Qualitative Comparison of Native and Machine-Translated Parliamentary Debates

Ajda Pretnar Žagar[1]

[1]*Institute of Contemporary History, Privoz 11, 1000 Ljubljana, Slovenia*

Abstract

Machine translation (MT) models have become increasingly accurate and widely accessible for multiple languages in recent years. They can potentially lift the barriers to applying NLP tools and methods to previously unsupported languages and boost comparative cross-lingual research in digital humanities. This study empirically contrasts results obtained with source and target Slovenian ParlaMint corpus of parliamentary debates on topic modelling. It qualitatively compares three steps in topic interpretation: topic description, topic significance in subcorpora, and marginal topic distribution. The results indicate that the topic modelling on the target corpus only partially replicates the topic modelling on the source corpus, but the overlap is sufficient to provide a starting point for the cross-country comparison.

Keywords

topic modelling, LDA, parliamentary data, machine translation, qualitative evaluation

1. Introduction

The proliferation of linguistically annotated, well-structured corpora enables in-depth philological, cultural, historical, and political analyses. When resources are available in several languages, as is the case of ParlaMint corpora on parliamentary speeches from 17 European countries [1], they also enable cross-country comparisons of discourses, topics, political agendas, and language development. However, comparative research of ParlaMint data requires language proficiency in more than a single language, which significantly limits transnational research.

Fortunately, machine translation (MT) models are increasingly accurate and freely available to the research community. They are a cost-efficient and fast method for converting almost any corpus to a language the researcher could understand. With state-of-the-art models approaching or sometimes even surpassing human accuracy [2], machine translation helps alleviate language barriers for comparative research in multilingual text collections.

The main research question of this paper is to what extent do bag-of-words results, specifically topic modelling, on machine-translated corpora correspond to the results on the native corpora. Given that topic modelling relies on word distributions and not on the order of words, proper grammar, and correct pronouns, in-context word-to-word translation accuracy is the most important requirement for MT, which is already relatively high with existing approaches.

Digital Parliamentary Data in Action (DiPaDA 2022) workshop, Uppsala, Sweden, March 15, 2022.

ajda.pretnar@inz.si (A. P. Žagar)

https://github.com/ajdapretnar (A. P. Žagar)

0000-0002-5927-4538 (A. P. Žagar)

Similar research has already shown that machine translations can successfully capture topics, similar to the source corpus [3], and that target corpora can be used in comparative research [4, 5].

We chose to work with a recently published ParlaMint corpus, which provides rich linguistic annotation and metadata. Moreover, we chose the Slovenian ParlaMint corpus as we are native speakers of the language and can accurately interpret the results. Slovenian generally has fewer language resources than English and is morphologically rich, resulting in poorer machine translation.

The paper qualitatively compares the outcomes of source topic models with their counterparts obtained from the target corpus. Topic overlap is estimated, not in terms of topic-term similarity, but on how similar the analytical results would be if using target corpus. Evaluation is done by comparing: a) topic interpretation in the source and target topic model, b) significant topics for pre-COVID and COVID period, and c) marginal topic distribution of both models.

2. Related work

Various techniques can be applied for a comparative analysis of topic models of multilingual corpora. Mimno et al. [6] propose Polylingual Topic Models (PTM), which can extract topics for corpora in many languages, but they require an initial set of comparable documents. PTM can be extended to unaligned documents, but not all corpora contain comparable documents. Boyd-Graber and Blei [7] further this idea by proposing multilingual topic models for unaligned documents. When the documents in different languages do not cover the same topics, which is often the case, Yang, Boyd-Graber and Resnik [8] propose a multilingual topic model to match the learned topics partially. However, all of these approaches require knowledge of the languages of the corpus.

The alternative is using machine-translated corpora and computing topic models on those. However, the results depend heavily on the quality of the translation. There are numerous approaches to automatically estimating machine translation quality. Most rely on quantitative assessment against a reference text, such as BLEU [9] and NIST [10]. Establishing community-accepted automatic scoring methods boosted research in machine translation models as it enabled fast and cost-effective evaluation of model improvements. That said, certain criticism has been raised against such evaluations. Turian, Shea and Melamed [11] argue that the correlation between human evaluation and MT quality estimates is low. Others point to the inability of such measures to capture translation improvements in syntactic and semantic quality [12]. Hence a qualitative estimation of topic model similarity between target and source texts can be a viable alternative to quantitative scores.

