Get to Know Your Parallel Data: Performing English Variety and Genre Classification over MaCoCu Corpora

Taja Kuzman and **Peter Rupnik**
Jožef Stefan Institute, Slovenia
taja.kuzman@ijs.si,
peter.rupnik@ijs.si

Nikola Ljubešić
Jožef Stefan Institute, Slovenia
Center za jezikovne vire in tehnologije
Univerze v Ljubljani, Slovenia
nikola.ljubesic@ijs.si

Abstract

Collecting texts from the web enables a rapid creation of monolingual and parallel corpora of unprecedented size. However, unlike manually-collected corpora, authors and end users do not know which texts make up the web collections. In this work, we analyse the content of seven European parallel web corpora, collected from national top-level domains, by analysing the English variety and genre distribution in them. We develop and provide a lexicon-based British-American variety classifier, which we use to identify the English variety. In addition, we apply a Transformer-based genre classifier to corpora to analyse genre distribution and the interplay between genres and English varieties. The results reveal significant differences among the seven corpora in terms of different genre distribution and different preference for English varieties.

1 Introduction

Collecting text corpora in an automatic manner, by crawling web pages, allows for quick gathering of large amounts of texts. With this approach, the MaCoCu[1] project (Bañón et al., 2022h) aims to provide some of the largest freely available monolingual and parallel corpora for more than 10 under-resourced European languages. However, in contrast to manual text collection methods, the disadvantage of automatic methods is that both the corpora creators and the users do not know what the overall quality of the dataset is, and what type of texts the collections consist of (Baroni et al., 2009). The MaCoCu corpora address this issue by providing rich metadata, including information on source URLs, paragraph quality, translation direction, English varieties, and genres. In this paper, we present two of the text classification methods, used to automatically enrich massive corpora with meaningful metadata: English variety classification and automatic genre identification. We show how they provide a better insight into the differences between corpora.

There is limited research on the use of British and American English in the non-native English-speaking countries. Previous findings show that these English varieties are preferred to a different extent in different educational systems (Forsberg et al., 2019), translation services (Forsyth and Cayley, 2022) and on different national webs (Atwell et al., 2007). However, to the best of our knowledge, there is no freely available classifier between American and British English which would allow easy identification of an English variety in large corpora, and thus allow for a corpus-based research of this phenomenon on a larger scale. To this end, we develop a fast and reliable classifier which is based on a lexicon of variety-specific spellings and words. In addition, we also compare the web corpora in terms of genres. Genres are text categories which are defined considering the author's purpose, common function of the text, and the text's conventional form (Orlikowski and Yates, 1994). Examples of genres are *News*, *Promotion*, *Legal*, etc. In addition to providing a valuable insight into the dataset content, information about the genre of the document was shown to be beneficial for various NLP tasks, including POS-tagging (Giesbrecht and Evert, 2009), machine translation (Van der Wees et al., 2018) and automatic summarization (Stewart and Callan, 2009).

The main contributions of our paper are the following:

1. We present a freely available American-British variety classifier that we make available as a Python package[2]. The classifier is based on a lexicon of variant-specific words and is thus reliable and fast. In contrast to deep neural models that are trained on varying

[1] https://macocu.eu/

[2] https://pypi.org/project/abclf/

Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023), pages 91–103

texts, deemed to represent different varieties, the classifier cannot be influenced by any bias in the training data, such as differences in topics or proper names, and its predictions are explainable. The classifier can be applied to any English corpus with texts of sufficient length and could be used for researching which English variety is preferred in different national web domains, official translations, school systems and so on.

2. We introduce a method of comparing large web corpora and obtaining additional insight into their contents based on English variety and genre information. We apply the English variety and genre classifier to 7 parallel web corpora harvested from the following European national webs: Icelandic, Maltese, Slovene, Croatian, Macedonian, Bulgarian and Turkish. The results show that English variety and genre information reveal differences between these datasets. These insights provide useful information to corpora creators and researchers that use the corpora for downstream tasks, such as training language models and machine translation models, as well as performing corpus linguistic studies on these corpora.

The paper is organized as follows. We first present the related work on English variety identification and automatic genre identification in Section 2. In Section 3, we present the web corpora to which we apply the classifiers, described in Section 4. The results in Section 5 show that these approaches reveal important differences between the corpora. The paper concludes with Section 6, where we summarize the main findings and present future work.

