Going Cross-Lingual: Computational Methods for Multilingual Text Analysis

Hauke Licht and Fabienne Lind

Gesis Training Course

Cologne

6-8.12.2023

Introduction

About us

Hauke Licht

- Post-Doc, Cologne Center for Comparative Politics, University of Cologne
- Research focus: computational political science, electoral competition and democratic representation, political rhetoric, and multilingual text analysis
- hauke.licht@wiso.uni-koeln.de

Fabienne Lind

- Post-Doc, Computational Communication Science Lab, University of Vienna
- Research focus: Multilingual automated text analysis, opinion mining, migration and media, knowledge gap, comparative research
- fabienne.lind@univie.ac.at

We will share insights from

Joint publication:

- Going cross-lingual: A guide to multilingual text analysis

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

Individual research:

- multilingual supervised text classification, dictionary analysis, and topic modeling
- deep learning applications to computational political text analysis

Your turn :)

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

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

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 philosophie

Topics are covered with

- Lecture style input
- Guided coded sessions
- Plenum and small group discussions

Interrupt, ask all kinds of questions!

Individual consultations on your projects

- Very informal opportunity to talk about your use cases and (initial) design (plans)
- Research question, Data, Methods, Current struggles
- Dedicated slot: Friday afternoon

Workshop repository

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

Today

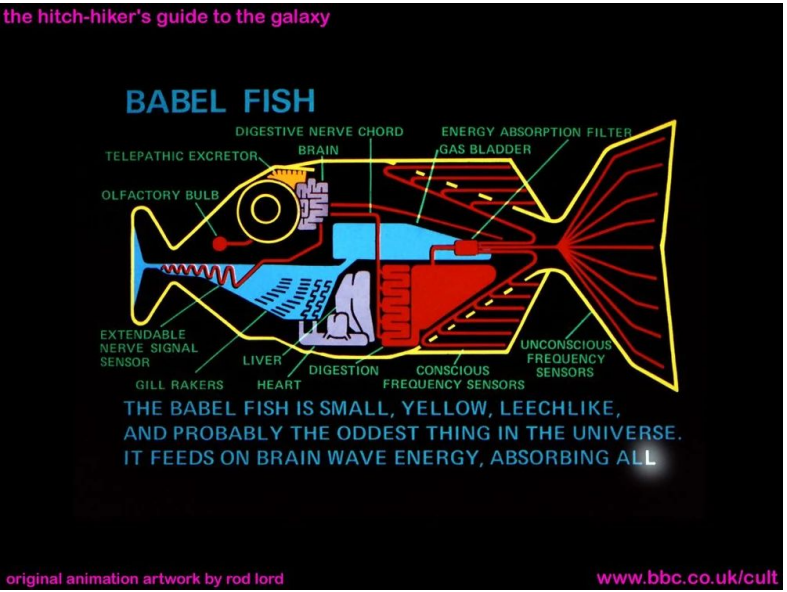
10:00 - 11:30	Introduction to the topic, overview about applications and main problems
11:30 - 11:45	*Coffee break*
11:45 - 13:00	Introduction to the main problems and solutions approaches, Validation framework
13:00 - 14:00	*Lunch break*
14:00 - 15:30	Valid corpus selection in multilingual & multi-context scenarios
15:30 - 15:45	*Coffee break*
15:45 - 17:00	Search string/keyword selection and testing, Creation of a validation benchmark

Introduction to the topic

What is the “Babel problem”?

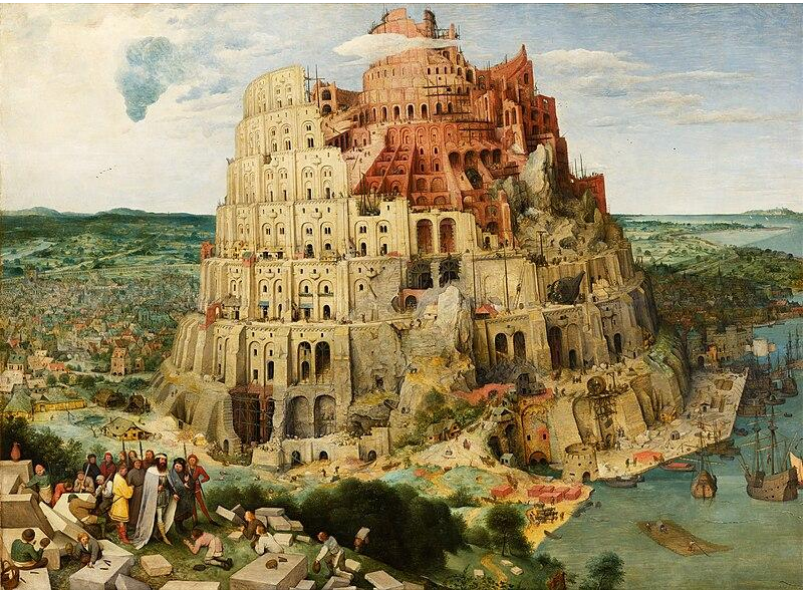
Anyone using this app?

“Probably the oddest thing in the Universe.”



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

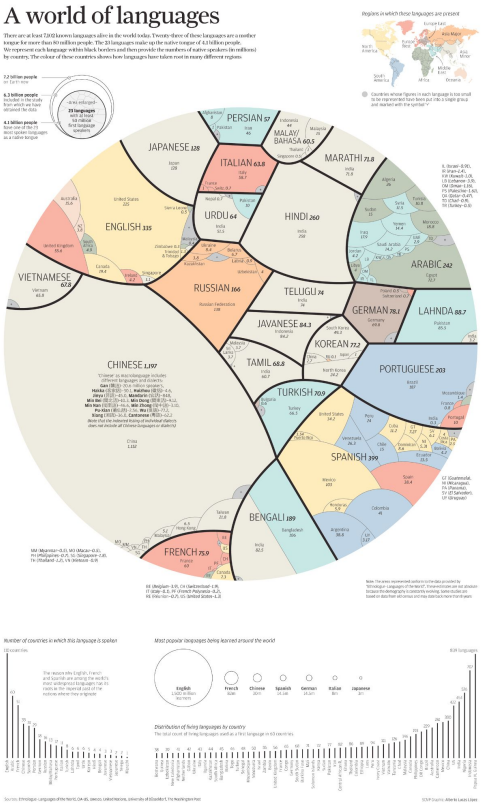
The tower of Babel



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

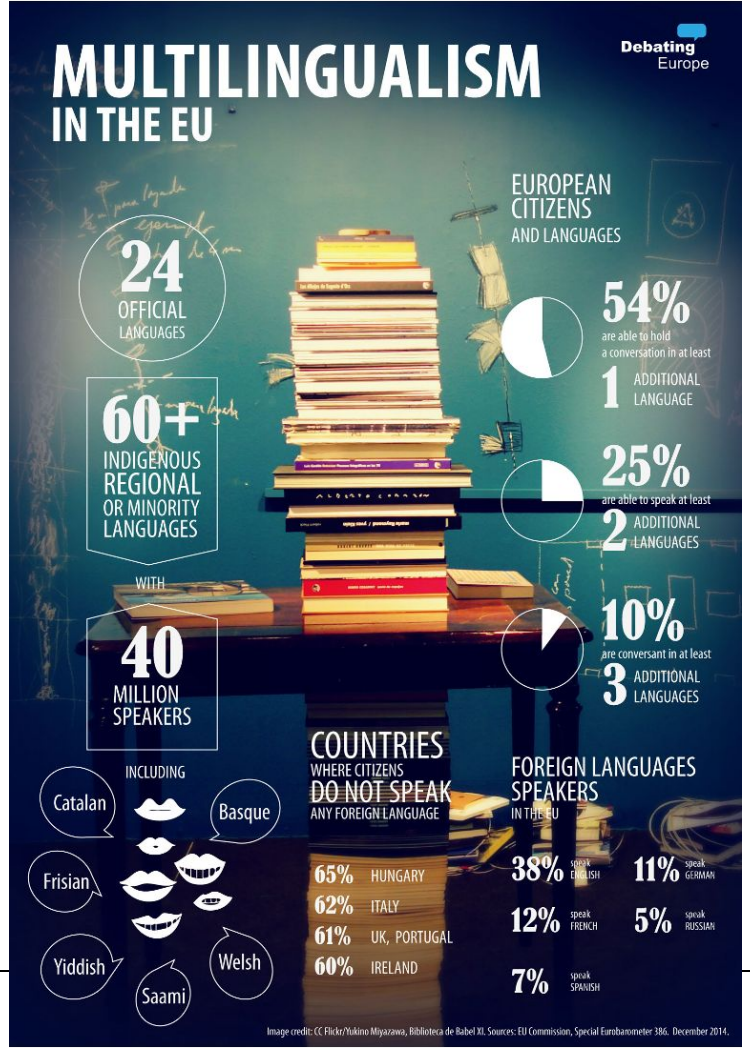
Global linguistic diversity is enormous…

- 7000+ languages are spoken globally today.



- 7.2 billion people on Earth
- 4.1 billion people have one of the 23 most spoken language as a native language

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.

Brainstorming

What languages should we include in our text analysis projects?

What are guiding criteria?

Overview about applications

Multilingual Computational Text Analysis?

