Assessing Comparability of Genre Datasets via Cross-Lingual and Cross-Dataset Experiments

Taja Kuzman[†] [*], **Nikola Ljubešić**[†], **Senja Pollak**[†]

[†]Department of Knowledge Technologies, Jožef Stefan Institute
taja.kuzman@ijs.si, nikola.ljubesic@ijs.si, senja.pollak@ijs.si

[*]Jožef Stefan International Postgraduate School

Abstract

This article explores comparability of an English and a Slovene genre-annotated dataset via monolingual and cross-lingual experiments, performed with two Transformer models. In addition, we analyze whether translating the Slovene dataset into English with a machine translation system improves monolingual and cross-lingual performance. Results show that cross-lingual transfer is possible despite the differences between the datasets in terms of genre schemata and corpora construction methods. Furthermore, the XLM-RoBERTa model was shown to provide good results in both settings already when learning on less than 1,000 instances. In contrast, the trilingual CroSloEngual BERT model was revealed to be less suitable for this text classification task. Moreover, the results reveal that although the English dataset is 40 times larger than the Slovene dataset, it provides similar or worse classification results.

1. Introduction

Texts in datasets can be grouped by genres based on their common function, form and the author's purpose (Orlikowski and Yates, 1994). Labeling texts with genres allows for a deeper insight into the composition and quality of a web corpus that was collected with automatic means, more efficient queries in information retrieval tools (Vidulin et al., 2007), as well as improvements of various language technologies tasks, such as part-of-speech tagging (Giesbrecht and Evert, 2009) and machine translation (Van der Wees et al., 2018). That is why automatic genre identification (AGI) has been a subject of numerous studies in the computational linguistics and information retrieval fields (e.g., see Egbert et al. (2015), Sharoff (2018)).

As in other text classification tasks, a large manually annotated dataset is required in AGI in order to train and test a classifier. While there exist some large English genre-annotated datasets, such as the Corpus of Online Registers of English (CORE) (Egbert et al., 2015) with 53,000 texts and the Leeds Web Genre Corpus (Asheghi et al., 2016) with 5,000 texts, for other languages there is either no dataset or mostly a small one, consisting of 1,000 to 2,000 texts, such as genre-annotated corpora for Russian (Sharoff, 2018), Finnish (Laippala et al., 2019), Swedish and French (Repo et al., 2021). This means that for obtaining a large dataset needed for genre identification of other languages, costly and time-consuming annotation campaigns are still needed, leaving most languages under-resourced in regard to the technologies based on the AGI.

However, it might be possible to overcome this obstacle by leveraging the cross-lingual transfer, applying models trained on high-resource languages to the low-resource languages. Recently, Repo et al. (2021) showed that it is possible to achieve good levels of cross-lingual transfer in AGI experiments. They performed experiments in zero-shot cross-lingual automatic genre identification by training multilingual Transformer-based models on the English CORE corpus (Egbert et al., 2015) and testing them on smaller Finnish, Swedish, and French datasets. Rönnqvist et al. (2021) extended this research, training the models on a multilingual dataset, created from the four corpora, which further improved the results.

These promising results stimulated creation of genre-annotated datasets for other languages, and for Slovene, a web genre identification corpus GINCO 1.0 (Kuzman et al., 2021) was created. Its genre schema was based on the CORE schema with the possibility of cross-lingual experiments in mind (see Kuzman et al. (2022)). However, a linguistic analysis of the categories (Biber and Egbert, 2018) and a low inter-annotator agreement, reported by Egbert et al. (2015) and Sharoff (2018), revealed some shortcomings of the CORE schema that could impact the reliability of the dataset. Thus, Kuzman et al. (2022) diverged from the original schema when annotating GINCO, striving towards a more reliably annotated dataset. In addition to this, the CORE and GINCO datasets were created following different corpora collection and annotation approaches (see Section 3.1.). Due to these differences, it remained unclear whether the datasets are comparable enough to allow cross-lingual transfer which would eliminate the need for extensive annotation campaigns of Slovene and other under-resourced languages of interest. This article provides first insight into this, exploring the comparability of the two datasets through cross-dataset and cross-lingual experiments.

2. Goal of the Paper

This paper analyzes comparability of two genre-annotated datasets, the Corpus of Online Registers of English (CORE) (Egbert et al., 2015) and the Slovene Web genre identification corpus GINCO 1.0 (Kuzman et al., 2021). We perform cross-dataset and cross-lingual automatic genre identification experiments to address the main research question (Q1): Is the CORE dataset comparable to the GINCO dataset enough to provide good cross-lingual transfer, as it was achieved by Repo et al. (2021) who

used comparably encoded Finnish, Swedish and French datasets?

To compare the corpora and to analyze their usefulness for monolingual as well as for cross-lingual automatic genre identification, first, labels from both corpora were mapped to a joint schema, the GINCORE schema. Then, multilingual pre-trained Transformer-based models were trained on the English CORE dataset with GINCORE labels (EN-GINCORE), the Slovene GINCO dataset with GINCORE labels (SL-GINCORE) and the SL-GINCORE dataset that was machine translated into English (MT-GINCORE). We conduct 1) monolingual in-dataset AGI experiments, training and testing on the same dataset, 2) cross-lingual and cross-dataset AGI experiments, training on one dataset and testing on the other. The machine-translated dataset is added to the comparison to explore two additional research questions: Q2) In monolingual in-dataset experiments, do multilingual models, which were pre-trained on more English than Slovene data, perform differently on Slovene dataset (SL-GINCORE) than on a Slovene dataset, machine-translated to English (MT-GINCORE)? and Q3) In cross-lingual cross-dataset experiments, does translating the training data (MT-GINCORE) into the language of test data (EN-GINCORE) provide better results than using training and testing data in different languages (SL-GINCORE and EN-GINCORE)?

