universität wien

Multilingual automated content analysis for comparative social science research

Fabienne Lind

Workshop DSI

University Zurich

13.02.2023

universität wien

Today

9:00-10:00	Input: Motivations, challenges, and strategies
10:30-10:50	*Coffee Break*
10:50-12:00	Case studies: Input and joint coding session
12:00-13:00	*Lunch*
13:00-14:45	Case studies: Input and joint coding session
14:45-15:00	Wrap-up

universität wien

Introduction

universität wien

About me

Fabienne Lind

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

universität wien

Sharing insights from

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

Collaboration with Hauke Licht

- Relevant today: working paper: Going cross-lingual: A guide to multilingual text analysis

universität wien

Your turn :)

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

universität wien

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

universität wien

Workshop philosophie

Topics are covered with

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

Interrupt, ask all kinds of questions

universität wien

Conversation about your projects

Very informal opportunity to talk about your text analysis use cases and (initial) design (plans)

- Research question, Data, Methods, Current struggles

Get in contact

universität wien

Workshop repository

https://github.com/fabiennelind/Workshop_Multilingual-Text-Analysis_and_Comparative-Research

universität wien

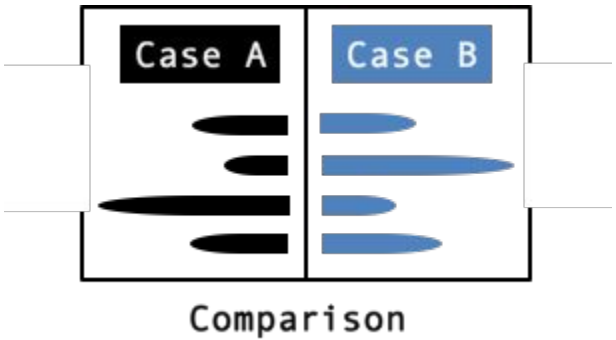
Comparative research and content analysis

Input 1

DALL.E

Comparative research

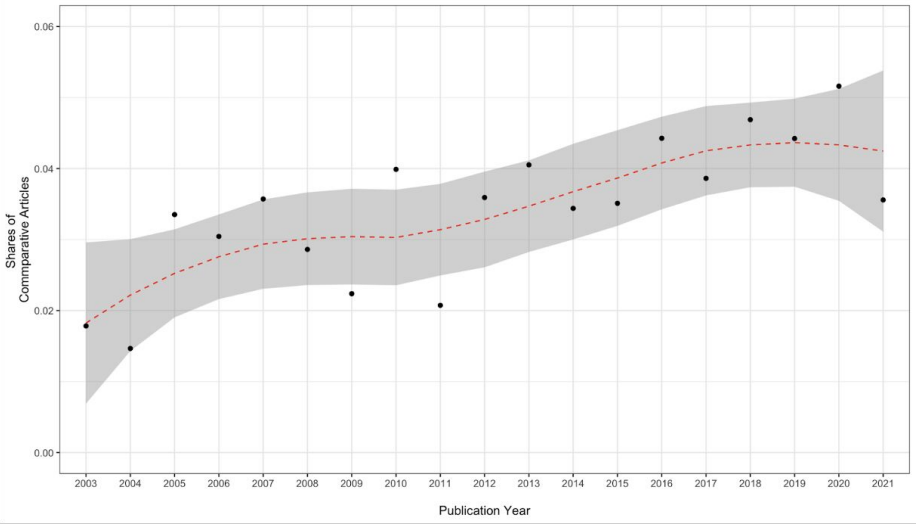
- Comparative research in social science involves comparisons between a minimum of two cases with at least one object of investigation relevant to social science research.
- Cases as we define them here are macro-level units such as systems, cultures, countries, and markets)



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

Increasing popularity of comparative research

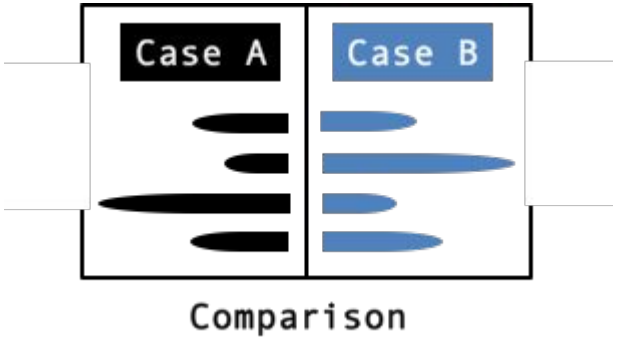
Shares of comparative articles in comm science journals, 2003-2021



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

Reasons to compare

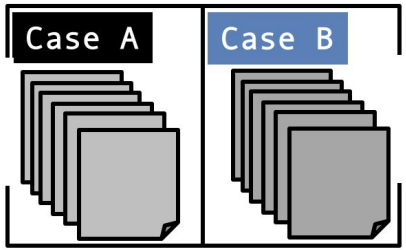
- insights into the differences and similarities of cases
- improved understanding and contextualization of the own case
- raised awareness for other cases
- the test and generalizability of theories across diverse settings
- the investigation of transnational processes across contexts



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

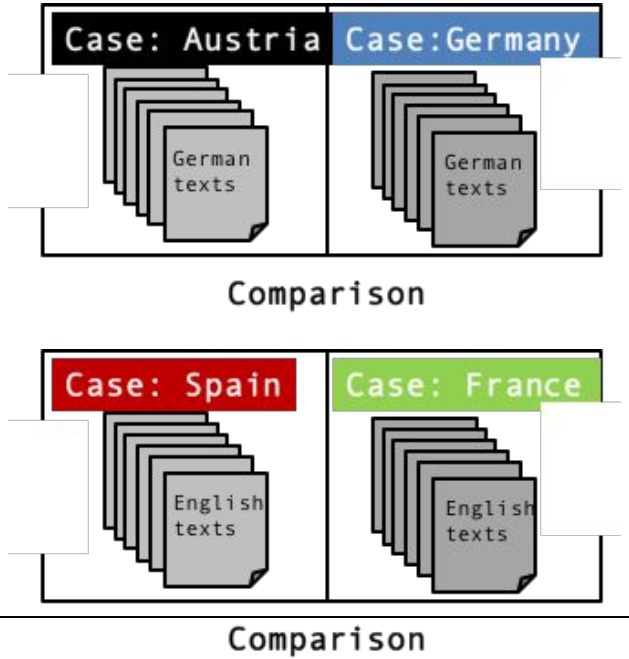


Boomgaarden & Song, 2019; McLeod & Blumler, 1987; Esser & Vliegenthart, 2017; Esser & Hanitzsch, 2012; Livingstone, 2003

