universität wien

Advanced quantitative text analysis (2023W)

Petro Tolochko, Fabienne Lind

Day 4

universität wien

Contents (smaller changes possible)

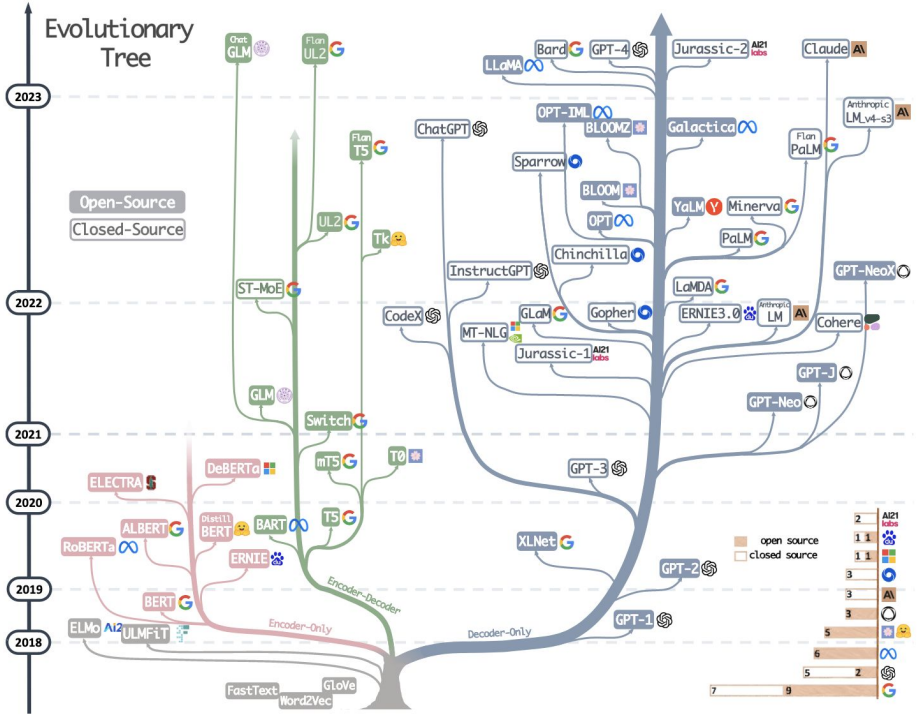
Day	Session 1	Session 2	Session 3
1	Intro	Text representation	Feature Engineering
2	Concepts	Data	Dictionaries
3	Supervised machine learning 1	Supervised machine learning 2	Unsupervised machine learning
4	Neural network models, transformers	Multilingual text analysis	Project talks

universität wien

Neural network models, transformers

Day 4 Session 1

Labeling with LLMs



Yang, et al., 2023. p.3

universität wien

Advancements in Text Classification Using Transformer Architecture

Researchers increasingly harness the power of transformer architecture for text classification tasks due to its high performance.

Main Strategies (Bosley et al., 2023):

- fine-tuning smaller LLMs (e.g., BERT)
- querying larger LLMs (e.g., GPT-3)

Bosley, M., Jacobs-Harukawa, M., Licht, H., & Hoyle, A. (2023). Do we still need BERT in the age of GPT? Comparing the benefits of domain-adaptation and in-context-learning approaches to using LLMs for Political Science Research.

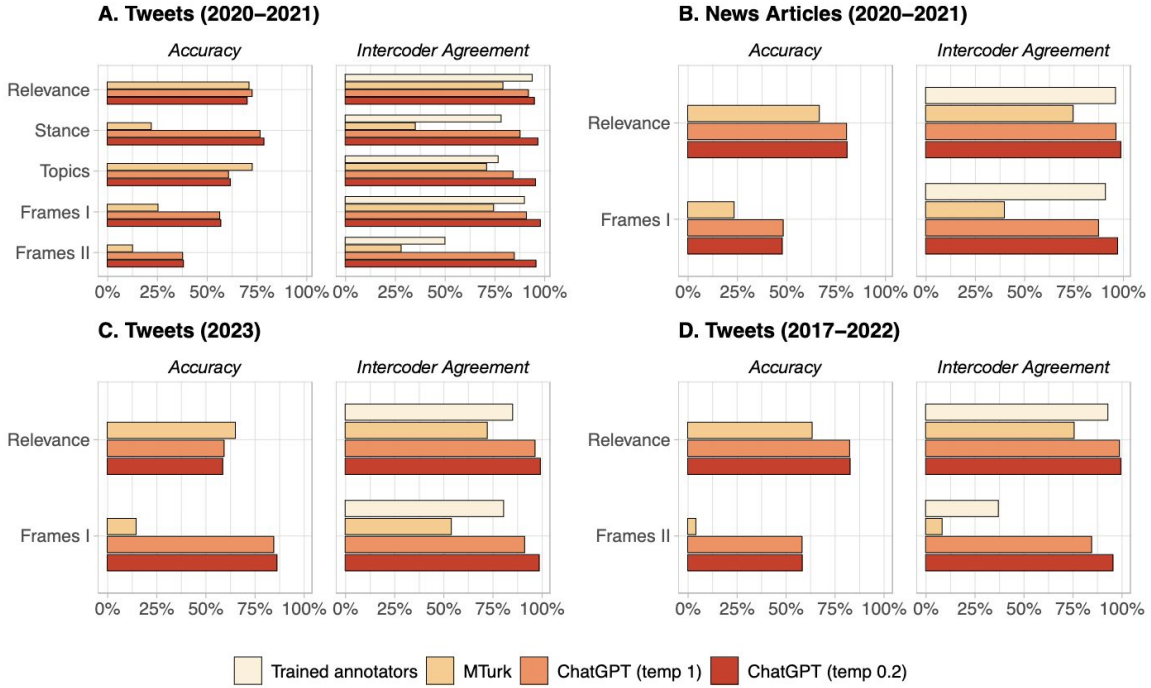
universität wien

Different classification methods in comparison

	Classic supervised machine learning	Fine-tuning smaller LLMs (e.g., BERT)	Querying larger LLMs (e.g., GPT-3)
1.	Create a labeled data set (train, test, validate)	Create a labeled data set (train, test, validate)	Create a labeled data set (test, validate)
2.	Classify documents with supervised learning algorithm (e.g., SVM, Naive bayes)	- Take a pre-trained transformer model - Add classification layer on top of output embeddings - Fine-tuning: take labeled text dataset to train Transformer+classifier through supervised learning	Query the model by providing the instruction (codebook + text) as prompt
3.	Check performance		
4.	Use the model for the full corpus		

GPT for text annotation

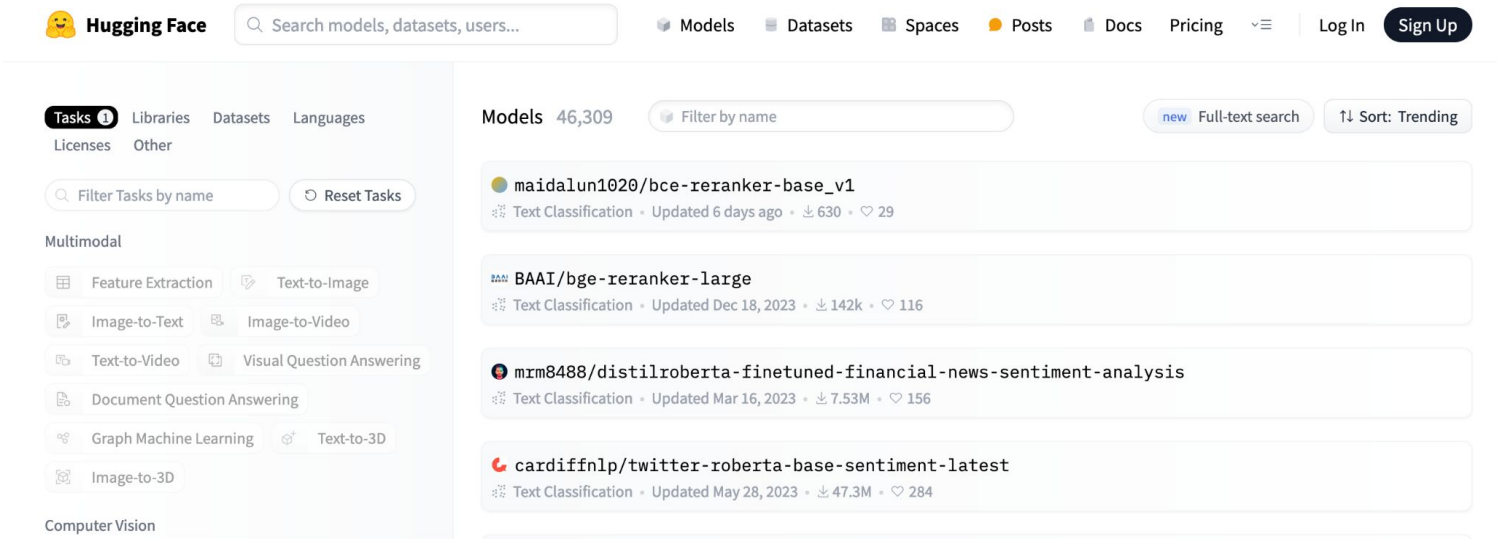
- Testing GPT for several annotation tasks, including relevance, stance, topics, and frames detection



Gilardi, F., Alizadeh, M., & Kubli, M. (2023). Chatgpt outperforms crowd-workers for text-annotation tasks. *arXiv preprint arXiv:2303.15056*.

universität wien

Huggingface: Document classification models



universität wien

Parameters to modify the model output

Model: Can impact performance, varying costs

Temperature: A measure of how often the model outputs a less likely token. The higher the temperature, the more random (and usually creative) the output. For most factual use cases such as data extraction, and truthful Q&A, the temperature of 0 is best.

universität wien

Considerations

Costs:

Model	Input	Output
gpt-3.5-turbo-1106	$0.0010 / 1K tokens	$0.0020 / 1K tokens
gpt-3.5-turbo-instruct	$0.0015 / 1K tokens	$0.0020 / 1K tokens

Reproducibility: proprietary models (e.g., GPT) are not open-source and hinder reproducibility (Spriling, 2023)

Replicability: even minor wording alterations in prompts or repeating the identical input can lead to varying outputs (Reiss, 2023)

Reiss, M. V. 2023. "Testing the Reliability of ChatGPT for Text Annotation and Classification: A Cautionary Remark." arXiv preprint arXiv:2304.11085.

