Multilingual Automated Content Analysis for Comparative Social Science Research

Fabienne Lind

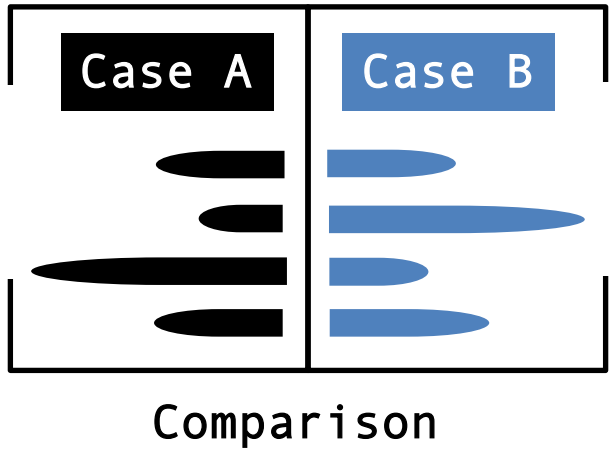
Computational Communication Science Lab,
Department of Communication, University of Vienna

Agenda

- Comparative social science & (automated) content analysis
- Step-by-step guidelines: Planning a research design
- Discussion round

Comparative social science

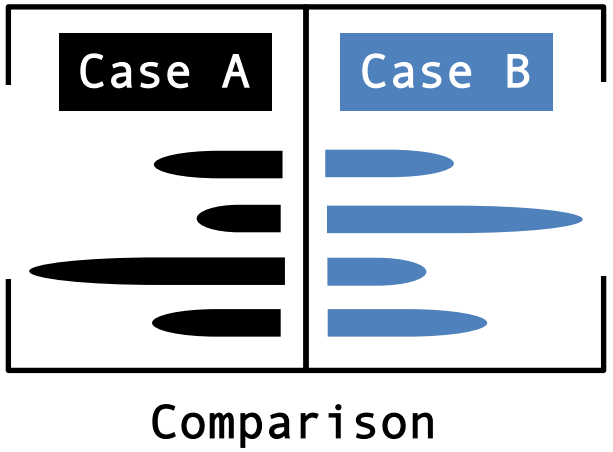
- **Comparative research in social science** involves comparisons between a minimum of two cases with at least one object of investigation relevant to social science research.
- **Cases** are macro-level units such as systems, cultures, countries, and markets)



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

Reasons to compare cases

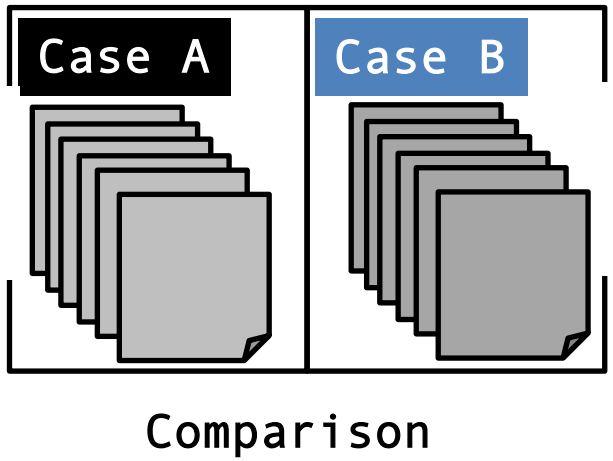
- insights into the differences and similarities of cases
- improved understanding and contextualization of the own case
- raised awareness for other cases
- the test and generalizability of theories across diverse settings
- the investigation of transnational process across contexts



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

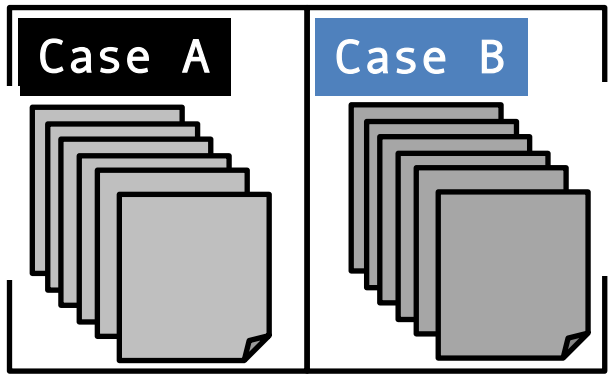
Comparison of cases with content analysis