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

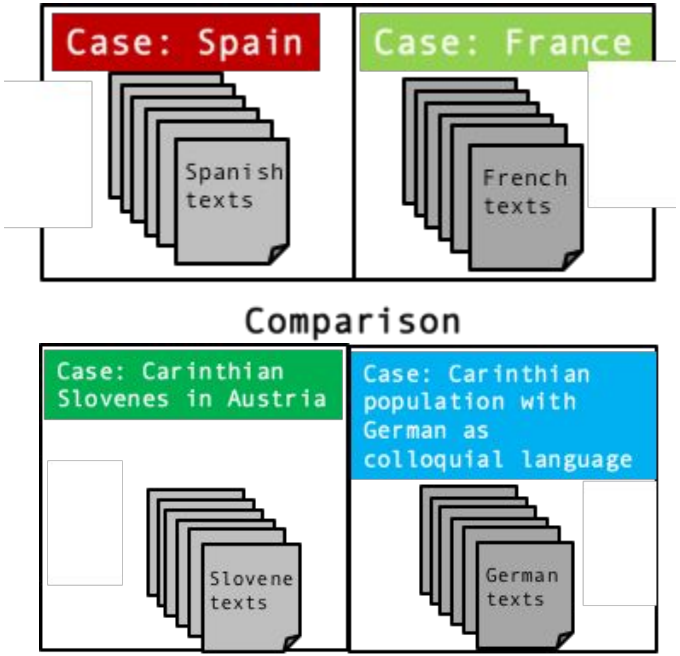
But the likely scenario is multilingual

- 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)



universität wien

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 MANIFESTO (Volkens et al., 2015)

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

universität wien

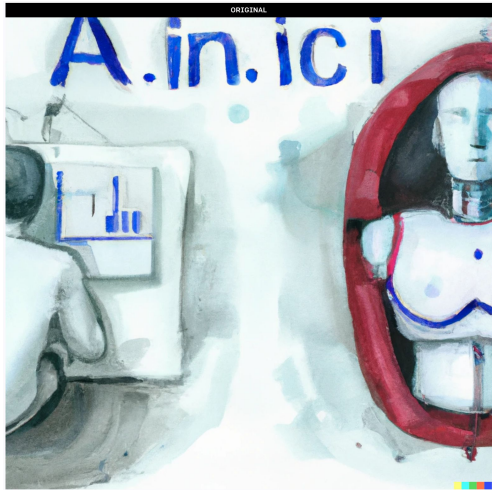
Purpose of obtaining measures for a large number of documents

- 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

The end of manual coding?

- Augmenting not replacing (Grimmer & Steward, 2013)
- Human input for quality control:
  - select, monitor, and test on the level of data, inputs, process, outputs
  - Even more important in projects with multiple cases and languages



DALL.E

universität wien

Some major challenges when working with large corpora

Big data, big bias?

The end of theory?

Generalizing from online to offline behavior

Ethical concerns (Guess, 2021)

universität wien

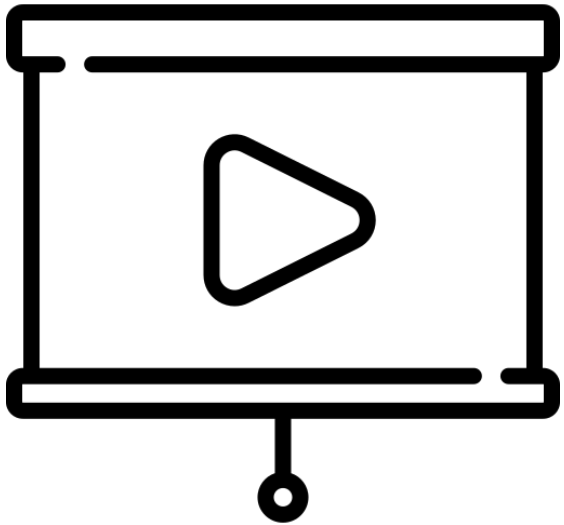
Multilingual automated text analysis methods

- 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.

Your (initial) text analysis project ideas?:)



by Freepik - Flaticon

by Freepik - Flaticon

universität wien

Main goal of the workshop

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

universität wien

Main goal of the workshop

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

universität wien

Handling multilingual corpora

Key challenge and solution approaches

universität wien

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 substantively more interesting patterns

universität wien

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

universität wien

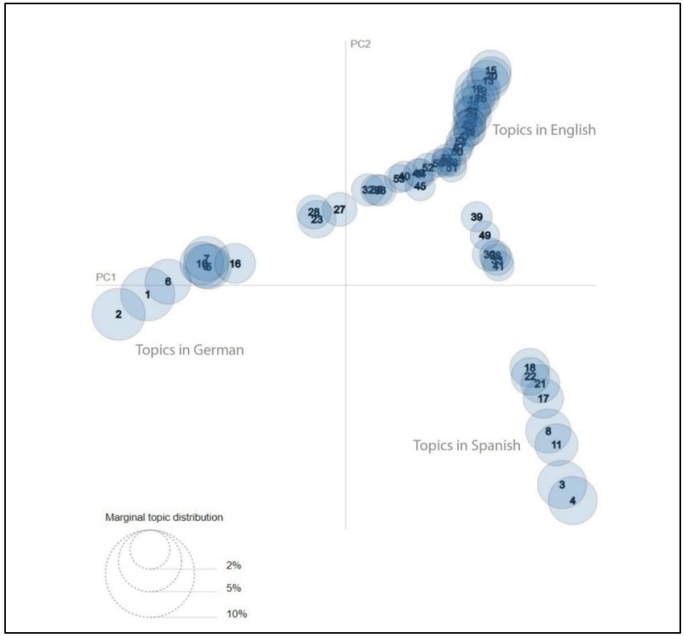
Illustration 1 (Part 2)

