Going Cross-Lingual: Computational Methods for Multilingual Text Analysis

Hauke Licht and Fabienne Lind

Gesis Training Course

Cologne

6-8.12.2023

Contents (smaller changes possible)

Day	Morning Session	Afternoon Session
1	Intro & Overview about applications, problems, and solutions, validation	Corpus and Input Selection
2	Obtaining Measures: machine translation and supervised classifications	Multilingual supervised classification and measurement validation
3	Multilingual topic modeling with BERTopic and input on tokenization and pre-processing	Individual consultations on your projects

Workshop repository

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

Today

09:30 - 11:00	Annotation with GPT, Multilingual topic modeling
10:00 - 11:15	*Coffee break*
11:15 - 12:30	More resources, Fostering globally more inclusive research, Wrap-up
12:30 - 13:30	*Lunch break*
13:30 - 15:00	Individual consultations on participants' projects. Time can also be used by participants to work on their projects. We further prepare case studies for participants who prefer to work on prepared datasets and questions.
15:00 - 15:15	*Coffee break*
15:15 - 16:30	Individual consultations on participants' projects. Time can also be used by participants to work on their projects or the prepared examples the instructors provide.

What about GPT?

Annotating multilingual data with GPT?

First working papers examine the performance:

- (Rathje et al., 2023) https://psyarxiv.com/sekf5/
  - Data: tweets and news headlines
  - ChatGPT (zero-shot) vs. dictionary (against manual baseline)
  - GPT can accurately detect psychological constructs (sentiment, discrete emotions, and offensiveness) across 12 languages: high-resource (English, Arabic, Indonesian, Turkish) and low-resource languages (Swahili, Amharic, Yoruba and Kinyarwanda).
  - Performance worse for low-resource languages

Prompt examples (Rathje et al., 2023)

Table 2. Prompt table

Sentiment analysis (categorical)	**Emotion detection (categorical)**	**Offensiveness**	**Sentiment analysis (Likert)**	**Emotion detection (Likert)**
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]*	Which of these four emotions - [list of emotions] - best represents the mental state of the person writing the following (Indonesian) text? Answer only with a number: 1 if [emotion1], 2 if [emotion2], [...]. Here is the text: *[Tweet text]*	Is the following (Turkish) post offensive? Answer only with a number: 1 if offensive, and 0 if not offensive. Here is the post: *[Tweet text]*	How negative or positive is this headline on a 1-7 scale? Answer only with a number, with 1 being 'very negative' and 7 being 'very positive.' Here is the headline: *[Headline text]*	How much [emotion] is present in this headline on a 1-7 scale? Answer only with a number, with 1 being 'no [emotion]' and 7 being 'a great deal of [emotion].' Here is the headline: *[Headline text]*

Shown are all the prompts used for each construct. Non-English prompts were derived from the English prompts by specifying the language the text was written in. Prompts in combination with the tweet or headline text were run for each text entry in the dataset using the GPT API.

https://psyarxiv.com/sekf5/

Annotating multilingual data with GPT?

- (Kuzman et al., 2023) http://doi.org/10.48550/ARXIV.2303.03953
  - Data: English and Slovenian web content
  - ChatGPT (zero-shot) vs. fine-tuned large language models (against manual baseline)
  - English prompt with English text, English prompt with Slovenian text and Slovenian prompt with Slovenian text.
  - Results: ChatGPT outperforms the fine-tuned LLM on English test set. ChatGPT’s performance on the Slovene dataset is no worse than on English, provided that the prompt is in English instead of Slovenian.

