Domain-Adapting BERT for Attributing Manuscript, Century and Region in Pre-Modern Slavic Texts

Piroska Lendvai
Dept. of Digital Humanities
Bavarian Academy of Sciences
Munich, Germany
piroska.lendvai@badw.de

Uwe Reichel
audEERING GmbH, Germany &
Hungarian Research Centre for Linguistics
Budapest, Hungary
ureichel@audeering.com

Anna Jouravel and **Achim Rabus** and **Elena Renje**
Department of Slavic Languages and Literatures
University of Freiburg, Germany
anna.jouravel,achim.rabus,elena.renje@slavistik.uni-freiburg.de

Abstract

Our study presents a stratified dataset compiled from six different Slavic bodies of text, for cross-linguistic and diachronic analyses of Slavic Pre-Modern language variants. We demonstrate unsupervised domain adaptation and supervised finetuning of BERT on these low-resource, historical Slavic variants, for the purposes of provenance attribution in terms of three downstream tasks: manuscript, century and copying region classification. The data compilation aims to capture diachronic as well as regional language variation and change: the texts were written in the course of roughly a millennium, incorporating language variants from the High Middle Ages to the Early Modern Period[1], and originate from a variety of geographic regions. Mechanisms of language change in relatively small portions of such data have been inspected, analyzed and typologized by Slavists manually; our contribution aims to investigate the extent to which the BERT transformer architecture and pretrained models can benefit this process. Using these datasets for domain adaptation, we could attribute temporal, geographical and manuscript origin on the level of text snippets with high F-scores. We also conducted a qualitative analysis of the models' misclassifications.

1 Introduction

One of the prerequisites of diachronic linguistic research is the chronological and geolocational attribution of historical texts. Establishing the provenance of textual material incorporates two interwoven research areas: language history and textual history. For language history, reliable provenance attribution enables determining and categorizing linguistic features corresponding to specific time periods that can thereby uncover language change; for textual history, it facilitates the tracking of the traditions of text creation (copying and handing down) employed in manuscripts, and thereby the reconstruction of a text's archetype.

Chronological and geolocational attribution of historical texts is a laborious process that can benefit from recent advances in natural language processing (NLP): to this end, in a collaborative project between Slavic studies and language technology, we apply domain adaptation and finetuning of BERT (Devlin et al., 2019) on historical Slavic data. Our focus material consists of six bodies of text that originate from medieval and early modern manuscripts and early printings, created in South-Eastern and Eastern Europe. They had been manually transcribed and dated between the 10th-18th centuries on the manuscript level, based on codicological, linguistic and paleographical aspects. The manuscripts and early printings we examined use Cyrillic script and non-normalized orthography[2]. They pertain to the written genre of non-vernacular language and to the broader domain of religion.

The texts encompass language varieties ranging from Old Church Slavic to its later recensions; these are known to have developed under influences of a.o. geographically constrained cultural areas. Variants were formed by factors that gave rise to orthographic, lexical and morphosyntactic changes, e.g. via modernising tendencies that adapted to the vernacular usage at the geographic area where the texts got copied and compiled, but also reverse ten-

[1] According to Western classification.

[2] Written in *scriptio continua* customary for that time, where spaces are occasionally used in an unsystematic way to mark breath pauses, but our transcribed texts are word segmentated either during transcription or during HTR.

Proceedings of the 4th Workshop on Computational Approaches to Historical Language Change, pages 15–21

dencies in the form of stylistic archaizing, reintroducing specific linguistic properties characteristic of South Slavic; this was in trend at the turn of the 14th/15th centuries in certain Rus'ian literary schools, called the Second South Slavic influence (Talev, 1973).

The above heterogeneity of change-inducing factors impacted various linguistic levels, as reflected by our historical data. This poses uncharted challenges to provenance attribution, which we tackled in three downstream text classification tasks: the attribution of the properties *manuscript*, *century* and *region* performed with BERT models on texts segmented into sentence-like snippets. We also used the data for domain adaptation of BERT models, evaluating its impact on the downstream tasks.

