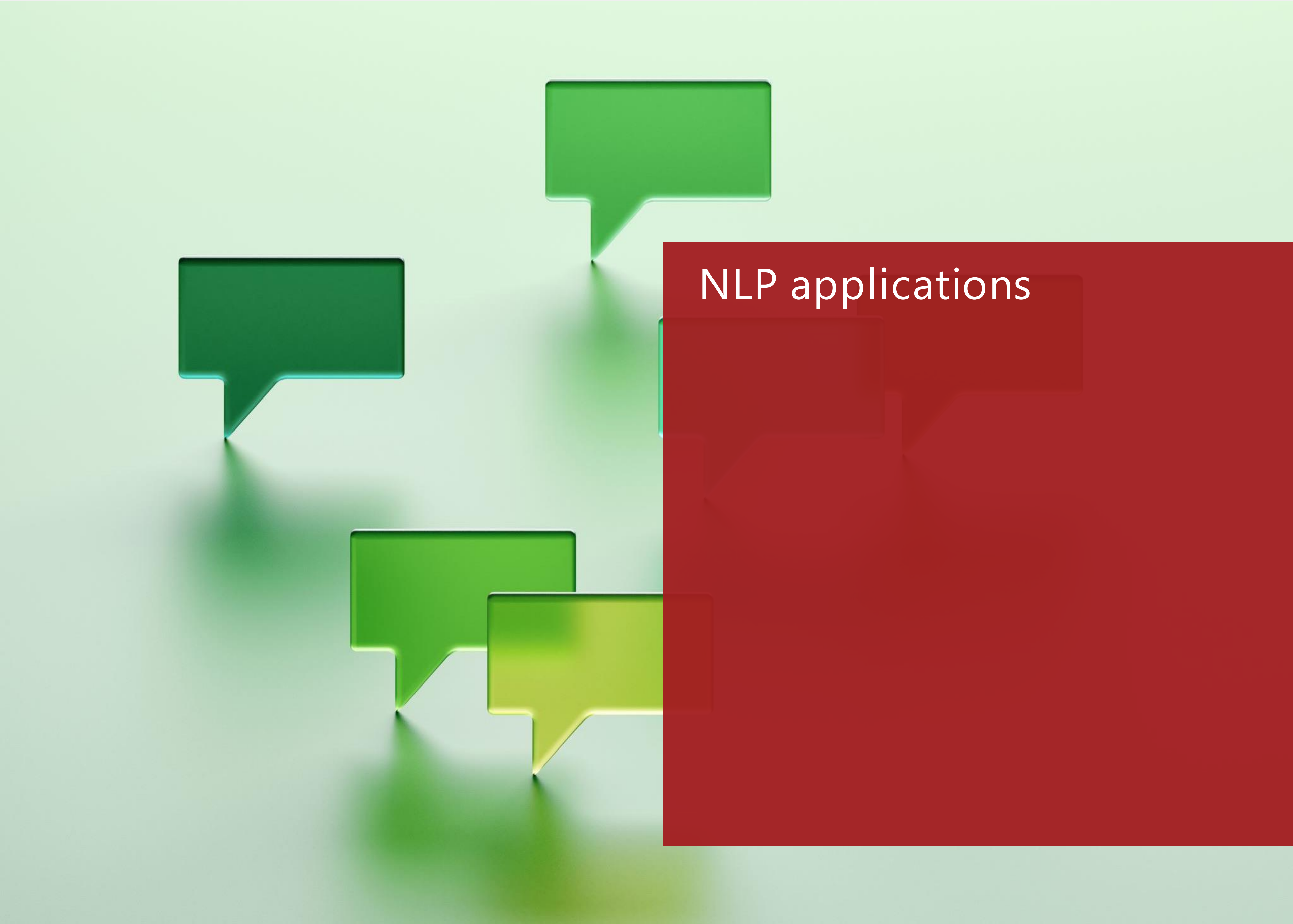
Cultural and Environmental Considerations in Natural Language Processing

University of Haifa
Information Systems Seminar
30 March 2022

Daniel Hershcovich

UNIVERSITY OF COPENHAGEN



NLP applications

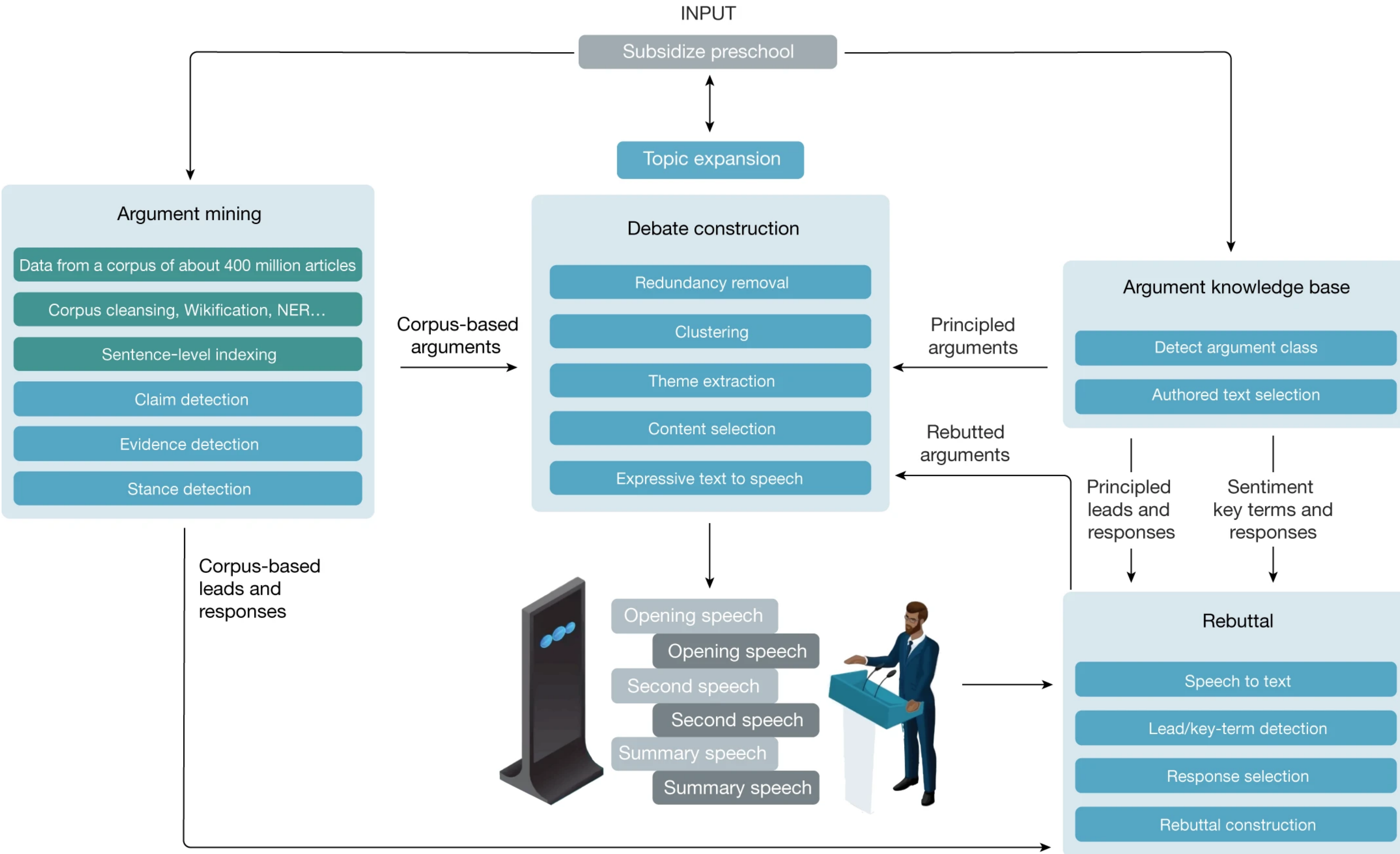
Debating

	Pre-debate: both sides receive the motion and prepare	15 min
	Moderator introduces the motion to the audience	
Opening speeches	Project Debater delivers the ‘government’ opening speech	4 min
	Human debater delivers the ‘opposition’ opening speech and replies	4 min
Second speeches	Project Debater offers rebuttal and additional points	4 min
	Human debater offers rebuttal and additional points	4 min
Summary speeches	Project Debater provides final rebuttal and closing statements	2 min
	Human debater provides final rebuttal and closing statements	2 min

[An autonomous debating system](#)

Slonim et al. (Nature 2021)

Debating



[An autonomous debating system](#)
Slonim et al. (Nature 2021)

Fact checking

Examples of disinformation

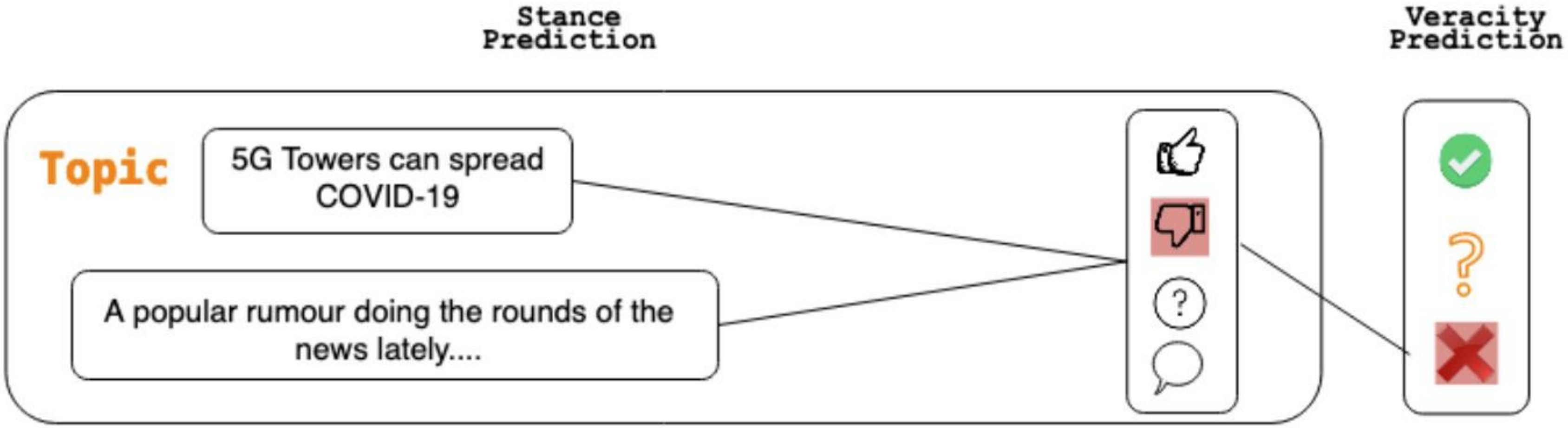
False claims such as ‘drinking bleach or pure alcohol can cure the coronavirus infections’: on the contrary, drinking bleach or pure alcohol can be very harmful. **Belgium’s Poison Control Centre has recorded an increase of 15% in the number of bleach-related incidents.**

Conspiracy theories, such as the claim that coronavirus is 'an infection caused by the world’s elites for reducing population growth'. The scientific evidence is clear: the virus comes from a family of viruses originating in animals that include other viruses such as SARS and MERS.

