Multi-National Topics Maps for Parliamentary Debate Analysis

55th Hawaii International Conference on System Sciences (HICSS)
3 – 7th of January 2021

Markus Schaal, Enno Davis, Roland M. Müller
Berlin School of Economics and Law

Corresponding Author Email: schaal@hwr-berlin.de

Agenda

1. Motivation
2. Design Science As Our Methodology
3. Multi-national Political Topic Modeling
   - The ParlSpeechV2 data set
   - Latent Dirichlet Allocation and Coherence Score
   - Preprocessing
   - Reference Model
   - Linking the Topics
   - Results
4. Conclusions

Motivation

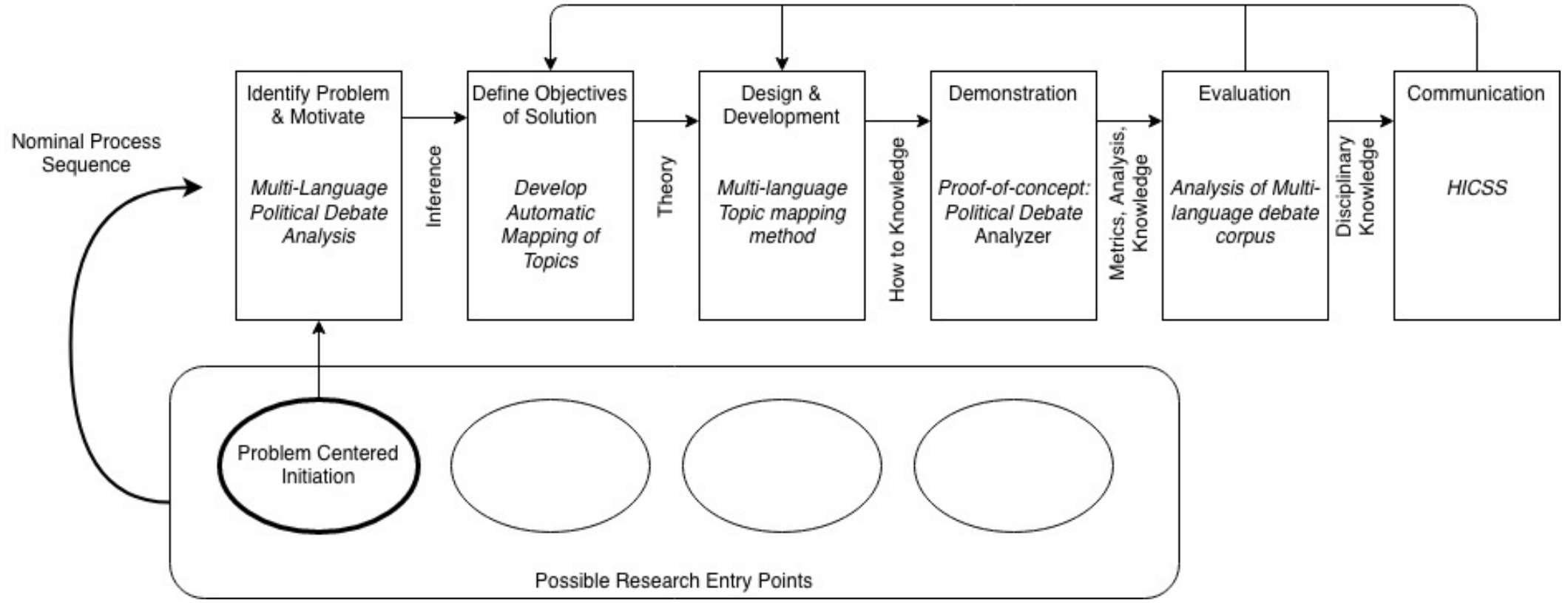
Linking parliamentary speech topics cross-nationally

Challenge:

- *different concepts in different languages*
- *addressed by latent topic models*
- *but linking unclear*