- 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

universität wien

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!

universität wien

How to jointly analyze documents in different languages?

universität wien

Two approaches

1. Separate analysis
2. Input alignment

1. Separate analysis

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

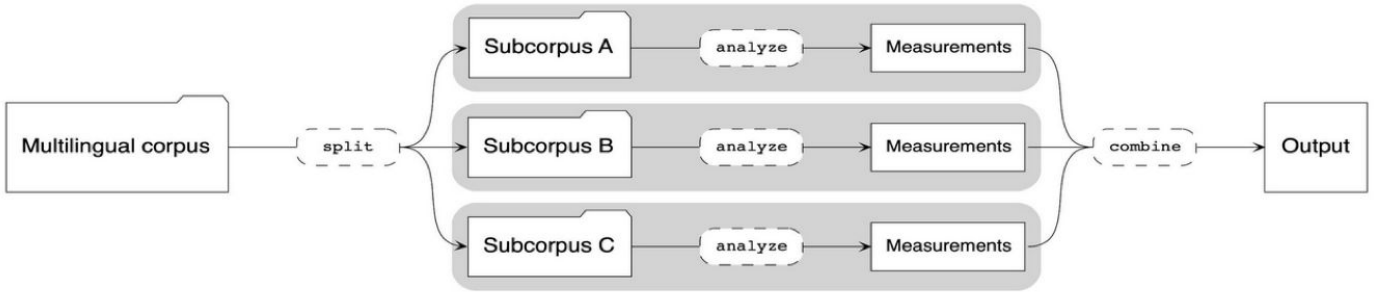


Figure 1 Illustration of the separate analysis strategy to multilingual text analysis.

Licht & Lind, forthcoming

universität wien

1. Separate analysis

Example

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

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

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

universität wien

2. Input alignment

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

2 options to implement the idea:

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

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

2. Input alignment

Option 1: (Machine) Translation

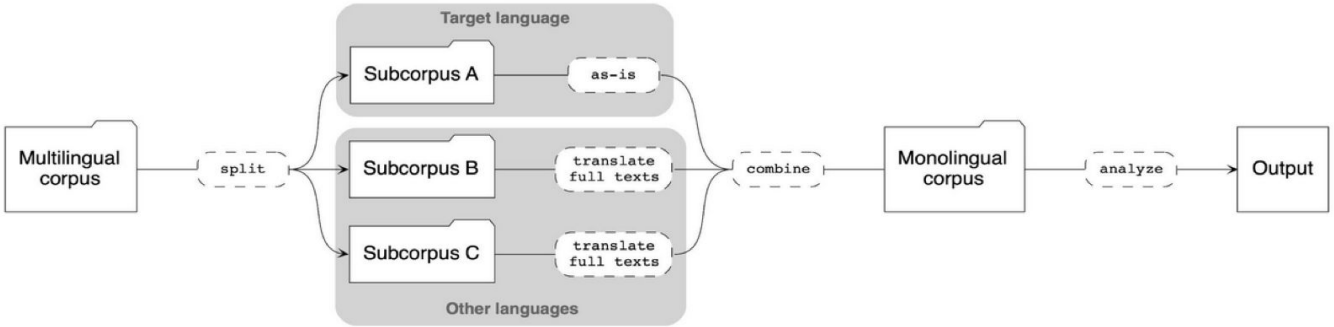


Figure 2 Illustration of the full-text translation approach to input alignment

Licht & Lind, forthcoming

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, forthcoming

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

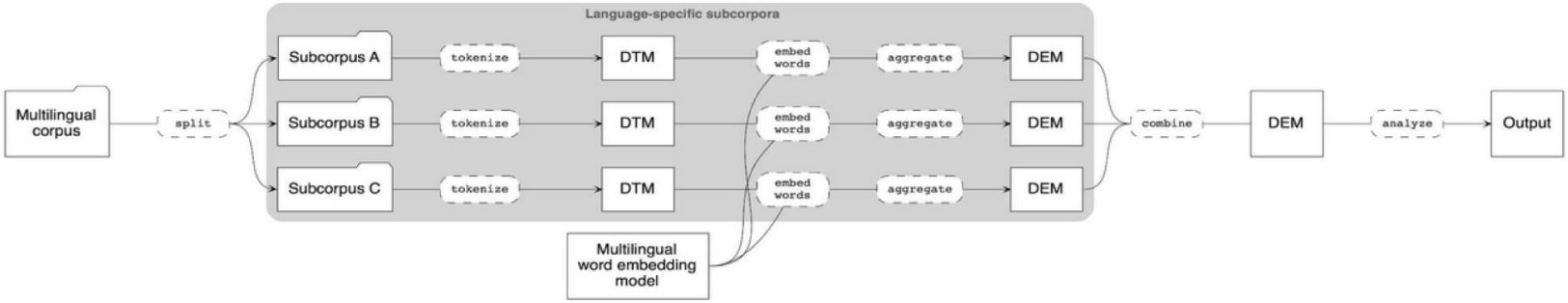


Figure 4 Illustration of the multilingual word embedding approach to input alignment.

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.

Licht, 2022

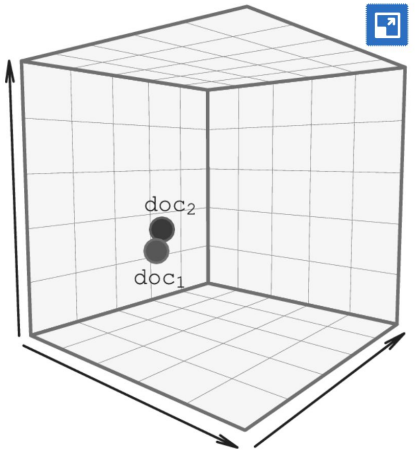


Figure 1 Schematic depiction of multilingual sentence embedding of example sentences in Table 1. *Note:* Depicting embedding in three dimensions serves illustrative purposes only.