2 Related Work

Diatopic variation, that is, variation among national varieties of the same language (Zampieri et al., 2020), can be approached similarly to variation between different languages. Two main approaches in language identification of English varieties are 1) corpus-based text classification and 2) lexicon-based text classification. In corpus-based classification, researchers use datasets which have a known origin of the texts as a reference based on which the classifiers are trained and evaluated, while lexicon-based classifiers identify varieties based on a list of variety-specific words or spelling variants.

Most previous studies on identification of English varieties were corpus-based (Lui and Cook, 2013; Utomo and Sibaroni, 2019; Cook and Hirst, 2012; Dunn, 2019; Simaki et al., 2017; Rangel et al., 2017). The advantage of corpus-based classification is that as the model is trained on actual text collections, it could show the differences in the varieties as they appear "in the wild", and researchers do not need a profound knowledge of lexical differences between the varieties that linguists are aware of. To obtain reference datasets that are large enough to be used for training the model, researchers most often used or constructed web corpora (Atwell et al., 2007; Lui and Cook, 2013), using the national top-level domains as indicators of the text origin (e.g., .uk for British English), journalistic corpora (Zampieri et al., 2014), national corpora (Lui et al., 2014; Utomo and Sibaroni, 2019), such as the British National Corpus (BNC) (Consortium et al., 2007) , and/or social media corpora (Dunn, 2019; Simaki et al., 2017; Rangel et al., 2017), consisting of texts from Twitter and Facebook, where the variety is assigned to texts based on the metadata about the post or its author.

However, one of the major drawbacks of this approach is that it is assumed that the texts from a specific top-level domain or posted to social media from a certain location are written by a native speaker of this variety, while this is hard or impossible to verify. In addition, web, national and journalistic corpora can contain cross citations and republications (e.g., a British text that was republished by an American newspaper website and vice versa). This was revealed for the DSL corpus collection, used in the Discriminating between Similar Languages (DSL) shared task 2014, where 25% of texts were discovered to be likely annotated with the wrong English variety (Zampieri et al., 2014). Another drawback of the corpus-based approach is that training on text collections can introduce various bias into the classification task. As no parallel corpus of English varieties exists, the classification is based on two or more separate collections of texts. The datasets which represent each variety do not differ only in language specificities, but also in content and style. This hinders learning truly representative differences between the varieties, and the classification models might learn to differentiate between the datasets based on other differences,

unrelated to varieties, such as topic (Kilgarriff and Kilgarri, 2001; Tiedemann and Ljubešić, 2012).

An alternative approach to the corpus-based classification is lexicon-based. It has been used by Lui and Cook (2013) who devised a "variant pair" classifier based on the VarCon lexicon of spelling variants (Atkinson and Titze, 2020). This approach does not introduce any biases, related to the corpora content, and the classification is explainable. However, its disadvantage is that as it relies on a specific list of words, if none of the words occur in a text, its variety is unknown, meaning that some texts in the corpus may remain unclassified.

While research on automatic identification of English varieties is rather limited, automatic genre identification (AGI) has been an established text categorization task ever since the advent of the world wide web. As genre information is very useful for obtaining better hits to a query in information retrieval tools, used on the web, there has been large interest for genre identification in the area of information retrieval (see Roussinov et al. (2001); Vidulin et al. (2007)). In addition, with the emergence of technologies for automatic collection of text corpora, an interest for tools for AGI emerged also in the field of corpora creation and curation. To this end, genre researchers devised sets of genre categories which aim to cover all of the diversity of texts found on the web, and provided manually annotated datasets (see Egbert et al. (2015); Sharoff (2018); Kuzman et al. (2022b)). Classification of genres was shown to be a hard task as texts can display characteristics of multiple genres (Sharoff, 2021), and most genre classification models were not able to generalize outside of the dataset on which they were trained (Sharoff et al., 2010). However, recent advances in deep neural technologies led to a breakthrough in this field, and Transformer-based language models (Vaswani et al., 2017), fine-tuned on manually-annotated genre datasets, showed the ability to identify genres in various web corpora and languages (see Rönnqvist et al. (2021); Kuzman et al. (2022a)). Following encouraging results, Transformer-based genre classifiers have started to be applied to web corpora to provide genre information as metadata. For instance, as part of newly available massive Oscar web corpora, 351 million documents in 14 languages were enriched with genre information (Laippala et al., 2022).

Dataset	Size	Text length
MaCoCu-tr-en	193,782	184
MaCoCu-hr-en	91,619	172
MaCoCu-sl-en	91,459	190
MaCoCu-bg-en	88,544	170
MaCoCu-mt-en	21,376	300
MaCoCu-mk-en	20,108	194
MaCoCu-is-en	11,639	201

Table 1: Comparison of English datasets, extracted from the parallel corpora, in terms of size (number of English texts) and median text length in words.

