Going Cross-Lingual: Computational Methods for Multilingual Text Analysis

Fabienne Lind

Workshop

TU Illmenau

26.04.2024

About me

Fabienne Lind

- Post-Doc, Computational Communication Science Lab, University of Vienna
- Research focus: Multilingual computational text analysis and comparative research methods; social inequality in the context of migration, knowledge gap, climate change
- fabienne.lind@univie.ac.at

I will share insights from

Joint publication with Hauke Licht:

- Going cross-lingual: A guide to multilingual text analysis https://doi.org/10.5117/CCR2023.2.2.LICH

REMINDER (2017-2019)

- Comparison of migration news discourse in 7 countries
- Relevant today: Publications covering multilingual methods for comparative research

OPTED (2020-2023)

- European infrastructure design for text analysis in pol. com
- Relevant today: Validation framework and tools for multilingual text analysis

Your turn :)

- Name
- Affiliation? Background?
- Experience with (automated) content analysis R & Python
- What are the expectations and wishes for the workshop and the workshop leader?

Course objectives

- Getting to know key strategies of multilingual text analysis for comparative designs
- Insight into practical challenges
- Critical reflection on the methods and their validation
- Inspiration for your own projects

Workshop philosophie

Topics are covered with

- Lecture style input
- Guided coded sessions
- Plenum discussions

Interrupt, ask all kinds of questions!

Workshop repository

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

Today

09:00 - 10:30	Introduction to the topic, overview about applications, validity, main problems and the main solutions approaches
10:30 - 10:45	*Coffee break*
10:45 - 12:00	Multilingual search string/keyword selection and testing
12:00 - 13:00	*Lunch break*
13:00 - 14:30	Machine translation
13:30 - 14:45	*Coffee break*
14:45 - 16:00	Overview about additional topics (e.g., multilingual topic models, LLMs), ressources
16:00 - 17:00	Time to discuss projects by participants 1:1

Time to discuss your projects

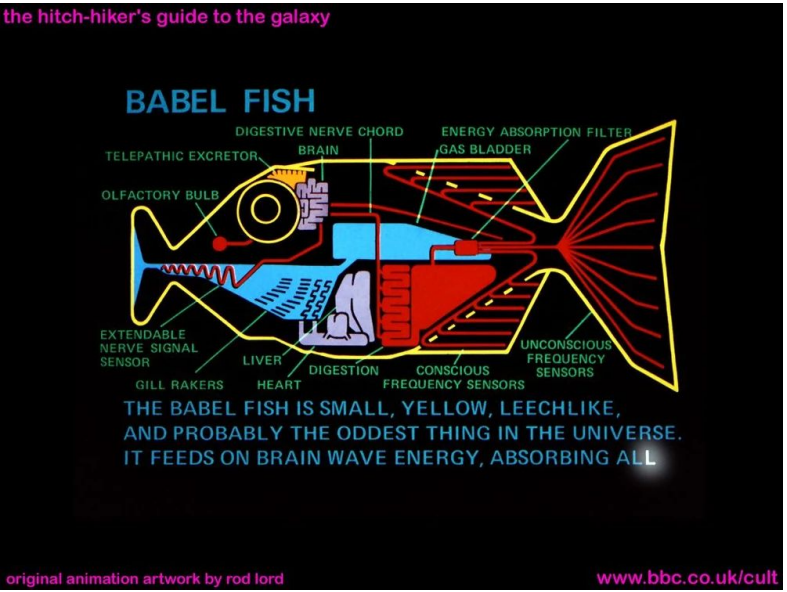
- Informal opportunity to talk about your use cases and (initial) design (plans)
- Research question, Data, Methods, Current struggles
- Dedicated slot: 4-5 pm

Introduction

What is the “Babel” problem?

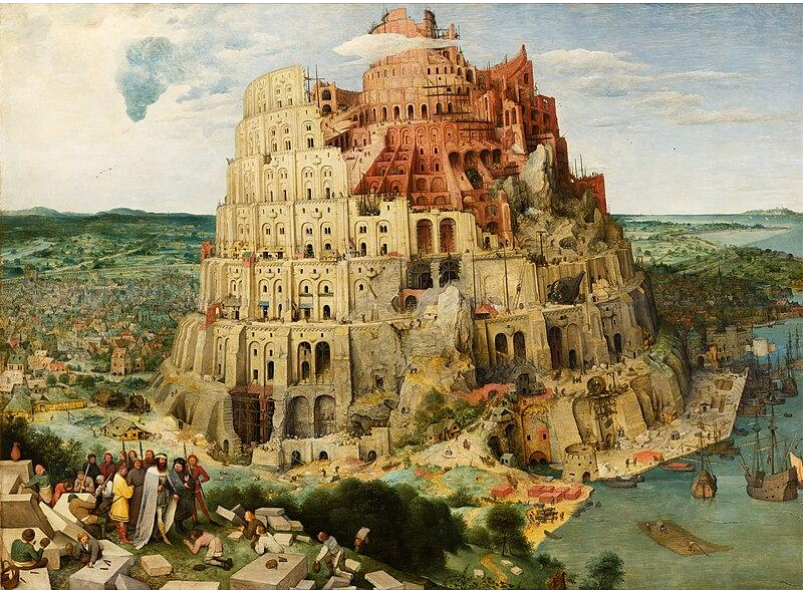
Anyone using this app?

“Probably the oddest thing in the Universe.”



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

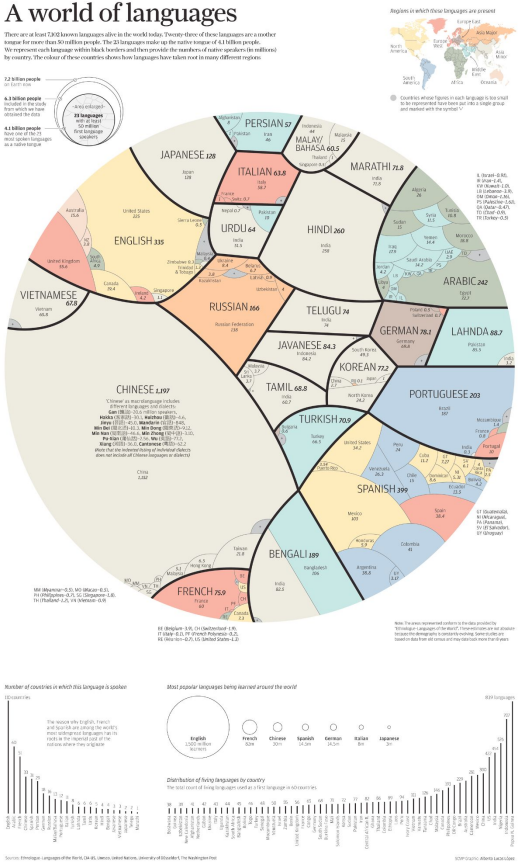
The tower of Babel



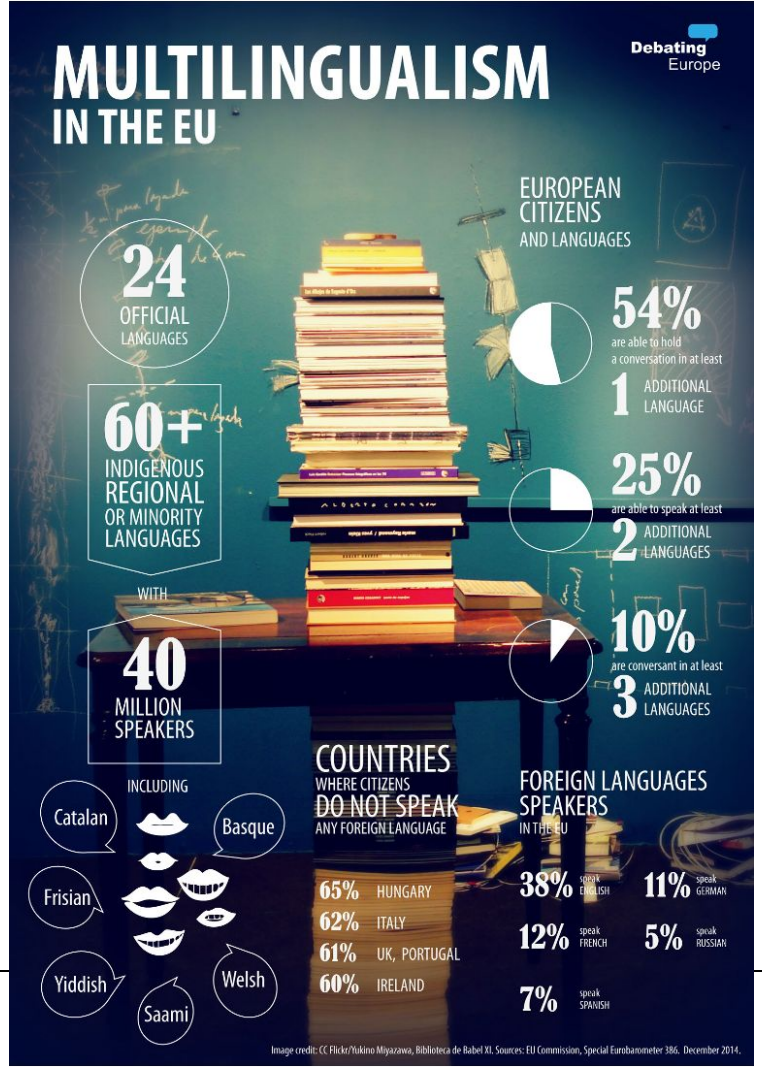
The Tower of Babel by Pieter Bruegel the Elder, 1563. (Wikimedia Commons)

