Multilingual Automated Content Analysis for Comparative Social Science Research

Workshop LMU Munich, 4.10.2022

Fabienne Lind

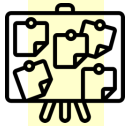
Computational Communication Science Lab,
Department of Communication, University of Vienna

Today

14:00 – 14:05 Introduction

14:05 – 14:15 Comparative social science & (automated) content analysis

14:15 – 14:45 Handling multilingual corpora:
Key challenge and three solution approaches



14:45 – 15:00 How to decide between the three approaches?

15:00 – 15:30 Coffee & Coding

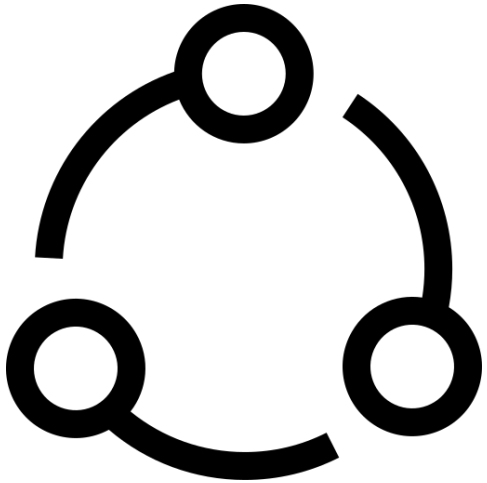
Today

15:30 – 15:45 Communication in social contexts

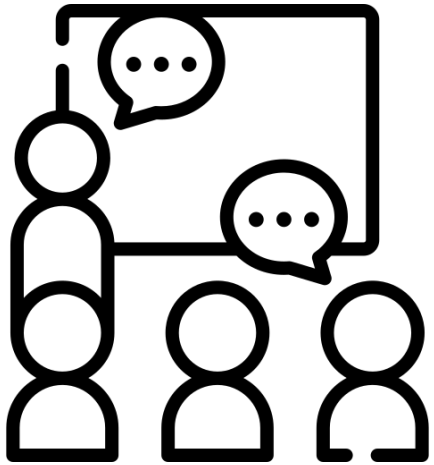
15:45 – 16:15 A language and context sensitive validation framework

16:15 – 16:30 Resources to get started

16:30 – 17:00 Themed tables & Wrap-up



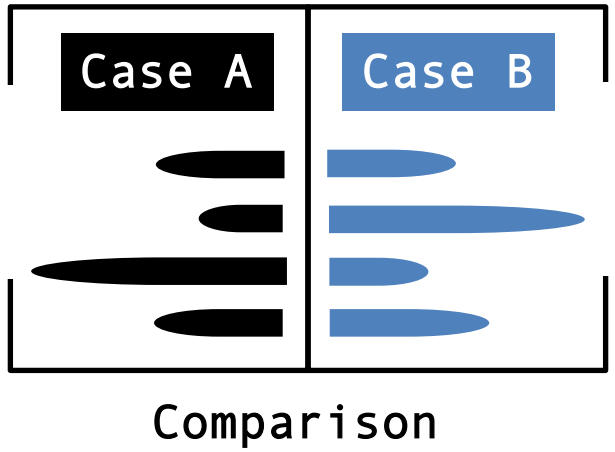
Introduction



Comparative social science & (automated) content analysis

Comparative social science

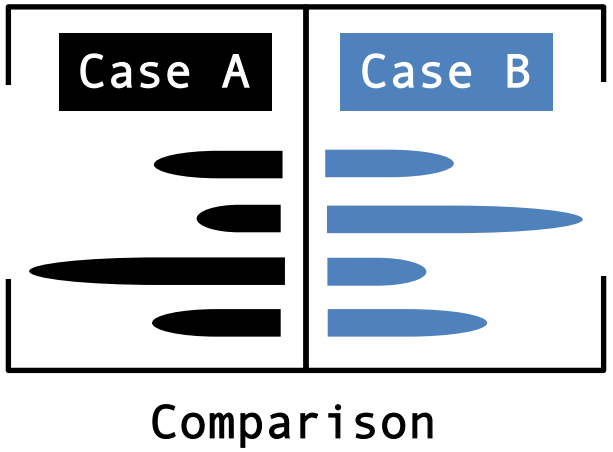
- **Comparative research in social science** involves comparisons between a minimum of two cases with at least one object of investigation relevant to social science research.
- **Cases** are macro-level units such as systems, cultures, countries, and markets)



definitions adapted from Esser & Hanitzsch, 2012, p.5

Reasons to compare cases

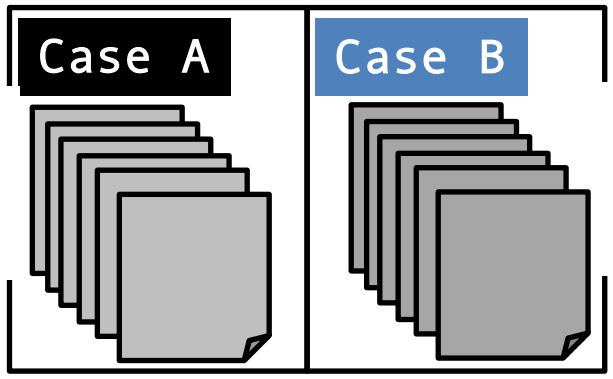
- insights into the differences and similarities of cases
- improved understanding and contextualization of the own case
- raised awareness for other cases
- the test and generalizability of theories across diverse settings
- the investigation of transnational processes across contexts



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

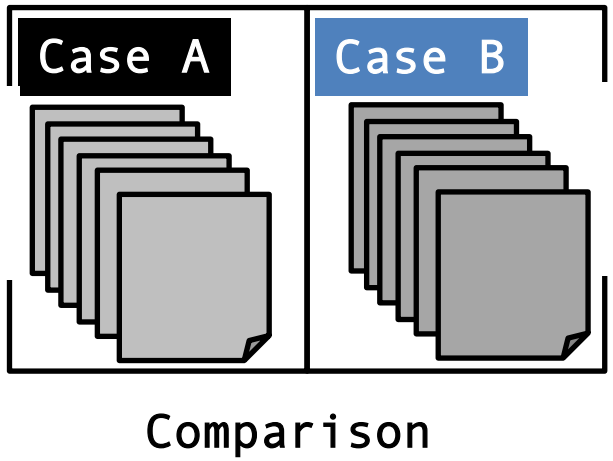
Comparison of cases with content analysis

- 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



Comparison of cases with content analysis

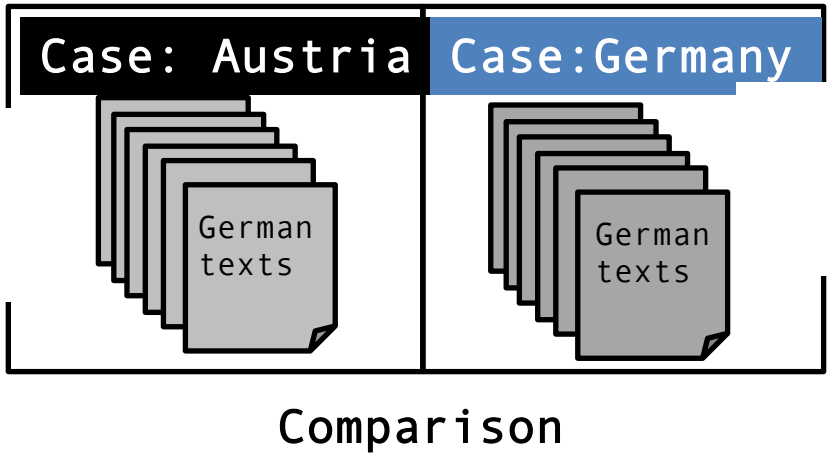
- Cases can be represented with documents
- Content analysis as one method to draw reproducible and valid inferences or meanings from documents (Krippendorff, 2004, p. 18)
- The goal of comparative content analysis is to compare cases via the comparison of specific contents in the documents



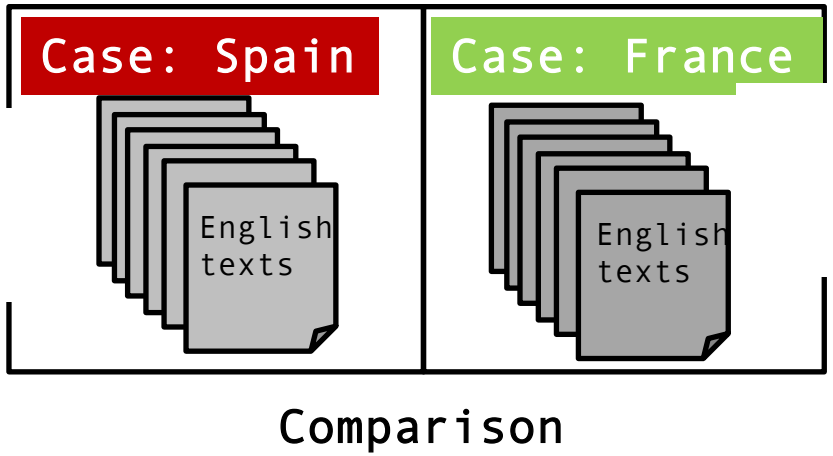
Comparison of cases & the language(s) of documents

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 & the language(s) of documents

universität wien

But the likely scenario is multilingual

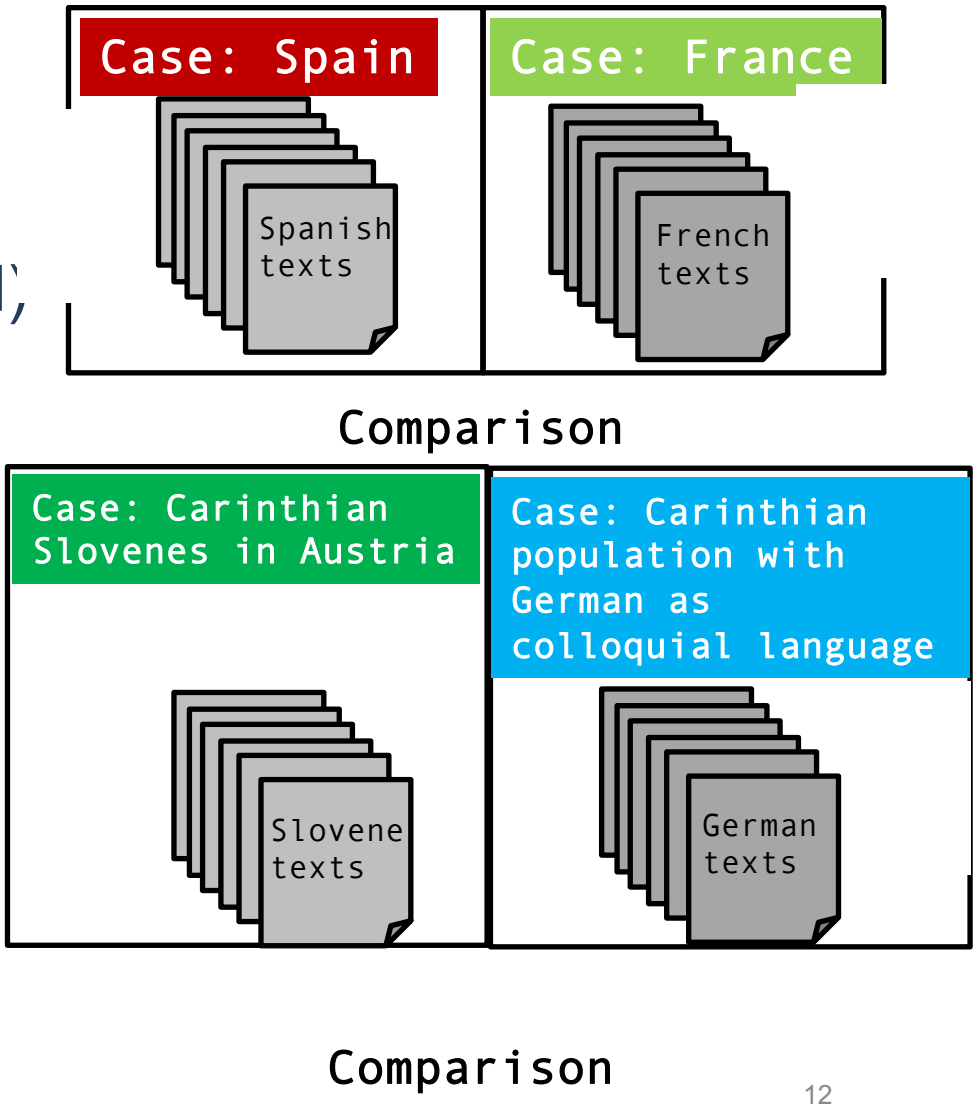
- human communication of at least two compared cases manifests in texts in different languages

