27th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES 2023)

FreCDo: A Large Corpus for French Cross-Domain Dialect Identification

Mihaela Găman[a], Adrian-Gabriel Chifu[b], William Domingues[b], Radu Tudor Ionescu[a,*]

[a]*University of Bucharest, Romania*
[b]*Aix-Marseille Université, France*

Abstract

We present a novel corpus for French dialect identification comprising 413,522 French text samples collected from public news websites in Belgium, Canada, France and Switzerland. To ensure an accurate estimation of the dialect identification performance of models, we designed the corpus to eliminate potential biases related to topic, writing style, and publication source. More precisely, the training, validation and test splits are collected from different news websites, while searching for different keywords (topics). This leads to a **Fre**nch **c**ross-**do**main (FreCDo) dialect identification task. We conduct experiments with four competitive baselines, a fine-tuned CamemBERT model, an XGBoost based on fine-tuned CamemBERT features, a Support Vector Machines (SVM) classifier based on fine-tuned CamemBERT features, and an SVM based on word n-grams. Aside from presenting quantitative results, we also make an analysis of the most discriminative features learned by CamemBERT. Our corpus is available in open-source format at https://github.com/MihaelaGaman/FreCDo.

Keywords: dialect identification; French corpus; cross-domain evaluation.

1. Introduction

Dialect identification is an actively studied topic in computational linguistics, one of the most significant drivers for research in this area being the yearly shared tasks on dialect identification organized at VarDial [4, 7, 11, 15, 16, 17]. Identifying dialects is important for analyzing and understanding the evolution of languages across time and geographical regions. The task is also useful in forensics analysis, where the dialect of a text sample might reveal some information about the author of the respective text. Despite the valuable applications, French dialect identification is not extensively studied in literature [2, 8, 13]. Since French is spoken across multiple countries across the globe, we believe that French dialect identification should be given more attention. To this end, we present a novel and large

---

* Corresponding author.
 E-mail address: raducu.ionescu@gmail.com

dialect identification corpus composed of over 400K French text samples representing four dialects: Belgian French, Canadian French, Hexagonal French and Swiss French.

Due to the high number of data samples, our corpus is suitable for the training and evaluation of modern deep language models, such as CamemBERT [12]. However, machine learning models are prone to overfitting the biases of the training data. To ensure a fair estimation of the generalization capacity of such models, we designed the corpus to mitigate potential biases related to topic, writing style, and publication source, which could lead to falsely increasing the accuracy of the evaluated models. This is achieved by separating the text samples into training, validation and test sets, such that the examples in each of the three splits are from different news websites than the other two splits. Moreover, the data samples in each split are collected from the results retrieved while searching for a set of keywords (topics) that are distinct from the other splits. Our data collection process leads to a **Fre**nch **c**ross-**do**main (FreCDo) dialect identification task.

We conduct experiments with four baselines, namely a fine-tuned CamemBERT model, an XGBoost [5] based on fine-tuned CamemBERT features, a Support Vector Machines (SVM) [6] classifier based on fine-tuned CamemBERT features, and an SVM based on word n-grams. Although our baselines are fairly competitive, the empirical results indicate that the French cross-domain dialect identification task is very challenging, the highest macro F_1 score being 39.67%. To determine if the top scoring model, the standalone CamemBERT, learned training set biases, we analyze the most discriminative features of the respective model.

In summary, our contribution is threefold:

- We introduce a novel large-scale corpus for French cross-domain dialect identification.
- We train and evaluate four competitive baselines, three of them being based on a modern deep language model which uses a transformer architecture, i.e. CamemBERT.
- We analyze the most discriminative features of CamemBERT, revealing the dialectal patterns learned by this state-of-the-art model.

2. Related Work

One of the most important research efforts in the area of dialect identification is represented by the DSL Corpus Collection (DLSCC)[1] [13], associated with the VarDial workshop. The corpus is at its fifth iteration and it contains short excerpts of journalistic texts. The latest version (4.0) contains six groups of languages with their corresponding dialects. For instance, the Hexagonal French and Canadian French represent one of the groups. Our data set contains two more French dialects (Swiss and Belgian) and is considerably larger (by an order of magnitude).