The experiments were performed with two multilingual Transformer-based pre-trained language models, massively multilingual XLM-RoBERTa model (Conneau et al., 2020), and the trilingual Croatian-Slovene-English CroSloEngual BERT model (Ulčar and Robnik-Šikonja, 2020). This provides an answer to the fourth research question (Q4): Does CroSloEngual BERT, pre-trained on a smaller number of languages, perform better in the cross-lingual AGI experiments than a massively multilingual XLM-RoBERTa model?

3. Data Preparation

3.1. Original Datasets

In this research, three datasets were used: the Corpus of Online Registers of English (CORE) (Egbert et al., 2015), the Slovene Web genre identification corpus GINCO 1.0 (Kuzman et al., 2021) and the GINCO 1.0 corpus, machine translated to English.

The CORE corpus consists of web texts that were extracted from the "General" part of the Corpus of Global Web-based English (GloWbE) (Davies and Fuchs, 2015). The GloWbE corpus was collected via Google searches with high frequency English 3-grams as the queries (Davies and Fuchs, 2015). After obtaining the texts, further cleaning was performed, more specifically, the boilerplate was removed with the Justext tool (Pomikálek, 2011).

The CORE corpus was annotated based on a hierarchical schema which consists of 8 main genre categories, such as *Narrative*, *Opinion*, *Spoken*, and 54 subcategories, e.g., *News Report/Blog*, *Instruction*, *Travel Blog*, *Magazine Article*. The annotation was single-label, i.e., each annotator, recruited through a crowd-sourcing platform, could assign one main category and one subcategory to a text. However, as each text was annotated by four annotators, that means that it can have up to four labels. The corpus that we obtained from the authors and used in this research consists of 48,415 texts, labeled with 8 main categories and 47 subcategories. The corpus was further cleaned by removing duplicated texts and texts with more than one assigned label, resulting in 41,502 texts.

The GINCO corpus (Kuzman et al., 2022) consists of a random sample of web texts from two Slovene web corpora, slWaC 2.0 corpus (Erjavec and Ljubešić, 2014) from 2014 and MaCoCu-sl 1.0 corpus (Bañón et al., 2022) from 2021. Both web corpora were created by crawling the Slovene top-level domain and some generic domains that are interlinked with the national domain. As in GloWbE, the boilerplate was removed with the Justext tool (Pomikálek, 2011). The GINCO corpus consists of two parts, the "suitable" part, annotated with genres, and "not suitable" part, consisting of texts not suitable for genre annotation, such as texts in other languages, machine-translated texts etc. In this research, only the suitable part, consisting of 1,002 texts, was used.

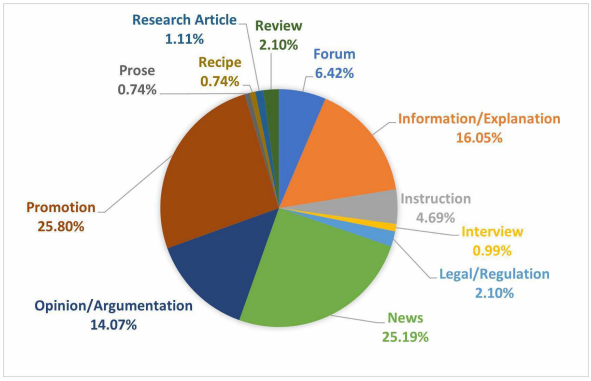
For the annotation, a GINCO schema was used, consisting of 24 labels, e.g., *News/Reporting*, *Opinion/Argumentation*, *Promotion of a Product*. The schema is based on the subcategory level of the CORE schema and on other schemata from previous genre studies. The texts were annotated by two annotators with the background in linguistics. In case of disagreement, final labels were determined at frequent meetings. Multi-label annotation was allowed, i.e., each text could be annotated with up to three classes which were ordered according to their prevalence in the text as a primary, secondary and tertiary label. However, in these experiments, only the primary labels are used. Each paragraph in the texts is accompanied with metadata (attribute `keep`) with information on whether it was manually identified to be a part of the main text and thus useful for the annotation. In this research, paragraphs not deemed to be useful were discarded.

The machine-translated GINCO corpus (MT-GINCO) was created by translating the Slovene GINCO 1.0 to English with the DeepL[1] machine translation system. The system is stated by its developers to be "3x more accurate" than its closest competitors, i.e., Google Translate, Amazon Translate and Microsoft Translator, based on internal blind tests (DeepL, nd). DeepL was confirmed to outperform Google Translate also in an independent study of Yulianto and Supriatnaningsih (2021). The GINCO corpus was translated into British English, as this variety seems to be more frequent than American English in the general part of the GloWbE corpus on which the CORE corpus is based (GloWbe, nd). The prevalence of the British variety in the CORE corpus was also confirmed with a lexicon-based American-British-variety Classifier (Rupnik et al., 2022) which identified 40% of texts to be British, 25% to be American, while the rest contain a mixture of both varieties or no signal for any of them.

