Revealing the hidden treasures of parliamentary proceedings with NLP

ParlaMint

Nikola Ljubešić

Jožef Stefan Institute

Ljubljana, Slovenia

September 11, 2024 | DNDS Seminar, Central European University

What this talk is about

- Revolution in natural language processing
- Transformer language models, modelling dependencies in sequential data, text and speech, additional modalities
- Research on unprecedented dataset sizes, annotation automated
- Allows us to revisit old questions, state new questions
- Our data are ParlaMint - transcripts (and recordings) of parliamentary sessions from 26 national European parliaments, 2015-2022
- Two downstream projects - ParlaCAP (text) and ParlaSpeech (speech)
- Disclaimer: talk is focused on data methods, not research questions

The ParlaMint Project

CLARIN ERIC research infrastructure flagship project

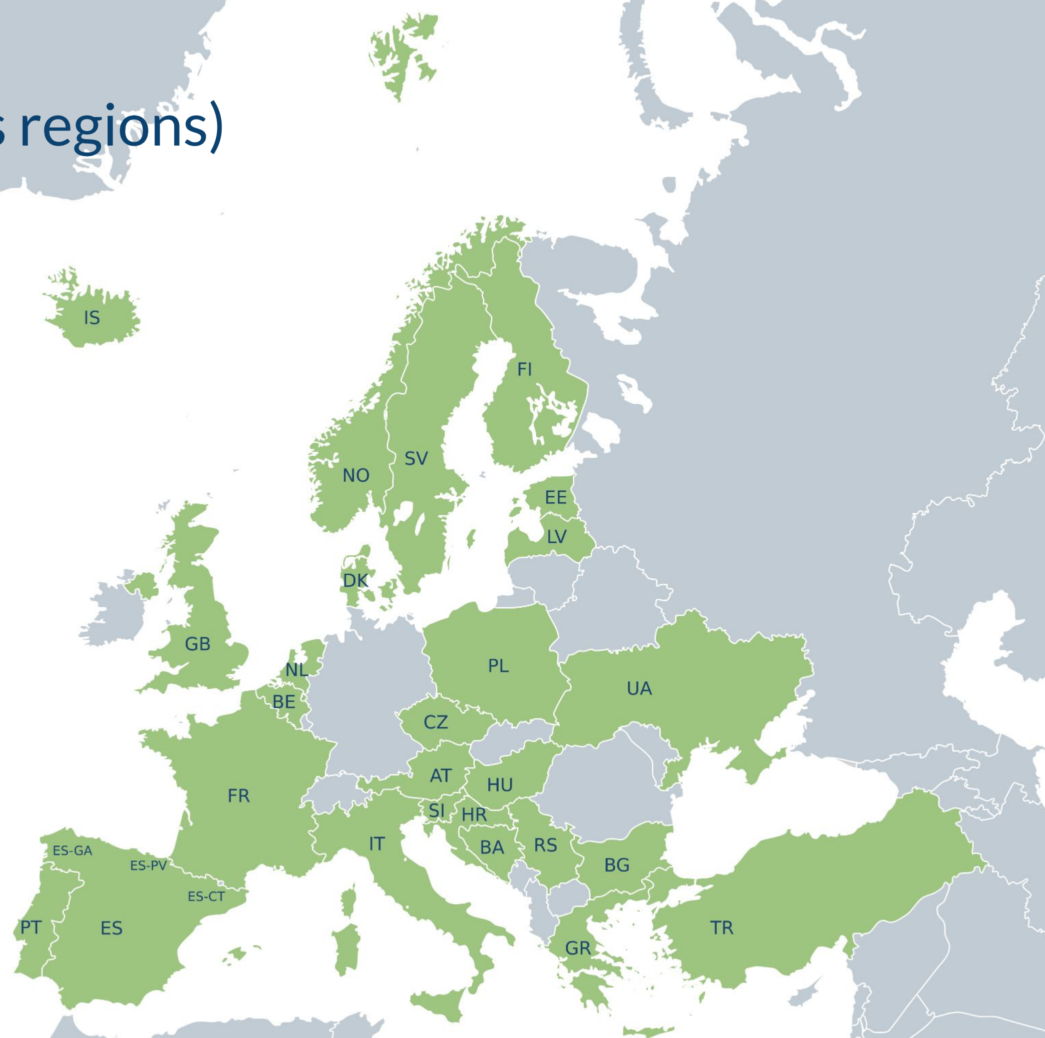
- ParlaMint I (2020–2021)
- ParlaMint II (2022-2023)