Global linguistic diversity is enormous…

- 7000+ languages are spoken globally today.



https://www.scmp.com/infographics/article/1810040/infographic-world-languages


MULTILINGUALISM
IN THE EU
Debating Europe
24
OFFICIAL LANGUAGES
60+
INDIGENOUS REGIONAL OR MINORITY LANGUAGES
WITH
40
MILLION SPEAKERS
INCLUDING
Catalan
Basque
Frisian
Yiddish
Saami
Welsh
EUROPEAN CITIZENS AND LANGUAGES
54%
are able to hold a conversation in at least
1 ADDITIONAL LANGUAGE
25%
are able to speak at least
2 ADDITIONAL LANGUAGES
10%
are conversant in at least
3 ADDITIONAL LANGUAGES
COUNTRIES WHERE CITIZENS DO NOT SPEAK ANY FOREIGN LANGUAGE
65% HUNGARY
62% ITALY
61% UK, PORTUGAL
60% IRELAND
FOREIGN LANGUAGES SPEAKERS IN THE EU
38% speak ENGLISH
11% speak GERMAN
12% speak FRENCH
5% speak RUSSIAN
7% speak SPANISH
Image credit: CC Flickr/Yukino Miyazawa, Biblioteca de Babel XI. Sources: EU Commission, Special Eurobarometer 386. December 2014.

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

- English
  - has a considerable head start in computational development
  - is academic lingua franca
  - has 1.5 million learners globally
  - etc.

Baden, C., Pipal, C., Schoonvelde, M., & van der Velden, M. A. G. (2022). Three gaps in computational text analysis methods for social sciences: A research agenda. *Communication Methods and Measures*, *16*(1), 1-18.

Language selection?

English is highly popular (speaker + technical development)

There is a broad diversity of languages

Discuss:

- What languages should we include in our text analysis projects?
- What are guiding criteria?

Overview about Applications

When are researchers going cross-lingual?

Applications

In the social sciences:

- Comparative research
- Towards more global, inclusive text analysis
- etc.

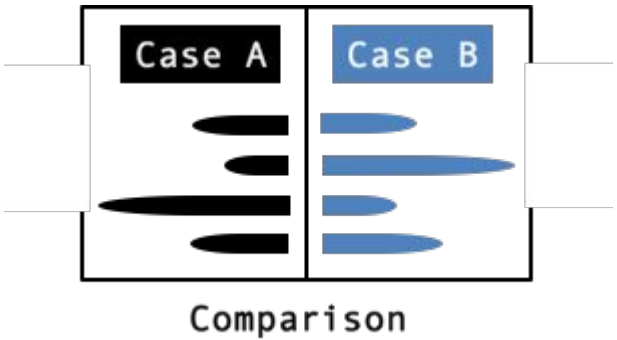
Comparative research

DALL.E

Comparative research

- **Comparative research in social science** involves comparisons between a minimum of two cases with at least one object of investigation relevant to social science research.

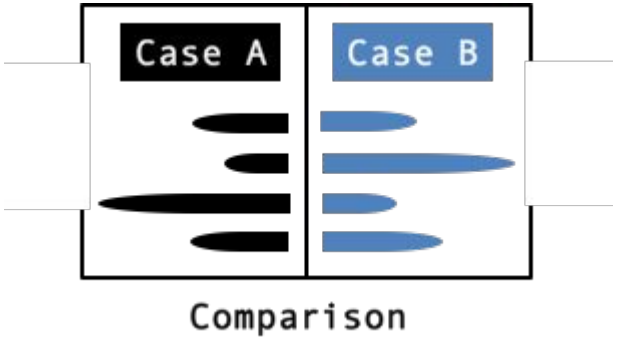
- **Cases** are defined as macro-level units such as systems, cultures, countries, and markets



Boomgaarden & Song, 2019; McLeod &
Blumler, 1987; Esser & Vliegenthart, 2017; Esser & Hanitzsch, 2012; Livingstone, 2003

Reasons to compare

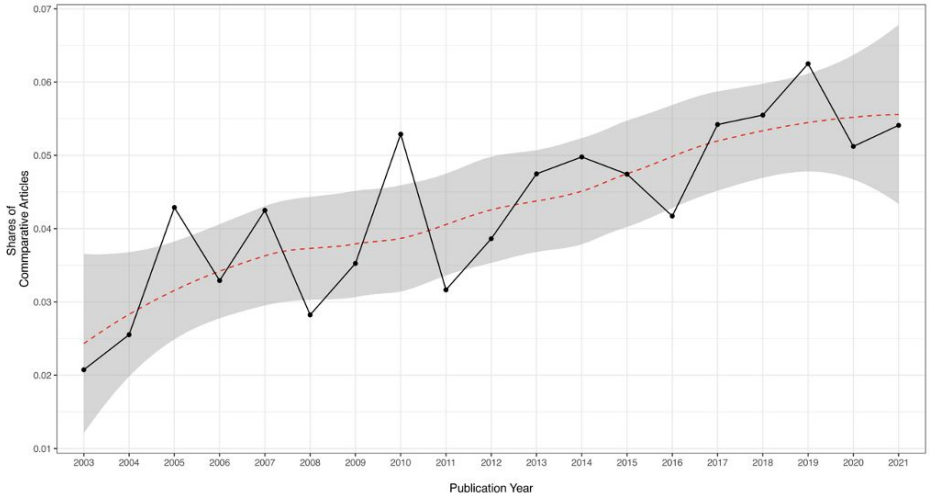
- Insights into the differences and similarities of cases
- Improved understanding and contextualization of the own case
- Raised awareness for other cases
- The test and generalizability of theories across diverse settings



Boomgaarden & Song, 2019; McLeod &
Blumler, 1987; Esser & Vliegenthart, 2017; Esser & Hanitzsch, 2012; Livingstone, 2003

Increasing popularity of comparative research

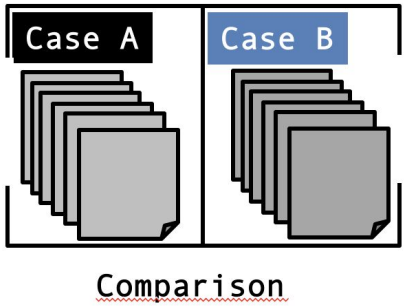
Shares of comparative articles in comm science journals, 2003-2021



Lind, Boomgaarden, Kathirgamalingam, Song, Syed Ali, & Vliegenthart (Working Paper).

Comparison of cases with content analysis

- Cases can be represented with documents
- Content analysis as one method to draw reproducible and valid inferences or meanings from documents (Krippendorff, 2004, p. 18)
- The goal of comparative content analysis is to compare cases via the comparison of specific contents in the documents

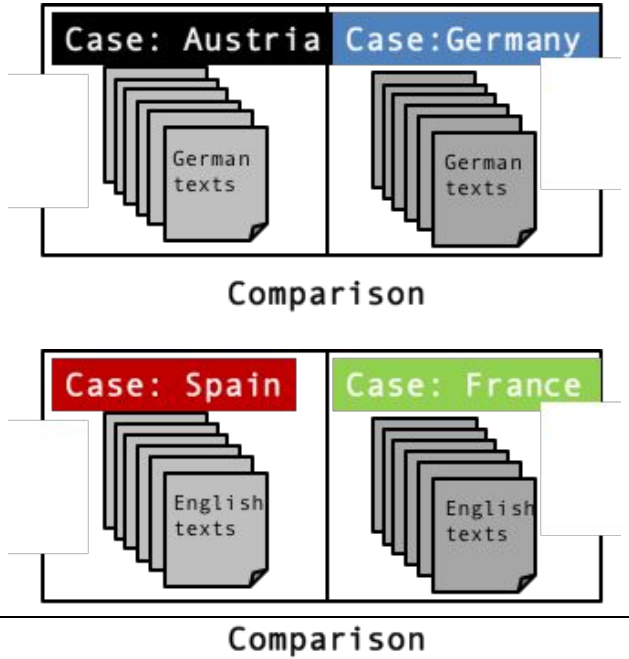


Boomgaarden & Song, 2019; McLeod &
Blumler, 1987; Esser & Vliegenthart, 2017; Esser & Hanitzsch, 2012; Livingstone, 2003

Comparison of cases & language(s) of documents

Monolingual scenarios exist

- compared documents are published for cases with very similar official languages (e.g., Austria, Germany, Switzerland; Gründl, 2020)
- Selecting text types that are available in the same language (e.g., English tweets, English international media reporting) for all cases (e.g., Abdelwahab et al., 2014)