Comparison of cases & the language(s) of documents

universität wien

Examples for multilingual scenarios:

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



Comparison 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)

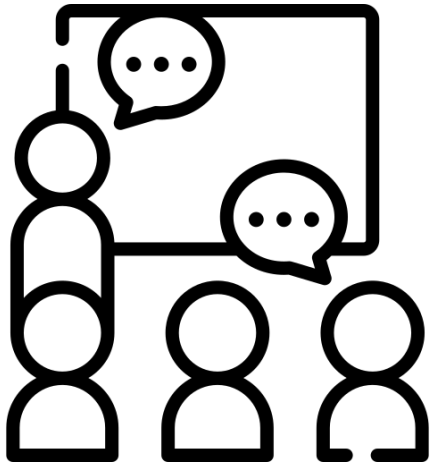
Automated content analysis as fast and reliable alternative to analyze large numbers of documents

Multilingual automated content analysis methods

- Methods to process and analyze large numbers of documents that are written in multiple languages
- Useful for comparative research designs where 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.



Handling multilingual corpora:
Key challenge and three solution approaches

What is the actually the problem?

- Moving from raw texts to quantitative text representations applying the same procedures as in monolingual scenarios is little useful

Illustration 1:

- Document 1: “I like my coats”
- Document 2: “Ich mag meinen Mantel”
- Document 3: “I like smoking”
- Document 4: “Ich mag meinen Smoking”

Can you name the documents that address the same concept? Which concept could it be?

What is the actually the problem?

universität wien

Illustration 1:

- representing the documents in a document term matrix, all lower case

	I	like	my	coats	Ich	mag	meinen	mantel	smoking
D1	1	1	1	1	0	0	0	0	0
D2	0	0	0	0	1	1	1	1	0
D3	1	1	0	0	0	0	0	0	1
D4	0	0	0	0	1	1	1	0	1

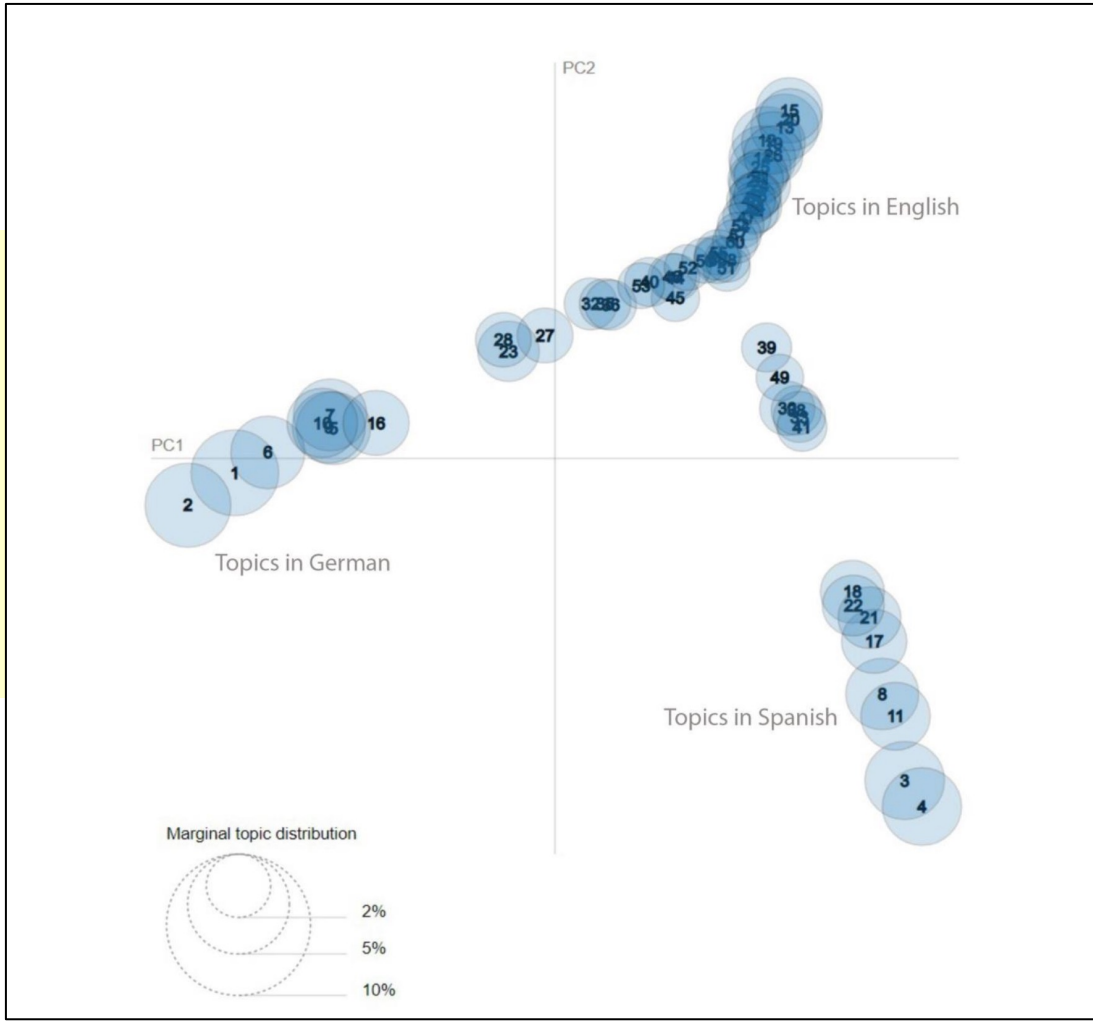
- different words express the same meaning
- identical words express different meanings

What is the actually the problem?

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

Goal

- Achieving cross-lingual measurement equivalence
  = 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?

universität wien

Image by Myriams-Fotos from Pixabay

3 approaches

universität wien

Seperate analysis

Input alignment

Automated bridging during analysis

Approach 1: Separate Analysis

- **Idea:** Process documents through language-specific pipelines

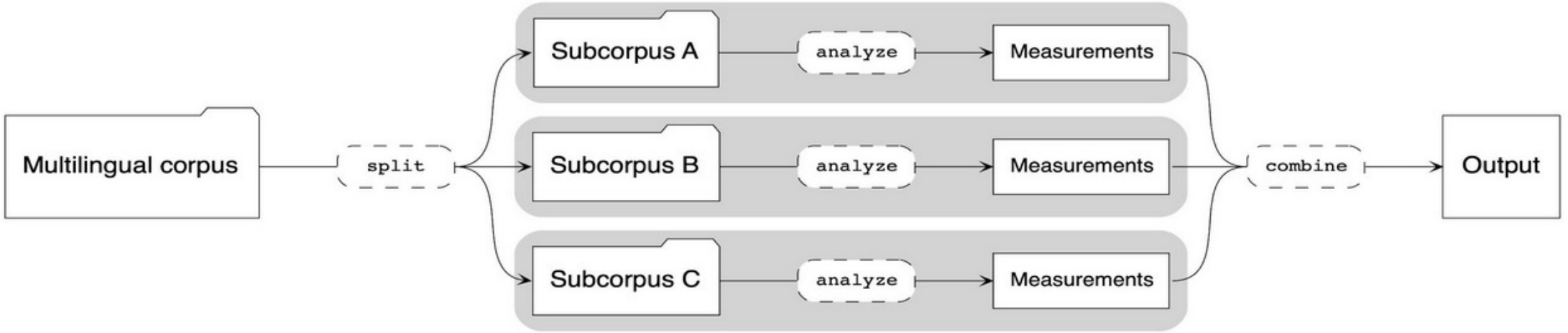


Figure 1 Illustration of the separate analysis strategy to multilingual text analysis.

Licht & Lind, working paper

Approach 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

Approach 2 Input Alignment

- Option 1: Machine translation

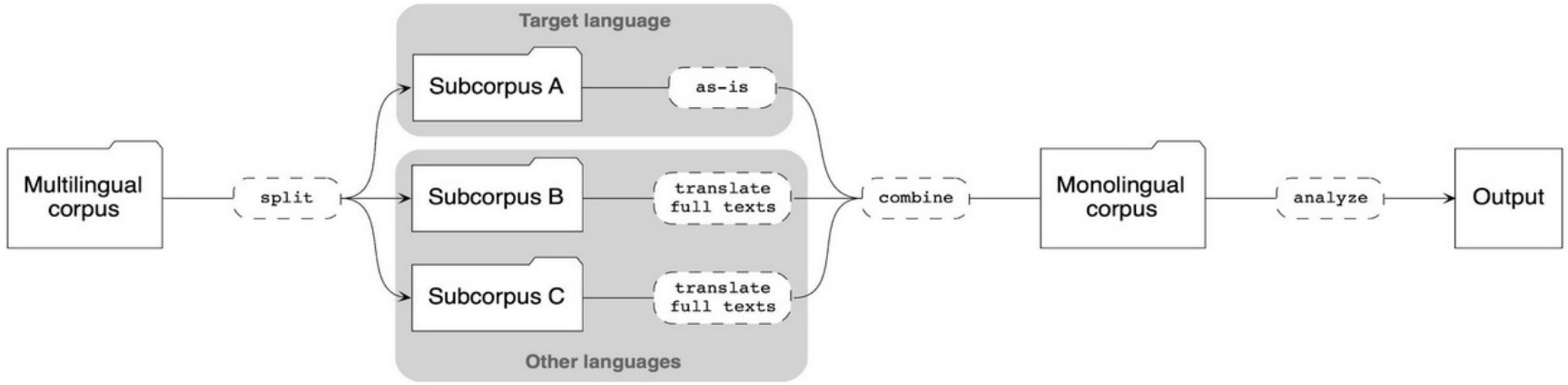


Figure 2 Illustration of the full-text translation approach to input alignment

Licht & Lind, working paper

Approach 2 Input Alignment

- Option 2: Multilingual embedding: 2 versions

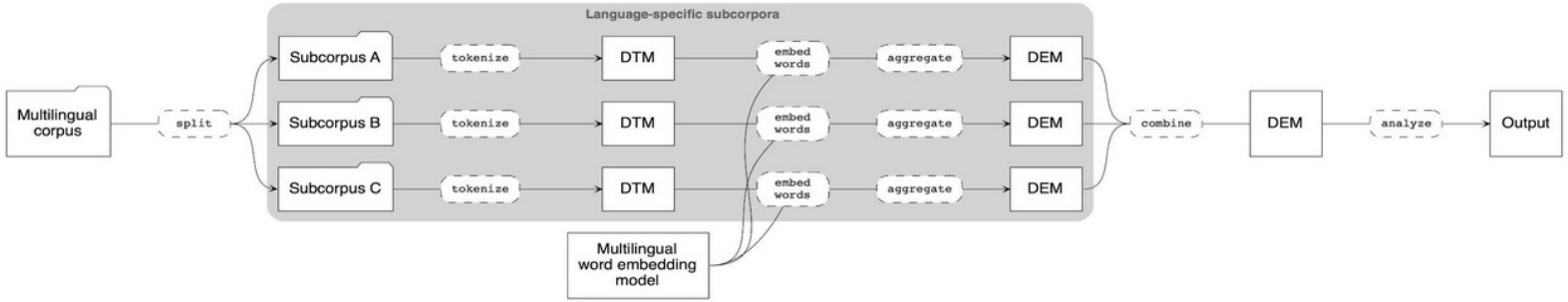


Figure 4 Illustration of the multilingual word embedding approach to input alignment.