CFPR[2] is a regional spoken French corpus containing voice recordings with their transcriptions. The corpus contains 388,123 words (30 hours of recordings) produced by 172 speakers from around the world. Most of the speakers (130) are from Europe, out of which 17 are from Belgium and 6 from Switzerland. There are 5 speakers from Canada, 17 from Northern Africa and 2 from Lebanon. Even though the corpus contains more dialects than our corpus, we observe that the CFPR corpus is very unbalanced. Moreover, it does not take into account the concept of topics, i.e. samples can also be discriminated based on the discussed topics.

Another considerable French resource is the French Web Corpus (frTenTen)[3], a part of the TenTen corpus family [8]. The latest iteration of this French corpus (frTenTen17, from 2017) contains 5.7 billion words and it was annotated by FreeLing (lemmatization, part-of-speech tagging, etc.). This corpus is not designed specifically for dialect identification, but the dialects could be inferred from the metadata containing the web domain. Of course, this requires some processing efforts. As the CFPR corpus, frTenTen is not focused on cross-topic or cross-source analysis either. There are other researchers focusing on French variants or dialects, but with a different scope. For instance, Blondeau et al. [2] have studied the Montreal French intrinsic evolution.

---

[1] http://ttg.uni-saarland.de/resources/DSLCC/
[2] https://cfpr.huma-num.fr
[3] https://www.sketchengine.eu/frtenten-french-corpus/#toggle-id-3

Split	Topics: FR (EN)	Country	Source	#Samples	#Tokens
Train	Guerre (War), Ukraine (Ukraine)	BE	dhnet	121,746	11,619,874
		CA	radio-canada	34,003	2,505,254
		CH	letemps	141,261	12,719,203
		FR	cnews	61,777	6,397,943
			Total:	358,787	33,242,274
Validation	Russie (Russia) États-Unis (United States)	BE	ln24	7,723	824,871
		CA	journal-metro	171	17,061
		CH	le-courrier	5,244	476,338
		FR	huffpost	4,864	434,547
			Total:	18,002	1,752,817
Test	Réchauffement climatique (Global warming) Covid (Covid)	BE	lavenir	15,235	1,227,263
		CA	science-presse	944	86,724
		CH	24heures	9,824	910,700
		FR	franceinfo	10,730	848,845
			Total:	36,733	3,073,532

Table 1. Our large-scale corpus is composed of more than 400K data samples, containing a total of more than 38M tokens. The corpus is divided into training, validation and test sets, such that the overlap between topics and publication sources across its subsets is reduced as much as possible.
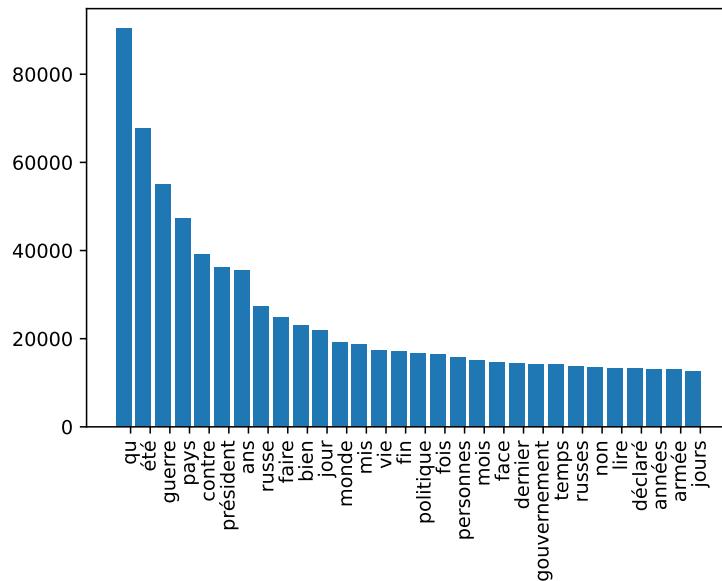

Fig. 1. A histogram of the most frequent 30 words from the training set, after eliminating stop words with spaCy. Best viewed in color.