Comparison of cases & language(s) of documents

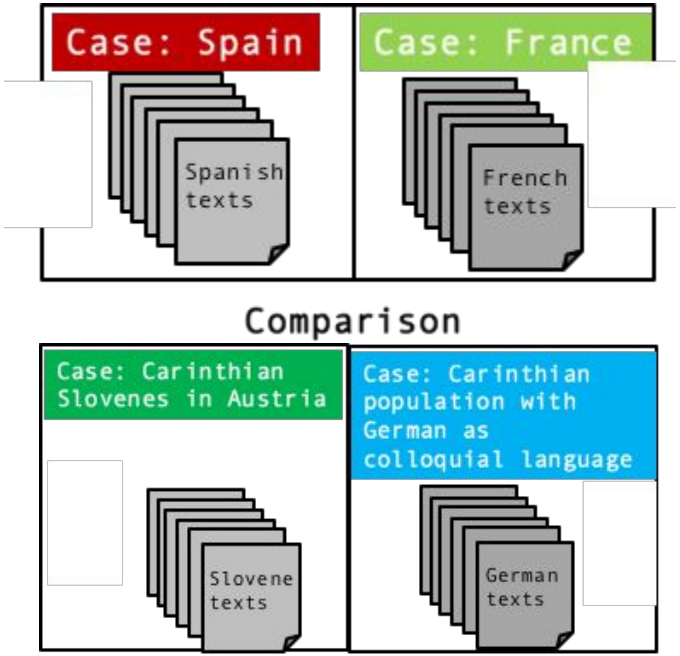
But the likely scenario is multilingual

- human communication of at least two compared cases manifests in texts in different languages

Comparison of cases & language(s) of documents

Examples for multilingual scenarios:

- Countries with different official languages
- Language areas (e.g., in Belgium or Switzerland)
- Sub-national regions such as the BasqueCountry and Catalonia in Spain, and Quebec in Canada)
- Minority languages in a country (e.g., Slovene in Austria)



Comparisons of cases with content analysis

Manual large-scale content analysis have been worthwhile mainly for a few selected topics with sufficient budgets. For example:

- media coverage of the European elections PIREDEU (Banducci et al., 2014)
- political news stories from 16 countries NEPOCS project (Hopmann et al., 2016)
- parties' electoral manifestos MARPOR (Volkens et al., 2015)

Computational text analysis as fast and reliable alternative to analyze large numbers of documents

Multilingual computational text analysis methods

- Methods to process and analyze large numbers of documents that are written in multiple languages
- Useful for comparative research designs when the human communication of at least two compared cases manifests in texts in different languages

Analysis goals (just as in monolingual content analysis)

- Classification, Topic Modeling, Scaling, etc.

Towards more global, inclusive text analysis

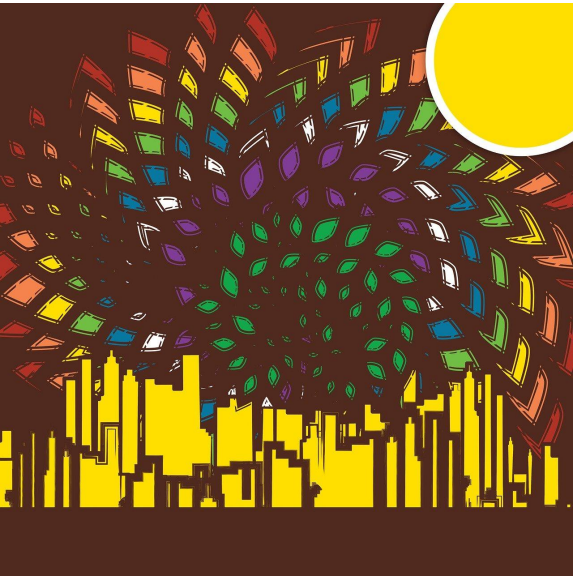


Image by ooceey from Pixabay

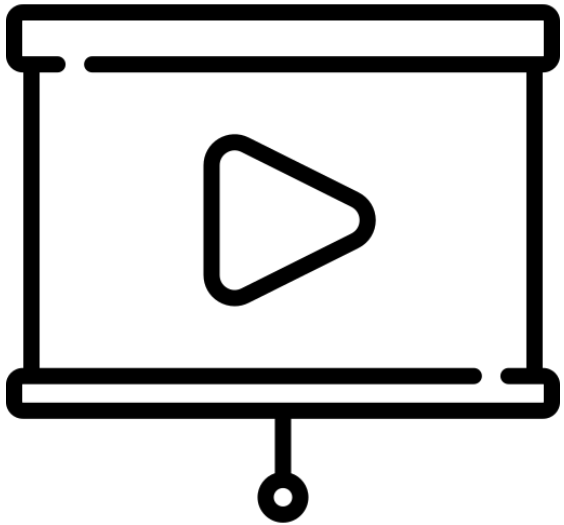
Progressing internationalisation

- 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

Your (initial) text analysis project ideas?:)



by Freepik - Flaticon

by Freepik - Flaticon

Main goal of the workshop

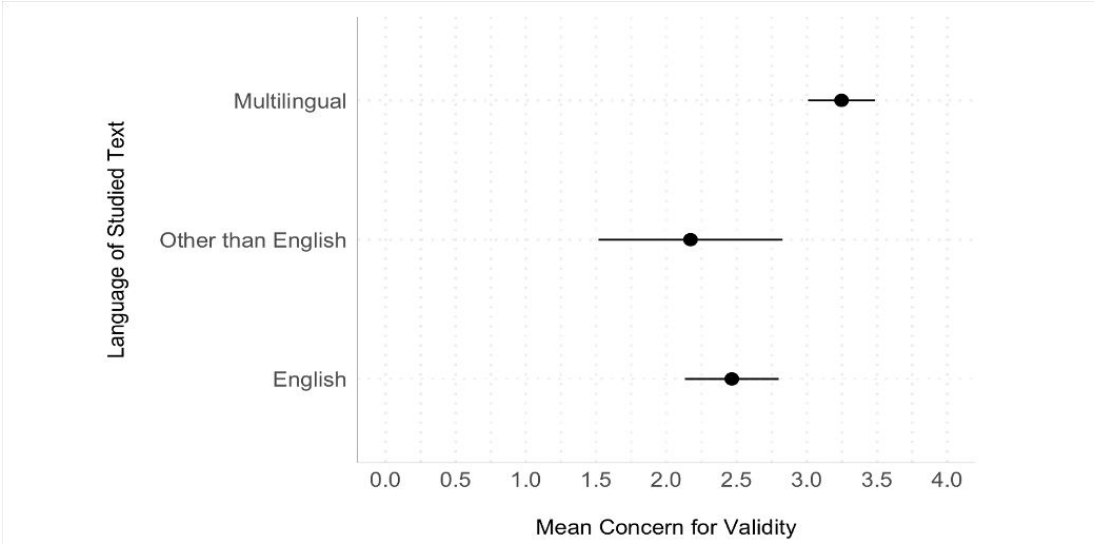
- Planning a research design for projects with multiple languages and multiple cases
- including validation strategies!

Equivalence and validity in multilingual computational text analysis

How to approach validation?

Validation concerns

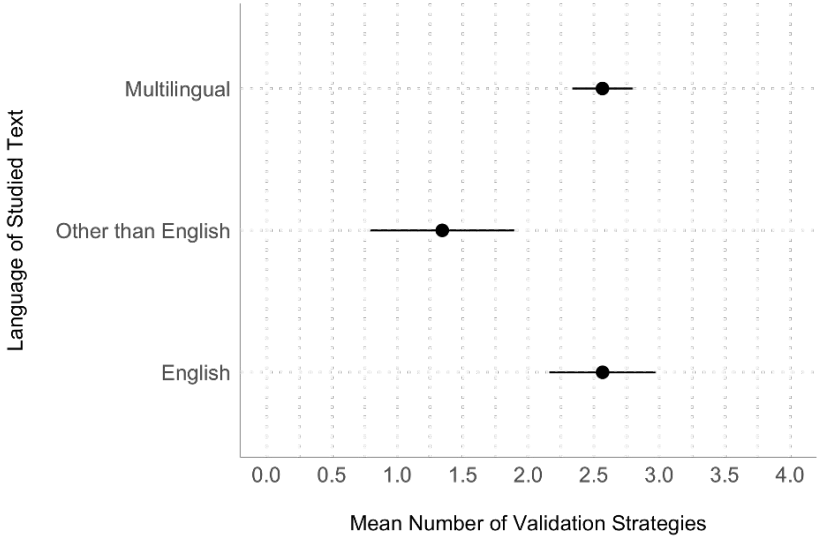
- Among researchers who published work that relies on quantitative text analysis, those who work in more than one language, express more concerns about the validity of findings from computational methods (Baden et al., 2022).



Baden, Dolinsky, Lind, Pipal, Schoonvelde, Shababo, & van der Velden, 2022

Validation strategies

- But this is not reflected in a more extensive focus on validation (Baden et al., 2022)



Baden, Dolinsky, Lind, Pipal, Schoonvelde, Shababo, & van der Velden, 2022

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

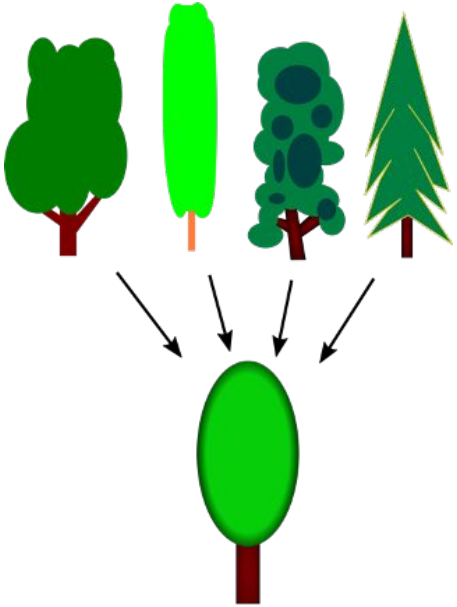
Emic and etic

- Two approaches to comparability
- Two ends of a continuum
- The positioning of the own research project (of constructs and measurement) on this continuum helps to plan the comparative research design and especially an appropriate validation method.