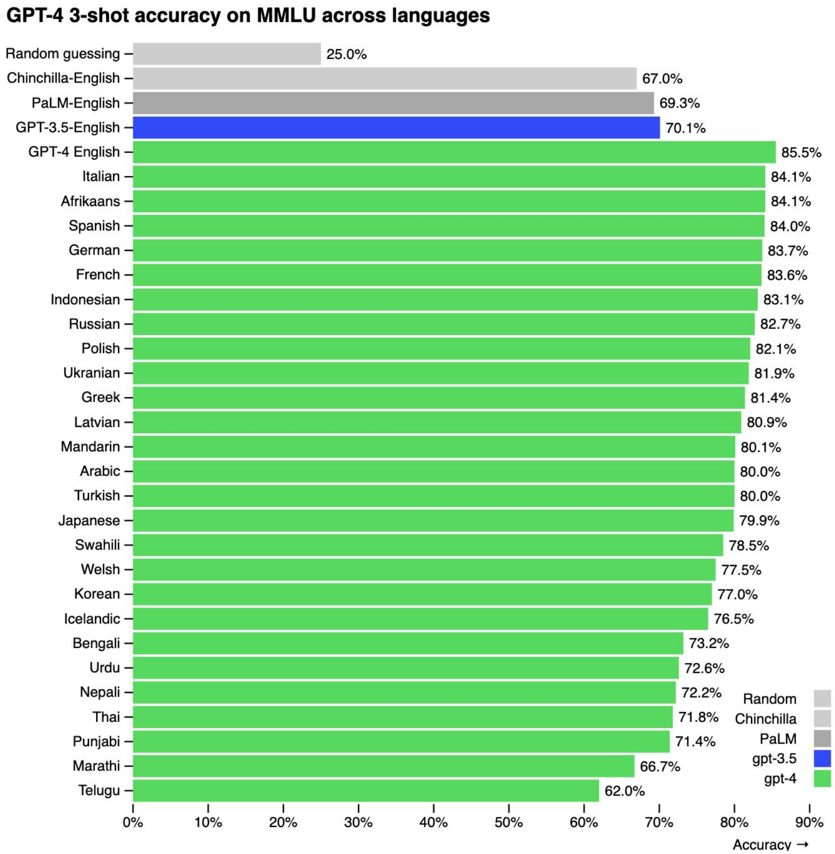


Figure 5. Performance of GPT-4 in a variety of languages compared to prior models in English on MMLU. GPT-4 outperforms the English-language performance of existing language models [2, 3] for the vast majority of languages tested, including low-resource languages such as Latvian, Welsh, and Swahili.

OpenAI. (2023). Technical Report.

Resources

Prompt writing help and best practices

- OpenAI "best practices"
- 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

Resources

Counting tokens

- https://help.openai.com/en/articles/4936856-what-are-tokens-and-how-to-count-them: Depending on the model used, requests can use up to 4097 tokens shared between prompt and completion. If your prompt is 4000 tokens, your completion can be 97 tokens at most.
- https://beta.openai.com/tokenizer
- https://github.com/openai/tiktoken

Let's code

- How to call the GPT API from R?
- How to prompt GPT and assess the performance against a benchmark?

Multilingual topic modeling

Input

Topic modeling in comparative research

- Objective: Topic extraction from document collections for comparative research

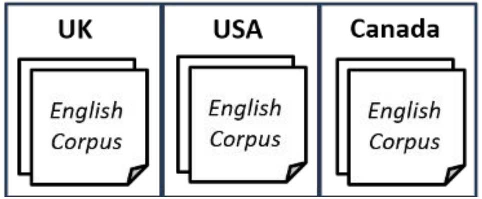
- Problem: Multilingual character of the data prevents direct application of “classic” topic modeling algorithms such as LDA

Aspects to consider

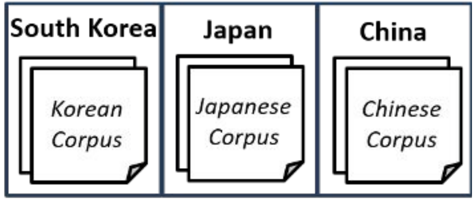
- Corpus
- Analysis Goal
- Comparability
- Resources

What is the corpus like?

Documents in one language

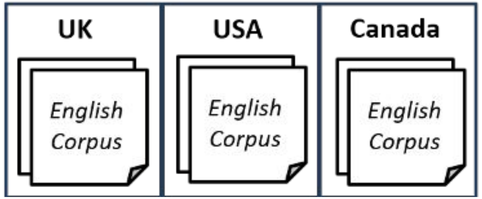


Documents in multiple languages



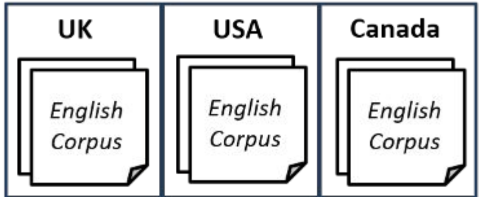
What is the analysis goal?

Identify case-specific topics



EMIC

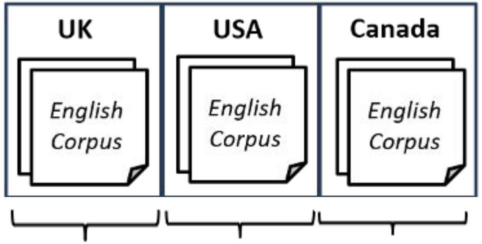
Identify meta-level topics across cases



ETIC

What is the analysis goal?

Identify case-specific topics

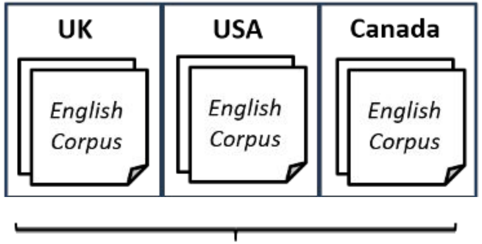


Run topic model for case 1

Run topic model for case 2

Run topic model for case 3

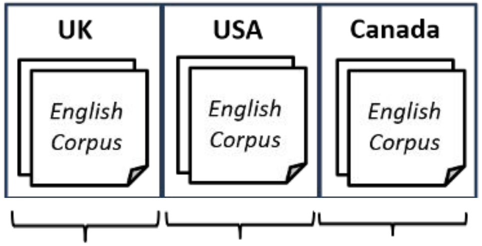
Identify meta-level topics across cases



Run one topic model for all cases together

How to compare the results?

