universität wien

Today

9:45-10:40	Input	
10:40-10:50	Coffee Break	
10:50-11:45	Coding	
11:45-12:45	Lunch	
12:45-13:40	Input: Multilingual text analysis	
13:40-13:50	Coffee Break	
13:50-14:45	Coding	

universität wien

Unsupervised classification

Day 4 Session 1

universität wien

Types of machine learning

1. Supervised
    - An outcome variable is defined
    - Focus is on prediction
2. Unsupervised
    - No outcome variable has been defined
    - Focus is on patterns

universität wien

How to use the ***supervised*** methods?

- Easy
- At least *conceptually*
- *Clear* ***objective function***

universität wien

How to use the *supervised* methods?

$$Y = (y_1, y_2, ..., y_n)$$
$$X = (x_1, x_2, ..., x_n)$$
$$(y_1, x_1), (y_2, x_2), ..., (y_n, x_n)$$

Task to predict $\hat{y}$ as close to y

universität wien

How to use the *supervised* methods?

$$L(y, \hat{y}) = (y - \hat{y})^2$$

$$\hat{y} = \underset{\theta}{\operatorname{argmin}} \; E\left[L(model(\mathbf{x}, \theta), y)\right]$$

universität wien

How to use unsupervised learning

- Objective function?

universität wien

How to use unsupervised learning

- Objective function?
- Quantity of interest?

universität wien

How to use unsupervised learning

- Objective function = ***your*** quantity of interest

universität wien

universität wien

How to use unsupervised learning

- Objective function = ***your*** quantity of interest
- This is difficult

universität wien

Focus is on Discovery

Grimmer et al. 2022

universität wien

Objectives

Descriptive analysis/Discriminating words:

- What are the characteristics of a corpus? How do some documents compare to each other
- Keyness, collocation analysis, readability scores, Cosine/Jaccard similarity

Clustering and scaling:

- What groups of documents are in the corpus? Can the documents be placed on a dimension?
- Cluster analysis, principal component analysis, wordfish..

Topic modeling:

- What are the main themes in a corpus?
- LDA, STM

Grimmer et al. 2022, Barberá, 2018

universität wien

K-Means Clustering

- Simple(ish) algorithmic method
- Partitions the data into K non-overlapping clusters

Setup

$$C_1, C_2, ..., C_k$$

$$C_1 \cup C_2 \cup \ldots \cup C_K = \{1, \ldots, n\}$$

$$C_k \cap C_{k'} = \emptyset \text{ for all } k \neq k'$$

universität wien

Assumption and task

- Optimal clustering solution is the one where ***within-cluster variation*** is ***as small as possible***

$$W(C_k)$$

$$\underset{C_1,\ldots,C_K}{\text{minimize}} \left\{ \sum_{k=1}^{K} W(C_k) \right\}$$

Within cluster variation

$$W(C_k) = \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2$$

$$\underset{C_1,\ldots,C_K}{\text{minimize}} \left\{ \sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 \right\}$$

universität wien

Algorithm

1. Randomly assign cluster numbers (1 through K) for each observation
2. Iterate until no further changes to the cluster assignment:
    a. For each cluster determine the ***centroid*** (average of all observations in the cluster)
    b. Re-assign observations to a cluster with the closest centroid (calculated with a distance metric).

universität wien

Guarantees convergence at a *local optimum*

- Cannot guarantee the best solution
- But are rather good one
- Sensitive to random assignment at the start

universität wien

Cluster Algorithms Validation

- Data assumptions (think data generation)
- Internal validity (best results for the data)
- External validity (matches with pre-existing understanding of data)
- Cross-validity (similar results across similar datasets)
- ***You are the validation method***

universität wien

Topic Modelling

- A model to discover latent topics
- ***Not*** synonymous with LDA
- LDA is one of topic models