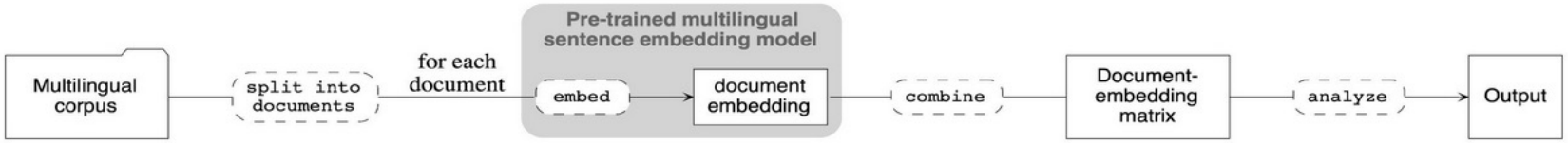


Figure 5 Illustration of the multilingual sentence embedding approach to input alignment

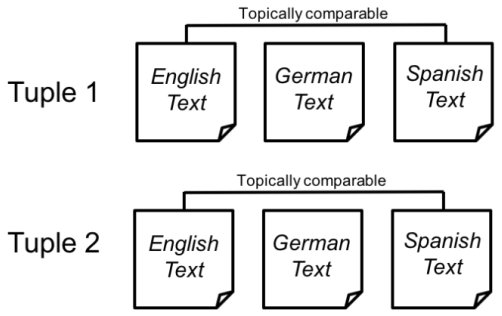
Licht & Lind, working paper

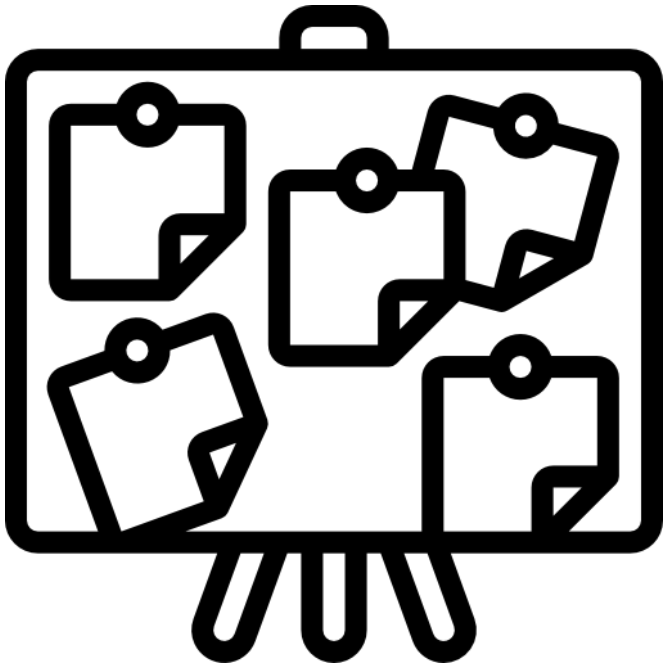
Approach 3 Automated bridging during analysis

- **Idea:** the information how to connect different language is considered automatically during the analysis

- 2 options to provide this information
  - Bilingual lexica
  - Parallel or comparable documents

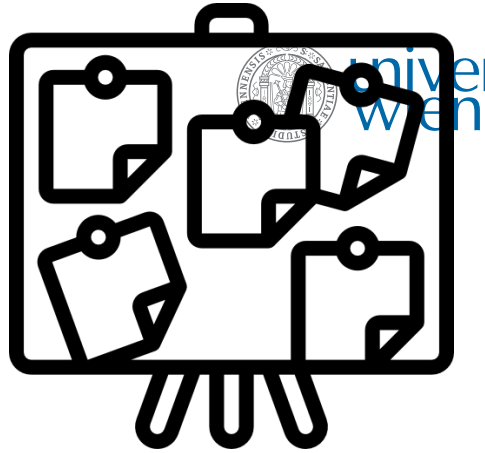
topically comparable documents





How to decide between the 3 approaches?

How to decide between the 3 approaches?



- Let‘s collect pros and cons for each approach
- Offline: Use a new paper for each idea
- Online: write the approach + idea in the chat

Seperate analysis

Input alignment

Automated bridging during analysis

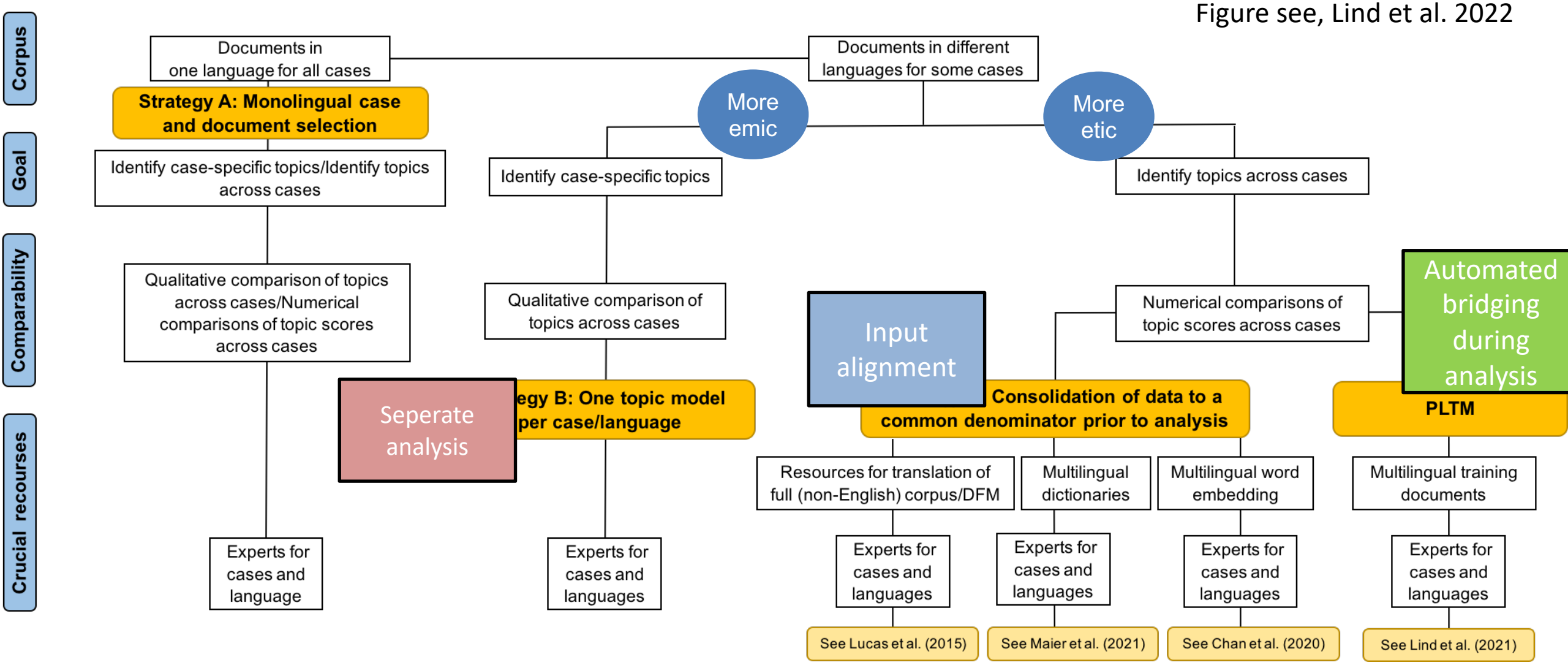
How to decide between the 3 approaches?

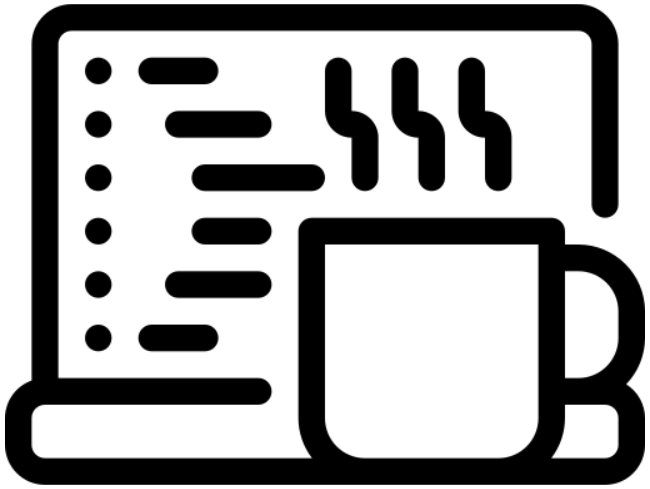
Some criteria

- skills
- availability of instruments (e.g., labeled data, dictionaries)
- type of analysis
- transparency
- replicability

Topic Modeling for Comparative Research

Figure see, Lind et al. 2022





Coffee and Coding

input alignment

- **deeplr** (Zumbach & Bauer, 2021)
  - Machine Translation
  - 26 languages
  - DeepL offers the DeepL API Free which allows a maximum of 500,000 characters/month to be translated for free
  - Tutorial: https://github.com/zumbov2/deeplr

- If dealing with sensitive/confidential text data, checking the data handling procedures of the external services is required

Implementing the input alignment approach

- Data, tasks, and code:

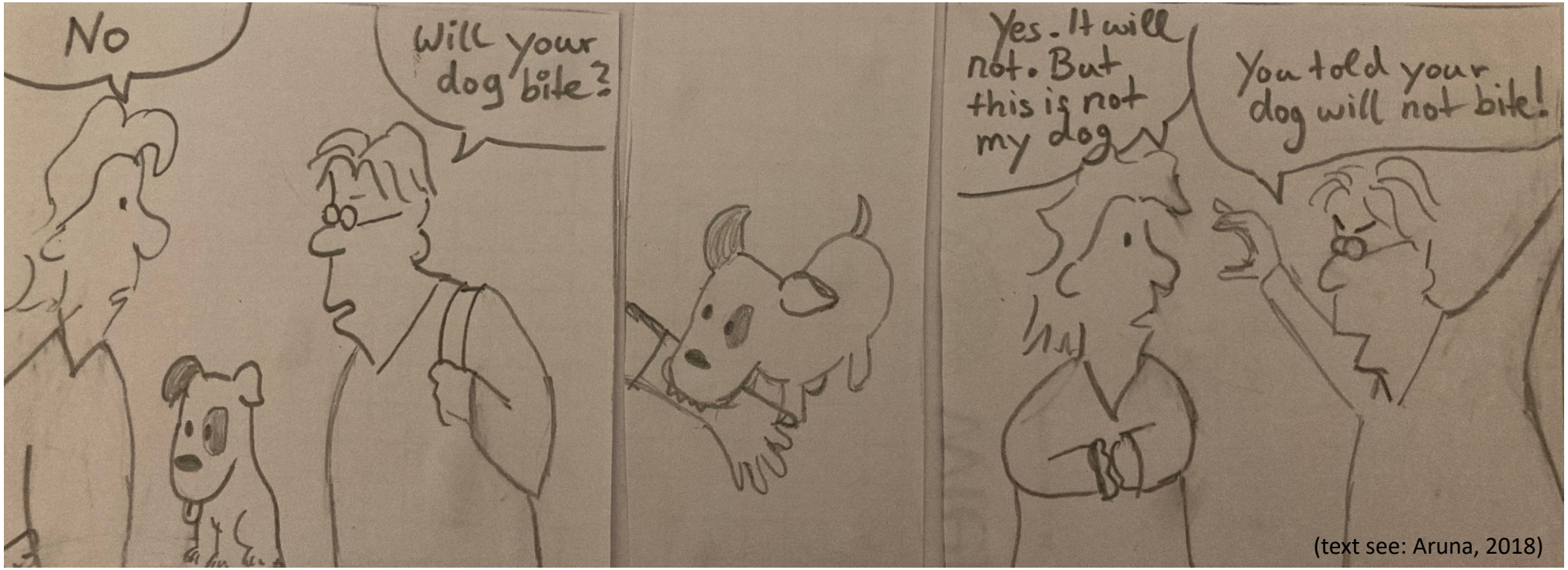
https://github.com/fabiennelind/Workshop_Multilingual-Text-Analysis_and_Comparative-Research