However, such studies are few. Reber [4] compares Google Translate and DeepL MT models on online discourses on climate change from Germany, the United Kingdom and the United States. The author uses Structural Topic Models on target corpus to compare topic prevalence in different national discourses. Maier et al. [13] empirically assess the difference in topic modelling results between machine-translated texts and multilingual dictionaries. They note the utility of both approaches but warn of method-specific differences in the results. Nevertheless, both studies demonstrate that it is possible to apply topic modelling in multilingual settings.

This contribution extends the paper from de Vries, Schoonvelde and Schumacher [3], which compare a gold standard human translation on *euparl* data set with a machine-translated corpus. They use topic modelling with LDA to compare both text sets via the generated term-document matrices, which explicitly shows that target corpora can be successfully used for extracting topics.

We likewise estimate the machine translation quality empirically, but in contrast to the above paper, we focus on a qualitative perspective. Instead, we leverage our native knowledge of the Slovenian language to estimate how close the interpretation of topic modelling of the target corpus would be to the topic modelling of the source corpus.

3. Data

The first data set[1] is ParlaMint-SI, a linguistically annotated corpus of parliamentary speeches from the Slovenian parliament from 2014 onward [14]. We will refer to it as the "source corpus". Corpus contains 414 transcribed recordings of parliamentary sessions, equipped with corresponding metadata on the parliamentary speakers and linguistic annotations of utterances, including lemmas, POS tags, and named entities.

We took the data from 2019-01-01 onward, encompassing about a year of pre-COVID and a year of COVID speeches. We parsed the corpus into 18,476 utterances, each representing a single speech given in a session. We kept only speeches given by regular MPs, as these would correspond best to topics discussed in the parliament. We also removed speeches (utterances) shorter than 50 words, as these would typically be procedural remarks [15]. In the end, the filtered corpus contained 6861 speeches.

The second data set[2] is a machine-translated version of ParlaMint-SI version 2.0. We will refer to it as the "target corpus". Machine translation was performed with *opus-mt-zls-en* model [3].

The source corpus already contains lemmas and POS tags attained with the CLASSLA pipeline [1]. We lemmatised the target corpus with the Lemmagen lemmatiser [16] and tagged it with the Averaged Perceptron Tagger from the NLTK library. The choice of lemmatiser and tagger undoubtedly introduces additional noise, resulting from imperfect preprocessing models and not the machine translation model [4].

We kept only lemmatised nouns, thus removing a large portion of tokens. The reasoning is that nouns sufficiently reflect topics in parliamentary speeches, and they are easier to interpret than, say, verbs [17]. When we tested the pipeline with nouns and verbs, there was always at least one topic with only verbs as characteristic topic words. We also removed tokens that appear in less than ten documents, as they are too niche and do not represent a topic sufficiently.

In the end, the source corpus retained 3695 types, while the target corpus had 5127. More than 30% difference in types in the target corpus is already a significant discrepancy. The difference can be attributed to the different lemmatiser (personal names were lemmatised correctly in the

[1] 10.6084/m9.figshare.19248812

[2] 10.6084/m9.figshare.19258814

[3] https://huggingface.co/Helsinki-NLP/opus-mt-zls-en

[4] A manual inspection of token differences between the source and target corpora revealed the difficulties of the lemmatiser to deal with Slovenian proper nouns and acronyms

source corpus, but not in the target one), different POS tagger (which tagged certain words erroneously), and, indeed, faulty machine translation model (which sometimes creates random repetitions of words) [5].

Compared with de Vries, Schoonvelde and Schumacher, we were stricter with the preprocessing. We kept only nouns from the source corpus, which empirically gave the best results[6] and is similar to related work [18]. It is necessary to note that our analysis is performed at the utterance level instead of the entire session transcription. Utterance-level models result in more coherent topics and enable later comparison between speakers.