- 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



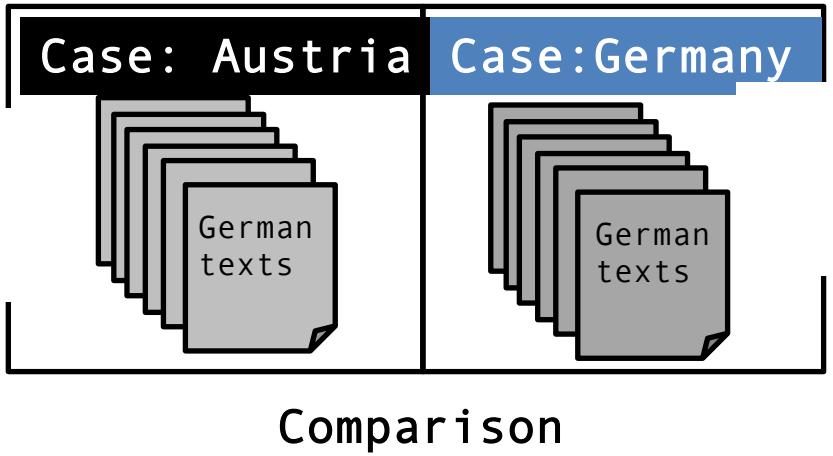
Comparison

For which topics could such a method be useful?

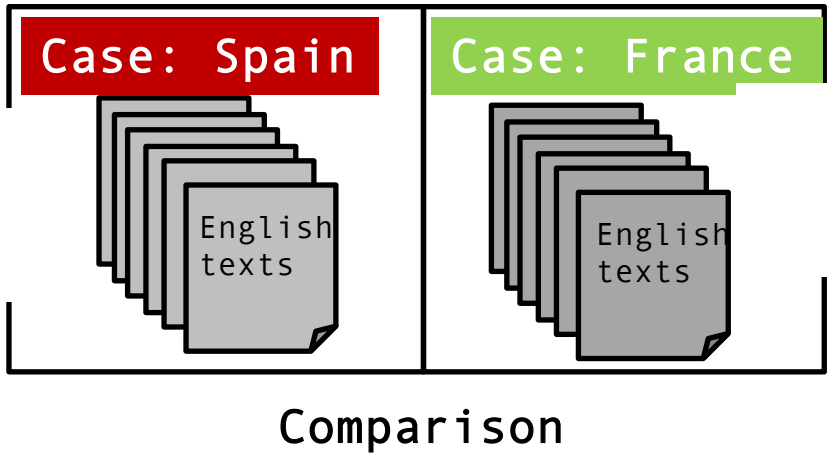
Comparison of cases & the language(s) of documents