Identify case-specific topics



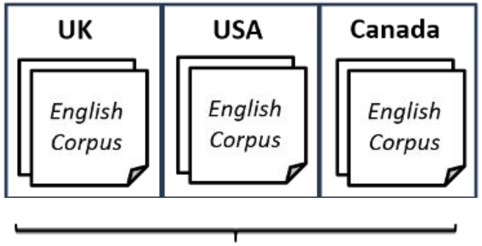
Run topic model for case 1

Run topic model for case 2

Run topic model for case 3

Qualitative comparison of the case-specific topics across cases

Identify meta-level topics across cases

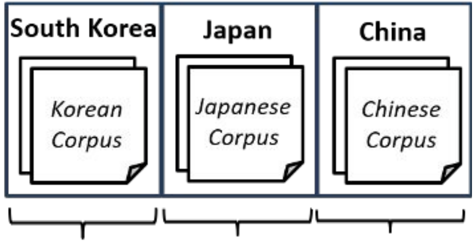


Run one topic model for all cases together

Numerical comparison of topic scores across cases

How to compare the results?

Identify case-specific topics



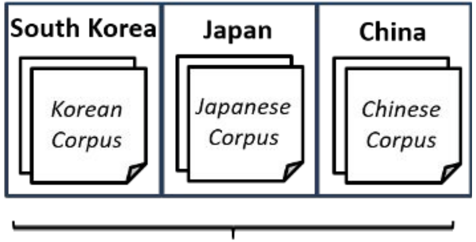
Run topic model for case 1

Run topic model for case 2

Run topic model for case 3

Qualitative comparison of the case-specific topics across cases

Identify meta-level topics across cases

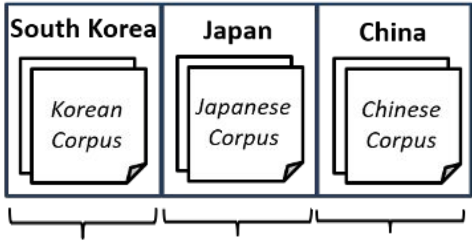


Run one topic model for all cases together

Numerical comparison of topic scores across cases

Crucial resources

Identify case-specific topics



Run topic model for case 1

Run topic model for case 2

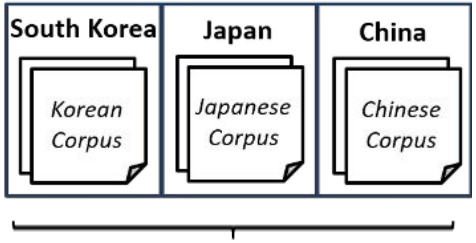
Run topic model for case 3

Qualitative comparison of the case-specific topics across cases

“Classic” topic modeling algorithms (e.g., LDA, BERTTopic)

Language and case experts for labeling and interpretati on

Identify meta-level topics across cases

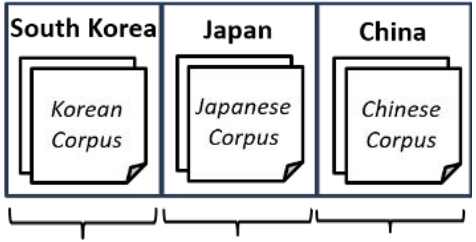


Run one topic model for all cases together

Numerical comparison of topic scores across cases

Crucial resources

Identify case-specific topics



Run topic model for case 1

Run topic model for case 2

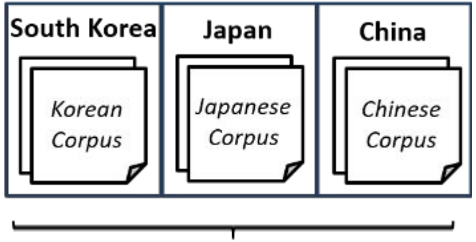
Run topic model for case 3

“Classic” topic modeling algorithms (e.g., LDA, BERTTopic)

Language and case experts

Qualitative comparison of the case-specific topics across cases

Identify meta-level topics across cases



Run one topic model for all cases together

Special linguistic resources necessary

Language and case experts

Numerical comparison of topic scores across cases

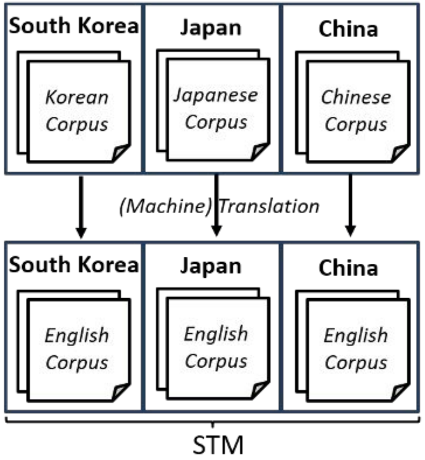
Techniques to identify meta-level topics across cases for a multilingual corpus