3. Corpus

The texts comprising the FreCDo corpus have been extracted from news websites from four countries: Belgium (BE), Canada (CA), Switzerland (CH) and France (FR). Since the texts are extracted from written news articles, the only connection to spoken French is coming from manual transcriptions of quotes or interviews included in the respective articles. Dialect-wise, written texts (including literary transcriptions of spoken language) are harder to identify than spoken utterances, since professional writing is more formal (standardized).

We have considered several general, country independent topics in order to be able to obtain a large volume of data, while also mitigating possible dialect-specific biases (since generic keywords are to be found in all the sources), and we have queried the news websites with them. The topics are: Guerre (War), Ukraine (Ukraine), Russie (Russia),
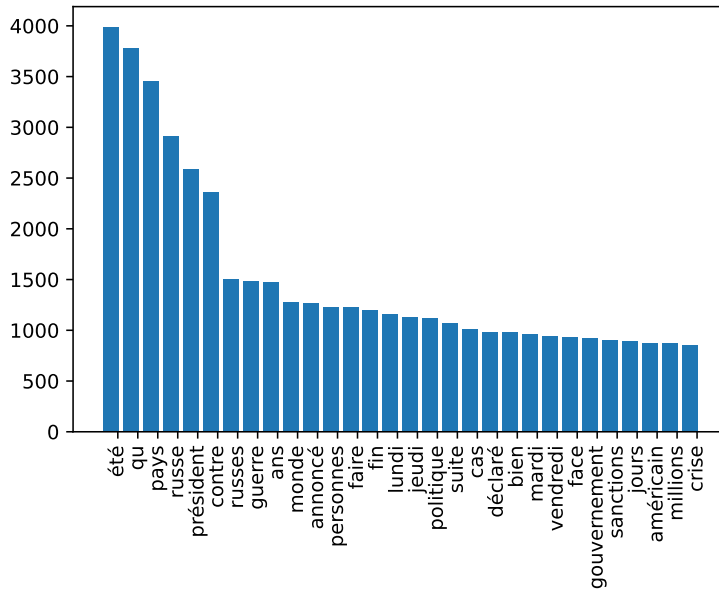
Fig. 2. A histogram of the most frequent 30 words from the validation set, after eliminating stop words with spaCy. Best viewed in color.
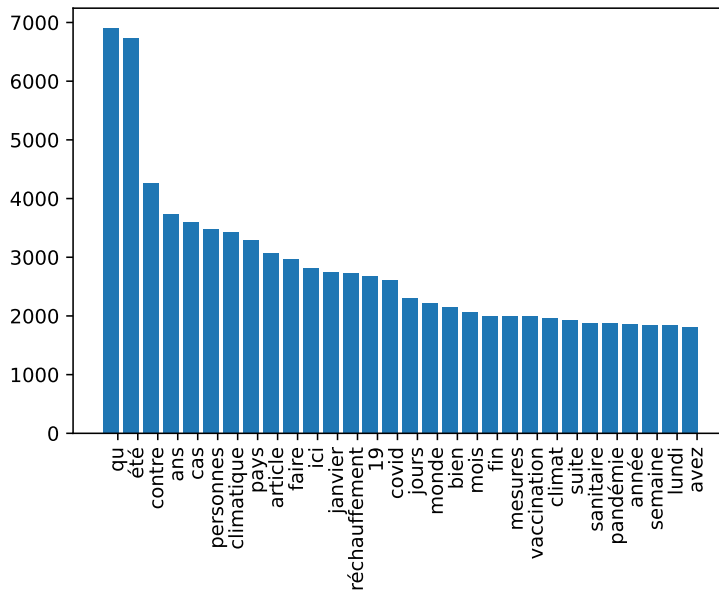

Fig. 3. A histogram of the most frequent 30 words from the test set, after eliminating stop words with spaCy. Best viewed in color.