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

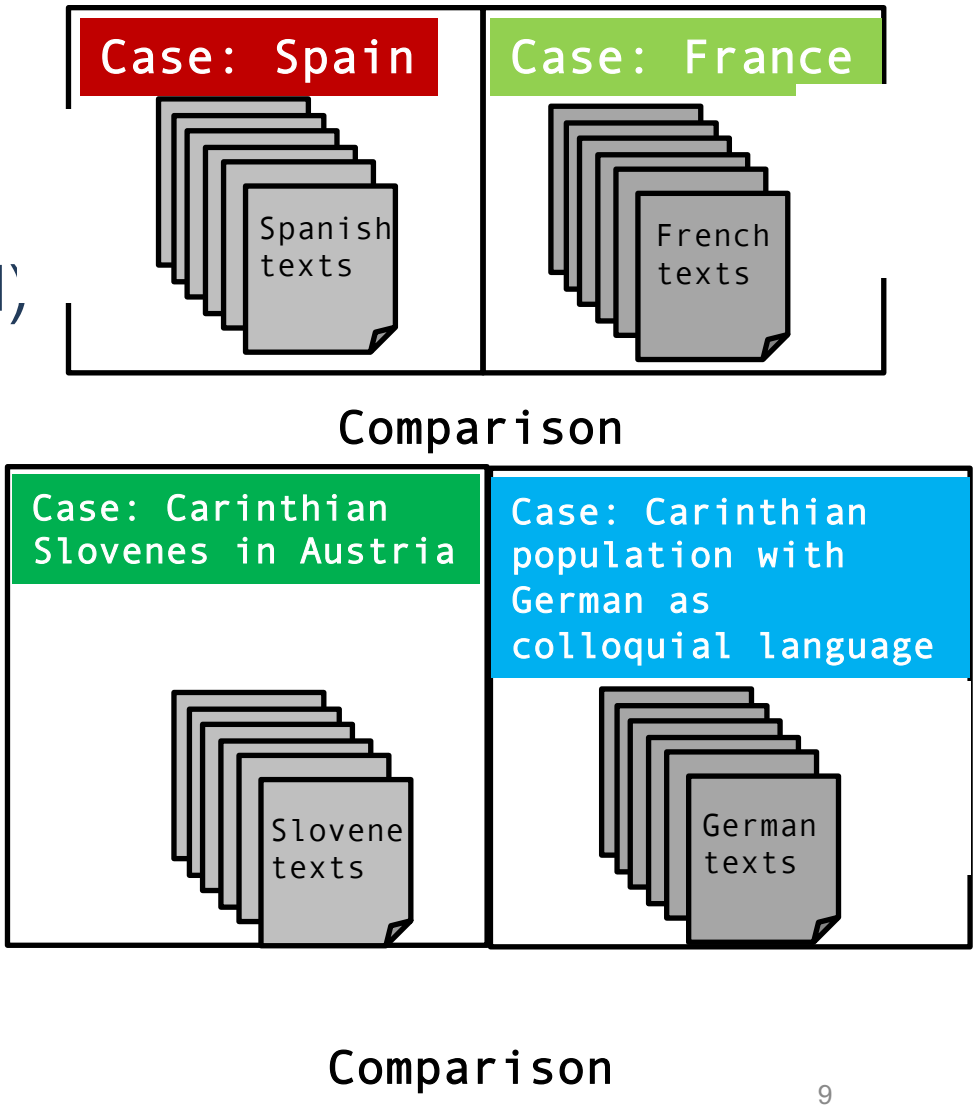
- human communication of at least two compared cases manifests in texts in different languages

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)

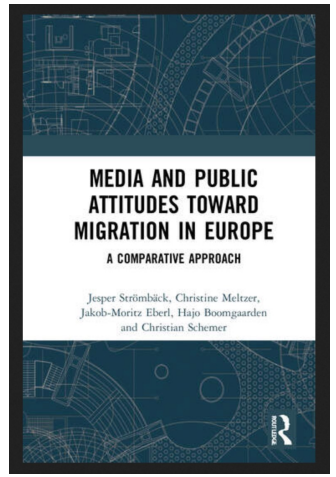
- Classification
- Topic Modeling
- Scaling
- etc.

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

Central tips to get started

1. **A must:** Experts for each case and language (for concept definition, instrument creation, ...)

2. Linguistic knowledge about the languages of your documents

3. Resources for validation

A main challenge: Equivalence

- 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.

- 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)

The effort to approach equivalence on four levels

Equivalence of

1. **document samples**
2. **constructs**
3. **measurements**
4. **procedures**

1. Sample equivalence

- ‘Sample equivalence’ relates to the challenge of finding units of analysis that are comparable across cases (Rössler, 2012, p. 461).