In related work in NLP, large language models and transformer architectures have been put to use for some historical languages (Bamman and Burns, 2020; Schweter et al., 2022; Gabay et al., 2022; Manjavacas, 2022; Lendvai and Wick, 2022), but we are not aware of studies using this technology for treating historical Slavic data; Kutuzov and Pivovarova (2021) reported on a shared task for assessing semantic change for selected lexical items but based on Modern Russian data starting from the 18th century. Use cases similar to ours are described in recent studies, e.g. on chronological attribution of text with deep learning methods on historical languages (Assael et al., 2019; Liebeskind and Liebeskind, 2020; Rastas et al., 2022). Further related downstream tasks include language identification, i.e. discriminating closely related languages or varieties, where studies report on the compilation of corpora specifically for this purpose and on methods that range from classical machine learning e.g. based on frequency of character n-grams, lexical frequency and exclusivity, part-of-speech and morphology information, to deep learning approaches, a.o. based on character embeddings (Islam et al., 2011; Zampieri et al., 2019; Wu et al., 2019; Bernier-Colborne et al., 2019).

Our contributions in this paper are the following: Introducing six Pre-Modern Slavic bodies of text (henceforth: datasets) and their employment in deep learning experiments with BERT (Section 2); Describing our experimental matrix in terms of BERT models, domain adaptation procedure and setup of downstream tasks (Section 3); Evaluating and analyzing the performance scores and misclassifications of the models and sketching ongoing work (Section 4); Discussing our pilot study in terms of limitations (Section 5).

2 Data and class labeling

Table 1 presents an overview of the six datasets we used. The first three columns correspond to our three downstream text classification tasks that each designate a small set of coarse-grained target labels. In effect, we partition the same data into different subsets along a specific property, the first one *manuscript*, where BERT needs to assign to each text snippet from which manuscript this snippet comes from. For attributing the *century*, we have three classes: '10–12', '15–16' and '18': we binned data from the first two datasets; resp. from the third and fourth, resp. from the last two. For attributing the property of *region* of the texts, two classes

Manuscript	Century	Region	Place of Copying	Language	Main genre	# Snippets
Codex Suprasliensis	10-11	South	Eastern Old Bulgaria	Old Church Slavic; South Slavic recension	hagiographical-homiletic	4,831
Cyril of Jerusalem's Catechetical Lectures	11-12	East	Kyivan Rus'	Old Church Slavic; South Slavic recension; Transmitted version used: East Slavic recension	dogmatic	4,282
Dionisio corpus (printed)	15-16	South	Serbia, Macedonia	Serbian Church Slavic; South Slavic recension	liturgical	10,685
Apostolos (from the Uspensky version of the Great Menaion Reader)	16	East	Muscovy	Russian Church Slavic; East Slavic recension	gospel	14,058
Sluzhabnik 'service book'	18	South	Serbia	Serbian Church Slavic; South Slavic recension	liturgical	3,350
Elizabeth Bible (printed)	18	East	Muscovy	Russian Church Slavic; East Slavic recension	Bible translation	11,796

Table 1: Data characteristics. Online information about each body of text is available by clicking on its name.

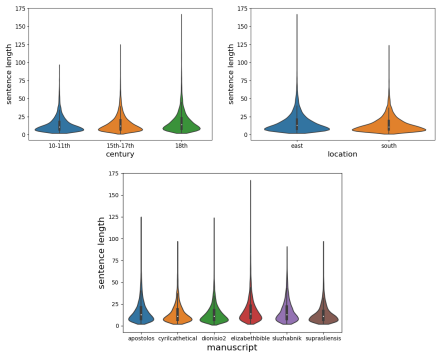

Figure 1: Violin plots showing the distribution of snippet lengths in the datasets per downstream task.

are distinguished, since the transmitted versions of manuscripts that we use have emerged either in the Southern Slavic or in the Eastern Slavic language area. It is important to see that partitioning along the spatial property (i.e., downstream task: *region* attribution) entails that the classes for that task will comprise temporally heterogeneous data (i.e., diachronic versions of the languages in that geographical area) and vice versa. In the downstream task of *manuscript* attribution, the data feature a specific combination of temporal and spatial properties that are unique to the given manuscript, etc.