- Tasks: input_alignment_task_description.Rdm
- Commented code to follow along: input_alignment_solution.Rmd

Illustration: Machine translated version of the data set

REMINDER

id	country	publication_date	headline	headline_mt
1	UK	2013-02-09	Asylum girl 'fed up' in UK; COURT	Asylum girl 'fed up' in UK; COURT
2	Spain	2005-06-04	Menores	Minors
3	Spain	2015-11-11	La Comisión considera altamente problemáticas las mec	The Commission considers highly problematic the key measures
4	UK	2012-03-16	Archbishop of Canterbury, Dr Rowan Williams: CV; Profi	Archbishop of Canterbury, Dr Rowan Williams: CV; Profile of the Archbishop of (
5	UK	2012-08-27	France's 'scandalous' expulsion of Roma camps resume	France's 'scandalous' expulsion of Roma camps resumes; French police on Mor
6	Spain	2002-03-13	ISRAEL TOMA LA 'CAPITAL' PALESTINA EN SU MAYOR (	ISRAEL TAKES THE PALESTINIAN 'CAPITAL' IN HIS MAJOR MILITARY OFFENSIV
7	Spain	2006-06-06	EL DRAMA DE LA INMIGRACION / La integracion. Valenc	THE DRAMA OF IMMIGRATION / Integration. Valencia protests the sending of 'w



(text see: Aruna, 2018)

Communication in social contexts

Communication in social contexts

- So far, we focused on semantic equivalence, on the comparability of the literal meaning of words, sentences or documents
- The cartoon implies the need to concentrate also on **pragmatics (= contextual meaning)** (Aruna, 2018)
- Pragmatics is concerned with the use of language in social contexts and the ways in which people comprehend meanings (Aruna, 2018)
- As social scientists, we are typically interested in communication that happens in social situations
- As social scientist interested in comparative research, we also care for social cultural economic contexts of the cases that we compare

Another example to illustrate the context dependency of language

“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

Pragmatics and comparative research

- How can we consider the pragmatic dimension of language when we compare e.g. countries based on corpora in different languages?
- I argue that we can try our best to include our communication science empowered contextual knowledge about the communicators and the audiences (per case) in each step when setting up the empirical design
- This way we can strive for equivalence across cases that considers also the pragmatic dimension of language

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)

- Two approaches to comparability:

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

(Livingstone, 2003)

Emic & Etic

- Emic and etic as 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.

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

Constructs in comparative research

- ‘Construct equivalence’ denotes the search for a shared understanding or interpretability of the construct to be studied with comparative content analysis.

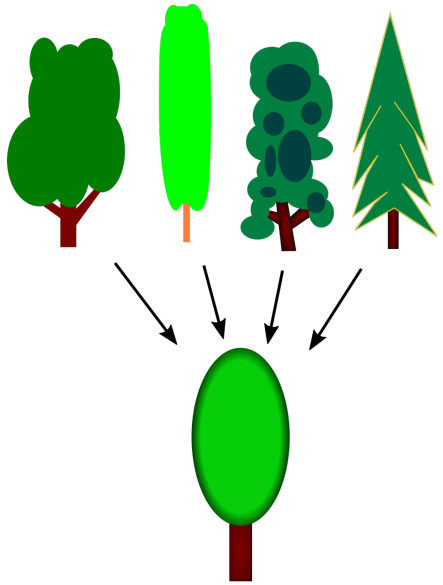
Selecting and defining constructs:

- Requires joint effort by case experts
- Is the concept translatable? Is it useful to be studied across cases?
- Decision to go for rather ‘emic’ or ‘etic’ definition

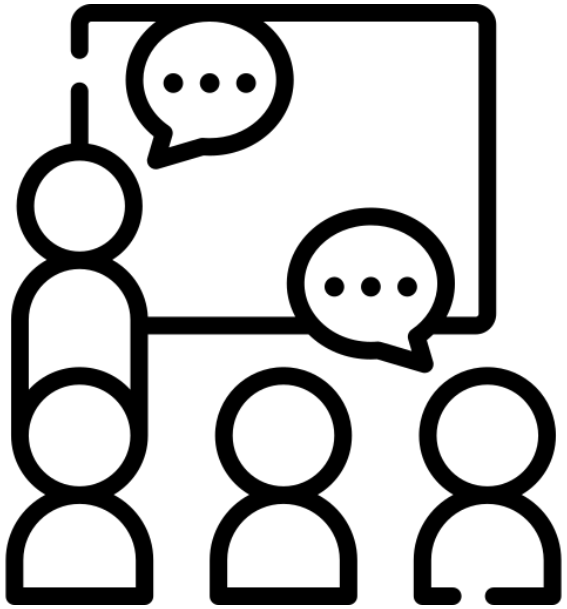
(Boomgaarden & Song, 2019, Esser & Vliegenthart, 2017)

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)



A language and context sensitive validation framework

Motivation to design a framework

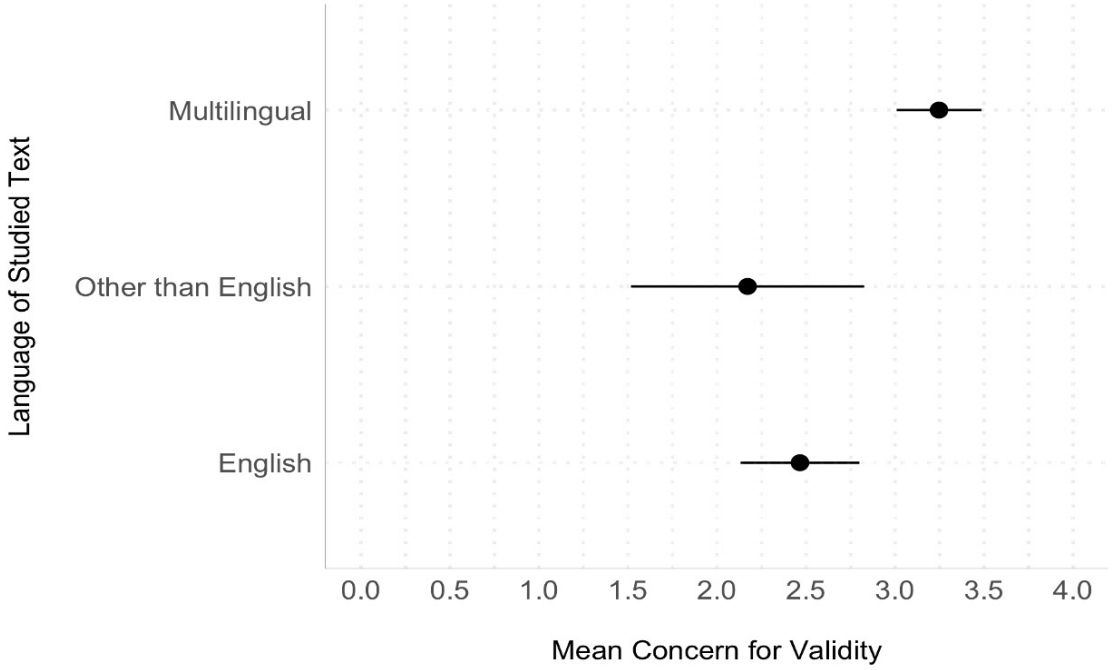
- insights from a **content analysis of published literature** in the social sciences and an **expert survey** with the respective authors

- both studies conduced within the OPTED project

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

Validation concern

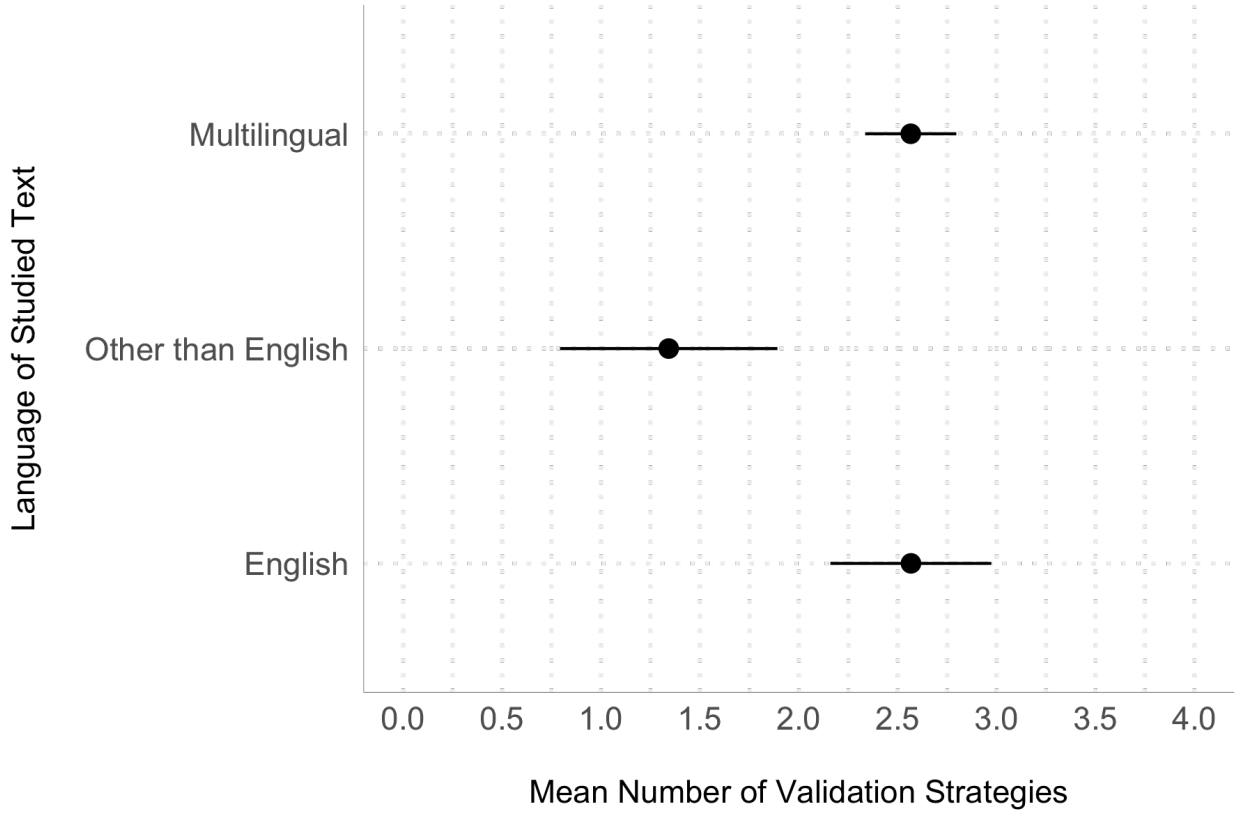
- Among researchers in our sample, those who work **in more than one language,** express more concerns about the validity of findings from computational methods



