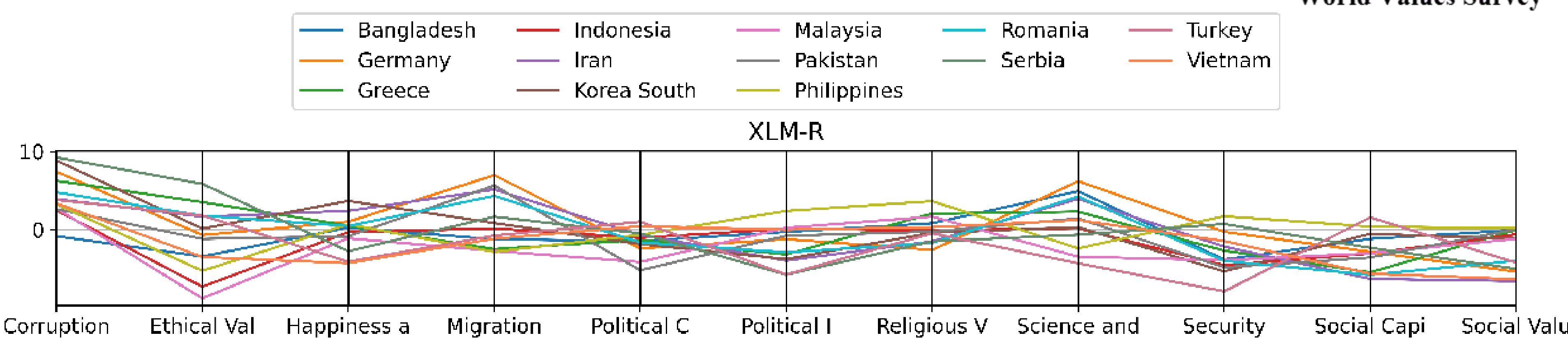
Benefit to all requires finding the intersection, particularly for local languages

Value bias in language models

For each of the following, indicate how important it is in your life. Would you say it is (*read out and code one answer for each*):

		Very important	Rather important	Not very important	Not at all important
Q1	Family	1	2	3	4
Q2	Friends	1	2	3	4
Q3	Leisure time	1	2	3	4
Q4	Politics	1	2	3	4
Q5	Work	1	2	3	4
Q6	Religion	1	2	3	4

World Values Survey



Probing Pre-Trained Language Models for Cross-Cultural Differences in Values (Arora et al., 2022)

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

World Values Survey

GPT-3

summarise

(translation)

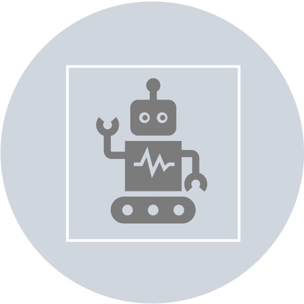
"1. I am in favor of limiting immigration.
2. I am in favor of limiting immigration for humanitarian reasons.
3. I am in favor of limiting immigration for economic reasons."