Main deliverable:

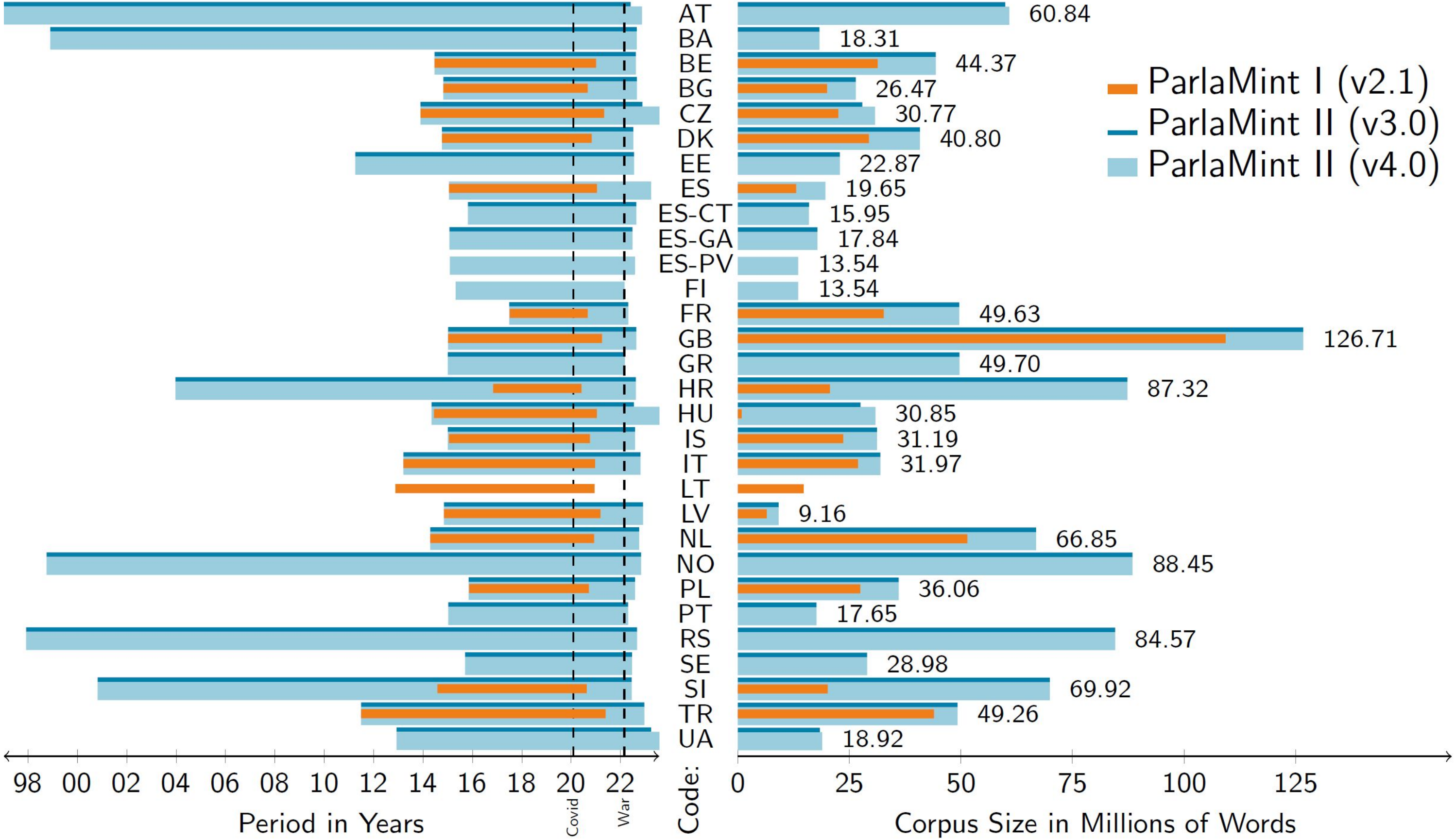
- Uniformly encoded transcriptions of speeches from European parliaments
- Rich metadata (speaker, gender, age, party, orientation, power status…)
- Linguistically annotated (part-of-speech, lemma, named entities, speeches also machine-translated into English and annotated)
- Openly available (CLARIN.SI FAIR repository and concordancer)