The texts were available to us in transcribed form. For sentence segmentation we used Stanza (Qi et al., 2020) with *Old Church Slavonic* set as language. The segmented material qualifies as text snippets rather than syntactically complete sentences: some contain only punctuation or are very short. We discarded snippets with character length (including whitespace) less than 15 in order to remove semantically rather unintelligible strings. In Figure 1 we show the resulting distribution of snippet lengths in the respective datasets per downstream task.

For all downstream tasks the aggregated dataset was split the same way into training, development, and test partition by the ratios 80/10/10. The split was stratified on the manuscripts and was made disjunct on manuscript paragraphs, aiming to reduce potential topic overlap between partitions. For the preceding domain adaptation step the training set was further split by 90/10 into a masked language modeling (MLM) training and development set.

3 BERT experiments

For the domain adaption and finetuning experiments we report on the usage of three pretrained models; all were available in the Hugging Face repository: the multilingual model *bert-base-multilingual-uncased*, and the specifically Cyrillic models *KoichiYasuoka/bert-base-slavic-cyrillic-upos* and *anon-submission-mk/bert-base-macedonian-bulgarian-cased*. We have run a matrix of 93 model trainings: as shown in Figure 2, we compared direct finetuning of the pretrained models (henceforth also referenced as the base models) on the downstream tasks vs. domain adapting the pretrained models plus their subsequent finetuning. The pretrained models serve as baseline for each downstream task, i.e. baseline results are obtained via the experiments along the right arrow.

3.1 Domain adaptation

Vocabulary extension For domain adaptation we extended the tokenizers' vocabularies with the lexical content of the manuscripts by adding the union of the 100 most frequent words of each manuscript of at least five characters that were yet unknown to the tokenizer. We restricted the vocabulary extension in order to avoid catastrophic forgetting in the subsequent masked language modeling task.

Masked Language Modeling Subsequently, each pretrained model was domain-adapted, i.e. finetuned on the MLM task. We added the standard *BertForMaskedLM* head provided by Hugging Face for the MLM training, in effect domain-adapting the encoder weights of each pretrained model. We trained the models on the MLM task in 10 epochs with a learning rate of $2e-5$, the AdamW optimizer with a Cross Entropy loss, and a batch size of 16. We kept the best model in terms of the lowest loss on the development set. We did not perform next sentence prediction (NSP) since our current downstream tasks do not require the understanding of sentence pair relations; classification operates on the level of single text snippets and we use mean pooling for the downstream tasks. For both masked LM and subsequent finetuning on the downstream tasks, we set the maximum number of tokens to 128.

3.2 Finetuning on downstream tasks

For each of the downstream tasks we finetuned the off-the-shelf as well as the domain-adapted (see above) variants of the three pretrained models in the same way: we added a classification head to the encoder consisting of one feed-forward hidden layer with a *tanh* activation function, and a final linear output projection layer to the respective number

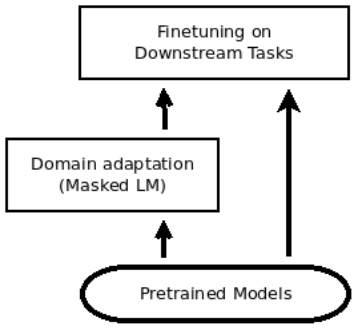

Figure 2: Experimental setup: we compared direct finetuning of the pretrained models on the downstream tasks vs. domain adapting the pretrained models and their subsequent finetuning.

of classes. Input to this head was the mean pooling over the hidden states of the last encoder layer to which we applied a dropout with probability 0.1.

Model finetuning was conducted in four epochs, by training on the training data and validating on the development data with a learning rate of $3e-5$, the AdamW optimizer with a weighted Cross Entropy loss, a batch size of 16 and without freezing the encoder layers. After the four epochs were completed we selected the model that performed best on the development set out of the four, in terms of Unweighted Average Recall (UAR), i.e. the mean value of the class-wise true positive rate; we subsequently evaluated this model on the held-out test data for the respective downstream task. Each such finetuning pass was repeated five times with different random seeds for each downstream task. Via the weighted loss for class balancing as well as via the UAR metric we aimed to address the imbalanced class distributions in our data.