Their common strategy: Consolidating the data to a common denominator prior to analysis

Crucial Resources:

- Machine Translation (Lucas *et al.*, 2015;)
- Multilingual Dictionaries (Maier *et al.*, 2021)
- Multilingual Word Embeddings (Chan *et al.*, 2020)
- Multilingual Transformers (Grootendorst, 2022)

Example: Consolidating via translation



Language as covariate

BERTopic

- (first?) Transformer-based topic model
- *not* a statistical model (like LDA), but a pipeline of data science techniques

Github: https://maartengr.github.io/BERTopic/api/bertopic.html

Documentation: https://maartengr.github.io/BERTopic/api/bertopic.html

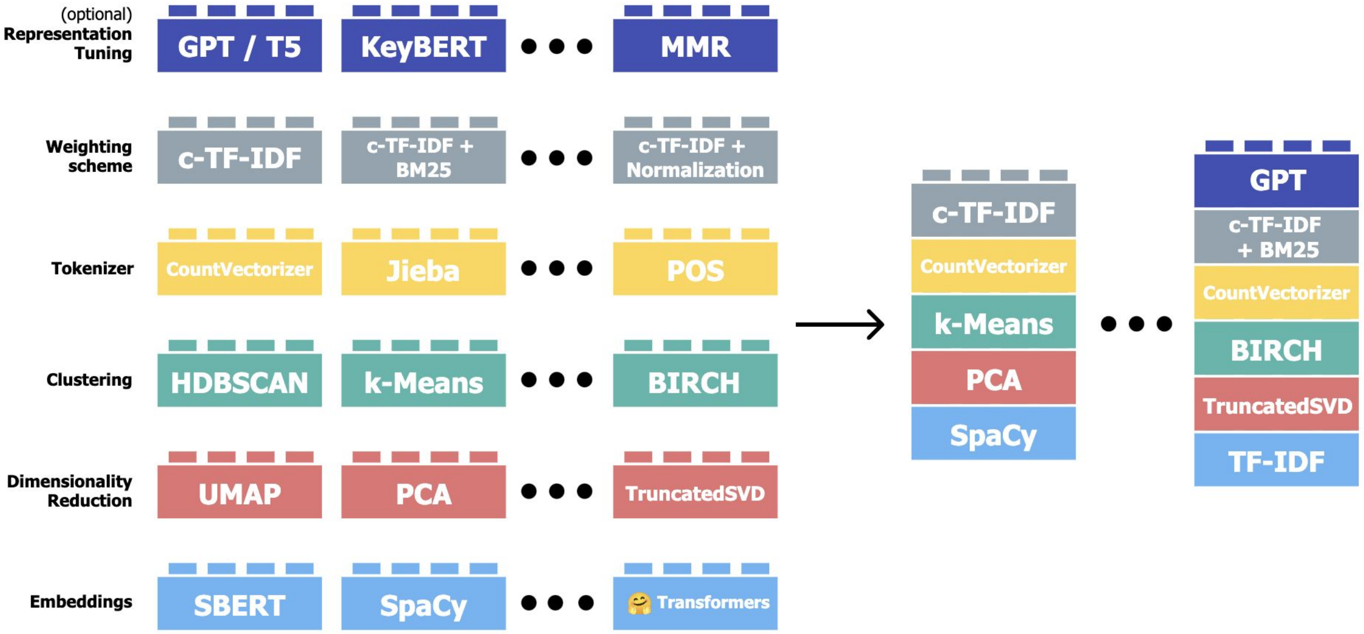
BERTopic

a modular **pipeline** of data science techniques

1. document/sentence embedding ⇒ from text to numeric vectors
2. dimensionality reduction ⇒ lower-dimensional doc. representation
3. un-/semi-supervised clustering ⇒ topic assignment
4. bag-of-words–based topic representation ⇒ returns topic-word scores

BERTopic

a **modular** pipeline



Let’s code!

notebook ‘code/bertopic_multilingual.ipynb’ on Github

Validating Topic Models

One strategy:

Coherence metric

REMINDER

Coherence i.e., close semantic relation of top words in one language and native speaker evaluation, see (Lau et al., 2014) for NPMI metric

Coherence (within language)