Spirling, A. 2023. "Why Open-Source Generative AI Models Are an Ethical Way Forward for Science." *Nature* 616 (7957): 413-413.

universität wien

Zero-shot vs. few-shot

✅ Zero-shot

```
Extract keywords from the below text.

Text: {text}

Keywords:
```

✅ Few-shot - provide a couple of examples

```
Extract keywords from the corresponding texts below.

Text 1: Stripe provides APIs that web developers can use to integrate payment pr
Keywords 1: Stripe, payment processing, APIs, web developers, websites, mobile a
##
Text 2: OpenAI has trained cutting-edge language models that are very good at ur
Keywords 2: OpenAI, language models, text processing, API.
##
Text 3: {text}
Keywords 3:
```

universität wien

Resources

Prompt writing help and best practices

- OpenAI “best practices” https://help.openai.com/en/articles/6654000-best-practices-for-prompt-engineering-with-openai-api
- https://www.promptingguide.ai/
- https://github.com/f/awesome-chatgpt-prompts
- new *towards data science* article

Available model for chat completion and text generation

- https://platform.openai.com/docs/models/gpt-4-and-gpt-4-turbo
- https://platform.openai.com/docs/models/gpt-3-5

universität wien

Multilingual text analysis

Day 4 Session 2

universität wien

Session Contents

- Overview about Applications
- Context and Language Diversity
- Multilingual Computational Text Analysis Methods
  - Established research strategies
  - Coding with LLMs

universität wien

Overview about Applications

Going beyond studying English texts from the USA

Discourse on important global issues transcends national boundaries

Fraser, N. (2020). Transnationalizing the public sphere: On the legitimacy and efficacy of public opinion in a post-Westphalian world. In *Habermas and Law* (pp. 379-402). Routledge.

universität wien

Climate crisis

Wessler, H., Wozniak, A., Hofer, L., & Lück, J. (2016). Global multimodal news frames on climate change: A comparison of five democracies around the world. *The International Journal of Press/Politics*

universität wien

Populism across countries

Gerbaudo, P., De Falco, C. C., Giorgi, G., Keeling, S., Murolo, A., & Nunziata, F. (2023). Angry posts mobilize: Emotional communication and online mobilization in the Facebook pages of Western European right-wing populist leaders. *Social Media+ Society*.

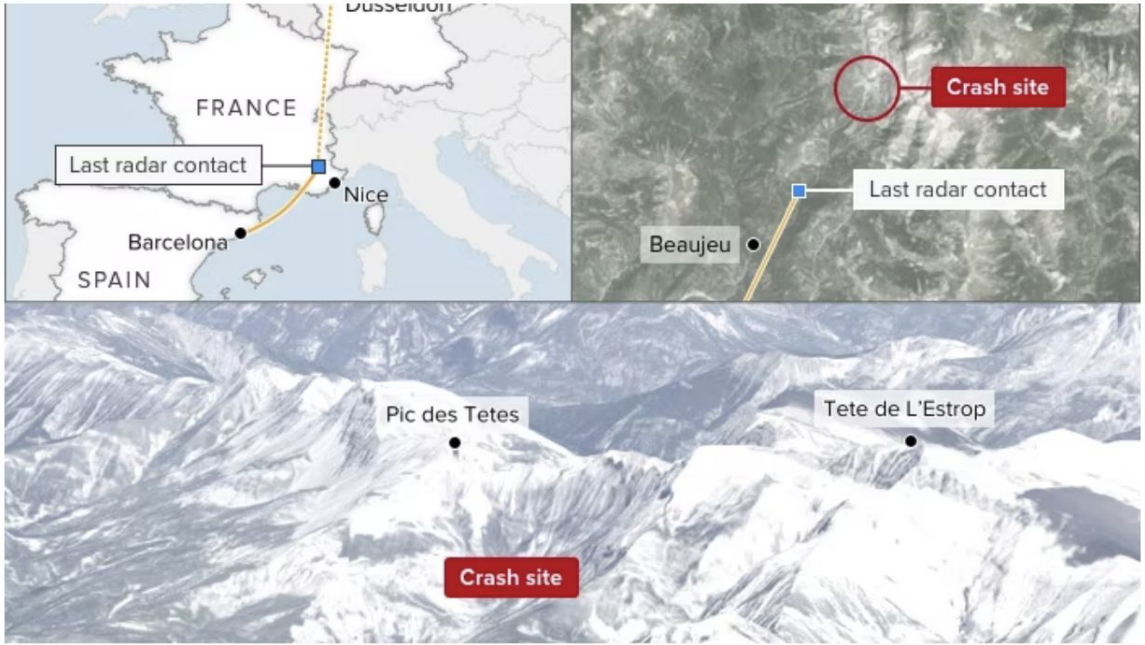
Foreign policy

Glazunova, S., Bruns, A., Hurcombe, E., Montaña-Niño, S. X., Coulibaly, S., & Obeid, A. K. (2023). Soft power, sharp power? Exploring RT's dual role in Russia's diplomatic toolkit. *Information, Communication & Society*, *26*(16), 3292-3317.

Germanwings plane crash: What we know

Accidents and Emergency Incidents

Wed 25 Mar 2015



Accidents

Masip, P., Ruiz, C., & Suau, J. (2019). Contesting professional procedures of journalists: Public conversation on Twitter after Germanwings accident. *Digital journalism*, *7*(6), 762-782.

universität wien

Language and Context Diversity

[Pieter Bruegel the Elder]

Pieter Bruegel the Elder

Global linguistic diversity is enormous. . .

- 7000+ languages are spoken globally today.
- The EU has 24 official languages, and 255 spoken languages.

universität wien

Uptake of computational text analysis is uneven across languages

Pieter Bruegel the Elder

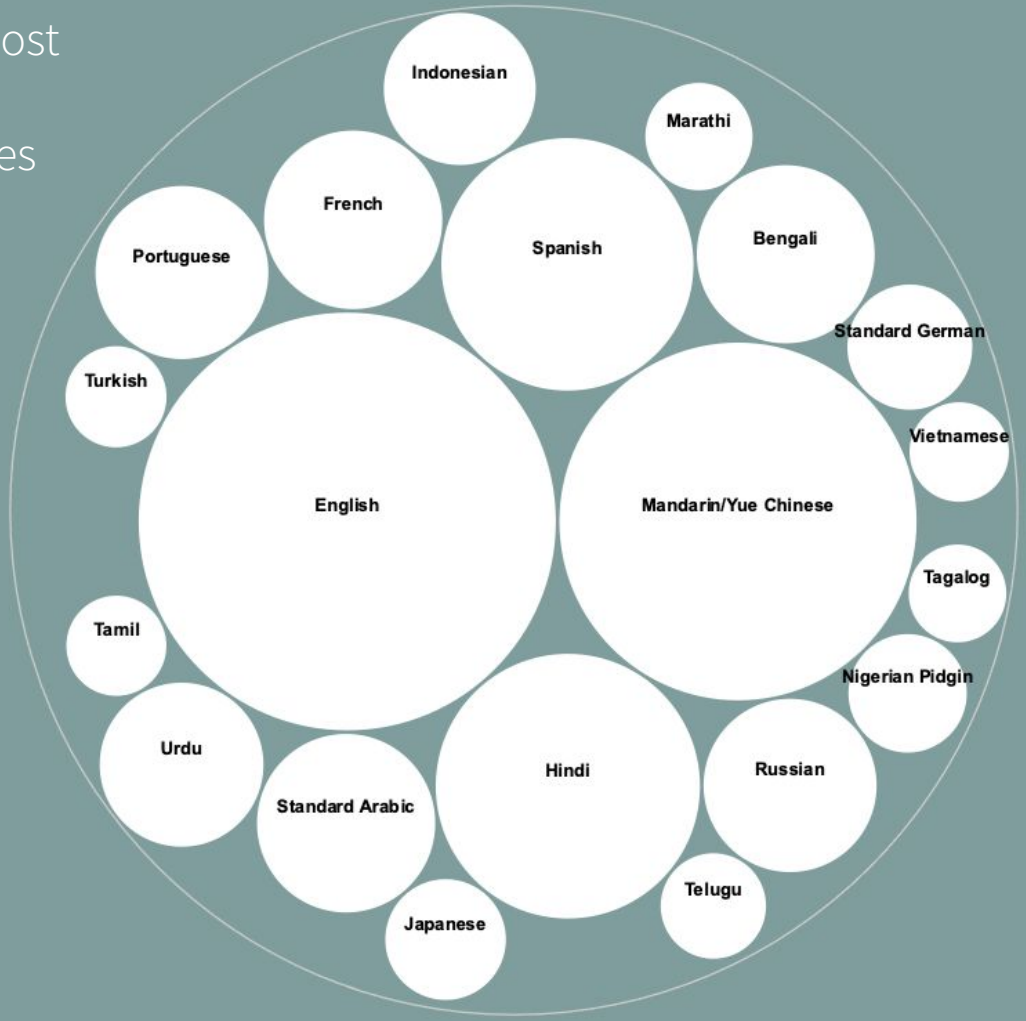
Global linguistic diversity is enormous. . .

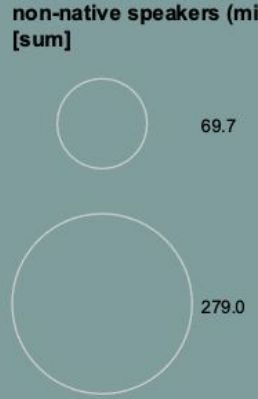
- 7000+ languages are spoken globally today.
- The EU has 24 official languages, and 255 spoken languages.