Claims that '5G installations would be spreading the virus'. These theories had no specific substantiation and led to attacks on masts.

https://ec.europa.eu/info/live-work-travel-eu/coronavirus-response/fighting-disinformation/tackling-coronavirus-disinformation_en

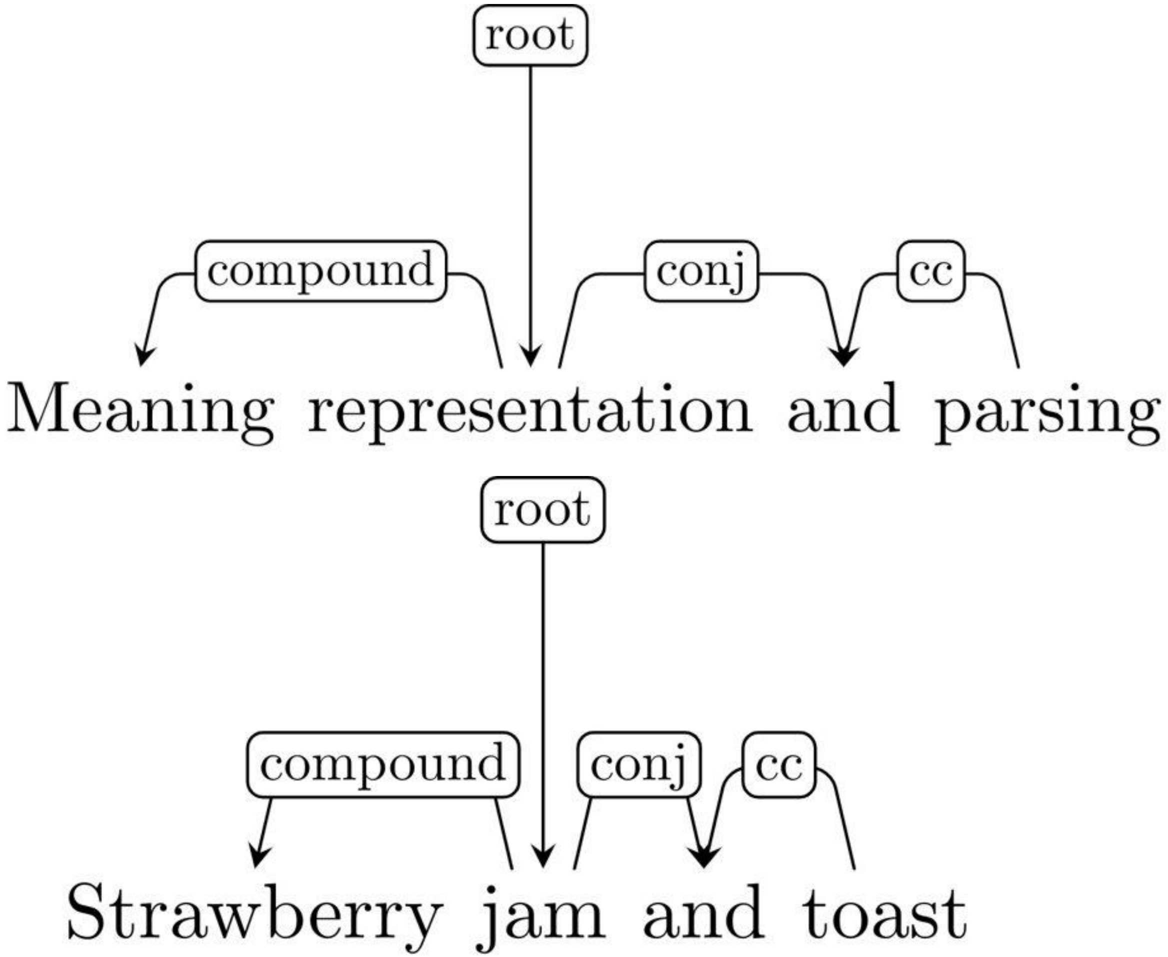
Fact checking



[A Survey on Stance Detection for Mis- and Disinformation Identification]
Hardalov et al. (2021).