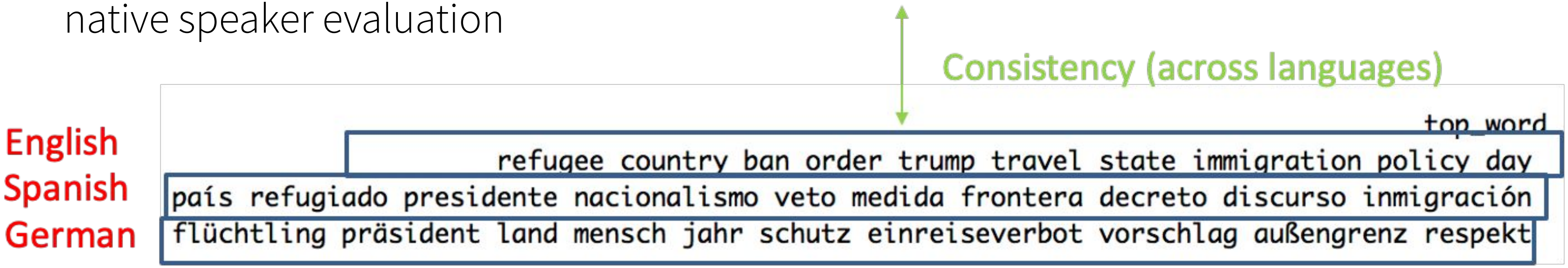
	top_word
English	refugee country ban order trump travel state immigration policy day
Spanish	país refugiado presidente nacionalismo veto medida frontera decreto discurso inmigración
German	flüchtling präsident land mensch jahr schutz einreiseverbot vorschlag außengrenz respekt

Example: Lind et al., 2022

Consistency metric

Consistency i.e., language specific representations of a multilingual topic relate to the same concept

- Machine Translation Accuracy metric (e.g., Boyd-Graber & Blei, 2009) and native speaker evaluation



Example: Lind et al., 2022

Illustration: Topic modeling

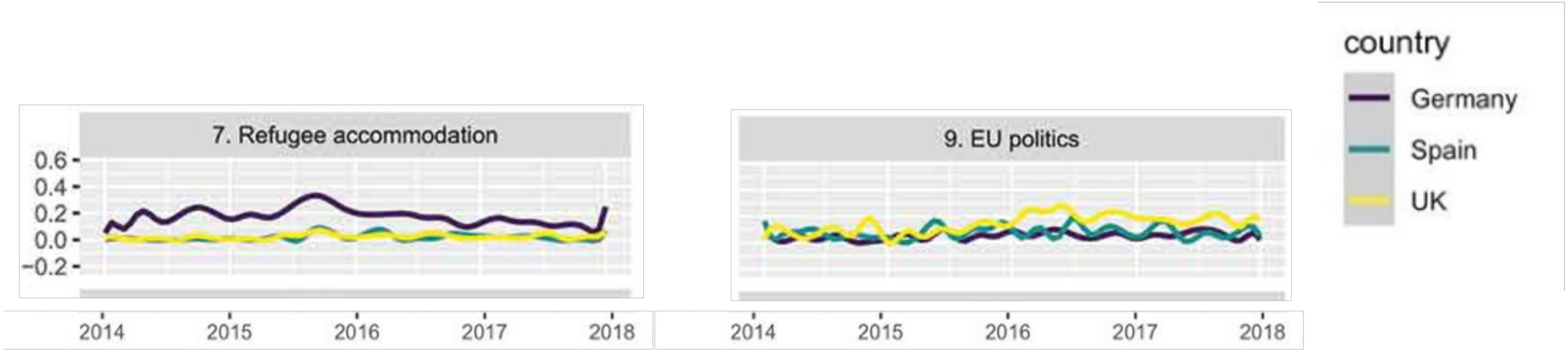
- Manual labeling of the final model

Topic	Lang.	Top 10 words
1.Welfare & jobs	EN	people worker immigration country migrant benefit report figure job health
	ES	persona número trabajador inmigración país millón inmigrante aumento cifra beneficio
	DE	zahl prozent land million migranten arbeit einwanderung bericht bevölkerung problem
2. Education	EN	school student child education project teacher university program class time
	ES	escuela estudiante niño proyecto educación joven clase universidad idioma programa
	DE	schule kind schüler projekt student universität arbeit lehrer sprache jugendliche
3. Election	EN	party election leader vote voter candidate campaign policy coalition poll
	ES	partido elección política campaña líder presidente voto votante candidato fiesta
	DE	partei wahl wähler abgeordnete stimme politik kandidat umfrage präsident rede
4. Security	EN	police time attack officer scene people crime security murder station
	ES	policía ataque hombre persona asesinato seguridad escena sospechoso grupo funcionario
	DE	polizei angriff polizist beamter anschlag szene täter mord opfer gruppe
5. Culture (film & theater)	EN	Film director series movie actor min love drama theater life
	ES	película director serie teatro actor cine comedia drama amor vida
	DE	film serie min schauspieler tv regisseur theater komödie leben drama
6. War	EN	war country attack force security government soldier camp terrorist city
	ES	guerra país ataque fuerza gobierno ejército seguridad presidente soldado arma
	DE	krieg land präsident soldat stadt angriff regierung kampf staat armee
7. Refugee accommodation	EN	refugee asylum seeker people accommodation country district situation office reception
	ES	refugiado asilo solicitante persona derecho alojamiento ayuda distrito oficina país
	DE	flüchtling asylbewerber unterkunft land nutzung hilfe zahl grenze syrer monat