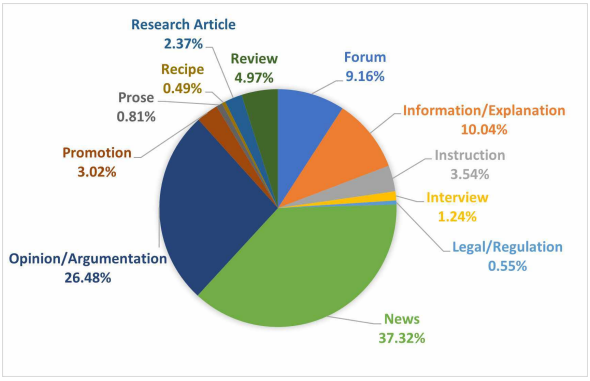
3.2. GINCORE Schema

To be able to perform cross-dataset experiments, the CORE and GINCO schemata were mapped to a joint

[1] `https://www.deepl.com/translator`



(a) SL-GINCORE and MT-GINCORE.



(b) EN-GINCORE.

Figure 1: The differences between the distributions of GINCORE labels in the GINCO corpora MT-GINCORE and SL-GINCORE, and in the EN-GINCORE (CORE corpus).

schema – the GINCORE schema. The schemata were mapped based on descriptions of categories in previous research, in the annotation guidelines for GINCO[2] and the guidelines for CORE, created for the needs of annotation of Finnish, French and Swedish corpora using the CORE schema[3] in further research (Laippala et al., 2019; Laippala et al., 2020). Furthermore, manual inspection of instances from the GINCO and CORE corpora was performed to analyze to which extent the annotations in the corpora match the guidelines. The basis of the GINCORE schema was the GINCO schema as it was shown to provide a more reliable annotation than CORE (see Kuzman et al. (2022)). Moreover, it is easier to map 54 CORE subcategories with a very high granularity to 24 broader GINCO categories than vice versa. The CORE schema consists of broad main categories and more specific subcategories. As the GINCO schema was based on the subcategories of the CORE schema, the subcategories level was used for the mapping from CORE to GINCORE.

Some of genre categories in both schemata are identical and can be directly mapped, namely *Recipe*, *Review*, *Interview* and *Legal/Regulation*. As the GINCO and CORE schemata differ in granularity, broader GINCORE labels were created which efficiently cover categories from both schemata. Some CORE categories were not included in the mapping, because a) these labels revealed to be very infrequent and there is no sufficient information about them available, b) the labels were too broad or problematic for annotators and as a result include instances that are too heterogeneous and cannot be mapped to just one GINCORE label. The resulting GINCORE schema[4] covers 43 CORE subcategories and all 24 GINCO categories by using 20 labels: 15 labels that are present in both corpora, and 5 labels, newly introduced by the GINCO schema and thus present only in the GINCO dataset.

3.3. GINCORE Datasets

For the purpose of performing cross-dataset experiments, only the GINCORE classes that have more than 5 instances in each of the datasets were used, resulting in a smaller set of 12 GINCORE labels: *News*, *Forum*, *Opinion/Argumentation*, *Review*, *Research Article*, *Information/Explanation*, *Promotion*, *Instruction*, *Prose*, *Interview*, *Legal/Regulation*, and *Recipe*. The texts annotated with other GINCORE labels were not included in the experiments. Thus, the final datasets are slightly smaller:

- the English CORE dataset with 12 GINCORE labels, henceforth referred to as the English GINCORE dataset (EN-GINCORE), consists of 33,918 texts;
- the Slovene GINCO dataset with 12 GINCORE labels, henceforth referred to as the Slovene GINCORE dataset (SL-GINCORE), consists of 810 texts;
- the machine-translated English GINCO dataset with 12 GINCORE labels, henceforth referred to as the Machine-Translated GINCORE dataset (MT-GINCORE), consists of 810 texts.

The text instances were not pre-processed, i.e. each instance is a running text as it was extracted from the original web page from which the boilerplate and HTML tags were removed. In GINCO datasets (SL-GINCORE and MT-GINCORE), the texts consist of paragraphs, which is indicated by the <p> tag, while in the CORE dataset (EN-GINCORE), the partitioning into paragraphs is not preserved. In addition to this, the datasets differ significantly in terms of length of the texts. In the CORE dataset, the median length is 649 words, while the minimum and maximum text length is 52 words and 118,278 words respectively. In the GINCO datasets, most texts are significantly shorter, with the median length of 198 words, minimum length of 12 words and maximum length of 4,134 words. As the Transformer models, used in the experiments, can

[2]The guidelines for GINCO are available here: https://tajakuzman.github.io/GINCO-Genre-Annotation-Guidelines/.

[3]The guidelines for the annotation campaigns using the CORE schema are available here: https://turkunlp.org/register-annotation-docs/.

[4]The final table with all the GINCORE mappings is available here: https://tajakuzman.github.io/GINCO-Genre-Annotation-Guidelines/genre_pages/GINCORE_mapping.html.

process maximum instance length of 512 tokens, this means that while the models will in most cases be trained on complete texts from the GINCO datasets, more than half of the texts from the CORE dataset will not be used in their entirety and the models will be trained only on the first part of these instances.