NLP methods

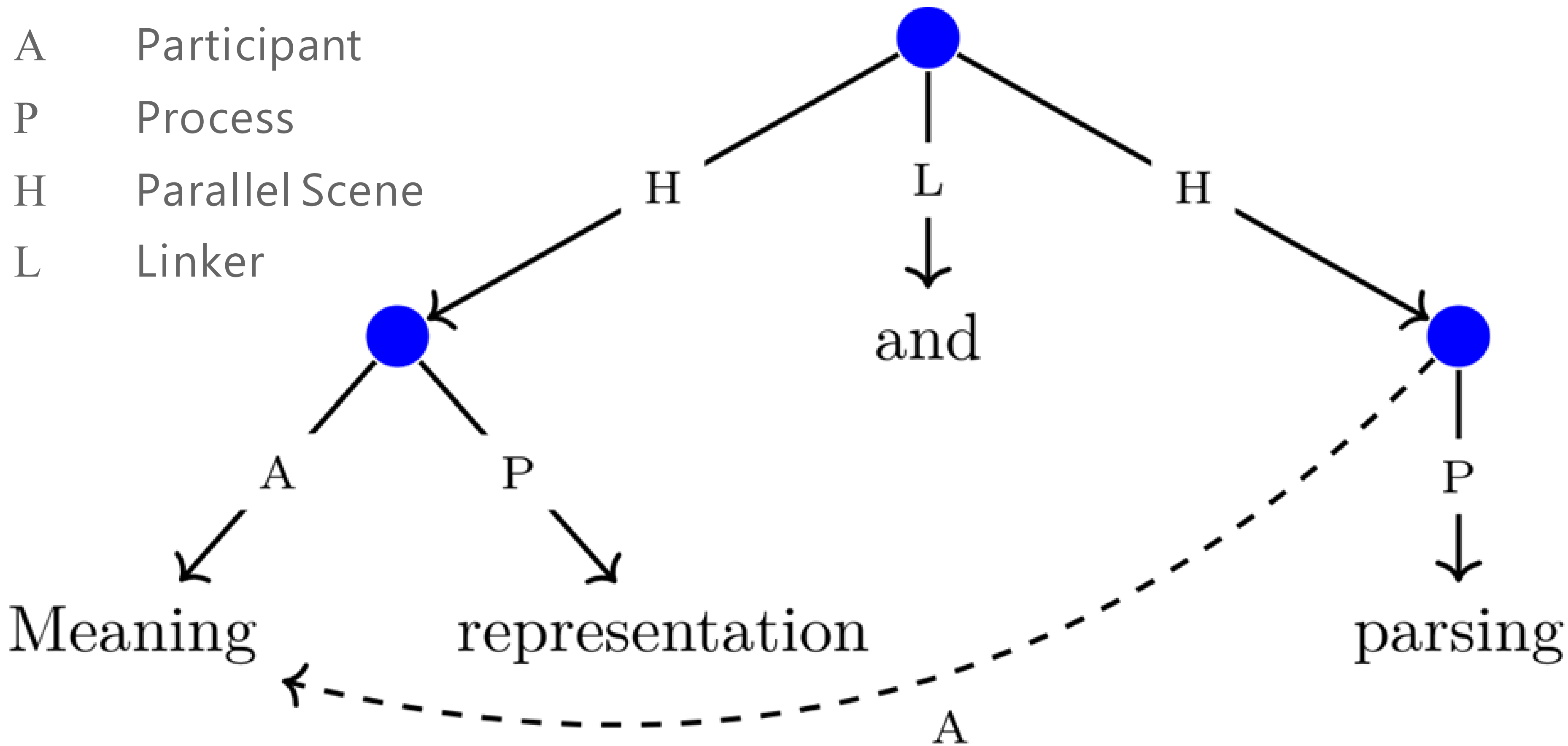
Syntactic dependency representation



Meaning representation

Example: UCCA

A	Participant
P	Process
H	Parallel Scene
L	Linker



Implicit meaning

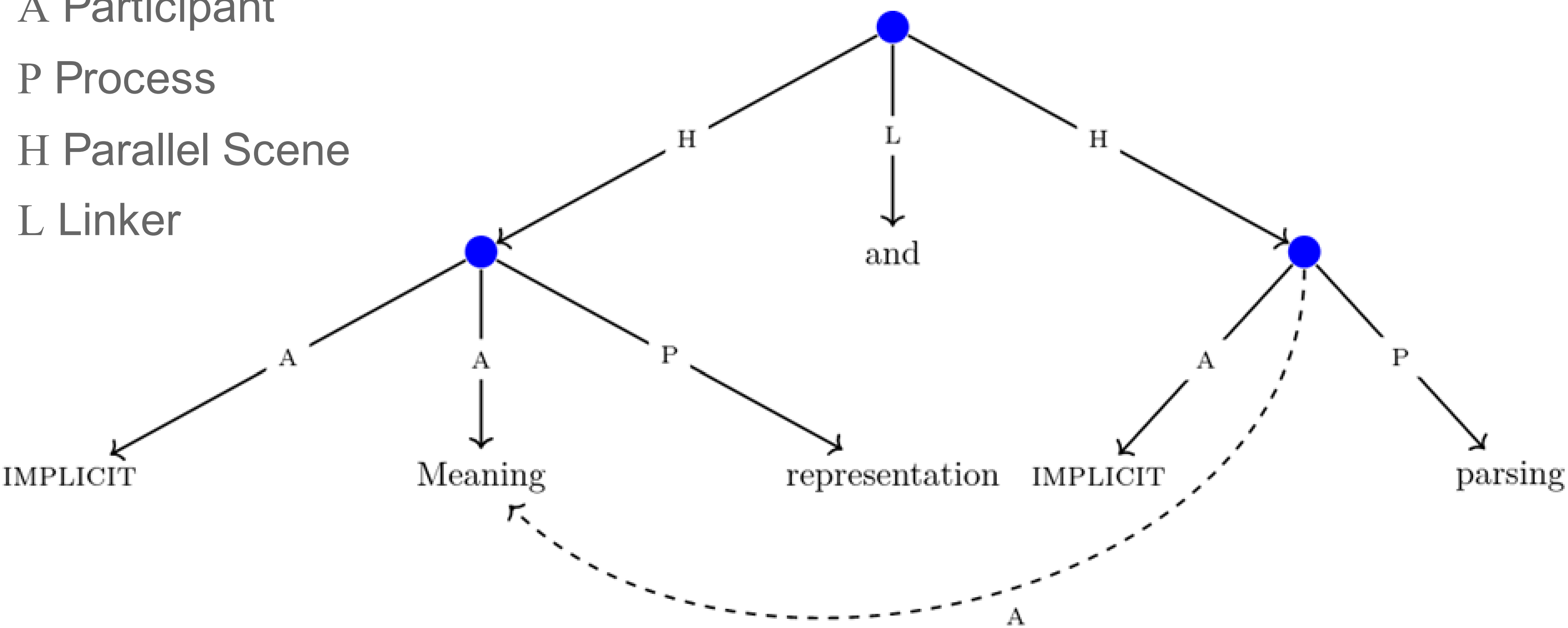
Example: UCCA

A Participant

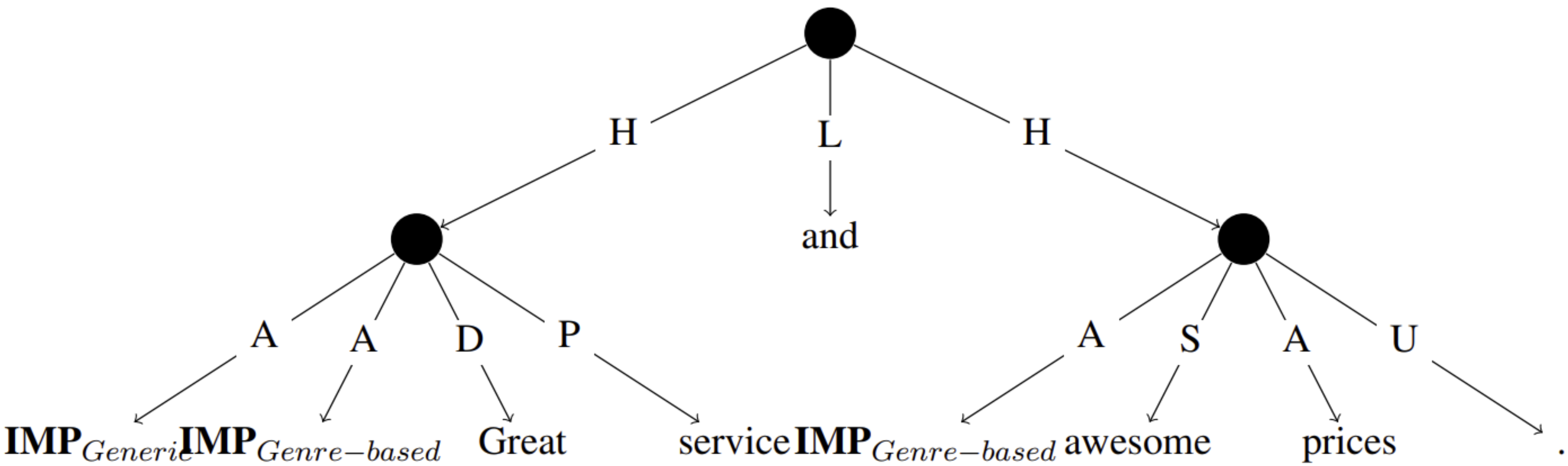
P Process

H Parallel Scene

L Linker



Implicit arguments

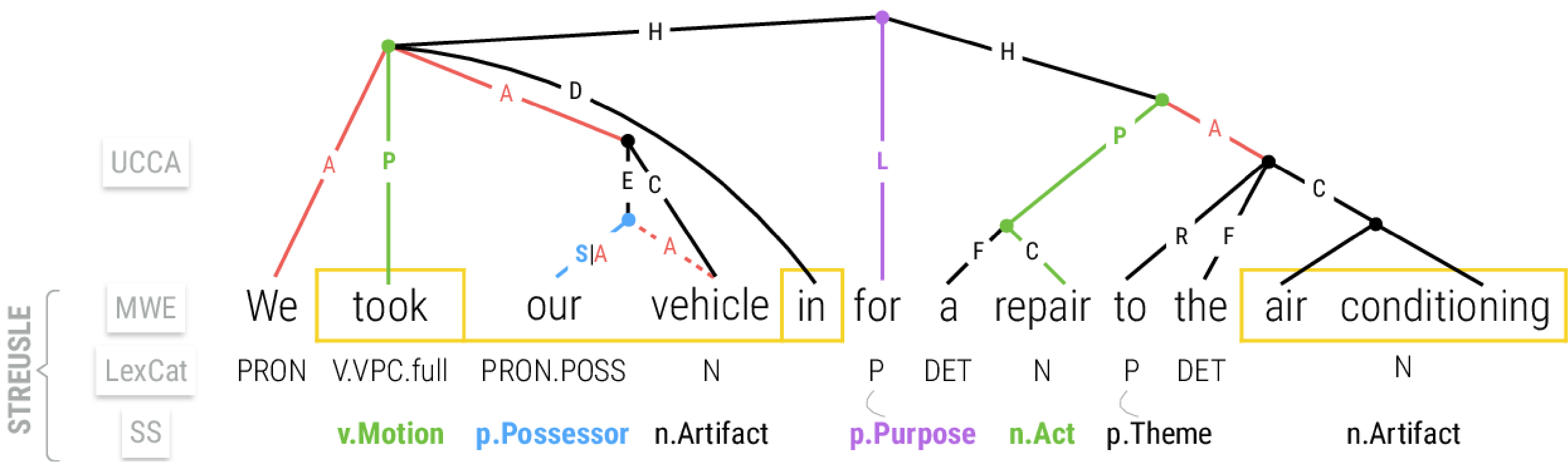


Great service and awesome prices.

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

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

Lexical and compositional meaning



[Comparison by Conversion: Reverse-Engineering UCCA from Syntax and Lexical Semantics]
(Hershcovich et al., COLING 2020)

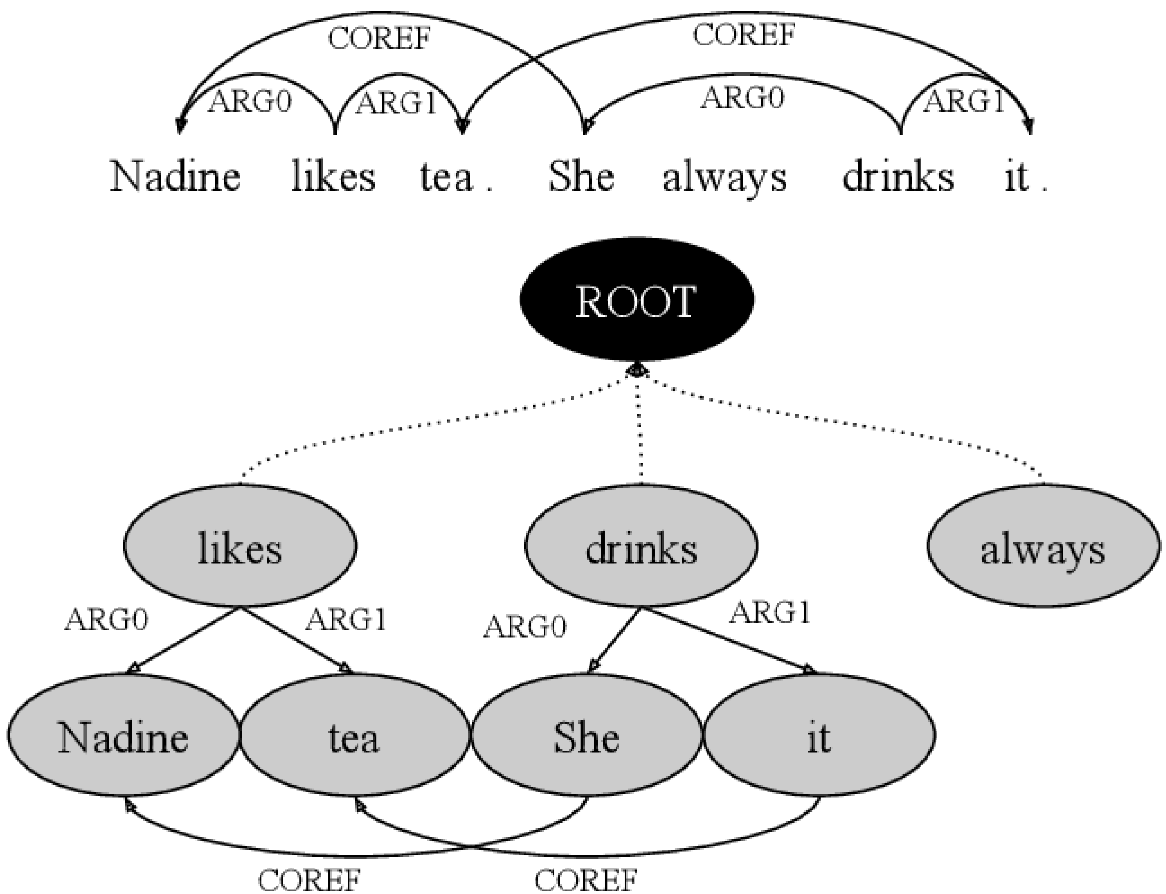
Coreference

[Lynyrd Skynyrd]$_1$ was formed in Florida$_2$. Other bands from [the Sunshine State]$_2$ include Fireflight and Marilyn Manson.

On March 19, Obama continued his outreach to the Muslim world, releasing a New Year's video message to the people and government of Iran. This attempt was rebuffed by the Iranian leadership.

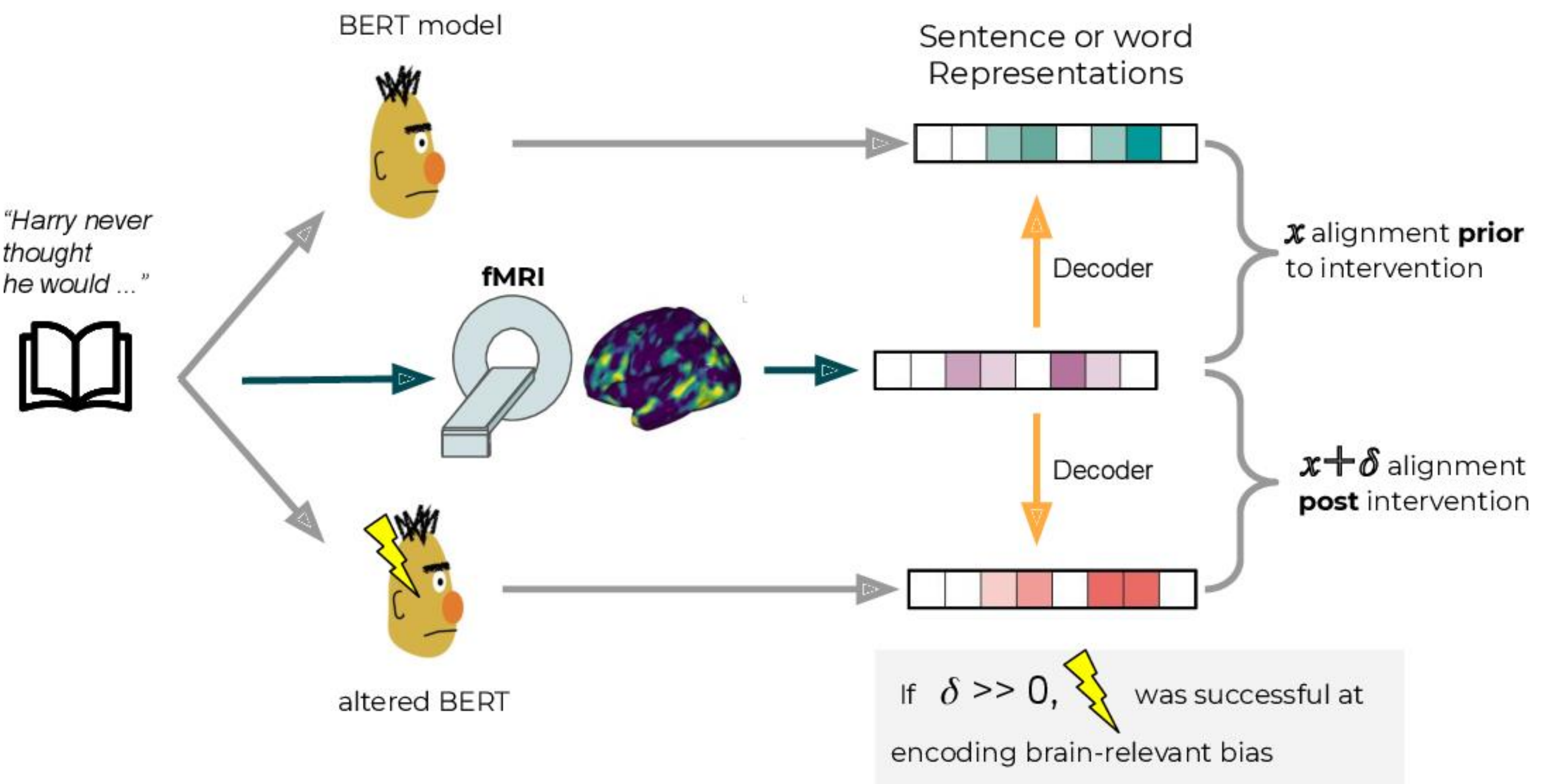
Rewarding Coreference Resolvers for Being Consistent with World Knowledge (Aralikatte et al., EMNLP 2019)

