universität wien

Advanced quantitative text analysis (2023W)

Petro Tolochko, Fabienne Lind

Day 5

universität wien

Contents (smaller changes possible)

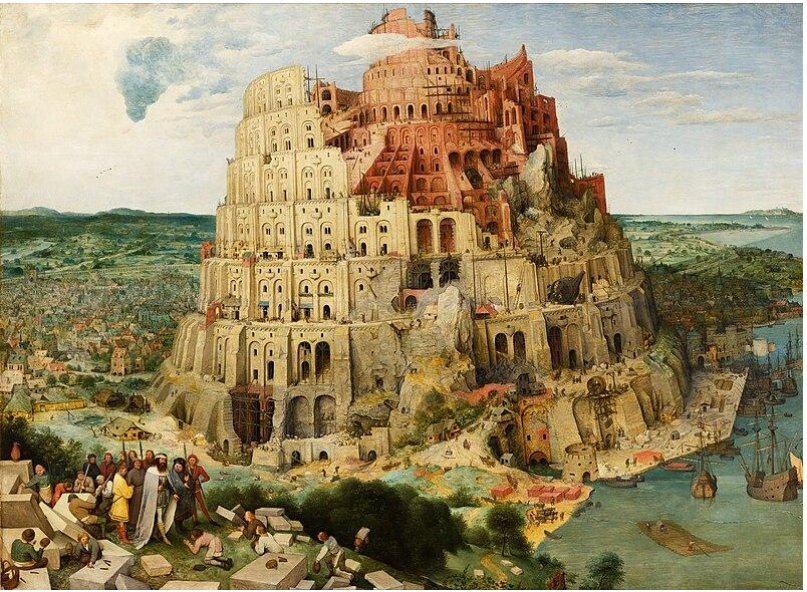
Day	Session 1	Session 2
1	Text as data, Text representation	Feature Engineering
2	Concepts & Data	Dictionaries
3	Supervised machine learning	Unsupervised machine learning
4	Neural network models, transformers	Using LLMs
5	Wrap-up	Project talks