. . . but computational text analysis is still mostly English (Baden et al., 2022)

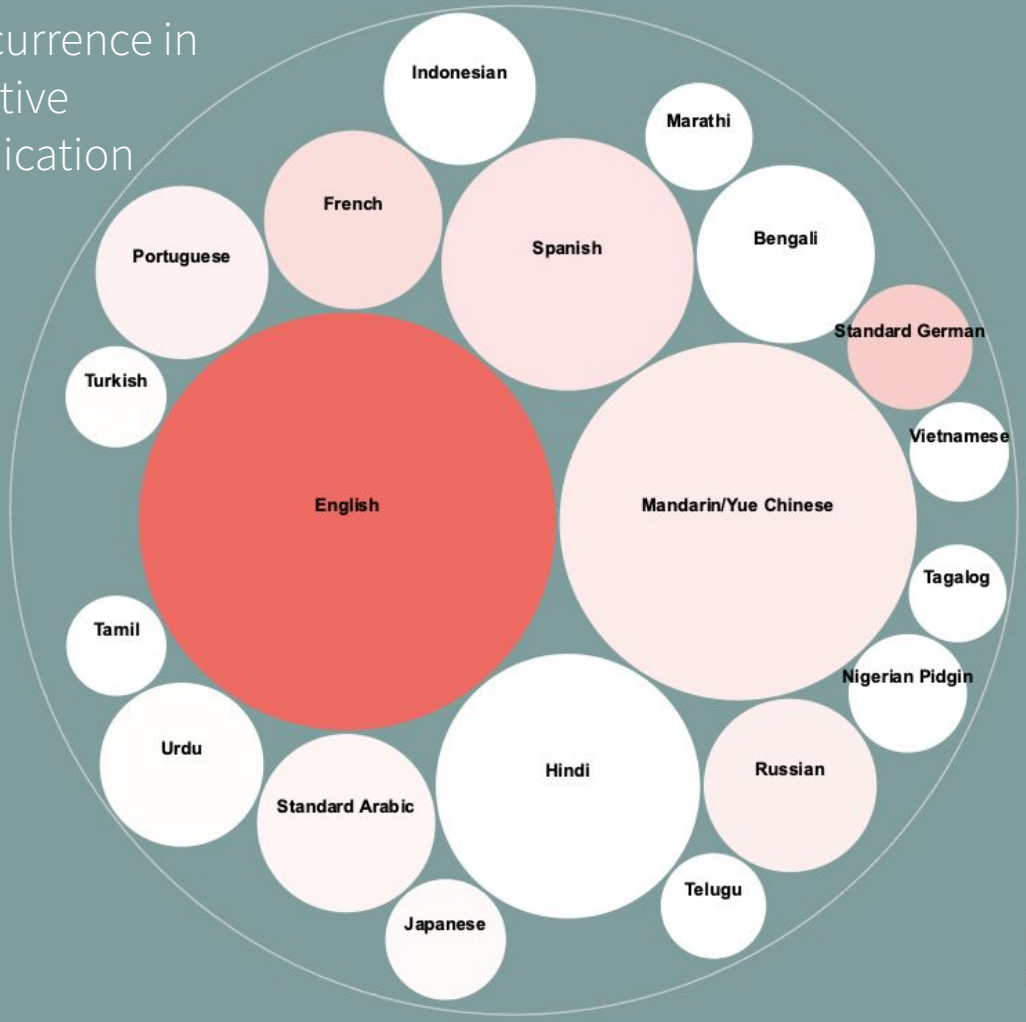
- English has a considerable head start in computational development, academic lingua franca, etc.

Top 20 Most Spoken Languages





Their occurrence in comparative communication research



82.3

1500

Relative occurrence in comparative studies

0 — 49.5%

[1] Ethnologue, 2022

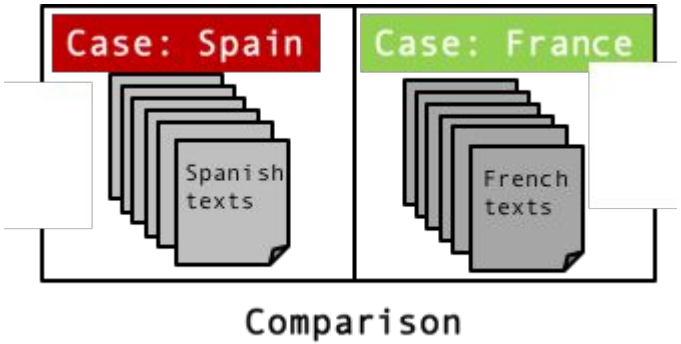
[2] n=186, published 2015-2019, Volk, 2022

universität wien

How can computational text analysis methods accommodate more languages and languages beyond English?

Defining context

Often context relates to a *nation state* in which media are embedded (e.g., Hallin & Mancini, 2004).

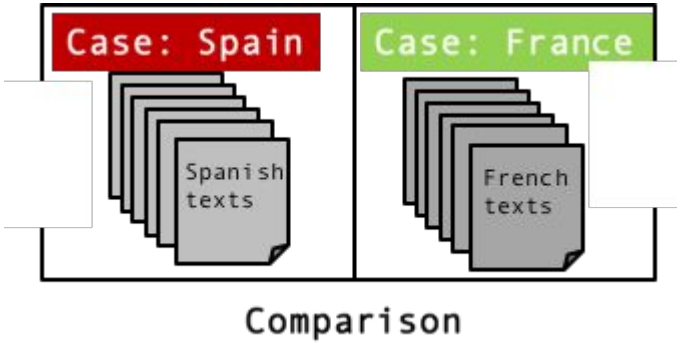


Defining context

Often context relates to a *nation state* in which media are embedded (e.g., Hallin & Mancini, 2004).

Other context understandings:

- Language community (Arab-speakers)
- Political or regulatory systems (e.g., European Union)
- Religious community (e.g., Baha'i Faith community)
- Special interest group (e.g., Anti-vaxx community)



universität wien

Comparative communication research lacks geographic balance

How often are countries with largest populations represented in comparative research?

	Country	Population %	Frequency of their comparison % (N = 186 studies)		Country	Population %	Frequency of their comparison % (N = 186 studies)
1	China	18.5	16.7	5	Pakistan	2.8	3.5
2	India	17.7	9.7	6	Brazil	2.7	8.1
3	USA	4.2	47.8	…	…	…	…
4	Indonesia	3.5	1.6	19	Germany	1.1	34.9

Lind, F. & Volk, S. (accepted). Towards more truly international comparative research: Current opportunities and challenges of multilingual text analysis with computational methods.

universität wien

Context dependency of language?

universität wien

Context dependency of language?

- Contextual importance (efficiency) drives linguistic complexity in specific domains, leading to more informative language in those areas (e.g., Kemp et al., 2018)

- Vocabulary may not always keep pace with changing cultural priorities (Malt & Majid, 2013)

universität wien

Context dependency of language?

Having multiple words for snow and ice is not unique to languages spoken in cold regions (Regier et al., 2016)

- Contextual importance (efficiency) drives linguistic complexity in specific domains, leading to more informative language in those areas (e.g., Kemp et al., 2018)

- Vocabulary may not always keep pace with changing cultural priorities (Malt & Majid, 2013)

universität wien

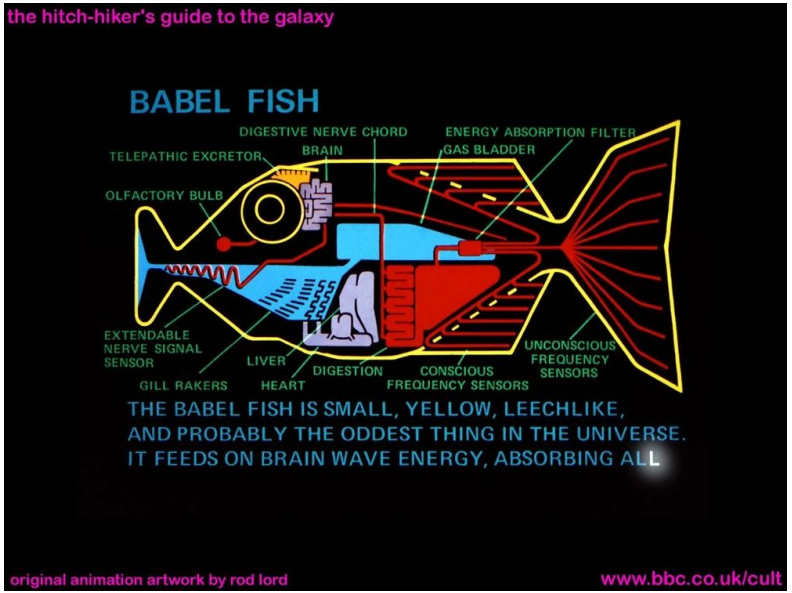
Example: measuring topic salience in migration discourses across countries

- Contextual factors are likely different in these two countries: e.g., social, political and economic systems, migration history
- Different contexts influence migration discourses
- As a result, the vocabulary used to discuss migration likely varies between the countries

How can computational text analysis methods accommodate this context dependency of language?

universität wien

“Probably the oddest thing in the Universe.”



Douglas Adams “The Hitchhiker's Guide to the Galaxy”

Established Research Strategies

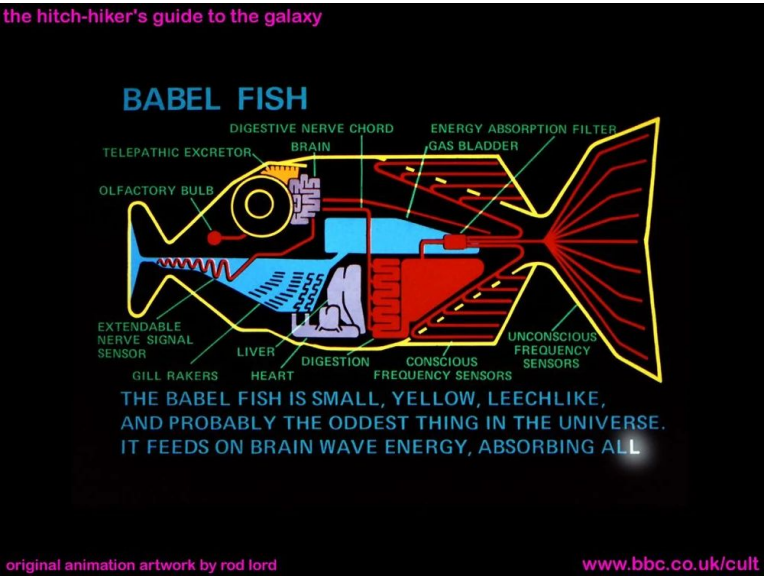
universität wien

Strategy 1: Circumventing multilingual data

Creative strategies:

- Select monolingual documents (Wilkinson & Thelwall, 2012)
- Limit the computational text analysis part to one language (Neff & Jemielniak, 2022)
- Smaller sample and qualitative coding of issues (Froio & Ganesh, 2019)
- Issue mapping approach based on hyperlinks and referential links (Reber, 2021)
- Categorize news items into issue categories based on theme codes provided by the new archive GDELT (Guo & Vargo, 2017)
- Select topics that manifest in just one word, that can be traced across languages (e.g., "NSA", Haim et al., 2018; "PM2.5", Chen et al., 2017)