Example: Lind et al., 2022

Face validity

assess expectations regarding the salience of individual topics in the different countries and at certain points in time with the topic visualization



Example: Lind et al., 2022

Convergent validity

- A comparison of the topic probabilities per document with external trusted measures for the same documents
- Lind et al., 2022: External measures obtained by keyword-based dictionaries designed to measure economy & budget, a security, and a welfare frame
- Results:
  - Economy & budget keywords most strongly related to the topic probabilities of topic 10 labeled "Economy."
  - Security keywords most strongly related to the topic probabilities of topic 4 labeled "Security".
  - Welfare keywords most strongly related to Topic 2 "Education" and 19 "Family"

Example: Lind et al., 2022

More Ressources

Examples

Annotation

Coding Tools

- Google sheets
- AnnoTinder
- docanno

Crowd-coding platform

- Prolific: you can use screeners to select coders based on language skills
- Cloud research

more slides from a past GESIS course here

Example: Baseline creation templates For search strings

Create sampling plan (goal: representative for universe of texts)

Database	Date	Outlet	Number of all articles published that day
APA	Mon 8.10.2018	Standard	136
APA	Tue 9.07.2019	Standard	87
APA	Wed 12.02.2020	Standard	89
APA	Thr 15.04.2021	Standard	98
APA	Fri 27.05.2022	Standard	94
Note: Ideally repeat this procedure for each outlet and case included; cover the full range of time period investigated			

Collect articles

Code manually

Search with search string

Article id	Date	Text	Manually perceived as relevant (1=yes, 0 = No)	Perceived as relevant by search string (1=yes, 0 = No)
1	Mon 8.10.2018	Kern ist an sich selbst gescheitert. Die SPÖ braucht jetzt mehr Geradlinigkeit und weniger Gockelhaftigkeit …	1	1
2	Mon 8.10.2018	Impressum und Offenlegung: Herausgeber: Oscar Bronner…	0	0
3	Mon 8.10.2018	Einseitiger Vorschlag. Zu viele Waffen in der Hand der Bürger sind gefährlich. Ein Blick in die USA zeigt, warum. Im Kern geht es…	0	1
…	…	…	…	

Calculate recall and precision

Multilingual computational text analysis resources for comparative research (selection) 1/2

Function	Name	Authors	Countries	Languages
Geographical classification of text	Newsmap R package	Watanabe, 2018	240	12
Language and Location Code Convertor	ISOcodes R package	Buchta & Hornik, 2022	249	7000+
Obtain typological information (e.g., Phonology, Lexical semantics) about a language	World Atlas of Language Structures (WALS)	Dryer & Haspelmath, 2022	-	2,676

Multilingual computational text analysis resources for comparative research (selection) 2/2

Function	Name	Authors	Countries	Languages
Named entity detection and extraction tools	SpaCy	Honnibal et al., 2023	-	72+
Open Source LLMs and datasets	Hugging Face	Hugging Face, Inc., 2023	-	200
Inventory of news source names, tools, datasets, organizations	Meteor	Balluff et al., 2022	34	164
Multilingual tokenization and pre-processing (in python)	nltk, stanza see code in our Github repo			