3 Datasets

In this paper, we compare seven parallel web corpora, created in the scope of the MaCoCu project (Bañón et al., 2022h): Croatian-English MaCoCu-hr-en (Bañón et al., 2022b), Slovene-English MaCoCu-sl-en (Bañón et al., 2022f), Bulgarian-English MaCoCu-bg-en (Bañón et al., 2022a), Macedonian-English MaCoCu-mk-en (Bañón et al., 2022d), Turkish-English MaCoCu-tr-en (Bañón et al., 2022g), Icelandic-English MaCoCu-is-en (Bañón et al., 2022c) and Maltese-English MaCoCu-mt-en (Bañón et al., 2022e) corpus. The corpora were created by crawling the national top-level domains, e.g. the Slovenian top-level domain `.si` for the English-Slovene dataset MaCoCu-sl-en. Important to note is that the crawl primarily focused on the top-level domain crawling, but was allowed to harvest data from generic domains (`.com`, `.net` etc.) if the domain proved to have enough data in the language being crawled. Websites containing the target language and English were identified and processed with the Bitextor[3] tool.

3.1 Preparation of Datasets

The corpora we analyse are available in a sentence-level and paragraph-level format. Based on the information on the URL of the original document and metadata on the position of the sentence in this document, we took English sentences from the sentence pairs in the sentence-level format and created a document-level corpus of English texts from each parallel corpus.

We applied the American-British variety classifier and genre classifier to the documents. Finally, as a post-processing step, we filtered out texts with noisy genre predictions, that is, based on manual

[3] https://github.com/bitextor/bitextor

inspection, we decided to remove texts that were annotated with two less reliable and very infrequent labels – Forum and Other –, and texts with labels where the confidence of the genre classifier was low. This post-processing step amounted to around 10% of documents being discarded from the final corpora. The code for the dataset preparation, enrichment and analysis of the results is published on GitHub for the purposes of reproducibility[4].

3.2 Final Datasets

The final sizes of datasets, used in our comparisons, are shown in Table 1. The Turkish corpus is the largest, consisting of almost 200,000 English texts, followed by the Croatian, Slovenian and Bulgarian corpora with around 90,000 English texts. The smallest corpora are Maltese, Macedonian and Icelandic with 10,000 to 20,000 English texts. Table 1 also shows differences between the median length of texts. While the median text length in most datasets is between 170 and 200 words, the Maltese stands out with longer texts with the median length of 300 words.

4 Enrichment of Datasets

4.1 American-British Variety Classifier

Although there exist numerous English varieties throughout the world, including Indian English, New Zealand English, Irish English, etc., in this paper, we focus on differentiating between American and British English, which are often considered as the main varieties of standard English (Quirk, 2014). To avoid topic-related and other biases that come with training a classifier on any reference corpora, we opted for the lexicon-based approach. At the same time, as the classifier is based on a lexicon of variety-specific words and spellings, it has a limited coverage: it cannot classify texts if they do not contain any of variety-specific words. However, to obtain reliable results, we opted for a high precision approach rather than high recall.

To create our classifier, we used the VarCon lexicon of different spellings and vocabularies (Atkinson and Titze, 2020) which is based on various dictionaries and resources on spelling differences. We extracted British and American variety-specific words from the lexicon. To improve the classifier's performance and reliability, a researcher with a translation background inspected the list. We discarded rare words and words that are specific for one variety solely when used as a certain part-of-speech type, e.g. *can* (noun, as opposed to *tin*, while the verb *can* is used in both varieties), or in a certain context, e.g., *rubber* (as opposed to *eraser*, while the material *rubber* is used in both varieties). Multiple English dictionaries were consulted as a reference, including Oxford Advanced Learner's Dictionary of Current English (Hornby, 1995) and the online Cambridge dictionaries[5].

The final size of the lexicon is 6,041 words. It includes spelling differences, such as "-our" versus "-or" (Br. *behaviour*, Am. *behavior*), "-ll-" versus "-l-" (Br. *bejewelled*, Am. *bejeweled*), "-ae-" versus "-a-" (Br. *anaemia*, Am. *anemia*), "-re" versus "-er" (Br. *theatre*, Am. *theater*). While the great majority of words in the lexicon are spelling variants, there are also some variety-specific words, such as Br. *lorry* and Am. *beltway*.