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

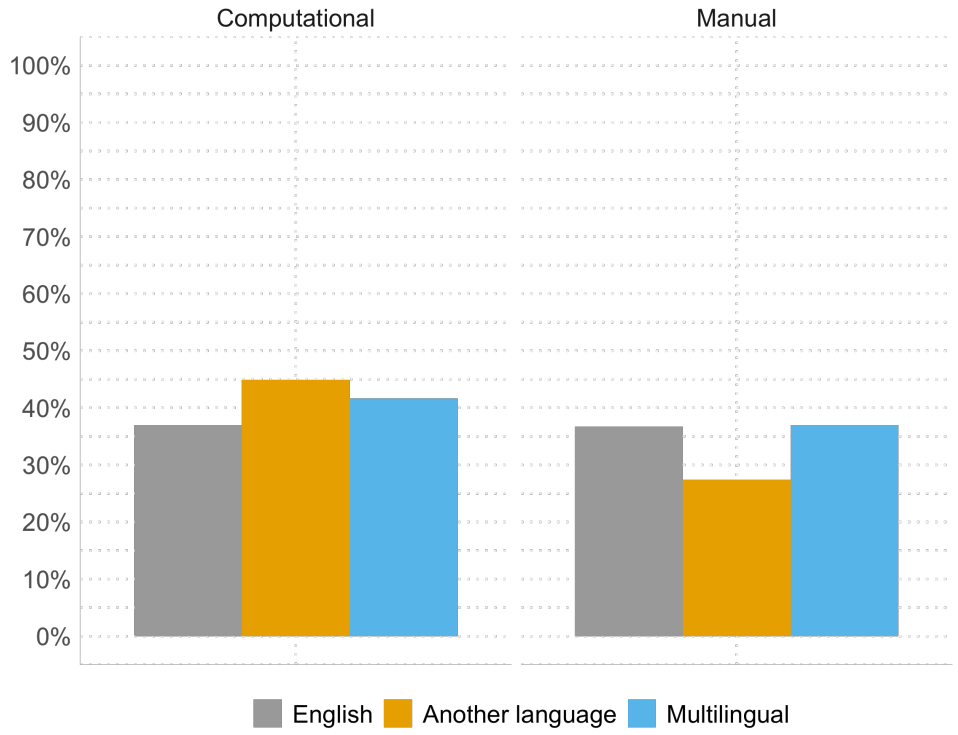
Validation strategies

- But this is not reflected **in a more extensive focus** on validation



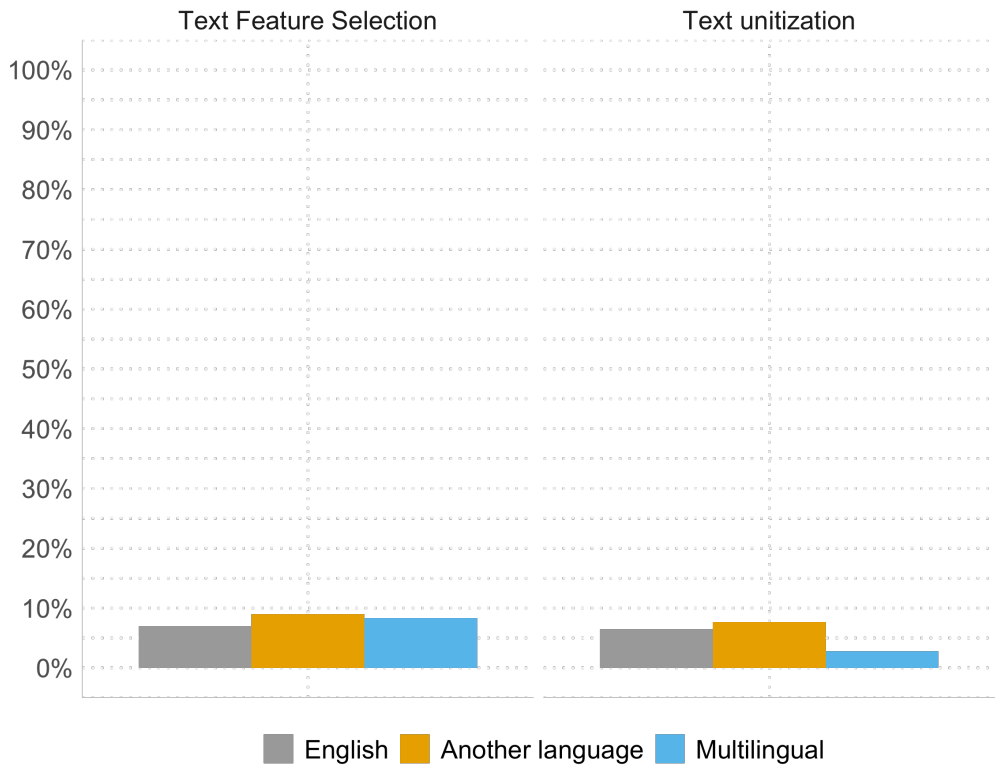
1. Data validation

- 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.



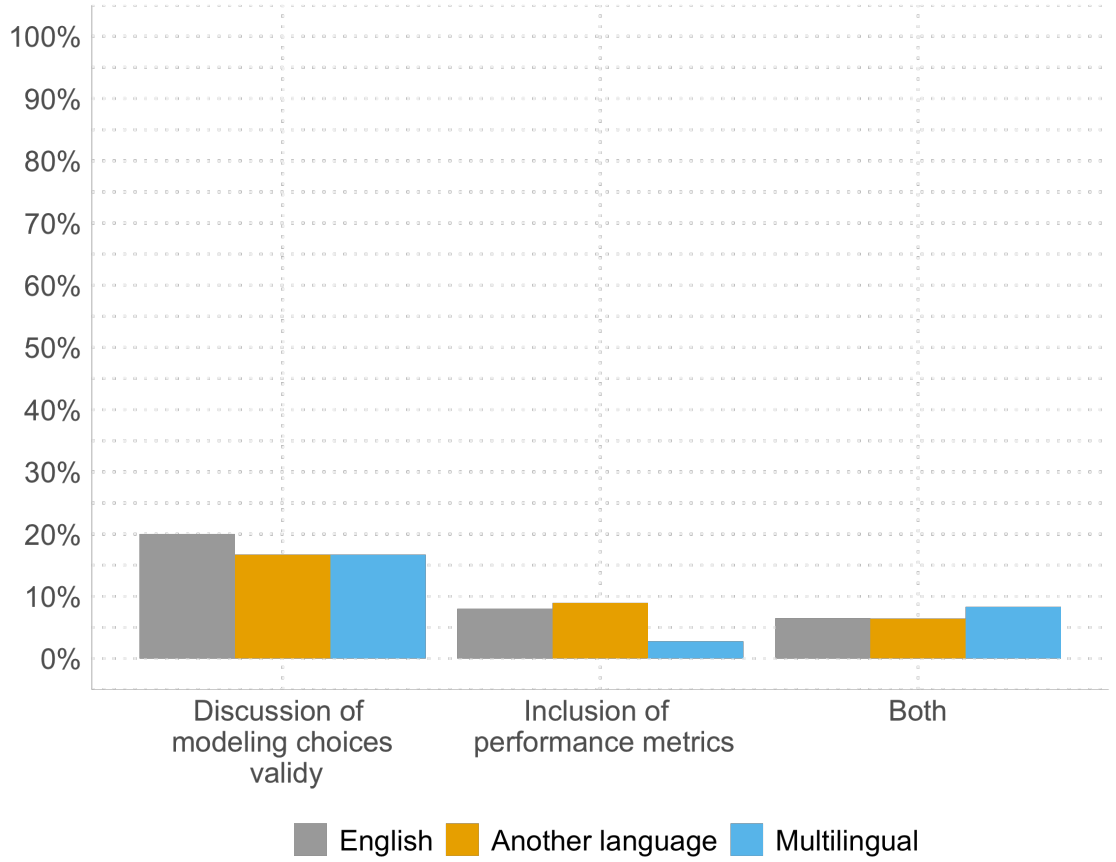
2. Input validation

- **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?*)



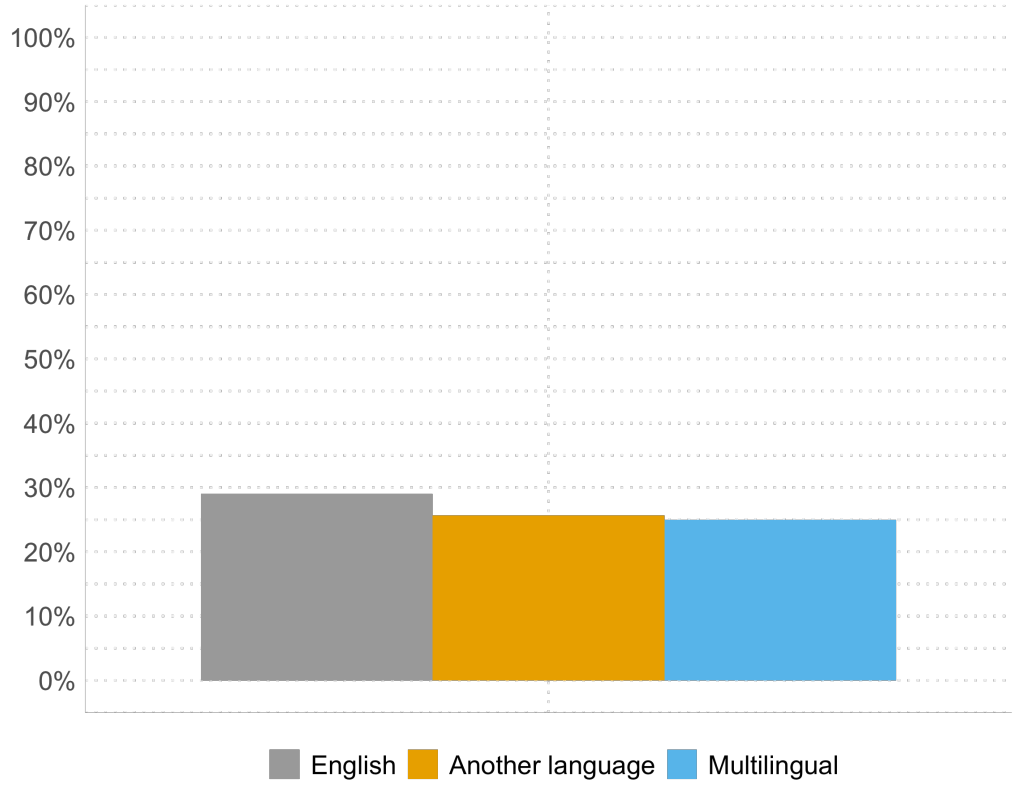
3. Process validation

- 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.