Finding equivalent document sources

- Text corpora by political organizations and Legislative text corpora: https://opted.eu/results/inventories/
- European media sources: https://wp3.opted.eu

Retrieval of equivalent documents

- 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

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

2. Construct equivalence

- ‘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 is useful to be studied across cases?
- Decision to go for rather ‘emic’ or ‘etic’ definition

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

Illustration: Construct definition

REMINDER

Process: Support by case 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)

3. Measurement equivalence

- ‘Measurement equivalence’ concerns the operationalization and design of appropriate research instruments.
  - Appropriateness is evaluated in respect to the decision to strive for rather ‘emic’ or rather ‘etic’ measurement

Measurement goals may include

- Classification
- Topic modeling
- Scaling

Classification methods

- **Dictionary methods:** (Multilingual) keywords to be searched in a (multilingual) corpus
- **Supervised Machine Learning:** (Multilingual) annotated documents for training, test and validation purposes

Why is multilingual in brackets?

- While it is possible to create instruments per case/language (e.g., Baden & Stalpouskaya, 2015), an alternative technique is to design instruments only for one language and (machine) translate the documents into this language (e.g., Benoit et al., 2012).

Illustration: Supervised text classification for multilingual corpora

- Document translation into English
- Document annotation with an English dictionary
- The dictionary is designed to incorporate case-specific and transnational aspects
- It is validated with a manually annotated subset (includes documents for all cases)