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

Strategies

Existing and potential ways to address the challenges

DATA



MODELS

TASKS

Data

Selection

Annotation

Projection

Culture-sensitive curation

Culturally diverse collection

Native data or culturally sensitive translation

Models

Training

Transfer

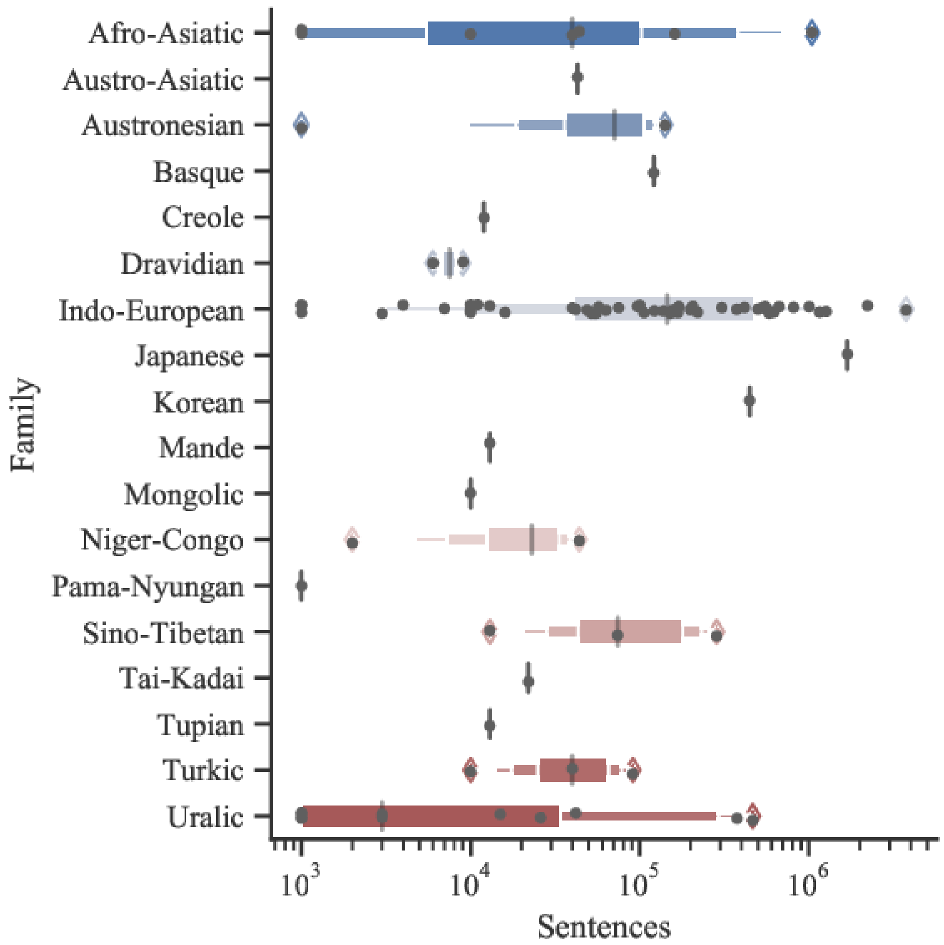
Pre-training

Robust optimisation

Balanced sampling

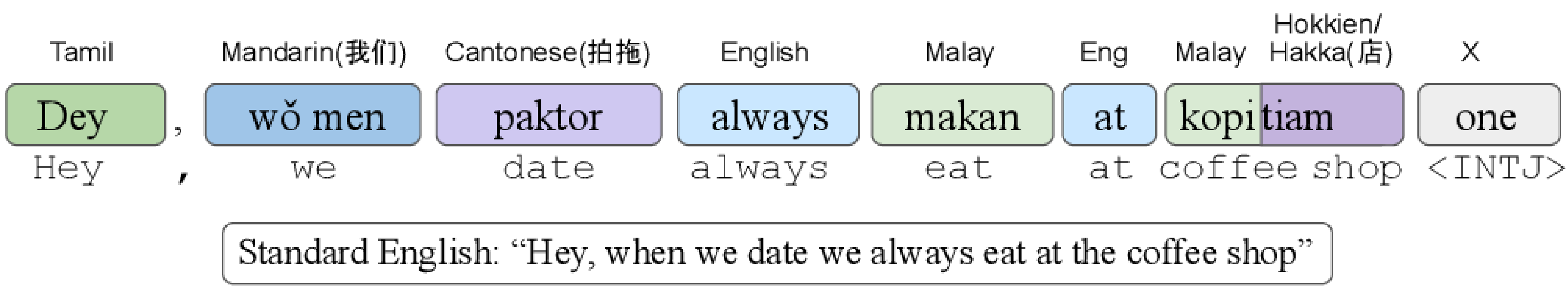
Measuring cultural bias

Robust optimisation



Minimax and Neyman–Pearson Meta-Learning for Outlier Languages
(Ponti et al., Findings 2021)

Robust optimisation does not always work



[On Language Models for Creoles](#)
(Lent et al., CoNLL 2021)