Semantic roles



[Joint Semantic Analysis with Document-Level Cross-Task Coherence Rewards]
(Aralikatte et al., AAAI 2021)

Investigating human language processing



[Does injecting linguistic structure into language models lead to better alignment with brain recordings?]()
(Abdou et al., 2021)

Question answering from knowledge bases

Lang.	Question
En	Did Lohengrin's male actor marry Margarete Joswig
He	האם השחקן הגברי של לוהנגרין התחתן עם מרגרט יוסוויג
Kn	ಲೋಹೆಂಗ್ರಿನ್ ಅವರ ಪುರುಷ ನಟ ವಿವಾಹವಾದರು ಮಾರ್ಗರೆಟ್ ಜೋಸ್ವಿಗ್
Zh	Lohengrin的男演员嫁给了Margarete Joswig吗

SPARQL Query:

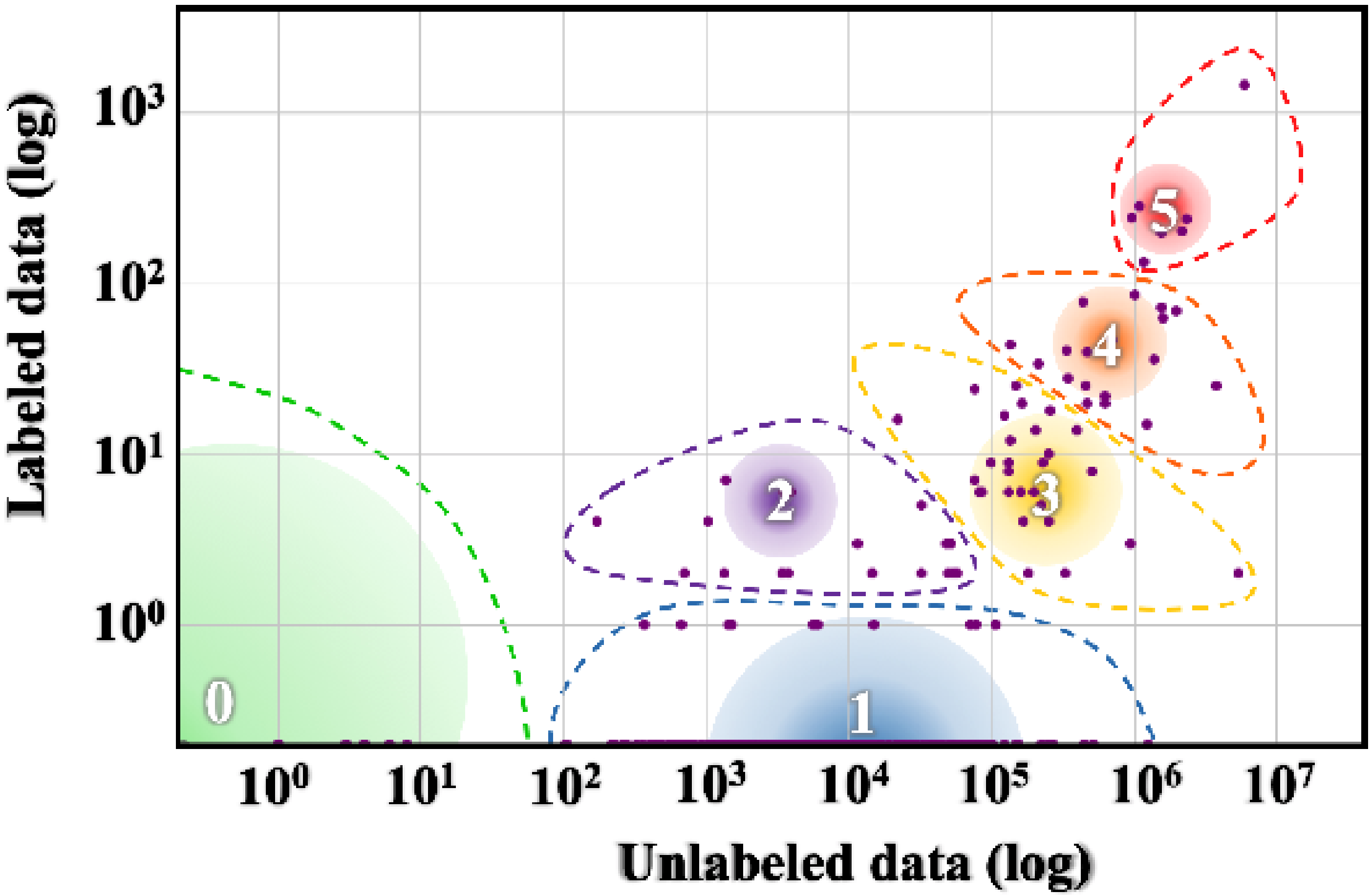
```
ASK WHERE { ?x0 wdt:P453 wd:Q50807639 . ?x0 wdt:P21
wd:Q6581097 . ?x0 wdt:P26 wd:Q1560129 . FILTER (
?x0 != wd:Q1560129 )}
```

[Multilingual Compositional Wikidata Questions](#)
(Cui et al., 2021)

WIKIDATA

Cultural considerations

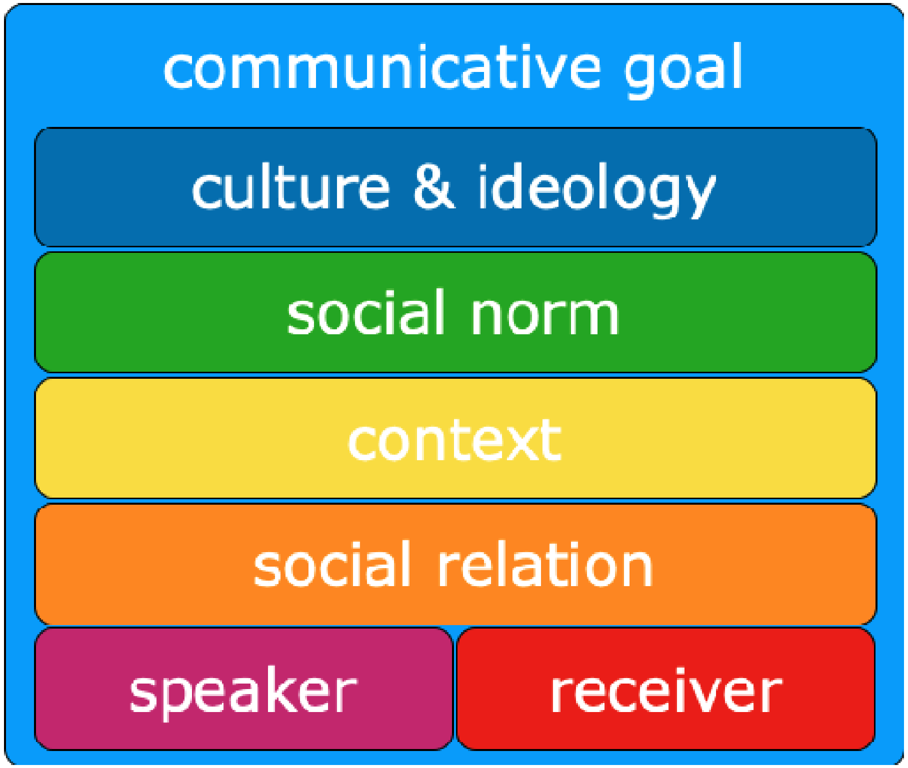
Resource disparity for languages



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

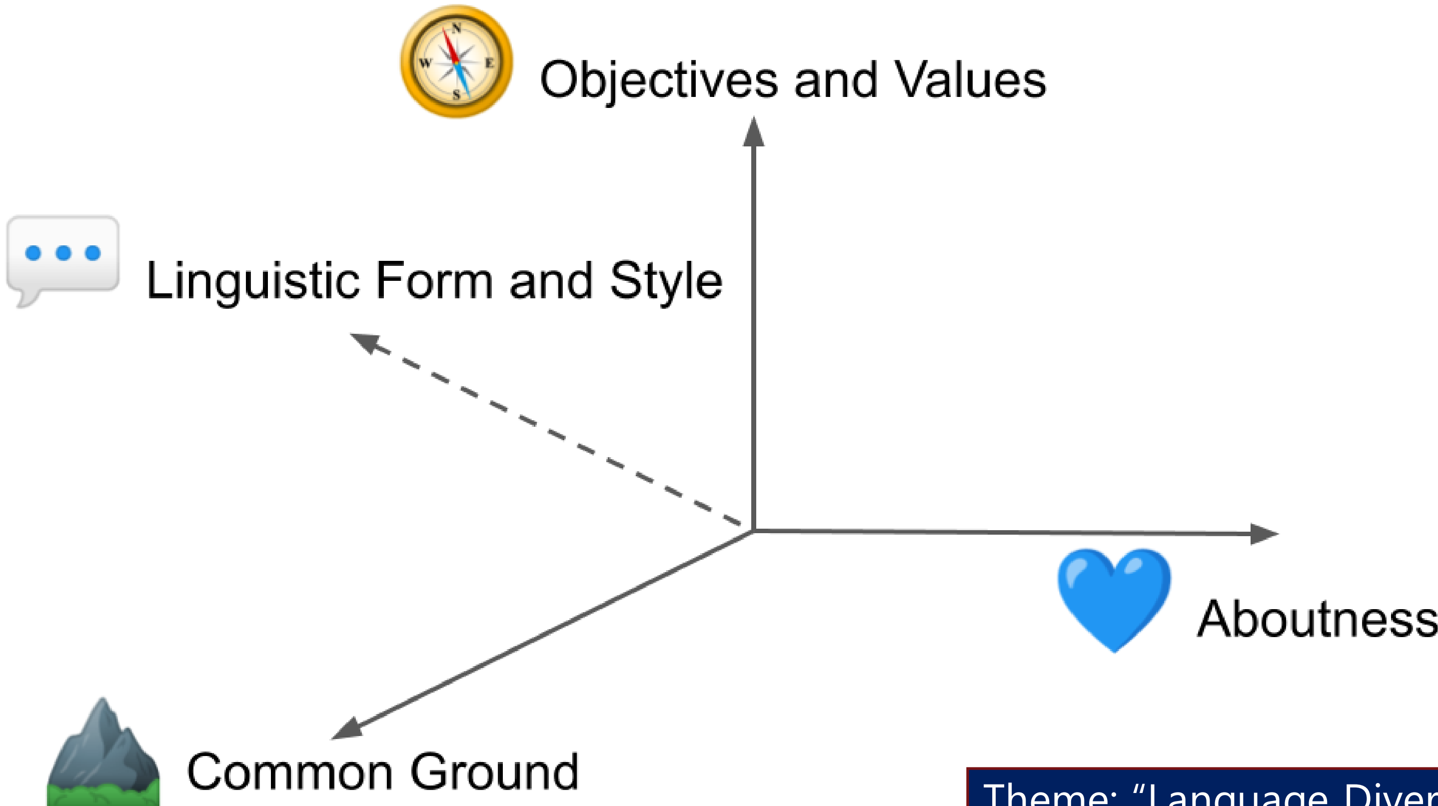
Social factors

NLP is for people (not just languages)



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

Dimensions of culture



Theme: "Language Diversity: from Low-Resource to Endangered Languages"

ACL 2022
22ND – 27TH MAY | 60TH MEETING | DUBLIN

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

Form

- *How* we express ourselves in language
- Morphosyntax, word choice…
- Stylistic aspects of linguistic form:

Directness

Formality

Politeness

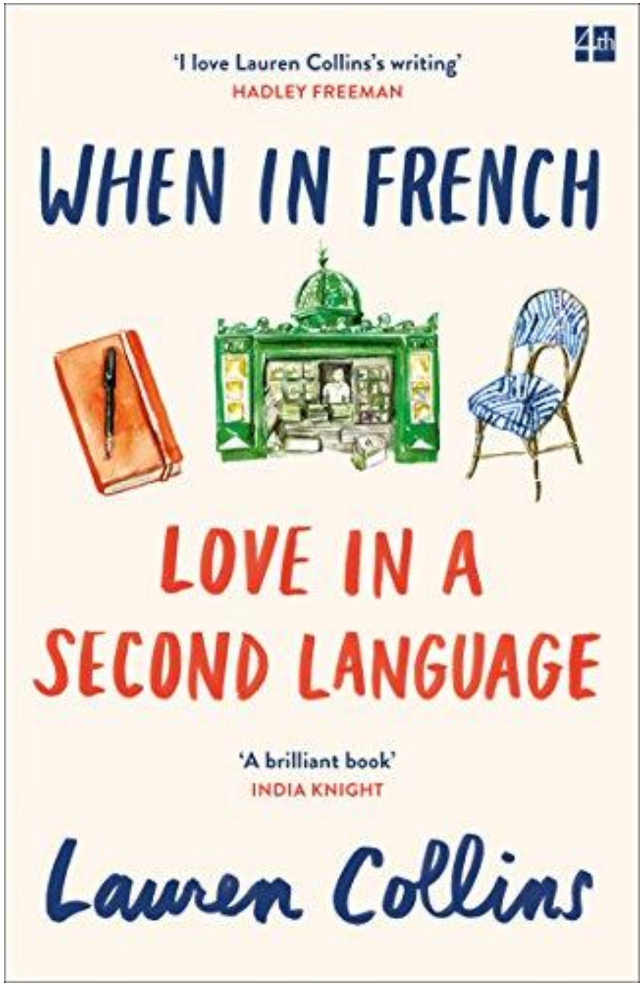
Emotional expression

Common ground

Shared knowledge based on which people reason and communicate

- Conceptualisation
- Commonsense

Common ground



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

Conceptualisation


Tamil
ஜல்லிக்கட்டு
jallikattu

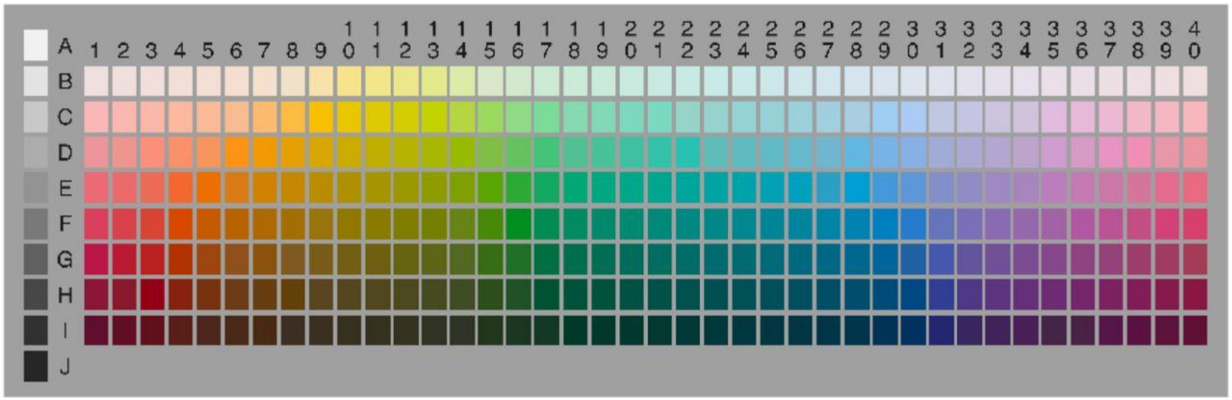
Swahili
leso


Visually Grounded Reasoning across Languages and Cultures

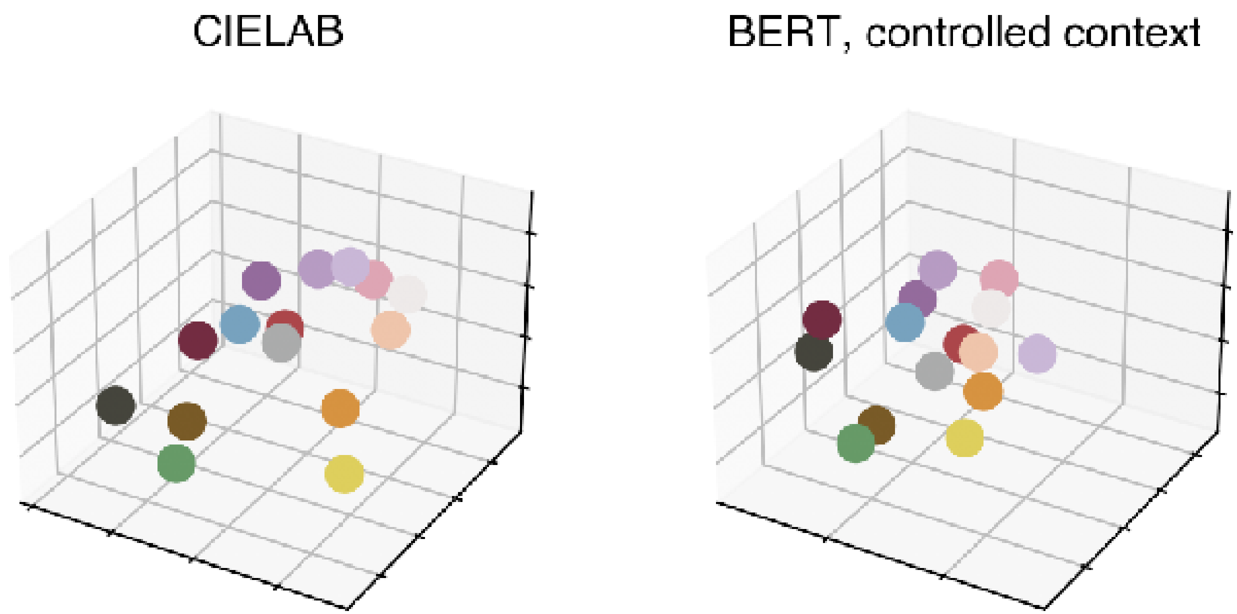
(Liu et al., EMNLP 2021)

Colour

World Colour Survey



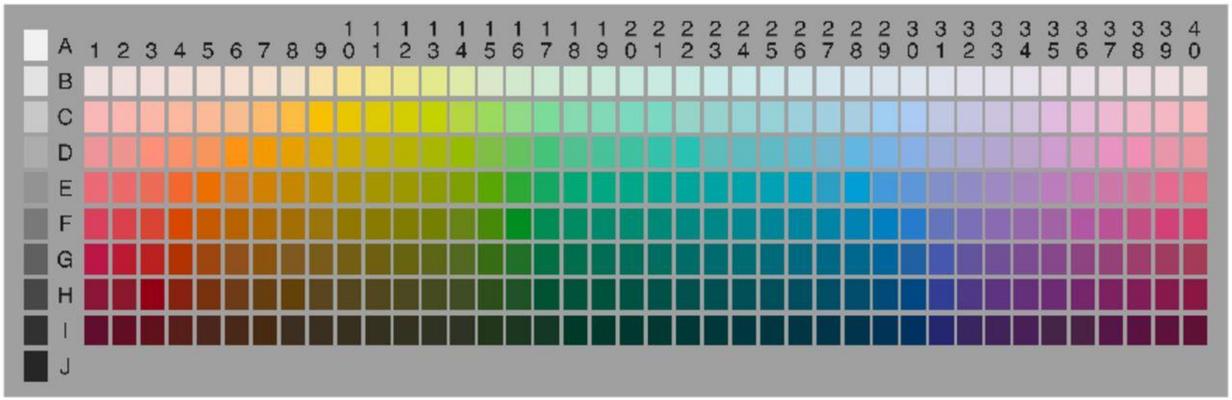
Probing colour



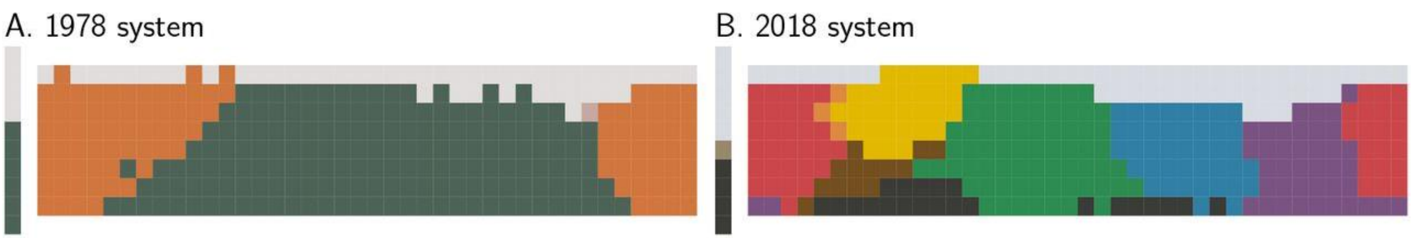
English BERT aligns with English-speaking Americans.
(What about others?)

[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

"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)

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)

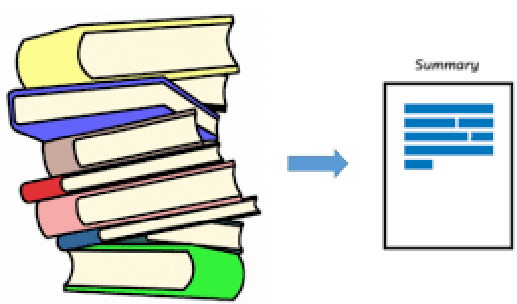
Commonsense

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

Aboutness

- What content do people *care about*?

- Related to topic/domain



Visual concepts

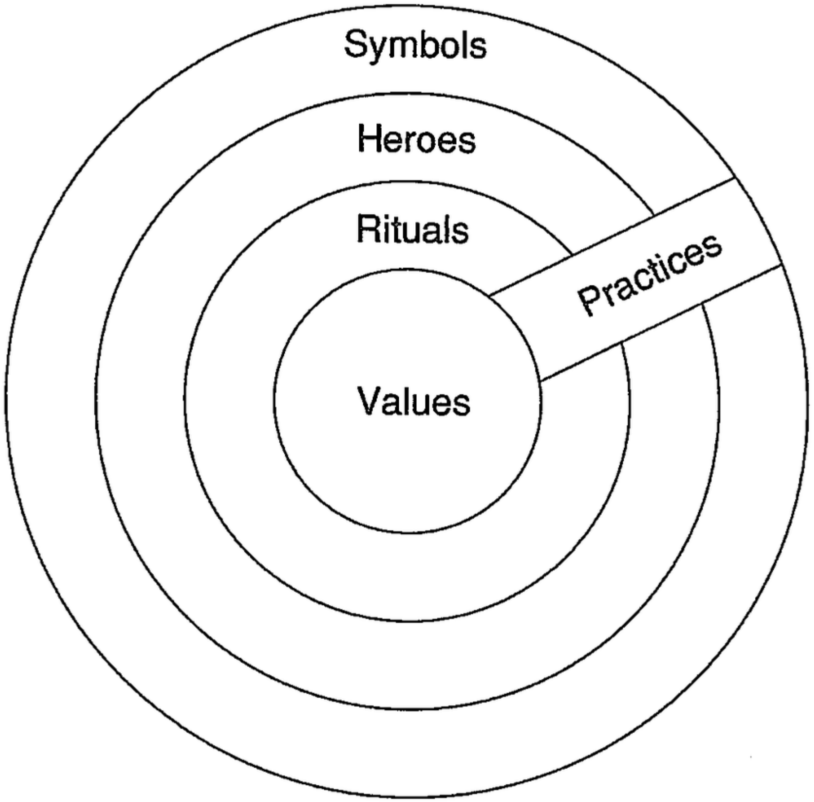
Beer reviews

News generation

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
- Common meta-objectives in *NLP research culture*