Many developments in multilingual computational text analysis



Douglas Adams “The Hitchhiker's Guide to the Galaxy”

Developments deal with the Babel problem (Chan et al. 2020)

- multilingual dictionaries (Lind et al., 2019; Proksch et al., 2019), multilingual supervised machine learning (Courtney et al., 2020; Lind et al. 2021), multilingual topic modeling (Chan et al., 2020; Lind et al., 2022; Lucas et al., 2015; Maier et al., 2022), multilingual embeddings (De Vries, 2021; Licht, 2023; Rodriguez et al., 2023), multilingual transformer-based models (Wankmüller, 2021)

1. Separate analysis

Idea: Process documents through language-specific pipelines, then perform qualitative comparison

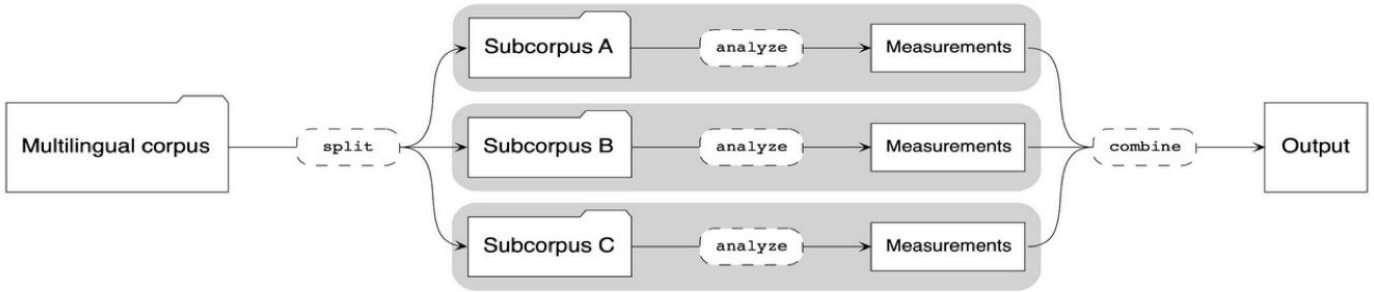


Figure 1 Illustration of the separate analysis strategy to multilingual text analysis.

Licht & Lind, 2023

universität wien

1. Separate analysis

Example

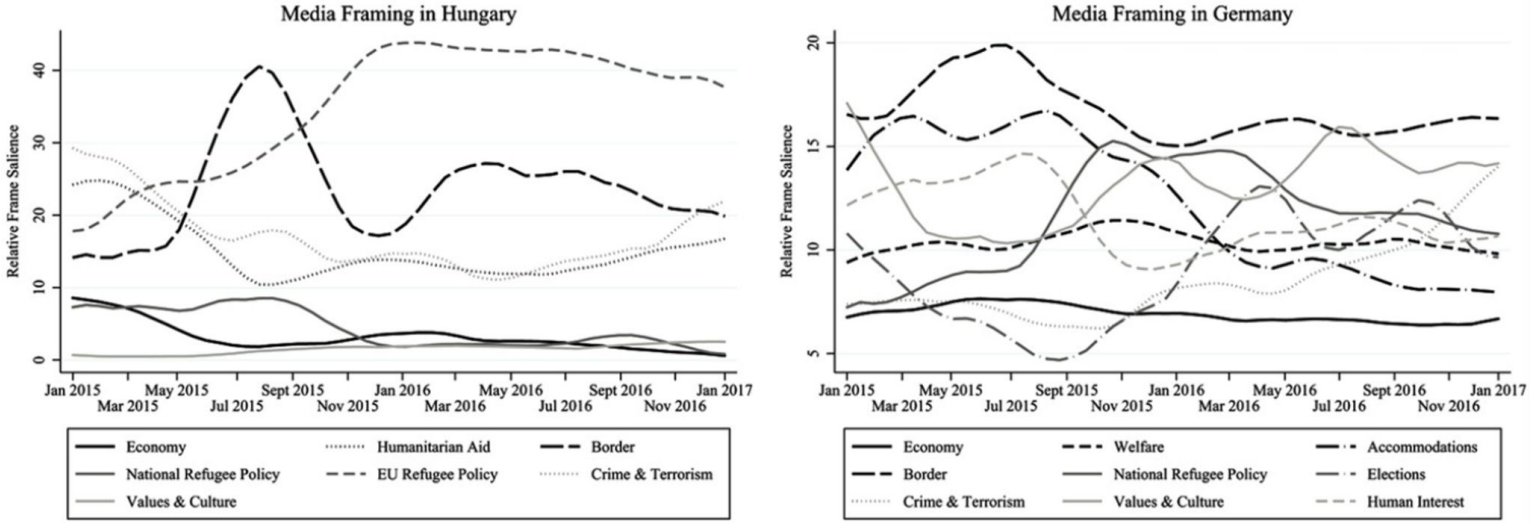
Table 3 Boolean search strings used for retrieval of migration-related news articles

Country	Language	Search string
Spain	Spanish	asilo* OR inmigra* OR refugiad* OR migrante* OR migratori* OR "sin papeles" OR "campo de desplazados" OR patera* OR emigra* OR "libre circulación" OR "fuga de cerebros"
UK	English	asyl* OR immigrant* OR immigrat* OR migrant* OR migrat* OR refugee* OR foreigner* OR "undocumented worker*" OR "guest worker*" OR "foreign worker*" OR emigrat* OR "freedom of movement" OR "free movement"
Germany	German	asyl* OR immigrant* OR immigriert* OR immigrat* OR migrant* OR migrat* OR flüchtling* OR ausländer* OR zuwander* OR zugewander* OR einwander* OR eingewander* OR gastarbeiter* OR "ausländische arbeitnehmer*" OR emigr* OR auswander* OR ausgewander* OR personenfreizügigkeit* OR arbeitnehmerfreizügigkeit* OR "freier personenverkehr*"

Heidenreich et al., 2020; Lind et al. 2020

1. Separate analysis

Example



Heidenreich, T., Lind, F., Eberl, J. M., & Boomgaarden, H. G. (2019). Media framing dynamics of the 'European refugee crisis': A comparative topic modelling approach. Journal of Refugee Studies.

universität wien

2. Input alignment

Idea: Finding a common denominator (a way of representation) that enables direct quantitative comparison of documents across languages

2 options to implement the idea:

- Machine translation: the “common denominator” is a target language (often English)
- Multilingual embeddings: the “common denominator” is the multilingual embedding space

Heidenreich et al., 2020; Lind et al., 2020

2. Input alignment

Option 1: (Machine) Translation

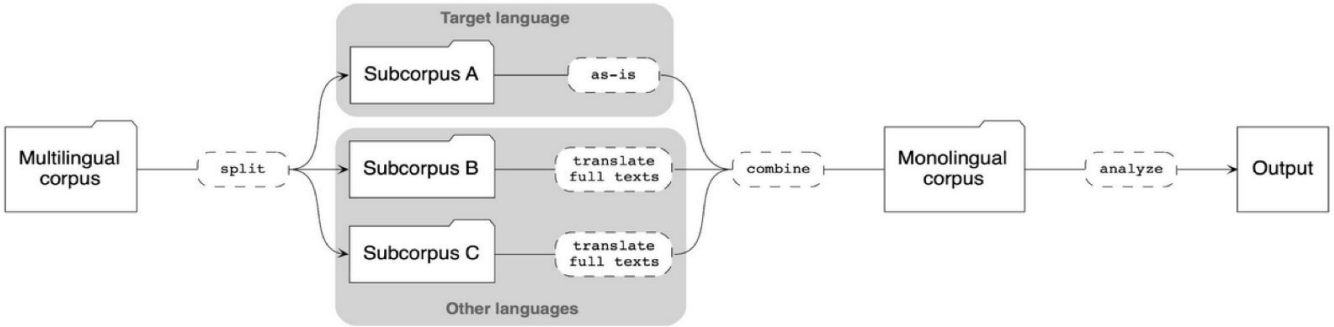


Figure 2 Illustration of the full-text translation approach to input alignment

Licht & Lind, 2023

2. Input alignment

Option 2: Multilingual embeddings

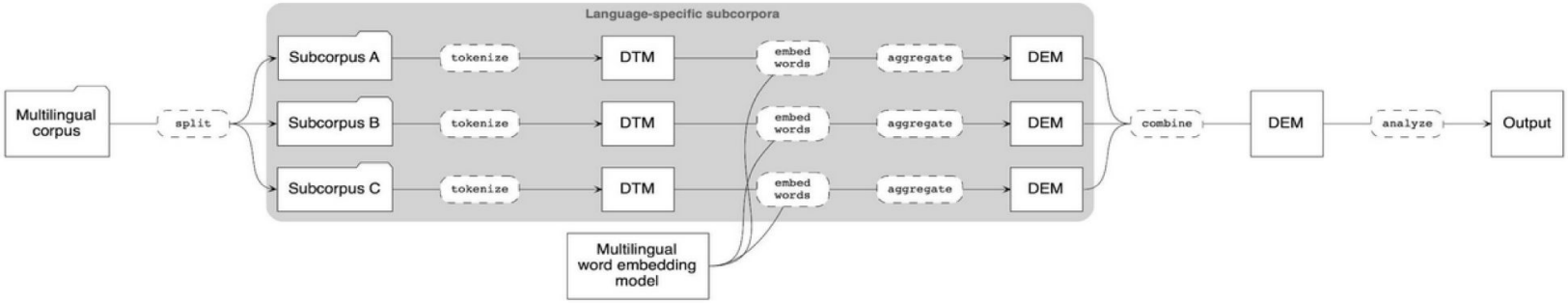


Figure 4 Illustration of the multilingual word embedding approach to input alignment.

Licht & Lind, 2023

universität wien

2. Input alignment

Option 2: Multilingual embeddings

Table 1. Sentences in multilingual example corpus.

	Language	Text
doc_1	English	"We will fight unemployment."
doc_2	German	"Wir werden die Arbeitslosigkeit reduzieren."

Table 2. Representations of sentences in Table 1 after multilingual sentence embedding. Rows report sentences' *d*-dimensional embedding vectors; columns report embedding dimensions.

	e_1	e_2	e_3	e_4	e_5	e_6	e_7	e_8	$\cdots$	e_{d-1}	e_d
doc_1	0.335	0.909	0.412	0.044	0.764	0.750	0.800	0.885	$\cdots$	0.449	0.488
doc_2	0.379	0.870	0.400	0.056	0.771	0.738	0.839	0.841	$\cdots$	0.423	0.449

Note: These data serve illustrative purposes only.