- Wherever you seek to obtain measures from large amounts of text data written in at least 2 languages

Applications

In the social sciences:

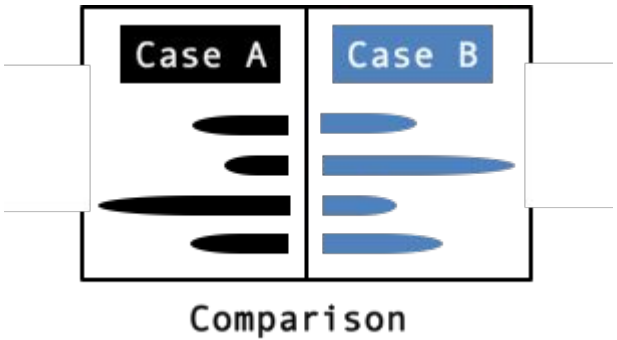
- Comparative research
- Towards more global, inclusive text analysis
- etc.

Comparative research

DALL.E

Comparative research

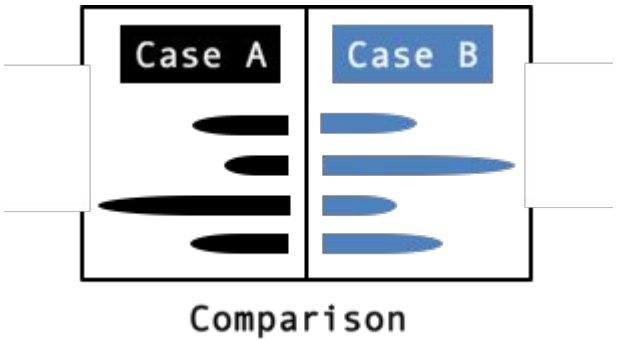
- **Comparative research in social science** involves comparisons between a minimum of two cases with at least one object of investigation relevant to social science research.



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

What are cases in comparative research?

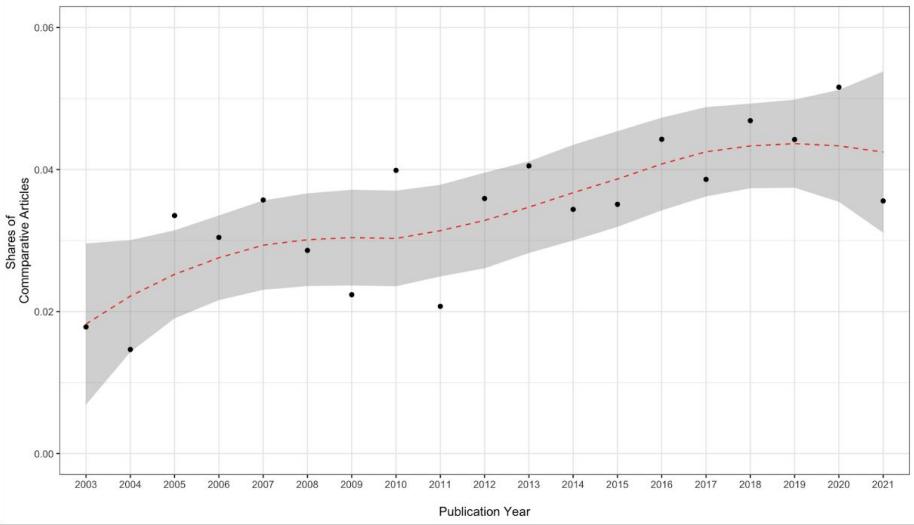
- **Cases** as we define them here are macro-level units such as systems, cultures, countries, and markets



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

Increasing popularity of comparative research

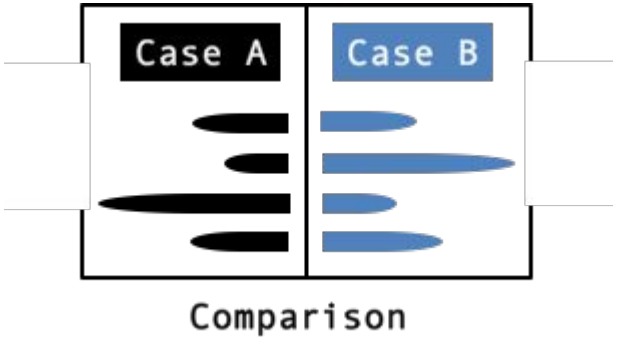
Shares of comparative articles in comm science journals, 2003-2021



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

Reasons to compare

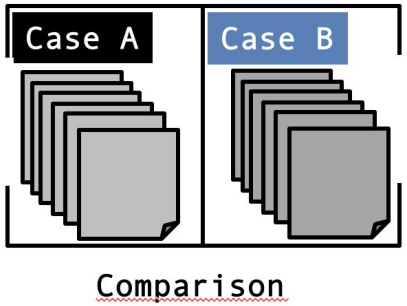
- Insights into the differences and similarities of cases
- Improved understanding and contextualization of the own case
- Raised awareness for other cases
- The test and generalizability of theories across diverse settings
- The investigation of transnational processes across contexts



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

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

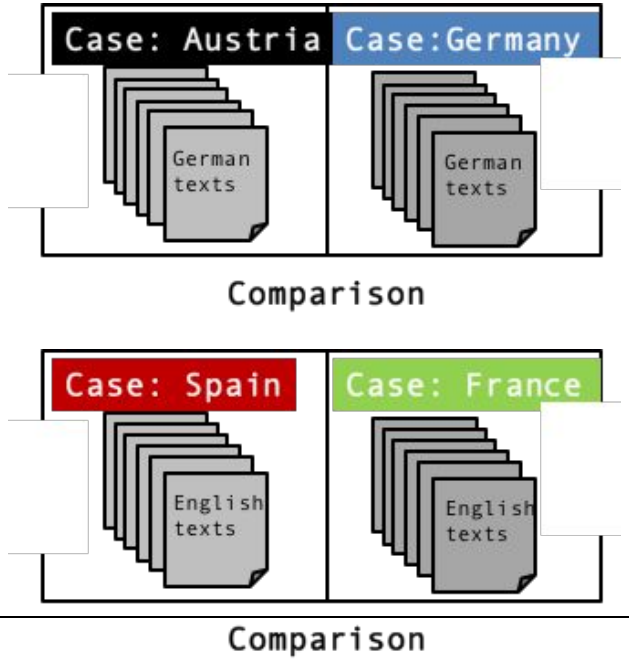


Boomgaarden & Song, 2019; McLeod &
Blumler, 1987; Esser & Vliegenthart, 2017; Esser & Hanitzsch, 2012; Livingstone, 2003

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

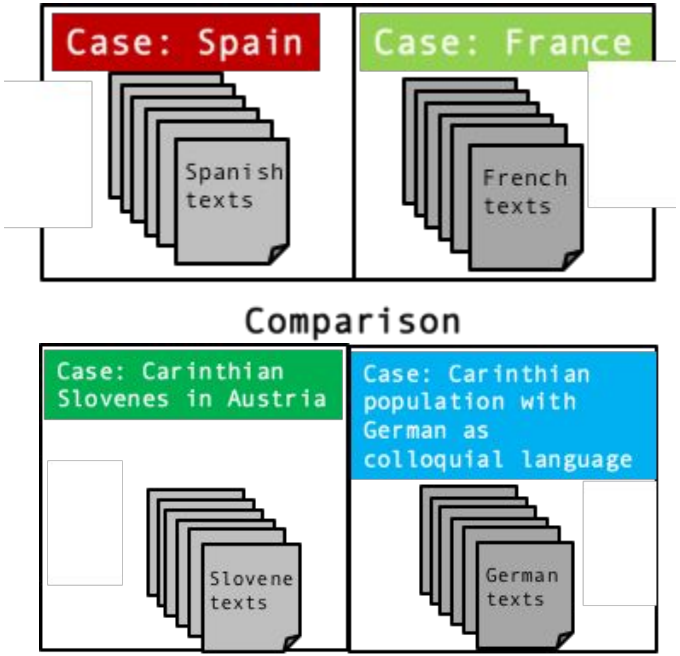
But the likely scenario is multilingual

- 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 only 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)

Automated content analysis as fast and reliable alternative to analyze large numbers of documents

Purpose of obtaining measures for a large number of documents

- Filter options
  - If measurements are available for a large number of data, other filtering options are possible. E.g.; linkage studies combination with media usage data
- Evidence-based policy making
  - Making the opinion of populations visible, holding politicians accountable