How to decide between the approaches? (15 min)

Let‘s form groups

Discuss and collect decision criteria, pros and cons

1. Separate analysis
2. Input alignment

universität wien

How to decide between the approaches

Some criteria

- Skills
- target concept
- availability of instruments (e.g., labeled data, dictionaries)
- transparency
- Replicability

universität wien

Handling multilingual corpora in (social) contexts

Key challenge and solution approaches

universität wien

What could the following sentence mean?

“You have a green light.”

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

universität wien

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

universität wien

What could the following sentences mean?

“Immigrants improve society by bringing in new ideas and cultures”

Braun et al., 2013

universität wien

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

universität wien

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**

universität wien

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!

universität wien

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.

universität wien

Equivalence and validity

Equivalence in comparative research

- Comparability or equivalence 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 valid comparison

DALL.E

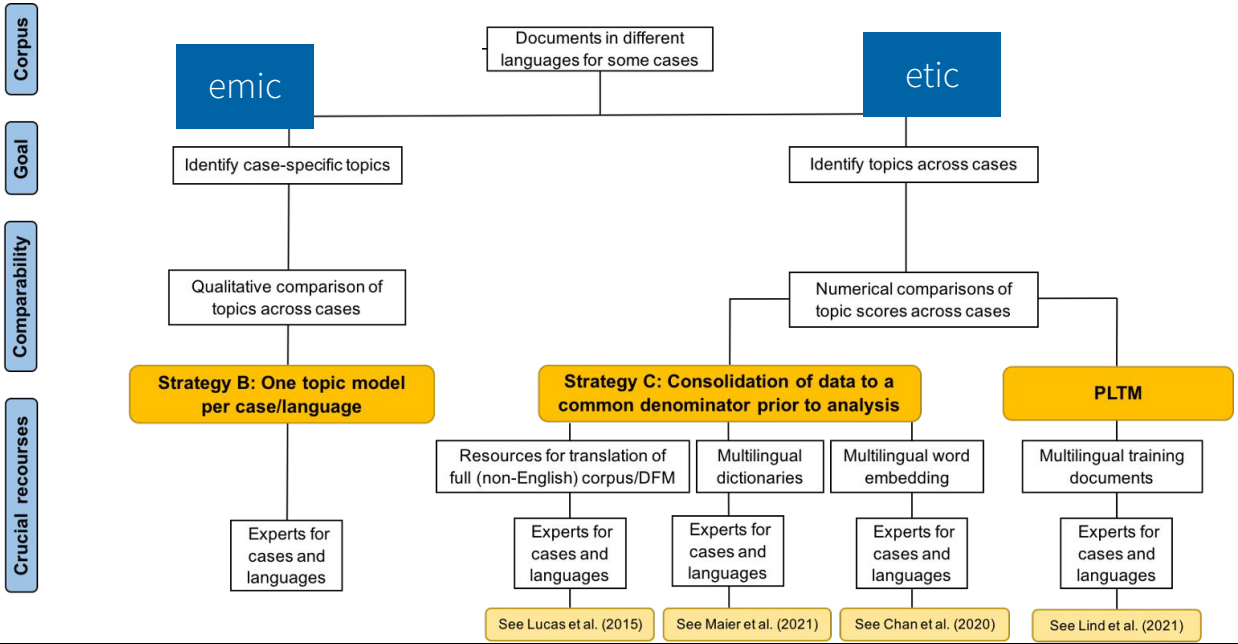
universität wien

Emic and etic

- Two approaches to comparability
- Two ends of a continuum
- The positioning of the own research project 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	measure construct with standardized instruments
may hinder comparison between cases	may overlook specific cultural perspectives

Emic and etic: Example topic modeling



Emic and etic

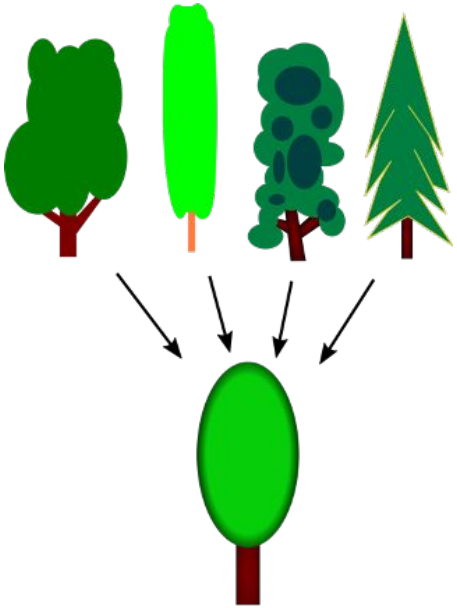
- Can you think of a research question where an emic approach and of another where an etic approach would be preferred?

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 comparison between cases	may overlook specific cultural perspectives

Constructs in comparative research

A common approach:

- 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

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

universität wien

A language and context sensitive validation framework

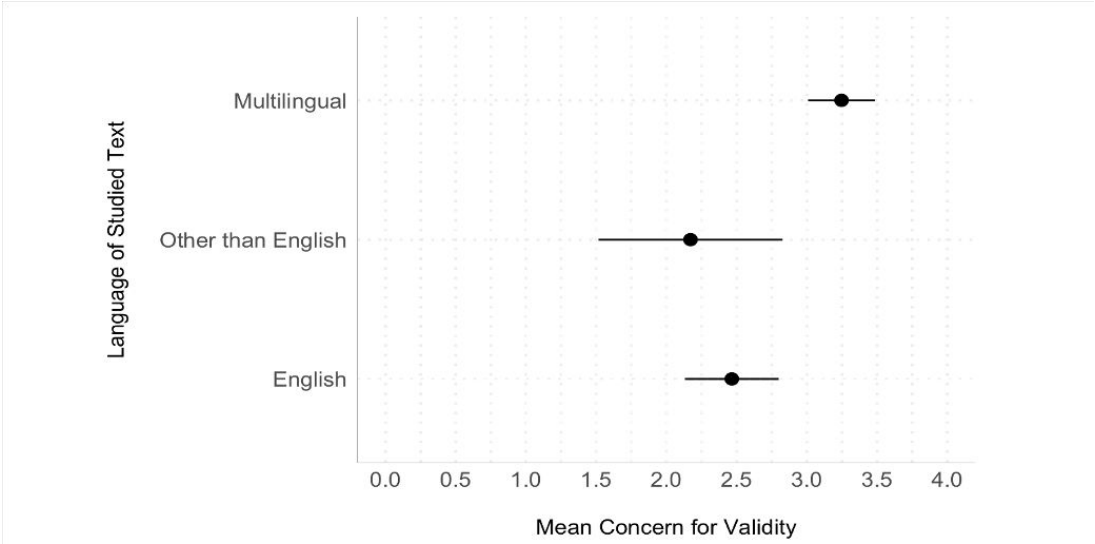
universität wien

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

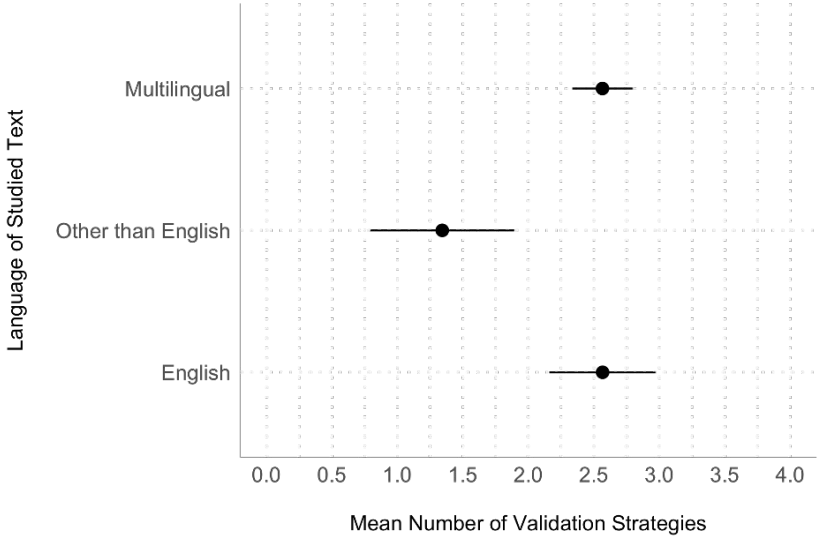
- Among researchers in our sample, those who work in more than one language, express more concerns about the validity of findings from computational methods



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

universität wien

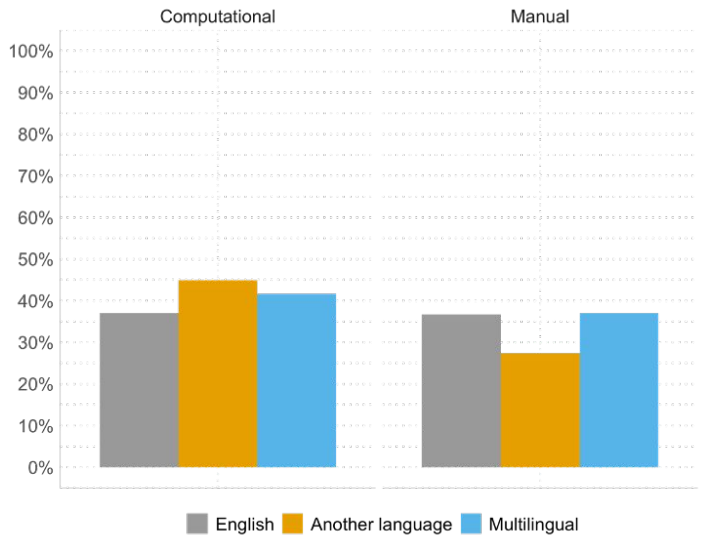
Validation strategies

- We focus on validation strategies on our four levels

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

Validation on data level

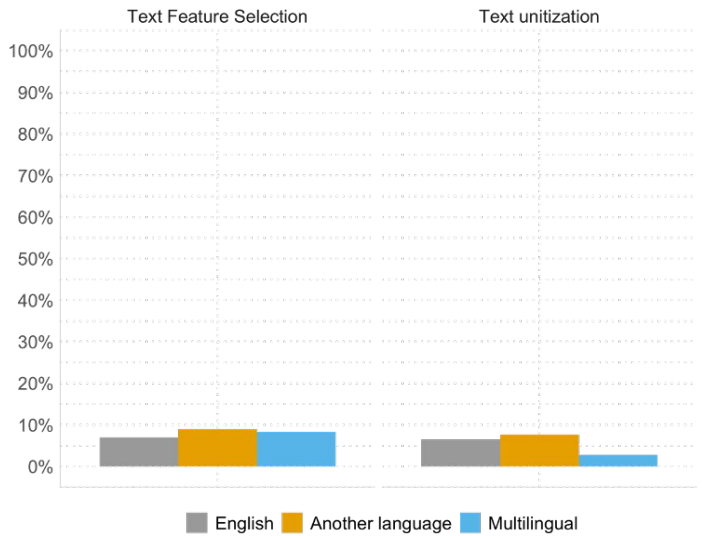
- Approximately 45% of articles that rely on computational methods and corpora in multiple languages report on data validation efforts, while about 35% of articles that rely on manual methods and corpora in multiple languages do the same.



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

Validation on input level

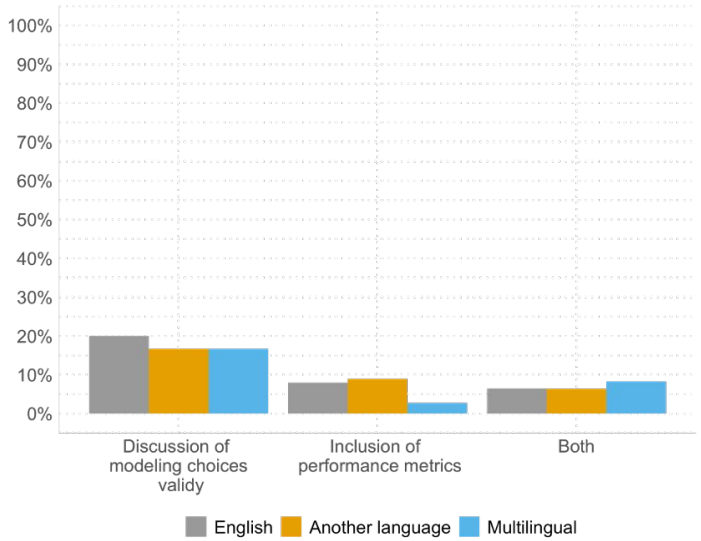
- Only few computational papers discuss text feature selection (*why this text representations and not some other?*) and text unitization (*why this unit of observation and not some other?*)