Emic	Etic
define a construct case-specific	reach a ‘meta-theoretical’ understanding of a construct
measure the construct with case-specific instruments and procedures	measure construct with standardized Instruments and procedures
may hinder comparison between cases	may overlook specific cultural perspectives

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

Example 1

Etic concept definition:

Migration

Migration' is understood as a generic term and thus stands equally for migration, emigration and immigration. 'Migrant/s', refers to people that explicitly changed, change or will/might change their place of residence from one country to another.".

Emic case sensitive measurement

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

Emic and etic

- Can you think of a research question where an emic approach and of another where an etic approach would be preferred?

Emic	Etic
define a construct case-specific	reach a ‘meta-theoretical’ understanding of a construct
measure the construct with case-specific instruments	measure construct with standardized instruments
may hinder quantitative comparison between cases	may overlook specific cultural perspectives

Overview about challenges and main solution approaches

Challenges and solutions

1. on a linguistic level
2. on a contextual level

Challenges and solutions

1. **on a linguistic level**
2. on a contextual level

A key challenge

- Moving from raw texts to quantitative text representations applying the same procedures as in monolingual scenarios is little useful for the subsequent process and output stage

- Why? We may pick up language differences instead of substantially more interesting patterns

Illustration 1 (Part 1)

- Four example sentences as illustration for a multilingual corpus

	text	target label
Doc1	Asylum seekers are no burden on the social system	welfare
Doc2	Asylsuchende belasten das Gemeinwesen nicht	welfare
Doc3	What we’re seeing is welfare tourism	welfare
Doc4	Das System der Einschüchterung führt zu mehr Gewalt	security

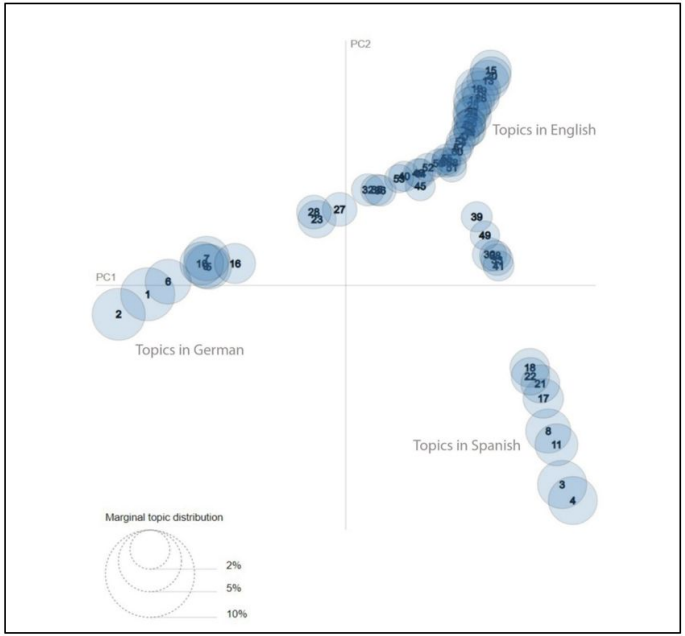
Illustration 1 (Part 2)

- bag-of-words representations of the four example sentences

	asylum	seekers	are	no	burden	on	the	social	system	asylsuchende	belasten	das	gemeinwesen	nicht	what	we‘re	seing	is	welfare	tourism	das	der	einschüchterung	führt	zu	mehr	gewalt
Doc1	1	1	1	1	1	1	1	1	1																		
Doc2										1	1	1	1	1													
Doc3															1	1	1	1	1	1							
Doc4									1												1	1	1	1	1	1	1

Illustration 2

- LDA topic model applied to English, Spanish, German documents
- Topics are very much clustered into languages
- Not useful to deliver topics that span across languages which allow the direct numerical comparison of cases



Lind et al., 2022, Appendix, p.6

Objective on measurement level

- Striving for measurement equivalence across languages

  = equivalence on a semantic level
- **Documents that indicate the same concept should receive sufficiently similar measurements independent from their language**

- Additional efforts are necessary!

How to jointly analyze documents in different languages?

Two solution approaches

1. Separate analysis
2. Input alignment

1. Separate analysis

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

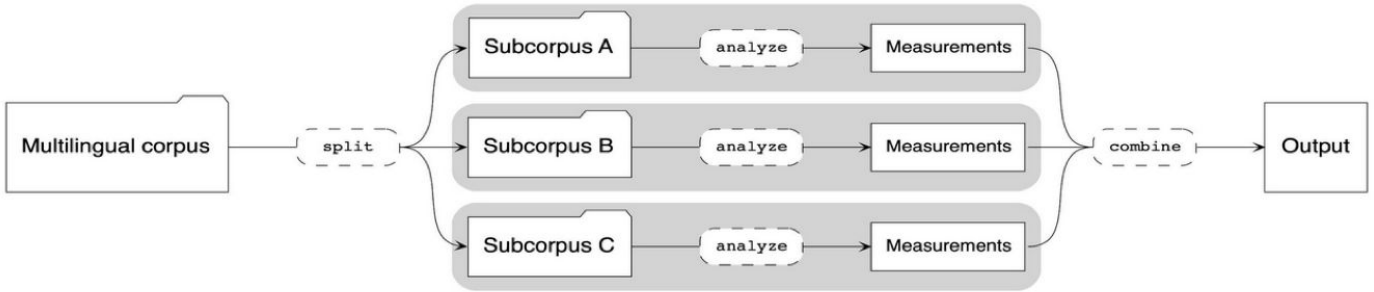


Figure 1 Illustration of the separate analysis strategy to multilingual text analysis.

Licht & Lind, 2023

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

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

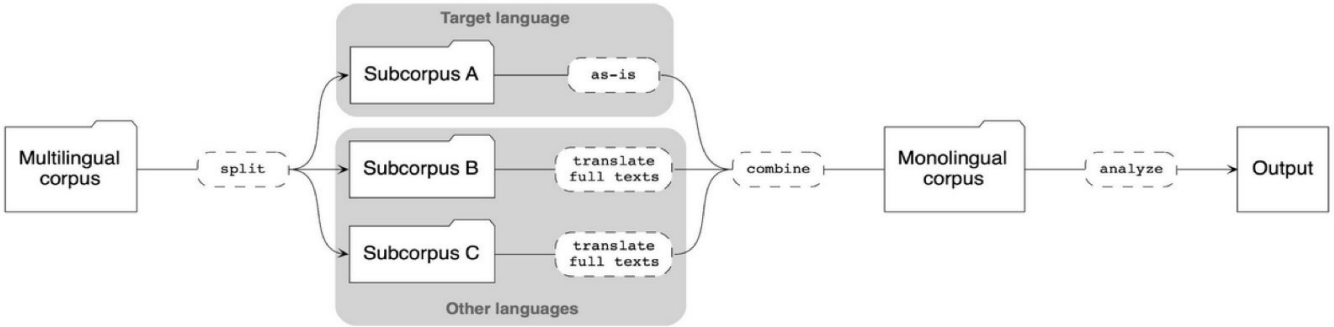


Figure 2 Illustration of the full-text translation approach to input alignment

Licht & Lind, 2023

2. Input alignment

Option 1: (Machine) Translation

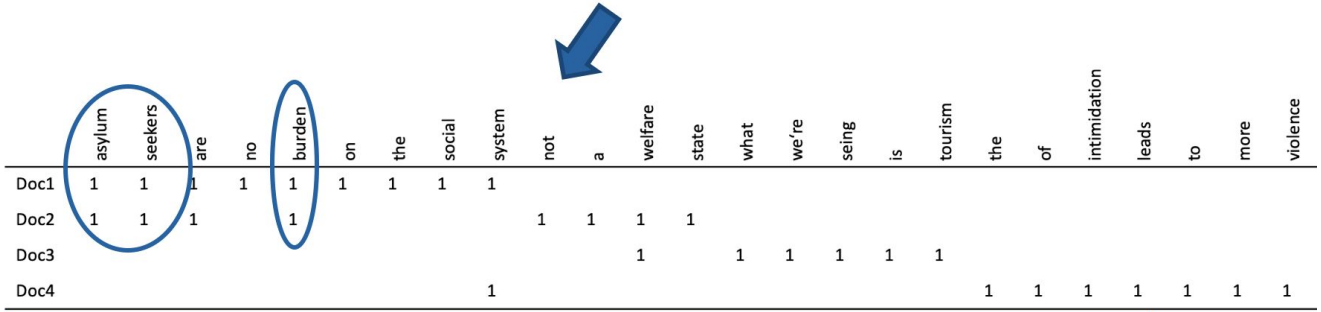
	text	text (english version)	target label
Doc1	Asylum seekers are no burden on the social system	Asylum seekers are no burden on the social system	welfare
Doc2	Asylsuchende belasten das Gemeinwesen nicht	Asylum seekers are not a burden on the welfare state	welfare
Doc3	What we're seeing is welfare tourism	What we're seeing is welfare tourism	welfare
Doc4	Das System der Einschüchterung führt zu mehr Gewalt	The system of intimidation leads to more violence	security