4. Research Design

We compare the practical efficiency of machine translations for comparative research of multilingual corpora on topic modelling results[7]. The choice of topic modelling is in line with related work. However, we extend this with a qualitative comparison of the results. Namely, we wish to determine whether machine-translated corpus would give similar results to the native corpus. We use the Latent Dirichlet Allocation (LDA) topic model, a generative model that extracts topics based on word distributions. We compare the topic interpretation of the source and target topic model, the ranking of topic significance for pre-COVID and COVID subcorpus, and the marginal topic distribution of the two topic models.

The tasks generally correspond to typical analytical workflows in topic modelling research, namely topic identification, contrasting topic frequencies in different periods or between parties, and estimation of topic importance [19, 15, 18].

Furthermore, we aim to explore the quality of the target topic model for a language with lesser resources, namely Slovene. With the proliferation of freely available yet high-quality MT models, such as the OPUS collection from Helsinki NLP group [20] and Facebook's MBart models [21], it is now possible to translate even smaller languages successfully. We intentionally use a freely available model from the Hugging Face repository to demonstrate open-source models' increasing accuracy and accessibility.

5. Results

We extracted 20 topics with Latent Dirichlet Allocation on TF-IDF weighted bag-of-word matrix. Twenty topics is a sufficiently large number to cover a wide array of topics that can be discussed in the parliament while also being sensibly moderate to allow interpretation. Zhao et al. [22] and Rosa, Gudowsky and Repo [23] corroborate the decision for 20 topics.

The results of topic modelling with the top 10 words describing each topic are detailed in Table 3 for the target corpus and in Table 4 for the source corpus (see Appendix). We manually

[5]Machine translation accuracy cannot be estimated on the ParlaMint corpus due to the lack of a gold standard. However, authors report a BLUE score of 25.6 and character n-gram F-score of 0.407 for Slovenian to English translation on Tatoeba corpus.

[6]We tried topic modelling on all tokens, NOUN+VERB+ADJECTIVE, NOUN+VERB and only NOUN and compared the results for 5, 10, 20 and 50 topics. The pipeline that yielded the best results was only nouns with 20 topics.

[7]The Orange data mining workflow for reproducing the analysis is available at 10.6084/m9.figshare.19248806.

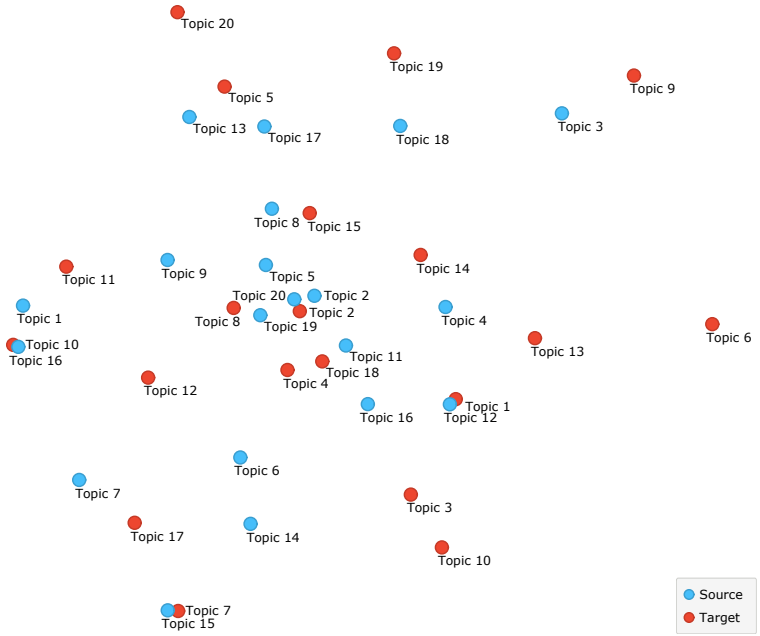


Figure 1: Topic similarity between the source and target corpus. The figure shows a t-SNE projection of topics, which were embedded on their 10 most descriptive words with a FastText word embedding model and aggregated by mean into document (topic) vectors.

translated the results into English and assigned topic names. We use SMALL CAPS to denote topics from the target corpus and **bold** to denote topics from the source corpus.