Design Science As Our Methodology

Research Methodology (DSRM)



c.f., Pfeffers, et al. 2007

Guidelines for Design Science Research

Guidelines		Our Application
Design as an Artifact	→	Towards multi-language political debate analysis
Problem Relevance	→	Political debates are important for democracy
Design Evaluation	→	Coherence score to evaluate topic alignment
Research Contribution	→	How to use LDA for cross-national transparency
Research Rigor	→	**Few and clear guidelines for design decisions**
Design as Search Process	→	Multiple alternatives are compared
Communication of Research	→	**This paper**

c.f., Hevner, et al. 2004

Multi-national Political Topic Modeling

ParlSpeechV2

– *"Full-text corpora of 6.3 million parliamentary speeches"*

c.f., Rauh and Schwalbach 2020

– Germany, United Kingdom, Spain
– Period: 1996-03-27 to 2018-12-14

  – Germany: 167943 Speeches
  – United Kingdom: 1381804 Speeches
  – Spain: 108214 Speeches

Multi-national Political Topic Modeling

Latent Dirchilet Allocation & Coherence Score

Latent Dirichlet Allocation (LDA):

Iterative procedure to find the best possible clusterings and probability distribution,

c.f., Blei et al. 2003

Coherence Score:

c-v as a novel coherence measure with the highest correlation with human ratings,

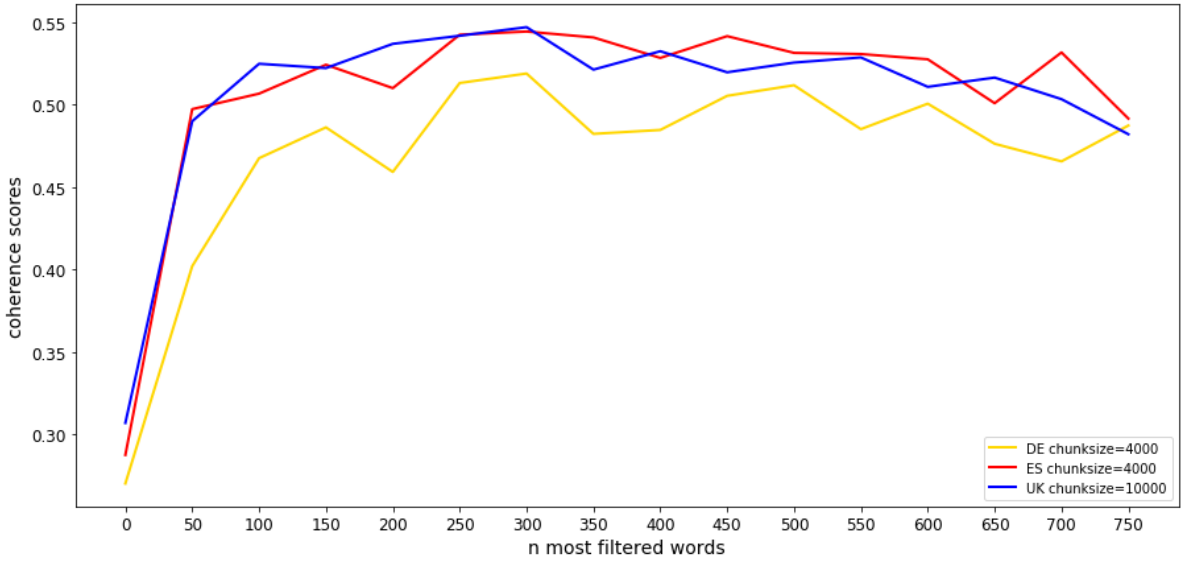
c.f., Röder et al. 2015

Preprocessing

1. Remove all text between brackets („(*text*)“)
   – Data set specifica
2. Remove punctuations
3. Remove numbers (digits 0-9)
4. Apply a Lemmatizer
5. Remove single-letter literals
6. Everything lowercase