González-Bailón, 2013

Multilingual automated 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

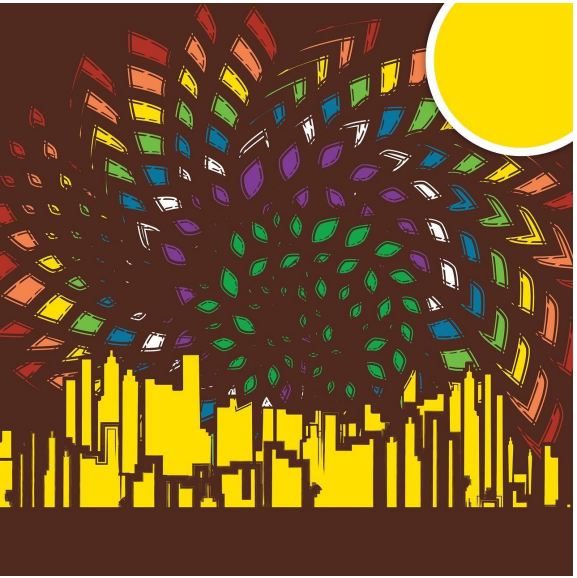


Image by ooceey from Pixabay

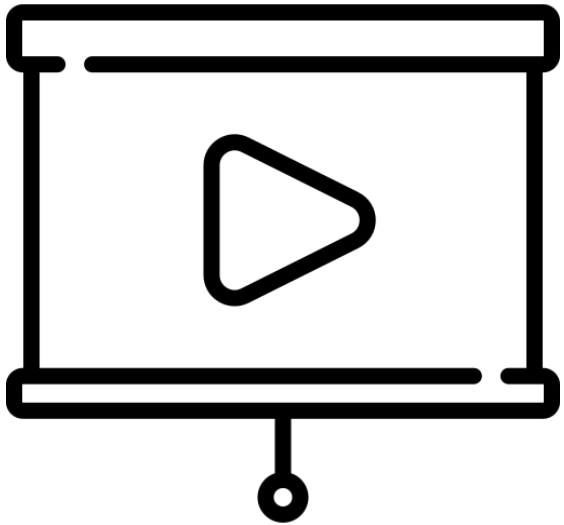
Progressing internationalisation

- 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 (including validation strategies!) for projects with multiple languages and multiple cases

Overview about challenges and main solution approaches

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

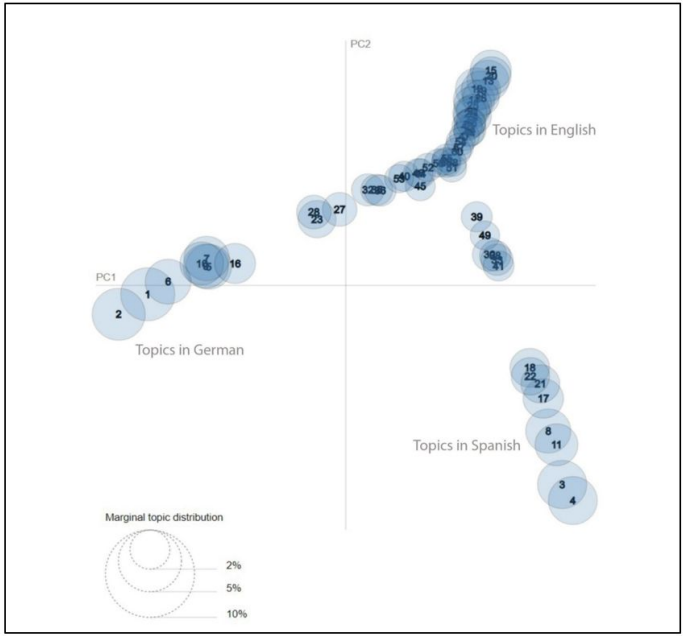
Illustration 1 (Part 2)

- 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 approaches

1. Separate analysis
2. Input alignment

1. Separate analysis

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

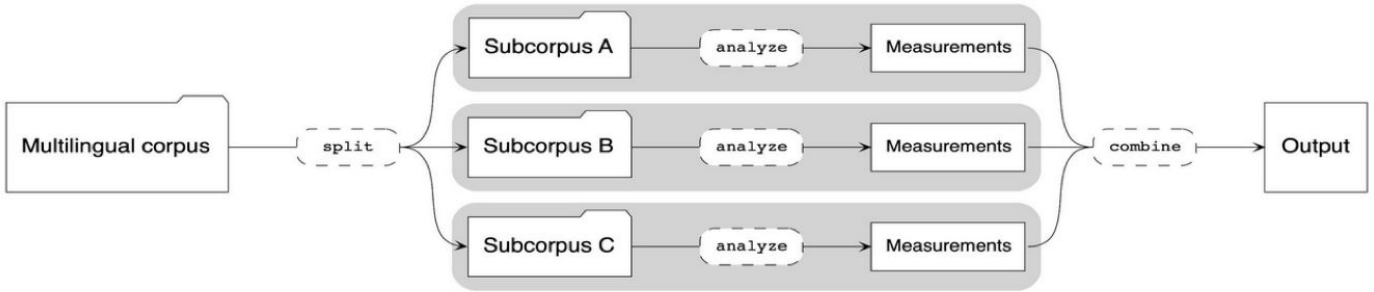


Figure 1 Illustration of the separate analysis strategy to multilingual text analysis.

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

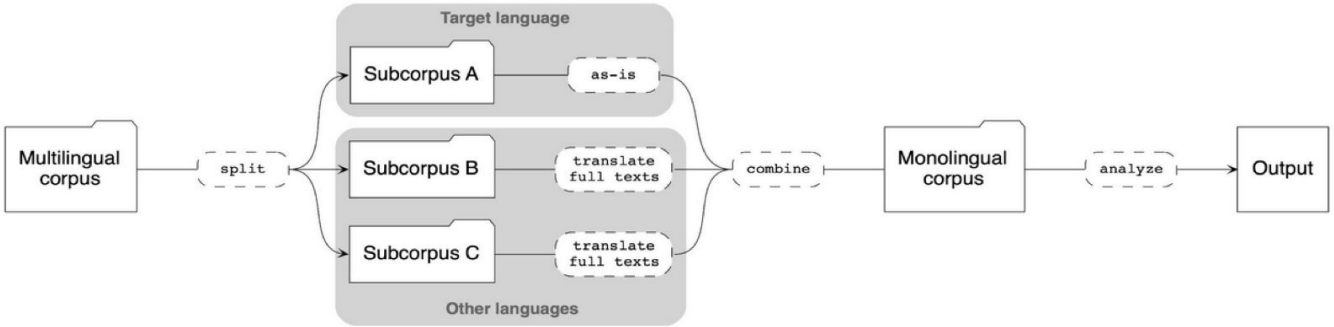


Figure 2 Illustration of the full-text translation approach to input alignment

Licht & Lind, 2023

2. Input alignment

Option 1: (Machine) Translation

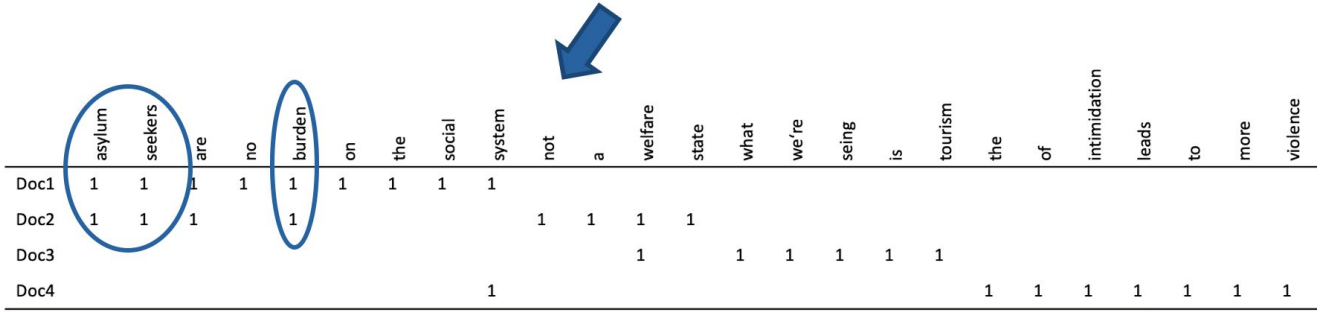
	text	text (english version)	target label
Doc1	Asylum seekers are no burden on the social system	Asylum seekers are no burden on the social system	welfare
Doc2	Asylsuchende belasten das Gemeinwesen nicht	Asylum seekers are not a burden on the welfare state	welfare
Doc3	What we're seeing is welfare tourism	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

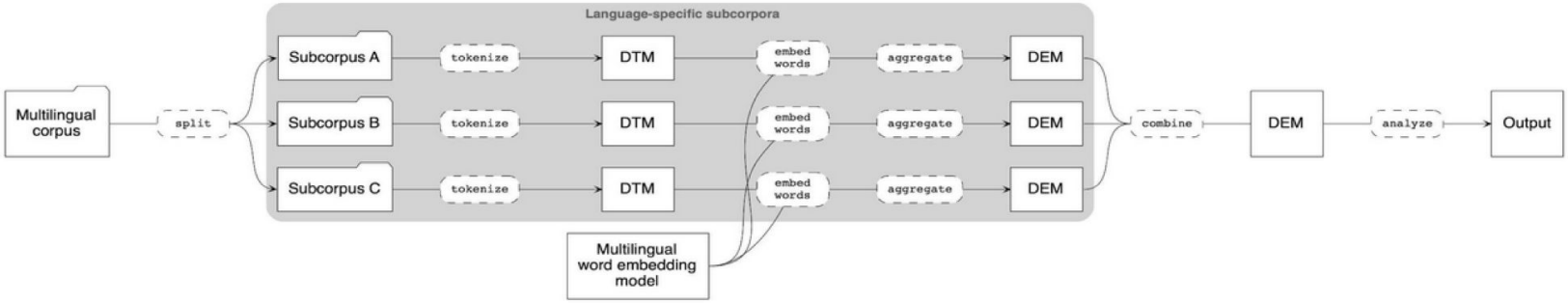


Figure 4 Illustration of the multilingual word embedding approach to input alignment.

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.