- Latent semantic analysis
- Singular value decomposition
- Even clustering methods (like the one we just discussed)

universität wien

Latent Dirichlet Allocation

- Bayesian generative hierarchical model
- First introduced as a way to simultaneously model traits and genes (Pritchard, 2000)
- Adjusted for text analysis ML applications (Blei et al., 2003)

universität wien

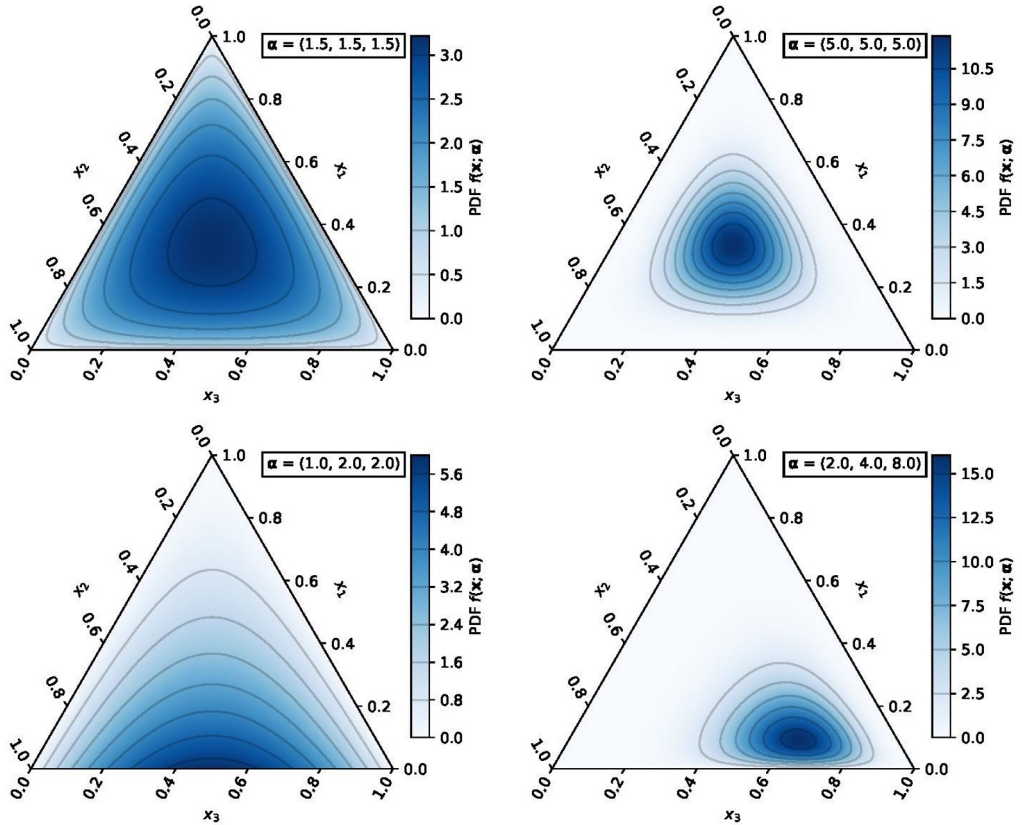
Latent Dirichlet Allocation

- Estimates a distribution of ***words*** across ***documents*** across ***latent topics***

universität wien

Hierarchical Models

$$W_i \; in \; D_d \sim Multinomial_K(1, \boldsymbol{\theta}_d)$$
$$T_k \; in \; D_d \sim Multinomial_V(1, \boldsymbol{\phi}_k)$$
$$\boldsymbol{\phi_k} \sim Dirichlet_V(\boldsymbol{\beta})$$
$$\boldsymbol{\theta_d} \sim Dirichlet_K(\boldsymbol{\alpha})$$


α = (1.5, 1.5, 1.5)
α = (5.0, 5.0, 5.0)
α = (1.0, 2.0, 2.0)
α = (2.0, 4.0, 8.0)
PDF f(x; α)
x1
x2
x3

universität wien

Tricky to estimate

$$\int_{\theta_j} P(\theta_j;\alpha)\prod_{t=1}^{N} P(Z_{j,t}\mid\theta_j)\,d\theta_j = \int_{\theta_j} \frac{\Gamma\left(\sum_{i=1}^{K}\alpha_i\right)}{\prod_{i=1}^{K}\Gamma(\alpha_i)}\prod_{i=1}^{K}\theta_{j,i}^{n^i_{j,(\cdot)}+\alpha_i-1}\,d\theta_j$$

$$= \frac{\Gamma\left(\sum_{i=1}^{K}\alpha_i\right)}{\prod_{i=1}^{K}\Gamma(\alpha_i)}\frac{\prod_{i=1}^{K}\Gamma(n^i_{j,(\cdot)}+\alpha_i)}{\Gamma\left(\sum_{i=1}^{K}n^i_{j,(\cdot)}+\alpha_i\right)}\int_{\theta_j}\frac{\Gamma\left(\sum_{i=1}^{K}n^i_{j,(\cdot)}+\alpha_i\right)}{\prod_{i=1}^{K}\Gamma(n^i_{j,(\cdot)}+\alpha_i)}\prod_{i=1}^{K}\theta_{j,i}^{n^i_{j,(\cdot)}+\alpha_i-1}\,d\theta_j$$