Pass Ia: Frequency Cap Optimization

1. *filter_n_most_frequent(remove n)* method of Gensim (c.f., Radim Řehůřek and Peter Sojka 2010)
   - Frequency = #Documents in which a term occurs
2. Apply LDA with a standard (50 Topics)
3. Evaluate with **c-v** coherence score
4. using the maximum (average) coherence per national corpus



→ absolute maximum for all three national corpora is 300

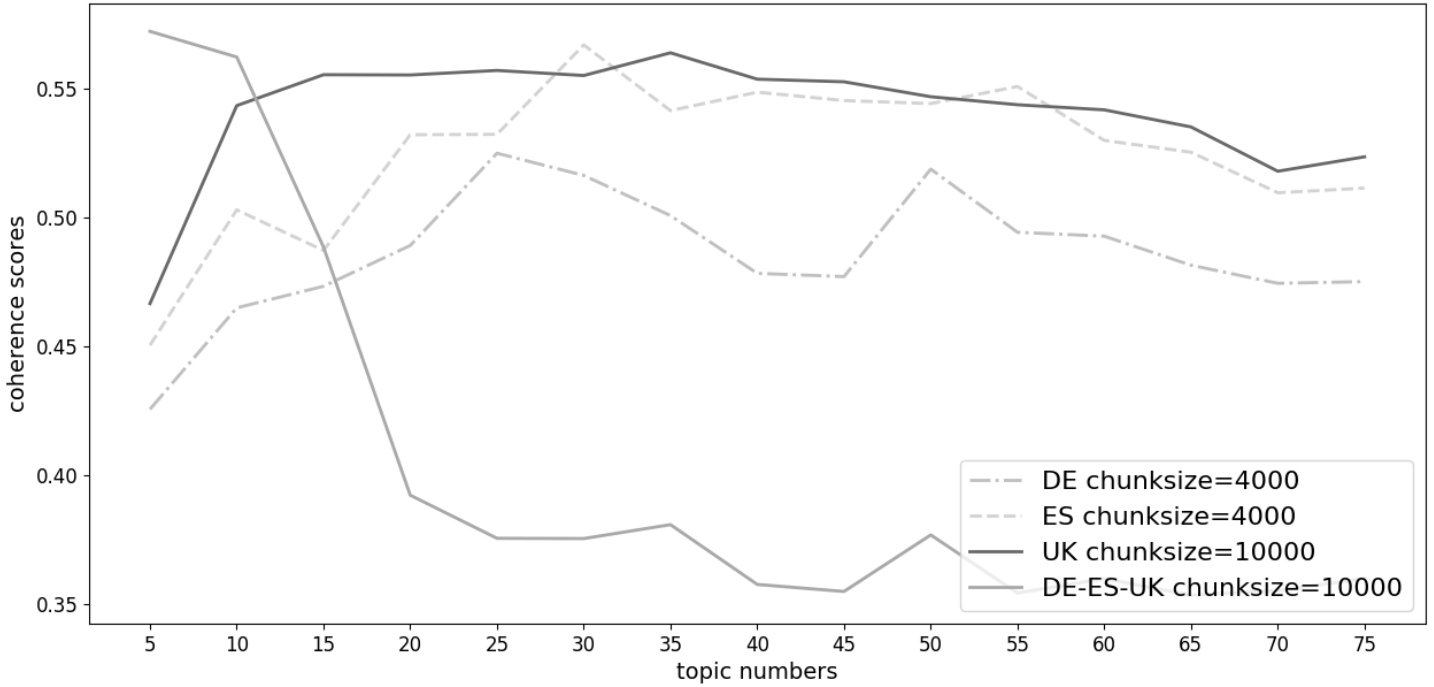
Pass Ib: Topic Number Optimization

Best Topic Models:

Germany: 25

Spain: 30

United Kingdom: 35



Multi-national Political Topic Modeling

Results: National Model - Germany

#	Topic terms	c-v
7	energie, klimaschutz, energiewende	0.738
EN	*energy, climate protection, energy transition*	
15	pflege, arzt, versorgung	0.736
EN	*care, doctor, supply*	
6	bundeswehr, einsatz, soldat	0.719
EN	*armed forces, use, soldier*	
11	türkei, syrien, menschenrechte	0.709
EN	*turkey, syria, human rights*	
19	projekt, infrastruktur, straße	0.665
EN	*project, infrastructure, street*	

Top-5 (out of 25) topics for Germany

Results: National Model - Spain

#	Topic terms	c-v
3	fiscal, imponer, impuesto	0.752
EN	*prosecutor, impose, tax*	
0	educativo, educación, formación	0.681
EN	*educational, education, training*	
8	aguar, andalucía, valenciano	0.673
EN	*water, andalusia, valencian*	
11	crecimiento, crisis, déficit	0.661
EN	*growth, crisis, deficit*	
9	justicia, judical, civil	0.652
EN	*justice, judical, civil*	

Top-5 (out of 30) topics for Spain

Results: National Model – United Kingdom

#	Topic terms	c-v
15	defence, armed, war	0.776
5	crime, prison, victim	0.731
0	care, nhs, hospital	0.727
6	international, security, foreign	0.725
33	rail, transport, train	0.721

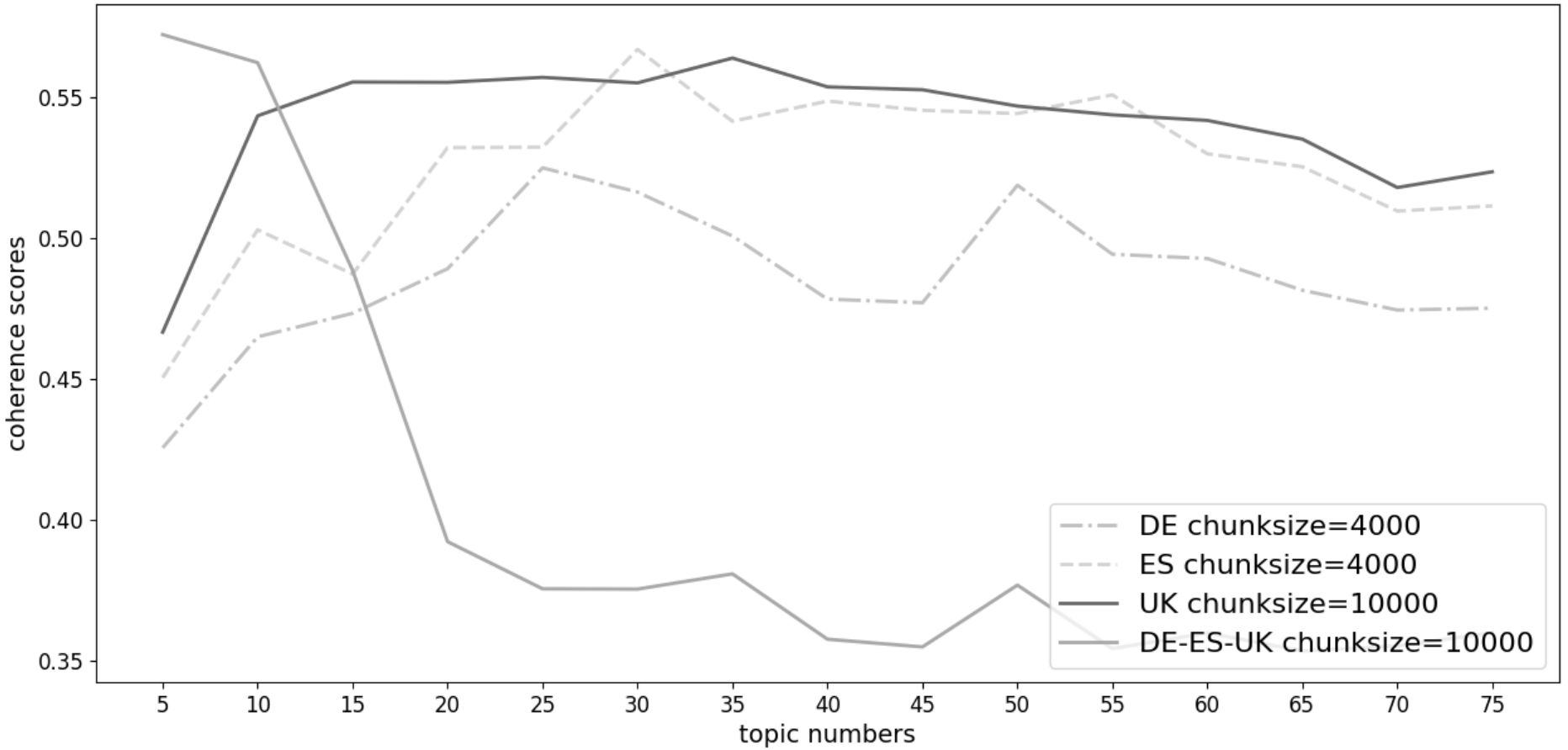
Top-5 (out of 35) topics for United Kingdom