5.1. Comparison of topic modelling results

Topics extracted from the source data are semantically more cohesive, as topics 1, 3, 18, and 19 from the target corpus represent a mix of two subtopics. For example, TOPIC 3 (Table 3) from the target corpus mixes discussions on family policy (child, family, parent, allowance) with those on electoral process (election, constituency, voter). Also, TOPIC 6 is a "junk" topic with unspecific words (i, t, something, someone).

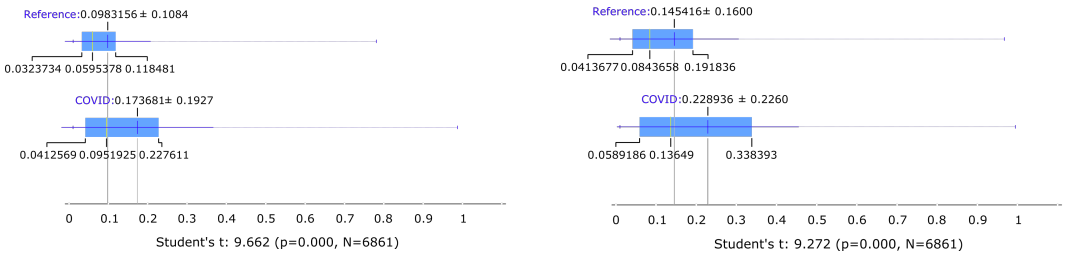
However, topic modelling on the target corpus is able to identify certain overarching topics, namely the discussions on Sunday working hours of shops, issues on education, and taxes. Other topics have partial overlap (Figure 1), such as *epidemic* (TOPIC 14 and **Topic 4**), *judiciary* (TOPIC 11 and **Topic 1**), *agriculture* (TOPIC 5 and **Topic 13**), and *credit management* (TOPIC 17 and **Topic 7**). There are two pairs of topics which display high similarity in t-distributed Stochastic Neighbour Embedding (t-SNE) projection, but we were unable to determine why they would be deemed similar, namely TOPIC 2 and **Topic 20**, and TOPIC 8 and **Topic 19**.

5.2. Evaluation of topic significance in subcorpora

Apart from determining topic overlap, we were interested in how the target topic model can replicate a more complex analytical result. Such a task can be comparing the differences between

Target	Source
Topic 14: *epidemic*	Topic 4: *epidemic*
Topic 7: *disabilities act*	Topic 3: *health care*
Topic 11: *judicial*	Topic 11: *firefighters*
Topic 15: *migration*	Topic 17: *migration*
Topic 18: *pensions & transport*	T1: *judicial*

Table 1
The top five most significant topics in the target and source corpus based on topic probabilities in the pre-COVID and COVID period.



(a) Topic 14 from target corpus

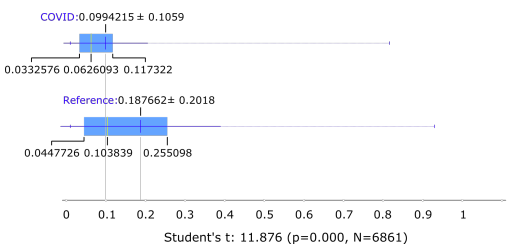
(b) Topic 4 from source corpus

Figure 2: Comparison of the topics with the highest t-test scores, computed on the reference (pre-COVID) and COVID subcorpora. Both topics describe the epidemic and share similar test statistic.

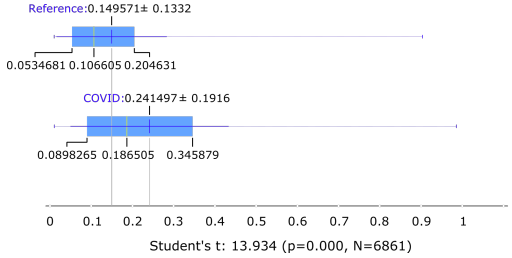
two subcorpora. The ParlaMint data set contains rich metadata with pre- and post-COVID speeches annotated. Thus we chose to compare topic prevalence in these two time periods. We determined which topics were more significant for the pre-COVID (label Reference) period and the pandemic period (label COVID) with the source corpus.