4. Output validation against human-coded benchmark

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



Validation framework (a first attempt)

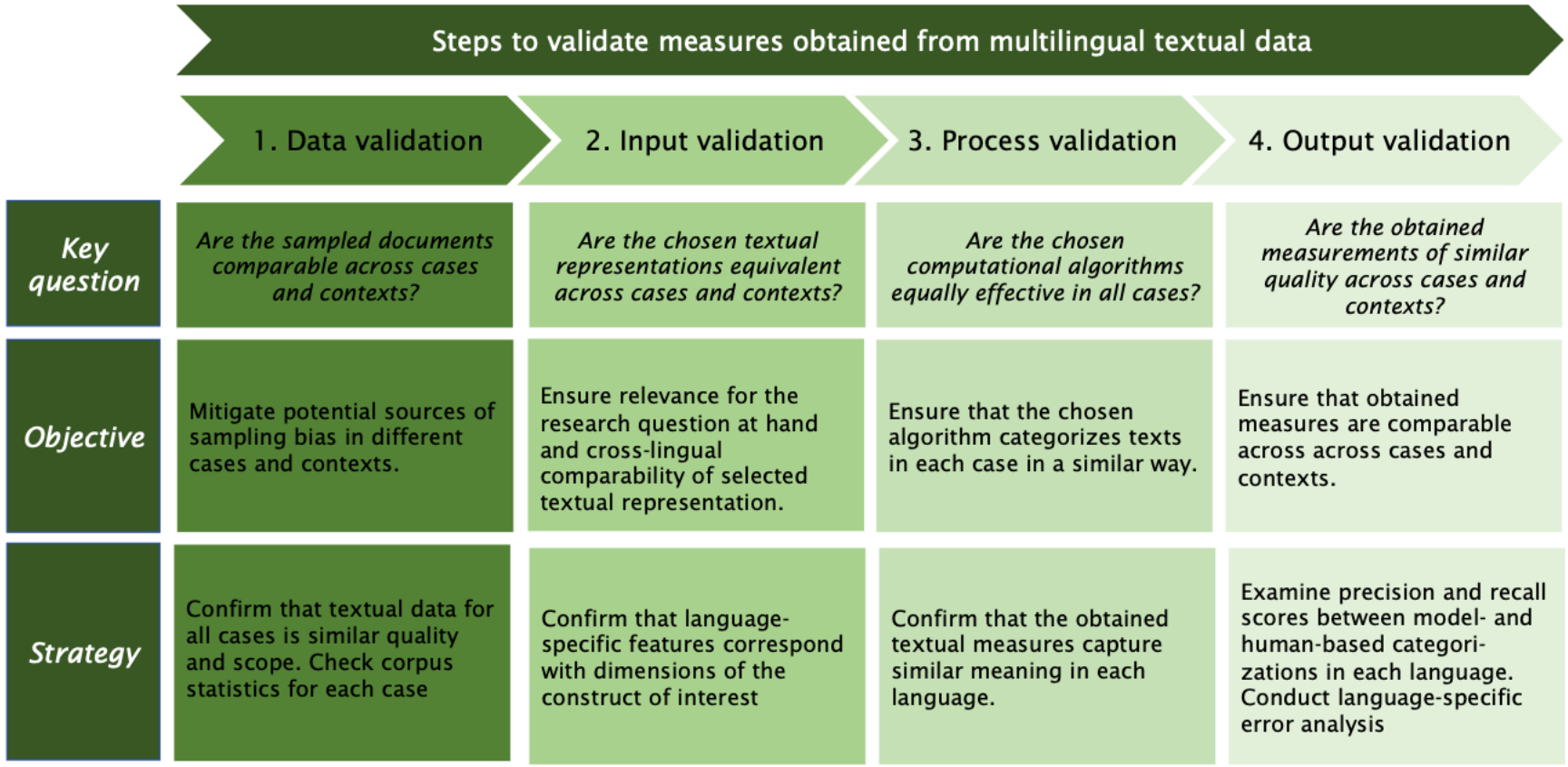


Figure 10 A VALIDATION FRAMEWORK FOR MEASUREMENTS OBTAINED FROM MULTILINGUAL TEXTUAL DATA TO BE USED FOR COMPARATIVE RESEARCH

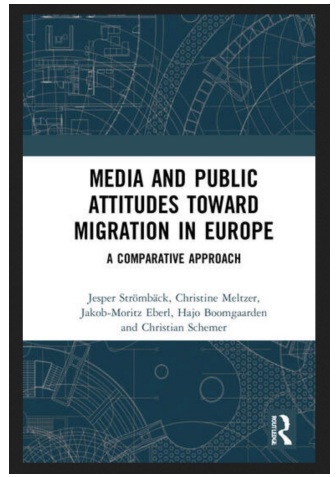
Baden, Dolinsky, Lind, Pipal, Schoonvelde, Shababo, & van der Velden, 2022

Planning a research design for a multilingual automated text analysis that can be used for a comparative social science question

Illustration with an example

REMINDER

REMINDER project: European media discourse about migration



Project goals:

- The measurement and comparison of frames and topics in migration news articles in 7 countries
- Prepare media content measures to be connected with panel survey data

Data:

- Multilingual annotated news article corpus (Heidenreich et al., 2020; Lind et al., 2020)
- For this workshop we use a subset of the data
  - Selected Cases: Germany, Spain, and the UK
  - Document Languages: German, Spanish, English

This project has received funding from the EU's Horizon2020 research and innovation program under the grant agreement no 727072

The effort to approach validity on four levels

1. data
2. input
3. process
4. output

1. Data validation

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

a) Finding equivalent document sources

- Text corpora by political organizations and Legislative text corpora: https://opted.eu/results/inventories/
- European media sources: https://meteor.opted.eu/

b) Retrieval of equivalent documents

- via multilingual search strings, validated (!)

Illustration: Creation and validation of a multilingual search string

REMINDER

Finding equivalent document sources:

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

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*"

Illustration: Multilingual news article sample (a snippet)

REMINDER

id	country	publication_date	source	source_type	headline
1	UK	2013-02-09	Daily Mirror	Print	Asylum girl 'fed up' in UK; COURT
2	Spain	2005-06-04	El Pais	Print	Menores
3	Spain	2015-11-11	El Pais	Print	La Comisión considera altamente problemáticas las medidas clave
4	UK	2012-03-16	telegraph.co.uk	Online	Archbishop of Canterbury, Dr Rowan Williams: CV; Profile of the Archbishop of Cant
5	UK	2012-08-27	telegraph.co.uk	Online	France's 'scandalous' expulsion of Roma camps resumes; French police on Monday
6	Spain	2002-03-13	El Pais	Print	ISRAEL TOMA LA 'CAPITAL' PALESTINA EN SU MAYOR OFENSIVA MILITAR EN VEIN
7	Spain	2006-06-06	El Mundo	Print	EL DRAMA DE LA INMIGRACION / La integracion. Valencia protesta por el envio de '

Lind et al., 2020

What about constructs?

- Reflection on emic/etic construct definitions

With case and language experts recommended

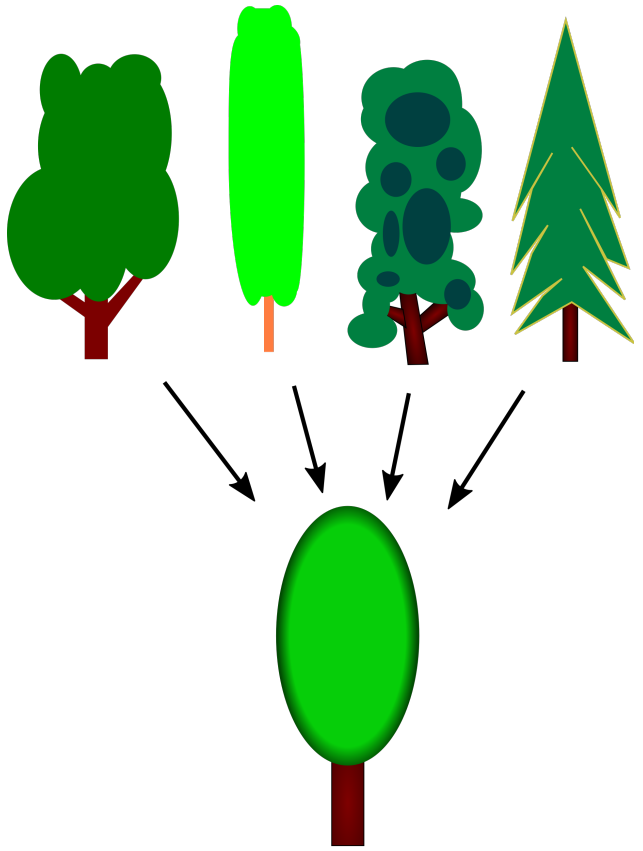


Illustration: Construct definition

REMINDER

Process: Support by case and language experts

Construct studied:

- **Frames:** 'Etic' construct definition, decision to search for four frames dominant in the literature (migration as 1. economy, 2. labor, 3. welfare, 4. security topic) (Definitions: Lind et al., 2020)
- **Topics:** 'Etic' construct definition, decision to search for the most prevalent topics for the entire multilingual collection (the European discourse on migration) and not case per case
- **Women and Men Migrant:** Decision to study this construct only for one case (Lind & Meltzer, 2020)

2. Input validation

Key questions: Are the chosen textual representations equivalent across cases and contexts? Are they relevante for the research question at hand?

For emic (=case-sensitive) measurments:

- Approach 1: try to preserve context per language/case during pre-processing

- Approach 2:
  - Machine translation: ideally add translation checks
  - 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)

3. Process validation

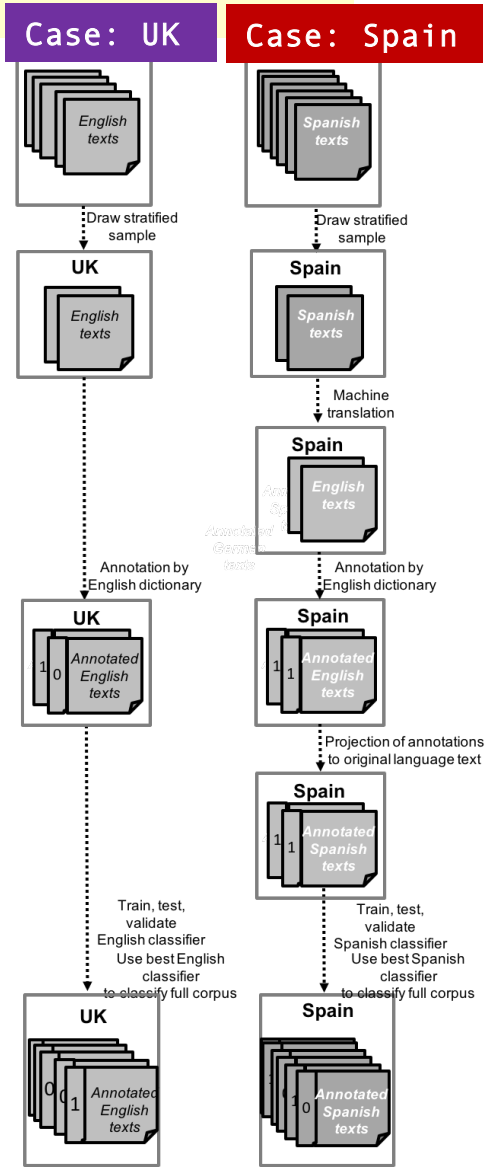
- Key questions: Are the chosen computational algorithms equally effective in all cases?

- Discuss 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 (e.g., lemmatizer, machine translation software, multilingual word embeddings)
- Use case-specific labelled documents for supervised classification

Illustration: Supervised text classification for multilingual corpora

- Document sample translation into English
- Document sample annotation with an English dictionary
- The dictionary is designed to incorporate case-specific and transnational aspects (‘etic’ definition)
- It is validated with a manually annotated subset (includes documents for all cases)
- The classifiers are trained with case-specific texts

- More details: https://github.com/Christoph/MultilingualTextAnalysis



4. Output validation

- Concerned with the quality of obtained measures and their equivalence across languages and across cases
- Assessed with a comparison of the estimates with an established benchmark
- 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)
- Comparison: examine recall and precision as well as the corresponding misclassifications
- In multilingual applications, output validation needs to be considered for each included language and case

Baseline 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: Quality of self-created baseline

REMINDER

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.