Licht & Lind, 2023

2. Input alignment

Option 1: (Machine) Translation

	text	text (english version)	target label
Doc1	Asylum seekers are no burden on the social system	Asylum seekers are no burden on the social system	welfare
Doc2	Asylsuchende belasten das Gemeinwesen nicht	Asylum seekers are not a burden on the welfare state	welfare
Doc3	What we're seeing is welfare tourism	What we're seeing is welfare tourism	welfare
Doc4	Das System der Einschüchterung führt zu mehr Gewalt	The system of intimidation leads to more violence	security



	asylum	seekers	are	no	burden	on	the	social	system	not	a	welfare	state	what	we're	seing	is	tourism	the	of	intimidation	leads	to	more	violence
Doc1	1	1	1	1	1	1	1	1	1																
Doc2	1	1	1		1					1	1	1	1												
Doc3												1		1	1	1	1	1							
Doc4									1										1	1	1	1	1	1	1

words from different languages that express the same meaning are now indicated by more similar numerical text representation

2. Input alignment

Option 2: Multilingual embeddings

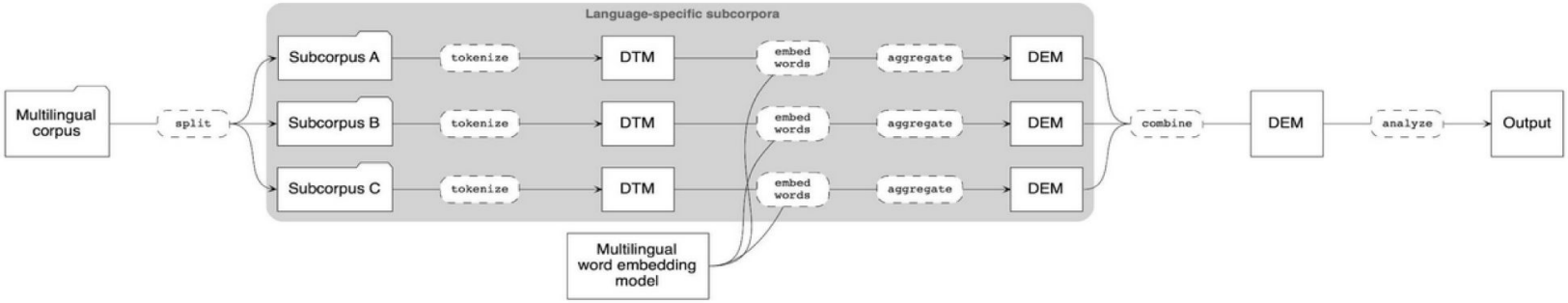


Figure 4 Illustration of the multilingual word embedding approach to input alignment.

Licht & Lind, forthcoming

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.

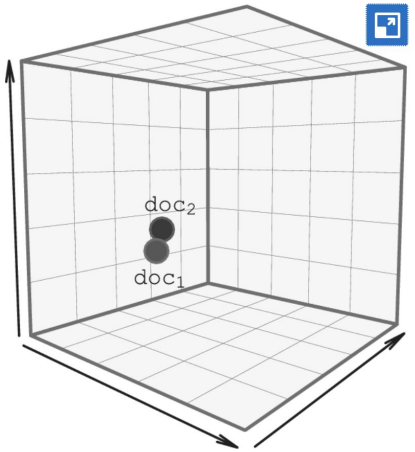


Figure 1 Schematic depiction of multilingual sentence embedding of example sentences in Table 1. *Note:* Depicting embedding in three dimensions serves illustrative purposes only.

Licht, 2022

How to decide between the approaches?

Discuss and collect decision criteria, pros and cons

1. Separate analysis
2. Input alignment

How to decide between the approaches

Some criteria

- Resource availability
  a. “human” resources: coders (e.g., for validation or collecting training data)
  b. “instruments:” dictionaries, pre-trained models
  c. computing (especially when using open-source models)
- Concept types
  a. context-dependent ⇒ separate analysis better to account for specifics
  b. latent concepts ⇒ (extra-textual) context matters

How to decide between the approaches

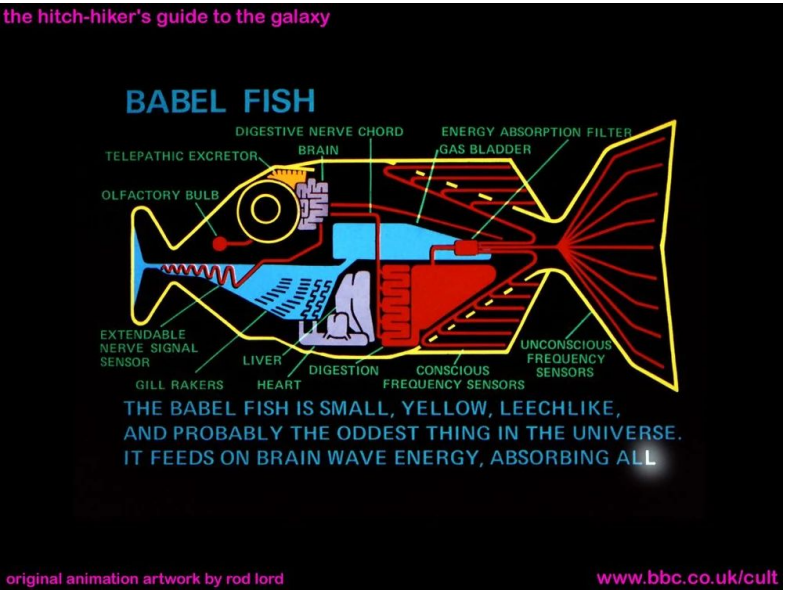
Some criteria

- Controllability and transparency: how much can we influence and validate several steps in the measurement process
  a. translation vs. embeddings
  b. open- vs. closed-source models
- Replicability

Challenges and solutions

1. on a linguistic level
2. **on a contextual level**

“Probably the oddest thing in the Universe.”



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

One interpretation:

- the babel fish is a parody by Adams of the implausibility of the translation machines described in science fiction literature
- In the novel, solving the language confusion with the Babel fish alone is not enough for mutual understanding (in the Galaxy there is no lack of wars)
- More is necessary: Context is key!

What could the following sentence mean?

“You have a green light.”

Example from: https://www.slideshare.net/miguelseura/pragmatic-linguistics

What could the following sentence mean?

“You have a green light.”

- It could mean that you have green ambient lighting
- It could mean that you have a green light while driving your car
- It could mean that you can go ahead with your project
- It could mean that you possess a light bulb that is tinted green
- Etc.

Example from: https://www.slideshare.net/miguelseura/pragmatic-linguistics

Semantics vs. pragmatics

Semantics = literal meaning of words, sentences or documents

Pragmatics = the contextual meaning of words, sentences or documents

- As social scientists, we are typically interested in communication that happens in social situations
- Thus, when setting up the empirical design, we ideally include our social science empowered contextual knowledge about the communicators and the audiences in each step

Aruna, 2018; Gurevitch and Blumler, 2003

Relevance of taking context into account

Example:

- Research goal: measure salience of (sub)topics in the national migration discourses in two countries
- contextual factors are likely different in these two countries: e.g., social, political and economic systems, migration history, immigration and emigration statistics
- As a consequence, the substance of the migration discourses in these countries likely differs, too. Thus, no fully congruent vocabulary would be used to indicate the concept in each country.

Objective on measurement level

- Striving for measurement equivalence across languages and across contexts

  = equivalence on a semantic level and on a pragmatic level
- **Documents that indicate the same concept should receive sufficiently similar measurements independent from their language and from their contexts**

- Additional efforts are necessary!

Main solution approach

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

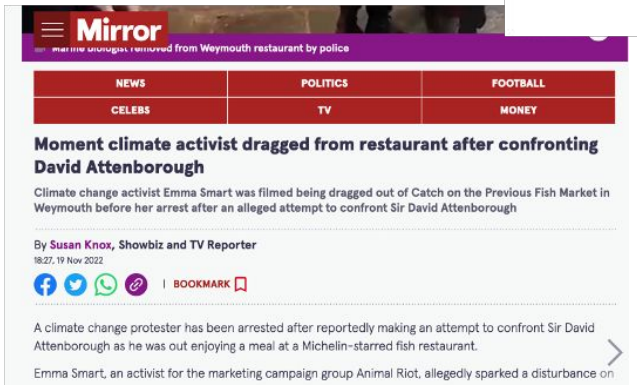
When:

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

Multilingual search string/keyword selection and testing

Case study: Climate Activism

- Comparing the media discourse about climate activism across several countries



Process

1. Evaluation of human expertise
2. Concept definition
3. Equivalent data selection
4. Search string validation

Process

1. Evaluation of human expertise

2. Concept definition

3. Equivalent data selection

4. Search string validation

Evaluation of human expertise

- Language expertise
- Case expertise
- Domain expertise

Within the team? Budget to hire experts?

Case study: Evaluation of human expertise

- Language expertise: French, German, English?
- Case expertise: France, Germany, Switzerland, UK?
- Domain expertise: climate change?