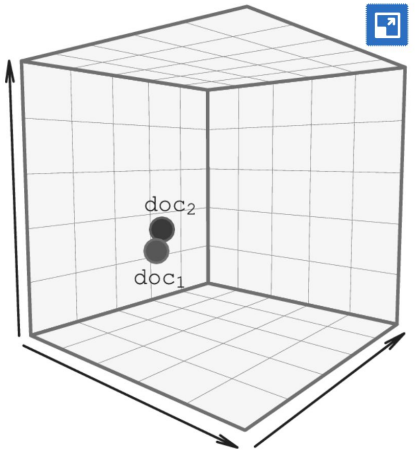


Figure 1 Schematic depiction of multilingual sentence embedding of example sentences in Table 1. *Note:* Depicting embedding in three dimensions serves illustrative purposes only.

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

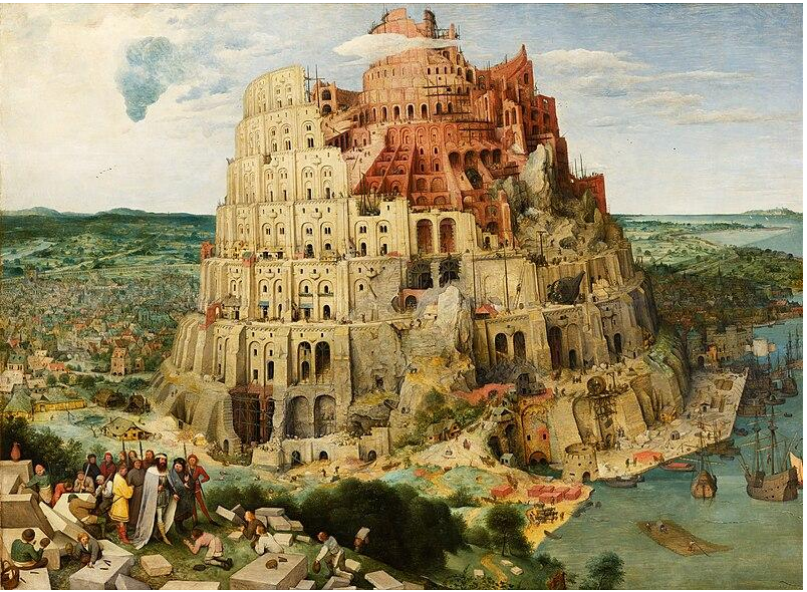
- resource availability
  a. “human” resources: coders (e.g., for validation or collecting training data)
  b. “instruments:” dictionaries, pre-trained models
  c. computing (especially when using open-source models)
- hard concepts
  a. context-dependent ⇒ separate analysis better to account for specifics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

How to decide between the approaches

Some criteria

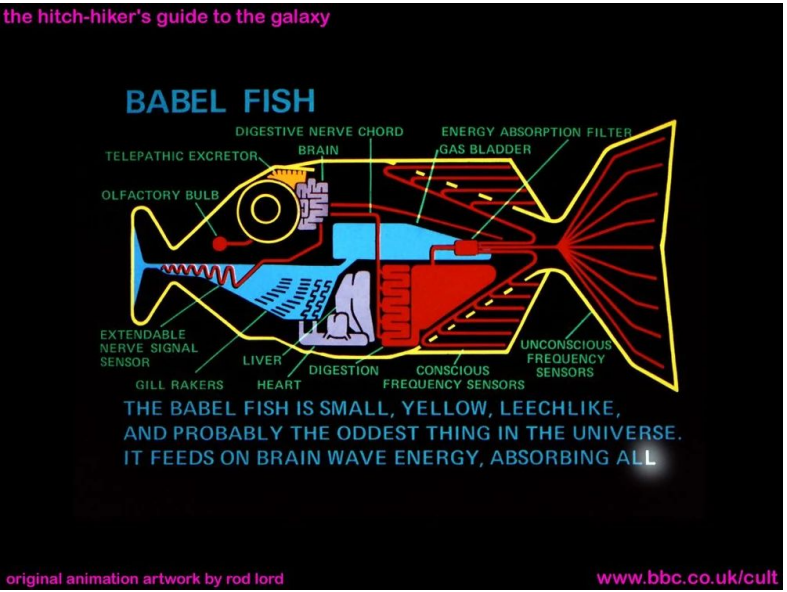
- controllability and transparency: how much can we influence and validate several steps in the measurement process
  a. translation vs. embeddings
  b. open- vs. closed-source models
- Replicability

The tower of Babel



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

“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!

Handling multilingual corpora in (social) contexts

Key challenge and solution approaches

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

What could the following sentences mean?

"I support socialized medicine."

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

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**

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!

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.

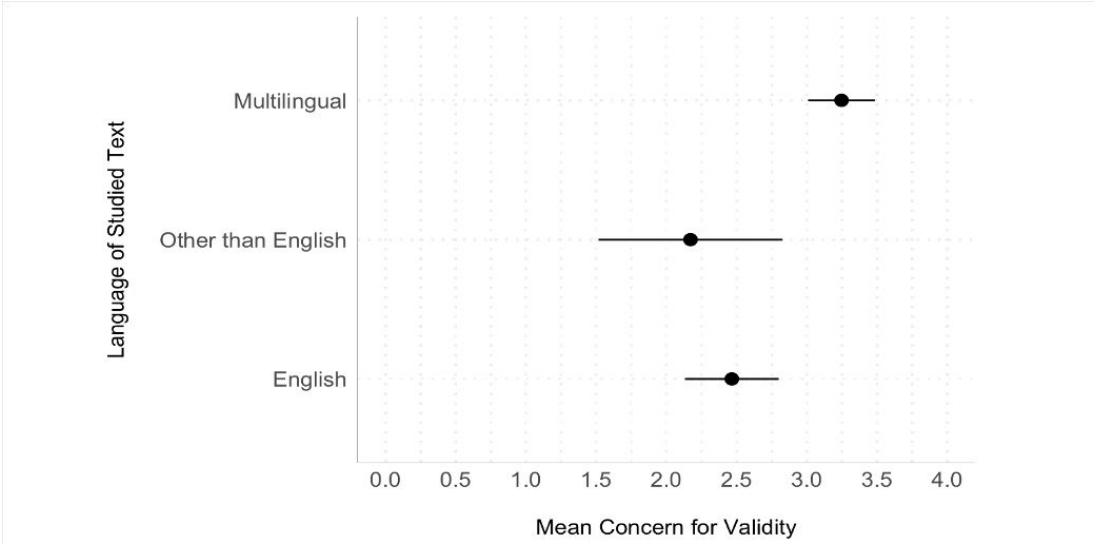
Equivalence and validity in multilingual computational text analysis: A validation framework

Motivations to design a framework

- insights from a content analysis of published literature in the social sciences and an expert survey with the respective authors (Baden et al., 2022)
- both studies conducted within the OPTED project

Validation concerns

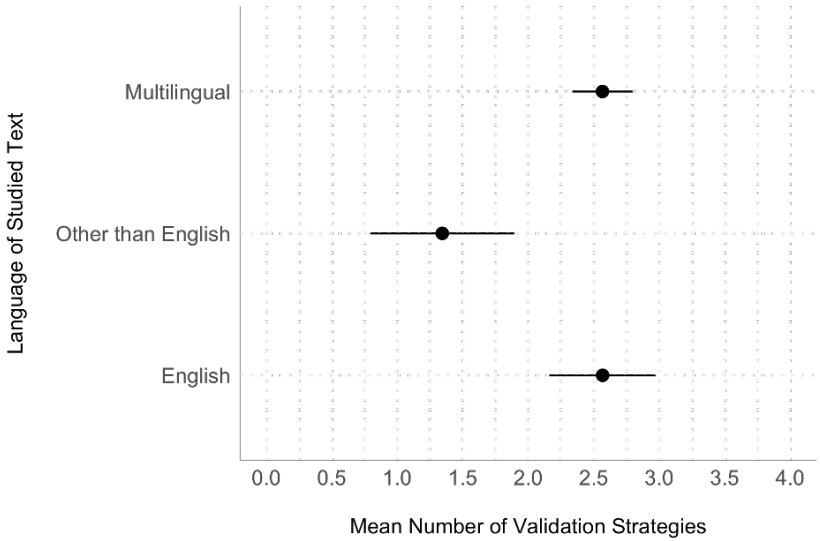
- Among researchers who published work that relies on quantitative text analysis, those who work in more than one language, express more concerns about the validity of findings from computational methods (Baden et al., 2022).



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

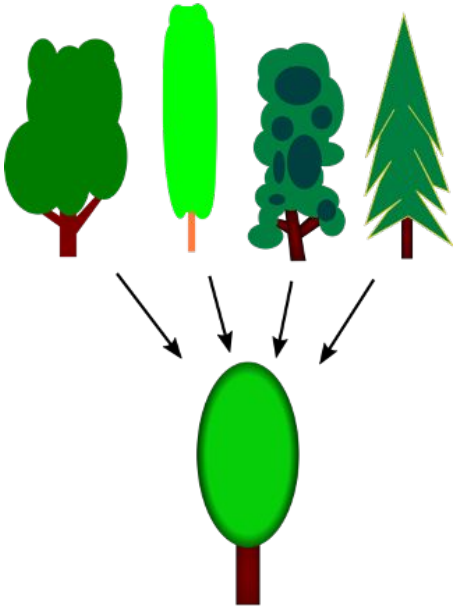
Emic and etic

- 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.".