Accuracy

Fairness

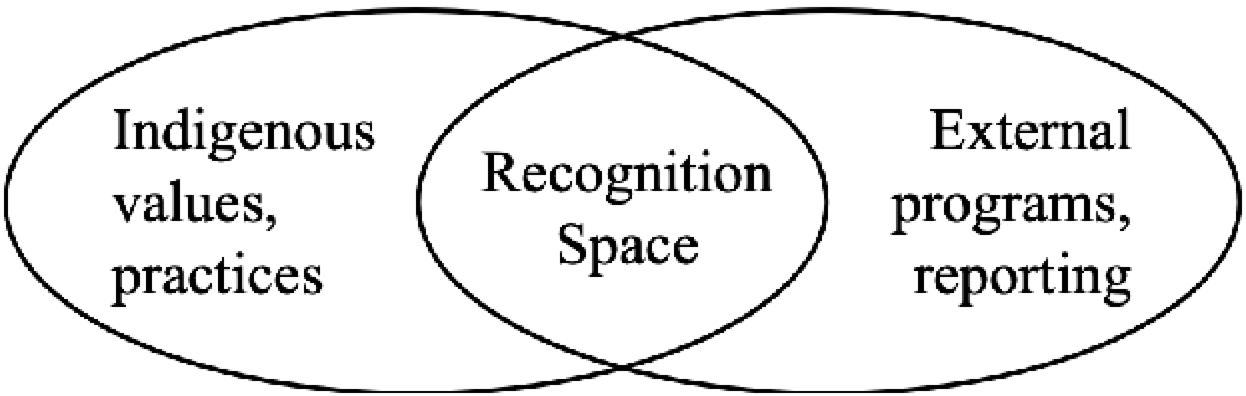
Robustness

Interpretability

Conflicting objectives?

- Researchers
- Practitioners
- End-users
- Regulators

Language technology for all (potential) users



Decolonising Speech and Language Technology
(Bird, COLING 2020)

Strategies

DATA

MODELS

TASKS

Data

- Selection
- Annotation
- Projection

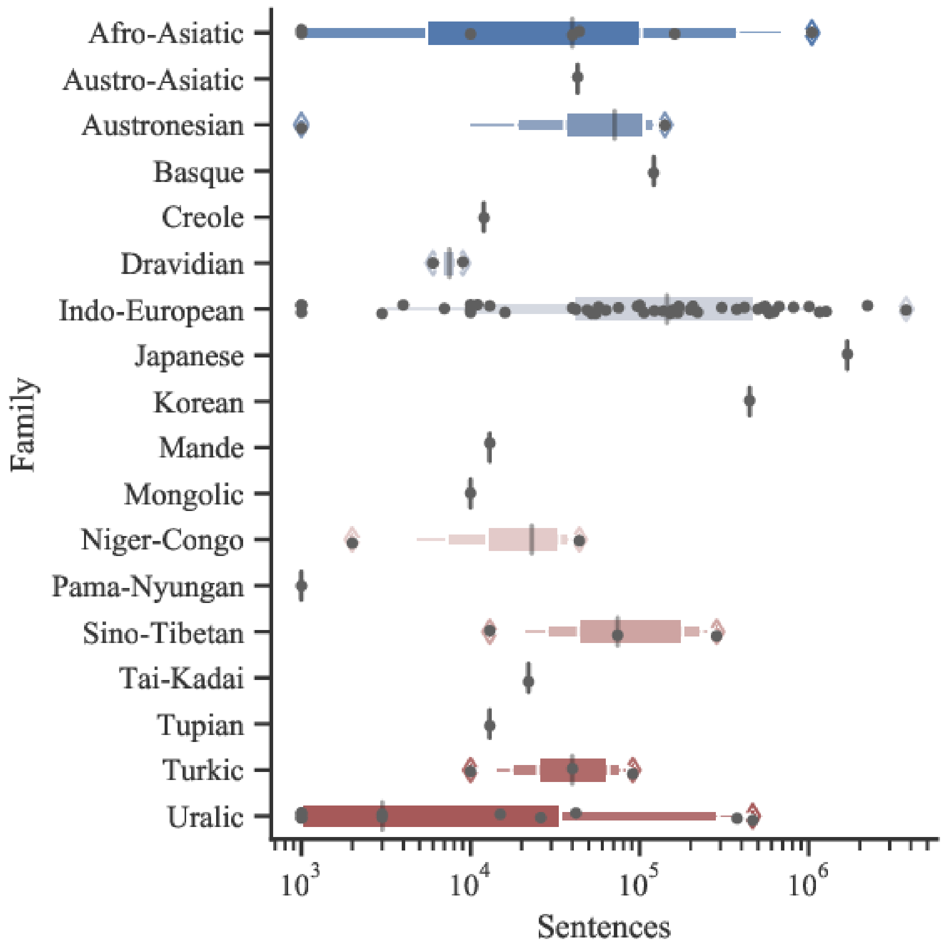
Models

TRAINING

TRANSFER

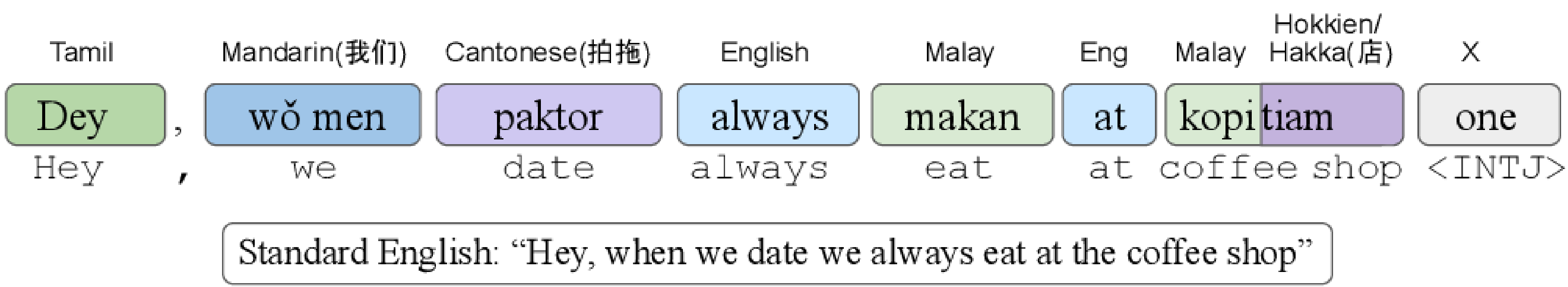
PRE-TRAINED LANGUAGE MODELS

Robust optimisation



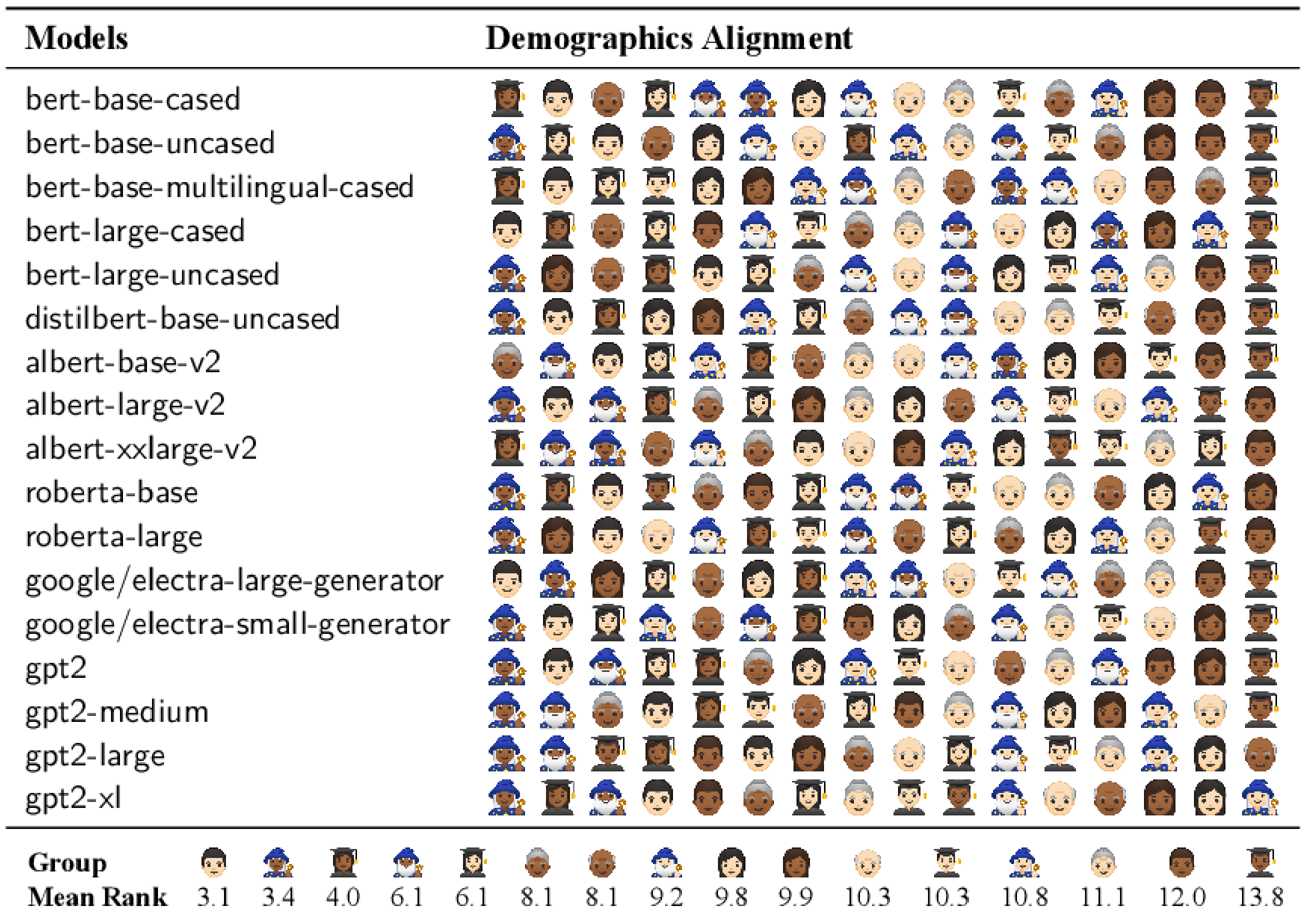
[Minimax and Neyman–Pearson Meta-Learning for Outlier Languages](#)
(Ponti et al., Findings 2021)

Robust optimisation does not always work



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

Social bias in language models



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

Knowledge bias in language models

	Query	Two most frequent predictions
en	X was created in MASK.	[Japan (170), Italy (56), . . .]
de	X wurde in MASK erstellt.	[Deutschland (217), Japan (70), . . .]
it	X è stato creato in MASK.	[Italia (167), Giappone (92), . . .]
nl	X is gemaakt in MASK.	[Nederland (172), Italië (50), . . .]