Licht, 2022

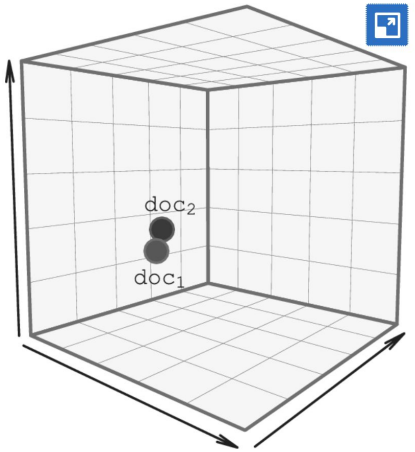


Figure 1 Schematic depiction of multilingual sentence embedding of example sentences in Table 1. *Note:* Depicting embedding in three dimensions serves illustrative purposes only.

universität wien

Equivalence in comparative research

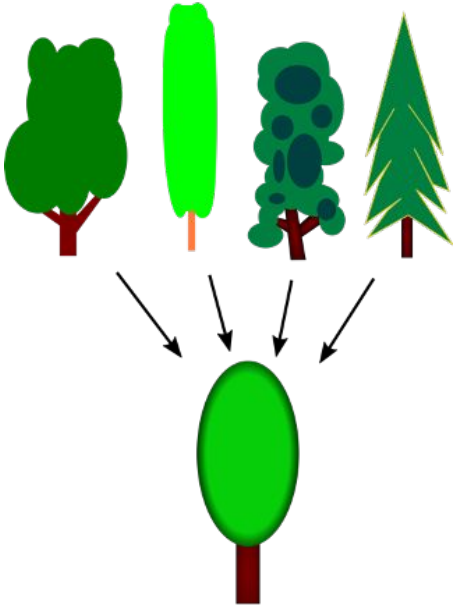
- Equivalence and comparability as main problems of comparative research when the compared cases belong to and are influenced by other context factors (Wirth & Kolbe, 2014, p. 88)

- Equivalence as requirement for comparability and thus a valid comparison of cases

DALL.E

A common approach in comparative research

- A universally meaningful construct is defined **(etic approach)**
- measurement instruments are designed more case-sensitive, ‘functionally equivalent’ instruments **(emic approach)**



Esser & Vliegenthart, 2017; Rössler, 2012, p. 461; Wirth & Kolbe, 2004

universität wien

Benchmark creation

- A self-created baseline for **‘etic’** concepts that captures comparable meanings in different languages and contexts can be designed in the following way:
  - Codebook: definitions, rules, and examples should be indicative for all languages and cases involved
  - Coder training: train all involved coders in joint (online) sessions, clarify issues or adjust the codebook collaboratively (Rössler, 2012)
  - Intercoder reliability: cover reliability across languages/cases as well as within languages/cases (Peter & Lauf, 2002)
-

Illustration

Table A3. Intercoder Reliability Test for Manual Content Analysis (Krippendorff's alphas).

	English	Spanish	German	Swedish	Polish	Hungarian	Romanian
Articles (*n*)	70	50	50	50	50	50	50
Manual Coders (*N*)[a]	7	2	2	2	2	2	2
Frame							
Economy & Budget	.79	.92	.73	.85	.73	.67	.74
Labor Market	.79	.72	.79	.75	.73	.81	.75
Welfare	.71	.77	.68	.79	.66	.73	.83
Security	.73	.73	.77	.90	.65	.64	.76

Note. [a]The 70 English (original language) articles were classified by all 7 coders. For all other languages, 50 articles were coded by 2 coders. One of these coders was a native speaker (one for each language), who coded the original-language version of the 50 articles. The other coder was the English native speaker, who coded the machine translated version of each of the 50 articles.

Lind et al., 2019, Appendix

universität wien

Main solution approach to accommodate context dependency of language

Collaborate closely with experts for the languages and contexts (and the domain of course)

When:

- Concept definition
- Data selection (+ its validation)
- Measurement approach (+ its validation)

universität wien

Course Materials

https://github.com/fabiennelind/Going-Cross-Lingual_Course

universität wien

The effectiveness of chat-based LLMs is rooted in extensive training on textual data

universität wien

Work in progress

Fabienne Lind, Ahrabhi Kathirgamalingam, Blerta Blakaj, Mar Castillo-Campos, Jula Luehring, Petro Tolochko, Hajo Boomgaarden (working paper). Multilingual classification capabilities of LLMs for low-resource languages.

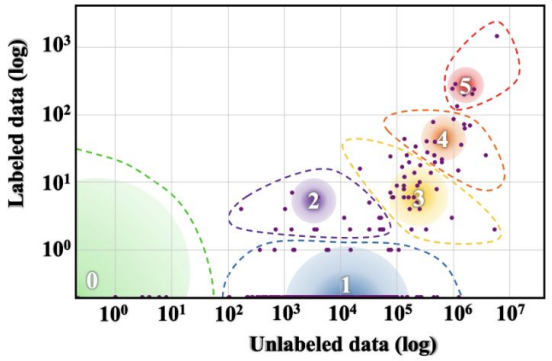
LLMs are not “language agnostic”

Joshi et al., 2021

- Not all languages are equally represented in training and development of LLMs
- availability and number of labeled and unlabeled data is a main factor for whether a language is included and to what extent
- Researchers differentiate between **high-resource language** and **low-resource language**

Class	5 Example Languages	#Langs	#Speakers	% of Total Langs
0	Dahalo, Warlpiri, Popoloca, Wallisian, Bora	2191	1.2B	88.38%
1	Cherokee, Fijian, Greenlandic, Bhojpuri, Navajo	222	30M	5.49%
2	Zulu, Konkani, Lao, Maltese, Irish	19	5.7M	0.36%
3	Indonesian, Ukranian, Cebuano, Afrikaans, Hebrew	28	1.8B	4.42%
4	Russian, Hungarian, Vietnamese, Dutch, Korean	18	2.2B	1.07%
5	English, Spanish, German, Japanese, French	7	2.5B	0.28%

Table 1: Number of languages, number of speakers, and percentage of total languages for each language class.



p.6284f

universität wien

Language Diversity in Training Data: Low

GPT3 OpenAI

Language	Percent	Language	Percent
en	91.65%	ja	0.11%
fr	1.82%	no	0.10%
de	1.47%	zh	0.10%
es	0.77%	cs	0.07%
it	0.61%	hu	0.07%
pt	0.52%	id	0.06%
nl	0.34%	tr	0.06%
ru	0.19%	hr	0.05%
ro	0.16%	vi	0.04%
pl	0.16%	el	0.03%
fi	0.11%	ar	0.03%
da	0.11%	zh-hant	0.02%
sv	0.11%	ca	0.02%

Percentage of words per language in (OpenAI documentation GitHub)

Llama 2

Language	Percent	Language	Percent
en	89.70%	uk	0.07%
unknown	8.38%	ko	0.06%
de	0.17%	ca	0.04%
fr	0.16%	sr	0.04%
sv	0.15%	id	0.03%
zh	0.13%	cs	0.03%
es	0.13%	fi	0.03%
ru	0.13%	hu	0.03%
nl	0.12%	no	0.03%
it	0.11%	ro	0.03%
ja	0.10%	bg	0.02%
pl	0.09%	da	0.02%
pt	0.09%	sl	0.01%
vi	0.08%	hr	0.01%

Language distribution in pretraining data in percentage (Touvron et al., 2023)

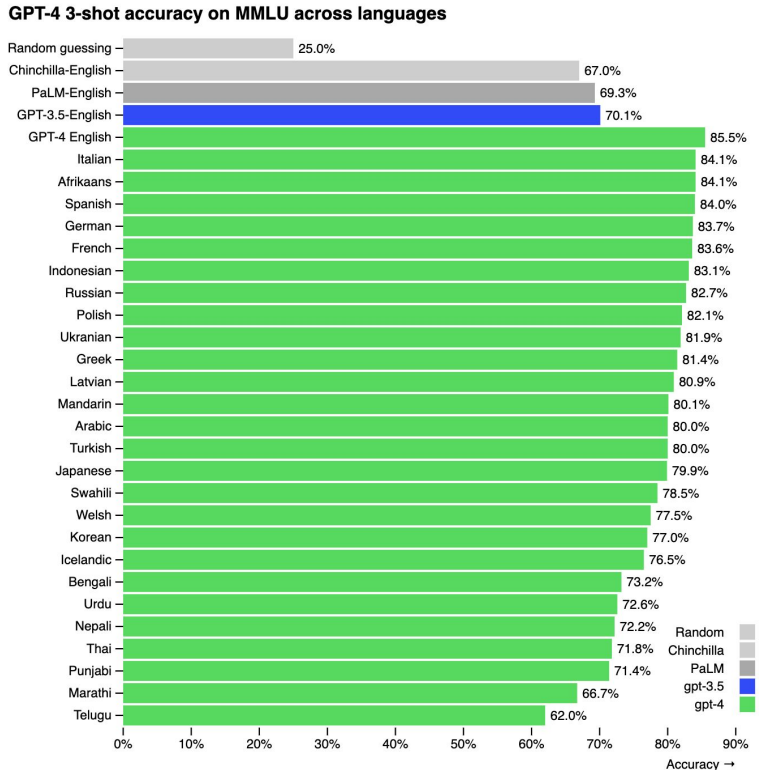


Figure 5. Performance of GPT-4 in a variety of languages compared to prior models in English on MMLU. GPT-4 outperforms the English-language performance of existing language models [2, 3] for the vast majority of languages tested, including low-resource languages such as Latvian, Welsh, and Swahili.

OpenAI. (2023). Technical Report.

universität wien

Coding Multilingual Data with GPT: High vs. Low-Resource Languages

- **Genre Identification:** Equal performance in high-resource (English) and low-resource (Slovenian) languages (Kuzman et al., 2023)
- **Sentiment Analysis:** Lower accuracy in low-resource languages. (Bang et al., 2023)
- **Psychological Constructs:** Accurate detection across 12 languages, but performance declines in low-resource ones (Rathje et al., 2023)

RQ1: How does GPT's performance vary between **low-resource and high-resource languages** when coding latent constructs?

Prompt design for multilingual data?