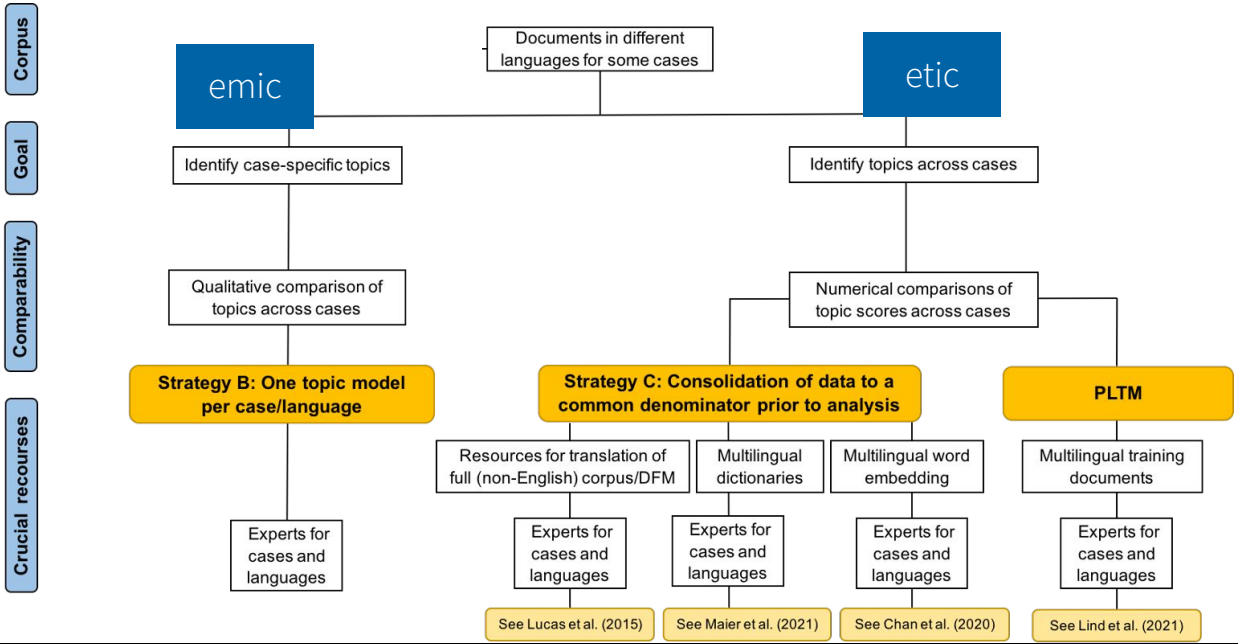
Emic case sensitive measurement

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: Example topic modeling approach



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

Wikimedia Commons

Illustrative example

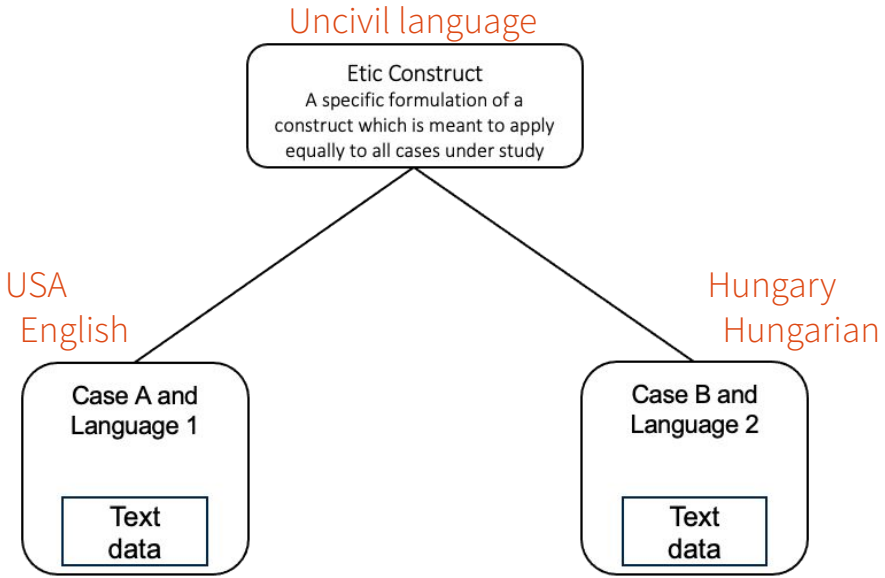
Wikimedia Commons

Wikimedia Commons

Example:

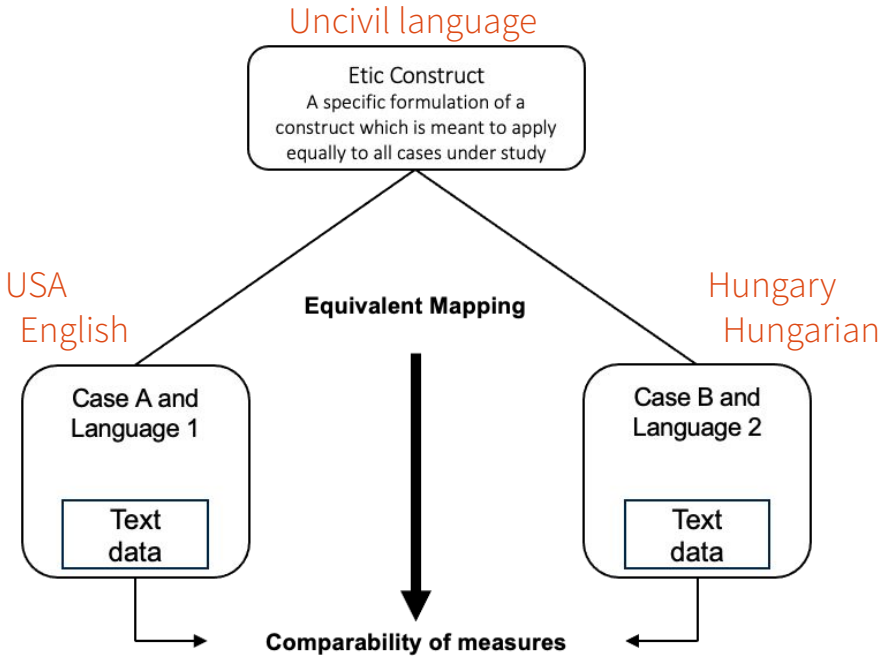
- Research objectives comparative research:
  - How similar/different are levels of uncivil language in Case A and Case B?
  - Do theories of deliberative democracy apply in Case A and in Case B?
- Etic concept definition: uncivil language in political communication as “disrespectful discourse that silences or derogates alternative views” (Jamieson et al. 2017, 206)