Here, it should be also noted that the CORE dataset and the GINCO datasets are characterized by a different distribution of GINCORE classes. Frequency of some classes, such as *Promotion*, is significantly different, as can be seen in Figure 1.

4. Machine Learning Experiments

4.1. Models

Experiments were performed with the Transformer-based pre-trained language models which were shown to perform well in the automatic genre identification task in a monolingual as well as a cross-lingual setting (Repo et al., 2021). More specifically, two models were used, the base-sized massively multilingual XLM-RoBERTa model (Conneau et al., 2020), and the trilingual Croatian-Slovene-English CroSloEngual BERT model (Ulčar and Robnik-Šikonja, 2020). The XLM-RoBERTa model was chosen because it was revealed to be the best performing model in cross-lingual automatic genre identification based on the CORE dataset (Repo et al., 2021), and to be comparable to the Slovene monolingual model SloBERTa (Ulčar and Robnik-Šikonja, 2021) in experiments, performed on GINCO (Kuzman et al., 2022). The CroSloEngual BERT model was revealed to achieve results comparable to the XLM-RoBERTa model or to even outperform the latter model in common monolingual and cross-lingual NLP tasks (Ulčar et al., 2021). Thus, it was included in these experiments to explore whether it achieves similar results on the AGI task as well.

4.2. Experimental Setup

The datasets were split into 60:20:20 train, dev and test splits, stratified according to the label distribution. The models were trained on the train split, consisting of 20,350 texts in the case of EN-GINCORE, and of 486 texts in the case of SL-GINCORE and MT-GINCORE, and tested on the test split, i.e., 6,784 texts in the case of EN-GINCORE and 162 texts in the case of SL-GINCORE and MT-GINCORE. The dev split, which is of the same size as the test split, was used for testing the hyperparameter optimization. When splitting the datasets, it was assured that the splits of SL-GINCORE and MT-GINCORE contain the same instances, so that they differ only in the language of the content.

The Transformer models are available at the Hugging Face repository and were trained using the Simple Transformers library. To find the optimal number of epochs and the learning rate, the hyperparameter search was performed separately for CroSloEngual BERT and XLM-RoBERTa. The maximum sequence length was set to 512 tokens and other hyperparameters were set to default values. As the EN-GINCORE dataset is more than 40 times larger than the SL-GINCORE and MT-GINCORE datasets, separate hyperparameter searches for each dataset were performed. Optimum learning rate was revealed to be 10^{-5}, while the optimum number of epochs varies based on the training dataset and the model, i.e., the optimum number of epochs when training on the EN-GINCORE with a) XLM-RoBERTa is 9, and b) CroSloEngual BERT is 6; while the optimum number of epochs when training on the SL-GINCORE and MT-GINCORE with a) XLM-RoBERTa is 60, and b) CroSloEngual BERT is 90.

We performed monolingual in-dataset experiments and cross-lingual cross-dataset experiments[5]. The monolingual experiments, described in Section 4.3.1., are in-dataset experiments, which means that the models were trained and tested on splits from the same dataset. In contrast to this, in cross-dataset experiments, presented in Section 4.3.2., the models are trained on one dataset and tested on the other. At the same time, these experiments are cross-lingual, as the original datasets are in different languages.

Three runs of each experiment were performed and average results are reported. The models used in monolingual and cross-lingual setups were evaluated via micro F1 and macro F1 scores to measure the instance-level and the label-level performance.

4.3. Results

4.3.1. Monolingual In-dataset Experiments

First, the datasets are compared via monolingual in-dataset experiments where the models were trained and tested on the splits of the same dataset. In addition to this, a dummy classifier which predicts the majority class was implemented as an illustration of the lower bound. The results, presented in Table 1, show that the mapping of the original labels into a joint schema was successful and that it is possible to achieve good results when learning Transformer models on GINCORE datasets. Transformer models are shown to be very effective at this task, achieving micro and macro F1 scores that are higher than the scores of the dummy model for at least 30 points. XLM-RoBERTa, which was revealed to be the best performing model, achieved relatively high results, with micro and macro F1 scores ranging between 0.72 and 0.84, even when trained on the two smaller datasets, which consist of less than 1,000 instances.

The results show that in a monolingual setting, the massively multilingual XLM-RoBERTa model outperforms the trilingual CroSloEngual BERT model. While Ulčar et al. (2021) showed that the trilingual model is comparable to the XLM-RoBERTa model at NLP tasks which are focused on classification of words or multiword units, such as named-entity recognition and part-of-speech tagging, these results reveal that CroSloEngual BERT is not as suitable as XLM-RoBERTa for automatic genre identification.

Among all monolingual experiments, the best micro and macro F1 results were achieved when the XLM-RoBERTa was trained and tested on the machine-translated MT-GINCO dataset, reaching average micro and macro F1 scores of 0.81 and 0.84 respectively. At the same time, the

[5]The code for data preparation and machine learning experiments is available here: `https://github.com/TajaKuzman/Cross-Lingual-and-Cross-Dataset-Experiments-with-Genre-Datasets`.