$$= \frac{\Gamma\left(\sum_{i=1}^{K}\alpha_i\right)}{\prod_{i=1}^{K}\Gamma(\alpha_i)}\frac{\prod_{i=1}^{K}\Gamma(n^i_{j,(\cdot)}+\alpha_i)}{\Gamma\left(\sum_{i=1}^{K}n^i_{j,(\cdot)}+\alpha_i\right)}.$$

universität wien

$$\int_{\boldsymbol{\varphi}} \prod_{i=1}^{K} P(\varphi_i;\beta) \prod_{j=1}^{M} \prod_{t=1}^{N} P(W_{j,t} \mid \varphi_{Z_{j,t}})\, d\boldsymbol{\varphi}$$

$$= \prod_{i=1}^{K} \int_{\varphi_i} P(\varphi_i;\beta) \prod_{j=1}^{M} \prod_{t=1}^{N} P(W_{j,t} \mid \varphi_{Z_{j,t}})\, d\varphi_i$$

$$= \prod_{i=1}^{K} \int_{\varphi_i} \frac{\Gamma\left(\sum_{r=1}^{V} \beta_r\right)}{\prod_{r=1}^{V} \Gamma(\beta_r)} \prod_{r=1}^{V} \varphi_{i,r}^{\beta_r - 1} \prod_{r=1}^{V} \varphi_{i,r}^{n^i_{(\cdot),r}}\, d\varphi_i$$

$$= \prod_{i=1}^{K} \int_{\varphi_i} \frac{\Gamma\left(\sum_{r=1}^{V} \beta_r\right)}{\prod_{r=1}^{V} \Gamma(\beta_r)} \prod_{r=1}^{V} \varphi_{i,r}^{n^i_{(\cdot),r} + \beta_r - 1}\, d\varphi_i$$

$$= \prod_{i=1}^{K} \frac{\Gamma\left(\sum_{r=1}^{V} \beta_r\right)}{\prod_{r=1}^{V} \Gamma(\beta_r)} \frac{\prod_{r=1}^{V} \Gamma(n^i_{(\cdot),r} + \beta_r)}{\Gamma\left(\sum_{r=1}^{V} n^i_{(\cdot),r} + \beta_r\right)}.$$

universität wien

Need Markov Chain Monte Carlo Simulation

universität wien

Structural Topic Models

Add additional distribution for additional structure in the text:

Allows to add additional covariates to the estimation

universität wien

Validating topic models

Tests of:

- topic semantic validity: assess the extent to which the keywords within each topic have a coherent underlying meaning, and how these meanings behave across topics (e.g., Quinn et al., 2010, p. 210)
- convergent validity: topic probabilities per document are compared with an external trusted variable like manual coding for the same documents (e.g., Guo et al., 2016)

A possible validation workflow proposed by Maier et al. (2018): combi of quantitative topic summary metrics (e.g., NPMI; Lau et al., 2014) and human expert evaluations.

universität wien

Text scaling methods

- Attempts to fit documents into a unidimensional space
- Documents are “scaled” based on the frequency of used terms
- Assume “discriminating” words have a Poisson distribution

- “Ideological” successor to log odds ratio we’ve seen

universität wien

Wordfish (Slapin and Proksch 2008)

$$w_{ik} \sim \text{Poisson}(\lambda_{ik})$$

$$\lambda_{ik} = exp(\alpha_i + \psi_k + \beta_k \times \theta_i)$$

α_i Text size from type i

ψ_k Frequency of word k

β_k Discrimination power of word k

θ_i Ideological position of type i

universität wien

Wordfish (Slapin and Proksch 2008)

$$w_{ik} \sim \text{Poisson}(\lambda_{ik})$$

$$\lambda_{ik} = exp(\alpha_i + \psi_k + \beta_k \times \theta_i)$$

α_i Text size from type i

ψ_k Frequency of word k

β_k Discrimination power of word k

θ_i Ideological position of type i

universität wien

Wordfish (Slapin and Proksch 2008)