États-Unis (United States), Réchauffement climatique (Global warming) and Covid (Covid). The timeframe for data collection is between November 2021 and March 2022, but the articles can be older than the start date. We queried the listed websites (see Table 1) with the aforementioned keywords and we have crawled the returned pages. For some queries, we obtained a large number of news articles from certain websites. In such cases, we used undersampling to preserve a relative class balance of the splits, randomly selecting articles from the very large retrieved lists, while keeping all returned articles from the other lists.

From each extracted article, we kept the title, the headline and the body of the article. We employed spaCy[4] to identify named entities and replace them with the special token NE, in order to avoid any topic or country specific biases. We next split the texts by paragraphs. A paragraph is defined by a group of three consecutive sentences (the sentences being delimited by punctuation). One paragraph represents a data sample (instance) in the data set. Following previous work in building corpora for dialect identification [3], the class labels are based on the website domains. Hence, human intervention was not used in the labeling process. The considered topics and their associated publication sources are summarized in Table 1. We underline that the overlap between topics and sources across the training and test splits is close to marginal. This is also confirmed by analyzing the most frequent words in each of the three subsets. Figure 1 indicates that topical words such "guerre" and "pays" are very frequent in the training set. By comparing Figures 1 and 2 , we notice some degree of overlap between the most frequent topical words in the training and validation sets. Examples include words such "guerre" and "russe". However, Figure 3 shows that there are no topical word overlaps between the training and test sets. Hence, models learning topical cues, author style or publication source patterns will not benefit from such biases when the evaluation is performed on the designated test set.

Out of the 413,522 samples, our data split leaves the training set with 358,787 samples (86.8%), and the validation and test sets with 18,002 (4.3%) and 36,733 (8.9%) samples, respectively. We notice that the CA dialect is significantly underrepresented in each of the three splits. We did not find sufficient public news articles from Canadian sources (most of the news articles in Canada are only accessible through subscription-based accounts), which explains the small number of samples that were collected from this country.

4. Experiments

4.1. Baseline Models

We publicly release the code to reproduce all baseline models along with the data set in our GitHub repository.

Fine-tuned CamemBERT. The first baseline is a fine-tuned CamemBERT model [12], which is based on the RoBERTa architecture [9]. In the original paper, Martin et al. [12] reported state-of-the-art results on a variety of French downstream tasks, hence our preference for using this architecture as baseline.

Our fine-tuning procedure requires us to encode each textual input into a list of token IDs. For this purpose, we use the CamemBERT tokenizer. Each token is translated into its corresponding 768-dimensional embedding vector. We then add a global average pooling layer to the current architecture, obtaining a Continuous Bag-of-Words (CBOW) representation for each text sequence. Then, we add a Softmax classification layer with four neural units, corresponding to each of the four French dialects targeted in this work, namely Belgium (BE), Canadian (CA), Swiss (CH) and French (FR). Given a text input, the final dialect label is obtained by applying *argmax* on the four probabilities. We fine-tune the model described above for 11 epochs (based on early stopping) on mini-batches of 32 samples, using the Adam with decoupled weight decay (AdamW) optimizer [10], with a learning rate of $5 \cdot 10^{-5}$ and an ϵ of 10^{-8}. We tuned the learning rate between 10^{-5} and 10^{-4} and tested two loss options, cross-entropy vs. negative log-likelihood. All models are trained on a GeForce GTX 1080Ti GPU.