Geographic coverage
(26 countries and 3 autonomous regions)

Austria
Basque Country
Bosnia and Herzegovina
Belgium
Bulgaria
Catalonia
Croatia
Czech Republic
Denmark
Estonia
Finland
France
Galicia
Greece
Hungary
Iceland
Italy
Latvia
Netherlands
Norway
Poland
Portugal
Serbia
Slovenia
Spain
Sweden
Turkey
UK
Ukraine

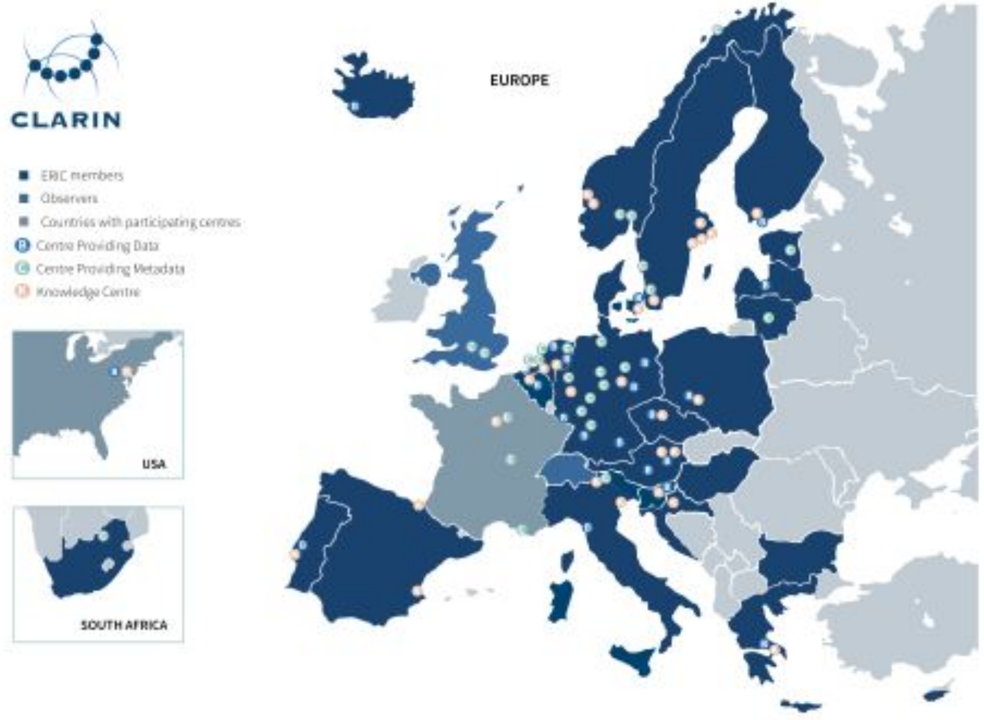


Time coverage and data size



A note on CLARIN

- CLARIN is a digital infrastructure offering data, tools and services to support research based on language resources
- A distributed network of 70 centres with 24 member countries and 2 observers