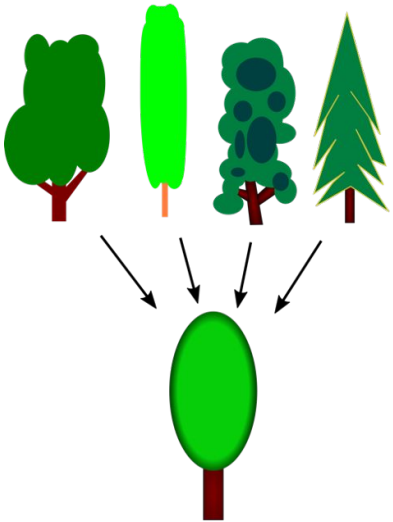
Process

1. Evaluation of human expertise
2. **Concept definition**
3. Equivalent data selection
4. Search string validation

Concept definition

Objective: selecting a target concept to be measured in text data

- Examples: Topics, sentiment, frames, uncivility…
- Reflection on emic/etic construct definitions
- With case and language experts recommended



Case Study: We are interested in …

Concept: climate activism

- Working definition: Climate activism can be defined as “mobilization of politically engaged participants—and other stakeholders to address climate challenges” (Bomberg, 2012, p.408)

Process

1. Evaluation of human expertise
2. Concept definition
3. **Equivalent data selection**
4. Search string validation

Equivalent data

Objective: finding units of analysis that are equivalent and thus comparable across cases (Rössler, 2012, p. 461).

Often two steps:

a) Finding equivalent document sources

b) Retrieval of equivalent documents

Illustration

a) Finding equivalent document sources

- Media sources selected on the basis of reach, genre, (and data availability)

Table A1.

Media Sources in the Data Set

Country	Source Type	Source
Germany	Print	Bild, Die Tageszeitung (taz), Die Welt, Frankfurter Allgemeine Zeitung, Frankfurter Rundschau, Süddeutsche Zeitung
	Online	bild.de, faz.net, spiegel.de, sueddeutsche.de, taz.de, welt.de, zeit.de
Hungary	Print	Magyar Hirlap, Magyar Idök, Nepszava
	Online	24.hu, blikk.hu, borsonline.hu, index.hu, magyarhirlap.hu, mno.hu, napi.hu, nepszava.hu, ripost.hu
Poland	Print	Dziennik Gazeta Prawna, Gazeta Wyborcza, Rzeczpospolita
	Online	fakt.pl, gazeta.pl, onet.pl, rp.pl, se.pl, wp.pl, wyborcza.pl
Romania	Print	Evenimentul Zilei, Jurnalul National, Romania Libera, Ziarul Financiar
	Online	adevarul.ro, click.ro, evz.ro, jurnalul.ro, libertatea.ro, romanialibera.ro, zf.ro, ziare.com

Case study: Selecting equivalent document sources

- Jointly discuss and select two main traditional news sources for the cases
- France, Germany, Switzerland, UK

Illustration

b) Retrieval of equivalent documents

- Approach: ‘Etic’ concept definition of’ migration
- Retrieval of a multilingual news article sample with ‘emic’ search string (i.e., a multilingual dictionary), selection of ‘functionally equivalent’ keywords

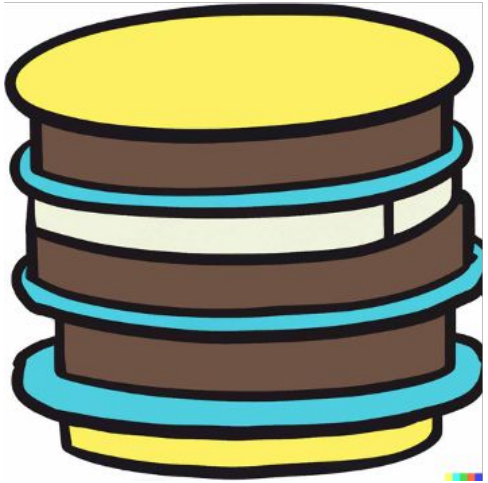
REMINDER

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*"

Case study: Selecting documents – Keywords per case

- Our goal is now to form search strings per case that are comparable across cases (and language).
- Ideas for relevant keywords that can be used in a search string to select news articles on climate activism
- What keywords should we select per case and language?



Let’s code

- Building search strings
- You find an R Markdown file “multilingual_dictionaire.Rdm” in the GitHub folder code.

Process

1. Evaluation of human expertise
2. Concept definition
3. Equivalent data selection
4. **Search string validation**

Equivalent output

Objective: Ensure that the obtained measures are equivalent across languages and across cases and of high quality

Strategies:

- Compare estimates with an established benchmark
- examine recall and precision as well as the corresponding misclassifications
- output validation needs to be considered for each included language and case

Benchmark types

- self-created baseline, often manually labeled documents (convergent validation)
- variables known to measure the same concept (convergent validation)
- variables known to measure concepts that differ (discriminant validation)

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

reliability across languages (we missed to do also a test across cases!)

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.

REMINDER

reliability within language/case (also not ideal: better 2 coders per case/language who are familiar with case and can code original language)

Lind et al., 2019, Appendix

Case Study: Develop an output evaluation strategy

- We will now label jointly and create a human labeled benchmark
- Usually, we would probably decide on a concept relevant to measure for an already selected climate activism corpus.
- In order for us to all work on the same dataset and concept, we will go a step back and code the concept “climate activism”. With our search string we basically designed already classification instruments, which we can now validate
- Our data is stored on a Google sheet:

Let’s code

- Assign labels manually (assess intercoder reliability)
- Compare manual and automated measures
- Calculate recall and precision
- “Manual_baseline_creation.Rdm”

Machine Translation

History

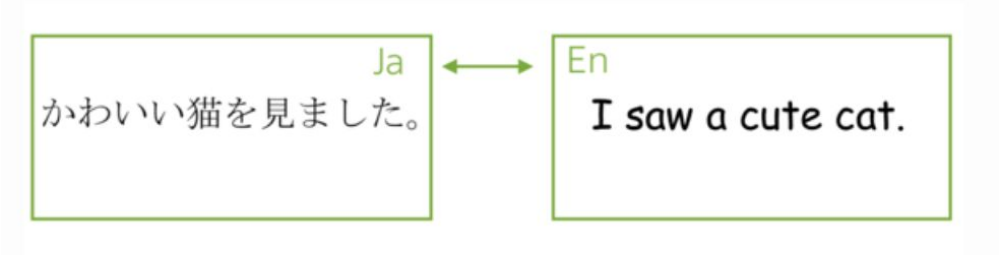
until recently, **machine translation** (MT) has been the default in multilingual computational text analysis

- several methods papers evaluate its reliability and validity for bag-of-words analysis (e.g., Lucas et al., 2015; de Vries et al., 2018, Mate et al., 2023) and embedding-based analysis (e.g., Mate et al., 2023)
- lots of applications in substantive comparative research

How it works (note: all illustrations by Lena voita)

MT is a **sequence-to-sequence** NLP task: we want to transfer

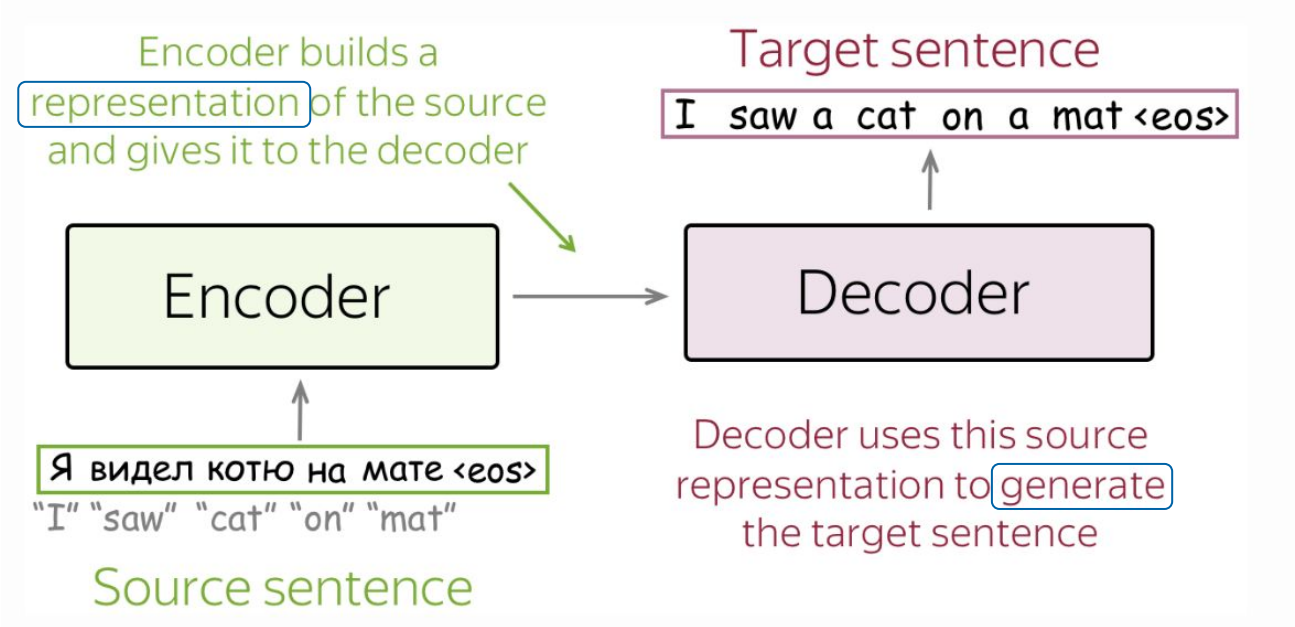
- a sequence of words in one language (the source)
  into
- a sequence of words in another language (the target)



How it works

The state-of-the-art is training a **decoder-encoder** neural network

- using many sentence pairs
- covering many languages
- Technical background see blog by Lena voita



Does it work for applied research?