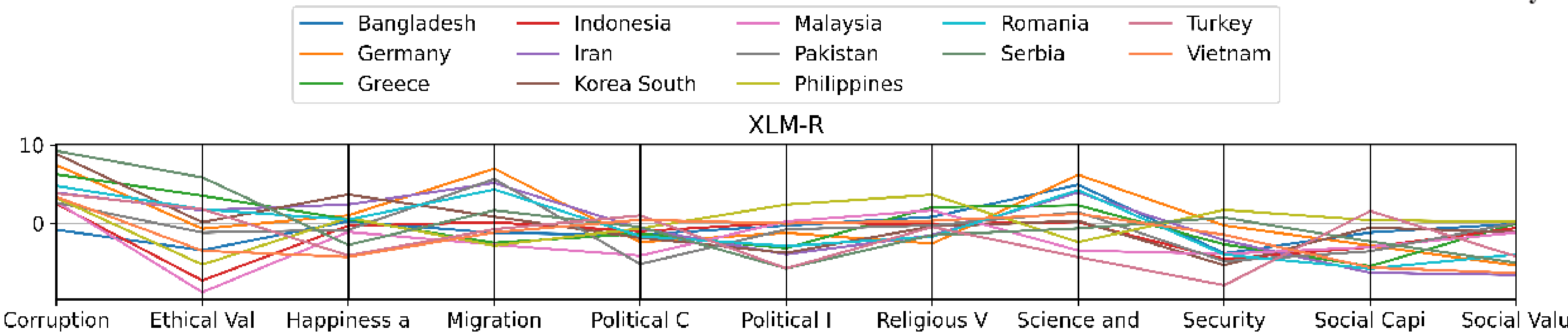
Multilingual LAMA: Investigating Knowledge in Multilingual Pretrained Language Models (Kassner et al., EACL 2021)

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

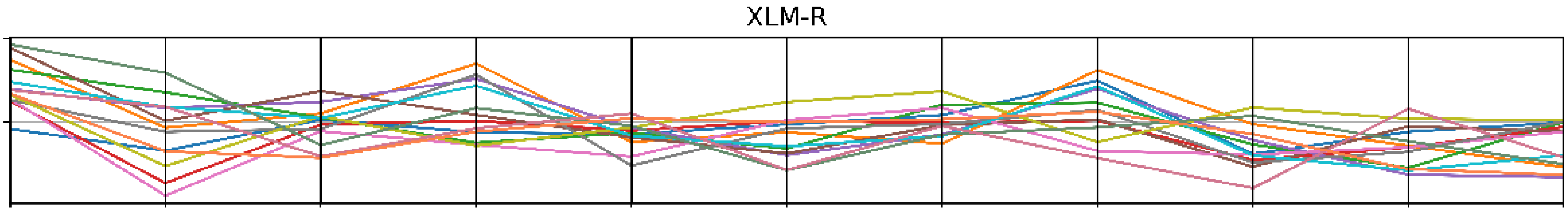
World Values Survey



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

Implicit context

For each of the following, indicate how important it is in your life.



Open question: who is the implied speaker?

Cross-cultural translation

Bridging between cultures *as a task*

"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)

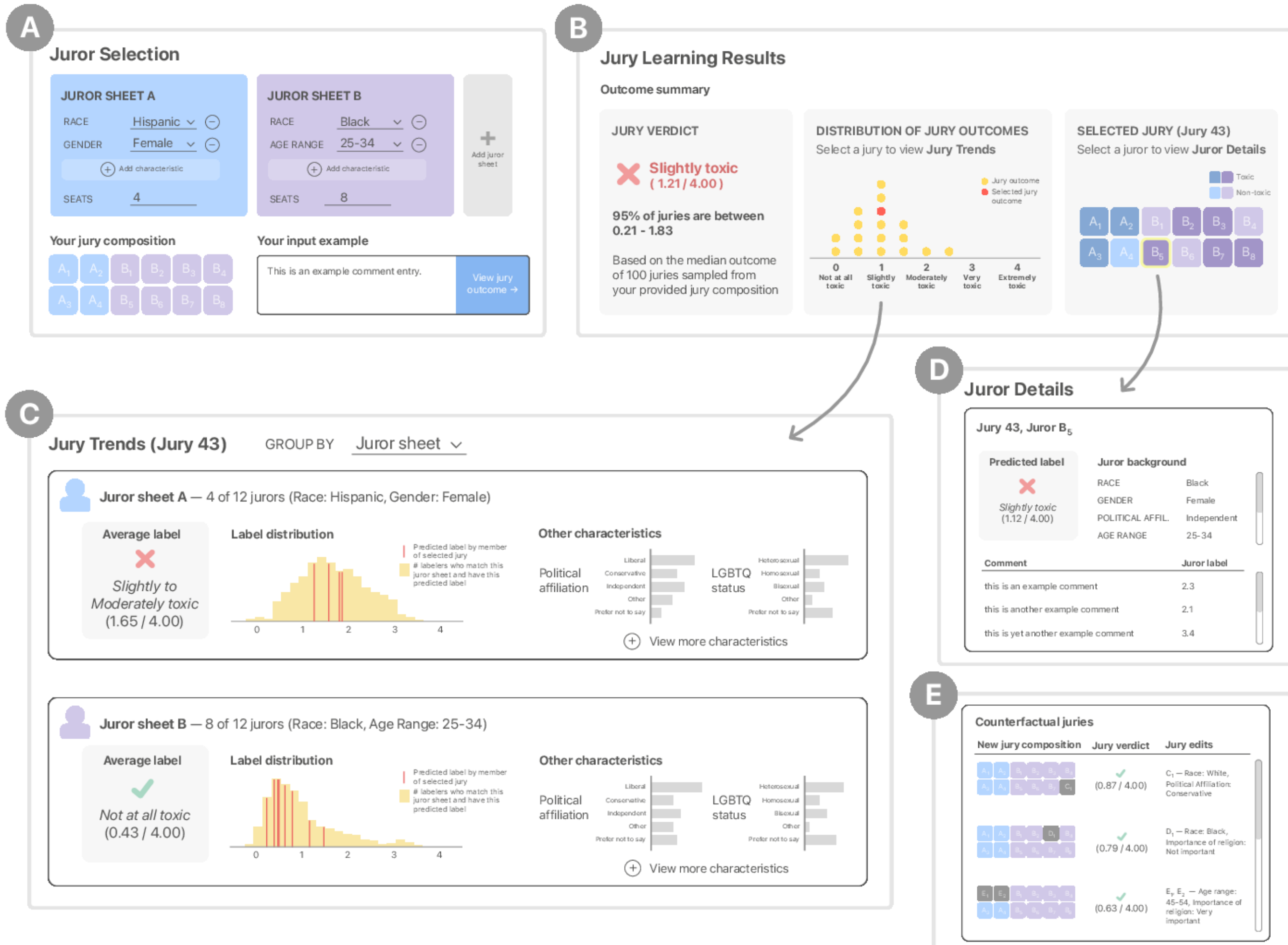
- Style transfer
- Entity adaptation
- Explanation by analogy

Levels of granularity

- Linguistic variation within a "language"
- Also applies to cultures

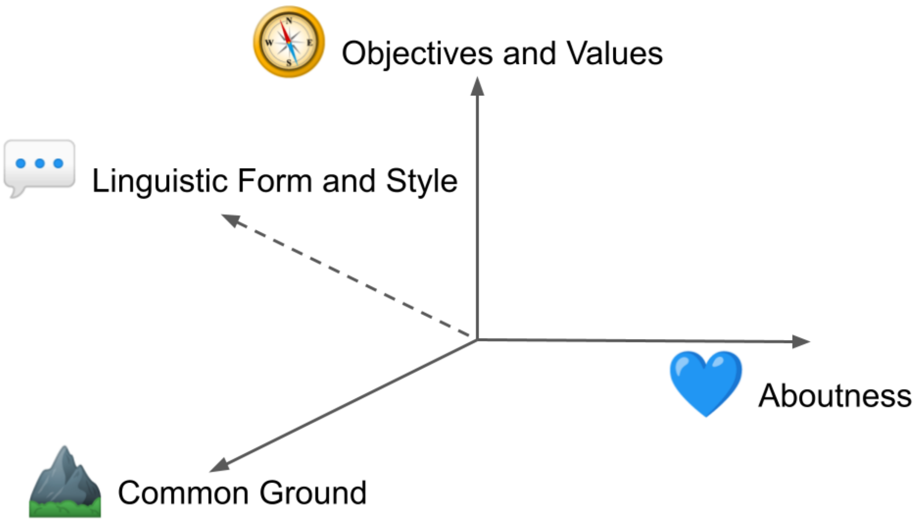
Idiolect	Sociolect, dialect	Standardised language	Language, language family
Individual	Social group or region	Country	

Multi-granularity adaptation



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

Interim summary



NLP is for people (not just languages)