universität wien

Last bonus materials

universität wien

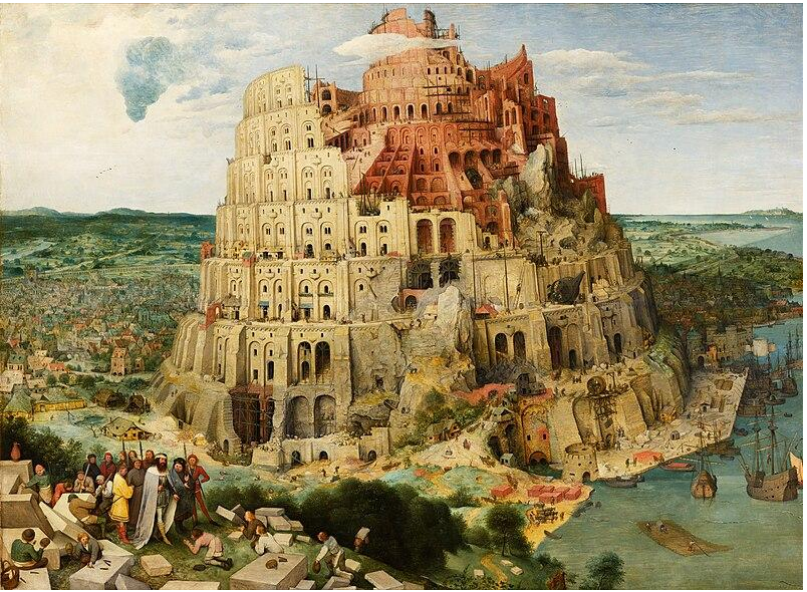
Course on Multilingual Text Analysis



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

universität wien

Course on Multilingual Text Analysis



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

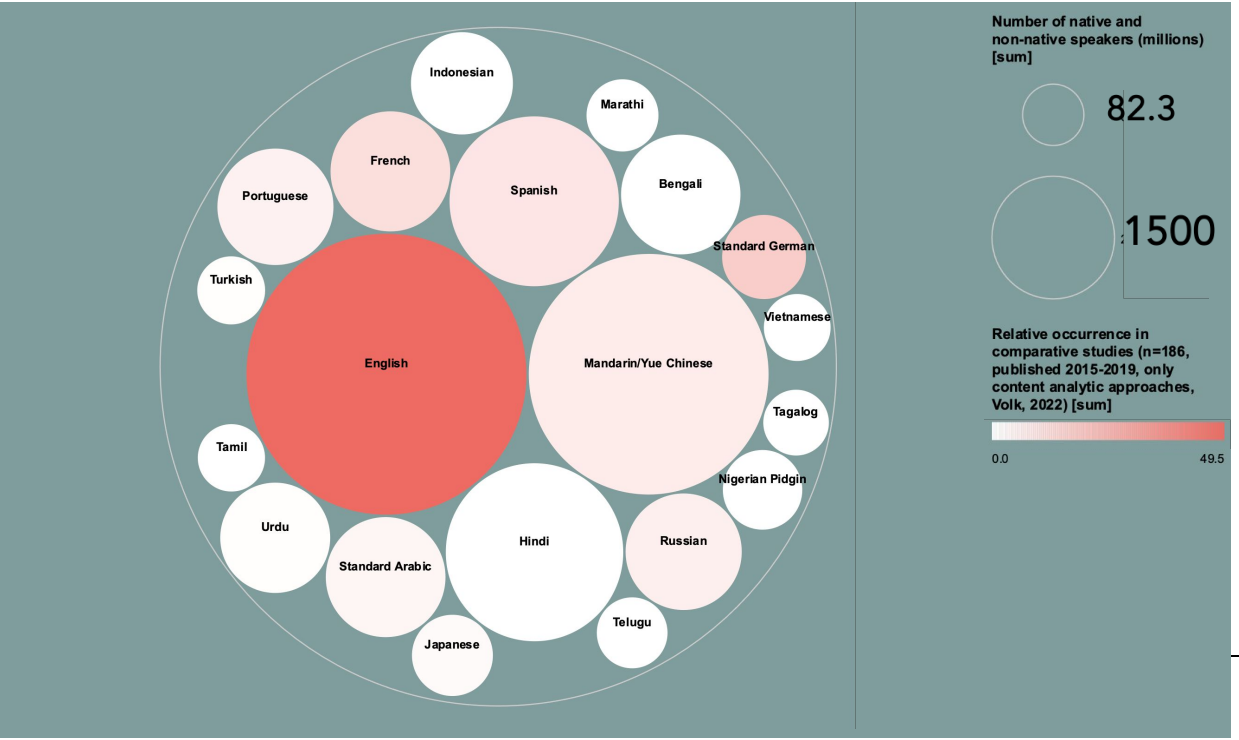
Paper: Licht, H., & Lind, F. (2023). Going cross-lingual: A guide to multilingual text analysis. *Computational Communication Research*, *5*(2), 1.

Our GitHub Course:

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

universität wien

Top 20 most spoken languages and their occurrence in comparative communication research



Lind, F. & Volk, S. (under review).

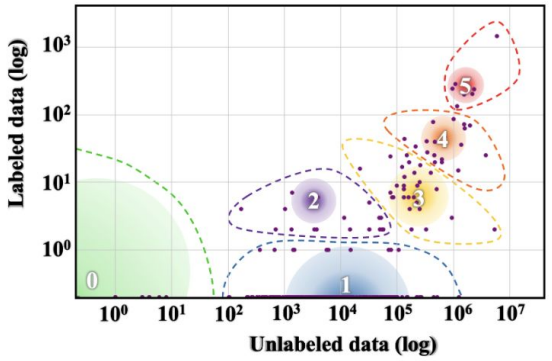
Questioning the “language agnostic” status of LLMs

Joshi et al., 2021 (see also Lauscher et al., 2020)

- LLMs rely on large amounts of labeled and unlabeled data for training
- not all languages are equally represented in training and development and the latest technologies
- availability and number of labeled and unlabeled data is a main factor for whether a language is included and to what extent
- in NLP literature, researchers differentiate between ‘low-resource’ languages and ‘high-resource’ languages