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

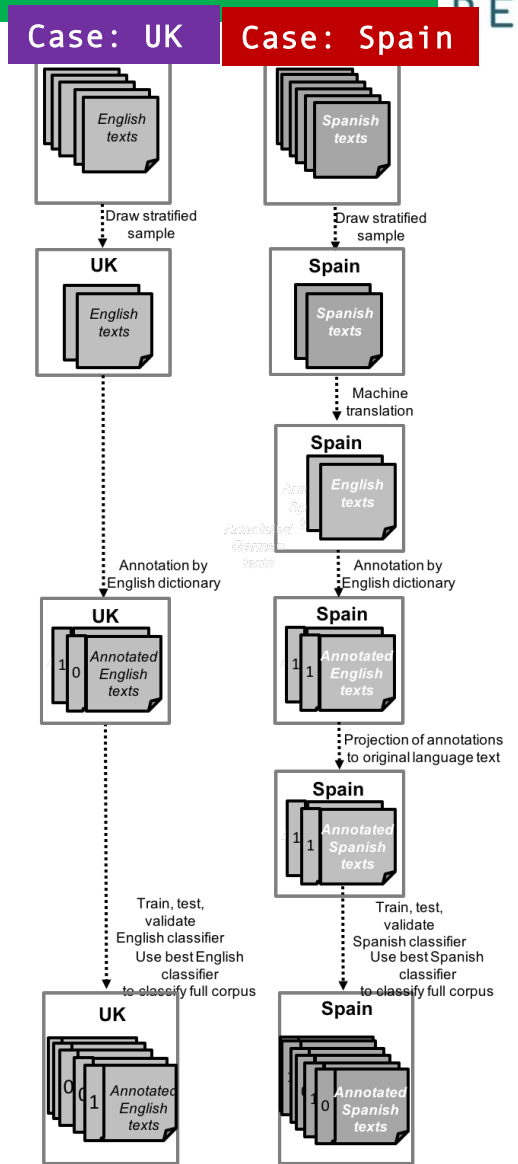


Illustration: Supervised text classification for multilingual corpora

REMINDER

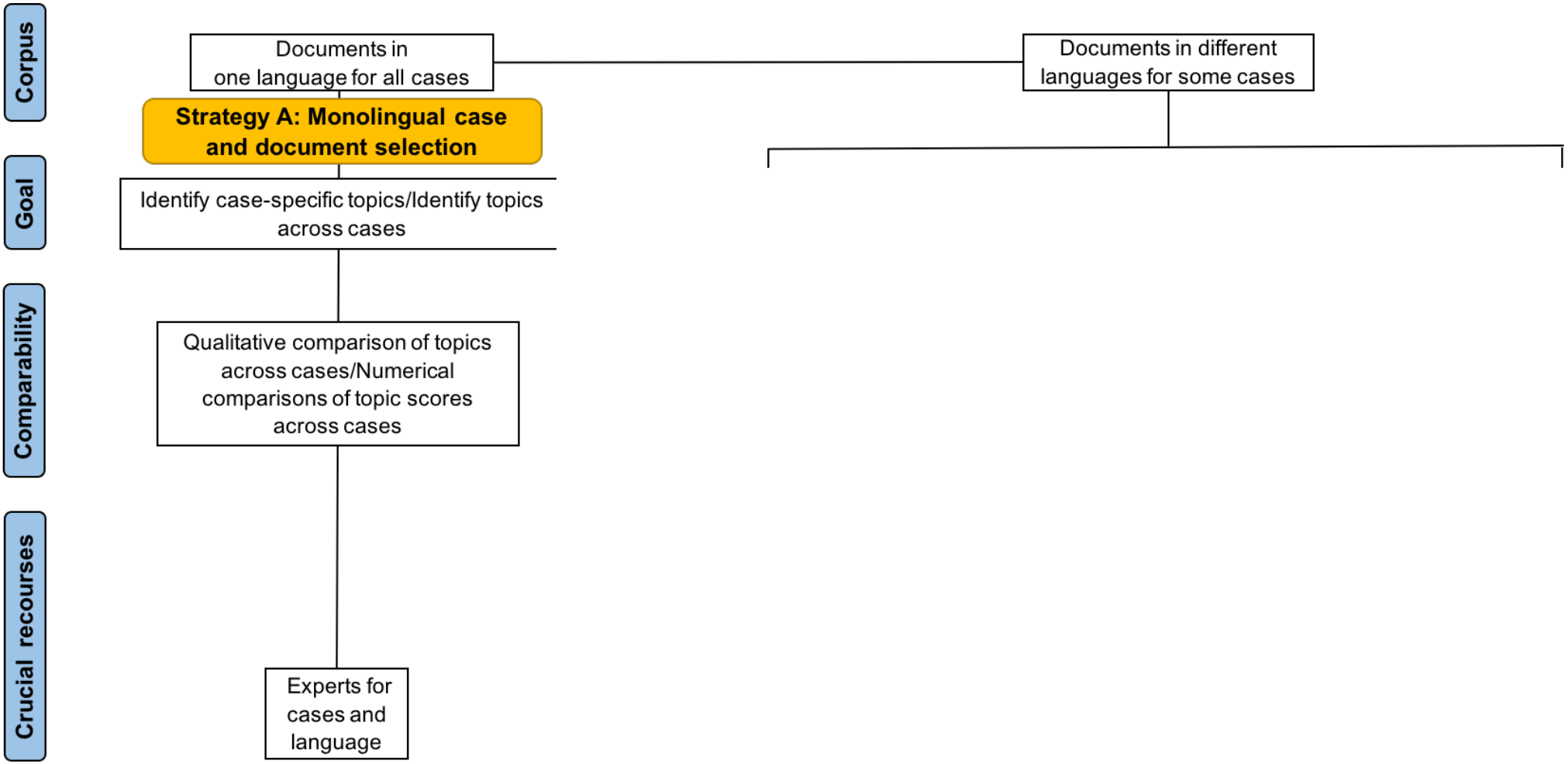
Frame measurement: migration as 1. economy, 2. labor, 3. welfare, 4. security topic

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)