Towards more global, inclusive text analysis

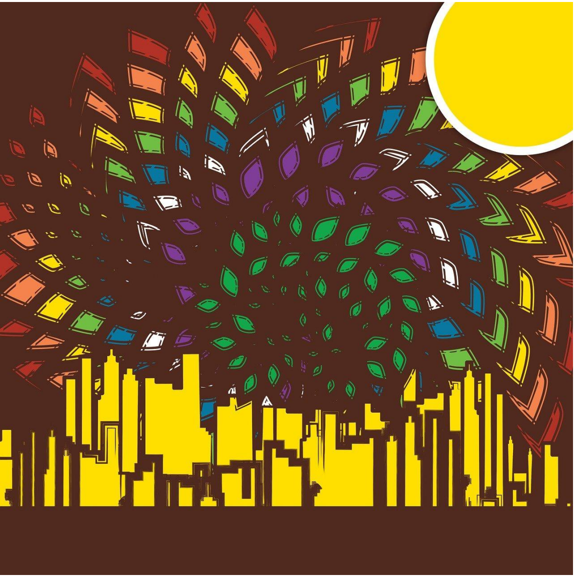
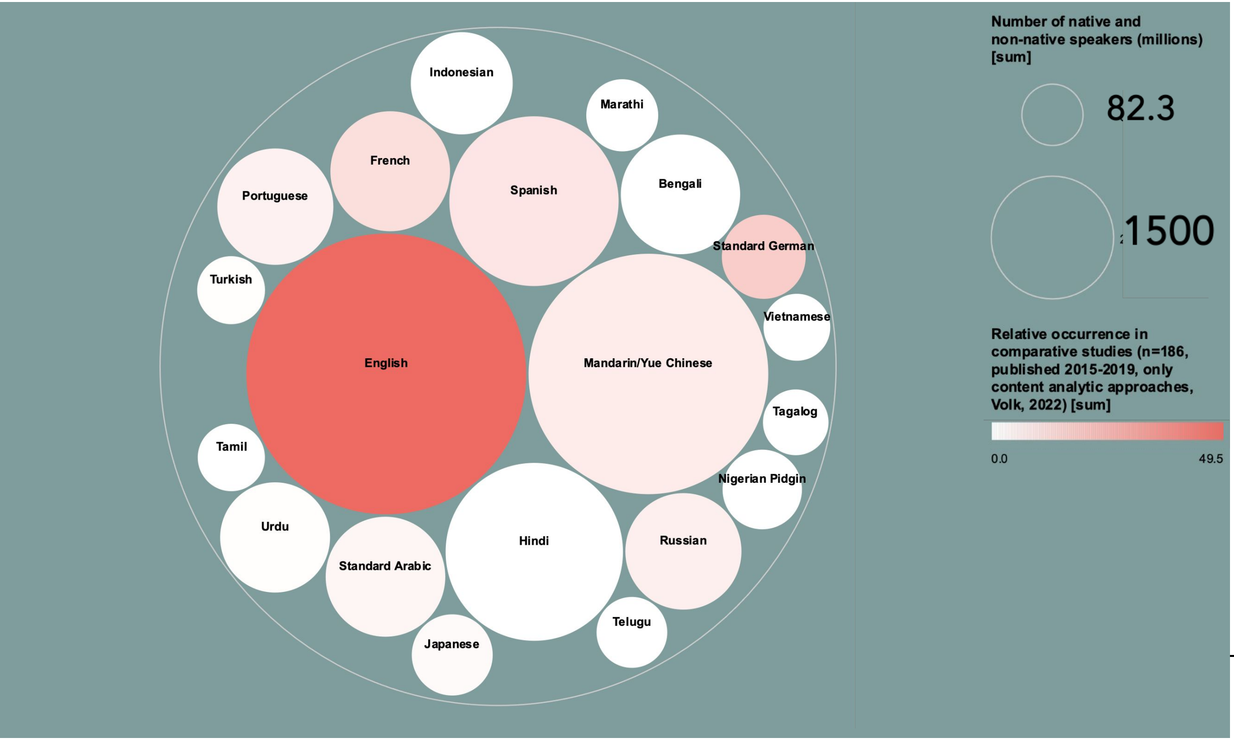


Image by ooceey from Pixabay

Progressing internationalisation

- Internationalisation of research and institutions in the social sciences has been picking up speed (e.g., Henriksen, 2016, Scharkow & Trepte, 2023).
- Growing awareness of the need to address persistent power asymmetries in the field (Demeter, 2022)
- Efforts to expand the focus of research beyond the dominance of Western, educated, industrialized, rich, and democratic (WEIRD) countries (Henrich et al., 2010)
    - **In text analysis:** Developing and employing methods for non-WEIRD countries and beyond English

Top 20 most spoken languages and their occurrence in comparative communication research



Lind, F. & Volk, S. (under review).

Top 10 countries with largest populations and their occurrence in comparative communication research

Country	Population %	Occurrence in comparative content analysis (% of 186 studies)
China	18.5	16.7
India	17.7	9.7
USA	4.2	47.8
Indonesia	3.5	1.6
Pakistan	2.8	3.8

Country	Population %	Occurrence in comparative content analysis (% of 186 studies)
Brazil	2.7	8.1
Nigeria	2.6	0.5
Bangladesh	2.1	1.1
Russia	1.9	8.1
Mexico	1.7	5.4

Lind, F. & Volk, S. (under review).

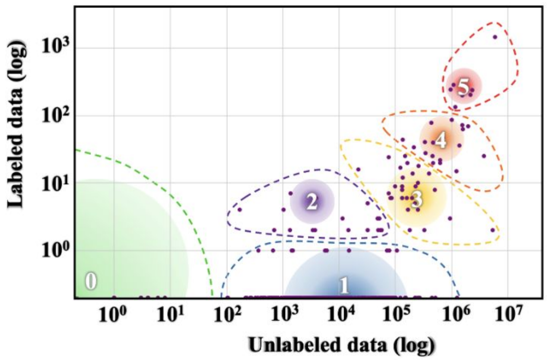
Questioning the “language agnostic” status of LLMs

Joshi et al., 2021 (see also Lauscher et al., 2020)

- LLMs rely on large amounts of labeled and unlabeled data for training
- not all languages are equally represented in training and development and the latest technologies
- availability and number of labeled and unlabeled data is a main factor for whether a language is included and to what extent
- in NLP literature, researchers differentiate between ‘low-resource’ languages and ‘high-resource’ languages