Since the spelling variant "-ise" (*apologise, criticise*, etc.) is specific for British English, while its alternative "-ize" is used in both American and British English, we included only the British "-ise" variants of these words. Consequently, the lexicon is unbalanced towards British. It consists of 4,368 British words and 1,673 American words. In this paper, we trained the classifier on the unbalanced lexicon. However, we also provide a balanced lexicon by discarding the British "-ise" spelling variants, and allow an option of using the classifier with the balanced lexicon. It consists of 3,268 words: 1,652 American and 1,616 British words. Both lexicons are made available along with the code of the classifier[6].

The American-British variety classifier transforms the input text into lower case and counts the number of variety-specific words from the lexicon that are present in the text. If no variety signal is present, the text is classified as "unknown". If one variety is at least twice as present than the other, the text is classified as the prevalent variety, either as British or American. If both varieties are present in a similar extent, the text is classified as a "mix". The resulting classifier is fast and explainable. It classifies a text of an average length from the MaCoCu corpora (190 words) in 0.25 ms and a text of 1,000 words in 1.2 ms.

We analysed the classifier's reliability by performing a manual analysis of the lexicon it is based

[4] https://github.com/TajaKuzman/Applying-GENRE-on-MaCoCu-bilingual

[5] https://dictionary.cambridge.org/dictionary/

[6] https://github.com/macocu/American-British-variety-classifier

Dataset	Size (texts)	Median length	Coverage	Accuracy	Mix
DSL-TL dev	599	30	12%	94%	0.1%
GloWbE + NOW	1,445	634	66%	90%	4.0%
PAN17 test	800	1,391	78%	94%	12.0%

Table 2: The size of datasets, used for testing the coverage and performance of American-British classifier, in terms of number of texts, and the median length of texts in terms of number of words. The coverage shows what percentage of texts was assigned a variety label (American or British) as opposed to the labels "unknown" or "mix". The accuracy is calculated only for the texts that were assigned a variety label. Mix shows the percentage of texts which include words from both varieties.

on, and by improving the lexicon by using the English dictionaries as a reference. We also evaluated the performance of the classifier on three datasets, annotated with English varieties: 1) the web corpora GloWbE (Davies, 2013) and NOW (Davies, 2016), 2) the manually-annotated news DSL-TL dataset, and 3) the Twitter PAN17 dataset (Rangel et al., 2017). We evaluated the classifier in two criteria: coverage – in what percentage of the texts it recognizes a variety instead of categorizing them as "UNK" (unknown) or "MIX" – and performance, calculated for the texts to which a British or American variety is assigned.

To test the classifier on web-corpus-like content, we applied it over samples of the Corpus of Global Web-based English (GloWbE) (Davies, 2013) and the News on the Web (NOW) corpus (Davies, 2016). The GloWbE corpus is a web corpus, collected by searching frequent n-grams on Google, while the NOW corpus consists of texts from web-based newspapers and magazines. While the corpora consist of texts from around 20 English-speaking countries, we used only texts from United Kingdom and United States. The sample is balanced between the two varieties and consists of around 1,400 texts. As shown in Table 2, our classifier identified a British or American variety in two thirds of texts (66%) and 90% of them were predicted correctly.

Similar results were obtained on the Twitter dataset[7] from the PAN 2017 shared task on author profiling (Rangel et al., 2017). The English part of the dataset comprises tweets, originating from the United States, Great Britain, Ireland, Canada, Australia and New Zealand. However, we used only texts from Great Britain and United States. For each author, 100 tweets were collected and concatenated into one text instance, and the assigned language variety was based on the location from which the author mostly posted tweets. We applied the American-British classifier on the test split of the dataset, which consists of 800 texts with the median text length of around 1,400 words. As shown in Table 2, the American-British variety classifier identified a variety in 78% of texts with accuracy of 94%. Out of the unidentified texts, 12% were revealed to consist of words from both varieties which might point toward lower reliability of this dataset.

In contrast, the classifier performed poorly when tested on the DSL-TL dataset[8]. The dataset is a subset of the DSLCC dataset (Zampieri et al., 2014) that was manually annotated with American and British English variety labels for the VarDial 2023 shared task on discriminating between similar languages. At the time of writing the paper, the test set with labels has not been published yet, so we tested our classifier on the development split. The dataset consists of excerpts from journalistic texts which are rather short – the median text length of the texts in the dev subset is only 30 words. The texts were shown to be too short to provide any signal of English varieties to our classifier. As shown in Table 2, it recognized English varieties in only 12% of texts. However, its accuracy on the labeled texts was high, reaching 94%.

The comparison of results on the three datasets shows a high reliability of the classifier on the texts that were predicted to be British or American. It also nicely shows its limitations, connected with the length of texts. Results in Table 2 show very clearly, but also very expectedly, that the longer the texts are, the bigger is the classifier's coverage.