Class	5 Example Languages	#Langs	#Speakers	% of Total Langs
0	Dahalo, Warlpiri, Popoloca, Wallisian, Bora	2191	1.2B	88.38%
1	Cherokee, Fijian, Greenlandic, Bhojpuri, Navajo	222	30M	5.49%
2	Zulu, Konkani, Lao, Maltese, Irish	19	5.7M	0.36%
3	Indonesian, Ukranian, Cebuano, Afrikaans, Hebrew	28	1.8B	4.42%
4	Russian, Hungarian, Vietnamese, Dutch, Korean	18	2.2B	1.07%
5	English, Spanish, German, Japanese, French	7	2.5B	0.28%

Table 1: Number of languages, number of speakers, and percentage of total languages for each language class.



p.6284f

Risks of LLMs for certain countries

- Bender et a., 2021
  - the environmental impact of training LLMs affects certain countries more than others
  - overrepresentation of hegemonic viewpoints encoded in LLMs and the resulting lack of diversity

https://www.buzzsprout.com/2126417

universität wien

Latent Semantic Scaling

- Uses Word Embeddings
- Unidimensional Position of Texts
- You decide what this dimension is using “seed words”

- LSX package in R (Watanabe, 2020)

universität wien

Wrap up

Day 5

universität wien

Central decision criteria

- Approach to measurement (is there a ‘truth’)?
- Concepts/Interests
- Availability of methods (dictionaries, labeled data, pretrained models)
- Time (and Importance for a specific RQ)
- Budget (Modeling, Validation)
- Skills (R, Python, Patience)
- Model implementation
- Replicability
- Interpretability
- Ethical and environmental considerations
- …

universität wien

Case study 1

You want to analyze whether and how the newspaper coverage of political parties has changed in the election campaign.

What potentials and what problems do you see approaching this task with Transformer models?