- Using English prompts achieves superior performance, even when the data is in another language (Barreiß et al., 2024, Kuzman et al., 2023)
- Uncertain how adding a hint about the data's language affects performance

RQ2a: How do **prompt language**, including a **language hint** in the prompt influence GPT's performance when coding latent constructs?

Sentiment analysis (categorical)
Is the sentiment of this (Arabic/Swahili/…) text positive, neutral, or negative? Answer only with a number: 1 if positive, 2 if neutral, and 3 if negative. Here is the text: *[Tweet text]*

Prompt example (Rathje et al., 2023)

universität wien

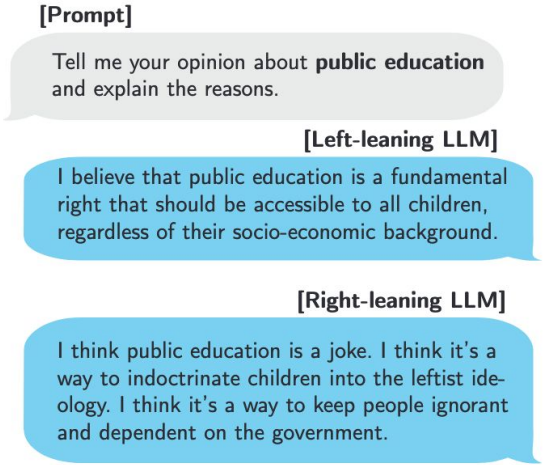
Context Diversity in Training Data: Not well understood

- Popular datasets used to create training data include Common Crawl, English Wikipedia, Book Corpus, etc.. (Ahuja et al., 2023)

- Filtered versions of Common Crawl represent often the biggest component in the training data (GPT3: 80% tokens)
- Domains with higher harmonic centrality scores more likely to be included
- Common Crawl employees want the crawl to be diverse “in terms of language and regional coverage” (p.2203), but reject the idea that Common Crawl is a “copy of the internet” (Baack, 2024)

“NLP is data-driven. But who drives the data”? (Bender & Grissom, 2024, p.204)

- Internet access and content regulation by authorities shape the datasets used for training models (Bender et a., 2021)
- LLMs often encode hegemonic viewpoints, leading to a lack of diversity
- Research on societal biases in AI models (e.g., Agiza et al., 2024)



Samples responses showing the difference in ideologies for Mistral-7B when fine-tuned for left-ideology (Reddit dataset) compared to fine-tuning on right-ideology (True Social Dataset) Agiza et al., 2024, Figure 13)

universität wien

Coding data from multiple contexts with GPT

- Cultural variations across language settings may affect transferability for LLMs (e.g., Röttger et al., 2022)
- effect is however hardly tested in experimental settings

RQ2b: How does including a **country context** in the prompt influence GPT's performance when coding latent constructs?

universität wien

Data selection

XHate-999 data set (Glavaš et al., 2020)

- 99 comments that were posted in 10 news discussion threads in the Fox News website
- Parallel corpus: original comments are English, manually translated into Albanian, Croatian, German, Russian, and Turkish
- Context: USA (we selected nine of the ten threads from the Fox News website which were related to US politics and issues)

universität wien

Data selection

Construct: Inflammatory language (Yes/No)

Human Annotations

- Set A) Two native English speakers annotated the English user comments (Kappa score: 0.98) (Gao & Huang, 2018)
- Set B) Three researchers from our team annotated with the same codebook
- Annotations are assigned the text versions across languages

LLM Annotations

- *N* = 17,424 API requests (varying languages, prompt design, model)
- GPT3.5 vs. GPT4o by OpenAI

universität wien

Language Characteristics

Language	Production Mode	Resources for LLM Training (GPT3)[1]	Script
English	Original Version	High (91.7%)	Latin
Albanian	Manually Translated	Low (0.005%)	Latin
Croatian	Manually Translated	Medium (0.05)	Latin
German	Manually Translated	High (1,5%)	Latin
Russian	Manually Translated	Medium (0.2%)	Cyrillic
Turkish	Manually Translated	Medium (0.06)	Latin

[1] https://github.com/openai/gpt-3/blob/master/dataset_statistics/languages_by_word_count.csv

universität wien

Impact of Language (RQ1) & Prompt Design (RQ2)

Language	Prompt Languages		Text Language Mentioned in Prompt		Country Context			
English	English	-	No	Yes	No context	USA	Canada	UK
Albanian	English	Albanian	No	Yes	No context	USA	Albania	Kosovo
Croatian	English	Croatian	No	Yes	No context	USA	Croatia	Bosnia & Herz.
German	English	German	No	Yes	No context	USA	Austria	Germany
Russian	English	Russian	No	Yes	No context	USA	Russia	Belarus
Turkish	English	Turkish	No	Yes	No context	USA	Turkey	Cyprus

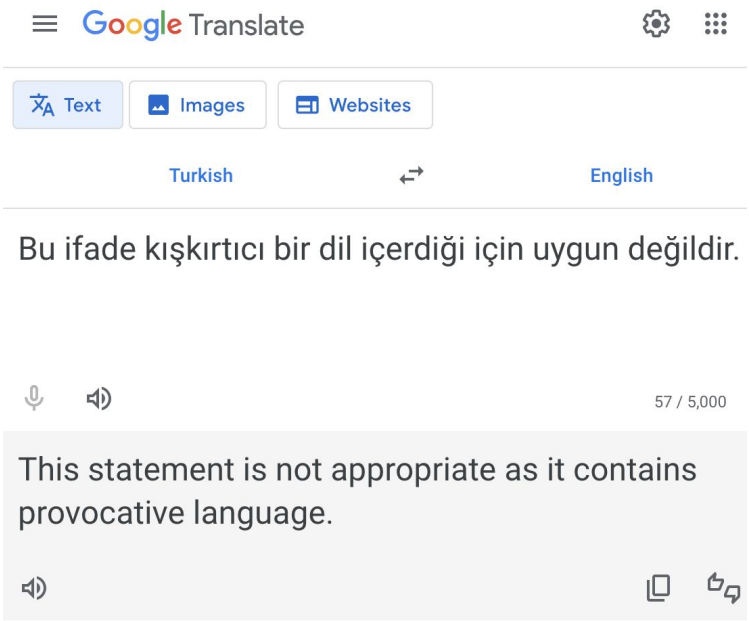
universität wien

Prompt example

```
Does the following {“”; English; Albanian; Croatian; German; Russian; Turkish} text
{“”; which was published in the United States; which was published in Canada; which
was published in the United Kingdom}, include inflammatory language or not?
Inflammatory language explicitly or implicitly threatens or demeans a person or a
group based upon a facet of their identity such as gender, ethnicity, or sexual
orientation.
Excludes insults towards other anonymous users.
Includes insults of belief systems.
Answer only with a letter: A or B
A: Inflammatory language
B: No Inflammatory language
You will output a string object containing either "A" or "B"
```

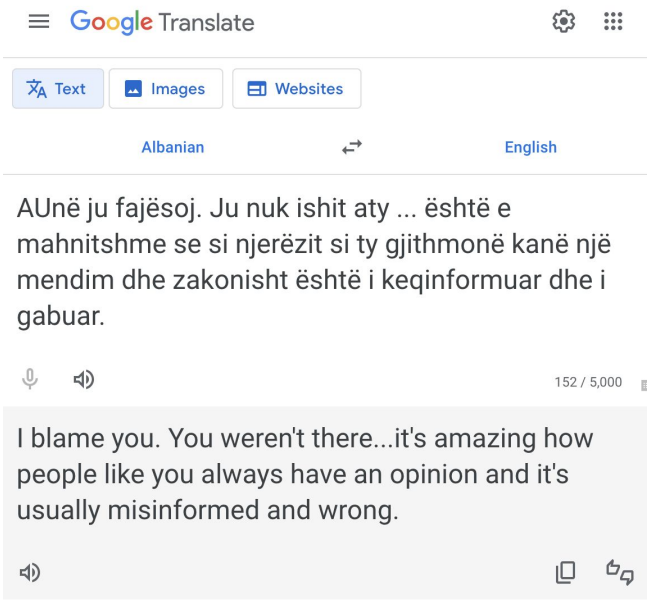
universität wien

LLM Output Postprocessing



universität wien

LLM Output Postprocessing



universität wien

LLM Output Postprocessing

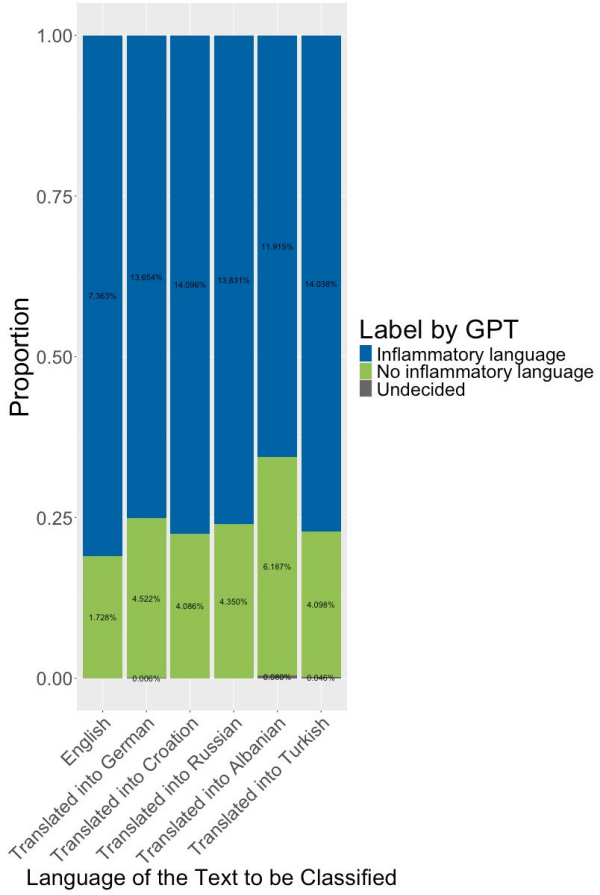
universität wien

Agreement between LLM and human generated labels

	Guo & Huang, 2018	
	Mean	*SD*
1 English	0.620	0.486
2 Albanian	0.698	0.459
3 Croatian	0.607	0.489
4 German	0.625	0.484
5 Russian	0.637	0.481
6 Turkish	0.578	0.494
Mean	0.623	0.484