XGBoost + fine-tuned CamemBERT features. XGBoost [5] is a parallel gradient boosting approach based on iteratively combining decision trees into a single model. We apply XGBoost over fine-tuned CamemBERT embeddings. Each text sample fed into XGBoost is represented by the *[CLS]* token returned by CamemBERT, which is a 768-dimensional embedding vector. For XGBoost, we use 400 estimators, with a maximum depth of a tree equal to 200, a learning rate of $7 \cdot 10^{-2}$, and $\gamma = 1$. These hyperparameters are obtained using grid search considering the following intervals: maximum depth between 10 and 400, number of estimators between 100 and 400, learning rate between 10^{-2} and $2 \cdot 10^{-1}$, and γ between 0 and 3. We set the L_2 and L_1 regularization weights to $\lambda = 0.1$ and $\alpha = 0.1$, without tuning.

---

[4] https://spacy.io (v3.2.2, "fr_core_news_sm" model)

Method	Validation		Test	
	Acc.	Macro F1	Acc.	Macro F1
Majority class	0.4290	0.1501	0.4147	0.1466
CamemBERT	**0.7352**	**0.4784**	**0.5584**	**0.3967**
XGBoost + emb	0.7280	0.4518	0.4909	0.3477
SVM + emb	0.6633	0.4534	0.4705	0.3439
SVM + n-grams	0.6560	0.4487	0.4514	0.3108

Table 2. Performance metrics obtained with each of the four baselines versus the majority class baseline, for both validation and test sets.
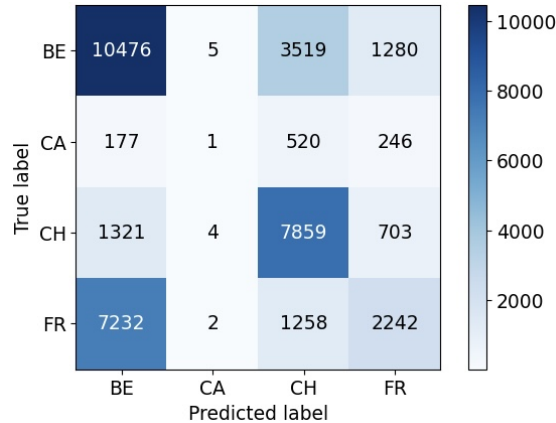

Fig. 4. Confusion matrix obtained after running the fine-tuned CamemBERT classifier on the test set. Best viewed in color.

SVM + fine-tuned CamemBERT features. We use a Support Vector Machines (SVM) [6] classifier as our third baseline, and apply it on 768-dimensional fine-tuned CamemBERT embeddings. In our experiments, we opt for the linear kernel and the one-vs-rest multi-class classification strategy. We tune the regularization penalty C, in a range of values from 10^{-4} to 10^4 using grid search. We obtain optimal results with $C = 10$.

SVM + word n-grams. As our fourth baseline, we include an SVM based on word n-grams, a conventional approach commonly used in dialect identification [1, 14]. We tune the n-gram length in the range 1-3, obtaining the best results with word bigrams. As for the third baseline, we also tune the penalty C of the SVM, considering values in the range 10^{-4} to 10^4. The optimal value for C is 1.

4.2. Results

The results reported in Table 2 show that the fine-tuned CamemBERT obtains both the best micro (accuracy) and macro F1 scores in the four-class French dialect classification task, on both validation and test sets from our corpus. The XGBoost model based on CamemBERT embeddings takes the second place, not too far from the best performing standalone CamemBERT. The SVM based on the fine-tuned embeddings falls further back compared to the other two methods based on deep CamemBERT features, both on the validation, as well as on the test set. The SVM based on word bigrams attains the lowest performance, confirming that the deep CamemBERT features are very useful. We underline that all our baselines obtain better results than the majority (dominant) class baseline. However, the task proves to be significantly challenging, leaving plenty of room for future improvements.

From what we can observe in Figure 4, our fine-tuned CamemBERT attains more accurate results for the Belgium and Swiss French dialects for which, in their majority, the samples tend to be correctly classified. Less represented in the training set, Canadian French registers the worst results in both evaluation subsets.