4 Results

Table 2 reports for each BERT model the performace in terms of unweighted average F-score (UAF), in particular its mean and standard deviation over the five random seeds. F-score is the standard evaluation metric in NLP for classification tasks, and UAF expresses the class-wise averaged harmonic mean of precision and recall. We observed that the ranking of the models is similar regardless of expressing the performance scores in terms of UAF or UAR metric, i.e. the trend stays the same: domain-adapted models outperform their underlying pretrained model, i.e. the baseline. Domain adaptation (expressed by the *From-Adapted* column in the table) proved beneficial for all tested language models. If we compare these results with those obtained by the baseline models (expressed by the *From-Pretrained* column), we see that all models profited from domain adaptation roughly to the same extent. The overall low standard deviation values indicate that the findings are independent of the seed and thus robust.

BERT reached top performance on the three attribution tasks that are complex and thus time-consuming for human Slavist experts. The universal model *bert-base-multilingual-uncased* yielded very high performance and in two out of three tasks the best results. It was not outperformed by the two other models that had been created specifically for Cyrillic texts. The universal model is likely highly competitive due to drawbacks of the two Cyrillic models: the uncased *bert-base-slavic-cyrillic-upos* model was trained for token classification (part-of-speech tagging), so it performed suboptimal on our downstream tasks which need to operate on the basis of sequence classification; *bert-base-macedonian-bulgarian-cased* is based on a cased tokenizer, but casing is not consistent in our historical datasets.

4.1 Analysis of misclassifications

We assessed the classification output qualitatively, manually inspecting misclassifications made by *bert-base-multilingual-uncased*. In terms of attributing *region*, we saw that text snippets from East Slavic datasets got misclassified as South Slavic when they contained a token – e.g. **вънезапоу** 'suddenly' – that already occurs in Old Church Slavic manuscripts dated to the 11th century, i.e. is of South Slavic origin, cf. Kurz (1958). Yet, what from a technical perspective is a misclassification, can have a significant value from the philological point of view: it might indicate – and in this particular case it indeed does – that a text snippet in a manuscript handed down in an East Slavic context has its roots in the South Slavic region. This is not surprising, given that the majority of Slavic religious texts were translations from Greek made on South Slavic soil and copied later in other regions.

In turn, a text snippet containing the token **тажє** ('the same') was misclassified into the East Slavic region, but this word is indeed seen in both East Slavic and South Slavic texts, even in the earliest manuscripts, cf. Kurz (1958), despite its diachronical variation between East and South Slavic. During linguistic-historical development, the Proto-Indo-European cluster **dj* changed its phonetic

Task	Model	From-Pretrained	From-Adapted
manuscript	KoichiYasuoka/bert-base-slavic-cyrillic-upos	0.922 (0.004)	0.941 (0.003)
manuscript	anon-submission/mk-bert-base-macedonian-bulgarian-cased	0.935 (0.002)	0.961 (0.001)
manuscript	bert-base-multilingual-uncased	0.945 (0.002)	**0.962** (0.003)
century	KoichiYasuoka/bert-base-slavic-cyrillic-upos	0.952 (0.002)	0.965 (0.001)
century	anon-submission/mk-bert-base-macedonian-bulgarian-cased	0.961 (0.001)	**0.977** (0.002)
century	bert-base-multilingual-uncased	0.959 (0.001)	0.976 (0.001)
region	KoichiYasuoka/bert-base-slavic-cyrillic-upos	0.96 (0.002)	0.976 (0.001)
region	anon-submission/mk-bert-base-macedonian-bulgarian-cased	0.968 (0.001)	0.984 (0.001)
region	bert-base-multilingual-uncased	0.979 (0.002)	**0.986** (0.001)

Table 2: Performance scores on the three downstream tasks on directly finetuned models (*From-Pretrained*) that we regard as baseline vs. domain-adapted and subsequently finetuned models (*From-Adapted*), in terms of Unweighted Average F-score arithmetic mean values and standard deviations (in brackets) obtained from five random seeds.

form, in East Slavic languages developing into the simple consonant ж [ʒ] – a voiced post-alveolar fricative as in viSion –, cf. Trunte (2001), p. 186, while in South Slavic languages it remained with the cluster, realized as жд [ʒd] so that in South Slavic manuscripts one encounters the form тажде but the form таже is similarly common there.