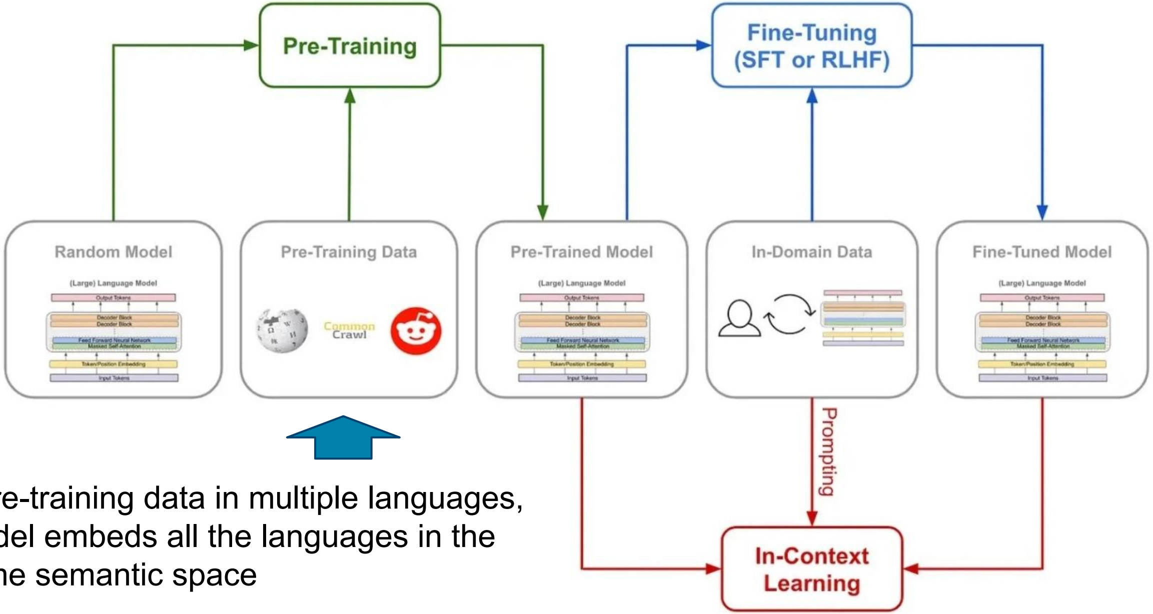


How to unlock the ParlaMint potential

- ParlaMint are primarily linguistic corpora, currently most useful to corpus and computational linguists
- Parliamentary data most relevant to social and political scientists, currently work on one of few parliaments due to data scarcity
- Social and political scientists less skilled in working with text
- “Text as data” paradigm - transform text into discrete values to be used in downstream analysis and modelling

Pre-trained language models

1. In-domain data in lang X
2. Model was pre-trained on langs X, Y, Z
3. Fine-tuned model on X will work on langs X, Y, Z.



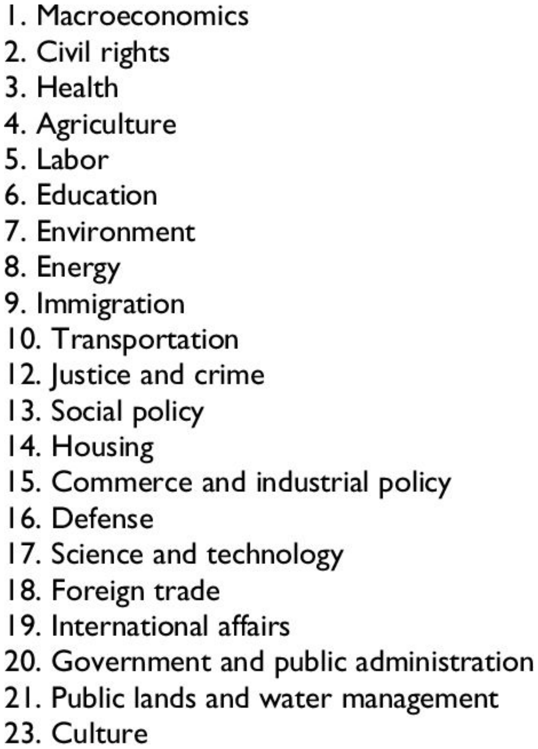
if pre-training data in multiple languages, model embeds all the languages in the same semantic space

https://medium.com/@bijit211987/the-evolution-of-language-models-pre-training-fine-tuning-and-in-context-learning-b63d4c161e49

ParlaCAP

- “Comparing agenda settings across parliaments via the ParlaMint dataset” - OSCARS Horizon Project, uptake of open science in Europe
- Cross-lingual language models to annotate more than 7 million ParlaMint speech transcripts from all 26 parliaments, 27 languages
- Annotations on topic and sentiment
- Topic schema from the Comparative Agendas Project
- Sentiment as a six-level ordinal schema, with dataset and model already developed https://huggingface.co/classla/xlm-r-parlasent

The CAP in ParlaCAP



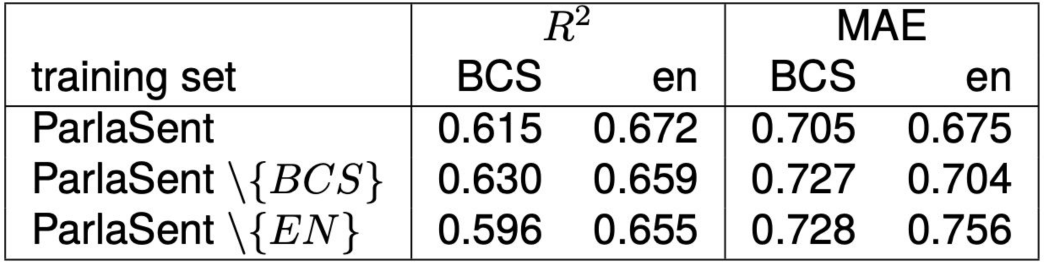
1. Macroeconomics
2. Civil rights
3. Health
4. Agriculture
5. Labor
6. Education
7. Environment
8. Energy
9. Immigration
10. Transportation
12. Justice and crime
13. Social policy
14. Housing
15. Commerce and industrial policy
16. Defense
17. Science and technology
18. Foreign trade
19. International affairs
20. Government and public administration
21. Public lands and water management
23. Culture