$$w_{ik} \sim \text{Poisson}(\lambda_{ik})$$

$$\lambda_{ik} = exp(\alpha_i + \psi_k + \beta_k \times \theta_i)$$

α_i Text size from type i

ψ_k Frequency of word k

β_k Discrimination power of word k

θ_i Ideological position of type i

universität wien

R time

universität wien

Multilingual text analysis

Day 4 Session 2

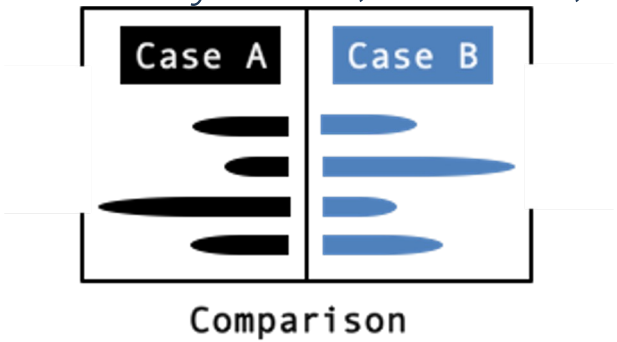
universität wien

Session Contents

- Motivation for multilingual text analysis
- Challenges
- Strategies
- Language and context sensitive validation framework
- Practical tips

Comparative research

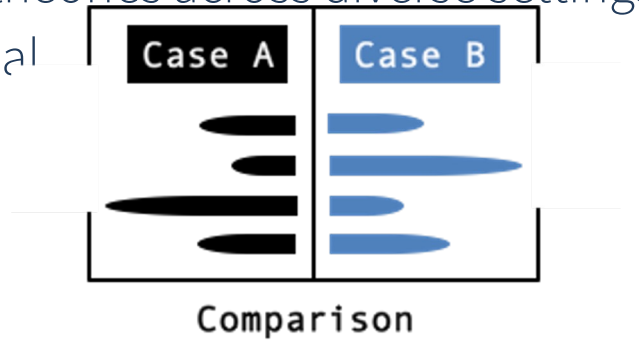
- Comparative research in social science involves comparisons between a minimum of two cases with at least one object of investigation relevant to social science research.
- Cases are macro-level units such as systems, cultures, countries, and markets)



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

Reasons to compare

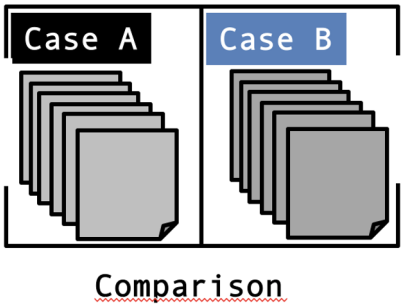
- insights into the differences and similarities of cases
- improved understanding and contextualization of the own case
- raised awareness for other cases
- the test and generalizability of theories across diverse settings
- the investigation of transnational processes across contexts



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

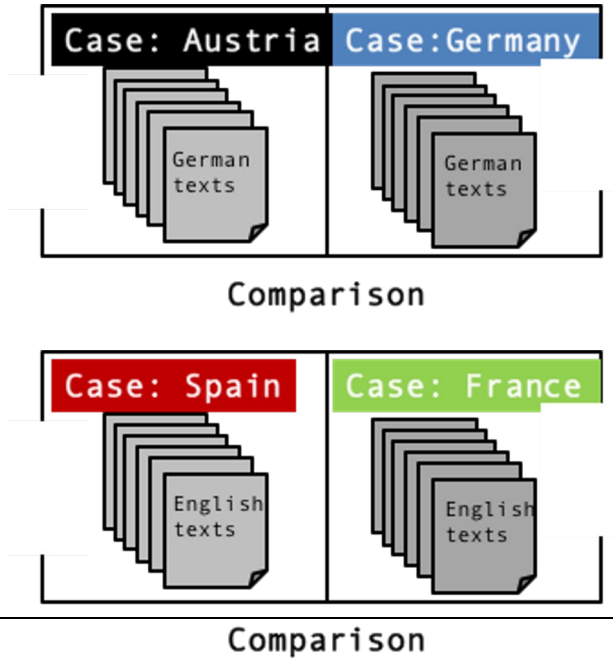


Boomgaarden & Song, 2019; McLeod & Blumler, 1987; Esser & Vliegenthart, 2017; Esser & Hanitzsch, 2012; Livingstone, 2003