Regarding chronological variation, 15th–16th and 18th century data misclassified as 10th–12th century contained phrasings (e.g. того ради и рече 'and it is that for/for this reason that he says'), which with regard to grammar and lexicon may actually be traced back to the 11th century. However, this specific string occurs with high-frequency and appears in numerous copied Church Slavic texts, and thereby has less profound interpretive implications. Concerning 11th century snippets misclassified as 15th-16th or 18th century material, we can exemplify the token пристоупиша ('they approached') that occurs in a snippet from a text translated in ca. 9th–10th cc., handed down in a manuscript hitherto dated to the last quarter of the 11th century and located in the Kyivan Principality. Since the orthography of the ending -а in the given grammatical form (*3PlAorIndAkt*) is more common in younger East Slavic manuscripts – the orthographical variant that had been in use in Old Church Slavic manuscripts was the *little yus'* grapheme ѧ, cf. Trunte (2001), p. 185 –, its attribution as 15th–16th century is comprehensible, but since this spelling was not unusual for manuscripts of the 11th–12th cc. either, the dating to the 15th-16th cc. cannot be postulated on the basis of this form.

Yet another example for variation involves the writing of the reflexive postfix -сѧ that can stand either directly adjacent to the word form or can be separated from it by a space; this variation however depends on modern editorial principles rather than on scribal usus, given the medieval *scriptio continua* practice. In particular, while adjacency is used in the contemporary edition of the 16th century *Apostolos* (ed. Besters-Dilger (2014)) as well as in the 18th century printed Elizabeth Bible (ed. 1751), likely influenced by its modern Russian (i.e. Eastern Slavic) continuation, we see that spacing is used in the modern edition of the *Codex Suprasliensis*, in line with typographical separation from the verbal stem in modern South Slavic languages. This orthographic discrepancy certainly implies some bias, implying that BERT's classification strategy is getting influenced by contemporary editorial principles represented in parts of the data.

4.2 Conclusions and future work

Our current pilot study set out to investigate the extent to which BERT can be used for provenance attribution on Pre-Modern Slavic manuscript data in terms of three coarse-grained text classification tasks that characterise temporal-spatial dimensions of historical, mainly liturgical and religious, language data. The aggregated dataset we employed in this study contains three axes of variation – time, region, manuscript –, allowing to perform analyses for identifying patterns between multiple variables that can play a role in language change. We experimented with domain adaptation of pretrained BERT models and reached overall high performance on the downstream text classification tasks. The results provide plausible insights into how BERT makes use of the data, even though we are aware that our initial approach bears limitations for comprehensive linguistic analyses: we showed examples that shed light on why temporal and regional variation in the texts lead to errors in the classification. For further studies on language change, we aim to make the trained models classify finer-grained phenomena and the labeled data more representative and then release these resources.

5 Limitations

Our current goal was to investigate the extent to which a generic BERT approach on the level of text snippets would be able to utilize data characteristics that encode in a heterogeneous way the provenance characteristics we are after. Such an approach is deliberately coarse-grained and is likely to be predominantly semantically-oriented. Our downstream tasks had classes that we were directly able to generate from the manuscript level. Since we lack ground truth provenance labels attributed on sub-manuscript level, we were aware that the current experimental setup would be suboptimal for acquiring results that would be describing linguistic specificities pointing out phonological, morphological, etc. features of linguistic change.

It is indeed the goal of our project to generate such expert labels in a data-driven way; for example, our task setup is getting extended to the token and to the character levels. We are also working on better token segmentation and expansion of the data in order to minimise potential manuscript biases in terms of orthography and content.

Ethics Statement

The authors fully acknowledge the ACL Ethics Policy and strongly commit to using their skills for the benefit of society, its members and the environment surrounding them.

Acknowledgements

The QuantiSlav project is funded from the EU's Recovery and Resilience Facility and by the Federal Ministry of Education and Research in accordance with the guidelines for funding projects to strengthen the data skills of young scientists (Grant number: 16DKWN123B).