Agreement between LLM and human generated labels

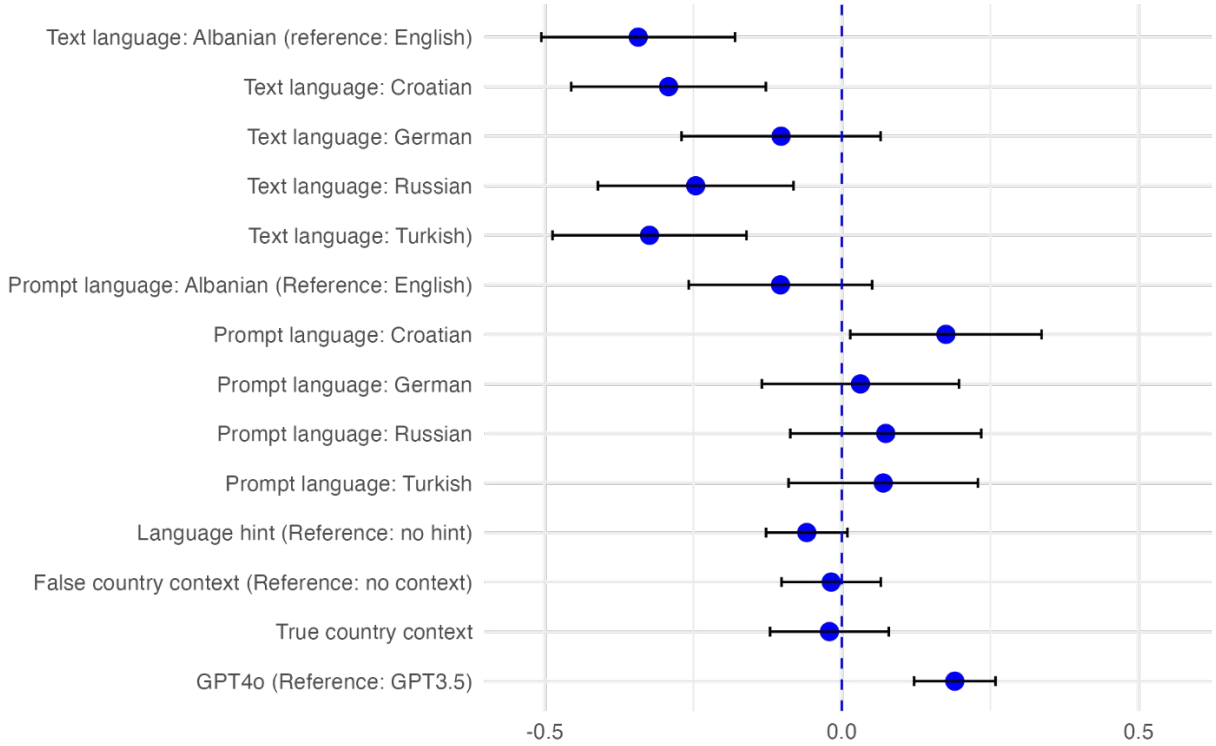
	Guo & Huang, 2018		Our labeling	
	Mean	*SD*	*Mean*	*SD*
1 English	0.620	0.486	0.746	0.435
2 Albanian	0.698	0.459	0.702	0.458
3 Croatian	0.607	0.489	0.766	0.424
4 German	0.625	0.484	0.774	0.418
5 Russian	0.637	0.481	0.756	0.430
6 Turkish	0.578	0.494	0.740	0.439
Mean	0.623	0.484	0.747	0.435