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%

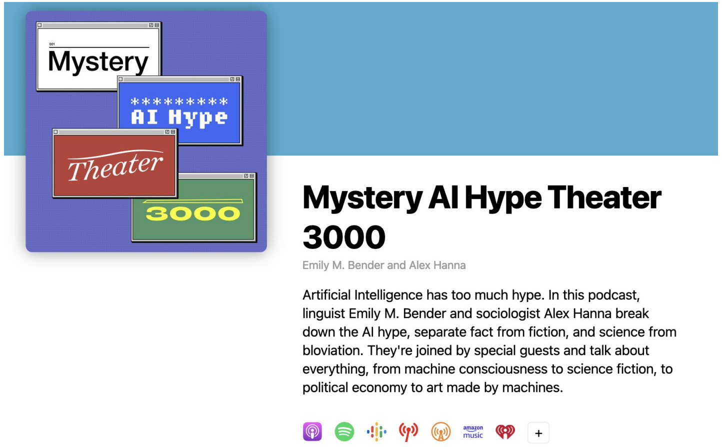
Table 1: Number of languages, number of speakers, and percentage of total languages for each language class.



p.6284f

Risks of LLMs for certain countries

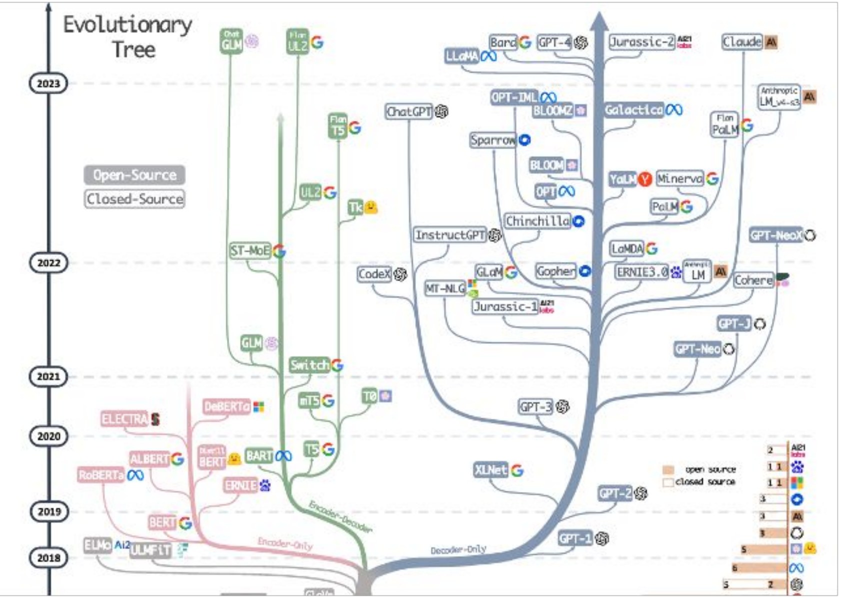
- Bender et a., 2021
  - the environmental impact of training LLMs affects certain countries more than others
  - overrepresentation of hegemonic viewpoints encoded in LLMs and the resulting lack of diversity



https://www.buzzsprout.com/2126417

Discussion points

Discussion



Yang et al., 2023, p. 3

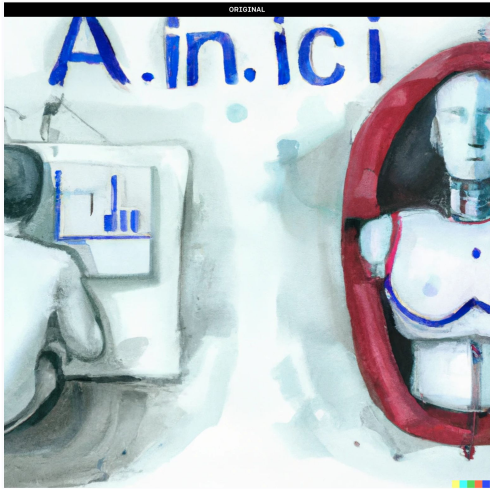
- The ways that we conduct multilingual CTA is changing drastically, but questions that we ask about validation do not
- Practical implementation of validation strategies requires significant resources
    - Research infrastructures, open science initiatives, international collaboration

Consequences for comparison of scores if validation reveals issues?

- report and reflect on the detected problems -> enables future research to build on better information on related measurements
- make a considered decision about the extent to which the measurements can be used to make substantive comparative statements about the cases
  - Are the measurements suitable for statements per case or for comparisons among a subset of the cases?
- explore error correction methods to account for misclassifications (i.e., Bachl & Scharkow, 2017; TeBlunthuis et al., 2023)

The end of manual coding?

- Augmenting not replacing (Grimmer & Steward, 2013)
- Human input for quality control:
  - select, monitor, and test on the level of corpus, data inputs, process, outputs
  - Even more important in projects with multiple cases and languages
  - Don’t trust numbers trust yourself and other human coders



DALL.E