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)



universität wien

Comparison of cases & language(s) of documents

But the likely scenario is multilingual

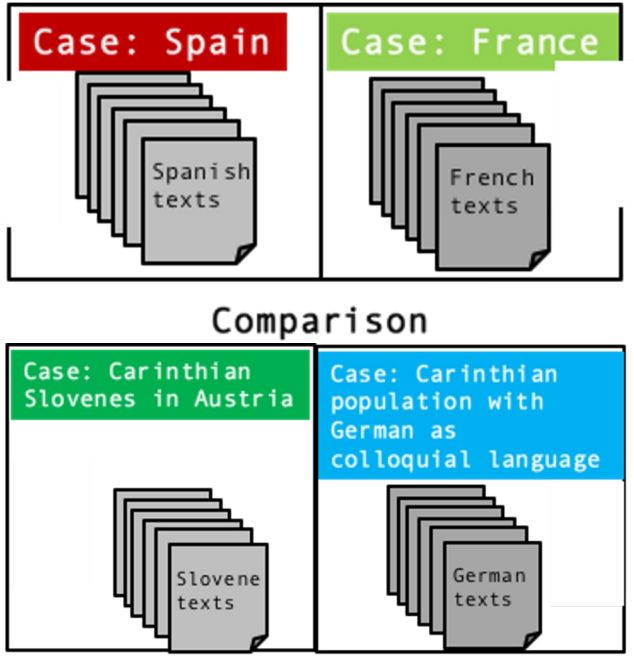
- human communication of at least two compared cases manifests in texts in different languages

universität wien

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)



Comparison

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

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.

universität wien

Handling multilingual corpora: Key challenge and three 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

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)

- Four example sentences as illustration for a multilingual corpus

	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

universität wien

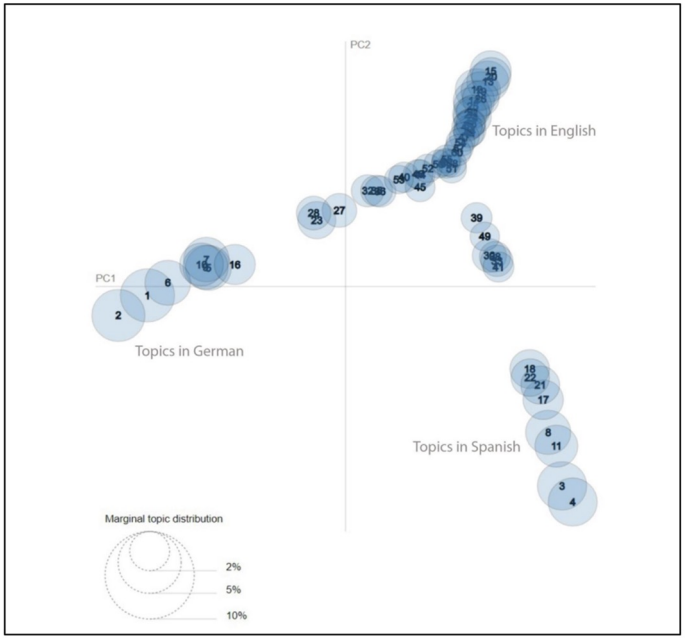
Illustration 1 (Part 2)