Tasks

Cross-cultural translation

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

↓

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

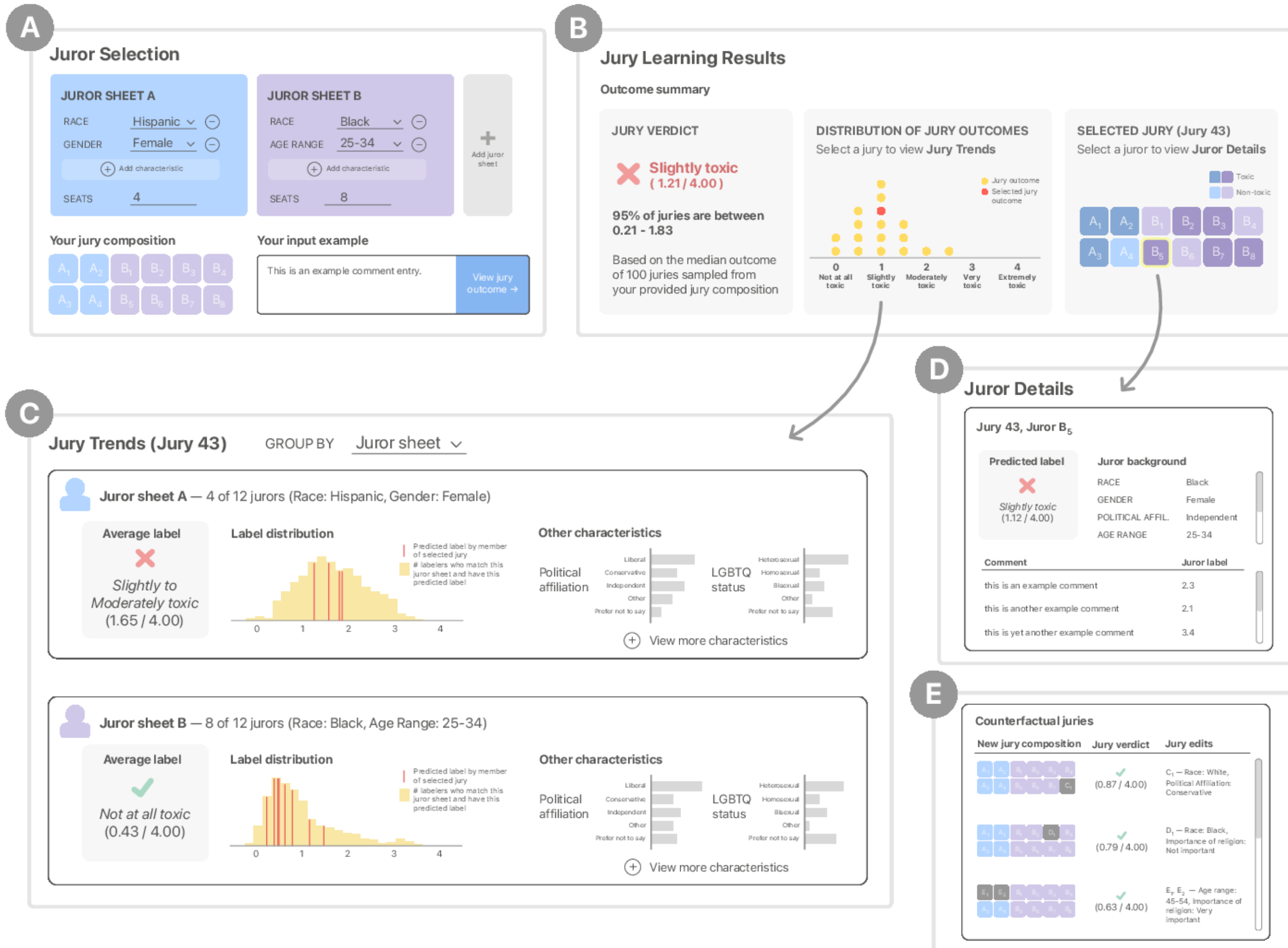
Adapting Entities across Languages and Cultures
(Peskov et al., Findings 2021)

Style transfer

Entity adaptation

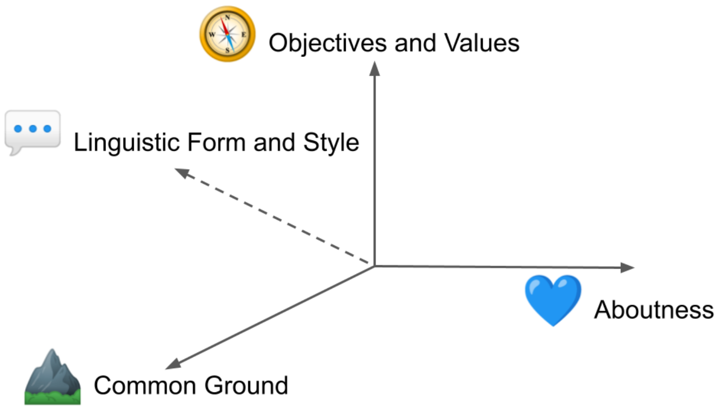
Explanation by analogy

Multi-granularity adaptation



[Jury Learning: Integrating Dissenting Voices into Machine Learning Models](#)
(Gordon et al., CHI 2022)

Summary



NLP is for people (not just languages)

Culture is multidimensional

Objectives may be in conflict

Generalisation-representation trade-off

danielhers.github.io
dh@di.ku.dk
@daniel_hers