Pass II: Unified Multi-National Corpus

1. Translate top 20 words non-englisch terms per topic
2. For each translation assign an ID
3. Multiple Translations (same source) → Keep most important
4. Multiple Translations (same target) → Unify ID
5. Multi-Term translations → Single token
6. Filter stop words
7. Replace terms in each corpus by their ID
8. Combine all corpora

Multi-national Political Topic Modeling

Pass II: Unified Multi-National Corpus



Linking the Topics

1. Only national topics with min 0.5 **c-v**
2. Cosine similarity between all possible topic pairs
3. Sort descending and cut off at 0.1 similarity
4. Highest similarity pair is a link

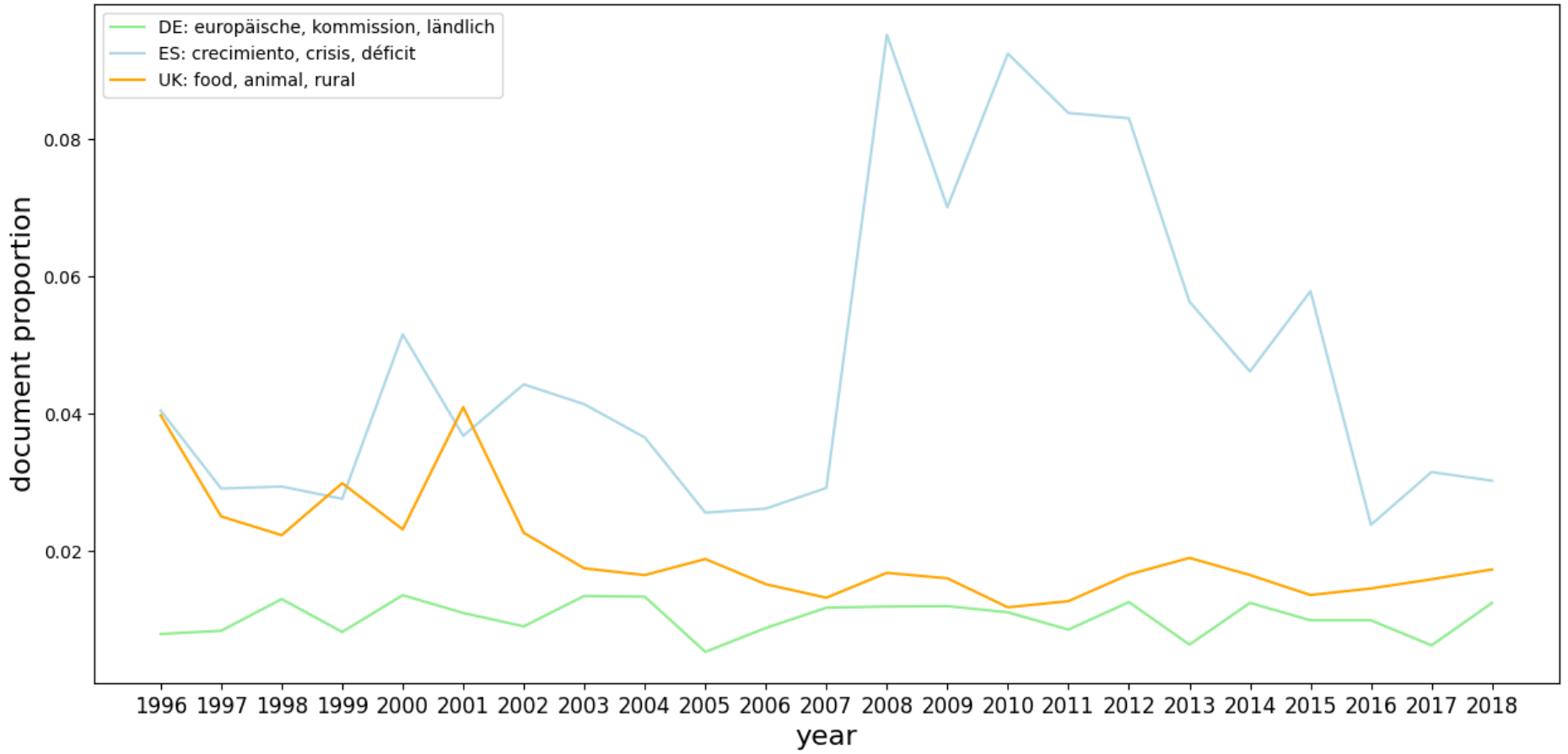
Multi-national Political Topic Modeling

Results: Reference Model

#	Germany	Spain	United Kingdom	Reference	c-v
2	zahlen, steuer, haushalt	fiscal, imponer, impeusto	tax, credit, universial	tax, bank, investment	0.357
	pay, tax, budget	*prosecuter, impose, tax*			
8		trabajador, laboral, formación	company, market, consumer	worker, contract, employer	0.36
		worker, labor, training			
11	projekt, infrastruktur, straße	inversión, sostenibilidad, obrar	rail, transport, train	investment, infrastructure, energy	0.344
	project, infrastructure, street	*investment, sustainability, work*			