- Four example sentences as illustration for a multilingual corpus

	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

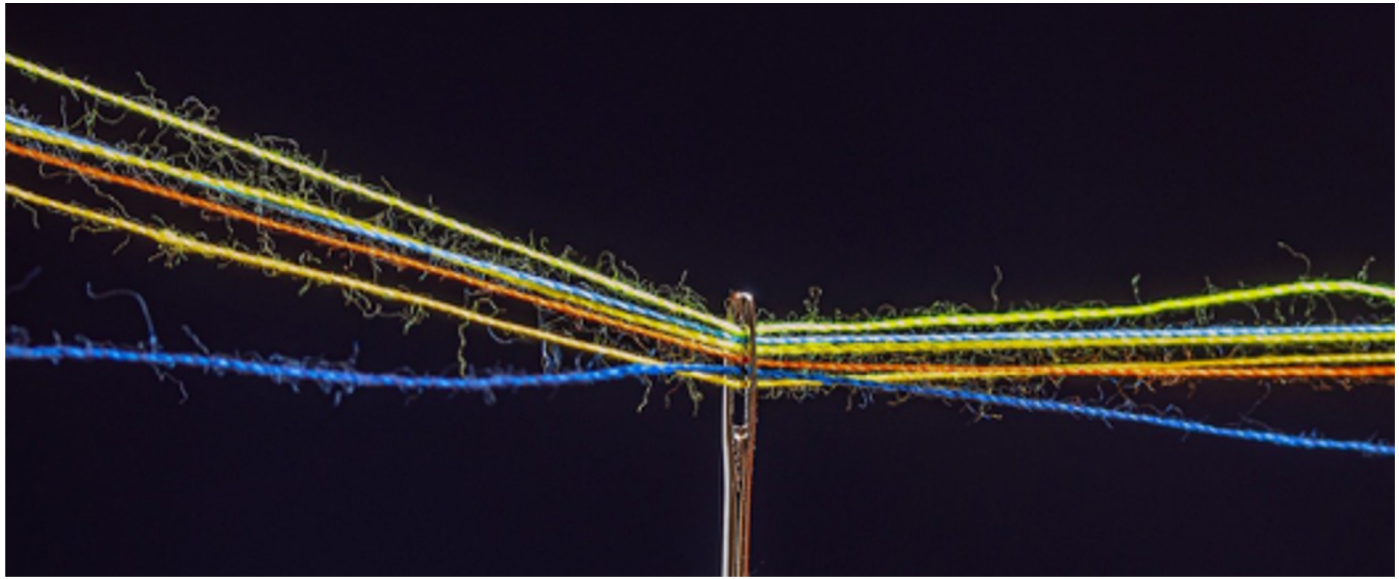
- Striving for measurement equivalence across languages

  = 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

Three approaches

1. Separate analysis
2. Input alignment
3. Anchoring

1. Separate analysis

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

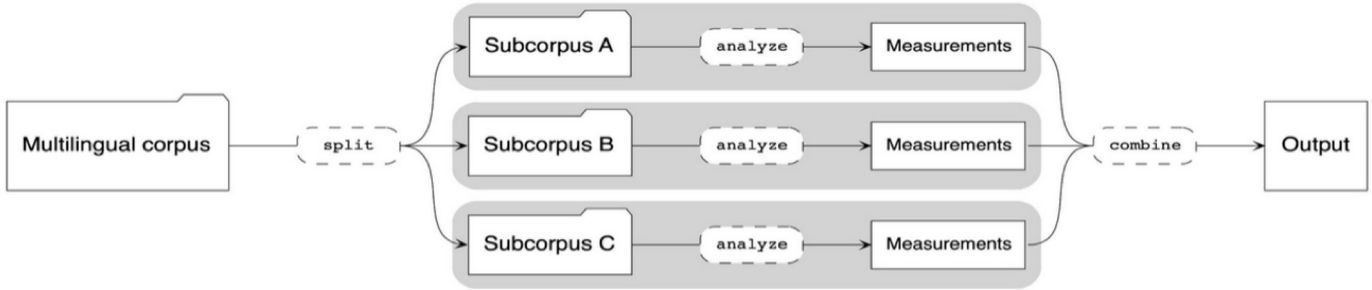


Figure 1 Illustration of the separate analysis strategy to multilingual text analysis.