We used Student's t-test to compare the differences in the topic distribution in the reference and COVID subcorpora. Topics with the highest test statistic denote more strongly represented topics in a specific period. For example, Figure 2a shows that in the target corpus Topic 14, which is about the epidemic, was more frequent in the COVID subcorpus compared to the reference subcorpus. The same is true for the source corpus, where Topic 4 represents the epidemic and is also the highest-ranked topic. The two topics even share a similar test statistic, which shows that, at least for this topic, relative word frequencies were successfully retained in the machine translation.

Topic ranks are listed in Table 1. Besides the epidemic, judiciary and migration topics were among the five highest-ranked topics. The overlap of topic ranks is partial. Three out of five top-ranked topics were identified in both the target and the source corpus. Health care formed a separate topic from the epidemic in the source corpus, while in the target corpus, pensions and transport were collated in a single topic—certainly, even small shifts in word distributions affect topic models and the results extracted from them.



(a) Topic 7 from target corpus



(b) Topic 3 from source corpus

Figure 3: Comparison of the topics with the second highest t-test scores, computed on the reference (pre-COVID) and COVID subcorpora. Topic 7 from target corpus describes the debate around the Pension and Disabilities Insurance Act, while Topic 3 covers health care.

5.3. Comparison of marginal topic frequencies

Finally, we observed marginal topic frequencies, showing which topics appear with greater probability in each corpus (Table 2). The target corpus suffers greatly from an overestimation of meaningless topics. At the top is Topic 16 containing procedural words, such as "law", "article", "amendment", and "draft". Topic 6 includes uninformative words, such as "t", "thing", "something", and "someone". Certain topics seem to be affected greatly by shifting word frequencies, such as *epidemic*, *judicial* and *budget allocation*, the latter completely disappearing from the target topic model.

Topic frequencies show a less promising picture of target topic models. As this particular target topic model seems to mix two topics often, topic frequencies will overlap less. The most represented topics in the source corpus correspond well to the parliamentary agenda, namely budget allocation, the COVID epidemic, infrastructure issues and pensions. Target topics do not reveal the same agenda, giving a less-than-clear picture of parliamentary discussions.

6. Conclusion

While certainly imperfect, machine translation can help researchers explore corpora in their non-native languages to some extent. The findings imply that machine-translated corpora can be used by researchers who are not fluent in a specific language but with limited success. In terms of topic modelling, LDA extracted topics, generally comparable with the source corpus, thus enabling a cross-country semantic comparison of parliamentary data sets in the English language.

On the example of ParlaMint-SI corpus, the topic model of the target corpus identified three topics, identical with the source topic model, while many topics at least partially overlapped. Certain topics in the target model were still relevant parliamentary topics, such as railway infrastructure, migration, and bank audits, even though they were not identified in the source corpus. Machine-translated word frequencies were different enough that LDA could not capture the same topics, but it did offer other relevant sub-topics.

That said, the target topic model reveals a high bias towards topics with generic words, overestimating the importance of procedural words in topic identification. It also merges

Target	Source
Topic 16: 0.0990851 (*procedural*)	Topic 15: 0.081623 (*budget allocation*)
Topic 6: 0.0767897 (*- no topic -*)	Topic 3: 0.070251 (*epidemic*)
Topic 17: 0.0706852 (*housing*)	Topic 5: 0.0692074 (*infrastructure*)
Topic 15: 0.0610067 (*migration*)	Topic 6: 0.0667152 (*pensions*)
Topic 10: 0.0589544 (*Sunday work*)	Topic 16: 0.064667 (*Sunday work*)
Topic 9: 0.0589093 (*media*)	Topic 1: 0.0602505 (*judicial*)
Topic 4: 0.0516202 (*economy*)	Topic 4: 0.0568706 (*health care*)
Topic 18: 0.049224 (*pension and transport*)	Topic 10: 0.0556865 (*procedural*)
Topic 8: 0.0482725 (*infrastructure and ecology*)	Topic 19: 0.0498313 (*inspection*)
Topic 12: 0.0481855 (*bank audit*)	Topic 7: 0.0488244 (*bank system*)
Topic 13: 0.0461726 (*health care*)	Topic 20: 0.046444 (*countryside*)
Topic 19: 0.0438473 (*army*)	Topic 9: 0.0451169 (*police*)
Topic 7: 0.0429779 (*disability act*)	Topic 17: 0.0449125 (*migration*)
Topic 11: 0.0422667 (*judicial*)	Topic 12: 0.044142 (*education*)
Topic 5: 0.0385346 (*agriculture*)	Topic 2: 0.0436232 (*regional development*)
Topic 1: 0.0362681 (*sport and education*)	Topic 18: 0.0416208 (*army*)
Topic 14: 0.0361004 (*epidemic*)	Topic 8: 0.0327569 (*hazardous waste*)
Topic 2: 0.0314906 (*regional development*)	Topic 13: 0.0271814 (*agriculture*)
Topic 20: 0.0288681 (*railways*)	Topic 11: 0.0266281 (*firefighters*)
Topic 3: 0.0268099 (*family and election*)	Topic 14: 0.0192786 (*railways*)