4.3. Discriminative Feature Analysis

For the discriminative feature analysis, we have considered a few correctly classified samples and we have analyzed the features for which CamemBERT has given high scores. From this manual analysis, we made a few observations. We noticed that a lot of short words and stop words have high CamemBERT scores. Regarding the discriminative terms, for one CH sample, there is the term "paraphe" (signature) occurring in the context of a petition signature. In Hexagonal French, one would say "signature" (signature) in the aforementioned context. For BE and CH samples, we observed the presence of numerals, such as "septante" (70) and "nonante" (90), as good indicators for the dialect. In Hexagonal French, the same numerals are based on the addition of 60 and 10 ("soixante-dix") to get 70, or the multiplication of 4 and 20 followed by an addition with 10 ("quatre-vingt-dix") to get 90. For CH samples, we also noticed the term "francs", which is the Swiss currency. This is quite specific for this country, since France and Belgium have the Euro, and Canada has the Canadian dollar. In conclusion, there are quite a few noticeable dialectal patterns learned by the model.

5. Conclusion

In this paper, we presented a novel corpus for cross-domain French dialect identification. To the best of our knowledge, FreCDo is the largest corpus for French dialect identification. We introduced three competitive baselines based on a state-of-the-art language model, CamemBERT, but the reported results showed that the proposed task is very challenging. We believe that a promising future direction for improving performance is to develop an effective approach to avoid overfitting training set biases.

Regarding the number of considered dialects, adding more dialects, e.g. from Northern Africa (e.g. Algeria) or Lebanon, would broaden our benchmark in the future. Unfortunately, when we started our data collection process, we did not find sufficient French news samples from other French speaking countries than those already included. Thus, perhaps the most important future direction is to extend our data set with samples representing French dialects from other regions, e.g. Northern Africa.

Ethics Statement

Our data processing steps ensure anonymity, since the named entities have been replaced with a special token. Hence, names of people have been removed from all samples. Moreover, our data collection process respects the intellectual property of the news articles, since the data samples are formed of three sentence excerpts (paragraphs) from news articles. The data samples are shuffled, so the original news articles cannot be reconstructed. Furthermore, we adhere to the European regulations[5] that allow researchers to use data in the public web domain for non-commercial research purposes. We thus release our corpus as open-source under a non-commercial share-alike license agreement.

We acknowledge that spaCy could have missed the names of some people. We will remove samples containing personal details about people, should we receive any specific complaints on this regard via e-mail.

References

[1] Benites, F., von Däniken, P., Cieliebak, M., 2019. TwistBytes - Identification of Cuneiform Languages and German Dialects at VarDial 2019, in: Proceedings of the Sixth Workshop on NLP for Similar Languages, Varieties and Dialects, pp. 194–201. URL: https://aclanthology.org/W19-1421, doi:10.18653/v1/W19-1421.

[2] Blondeau, H., Tremblay, M., Bertrand, A., Michel, E., 2021. A new milestone for the study of variation in montréal french: The hochelaga-maisonneuve sociolinguistic survey. Corpus doi:https://doi.org/10.4000/corpus.6221.

[3] Butnaru, A., Ionescu, R.T., 2019. MOROCO: The Moldavian and Romanian Dialectal Corpus, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 688–698. URL: https://aclanthology.org/P19-1068, doi:10.18653/v1/P19-1068.

---

[5] https://eur-lex.europa.eu/eli/dir/2019/790/oj

[4] Chakravarthi, B.R., Găman, M., Ionescu, R.T., Jauhiainen, H., Jauhiainen, T., Lindén, K., Ljubešić, N., Partanen, N., Priyadharahsini, R., Purschke, C., Rajagopal, E., Scherrer, Y., Zampieri, M., 2021. Findings of the VarDial Evaluation Campaign, in: Proceedings of the Eighth Workshop on NLP for Similar Languages, Varieties and Dialects, pp. 1–11. URL: https://aclanthology.org/2021.vardial-1.1.

[5] Chen, T., Guestrin, C., 2016. XGBoost: A scalable tree boosting system, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794. URL: https://doi.org/10.1145/2939672.2939785, doi:10.1145/2939672.2939785.