Validity in comparative research designs



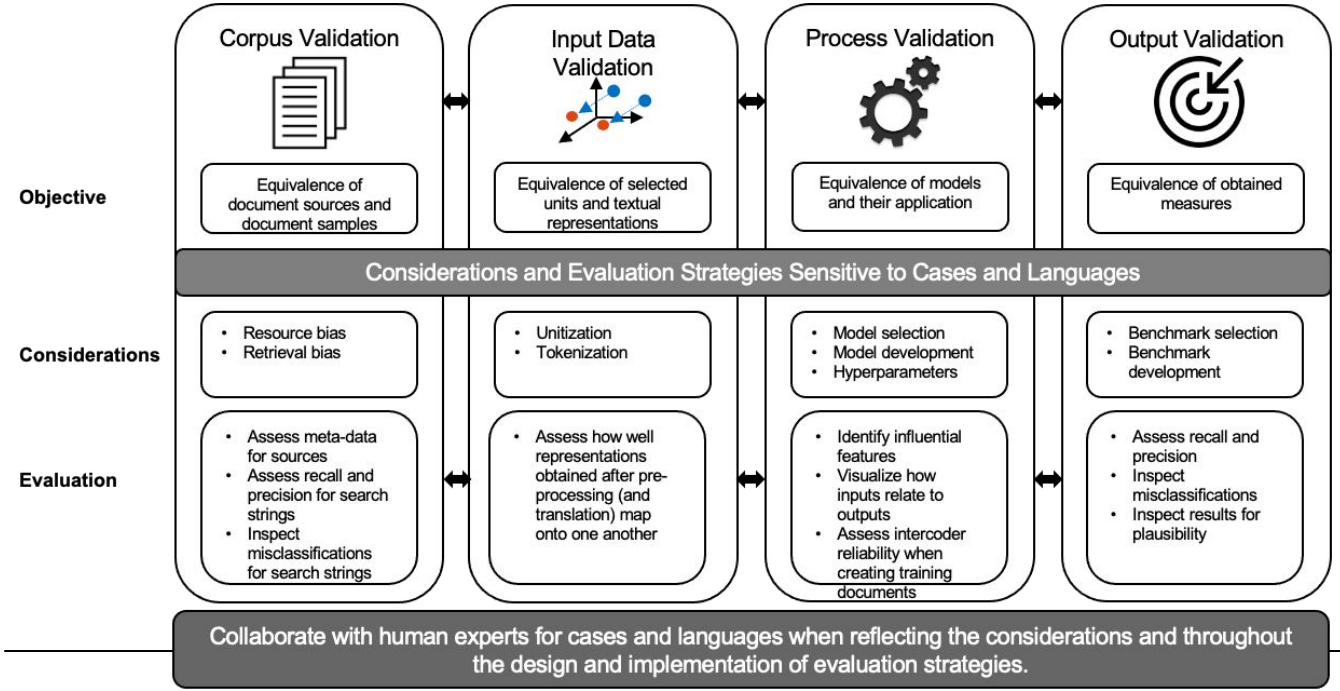
Lind et al., working paper, p.4

Validity in comparative research designs

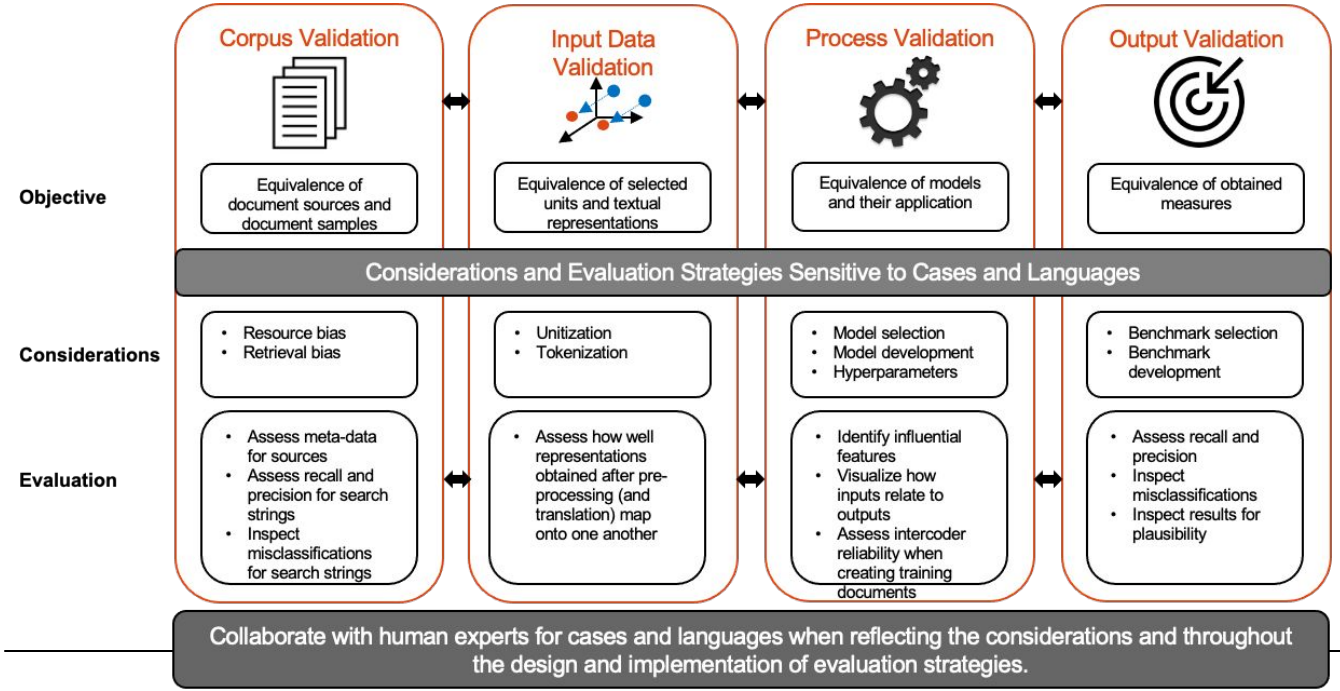


Lind et al., working paper, p.4

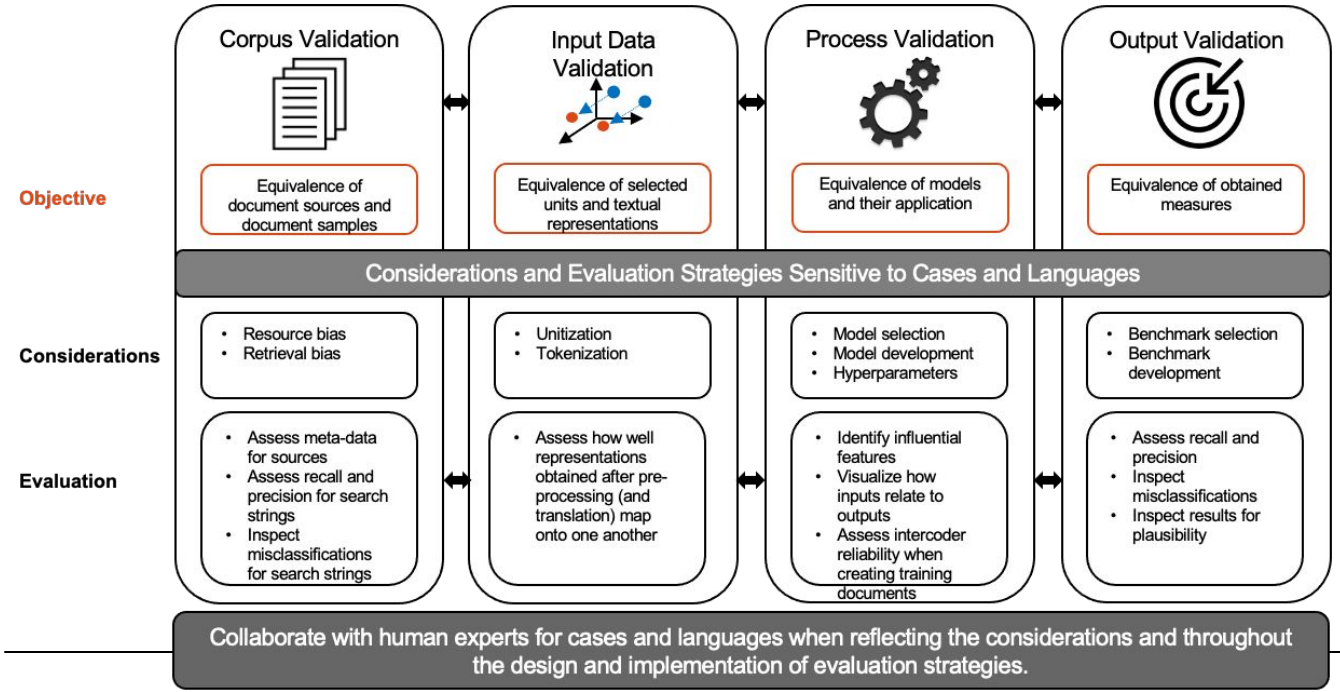
Validation framework



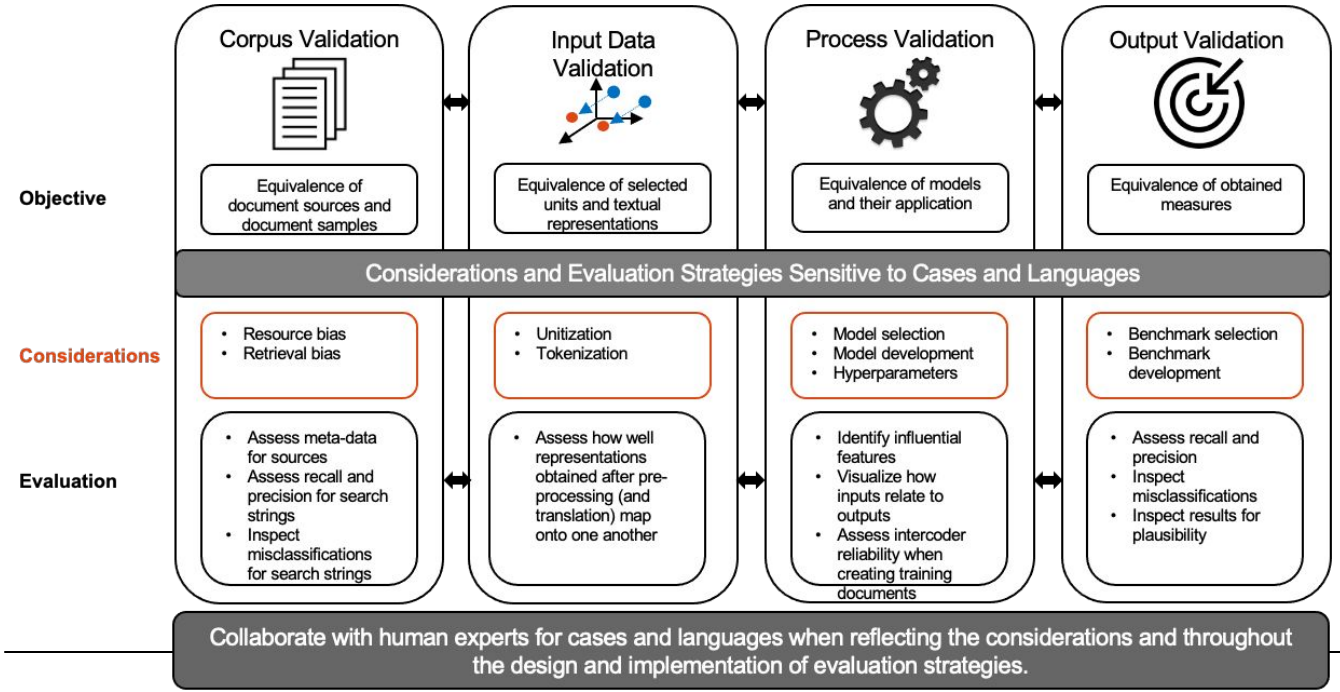
Validation framework



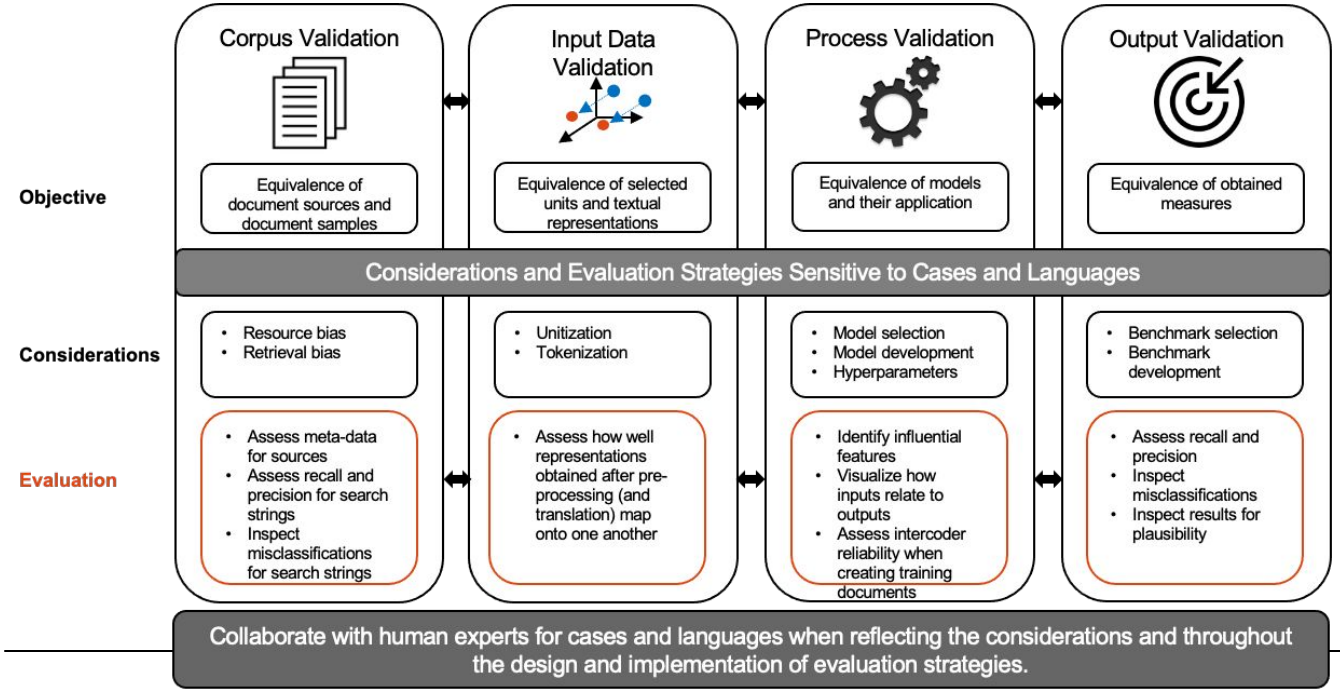
Validation framework



Validation framework

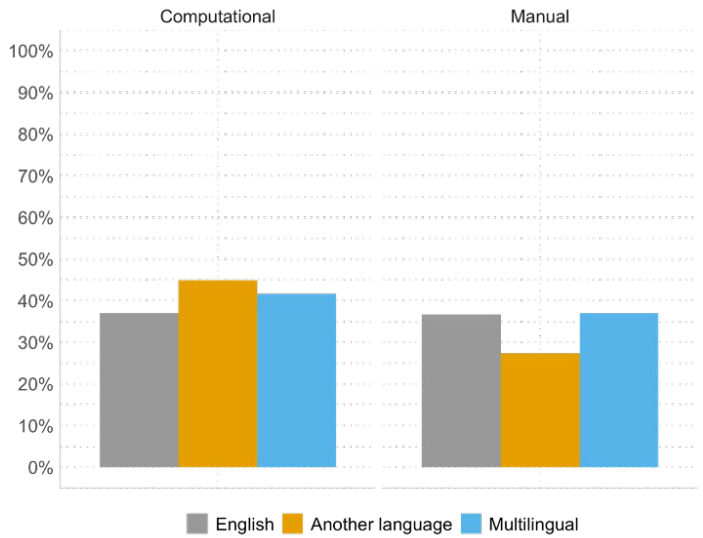


Validation framework



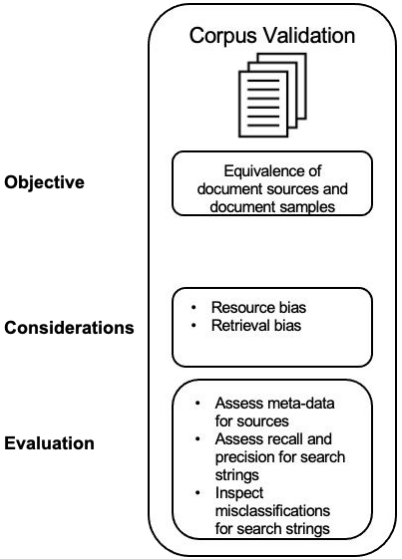
Validation on corpus level

- Approximately 45% of articles that rely on computational methods and corpora in multiple languages report on corpus validation efforts, while about 35% of articles that rely on manual methods and corpora in multiple languages do the same.