Datasets		Majority classifier		XLM-RoBERTa		CroSloEngual BERT	
Trained on	Tested on	Micro F1	Macro F1	Micro F1	Macro F1	Micro F1	Macro F1
SL-GINCORE	SL-GINCORE	0.259	0.027	0.782±0.02	0.725±0.01	0.738±0.01	0.599±0.06
MT-GINCORE	MT-GINCORE	0.259	0.027	**0.807**±0.01	**0.841**±0.03	0.714±0.00	0.501±0.05
EN-GINCORE	EN-GINCORE	0.363	0.036	0.768±0.00	0.715±0.00	**0.761**±0.00	**0.706**±0.00
SL-GINCORE	EN-GINCORE	0.029	0.004	**0.639**±0.01	0.539±0.01	0.547±0.02	0.391±0.02
MT-GINCORE	EN-GINCORE	0.029	0.004	0.625±0.01	0.521±0.01	0.585±0.01	0.409±0.01
EN-GINCORE	SL-GINCORE	0.253	0.027	0.603±0.02	0.575±0.03	0.566±0.02	0.510±0.03
EN-GINCORE	MT-GINCORE	0.253	0.027	0.630±0.02	**0.663**±0.03	**0.630**±0.01	**0.543**±0.01

Table 1: Results of monolingual and cross-lingual experiments performed with XLM-RoBERTa and CroSloEngual BERT models, reported via micro and macro F1 scores (averaged over three runs). As a baseline, the scores of a majority classifier are added. The best scores for each of the two Transformer models for each of the two setups (in-dataset experiments and cross-dataset experiments) are shown in bold.

lowest scores, i.e., micro F1 of 0.71 and macro F1 of 0.50, were obtained on the same dataset in combination with the CroSloEngual BERT. Similarly, while XLM-RoBERTa achieved the worst results when trained and tested on the EN-GINCORE, CroSloEngual BERT achieved the best results on this dataset. The difference between the results on the same datasets shows the importance of analyzing the output of multiple models before reaching any conclusion regarding the datasets – if only XLM-RoBERTa would be used, one could assume that the EN-GINCORE dataset is less suitable for automatic genre identification experiments. However, after performing experiments with both models, we can see that no dataset consistently provides the best results.

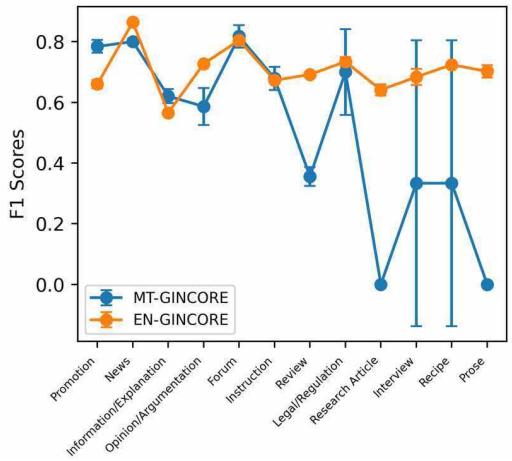


Figure 2: F1 scores per labels (averaged over three runs) in in-dataset experiments with MT-GINCORE and EN-GINCORE, performed with CroSloEngual BERT. Labels are ordered according to their frequency in the smallest of the two datasets, MT-GINCORE.

If we compare experiments, performed with the same model, we can observe that the largest differences between the datasets are in terms of macro F1 scores which are calculated on the level of labels. As shown in Figure 2, the biggest differences between the F1 scores per labels occur in cases of labels that are represented by a very small number of instances in the smaller datasets, SL-GINCORE and MT-GINCORE. Half of the labels, i.e., *Review*, *Legal/Regulation*, *Research Article*, *Interview*, *Recipe* and *Prose*, are represented by solely 4 instances or less in SL-GINCORE and MT-GINCORE test splits. One should be aware that this means that a correct or incorrect prediction of such a small number of instances per labels has a large impact on the macro F1 score. Furthermore, a correct prediction of labels with only one or two instances in the test split might happen due to chance or a similarity of texts in the train and test split. Thus, the F1 scores of these labels are not reliable. As shown in Figure 2, in the three runs, the F1 scores of *Interview* and *Recipe*, which are represented by only 1 instance in the SL-GINCORE and MT-GINCORE test sets, were either 0 or 1, which has a large impact on a macro F1. These results also show how important it is to repeat each experiment multiple times, to ascertain stability and reliability of results.

If we compare the three datasets based on micro F1 scores, there are small differences between them, i.e., a difference of 4 points between the lowest and highest scores when XLM-RoBERTa was used and a difference of 5 points when CroSloEngual BERT was used. Interestingly, although the EN-GINCORE is 40 times larger than the SL-GINCORE and MT-GINCORE, it does not provide higher results than the other two datasets when the XLM-RoBERTa model is used for training. Similar results were revealed in previous work (see Repo et al. (2021)) where they performed monolingual experiments with XLM-RoBERTa on the CORE dataset and three smaller genre-annotated datasets, Finnish FinCORE, French FreCORE and Swedish SweCORE datasets. Although the non-English datasets were annotated with the CORE schema, the annotation procedure and dataset collection methods are more similar to the GINCO approach than CORE. Their experiments showed that the XLM-RoBERTa and other Transformer models perform similarly or better when trained on datasets which consisted of 1,800 to 2,200 instances than when trained on the CORE dataset.