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 for multilingual corpora

REMINDER

Frame measurement: 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 mec	0	0	0	0
4	UK	2012-03-16	telegraph.co.uk	Online	Archbishop of Canterbury, Dr Rowan Williams: CV; Profi	0	0	0	0
5	UK	2012-08-27	telegraph.co.uk	Online	France's 'scandalous' expulsion of Roma camps resume	0	1	1	0
6	Spain	2002-03-13	El Pais	Print	ISRAEL TOMA LA 'CAPITAL' PALESTINA EN SU MAYOR (	0	0	0	0
7	Spain	2006-06-06	El Mundo	Print	EL DRAMA DE LA INMIGRACION / La integracion. Valenc	0	0	1	0

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

Illustration: Evaluation of 1 example topic

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

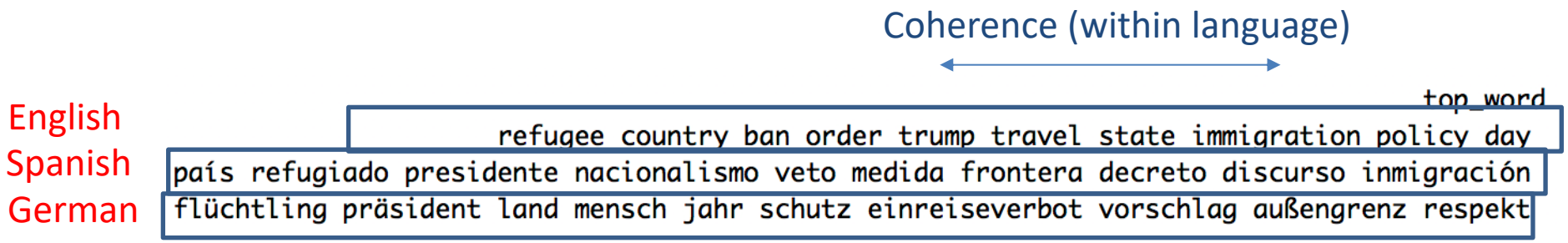


Illustration: Evaluation of 1 example topic

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

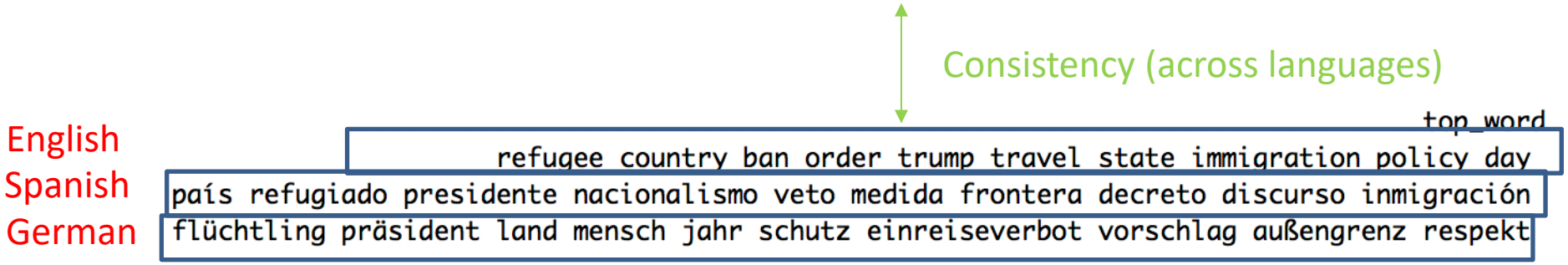


Illustration: Manual labeling of the final model

REMINDER

Table 1. 20 multilingual topics

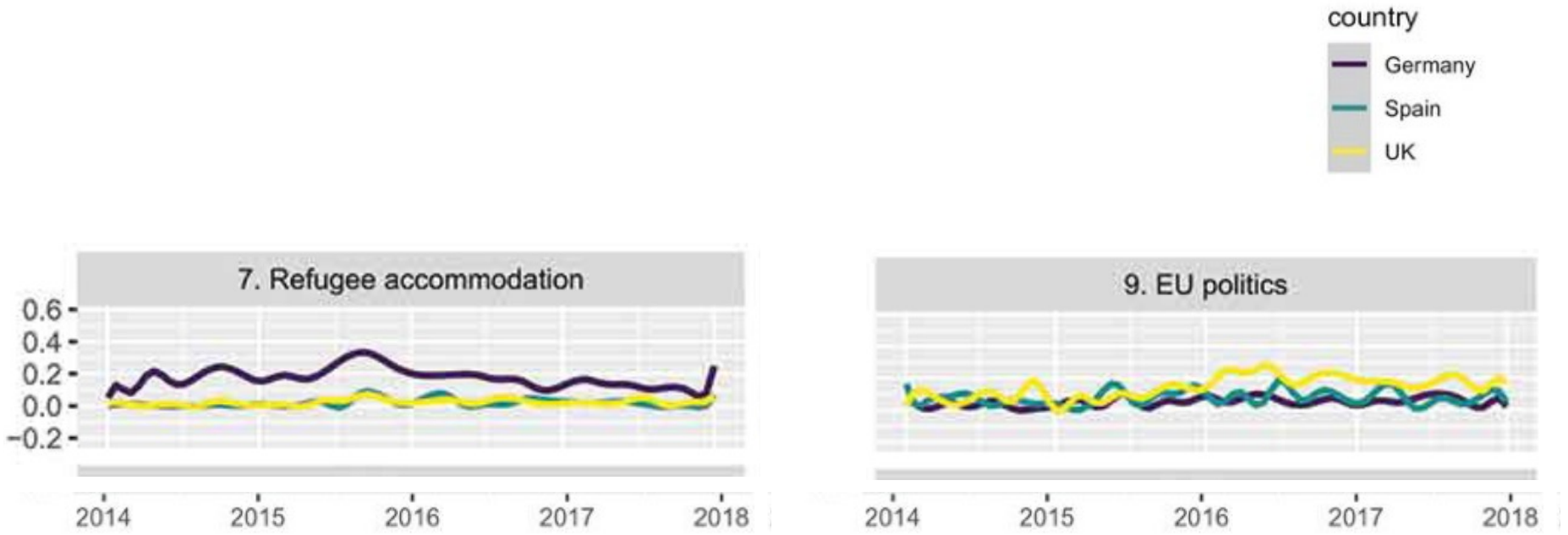
Topic	Lang.	Top 10 words
1.Welfare & jobs	EN	people worker immigration country migrant benefit report figure job health
	ES	persona número trabajador inmigración país millón inmigrante aumento cifra beneficio
	DE	zahl prozent land million migranten arbeit einwanderung bericht bevölkerung problem
2. Education	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

(An excerpt)

Lind et al., 2022

Illustration: Evaluating the validity of the final model (1/2)

- **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

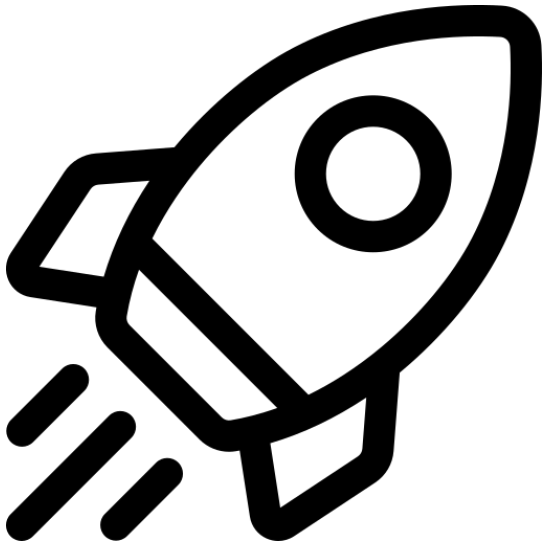
Ilustration: Evaluating the validity of the final model (2/2)