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

Validation on process level

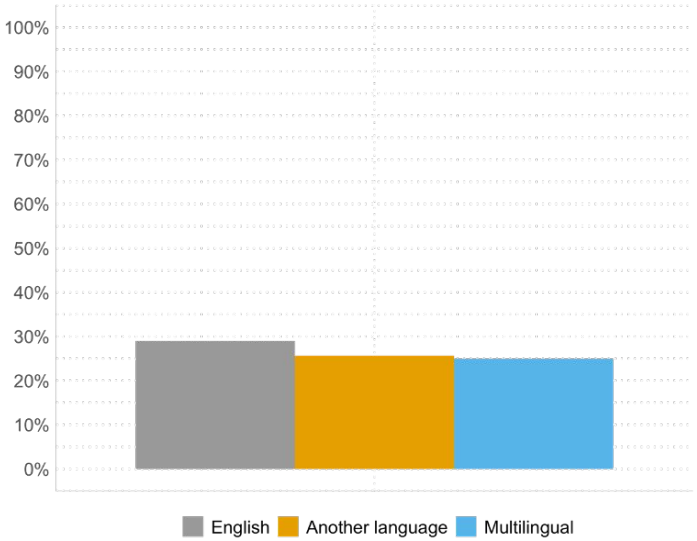
- Among computational papers modeling choices (e.g., the choice of hyperparameters or the number of topics in topic models) is present in among 20% of papers.



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

Validation 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

universität wien

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

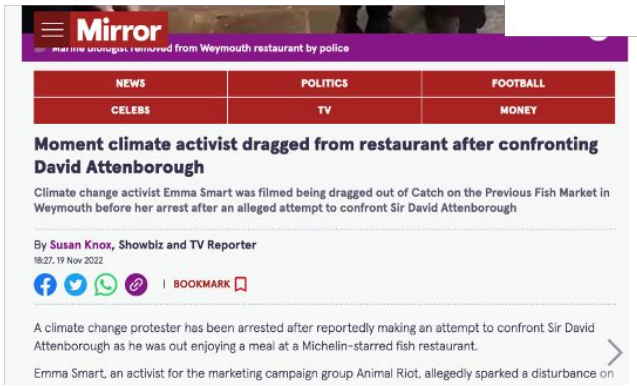
universität wien

Planning a research design along the framework

universität wien

Case study: Climate Activism

- Comparing the media discourse about climate activism across several European countries



universität wien

First steps

Evaluation of human expertise in respect to

- Language expertise
- Case expertise
- Domain expertise

Concept definition

universität wien

Mentimeter

I have native language like skills for

universität wien

Mentimeter

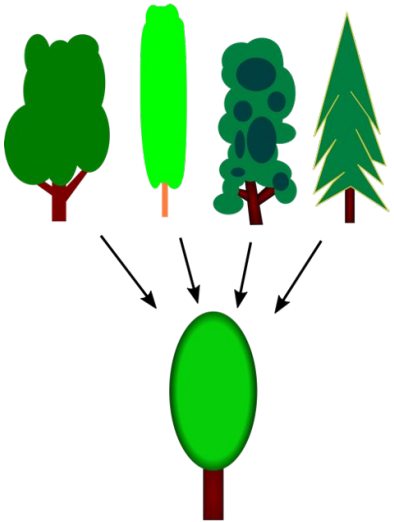
I am a case expert for

universität wien

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



universität wien

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).

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

Illustration

REMINDER

b) Retrieval of equivalent documents

- Approach: ‘Etic’ concept definition of’ migration
- Retrieval of a multilingual news article sample with 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)

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 – Testing 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".

Equivalent input

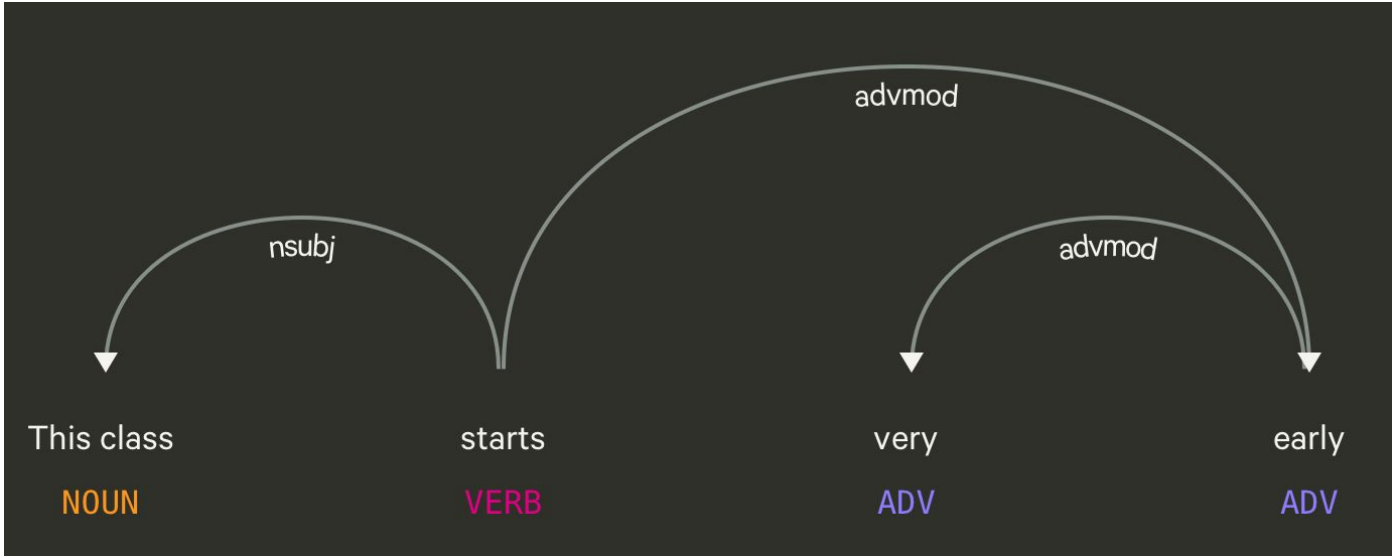
Objective: Select textual representations that are equivalent across languages and contexts & relevant for the research question at hand