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

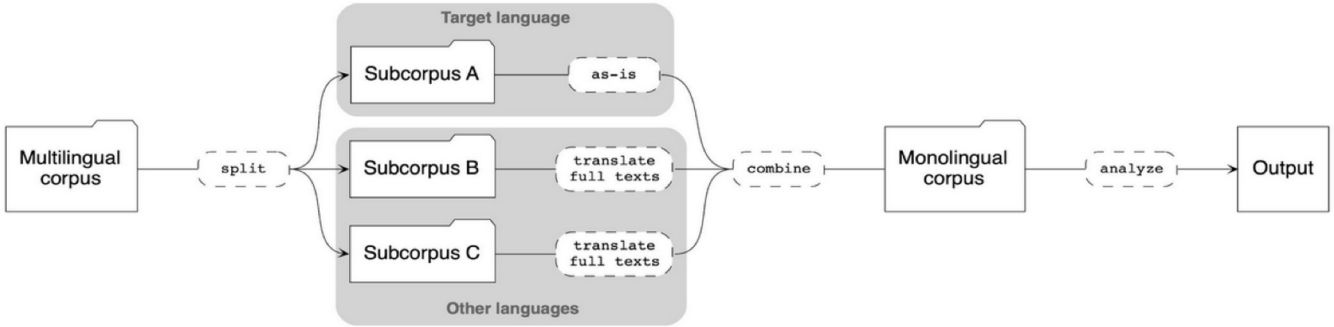


Figure 2 Illustration of the full-text translation approach to input alignment

Licht & Lind, forthcoming

2. Input alignment

Option 1: (Machine) Translation

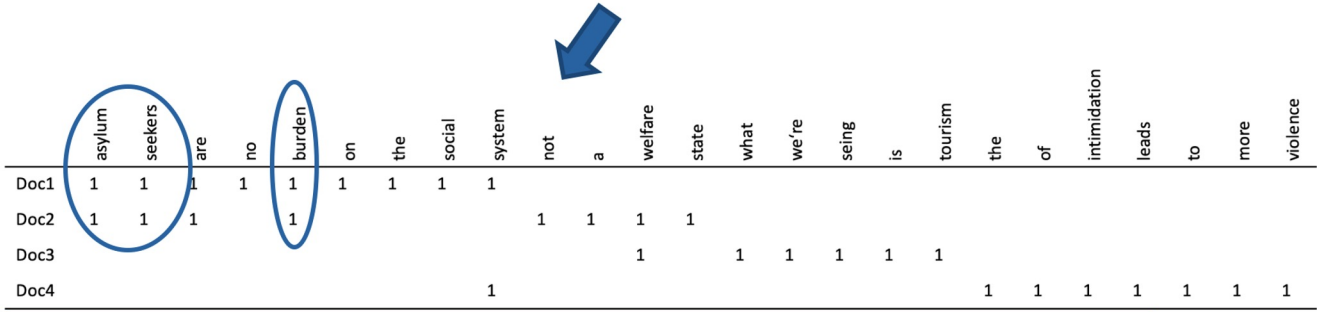
	text	text (english version)	target label
Doc1	Asylum seekers are no burden on the social system	Asylum seekers are no burden on the social system	welfare
Doc2	Asylsuchende belasten das Gemeinwesen nicht	Asylum seekers are not a burden on the welfare state	welfare
Doc3	What we're seeing is welfare tourism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

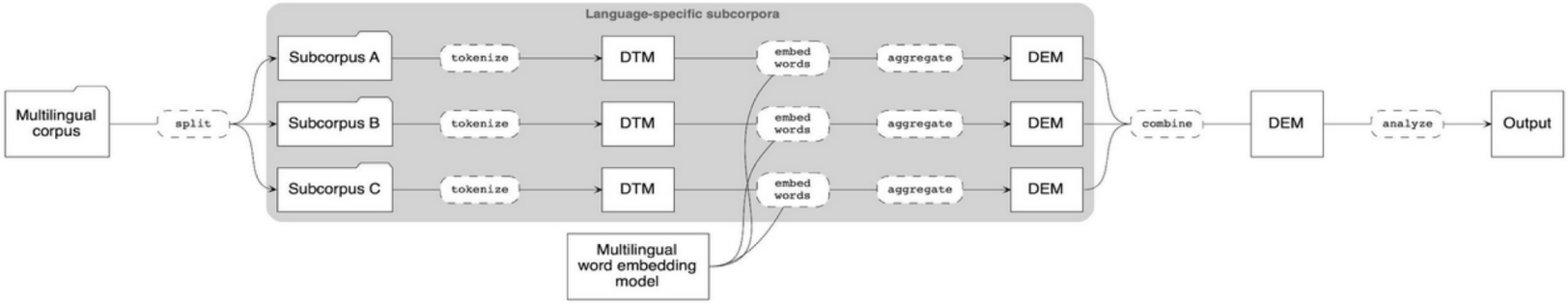


Figure 4 Illustration of the multilingual word embedding approach to input alignment.