- **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



Resources to get started

Central components

1. Experts for each language and case (for concept definition, instrument creation, ...)

2. Linguistic knowledge about the languages of your documents

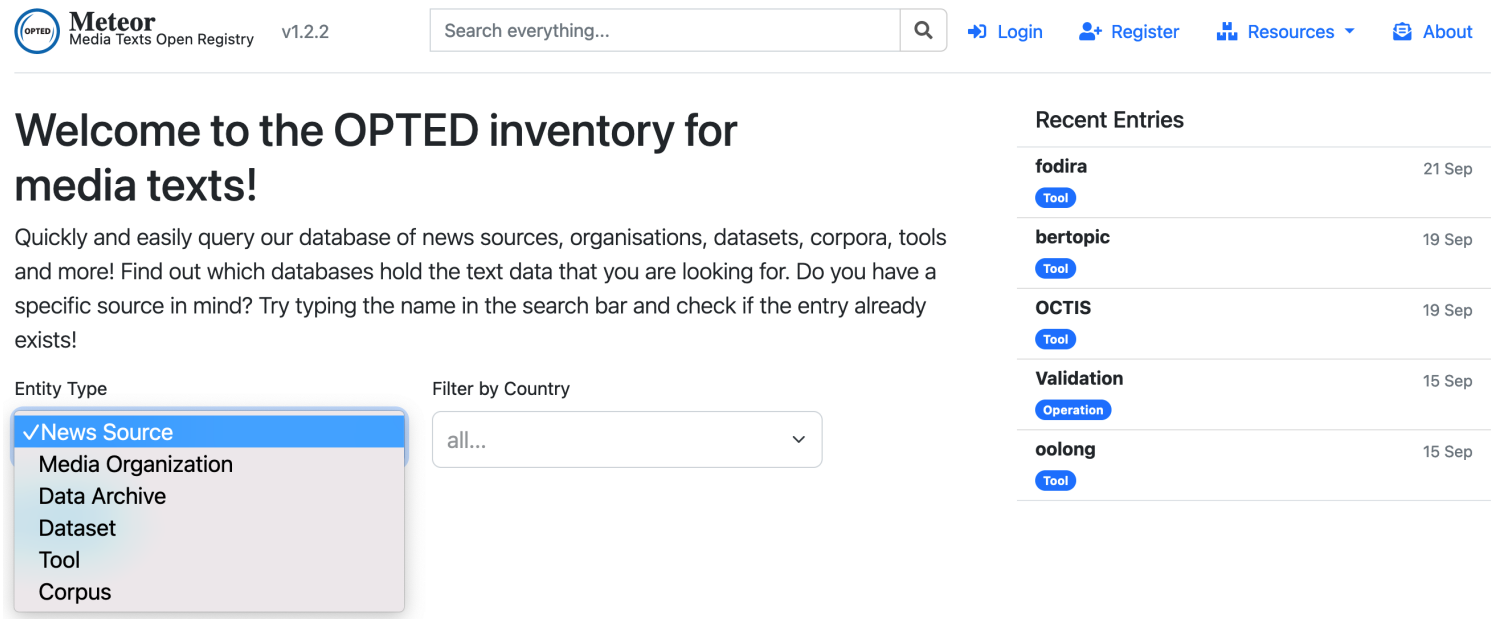
3. Budget for validation

OPTED Services

- Resources specifically for political text analysis www.opted.eu

One example: Meteor

- https://meteor.opted.eu/
- Inventory for news sources tools, datasets and more

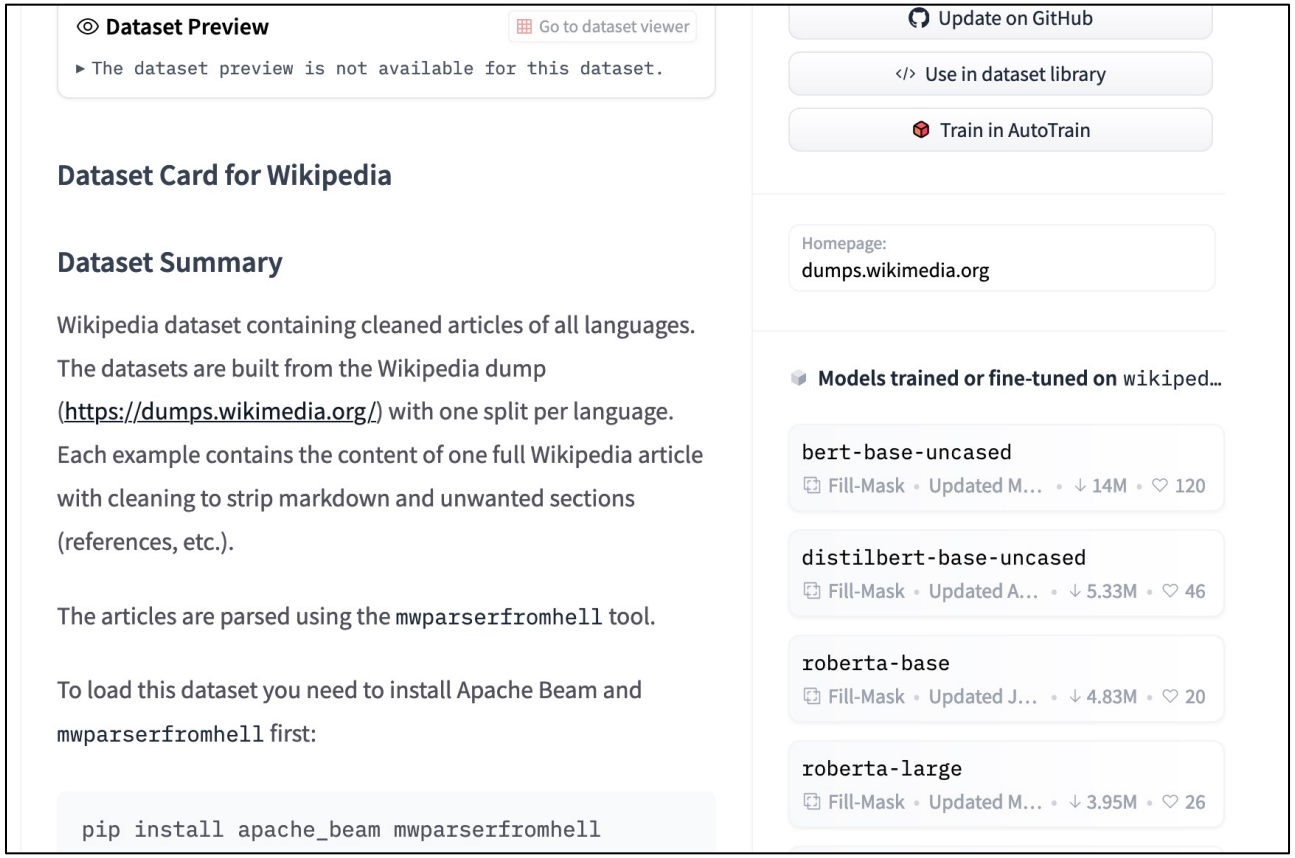


Resources: Hugging Face

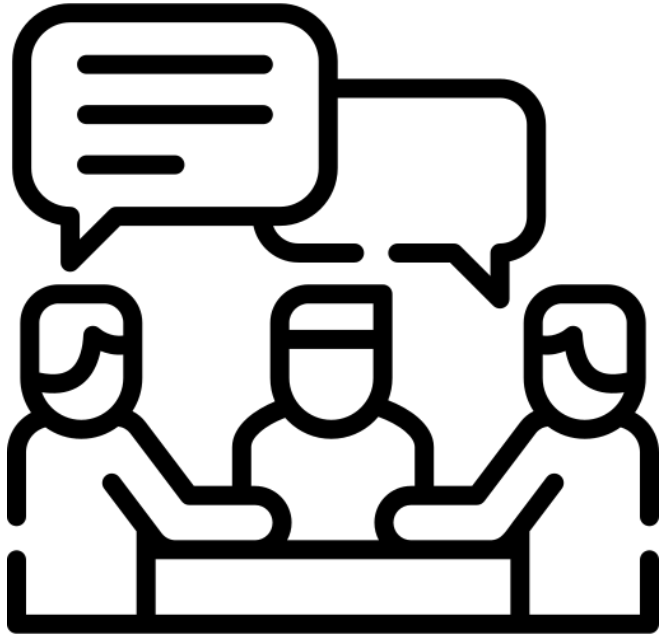
- **AI community** with numerous multilingual resources such as datasets and language models
- https://huggingface.co/

Hugging Face

Example Dataset



https://huggingface.co/datasets/wikipedia



Themed tables & Wrap-up

Topics

How useful are the measures obtained? What conclusions can be drawn in the end?

Multilingual embeddings and transformer architecture. Experiences and reflections on validation

???

What scenarios recommend a (partly) automated analysis? when is a manual analysis the better choice?

Space to chat about your own research project (ideas)?

Thank you very much for joining

Happy to discuss project ideas and their implementation, feel free to send a message fabienne.lind@univie.ac.at or @FabienneLind

Further readings

- Esser, F., & Vliegenthart, R. (2017). Comparative research methods. *The International Encyclopedia of Communication Research Methods.* Link.
- Lind, F. (2021). *Multilingual Automated Content Analysis for Comparative Communication Research*. (Doctoral Dissertation, University of Vienna). Happy to send you a copy.
- Livingstone, S. (2003). On the challenges of cross-national comparative media research. *European Journal of Communication, 18(4*), 477–500. Link.
- Lucas, C., Nielsen, R. A., Roberts, M. E., Stewart, B. M., Storer, A., & Tingley, D. (2015). Computer-assisted text analysis for comparative politics. *Political Analysis, 23*(2), 254–277. Link.
- Rössler, P. (2012). Comparative content analysis. In F. Esser & T. Hanitzsch (Eds.), The handbook of comparative communication research (pp. 481-490). Routledge.

References

- Abdelwahab, A., Robles, J., Chiru, C. G., & Rebedea, T. (2014). Tweets topic modelling across different countries. eLearning & Software for Education, 1, 134–141.
- Aruna, U. (2018). Pragmatic Equivalence in Translation. Journal of Emerging Technologies and Innovative Research, 5(10). https://www.jetir.org/papers/JETIR1810810.pdf
- Baden, C. Dolinsky,A. Lind, F., Pipal,C., Schoonvelde, M., Shababo, G., & van der Velden, M. (2022). Integrated standards and context-sensitive recommendations for the validation of multilingual computational text analysis. OPTED Deliverable.
- Baden, C., & Stalpouskaya, K. (2015). Common methodological framework: Content Analysis—A mixed methods strategy for comparatively, diachronically analyzing conflict discourse (INFOCORE Working Paper 2015/10). http://www.infocore.eu/wpcontent/uploads/2016/02/Methodological-Paper-MWGCA_final.pdf
- Benoit, K., Schwarz, D., & Traber, D. (2012, June). The sincerity of political speech in parliamentary systems: A comparison of ideal points scaling using legislative speech and votes. Paper presented at the Second Annual Conference of European Political Science Association, Berlin, Germany.
- Chan, C. H., Zeng, J., Wessler, H., Jungblut, M., Welbers, K., Bajjalieh, J. W., ... & Althaus, S. L. (2020b). Reproducible extraction of cross-lingual topics (rectr). Communication Methods and Measures, 14(4), 285–305. https://doi.org/10.1080/19312458.2020.1812555
- Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Esser, F., & Hanitzsch, T. (2012). On the why and how of comparative inquiry in communication studies. In F. Esser & T. Hanitzsch (Eds.), Handbook of comparative communication research (pp. 3-22). Routledge.
- Esser, F., & Vliegenthart, R. (2017). Comparative research methods. The International Encyclopedia of Communication Research Methods. https://doi.org/10.1002/9781118901731.iecrm0035
- Gründl, J. (2020). Populist ideas on social media: A dictionary-based measurement of populist communication. New Media & Society. https://doi.org/10.1177/1461444820976970

References

- Heidenreich, T., Lind, F., Eberl, J.-M., Galyga, S., Edie, R., Herrero-Jiménez, B., Gómez Montero, E.L., Berganza, R., & Boomgaarden, H.G. (2020). REMINDER: Short term media analysis on migration 2017-2018 (OA edition), [Data set and documentation]. AUSSDA Dataverse. https://doi.org/10.11587/T86DVG
- Krippendorff, K. (2004). Reliability in content analysis. Human Communication Research, 30(3), 411–433.
- Licht, Hauke (2021) "Language-agnostic supervised classification of political texts using multilingual embeddings". Paper presented at the Annual Conference of the Swiss Political Science Association, 4th and 5th of February 202
- Lind, F., Eberl, J. M., Eisele, O., Heidenreich, T., Galyga, S., & Boomgaarden, H. G. (2021). Building the Bridge: Topic Modeling for Comparative Research. *Communication Methods and Measures*, 1-19. https://doi.org/10.1080/19312458.2021.1965973
- Lind, F., Heidenreich, T., Eberl, J.-M., Galyga, S., Edie, R., Herrero-Jiménez, B., Gómez Montero, E.L., Berganza, R., & Boomgaarden, H.G. (2020). REMINDER: Historical media analysis on migration 2003-2017 (OA edition), [Data set and documentation].
- Lind, F., & Meltzer, C. E. (2020). Now you see me, now you don't: Applying automated content analysis to track migrant women's salience in German news. *Feminist Media Studies*, 1-18.
- AUSSDA Dataverse. https://doi.org/10.11587/IEGQ1B
- Livingstone, S. (2003). On the challenges of cross-national comparative media research. European Journal of Communication, 18(4), 477–500. https://doi.org/10.1177/0267323103184003
- Lucas, C., Nielsen, R. A., Roberts, M. E., Stewart, B. M., Storer, A., & Tingley, D. (2015). Computer-assisted text analysis for comparative politics. Political Analysis, 23(2), 254–277.
- Maier, D., Baden, C., Stoltenberg, D., De Vries-Kedem, M., & Waldherr, A. (2021). Machine Translation Vs. Multilingual Dictionaries Assessing Two Strategies for the Topic Modeling of Multilingual Text Collections. *Communication Methods and Measures*, 1-20. https://doi.org/10.1080/19312458.2021.1955845
- McLeod, J., & Blumler, J. (1987). The macrosocial level of communication science. In C. R. Berger & S. H. Chaffee (Eds.), Handbook of communication science (pp. 271-322). Sage.
- Mimno, D., Wallach, H., Naradowsky, J., Smith, D. A., & McCallum, A. (2009). Polylingual topic models. In P. Koehn & R. Mihalcea (Eds.), Proceedings of the 2009 conference on empirical methods in natural language processing (pp. 880-889). Association for Computational Linguistics. https://www.aclweb.org/anthology/D09-1000

References

- Rössler, P. (2012). Comparative content analysis. In F. Esser & T. Hanitzsch (Eds.), The handbook of comparative communication research (pp. 481-490). Routledge.
- Song, H., Tolochko, P., Eberl, J.-M., Eisele, O., Greussing, E., Heidenreich, T., ... & Boomgaarden, H. G. (2020). In validations we trust? The impact of imperfect human annotations as a gold standard on the quality of validation of automated content analysis. Political Communication, 37(4), 550–572. https://doi.org/10.1080/10584609.2020.1723752
- Stryker, J. E., Wray, R. J., Hornik, R. C., & Yanovitzky, I. (2006). Validation of database search terms for content analysis: The case of cancer news coverage. *Journalism & Mass Communication Quarterly*, *83*(2), 413-430. https://doi.org/10.1177/107769900608300212
- Swanson, D. (1992). Managing theoretical diversity in cross-national studies of political communication. In J. G. Blumler, J. M. McLeod, & K. E. Rosengren (Eds.), Comparatively speaking: Communication and culture across space and time (pp. 19-34). Sage.
- Watanabe, K. (2020). Latent Semantic Scaling: A semisupervised text analysis technique for new domains and languages. Communication Methods and Measures. https://doi.org/10.1080/19312458.2020.1832976
- Wijffels, J., Straka, M., & Strakov, J. (2019). Udpipe: Tokenization, parts of speech tagging, lemmatization and dependency parsing with the UDPipe NLP toolkit. (R Package Version 0.6). [Computer software]. https://cran.rproject.org/web/packages/udpipe/index.html
- Zumbach, D. & Bauer, P.C. (2021). deeplr: Interface to the 'DeepL' Translation API. [Computer software]. https://CRAN.R-project.org/package=deeplr

Credits

- Icons: www.flaticon.com