Table 2
Topic modelling results for target and source data, ranked by their marginal topic probabilities.

specific topics into one, which makes the identified topic difficult to interpret (i.e. *family and election*, *sport and education*). Stronger preprocessing could be applied to the target corpus to remove key MT errors, such as duplicating words and erroneous translation of personal names.

In the future, we plan to provide annotated machine-translated ParlaMint corpora for all 16 languages (the UK corpus is already in English). The annotated corpus will enable a more accurate comparison of the results with identical preprocessing. Nevertheless, machine translation can be a viable first option for non-fluent researchers even in its imperfect current form.

Acknowledgments

The work described in this paper was funded by the Slovenian Research Agency research programme P6-0436: Digital Humanities: resources, tools and methods (2022-2027) and the Research Infrastructure CLARIN ERIC flagship project ParlaMint (2020-2023).

References

[1] T. Erjavec, M. Ogrodniczuk, P. Osenova, N. Ljubešić, K. Simov, V. Grigorova, M. Rudolf, A. Pančur, M. Kopp, S. Barkarson, S. Steingrímsson, H. van der Pol, G. Depoorter, J. de Does, B. Jongejan, D. Haltrup Hansen, C. Navarretta, M. Calzada Pérez, L. D. de Macedo, R. van Heusden, M. Marx, Ç. Çöltekin, M. Coole, T. Agnoloni, F. Frontini, S. Montemagni,

V. Quochi, G. Venturi, M. Ruisi, C. Marchetti, R. Battistoni, M. Sebők, O. Ring, R. Darģis, A. Utka, M. Petkevičius, M. Briedienė, T. Krilavičius, V. Morkevičius, S. Diwersy, G. Luxardo, P. Rayson, The parlamint corpora of parliamentary proceedings, 2022. (in press).

[2] M. Popel, M. Tomkova, J. Tomek, Ł. Kaiser, J. Uszkoreit, O. Bojar, Z. Žabokrtskỳ, Transforming machine translation: a deep learning system reaches news translation quality comparable to human professionals, Nature communications 11 (2020) 1–15.

[3] E. de Vries, M. Schoonvelde, G. Schumacher, No longer lost in translation: Evidence that google translate works for comparative bag-of-words text applications, Political Analysis 26 (2018) 417–430. URL: https://www.jstor.org/stable/26563863.

[4] U. Reber, Overcoming language barriers: Assessing the potential of machine translation and topic modeling for the comparative analysis of multilingual text corpora, Communication Methods and Measures 13 (2019) 102–125. URL: https://doi.org/10.1080/19312458.2018.1555798. doi:10.1080/19312458.2018.1555798. arXiv:https://doi.org/10.1080/19312458.2018.1555798.

[5] J. Schwalbach, C. Rauh, Collecting large-scale comparative text data on legislative debates, in: H. Bäck, M. Debus, J. M. Fernandes (Eds.), The Politics of Legislative Debates, Oxford University Press, Oxford, 2021, pp. 91–109.

[6] D. Mimno, H. M. Wallach, J. Naradowsky, D. A. Smith, A. McCallum, Polylingual topic models, in: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Singapore, 2009, pp. 880–889. URL: https://aclanthology.org/D09-1092.

[7] J. Boyd-Graber, D. M. Blei, Multilingual topic models for unaligned text, in: Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence, UAI '09, AUAI Press, Arlington, Virginia, USA, 2009, p. 75–82.