Several pol/comm sci papers have evaluated NMT for bag-of-words analysis

Reference	Task	Domain	Translation service	Source language(s)	Target lang.
Lucas et al. (2015)	Topic modeling (STM)	Citizen-produced social media	Google Translate	Arabic, Chinese	English
Vries et al. (2018)	Topic modeling (LDA)	Parliamentary speech	Expert translations, Google Translate	Danish, French, German, Spanish, Polish	English
Reber (2019)	Topic modeling (LDA)	Web pages of (I)NGOs	Google Translate, DeepL	German	English
Windsor et al. (2019)	Dictionary analysis (LWIC)	UN plenary speeches	Google Translate	English, French, German, Russian, Chinese, Arabic	English
Düpont and Rachuj (2021)	Textual similarity	Party manifestos	Google Translate	12 languages [a]	English
Courtney et al. (2020)	Supervised classification	News article paragraphs	Google Translate	German, Spanish	English
Lind et al. (2021)	Supervised classification	News articles	Google Translate	German, Hungarian, Polish, Romanian, Spanish, Swedish	English
Licht (2023)	Supervised classification	Party manifestos	M2M (Fan et al. 2021)	12 languages [b]	English

[a] Catalan, Danish, Dutch, Finnish, French, Galician, German, Italian, Norwegian, Portuguese, Spanish, and Swedish

[b] same as Düpont and Rachuj (2021) by pooling their and data by Lehmann and Zobel (2018)

Does it work for applied research?

topic modeling:

- on average, document-topic and topic-word representations are very similar when comparing LDA topic models fitted to human- and machine-translated (*Google*) topics, respectively (de Vries *et al.*, 2018)
- Reber (2019) also evaluate DeepL and Google and finds similarly encouraging results

Does it work for applied research?

dictionary analysis:

- english dictionary applied to machine-translated corpus gives similar measurements as if applied to human-translated documents (Windsor *et al.*, 2019)
- but machine-translating of keywords is not a good idea (experts should check them; see Lind *et al.*, 2019; Proksch *et al.*, 2019)

Does it work for applied research?

supervised classification

- language-specific classifiers trained on machine-translated texts (Google) perform as well as classifiers trained on texts in their source languages (holding dataset constant; Courtney *et al.*, 2020)
- Transformers fine-tuned on machine-translated labeled texts perform as well as multilingual Transformers (Mate *et al.*, 2023, Table 3) (but only evaluated for Hungarian)

Does it work for applied research?

- only minor differences between the measurements generated by different methods when applied to corpora translated with open-source models and commercial service (Licht et al., working paper)

Licht, H., Sczepanski, R., Laurer, M., & Bekmuratovna, A. (2024). *No more cost in translation: Validating open-source machine translation for quantitative text analysis* (No. 276). University of Bonn and University of Cologne, Germany.

How to

two options

- commercial services (Google, DeepL, AWS, Microsoft, etc.)
- “free” open-source NMT models:
  - Helsinki NLP’s OPUS-MT
  - Facebook Research’s M2M
  - some others, see here for example

How to

two options

- commercial services (Google, DeepL, AWS, Microsoft, etc.)
- “free” open-source NMT models:

Discuss: what are these options pros and cons?

Two options – pros and cons

Commercial services

pros:

- high-quality translations
- often many translation directions (especially Google)
- usability (API, interface available)
- language coverage

cons:

- costs (you pay per character, 1 million characters ~= 20 US Dollars)
- limit replicability
- Replicability (there might be an update to the model)
- Problematic for sensitivity data

Two options – pros and cons

Open-source models

pros:

- freely available for download and use ⇒ replicable
- transparent: we know what training corpora have been used
- usually good-enough translation quality

cons:

- limited translation directions (problem with “low-resource” languages)
- you need to know how to code (but that’s why you’re here ;)
- need access to a GPU

So, let's code

Notebook 'translation.Rdm" in
https://github.com/fabiennelind/Going-Cross-Lingual_Workshop

For code in Python (includes translation with open source models) see notebook 'translation_basics.ipynb' in code/ at https://github.com/fabiennelind/Going-Cross-Lingual_Course/tree/main

Multilingual topic modelling

Topic modeling in comparative research

- Objective: Topic extraction from document collections for comparative research

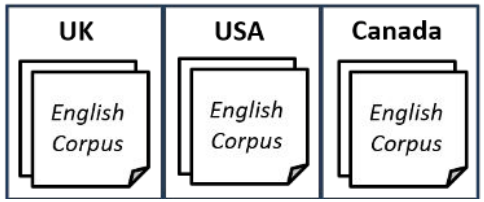
- Problem: Multilingual character of the data prevents direct application of “classic” topic modeling algorithms such as LDA

Aspects to consider

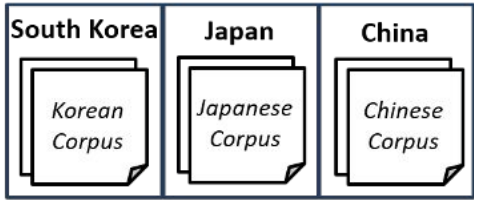
- Corpus
- Analysis Goal
- Comparability
- Resources

What is the corpus like?

Documents in one language

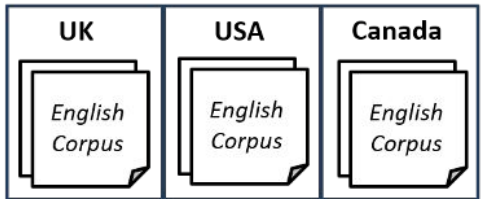


Documents in multiple languages



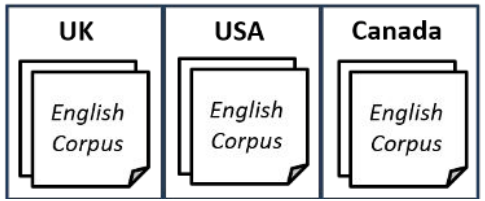
What is the analysis goal?

Identify case-specific topics



EMIC

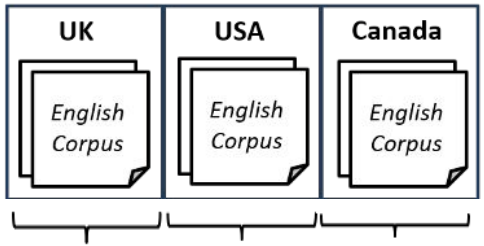
Identify meta-level topics across cases



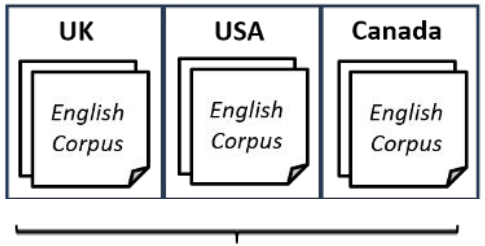
ETIC

What is the analysis goal?

Identify case-specific topics

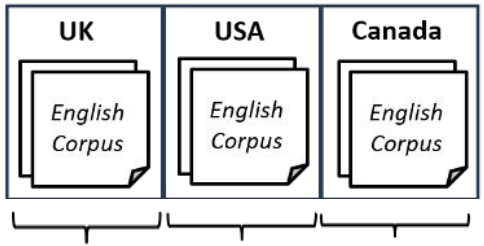


Identify meta-level topics across cases



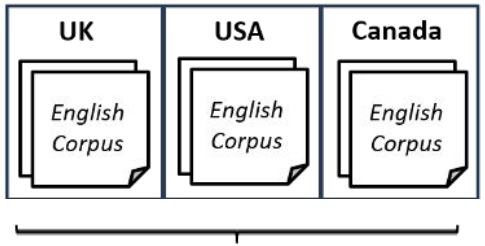
How to compare the results?

Identify case-specific topics



Qualitative comparison of the case-specific topics across cases

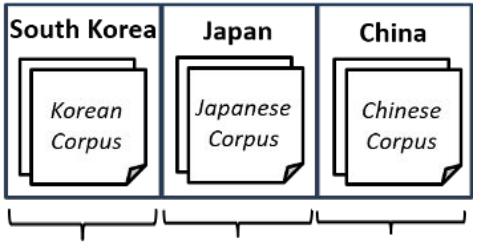
Identify meta-level topics across cases



Numerical comparison of topic scores across cases

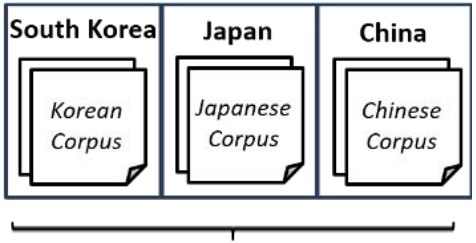
How to compare the results?