Licht & Lind, forthcoming

universität wien

3. Anchoring

Idea: cross-lingual comparability and alignment of some model inputs (= bridging observations) is used to explicitly incentivize alignment in model outputs

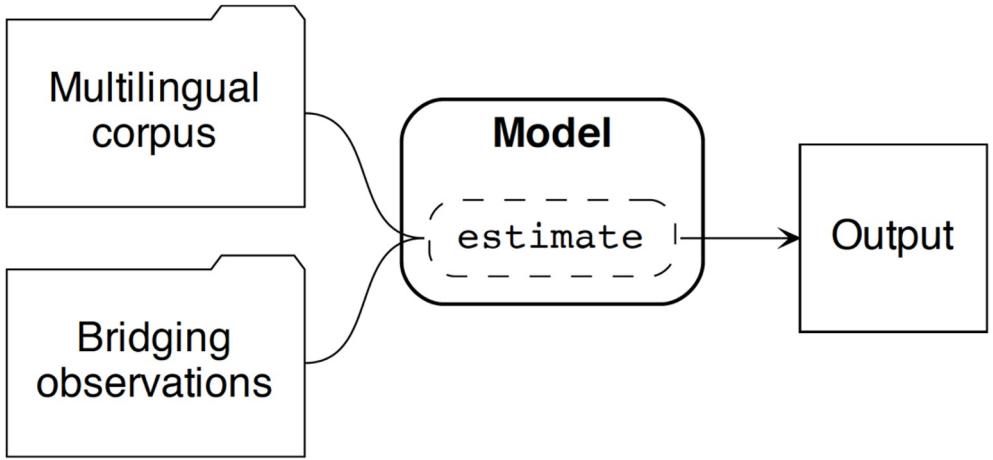
Ideas for bridging information

- Bilingual lexica
- Parallel or comparable documents

Licht & Lind, forthcoming

3. Anchoring

- so far, developed for topic modeling (Mimno et al., 2011); implementations (Lind et al. 2022; Pruss et al., 2019)



Licht & Lind, forthcoming

universität wien

How to decide between the three approaches

Let‘s collect pros and cons for each approach

1. Separate analysis
2. Input alignment
3. Anchoring

universität wien

Objective

- 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

Objective

- 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

Communication in social contexts

- In comparative research, next to semantic (literal meaning) there is also the need to concentrate 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 scientist interested in comparative research, we also care for social political cultural economic contexts of the cases that we compare (Gurevitch and Blumler, 2003)

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

A language and context sensitive validation framework

Motivation for designing a validation framework

- Lack of frameworks

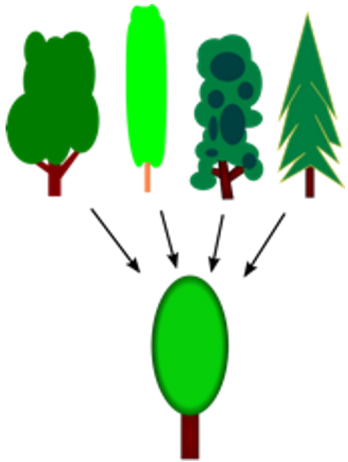
- Insights from a content analysis of published literature in the social sciences and from an expert survey with the respective authors (Baden et al. 2022)

  Main result: Validation concern higher among those working with multiple languages but concern is not reflected in higher focus on validation in published work

Scope of framework

- Measurement of universal type constructs

- Thus, assumption that information drawn from different cases can correspond on a conceptual level (Goertz, 2006; Adcock & Collier, 2001).



The proposed validation framework

- to make valid comparisons between cases, equivalence must be established and ideally demonstrated
- on two levels:
  - Semantics (valid across languages)
  - Pragmatics (valid across contexts)
- at four different steps of the data analysis pipeline:
  - Data, input, process, output

universität wien

Validation framework

	1. Data Validation	2. Input validation	3. Process validation	4. Output validation
Key question	Are the samples documents equivalent across languages and contexts?	Are the chosen textual representations equivalent across languages and contexts?	Are the algorithms equally effective in all languages and contexts?	Are the obtained measurements of similar quality and equivalent across languages and contexts
Objective	Mitigate potential sources of sampling bias	Ensure relevance for research question at hand and language and context equivalence of text representations	Ensure that the algorithms categorizes texts in each language and context in a similar way	Ensure that obtained measurements are equivalent across languages and contexts
Strategy	Expert based selection of equivalent sources; validate search strings across languages and contexts	Confirm that the language and context specific features correspond with the construct of interest	Attempt to make the models transparent	Calculate recall, precision via comparison with benchmark per language and context; Conduct language and case specific error analysis

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

universität wien

Implementing the input alignment approach

- Step 1: Set up a DeepL Account
- Step 2: Translate headlines into English via DeepL API
- Step 3: Lemmatize the English version of the headlines