[8] W. Yang, J. Boyd-Graber, P. Resnik, A multilingual topic model for learning weighted topic links across corpora with low comparability, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 1243–1248. URL: https://aclanthology.org/D19-1120. doi:10.18653/v1/D19-1120.

[9] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: a method for automatic evaluation of machine translation, in: Proceedings of the 40th annual meeting of the Association for Computational Linguistics, 2002, pp. 311–318.

[10] G. Doddington, Automatic evaluation of machine translation quality using n-gram co-occurrence statistics, in: Proceedings of the second international conference on Human Language Technology Research, 2002, pp. 138–145.

[11] J. P. Turian, L. Shea, I. D. Melamed, Evaluation of machine translation and its evaluation, Technical Report, NEW YORK UNIV NY, 2006.

[12] J. Giménez, L. Márquez, Linguistic measures for automatic machine translation evaluation, Machine Translation 24 (2010) 209–240. URL: http://www.jstor.org/stable/41410948.

[13] D. Maier, C. Baden, D. Stoltenberg, M. D. Vries-Kedem, A. Waldherr, Machine translation vs. multilingual dictionaries assessing two strategies for the topic modeling of multilingual text collections, Communication Methods and Measures 0 (2021) 1–20. URL: https://doi.org/10.1080/19312458.2021.1955845. doi:10.1080/19312458.2021.

1955845. arXiv:https://doi.org/10.1080/19312458.2021.1955845.
[14] T. Erjavec, M. Ogrodniczuk, P. Osenova, N. Ljubešić, K. Simov, V. Grigorova, M. Rudolf, A. Pančur, M. Kopp, S. Barkarson, S. Steingrímsson, H. van der Pol, G. Depoorter, J. de Does, B. Jongejan, D. Haltrup Hansen, C. Navarretta, M. Calzada Pérez, L. D. de Macedo, R. van Heusden, M. Marx, Ç. Çöltekin, M. Coole, T. Agnoloni, F. Frontini, S. Montemagni, V. Quochi, G. Venturi, M. Ruisi, C. Marchetti, R. Battistoni, M. Sebők, O. Ring, R. Darģis, A. Utka, M. Petkevičius, M. Briedienė, T. Krilavičius, V. Morkevičius, S. Diwersy, G. Luxardo, P. Rayson, Multilingual comparable corpora of parliamentary debates ParlaMint 2.1, 2021. URL: http://hdl.handle.net/11356/1432, slovenian language resource repository CLARIN.SI.
[15] B. Curran, K. Higham, E. Ortiz, D. Vasques Filho, Look who's talking: Two-mode networks as representations of a topic model of new zealand parliamentary speeches, PloS one 13 (2018) e0199072.
[16] M. Juršič, I. Mozetič, T. Erjavec, N. Lavrač, Lemmagen: Multilingual lemmatisation with induced ripple-down rules, Journal of Universal Computer Science 16 (2010) 1190–1214.
[17] F. Martin, M. Johnson, More efficient topic modelling through a noun only approach, in: Proceedings of the Australasian Language Technology Association Workshop 2015, 2015, pp. 111–115.
[18] M. Moilanen, S. Østbye, Doublespeak? sustainability in the arctic—a text mining analysis of norwegian parliamentary speeches, Sustainability 13 (2021) 9397.
[19] T. Sakamoto, H. Takikawa, Cross-national measurement of polarization in political discourse: Analyzing floor debate in the u.s. the japanese legislatures, 2017 IEEE International Conference on Big Data (Big Data) (2017) 3104–3110.
[20] J. Tiedemann, S. Thottingal, Opus-mt–building open translation services for the world, in: Proceedings of the 22nd Annual Conference of the European Association for Machine Translation, 2020, pp. 479–480.
[21] A. Conneau, G. Lample, Cross-lingual language model pretraining, Advances in Neural Information Processing Systems 32 (2019) 7059–7069.
[22] W. Zhao, J. J. Chen, R. Perkins, Z. Liu, W. Ge, Y. Ding, W. Zou, A heuristic approach to determine an appropriate number of topics in topic modeling, in: BMC bioinformatics, volume 16, Springer, 2015, pp. 1–10.
[23] A. B. Rosa, N. Gudowsky, P. Repo, Sensemaking and lens-shaping: Identifying citizen contributions to foresight through comparative topic modelling, Futures 129 (2021) 102733.