universität wien

Agreement between the LLM and human generated label

Results

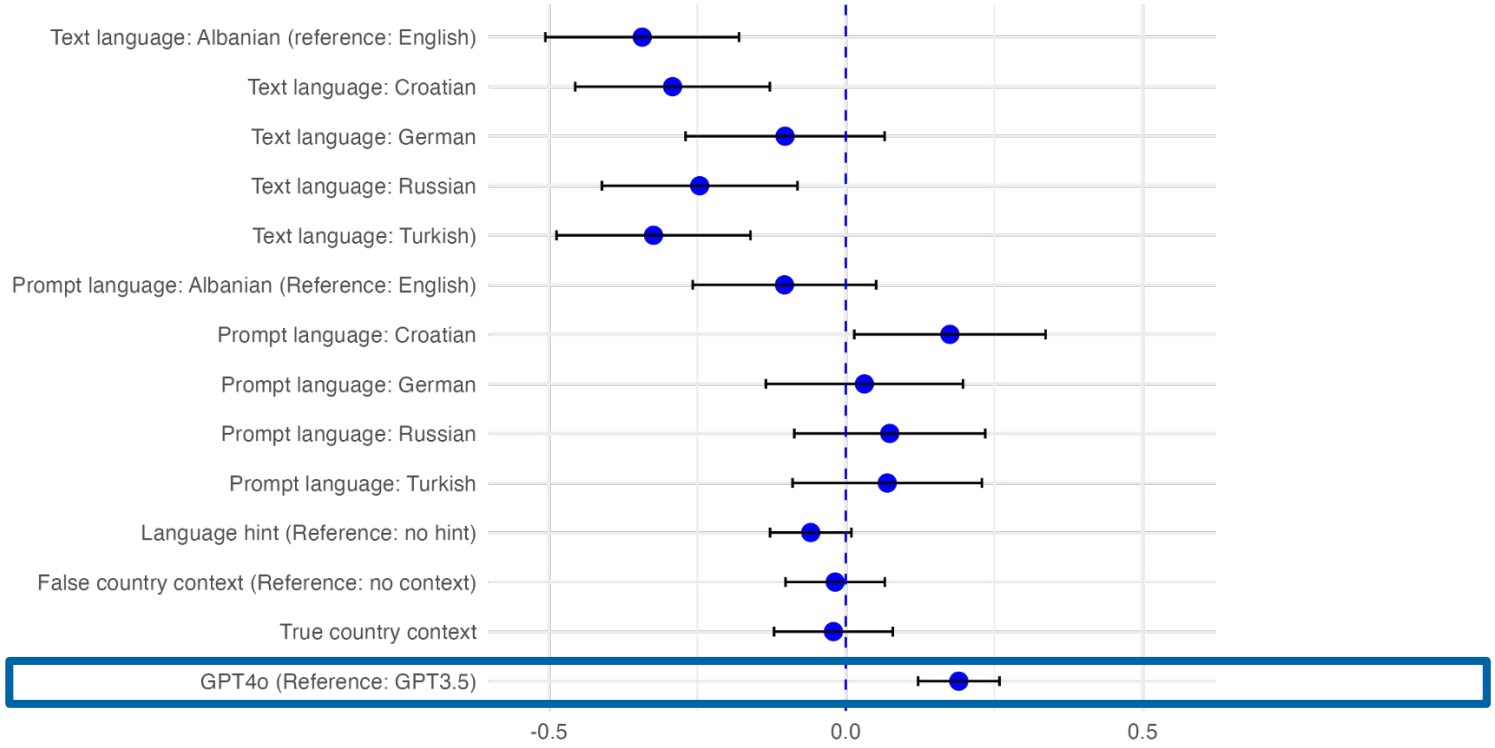


N = 17.424

universität wien

Agreement between the LLM and human generated label

Results

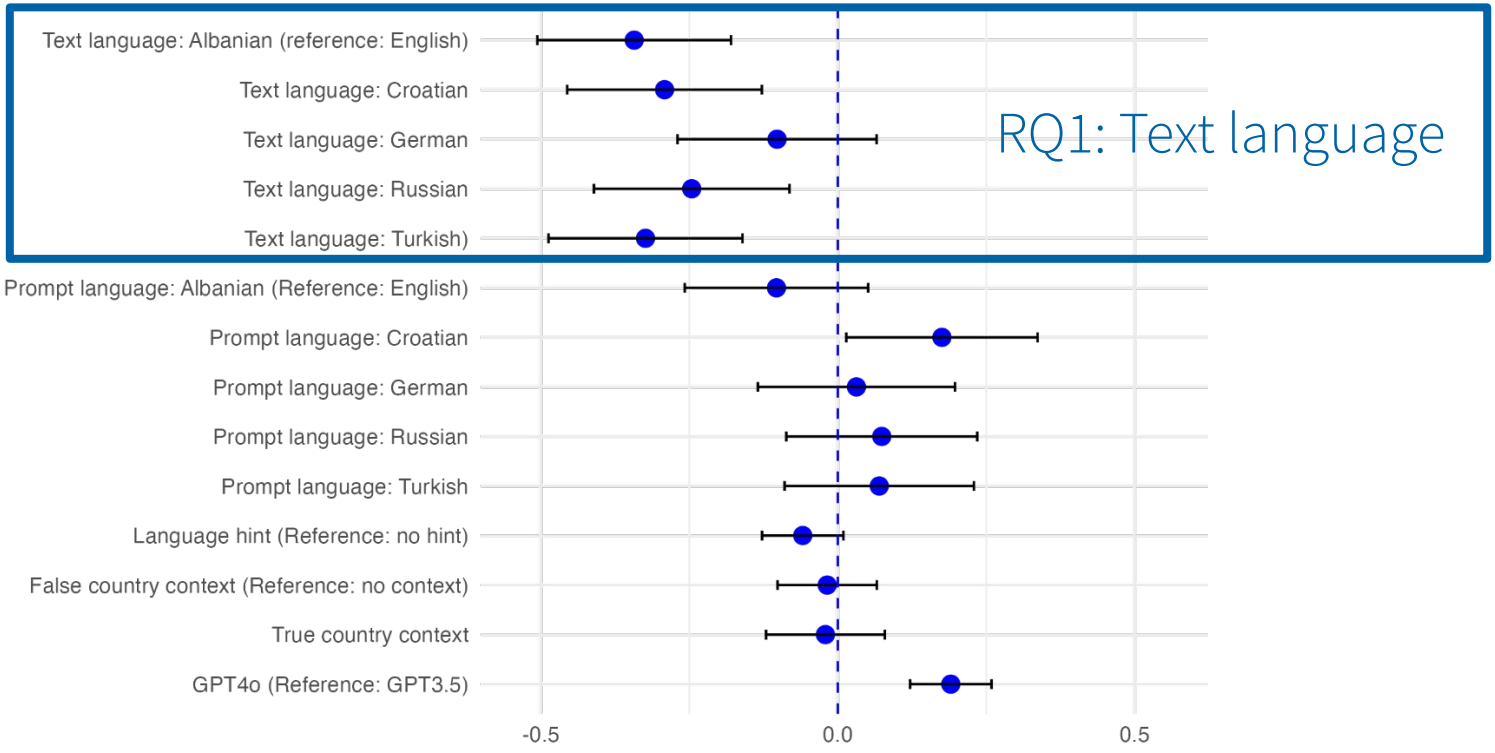


N = 17.424

universität wien

Agreement between the LLM and human generated label

Results



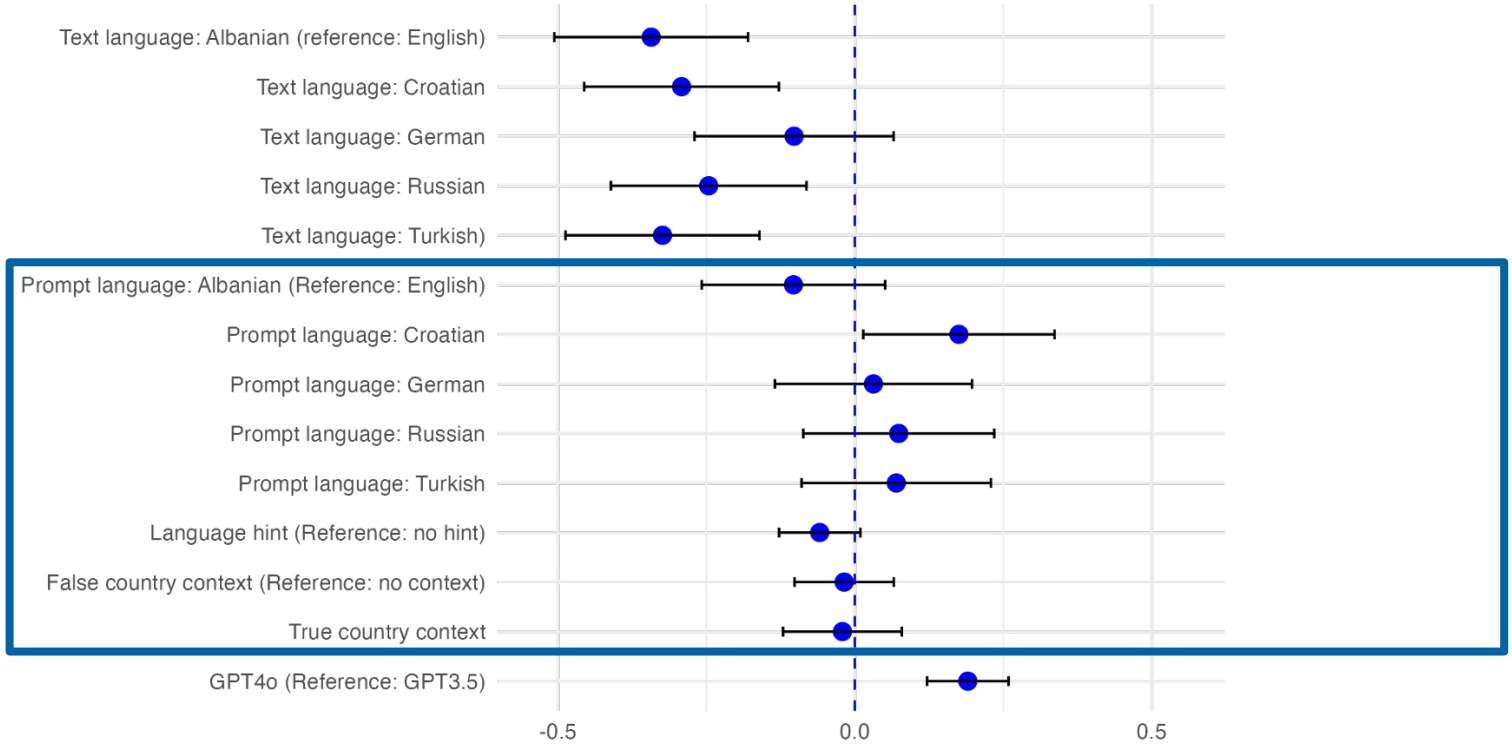
N = 17.424

universität wien

Agreement between the LLM and human generated label

Results

RQ2:
Prompt design



N = 17.424

universität wien

Take-aways

- LLM Coding performance is equally high for high-ressource languages German and English, but significantly lower for languages less represented in the training materials for LLMs
- Coding quality of LLMs is not affected by
  - prompt language
  - providing a language hint
  - providing country context information
- Validation for comparative studies: language and context sensitive coders

universität wien

Extending LLM Annotation Research

Examine the research questions with additional prompt experiments

- Add open source LLMs (e.g., BLOOM)
- Work with other datasets (more languages, other more context dependent concepts)

Study the impact of contexts (i.e., regional representation) of LLMs with other research designs

- Fine-tuning with context-specific datasets
- Systematic research on LLM documentation/set-up of large data sets (e.g., Common Crawl)

universität wien

Wrap up

Day 4

universität wien

Central decision criteria

- Approach to measurement (is there a ‘truth’)?
- Concepts/Interests
- Availability of methods (dictionaries, labeled data, pretrained models)
- Time (and Importance for a specific RQ)
- Budget (Modeling, Validation)
- Skills (R, Python, Patience)
- Model implementation
- Replicability
- Interpretability
- Ethical and environmental considerations
- …

universität wien

Case study 1

You want to analyze whether and how the newspaper coverage of political parties has changed in the election campaign.

What potentials and what problems do you see approaching this task with Transformer models?

universität wien

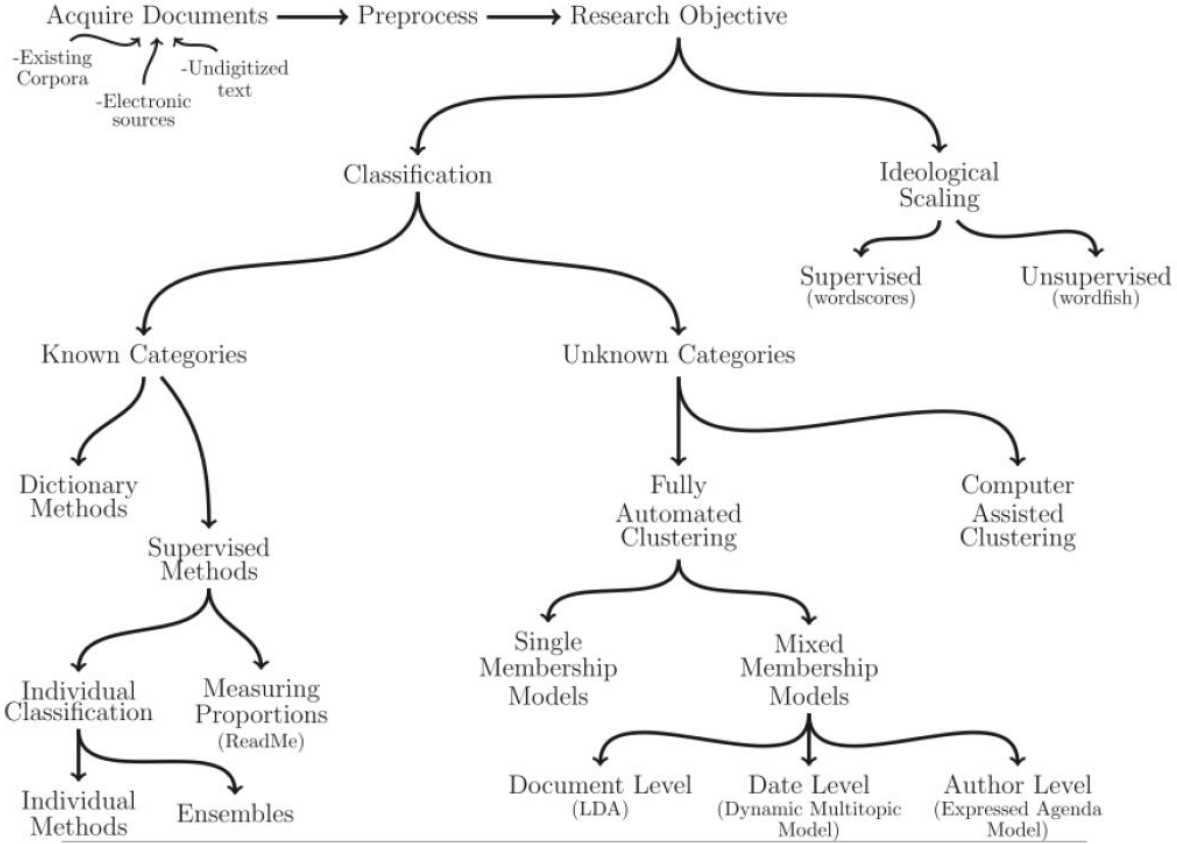
Case study 2 Data selection

You plan to study the emotional reactions of Austrian citizens in response to the attacks on work of arts.

What text data could you select?

How would you select only relevant content?

And what about validation?



Grimmer & Stuart, 2013, Fig 1

universität wien

Course assessment

Participation in class (20%)

Final paper: application of one or several automated text analysis methods on a topic related to the PhD thesis or a topic of free choice (80%)

- Contents: short motivation, analysis (commented code), description and interpretation of results (about 10 pages)
- Format: R Markdown
- Deadline: February 15th, 2024
- Send it to both of us via mail

universität wien

Final paper evaluation

Max. 80 points

Consistency (20 points): Motivation (objectives), analysis and interpretations of results are closely related

Readability/Format (20 points): Code is commented, with text

‘Correct’ application of method (20 points):

Critical reflection (20 points): some reflections for bigger decisions in the design and in respect to the result

Comptext 2025 (April 24-26, Vienna)

- Present your work and get feedback, only 250 word abstract needed, submit until January 15
- Workshops on many more computational topics
- https://www.comptextconference.org/7th-annual-comptext-conference-2025/

universität wien

Course evaluation

universität wien

Individual feedback

universität wien

Feedback for us

- Level of difficulty?

universität wien

Feedback for us

- Coverage of the field (prefer less topics more in depth or even more topics)?

universität wien

Feedback for us

- Application scenarios in your discipline?

universität wien

Feedback for us

- Data sources in your discipline?

universität wien

Feedback for us

- More time for working on coding challenges (without initial guidance)?

universität wien

Feedback for us

- What could we improve for the class next year?

universität wien

Pitch your projects

Very informal opportunity to pitch your text analysis use case and (initial) design

- Research question
- Data
- Methods
- Current struggles

And to receive some feedback (no grades, points, etc. just free brainstorming opportunity)

universität wien

Thank you very much