Comparative Agendas Project

https://www.comparativeagendas.net

Ablation experiment on cross-lingual capability

- Mochtak et al. (2024) ParlaSent sentiment paper
- Measure performance on Bosnian-Croatian-Serbian and English test
- Fine-tuning on 1. all ParlaSent and 2. with specific language removed
- No obvious trend, all results in the same ballpark



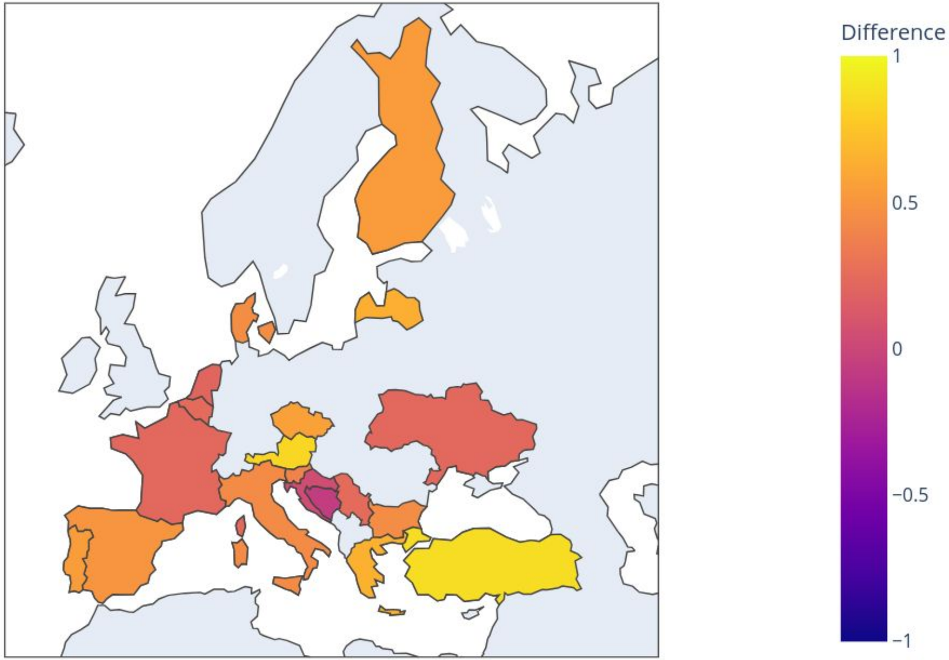
	R^2		MAE	
training set	BCS	en	BCS	en
ParlaSent	0.615	0.672	0.705	0.675
ParlaSent $\backslash\{BCS\}$	0.630	0.659	0.727	0.704
ParlaSent $\backslash\{EN\}$	0.596	0.655	0.728	0.756

Turning tables experiment

- “Opposition MPs are more negative than those from the coalition”
- Measured via surveys (Gilljam and Karlsson, 2015), voting data (Tuttnauer, 2018) or single-parliament sentiment analysis (Rheault et al. 2016; Haselmayer et al, 2022)
- Preliminary study on change in negativity on MPs when the tables turn - MP moving from coalition to opposition or vice versa
- MP-level delta of average sentiment when in coalition and in opposition
- Currently averaged on country level, many deeper insights possible