For an emic (=case-sensitive) measurement goal:

- Approach 1: try to preserve context per language/case during pre-processing
- Approach 2:
  - Machine translation: ideally add translation checks
  - Multilingual Embedding: use context dependent word embeddings, pre-trained models with domain similar documents and/or fine-tuning with the labeled documents
- Approach 3: use domain and context related external resources (dictionaries, parallel or aligned corpora)

universität wien

Part of Speech Tagging



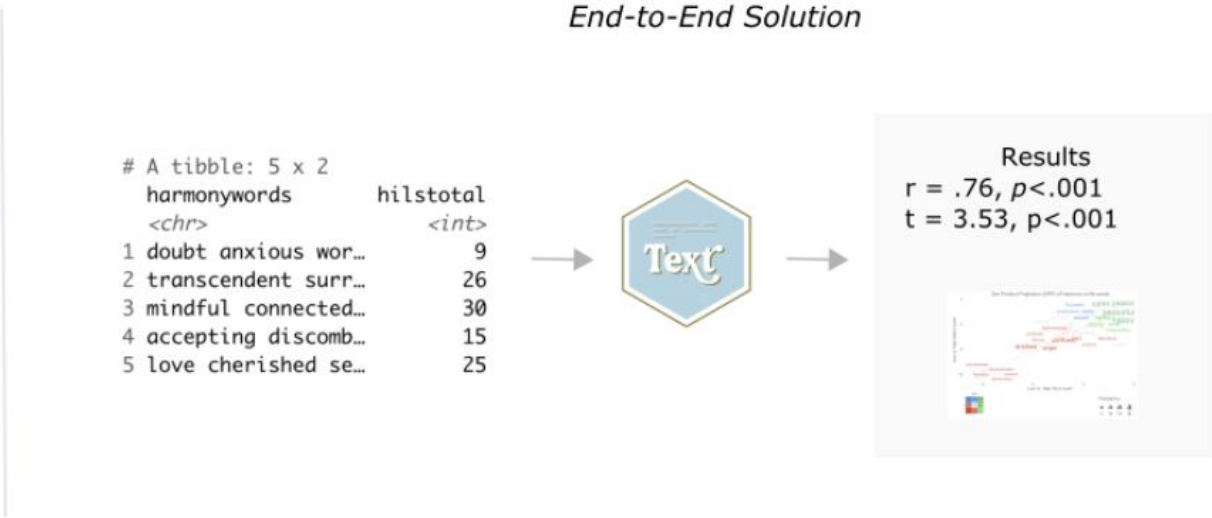
universität wien

Named Entity Recognition

John PERSON and Mary PERSON sure like to generate texts about bananas.
New York Times ORG has paid them 5 million dollars MONEY for an article recently. It was written in
English LANGUAGE .

universität wien

R package “text”



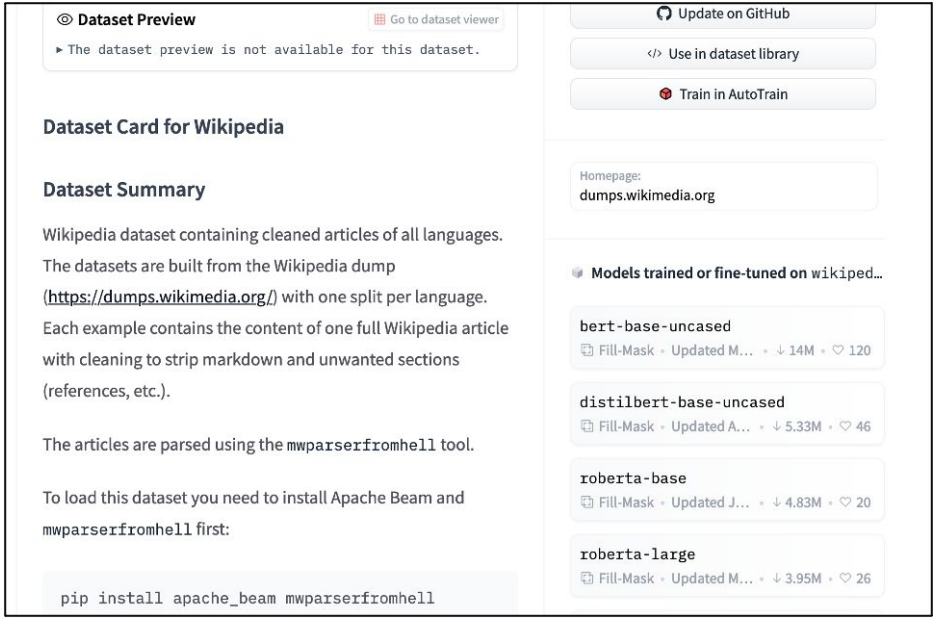
https://r-text.org/

universität wien

Hugging Face

- AI community with numerous multilingual resources such as datasets and language models
- https://huggingface.co/

Hugging Face



universität wien

Implementing the input alignment approach

Task: Step 1: Set up a DeepL Account

- Step 2: Translate headlines into English via DeepL API

universität wien

Feature selection

- Lemmatize text and select only nouns
- Extract named entities

Input selection for case study (15 min)

- Go back in your concept related groups
- Discuss and decide together what inputs you would use for the measurement of your target concept
- Be ready to explain your selection to the plenum

Input selection (30 min)

- You can now continue with the case study and try to implement the input selection decision of your group. You might be able to use parts of the code examples input_solution.Rdm
- If you like to work on another case (immigration news coverage), you can work more closely along prepared code: A task and solutions are available as input_task & input_solution.Rmd
-

Equivalent processes

Objective: Ensure that the chosen computational algorithms equally effective in all cases

Strategies:

- Discuss and compare e.g., the choice of hyperparameters or the number of topics in topic models
- Recommended: Knowledge about the (language-specific) quality of tools and algorithms used
- Use case-specific labelled documents for supervised classification

universität wien

Deep learning

Why use it?