We have two hypotheses why this is the case: 1) It might be that due to high capacities of Transformer models, their performance on this task plateaus already at a few thousand instances and contributing bigger datasets does not significantly improve the results. 2) Or this could indicate that

the CORE dataset is less suitable for AGI machine learning experiments. The reason for that could be that as crowdsourcing was used for the annotation of the dataset, the assigned labels are less reliable and the classes are consequently fuzzier. Poor reliability of the dataset was also confirmed by low inter-annotator agreement. The authors of the dataset reported that there was no agreement between at least three of four annotators on the subcategory of 48.98% of texts (Egbert et al., 2015). When the schema and approach was used by Sharoff (2018) on another corpus, he reported nominal Krippendorff's alpha of 0.53 on the level of subcategories, which is below the acceptable threshold of 0.67, as defined by Krippendorff (2018). In contrast to this, the GINCO dataset was reported to achieve Krippendorff's alpha of 0.71, confirming much higher reliability of annotations.

4.3.2. Cross-lingual Cross-dataset Experiments

To assess comparability of the English CORE dataset and the Slovene GINCO dataset, we performed cross-lingual cross-dataset experiments by training the Transformer models on one dataset and testing them on another. In addition to experimenting with cross-lingual transfer from Slovene to English dataset and vice versa, we also explored whether translating the Slovene dataset into English with a machine translation system improves the results of cross-dataset experiments.

The results, shown in Table 1, reveal that the trilingual CroSloEngual BERT model performs worse than the massively multilingual XLM-RoBERTa model in the cross-lingual experiments with a difference of 12 points between the highest macro F1 scores obtained by the models and a much slighter difference between the highest micro F1 scores (0.009).

In general, results obtained in the cross-lingual experiments are significantly lower than the results from the monolingual experiments. If we compare experiments performed with XLM-RoBERTa, there are differences in 13–18 points in micro F1 and 5–32 points in macro F1 between testing the model on the same dataset as it was trained on (monolingual experiments) and on another dataset (cross-lingual experiments). In case of CroSloEngual BERT, the differences between testing on the same dataset versus testing on the other dataset were in 13–20 points in micro F1 and 9–20 points in macro F1.

Nevertheless, the XLM-RoBERTa scores, which range between 0.6–0.64 and 0.52–0.66 for micro and macro F1 respectively, are a promising indicator that cross-lingual transfer could be possible in this task for Slovene as well. Furthermore, the results are comparable to the results of cross-lingual experiments with the CORE corpora, reported by Repo et al. (2021). When they trained the XLM-RoBERTa model on the CORE corpus and tested it on Finnish, Swedish and French datasets, annotated with the CORE schema, the micro F1 scores ranged from 0.61 to 0.69. Here it needs to be noted that they used a large-sized model which was shown to significantly outperform the base-sized model used by us (Conneau et al., 2020), and that they used 8 labels, while we used 12. Considering this, the results of learning on CORE, mapped to the GINCORE schema, and testing on SL-GINCORE, which reached 0.60 micro F1 with the base-sized XLM-RoBERTa model, are promising, showing that mapping to the GINCORE schema gives comparable results to using the CORE schema.

To obtain a deeper insight into the comparability of the GINCO and CORE corpora, we can compare how the F1 scores per labels change when we test the model on another corpus versus when we test it on the same dataset. Figure 3 shows a comparison between the F1 scores per labels for in-dataset experiments with SL-GINCORE and cross-dataset experiments from SL-GINCORE to EN-GINCORE, performed with XLM-RoBERTa. An analysis of these experiments, performed with CroSloEngual BERT, confirmed that differences between label scores occur when learning with any of the two models, and do not depend on the model. The same differences in label scores were also observed in experiments where MT-GINCORE is used instead of SL-GINCORE, which indicates that the language of the dataset does not seem to have a large impact on the results per labels.

As shown in Figure 3, the F1 scores for *News* and *Opinion/Argumentation* are almost the same in both setups, which shows that in regard to these genres, the datasets are comparable enough for the model to generalize from one dataset to the other. The F1 scores are significantly lower in cross-lingual experiments in case of *Promotion*, *Information/Explanation*, *Forum* and *Instruction*. For the labels that are under-represented in the SL-GINCORE, i.e., labels that are on the right side of *Review* in the Figure, it is not possible to ascertain whether the differences between the scores are an indicator that the datasets are not comparable in regard to these labels or that the differences occurred due to chance.

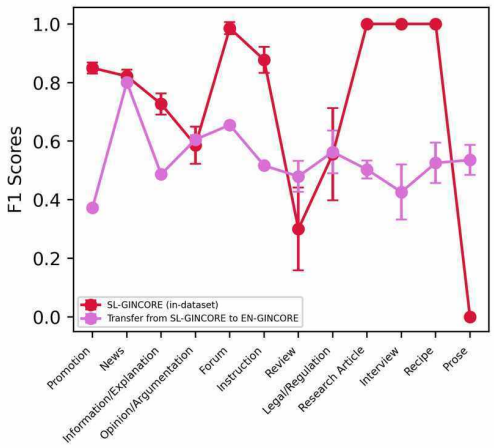


Figure 3: Comparison of average F1 scores per labels between in-dataset experiments and cross-dataset experiments with XLM-RoBERTa. The models were trained on SL-GINCORE, and tested on a) SL-GINCORE (in-dataset experiments) and b) EN-GINCORE (cross-dataset experiments). Labels are ordered according to their frequency in the smallest of the datasets, SL-GINCORE.