Topic Modeling for Comparative Research

universität wien



Topic Modeling for Comparative Research

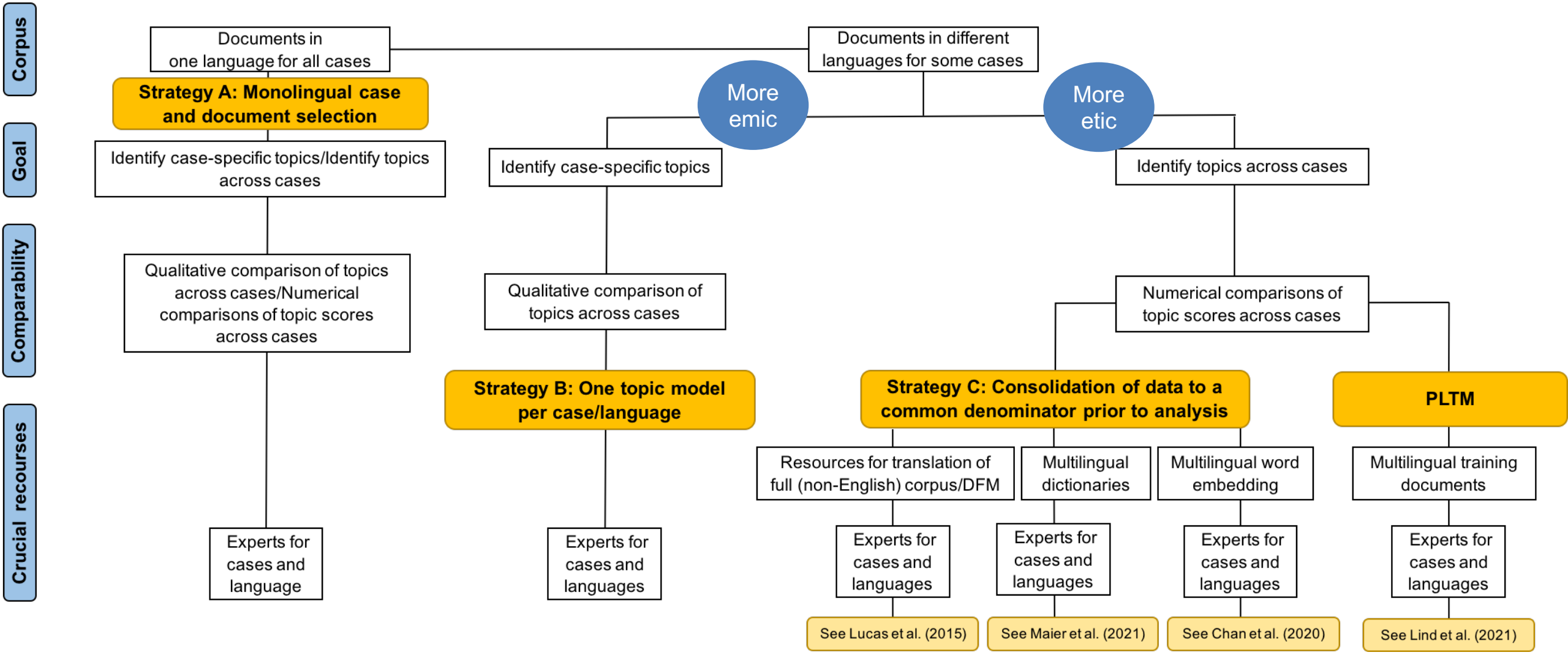


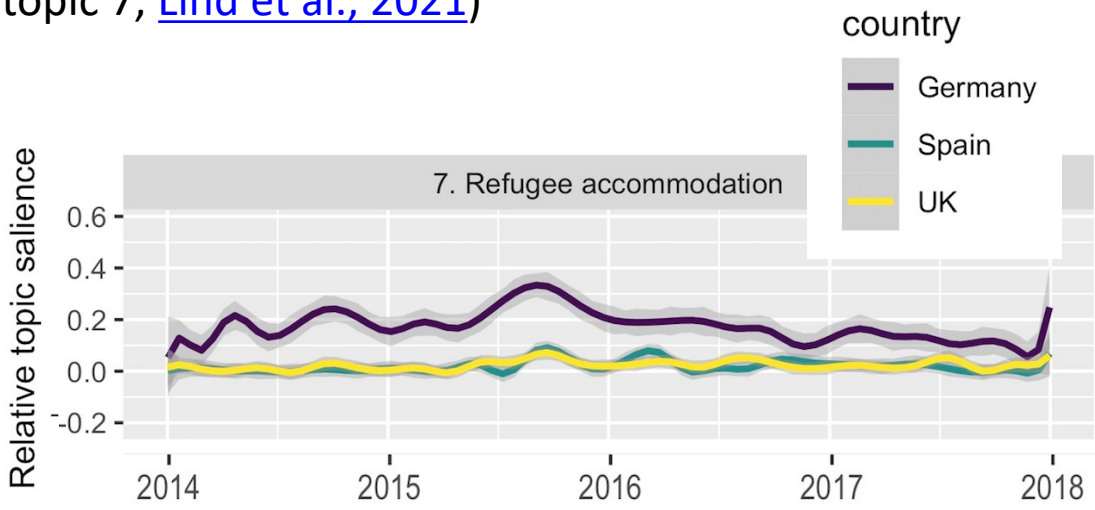
Illustration: PLTM Topic Modeling

REMINDER

- PLTM tutorial based on Mimno et al., 2009: https://github.com/fabiennelind/Topic-Modeling-for-Comparative-Research