Use python

Karas, pytorch, sklearn, spacy

Equivalent output

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

Strategy:

- 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

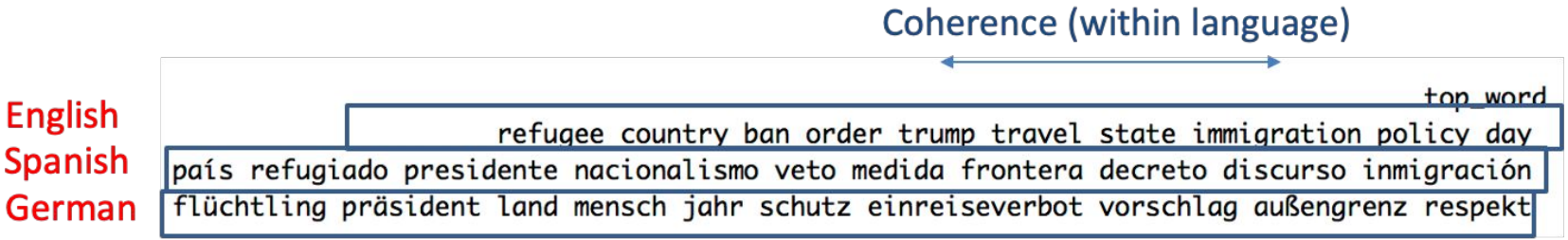
Annotations were performed on the basis of the full texts (not just the headlines)

Illustration: Topic modeling

REMINDER

Coherence i.e., close semantic relation of words in one language

NPMI metric (Lau et al., 2014) and native speaker evaluation



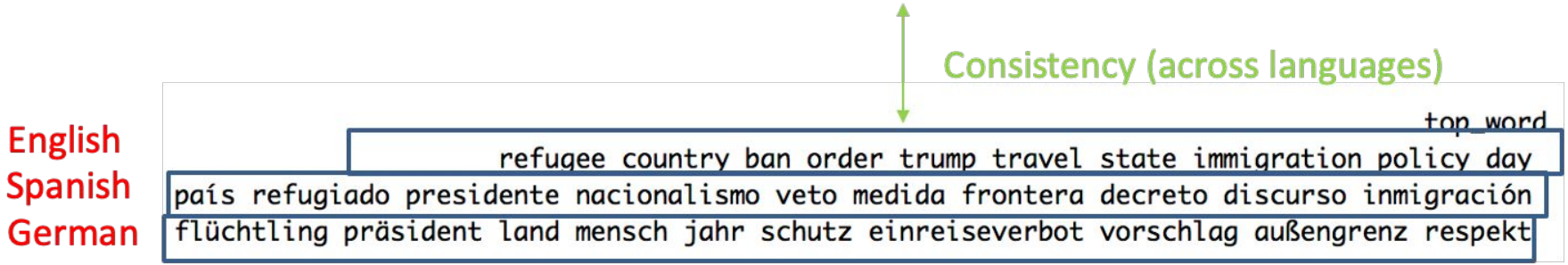
Lind et al., 2022

Illustration: Topic modeling

REMINDER

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

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



Lind et al., 2022

universität wien

Data | Input | Process | Output

Illustration: Topic modeling

- Manual labeling of the final model

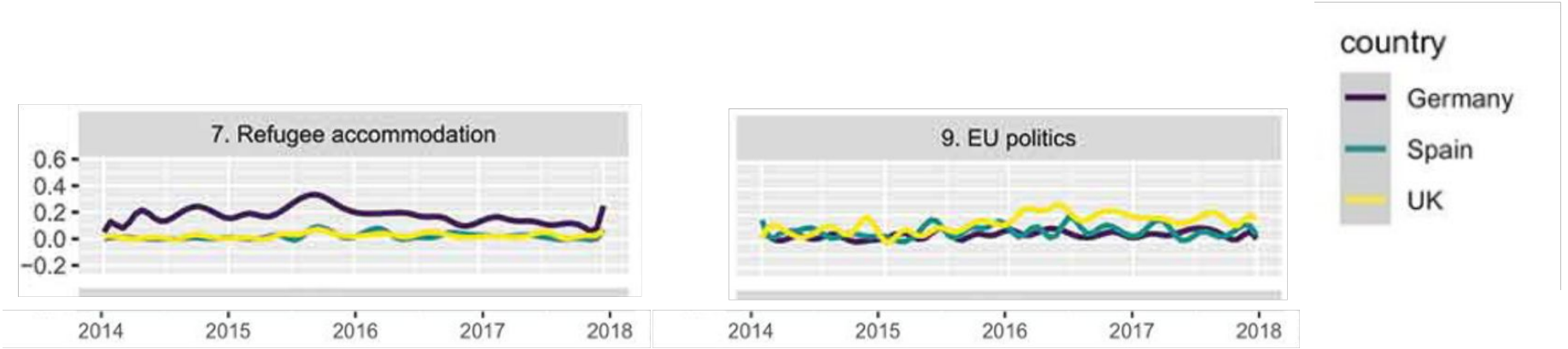
Topic	Lang.	Top 10 words
1.Welfare & jobs	EN	people worker immigration country migrant benefit report figure job health
	ES	persona número trabajador inmigración país millón inmigrante aumento cifra beneficio
	DE	zahl prozent land million migranten arbeit einwanderung bericht bevölkerung problem
2. Education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

Lind et al., 2022

Illustration: Topic modeling

REMINDER

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



Lind et al., 2022

Illustration: Topic modeling

REMINDER

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

Lind et al., 2022

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 Google sheets: https://docs.google.com/spreadsheets/d/1OcCvBD4iMSq6fzQZPbAIwgtdfvY9gwoOZwc8p7tPkSA/edit?usp=sharing

universität wien

Data | Input | Process | Output

Implement output evaluation strategy (30min)

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

universität wien

Wrap up

universität wien

What are open questions?

Validation

Human understanding of text as gold standard, even more so in multilingual comparative contexts

Don’t trust numbers trust yourself

And other human coders

universität wien

Upcoming

- Special issue on multilingual text analysis in Computational Communication Research
- Opted.eu

universität wien

Thank you very much