As in the in-dataset experiments, experiments with the two Transformer models show that while one dataset combination seems to achieve the best results with one model,

it performs differently with the other model. These results once again show the importance of using multiple models on multiple datasets in the experiments to see whether conclusions obtained from experiments with one model are still supported when using another, yet similar model, and how the performance of the models depends on the datasets. While results in terms of micro F1, achieved with XLM-RoBERTa, point to a conclusion that transfer from SL-GINCORE to EN-GINCORE achieves better results than the other direction, macro F1 scores, achieved with XLM-RoBERTa, and both F1 scores, achieved with CroSloEngualBERT, show transfer direction from English to Slovene to be better. However, although the EN-GINCORE dataset is 40 times larger than SL-GINCORE, the transfer from EN-GINCORE to SL-GINCORE does not achieve significantly higher results than the transfer in the other direction when the Slovene dataset is used.

In addition to this, the results show that machine-translating the dataset into English can in some cases improve the results of cross-lingual experiments. In cases where the model was trained on the GINCO datasets, i.e., SL-GINCORE or MT-GINCORE, and tested on the EN-GINCORE dataset, the setup with the machine-translated text achieved slightly lower results than the setup with the original Slovene dataset, SL-GINCORE, in case of XML-RoBERTa, and slightly better results in case of CroSloEngual BERT. However, when the transfer was applied in the other direction, that is, from EN-GINCORE to SL- or MT-GINCORE, machine translating the test instances from Slovene into English resulted in improvements of macro F1 scores, achieved with XLM-RoBERTa, and both micro and macro F1 scores, obtained with CroSloEngual BERT.

5. Conclusions

Following Repo et al. (2021) who showed that good levels of cross-lingual transfer can be achieved by training Transformer models on a large English genre dataset and applying them to datasets in other languages, the goal of this study was to explore whether it is possible to achieve similar results on the Slovene genre dataset. The results revealed to be promising, as despite using a smaller Transformer model and a different schema with more labels than previous work, the results are rather comparable, showing that the English CORE and Slovene GINCO datasets are comparable enough to allow cross-dataset experiments. The XLM-RoBERTa scores, which range between 0.6–0.64 and 0.52–0.66 in terms of micro and macro F1 scores respectively, are a promising indicator that cross-lingual transfer could be possible in the automatic genre identification task for Slovene as well. Furthermore, high F1 scores achieved with XLM-RoBERTa in monolingual experiments show that automatic genre identification is feasible already with a very small dataset, and that using the GINCORE schema on all datasets gives good results. Moreover, despite the fact that the CORE dataset is 40 times larger than the GINCO dataset, it did not provide consistently significantly better results than the GINCO dataset in either of the setups. We plan to analyze this further by exploring what results can be achieved when smaller portions of CORE are used for training, and by extending the GINCO dataset to analyze whether this further improves the results.

As recently developed trilingual Croatian-Slovene-English CroSloEngual model was shown to be comparable to massively multilingual XLM-RoBERTa model in numerous NLP tasks (see Ulčar et al. (2021)), both models were used in the experiments to analyze their performance in the AGI tasks. The results of both monolingual and cross-lingual experiments showed that despite achieving high results in other common NLP tasks, CroSloEngual BERT seems to be less suitable than XLM-RoBERTa for automatic genre identification.

To improve monolingual and cross-lingual results, we also experimented with translating the Slovene GINCO dataset into English, which is the main language on which the Transformer models were pre-trained. In regard to monolingual experiments, there were no consistent results which would confirm that using an English dataset improves classification. However, when the models were trained on the English EN-GINCORE and tested on MT-GINCORE, i.e., a Slovene dataset, machine-translated into English, this led to improvement of macro F1 scores, achieved with XLM-RoBERTa, and both micro and macro F1 scores for CroSloEngual BERT. This means that machine translating the dataset into the language of another dataset might be beneficial in cross-lingual cross-dataset experiments.

Although monolingual and cross-lingual experiments showed good results also when the models were trained on SL-GINCORE and MT-GINCORE, consisting of less than 1,000 instances, comparisons of F1 scores, reported for each label in different runs and setups, showed that some labels are represented by too few instances to provide reliable results. In the future, we plan to extend the GINCO dataset to assure more reliable results and to further improve the classifiers' performance.

In addition to this, recent work by Rönnqvist et al. (2021) showed that multilingual modeling, where the model was trained on CORE datasets in various languages, resulted in significant gains over cross-lingual modeling, where the model was trained solely on the English CORE dataset. As our research revealed that the CORE and GINCO labels can be successfully mapped to a joint schema, in the future, we plan to extend the experiments to multilingual modeling by training the model on a combination of all CORE datasets and the Slovene GINCO dataset.

Acknowledgments

This work has received funding from the European Union's Connecting Europe Facility 2014-2020 - CEF Telecom, under Grant Agreement No. INEA/CEF/ICT/A2020/2278341. This communication reflects only the author's view. The Agency is not responsible for any use that may be made of the information it contains. This work was also funded by the Slovenian Research Agency within the Slovenian-Flemish bilateral basic research project "Linguistic landscape of hate speech on social media" (N06-0099 and FWO-G070619N, 2019–2023) and the research programme "Language resources and technologies for Slovene" (P6-0411).

6. References

Noushin Rezapour Asheghi, Serge Sharoff, and Katja Markert. 2016. Crowdsourcing for web genre annotation. *Language Resources and Evaluation*, 50(3):603–641.