Table 3
LDA with 20 topics on target corpus

Topic 1	sport, school, student, child, education, athlete, parent, organisation, science, holiday
Topic 2	perspective, development, cohesion, policy, kilometer, home, union, minister, debate, variant
Topic 3	child, election, family, constituency, voter, van, parent, allowance, compensation, cost
Topic 4	budget, trader, government, coalition, shop, opposition, party, money, economy, sarca
Topic 5	animal, culture, food, agriculture, wood, park, book, technology, hunt, conservation
Topic 6	t, thing, i, virus, something, someone, m, everything, money, nothing
Topic 7	tax, income, right, ombudsman, relief, class, rate, veteran, deaf, contribution
Topic 8	project, water, waste, infrastructure, construction, municipality, packaging, investment, development, fund
Topic 9	tv, programme, interpretation, decision, court, janša, minister, member, rtv, janez
Topic 10	worker, sunday, work, trade, package, employee, employer, measure, job, day
Topic 11	investigation, candidate, judge, court, justice, prosecutor, prosecution, branch, crime, prison
Topic 12	bank, auditor, commission, investigation, procurement, dutb, court, report, audit, business
Topic 13	health, insurance, doctor, care, patient, heart, system, surgery, hospital, wait
Topic 14	equipment, mask, fan, purchase, reserve, march, vaccination, government, infection, minister
Topic 15	referendum, migrant, migration, woman, border, freedom, right, country, citizen, violence
Topic 16	law, article, amendment, draft, group, rule, proposal, procedure, provision, service
Topic 17	housing, fund, investment, apartment, budget, estate, credit, debt, crisis, guarantee
Topic 18	pension, transport, tachograph, vehicle, pensioner, driver, road, energy, safety, disability
Topic 19	army, defence, app, weapon, arm, security, police, application, border, soldier
Topic 20	railway, plant, road, traffic, station, transport, passenger, rail, crossing, land

Table 4
LDA with 20 topics on source data

Topic 1	court, constitution, procedure, (judicial) decision, judge, authority, (making a) decision, law, justice, protection
Topic 2	perspective, candidate (m), candidate (f), drawings (of funds), culture, minister (f), cohesion, program, means, development
Topic 3	opposition, government, measure, human, epidemic, crisis, economy, TV show, medium, virus
Topic 4	doctor, equipment, healthcare, institution, health, mask, purchase, minister (male), hospital, patient
Topic 5	project, infrastructure, road, apartment, investment, municipality, source, construction, axis, supply
Topic 6	pension, insurance, pensioner, treasury, system, insurance company, euro, abolition, period, year
Topic 7	bank, fund, asset, investment, credit, management, obligation, claim, law, company
Topic 8	waste, medicinal product, park, agency, transport, product, society, tachograph, use, ton
Topic 9	weapon, act, punishment, police, prison, information, authorisation, prevention, victim, authority
Topic 10	article, amendment, committee, law, proposal, assembly, rules of procedure, session, opinion, job
Topic 11	vehicle, firefighter, driver, driving, society, exam, category, accident, training, centre
Topic 12	school, child, sport, education (process), parent, program, education (system), athlete, kindergarten, financing
Topic 13	animal, directive, (food) product, being, crop, power plant, energy, food, agriculture, law
Topic 14	passage, vaccination, track, train, VAT, harmonisation, service, railway, transport, book
Topic 15	budget, supplementary budget, tax, million, euro, paycheck, means, billion, municipality, income
Topic 16	worker, work, store, supplement, Sunday, employer, compensation, paycheck, student, family
Topic 17	water, border, migrant, human, migration, problem, area, country, land, nation
Topic 18	army, defence, soldier, member, minister, unit, interpellation, system, chief (female), commander
Topic 19	committee, member, organisation, report, ombudsman, medium, investigation, control, board, recommendation
Topic 20	accession, resolution, holiday, digitalisation, agriculture, homeland, technology, countryside, politics, return