[6] Cortes, C., Vapnik, V., 1995. Support-vector networks. Machine Learning 20, 273–297. doi:10.1007/BF00994018.

[7] Găman, M., Hovy, D., Ionescu, R.T., Jauhiainen, H., Jauhiainen, T., Lindén, K., Ljubešić, N., Partanen, N., Purschke, C., Scherrer, Y., Zampieri, M., 2020. A Report on the VarDial Evaluation Campaign 2020, in: Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects, pp. 1–14. URL: https://aclanthology.org/2020.vardial-1.1.

[8] Jakubíček, M., Kilgarriff, A., Kovář, V., Rychlỳ, P., Suchomel, V., 2013. The TenTen corpus family, in: Proceedings of the 7th International Corpus Linguistics Conference CL, pp. 125–127. URL: https://www.sketchengine.eu/wp-content/uploads/The_TenTen_Corpus_2013.pdf.

[9] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V., 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv preprint arXiv:1907.11692 URL: https://arxiv.org/abs/1907.11692.

[10] Loshchilov, I., Hutter, F., 2019. Decoupled Weight Decay Regularization, in: Proceedings of the International Conference on Learning Representations. URL: https://openreview.net/pdf?id=Bkg6RiCqY7.

[11] Malmasi, S., Zampieri, M., Ljubešić, N., Nakov, P., Ali, A., Tiedemann, J., 2016. Discriminating between Similar Languages and Arabic Dialect Identification: A Report on the Third DSL Shared Task, in: Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects, pp. 1–14. URL: https://aclanthology.org/W16-4801.

[12] Martin, L., Muller, B., Ortiz Suárez, P.J., Dupont, Y., Romary, L., de la Clergerie, É., Seddah, D., Sagot, B., 2020. CamemBERT: a Tasty French Language Model, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 7203–7219. URL: https://aclanthology.org/2020.acl-main.645, doi:10.18653/v1/2020.acl-main.645.

[13] Tan, L., Zampieri, M., Ljubešic, N., Tiedemann, J., 2014. Merging Comparable Data Sources for the Discrimination of Similar Languages: The DSL Corpus Collection, in: Proceedings of the 7th Workshop on Building and Using Comparable Corpora, pp. 11–15. URL: https://aclanthology.org/www.mt-archive.info/10/BUCC-2014-Tan.pdf.

[14] Wu, N., DeMattos, E., So, K.H., Chen, P.z., Çöltekin, Ç., 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, pp. 54–63. URL: https://aclanthology.org/W19-1406, doi:10.18653/v1/W19-1406.

[15] Zampieri, M., Malmasi, S., Ljubešić, N., Nakov, P., Ali, A., Tiedemann, J., Scherrer, Y., Aepli, N., 2017. Findings of the VarDial Evaluation Campaign 2017, in: Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects, pp. 1–15. URL: https://aclanthology.org/W17-1201, doi:10.18653/v1/W17-1201.

[16] Zampieri, M., Malmasi, S., Nakov, P., Ali, A., Shon, S., Glass, J., Scherrer, Y., Samardžić, T., Ljubešić, N., Tiedemann, J., van der Lee, C., Grondelaers, S., Oostdijk, N., van den Bosch, A., Kumar, R., Lahiri, B., Jain, M., 2018. Language Identification and Morphosyntactic Tagging: The Second VarDial Evaluation Campaign, in: Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects, pp. 1–17. URL: https://aclanthology.org/W18-3901.

[17] Zampieri, M., Malmasi, S., Scherrer, Y., Samardžić, T., Tyers, F., Silfverberg, M., Klyueva, N., Pan, T.L., Huang, C.R., Ionescu, R.T., Butnaru, A.M., Jauhiainen, T., 2019. A Report on the Third VarDial Evaluation Campaign, in: Proceedings of the Sixth Workshop on NLP for Similar Languages, Varieties and Dialects, pp. 1–16. URL: https://aclanthology.org/W19-1401, doi:10.18653/v1/W19-1401.