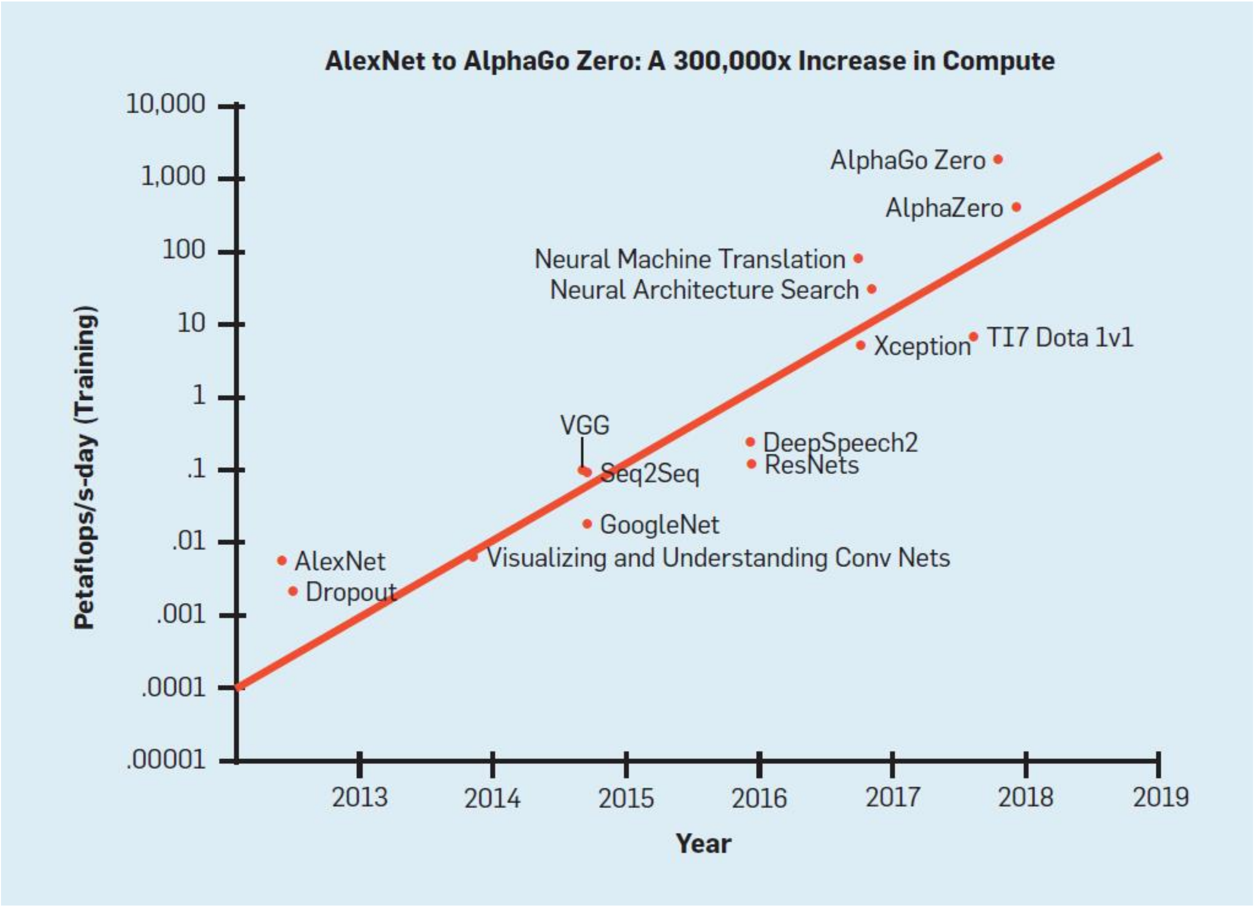
Culture is multidimensional

Objectives are conflicting

Generalisation-representation trade-off

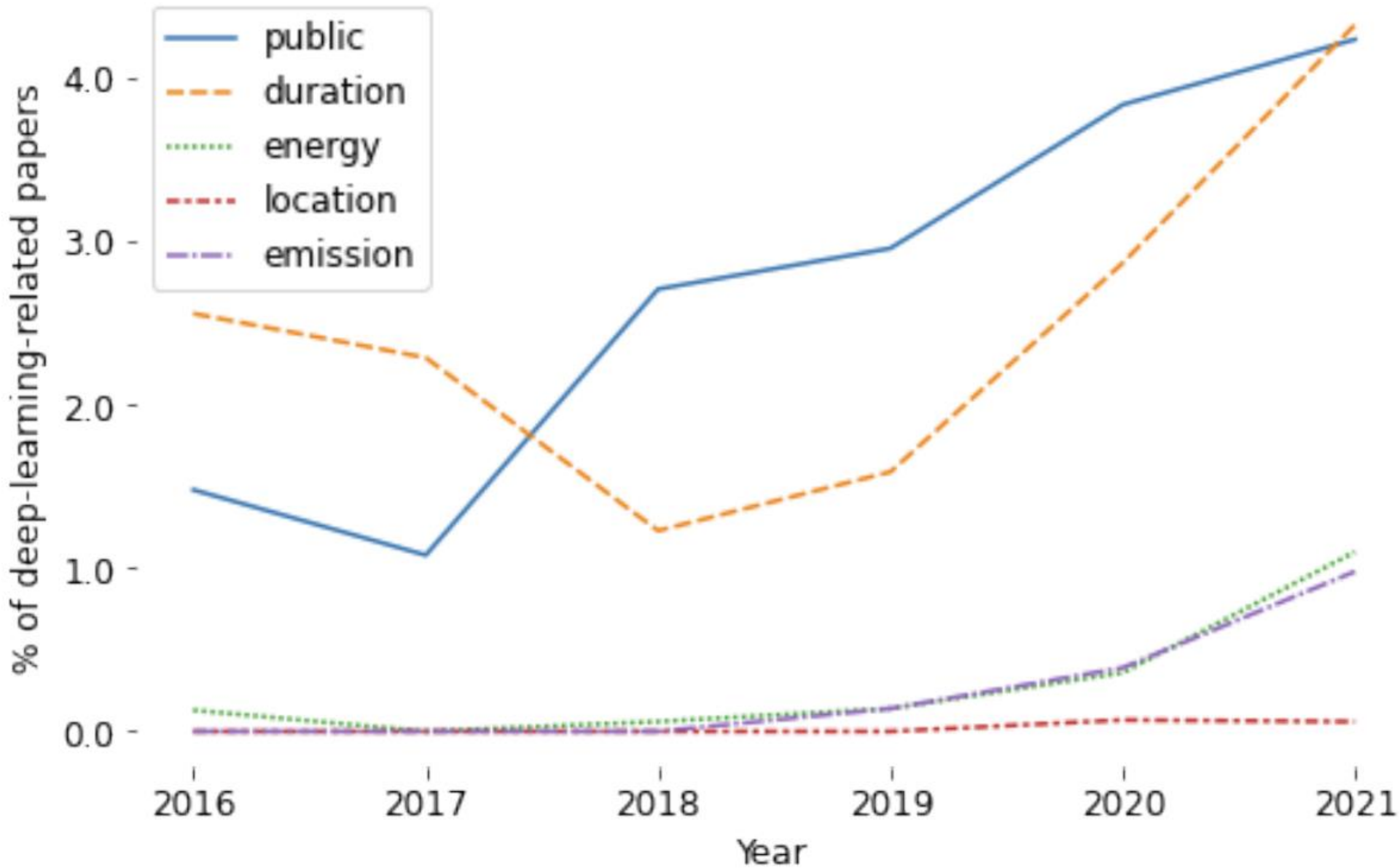
Environmental considerations

Green NLP



[Green AI]
(Schwartz et al., Communications of the ACM 2020)

Climate awareness



[Towards Climate Awareness in NLP Research]()

(under review)

Climate awareness

climatebert/**distilroberta-base-climate-f**

Fill-Mask | PyTorch | Transformers | en | arxiv:2110.12010

Model card | Files and versions | Settings

Climate performance model card

Minimum card

1. Is the resulting model publicly available?	Yes
2. How much time does the training of the final model take?	8 hours
3. How much time did all experiments take (incl. hyperparameter search)?	288 hours
4. What was the energy consumption (GPU/CPU)?	0.7 kW
5. At which geo location were the computations performed?	Germany

Extended card

6. What was the energy mix at the geo location?	470 gCO2eq/kWh
7. How much CO2eq was emitted to train the final model?	2.63 kg
8. How much CO2eq was emitted for all experiments?	94.75 kg
9. What is the average CO2eq emission for the inference of one sample?	0.62 mg

Towards Climate Awareness in NLP Research
(under review)

Climate awareness

Principles:

- Relevance
- Completeness
- Consistency
- Transparency
- Accuracy

\+ Positive impact

Towards Climate Awareness in NLP Research
(under review)

The collective action paradox

- Individual action cannot save the planet
- How can we make a difference?

Positive impact

Greenwashing detection

- Financial reporting increasingly discuss climate
- Claims are false or not specific

Claim:	97% consensus on human-caused global warming has been disproven.
Evidence: REFUTE	In a 2019 CBS poll, 64% of the US population said that climate change is a ""crisis"" or a ""serious problem"", with 44% saying human activity was a significant contributor.
Claim:	The melting Greenland ice sheet is already a major contributor to rising sea level and if it was eventually lost entirely, the oceans would rise by six metres around the world, flooding many of the world's largest cities.
Evidence : SUPPORT	The Greenland ice sheet occupies about 82% of the surface of Greenland, and if melted would cause sea levels to rise by 7.2 metres.

ClimateBert: A Pretrained Language Model for Climate-Related Text
(Webersinke et al., 2021)

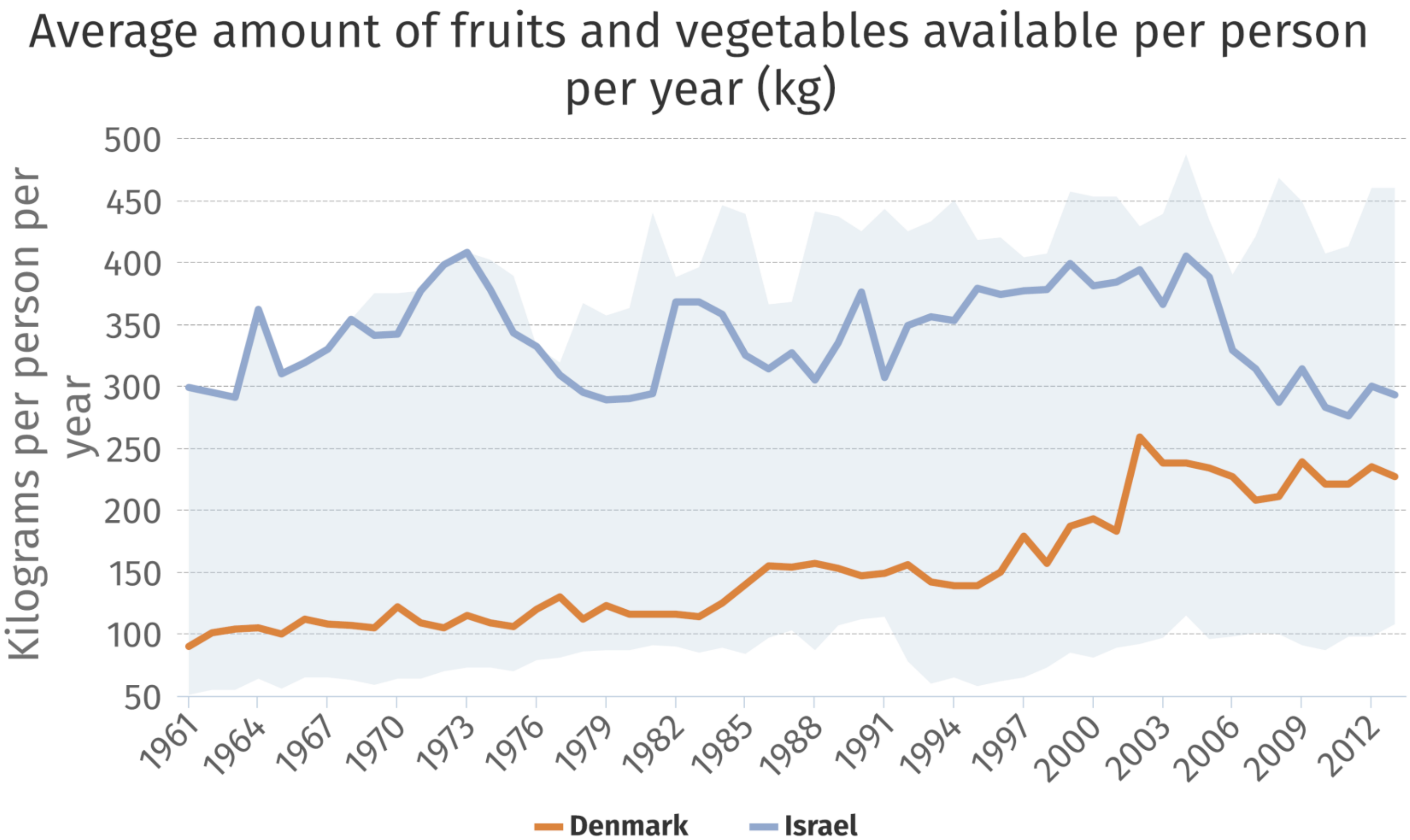
Cheap talk and cherry-picking: What ClimateBert has to say on corporate climate risk disclosures
(Bingler et al., Finance Research Letters 2022)

Government funding for plant-based foods

- Denmark: climate agreement for food and agriculture
- "Central element in the green transition"
- Consumer demand is an essential component

https://fm.dk/media/25215/aftale-om-groen-omstilling-af-dansk-landbrug.pdf

Historical Developments



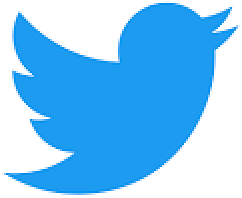
https://gateway.euro.who.int/en/indicators/hfa_446-3240-average-amount-of-fruits-and-vegetables-available-per-person-per-year-kg/visualizations/#id=19474

Historical precedence

Argument mining for green nutrition

- Dietary preferences are hard to change:
  - Perceptions of taste
  - Knowledge and skills
- Existing interventions:
  - Guidelines and policies
  - Everyday habits and convenience

- Work in progress: dataset and models for mining arguments on Twitter



- Potential applications:
  - Surveying public attitudes and exposing misinformation
  - Generation of convincing arguments and nudging

Opportunity for integration

- Cultural differences in consumer preferences

Source: https://www.seges.dk/innovation-og-udvikling/landbrug_og_klima/white_paper

Summary

Climate awareness will make efficient NLP mainstream

Efficient is not enough – net impact should be positive

Compliance and consumer

Thank you

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