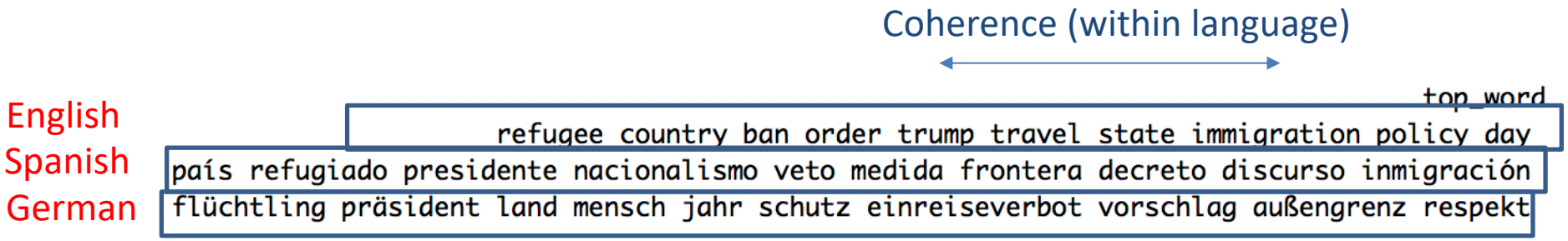
Example for a multilingual topic (here topic 7, Lind et al., 2021)

Topic Label	Language	Top 10 words
7. Refugee accommodation	English	refugee asylum seeker people accommodation country district situation office reception
	German	refugiado asilo solicitante persona derecho alojamiento ayuda distrito oficina país
	Spanish	flüchtling asylbewerber unterkunft land nutzung hilfe zahl grenze syrer monat

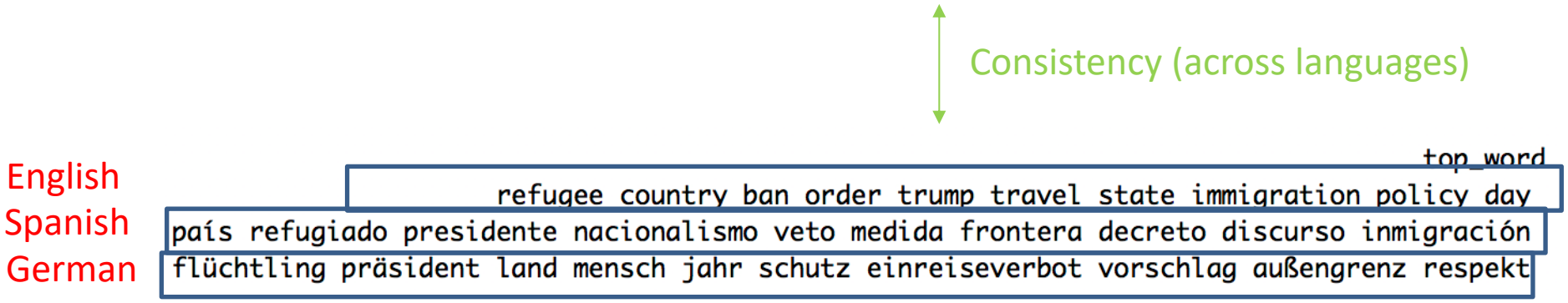


Test convergent validity: comparison of topic probabilities per document with dictionary coding

Evaluation: 1 example topic



Evaluation: 1 example topic



4. Procedure equivalence

- ‘Procedure equivalence’ deals with the standardized application of the content analysis instruments across cases

- Reliability of measurements:
  - Automated-coding has an advantage in terms of consistent application of a research instrument (Rössler, 2012, p. 462)
  - As automated methods are oftentimes validated with human-coded data, the establishment of reliable manual coding is relevant (Song et al., 2020).