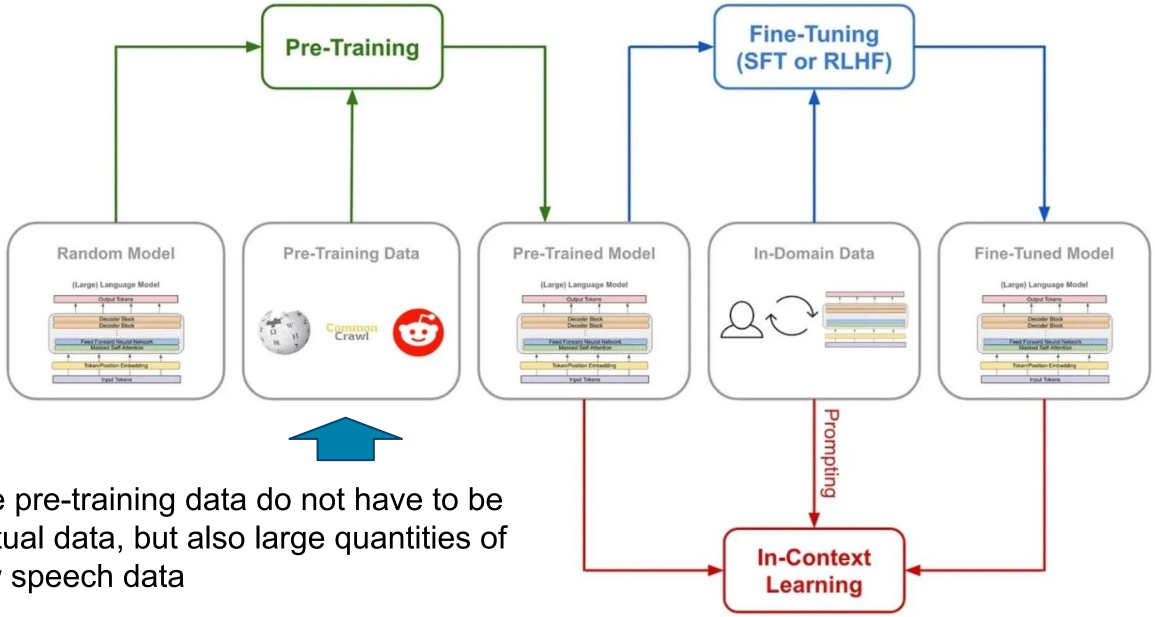
Turning tables experiment

- Measurements possible on 19 / 26 parliaments
- All 19 parliaments have difference > 0!
- Trends to be further investigated - Austria and Turkey, countries of Ex-Yugoslavia



Pre-trained language models on speech data

With small amount of labeled data, the model can learn how to identify various phenomena in speech (transcription, emotion, background sounds)



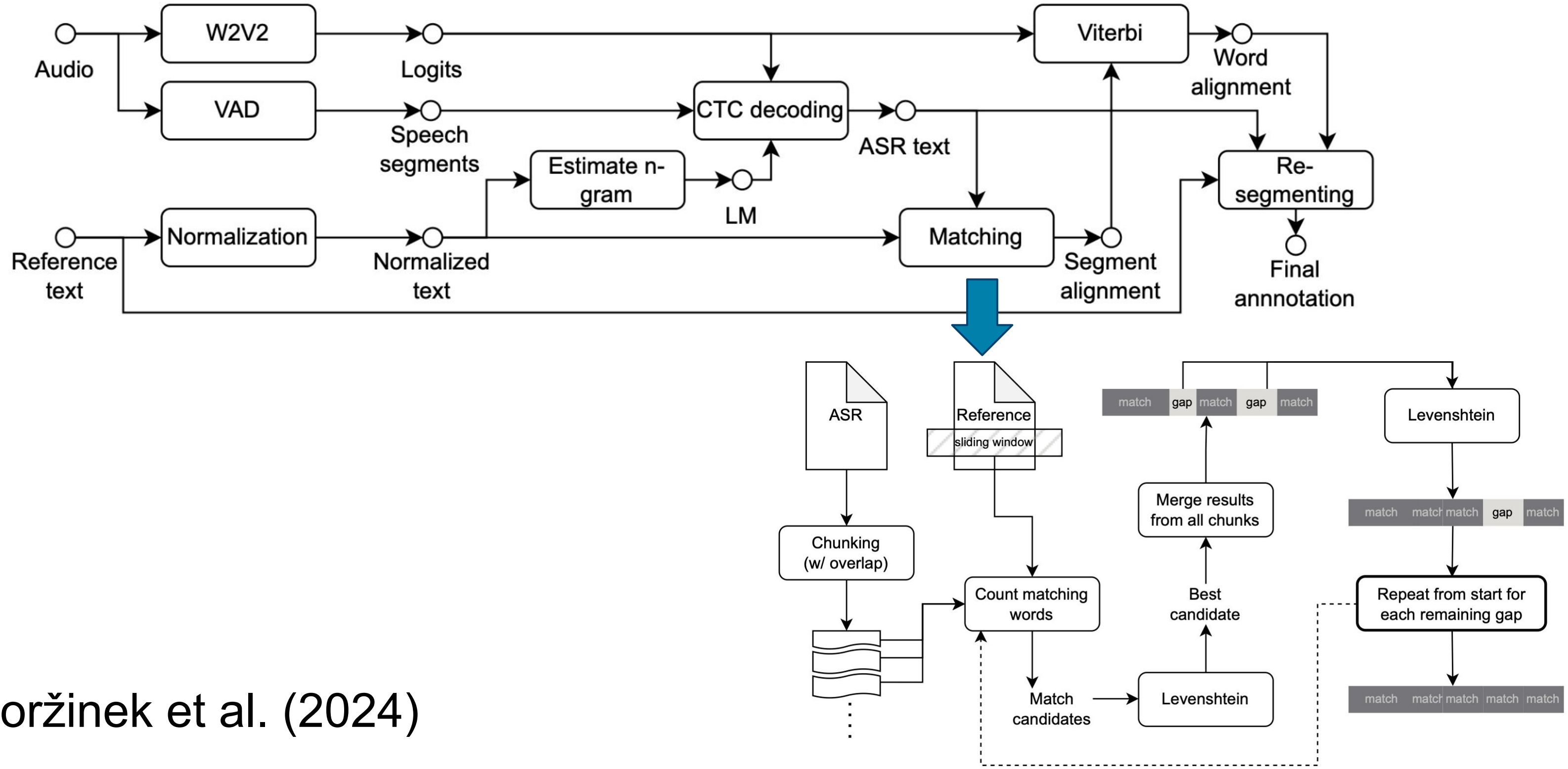
The pre-training data do not have to be textual data, but also large quantities of raw speech data