universität wien

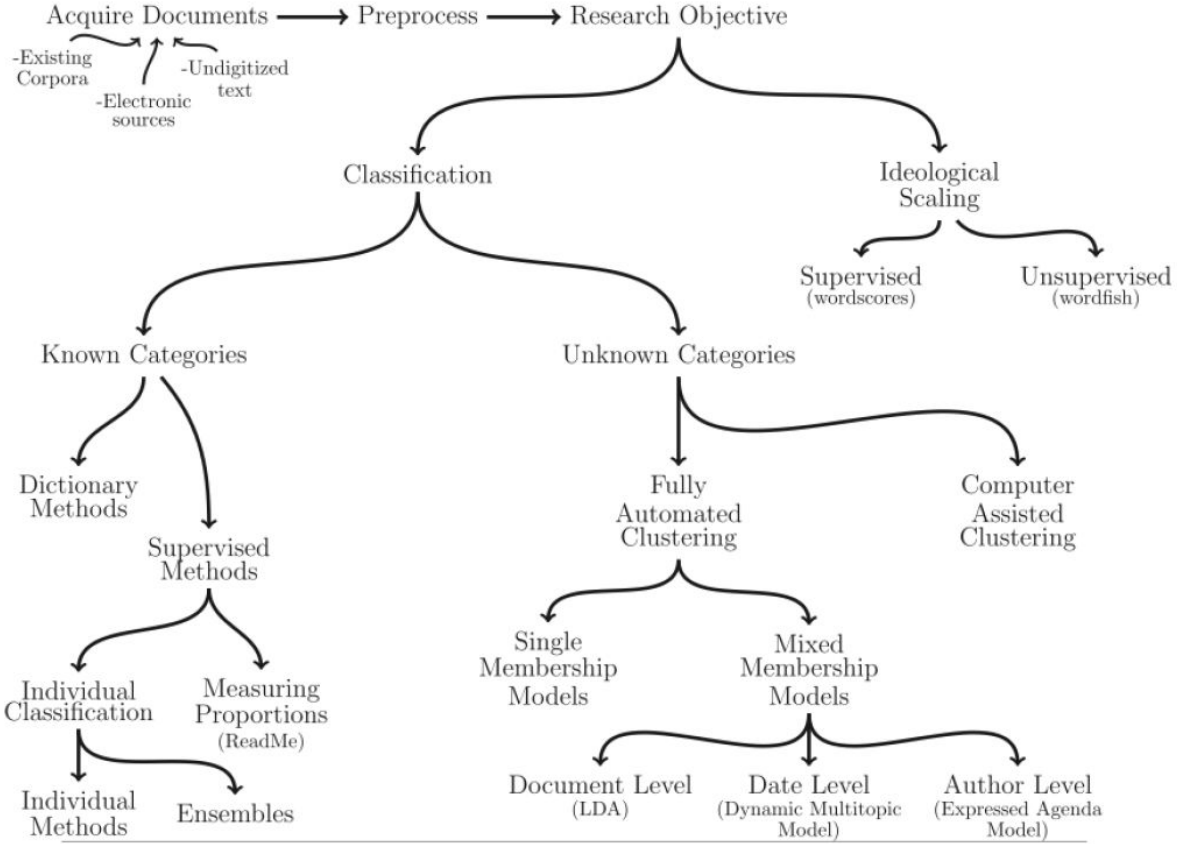
Case study 2 Data selection

You plan to study the emotional reactions of Austrian citizens in response to the attacks on work of arts.

What text data could you select?

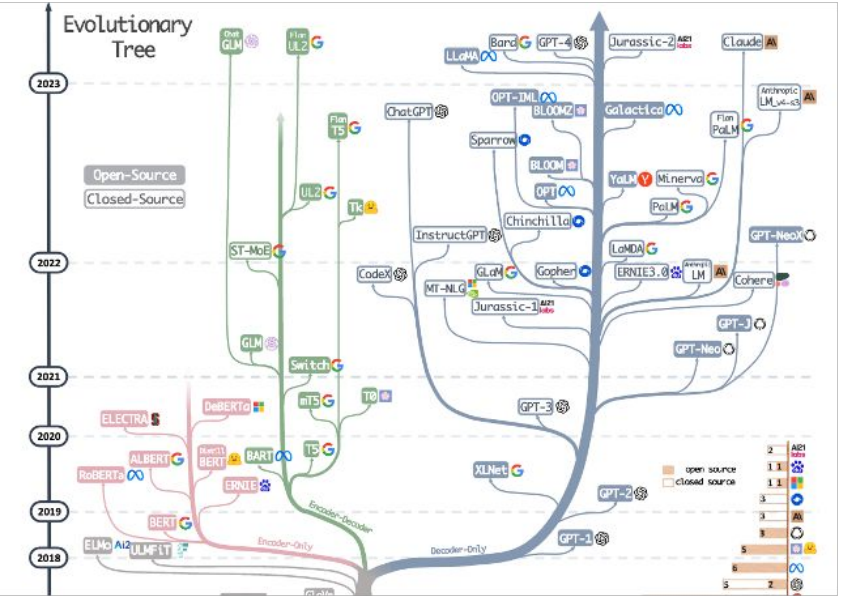
How would you select only relevant content?

And what about validation?



Grimmer & Stuart, 2013, Fig 1

Discussion



Yang et al., 2023, p. 3

- The ways that we conduct multilingual CTA is changing drastically, but questions that we ask about validation do not
- Practical implementation of validation strategies requires significant resources
  - Research infrastructures, open science initiatives, international collaboration

universität wien

Course assessment

Participation in class (20%)

Final paper: application of one or several automated text analysis methods on a topic related to the PhD thesis or a topic of free choice (80%)

- Contents: short motivation, analysis (commented code), description and interpretation of results (about 10 pages)
- Format: R Markdown
- Deadline: February 15th, 2024
- Send it to both of us via mail

universität wien

Final paper evaluation

Max. 80 points

Consistency (20 points): Motivation (objectives), analysis and interpretations of results are closely related

Readability/Format (20 points): Code is commented, with text

‘Correct’ application of method (20 points):

Critical reflection (20 points): some reflections for bigger decisions in the design and in respect to the result

universität wien

Course evaluation

universität wien

Pitch your projects

Very informal opportunity to pitch your text analysis use case and (initial) design

- Research question
- Data
- Methods
- Current struggles

And to receive some feedback (no grades, points, etc. just free brainstorming opportunity)

universität wien

Thank you very much