- Recommended: Knowledge about the (language-specific) quality of tools used (e.g., lemmatizer, machine translation software, multilingual word embeddings)

Pre-processing tools for multilingual text analysis in R

universität wien

- Lemmatization with **udpipe** (Wijffels, 2021)
  - more than 65 languages, free of charge
  - Tutorial: https://cran.r-project.org/web/packages/udpipe/vignettes/udpipe-annotation.html

- Machine Translation with **deeplr** (Zumbach & Bauer, 2021)
  - 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 confidential text data, checking the data handling procedures of the external services is required

Illustration: Machine translated version of the data set

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

Validation Considerations for a comparative design

Ideally: post-hoc assessments regarding the equivalence of

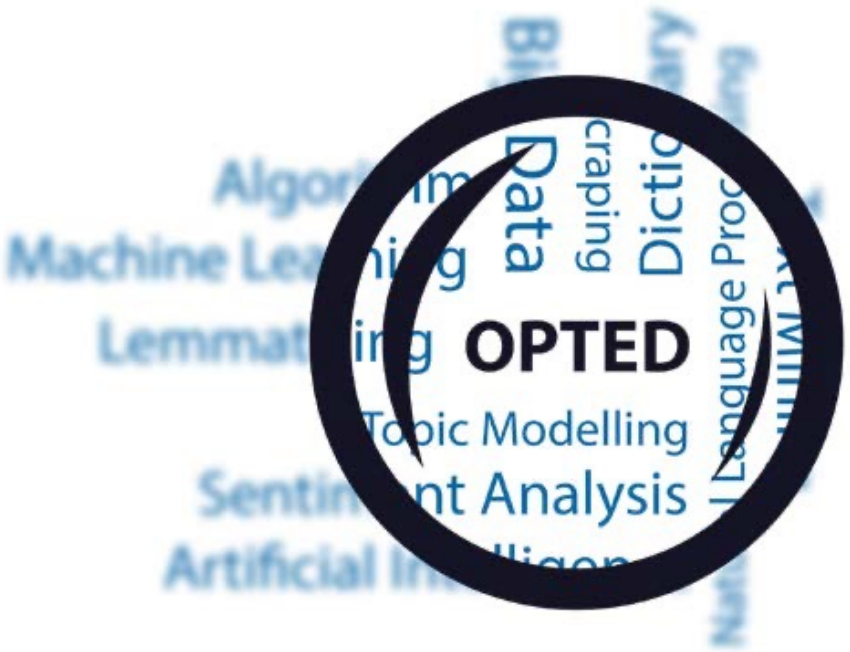
- Document samples
- Construct definitions
- Obtained automated measures
- The implemented procedures

Methods:

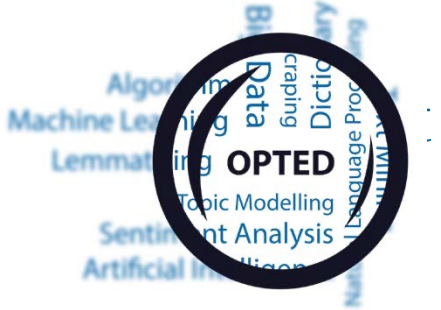
- Design tests with the involvement of case/language experts
- Measurement validation: Human-coded reference points or topic model interpretation and validation method needed for all languages for which the automated method produces a result; approach to comparability has to be taken into account

More resources: www.opted.eu

OPTED = Observatory for Political Texts in European Democracies



Planned:



OPTED Living hub and knowledge base for multilingual computational text analysis (see WP6)

Overview about Multilingual Named Entity Recognition Tagging (Balluff et al., 2021)

Community building @OPTED_H2020

Open round

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

How can the methodology benefit from collaborations across labs and disciplines?

Any resources that you like to share with the group?

What challenges do you see for 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.
- 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
- 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.
- 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
- 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