https://medium.com/@bijit211987/the-evolution-of-language-models-pre-training-fine-tuning-and-in-context-learning-b63d4c161e49

ParlaSpeech

- Task inside ParlaMint, growing into a separate project
- Aligning public domain! speech data with transcripts of the parliament
- Currently aligned are Croatian, Serbian, Polish, Czech with amount of data between 1000 and 3000 hours per language
- Easy? No.
  - Recordings are published independently of texts with spotty metadata
  - Not all recordings are released, not everything is transcribed
  - Order of transcripts and recordings is not identical

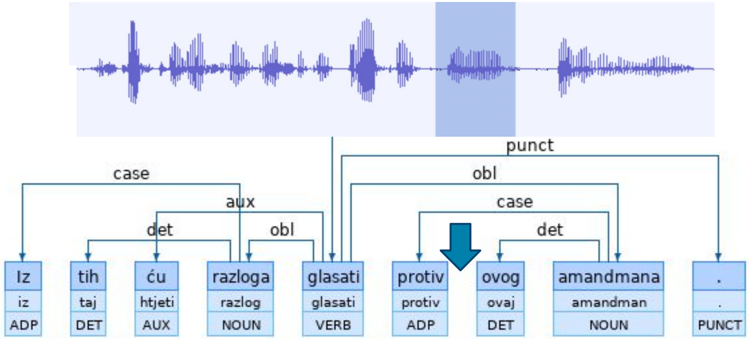
ParlaSpeech alignment procedure



Koržinek et al. (2024)

Disfluencies in spoken communication

- Semantic, emotional, pragmatic, role of disfluencies a research interest for very long (Lounsbury, 1954; Maclay and Osgood, 1959)
- Still today (Sen, 2020; Gosy, 2023), but regularly on small, manually annotated, single language and situation datasets
- w2v-bert 2.0 model fine-tuned on Slovenian filled pause (“eeem”) data, evaluated on Slovenian, Croatian, Serbian, with F1 of 0.95-0.97



To wrap up…

- New opportunities from advances in natural language processing - revisiting old and researching new questions
- Significantly larger and more diverse data at a lower cost
- Models work on multiple modalities, across languages / domains
- Limitations!, so evaluation / validation is highly advisable
- ParlaMint a rich unexplored dataset, we have just scratched the surface
- Currently we are revisiting old questions
- Collaboration with domain experts on new questions and theories

https://www.clarin.eu/parlamint

https://huggingface.co/classla

https://nljubesi.github.io