References

Yannis Assael, Thea Sommerschield, and Jonathan Prag. 2019. Restoring ancient text using deep learning: a case study on Greek epigraphy. 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)*, pages 6368–6375, Hong Kong, China. Association for Computational Linguistics.

David Bamman and Patrick J. Burns. 2020. Latin BERT: A contextual language model for classical philology.

Gabriel Bernier-Colborne, Cyril Goutte, and Serge Léger. 2019. Improving cuneiform language identification with BERT. In *Proceedings of the Sixth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 17–25, Ann Arbor, Michigan. Association for Computational Linguistics.

Juliane Besters-Dilger. 2014. *Kommentierter Apostolos. Volume 1.* Otto Sagner, Freiburg i. Br.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Simon Gabay, Pedro Ortiz Suarez, Alexandre Bartz, Alix Chagué, Rachel Bawden, Philippe Gambette, and Benoît Sagot. 2022. From FreEM to D'AlemBERT: a Large Corpus and a Language Model for Early Modern French.

Md. Zahurul Islam, Roland Mittmann, and Alexander Mehler. 2011. Multilingualism in ancient texts: Language detection by example of old high german and old saxon. In *GSCL conference on Multilingual Resources and Multilingual Applications*.

Joseph Kurz. 1958. *Slovník jazyka staroslovenského: Lexikon linguae palaeoslovenicae*. Nakladatelství Československé Akademie ved, Prague.

Andrey Kutuzov and Lidia Pivovarova. 2021. RuShiftEval: a shared task on semantic shift detection for Russian. In *Computational linguistics and intellectual technologies: Papers from the annual conference Dialogue*.

Piroska Lendvai and Claudia Wick. 2022. Finetuning Latin BERT for Word Sense Disambiguation on the Thesaurus Linguae Latinae. In *Proc. of the Workshop Cognitive Aspects of the Lexicon (CogALex), co-located with Asia-Pacific Chapter of the Association for Computational Linguistics (AACL) and 10th International Joint Conference on Natural Language Processing (IJCNLP)*.

Chaya Liebeskind and Shmuel Liebeskind. 2020. Deep learning for period classification of historical hebrew texts. *Journal of Data Mining & Digital Humanities*, 2020.

E. M. A. Manjavacas. 2022. Adapting vs. pre-training language models for historical languages. *Journal Of Data Mining & Digital Humanities, NLP4DH*, pages 1–19.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Iiro Rastas, Yann Ciarán Ryan, Iiro Tiihonen, Mohammadreza Qaraei, Liina Repo, Rohit Babbar, Eetu Mäkelä, Mikko Tolonen, and Filip Ginter. 2022. Explainable publication year prediction of eighteenth century texts with the BERT model. In *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change*, pages 68–77, Dublin, Ireland. Association for Computational Linguistics.

Stefan Schweter, Luisa März, Katharina Schmid, and Erion Çano. 2022. hmBERT: Historical multilingual language models for named entity recognition.

Ilya Talev. 1973. *Some Problems of the Second South Slavic Influence in Russia*. Otto Sagner.

Nicolina Trunte. 2001. *Slavenskij jazyk: ein praktisches Lehrbuch des Kirchenslavischen in 30 Lektionen; zugleich eine Einführung in die slavische Philologie. Bd. 2, Mittel- und Neukirchenslavisch*. Slavistische Beiträge (494), Munich.

Nianheng Wu, Eric DeMattos, Kwok Him So, Pin-zhen Chen, and Çağrı Çöltekin. 2019. Language discrimination and transfer learning for similar languages: Experiments with feature combinations and adaptation. In *Proceedings of the Sixth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 54–63, Ann Arbor, Michigan. Association for Computational Linguistics.

Marcos Zampieri, Shervin Malmasi, Yves Scherrer, Tanja Samardžić, Francis Tyers, Miikka Silfverberg, Natalia Klyueva, Tung-Le Pan, Chu-Ren Huang, Radu Tudor Ionescu, Andrei M. Butnaru, and Tommi Jauhiainen. 2019. A report on the third VarDial evaluation campaign. In *Proceedings of the Sixth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 1–16, Ann Arbor, Michigan. Association for Computational Linguistics.