Marta Bañón, Miquel Esplà-Gomis, Mikel L. Forcada, Cristian García-Romero, Taja Kuzman, Nikola Ljubešić, Rik van Noord, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Peter Rupnik, Vít Suchomel, Antonio Toral, Tobias van der Werff, and Jaume Zaragoza. 2022. Slovene web corpus MaCoCu-sl 1.0. Slovenian language resource repository CLARIN.SI.

Douglas Biber and Jesse Egbert. 2018. *Register variation online*. Cambridge University Press.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Mark Davies and Robert Fuchs. 2015. Expanding horizons in the study of World Englishes with the 1.9 billion word Global Web-based English Corpus (GloWbE). *English World-Wide*, 36(1):1–28.

DeepL. n.d. Why DeepL? `https://www.deepl.com/en/whydeepl`.

Jesse Egbert, Douglas Biber, and Mark Davies. 2015. Developing a bottom-up, user-based method of web register classification. *Journal of the Association for Information Science and Technology*, 66(9):1817–1831.

Tomaž Erjavec and Nikola Ljubešić. 2014. The slWaC 2.0 corpus of the Slovene web. *T. Erjavec, J. Žganec Gros (ur.) Jezikovne tehnologije: zbornik*, 17:50–55.

Eugenie Giesbrecht and Stefan Evert. 2009. Is part-of-speech tagging a solved task? An evaluation of POS taggers for the German web as corpus. In: *Proceedings of the fifth Web as Corpus workshop*, pages 27–35.

GloWbe. n.d. Corpus of Global Web-Based English (GloWbE): Texts. `https://www.english-corpora.org/glowbe/`.

Klaus Krippendorff. 2018. *Content analysis: An introduction to its methodology*. Sage publications.

Taja Kuzman, Mojca Brglez, Peter Rupnik, and Nikola Ljubešić. 2021. Slovene web genre identification corpus GINCO 1.0. Slovenian language resource repository CLARIN.SI.

Taja Kuzman, Peter Rupnik, and Nikola Ljubešić. 2022. The GINCO Training Dataset for Web Genre Identification of Documents Out in the Wild. In: *Proceedings of the Language Resources and Evaluation Conference*, pages 1584–1594, Marseille, France. European Language Resources Association.

Veronika Laippala, Roosa Kyllönen, Jesse Egbert, Douglas Biber, and Sampo Pyysalo. 2019. Toward multilingual identification of online registers. In: *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 292–297.

Veronika Laippala, Samuel Rönnqvist, Saara Hellström, Juhani Luotolahti, Liina Repo, Anna Salmela, Valtteri Skantsi, and Sampo Pyysalo. 2020. From web crawl to clean register-annotated corpora. In: *Proceedings of the 12th Web as Corpus Workshop*, pages 14–22.

Wanda J Orlikowski and JoAnne Yates. 1994. Genre repertoire: The structuring of communicative practices in organizations. *Administrative science quarterly*, pages 541–574.

Jan Pomikálek. 2011. *Removing boilerplate and duplicate content from web corpora*. Ph.D. thesis, Masaryk university Faculty of informatics, Brno, Czech Republic.

Liina Repo, Valtteri Skantsi, Samuel Rönnqvist, Saara Hellström, Miika Oinonen, Anna Salmela, Douglas Biber, Jesse Egbert, Sampo Pyysalo, and Veronika Laippala. 2021. Beyond the English web: Zero-shot cross-lingual and lightweight monolingual classification of registers. In: *16th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop, EACL 2021*, pages 183–191. Association for Computational Linguistics (ACL).

Samuel Rönnqvist, Valtteri Skantsi, Miika Oinonen, and Veronika Laippala. 2021. Multilingual and zero-shot is closing in on monolingual web register classification. In: *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 157–165.

Peter Rupnik, Taja Kuzman, and Nikola Ljubešić. 2022. American-British-variety Classifier. `https://github.com/macocu/American-British-variety-classifier`.

Serge Sharoff. 2018. Functional text dimensions for the annotation of web corpora. *Corpora*, 13(1):65–95.

Matej Ulčar and Marko Robnik-Šikonja. 2020. CroSloEngual BERT 1.1. Slovenian language resource repository CLARIN.SI.

Matej Ulčar and Marko Robnik-Šikonja. 2021. Slovenian RoBERTa contextual embeddings model: SloBERTa 2.0. Slovenian language resource repository CLARIN.SI.

Matej Ulčar, Aleš Žagar, Carlos S Armendariz, Andraž Repar, Senja Pollak, Matthew Purver, and Marko Robnik-Šikonja. 2021. Evaluation of contextual embeddings on less-resourced languages. *arXiv:2107.10614*.

Marlies Van der Wees, Arianna Bisazza, and Christof Monz. 2018. Evaluation of machine translation performance across multiple genres and languages. In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Vedrana Vidulin, Mitja Luštrek, and Matjaž Gams. 2007. Using genres to improve search engines. In: *1st International Workshop: Towards Genre-Enabled Search Engines: The Impact of Natural Language Processing*, pages 45–51.

Ahmad Yulianto and Rina Supriatnaningsih. 2021. Google Translate vs. DeepL: A quantitative evaluation of close-language pair translation (French to English). *AJELP: Asian Journal of English Language and Pedagogy*, 9(2):109–127.