Identify case-specific topics



Qualitative comparison of the case-specific topics across cases

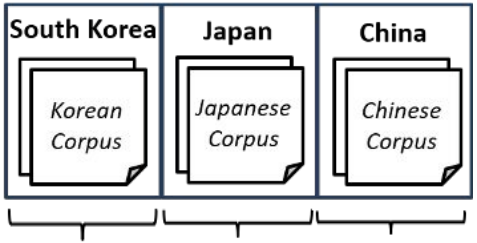
Identify meta-level topics across cases



Numerical comparison of topic scores across cases

Crucial resources

Identify case-specific topics



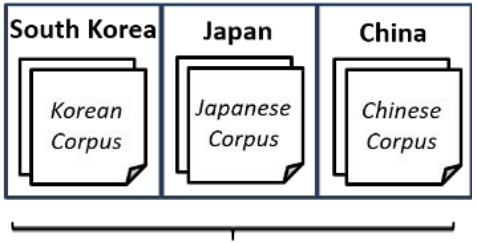
Run topic model for case 1 | Run topic model for case 2 | Run topic model for case 3

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

Language and case experts for labeling and interpretation

Qualitative comparison of the case-specific topics across cases

Identify meta-level topics across cases

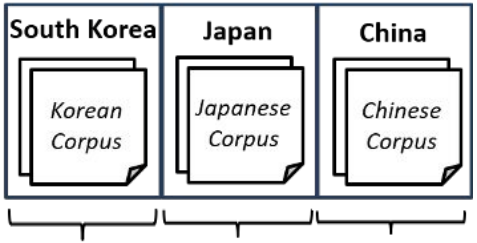


Run one topic model for all cases together

Numerical comparison of topic scores across cases

Crucial resources

Identify case-specific topics

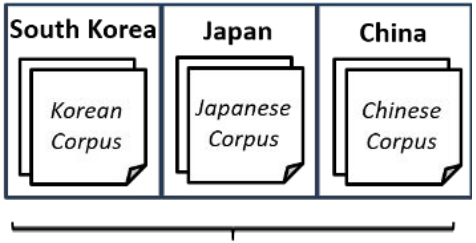


"Classic" topic modeling algorithms (e.g., LDA, BERTTopic)

Language and case experts

Qualitative comparison of the case-specific topics across cases

Identify meta-level topics across cases



Special linguistic resources necessary

Language and case experts

Numerical comparison of topic scores across cases

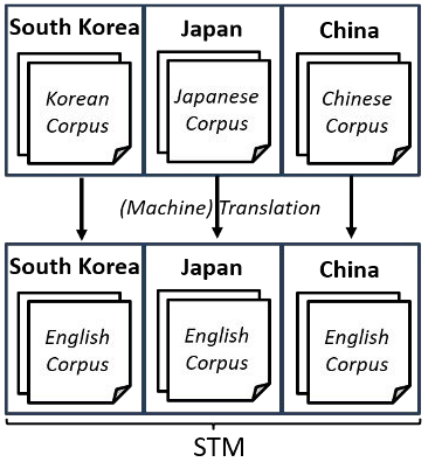
Techniques to identify meta-level topics across cases for a multilingual corpus

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

Code example Python: notebook ‘code/bertopic_multilingual.ipynb’ on https://github.com/fabiennelind/Going-Cross-Lingual_Course

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.

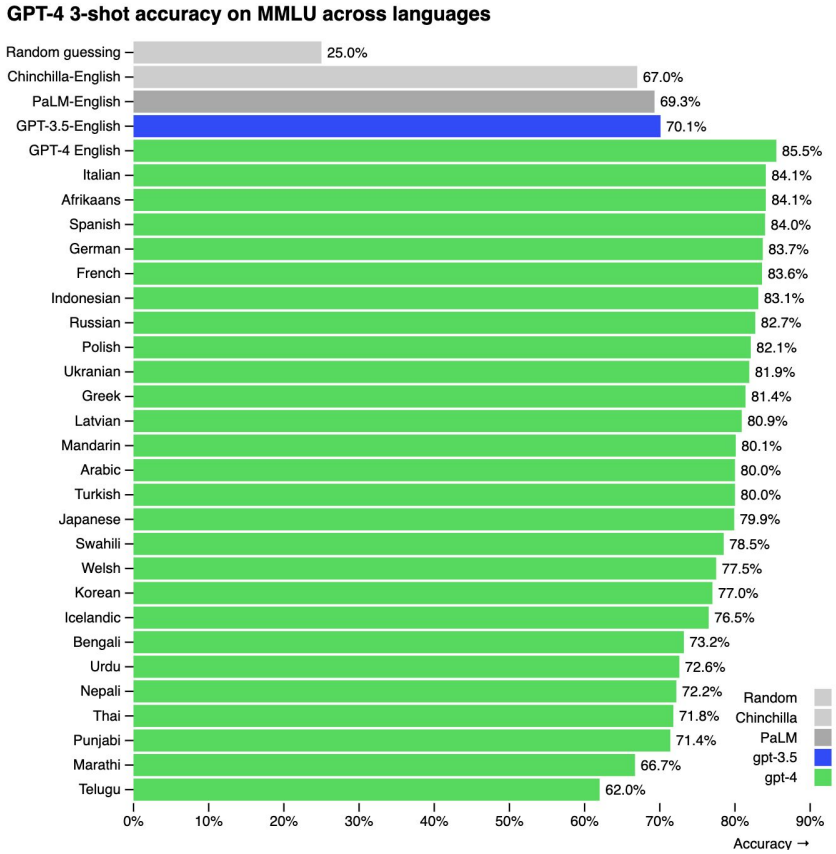


Figure 5. Performance of GPT-4 in a variety of languages compared to prior models in English on MMLU. GPT-4 outperforms the English-language performance of existing language models [2, 3] for the vast majority of languages tested, including low-resource languages such as Latvian, Welsh, and Swahili.

OpenAI. (2023). Technical Report.

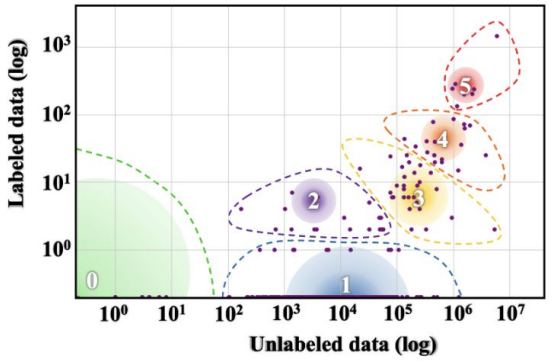
Questioning the “language agnostic” status of LLMs

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%

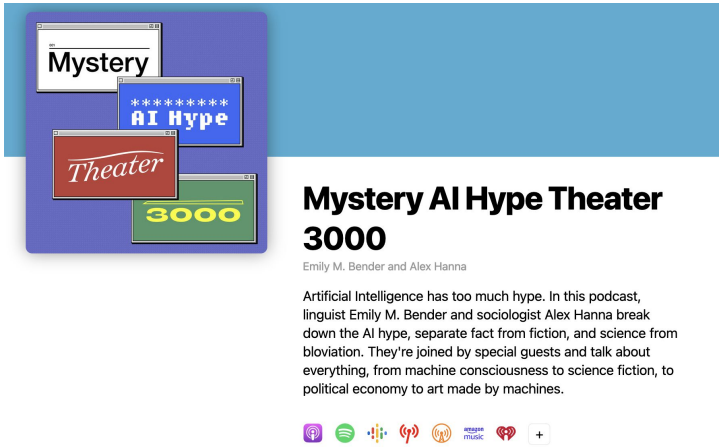
Table 1: Number of languages, number of speakers, and percentage of total languages for each language class.



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

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

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 GESIS course here

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

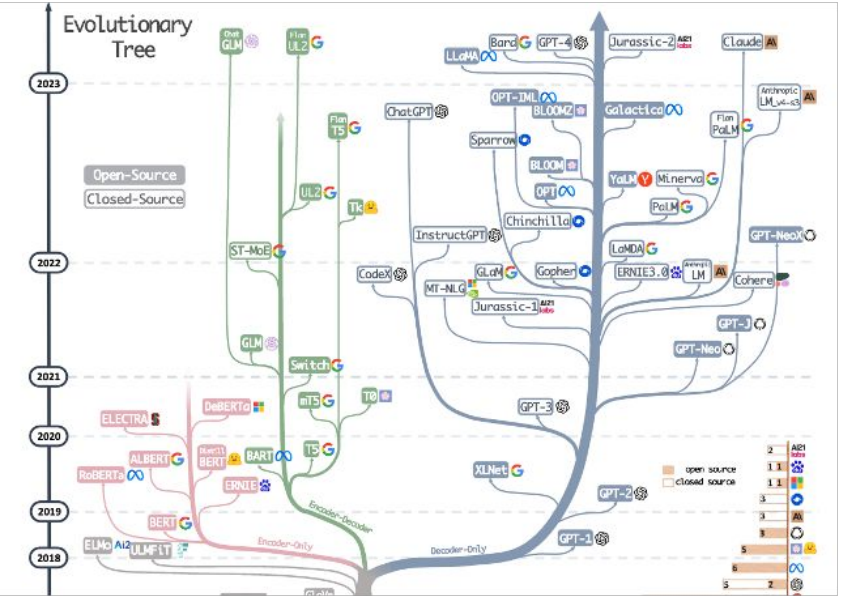
Function	Name	Authors	Countries	Languages
Geographical classification of text	Newsmap R package	Watanabe, 2018	240	12
Language and Location Code Convertor	ISOcodes R package	Buchta & Hornik, 2022	249	7000+
Obtain typological information (e.g., Phonology, Lexical semantics) about a language	World Atlas 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
Database that helps to link datasets on political parties	Partyfacts	Bederke, Döring, Regel	224	-

Conclusion

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

Comments, questions, thoughts

Time to discuss projects by participants 1:1

Thank you very much!