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

Validation framework: Corpus



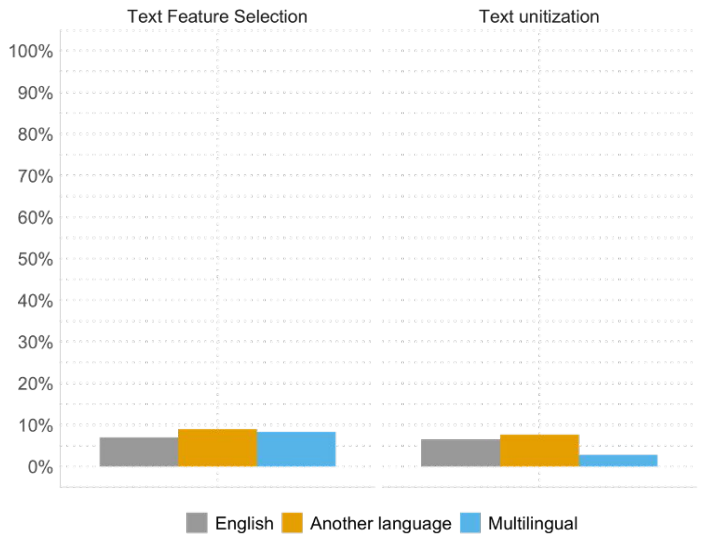
Example:

- Comparison of market shares of news outlets in Hungary and the US (Heft et al., 2023)

- News outlets may have distinct norms regarding the acceptance of specific slurs in their texts
  - Consequences for Hungarian and English search string design

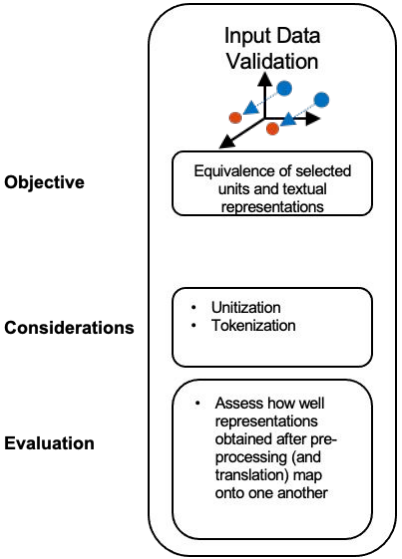
Validation on data input level

- Only few computational papers discuss text feature selection (*why these features and not some other?*) and text unitization (*why this unit of observation and not some other?*)



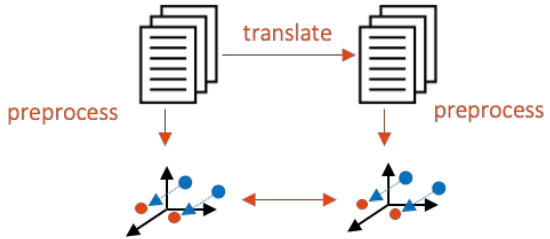
Baden, Dolinsky, Lind, Pipal, Schoonvelde, Shababo, & van der Velden, 2022

Validation framework: Data input



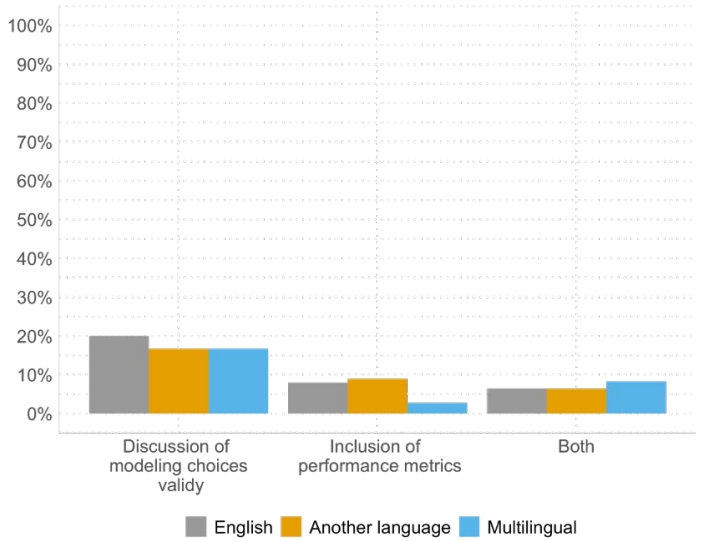
Example:

- examining uncivil rhetoric at the paragraph level
  - Reflecting about typical content of a paragraph in each case
- incivility might manifest through the use of explicit swear words, while in others, it might be expressed through specific multi-word phrases



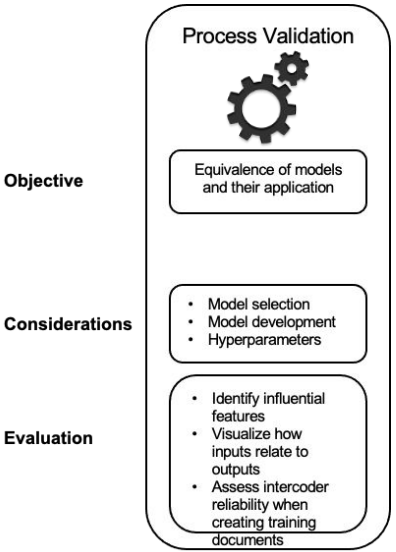
Validation on process level

- Among computational papers modeling choices (e.g., the choice of hyperparameters or the number of topics in topic models) is present in among 20% of papers.



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

Validation framework: Process



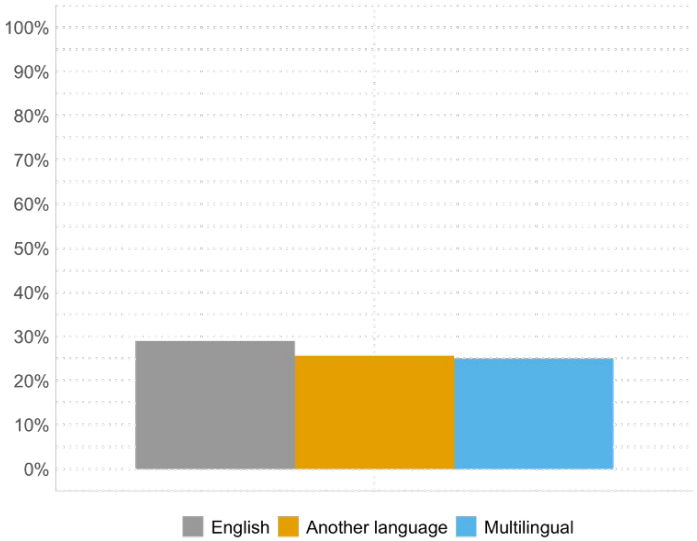
Example:

- Different languages may have different swearing cultures
  - Rutte is a ‘pancake’, Macron is a ‘crêpe’, ‘Trump is a ‘pancake’, ‘Orbán is a ‘palacsinta’, ‘Scholz is a ‘Pfannkuchen’

- When using ChatGPT to annotate incivility in text, test if the superiority of English as prompt language applies (Kuzman et al., 2023)

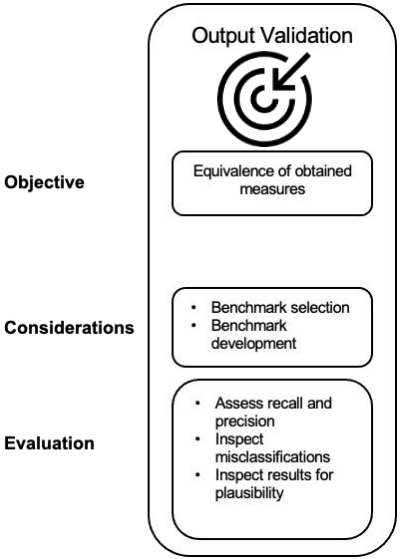
Validation on output level

- About one in three computational papers validates obtained measures against a human-coded benchmark (but the details of the coding are not always transparent)



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

Validation framework: Output



Example:

- •Benchmark that approaches the concept definition “disrespectful discourse that silences or derogates alternative views” (Jamieson et al. 2017, 206)

- Do we see a higher recall or precision in some languages and some cases than in others?

Summary

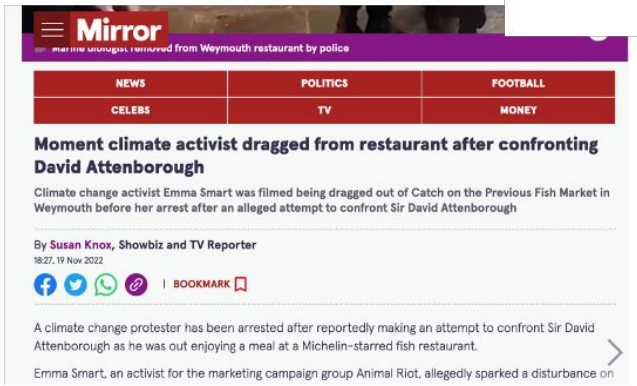
- Considering the fact that our analysis focused on studies published in top-ranking social science journals, it seems rather remarkable how few papers document an explicit effort to ensure the validity of measurement at any stage of the research process

- We don’t find dramatic language-specific differences but note that validation requires additional steps

Planning a research design along the framework

Case study: Climate Activism

- Comparing the media discourse about climate activism across several countries



First steps

1. Evaluation of human expertise in respect to
   - Language expertise
   - Case expertise
   - Domain expertise

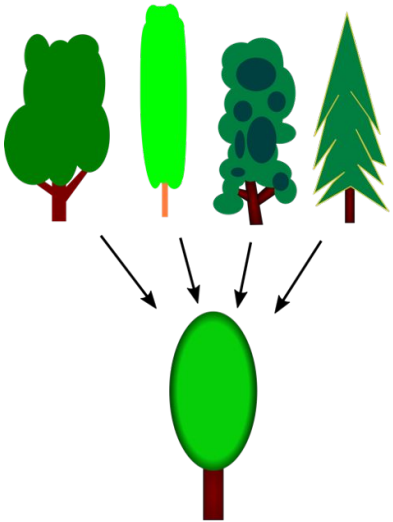
2. Concept definition

Mentimeter

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



Defining concepts (10min)

- Form groups – ideally one language/case expert per group
- Brainstorming task
- What could be an interesting concept to measure with an automated content analysis in the media discourse about climate activism?
- What concept would be interesting/relevant to compare across several European countries

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

REMINDER

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

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
- Search string validation: Case experts/native speakers code an artificial week (migration: yes/no), joint coder training (see Stryker et al., 2006)

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: We are interested in …

- Material type: Newspaper articles
- Topic: 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)

Selecting document sources (15 min)

- Form groups based on case expertise
- Jointly discuss and select two main traditional news sources for your case
- Ideas to start:
  - Open brainstorming
  - Source list: www.opted.eu
- Be ready to present and explain your selection to the plenum

Selecting documents –Keywords across cases

- Form groups based on language and case expertise
- Jointly discuss and identify relevant keywords that can be used in a search string to select news articles on climate activism
- Ideas to start:
  - Open brainstorming
  - documents examples
  - ChatGPT
- Be ready to present and explain your selection to the plenum

Selecting documents – Keywords per case (15min)

- Plenum discussion: Our goal is now to form search strings per case that are comparable across cases (and language).
- What keywords should we select per case and language?
- Since this is just an exercise: Let's keep it simple

Selecting documents – Building search strings

- You find an R Markdown file “data_task.Rdm” in the GitHub folder code.
- The solutions (commented code) can be found in “data_solution.Rdm”.

Valid outputs in multilingual and multi-context scenarios

Equivalent output

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

Strategyies:

- 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

Illustration: Supervised text classification result

REMINDER

Concepts: migration as 1. Economy, 2. Labor, 3. Welfare, 4. Security topic

id	country	publication_date	source	source_type	headline	m_fr_eco	m_fr_lab	m_fr_wel	m_fr_sec
1	UK	2013-02-09	Daily Mirror	Print	Asylum girl 'fed up' in UK; COURT	0	0	0	1
2	Spain	2005-06-04	El Pais	Print	Menores	0	0	1	0
3	Spain	2015-11-11	El Pais	Print	La Comisión considera altamente problemáticas las m	0	0	0	0
4	UK	2012-03-16	telegraph.co.uk	Online	Archbishop of Canterbury, Dr Rowan Williams: CV; Pro	0	0	0	0
5	UK	2012-08-27	telegraph.co.uk	Online	France's 'scandalous' expulsion of Roma camps resur	0	1	1	0
6	Spain	2002-03-13	El Pais	Print	ISRAEL TOMA LA 'CAPITAL' PALESTINA EN SU MAYO	0	0	0	0
7	Spain	2006-06-06	El Mundo	Print	EL DRAMA DE LA INMIGRACION / La integracion. Vale	0	0	1	0

Lind et al., 2021, Appendix

Annotations were performed on the basis of the full texts (not just the headlines)

Creation of a validation benchmark

Group exercise

Develop an output evaluation strategy (15 min)

- 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:

Implement output evaluation strategy (30min)

- Assign labels manually (assess intercoder reliability)
- Compare manual and automated measures
- Calculate recall and precision

Comments, questions, thoughts

Thank you very much