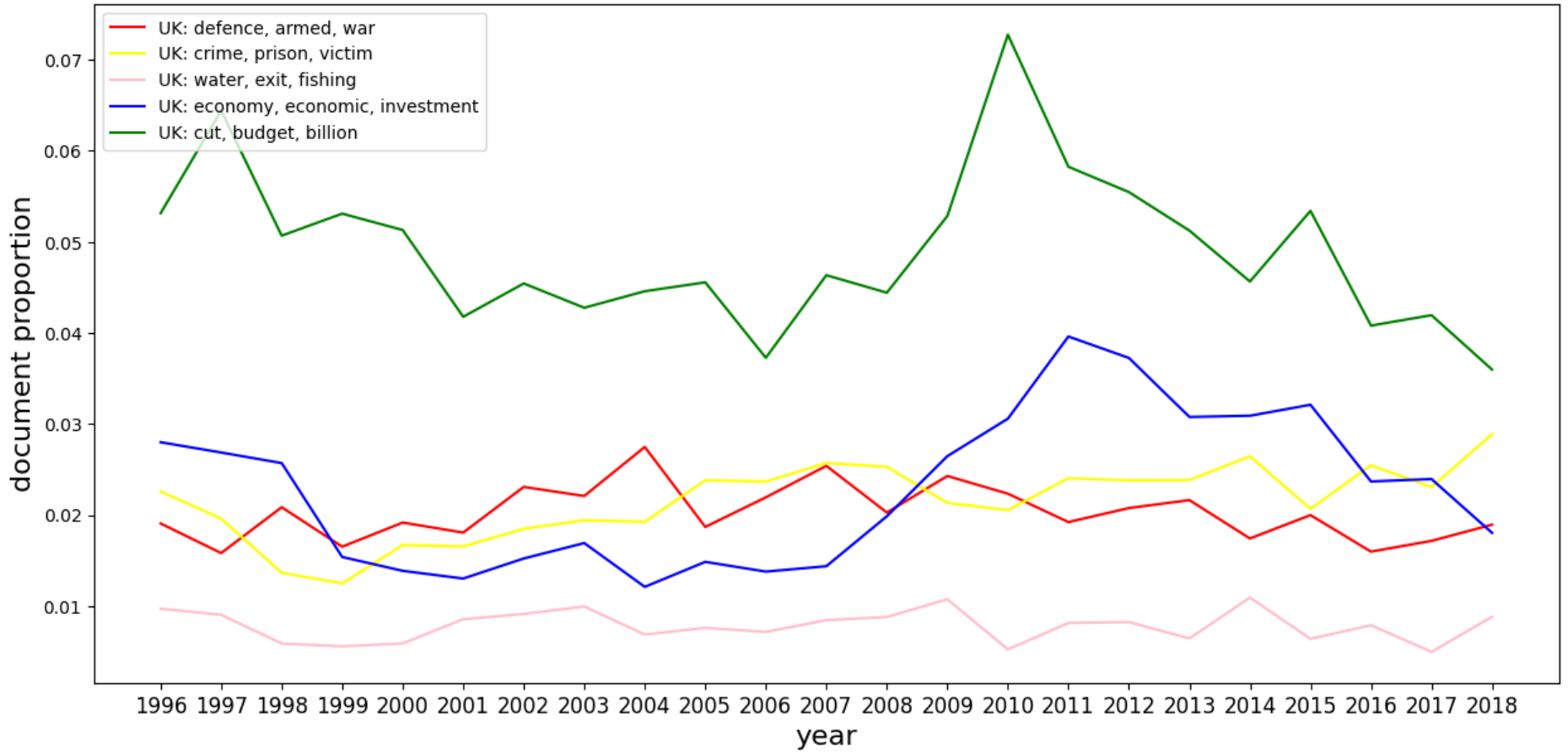
Multi-national Political Topic Modeling

Results: Reference Model



Multi-national Political Topic Modeling

Results: National Model



Conclusions

Our LDA-based process to create multi-national topics models …

1. Data-driven general approach for filtering stop words with LDA
2. Method to join multi-language corpora for probabilistic topic modeling
3. Method for topic linking

References

Pfeffers, et al. 2007
K. Peffers, T. Tuunanen, M. A. Rothenberger, and S. Chatterjee, “A design science research methodology for information systems research,” Journal of Management Information Systems, vol. 24, p. 45–77, Dec 2007.

Hevner, et al. 2004
A. R. Hevner, S. T. March, J. Park, and S. Ram, “Design science in information systems research,” MIS Quarterly, vol. 28, no. 1, p. 75–105, 2004

Rauh and Schwalbach 2020
C. Rauh and J. Schwalbach, “The ParlSpeech V2 data set: Full-text corpora of 6.3 million parliamentary speeches in the key legislative chambers of nine representative democracies,” 2020. https://doi.org/10.7910/DVN/L4OAKN.

Blei, et al. 2003
D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent dirichlet allocation,” Journal of Machine Learning Research, vol. 3, p. 993–1022, Mar. 2003.

Röder, et al. 2015
M. Röder, A. Both, and A. Hinneburg, “Exploring the space of topic coherence measures,” in Proceedings of the Eighth ACM International Conference on Web Search and Data Mining, WSDM ’15, (New York, NY, USA), p. 399–408, Association for Computing Machinery, 2015.

Radim Řehůřek and Peter Sojka 2010
R. Radim Řehůřek and P. Sojka, “Software Framework for Topic Modelling with Large Corpora,” in Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, (Valletta, Malta), pp. 45–50, ELRA, May 2010