[7]The PAN17 dataset is available at https://zenodo.org/record/3745980#.ZBxM3HbMI2w.

[8]The dataset is available at https://github.com/LanguageTechnologyLab/DSL-TL

4.2 Genre Classifier

To obtain information on genres in the corpora, we used the X-GENRE classifier[9], a multilingual classifier which categorizes texts into genres. It uses the following genre categories: *Information/Explanation*, *Instruction*, *News*, *Legal*, *Promotion*, *Opinion/Argumentation*, *Prose/Lyrical*, *Forum* and *Other* (see the description of the labels in Appendix A). The classifier is based on the base size multilingual XLM-RoBERTa Transformer-based model (Conneau et al., 2020). It was fine-tuned on a combination of three datasets, manually annotated with genre labels: English CORE (Egbert et al., 2015), English FTD (Sharoff, 2018) and Slovene GINCO (Kuzman et al., 2022b) dataset. Each of the datasets has their own set of categories, which were mapped into a joint schema. The reason for using multiple datasets instead of just one is to assure better generalization of the model to new datasets and languages.

We manually annotated around 150 English texts from the Slovene MaCoCu-sl-en corpus to analyse the reliability of the genre classifier on the MaCoCu datasets. Based on that, the genre classifier reached macro F1 of 0.73 and micro F1 of 0.88. Analysis showed that we can eliminate some noisy predictions by removing texts, annotated as *Forum* and *Other*, and texts, predicted with low confidence level, obtained from the raw output. As the main goal of this study is to analyse global differences between MaCoCu datasets, we decided to remove less reliably predicted instances, as described in Section 3.1, to perform comparison only on the most reliable data. With this intervention, while sacrificing the model's coverage a bit, we obtained a much higher classifier's performance, reaching 0.92 in terms of micro and macro F1 score.

We applied the genre classifier to each of the seven English datasets from the parallel MaCoCu corpora. Prediction took approximately 6 hours per 100,000 texts which amounted to around 35 hours on one NVIDIA V100 GPU. Afterwards, we post-processed the data, discarding noisy genre predictions. In the next section, we compare the resulting datasets in terms of English variety and genre distribution.

[9] https://huggingface.co/classla/xlm-roberta-base-multilingual-text-genre-classifier

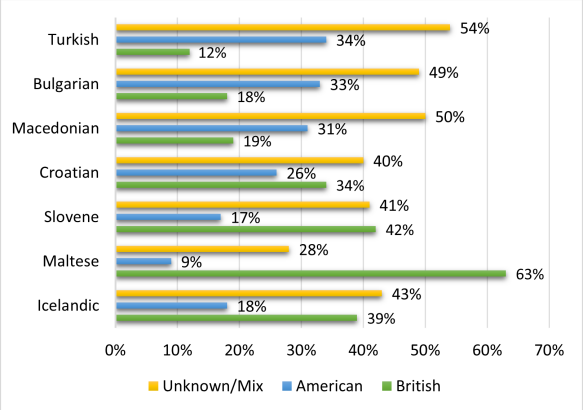

Figure 1: Distribution of American and British English in the English parts of the Icelandic, Maltese, Slovene, Croatian, Macedonian, Bulgarian and Turkish parallel web corpora.

5 Results

5.1 English Variety Distribution

By using our American-British variety classifier, more specifically, the unbalanced version, we identified the predominant English variety in each English text in the MaCoCu parallel corpora. If there were equal amounts of American and British variety-specific words in a text, the text was annotated as a "mix", and if there were no variety-specific words, the text was labeled as "unknown". The results, presented in Figure 1, show the distribution of British and American English in analysed corpora. The analysis shows that a variety was identified in mostly over 50% of texts in a corpus. Rather large amounts of unlabeled texts are not surprising, because most of the texts are quite short, with the median length of 170 to 300 words.

Figure 1 also shows that web corpora, obtained from different national top-level web domains, display different preference towards British and American English variety. The Maltese corpus was shown to have an overwhelming preference towards British English, with 63% texts classified as British, and only 9% classified as American. One of possible reasons for a strong influence of British English is Malta's close connection to the United Kingdom. The country is a former British colony and a member of the Commonwealth of Nations (Busuttil and Briguglio, 2023). Secondly, an inspection of the most frequent domains in the Maltese corpus revealed that half of the 10 most frequent domains are websites from the European Union, e.g. *europarl.europa.eu, eur-lex.europa.eu,*

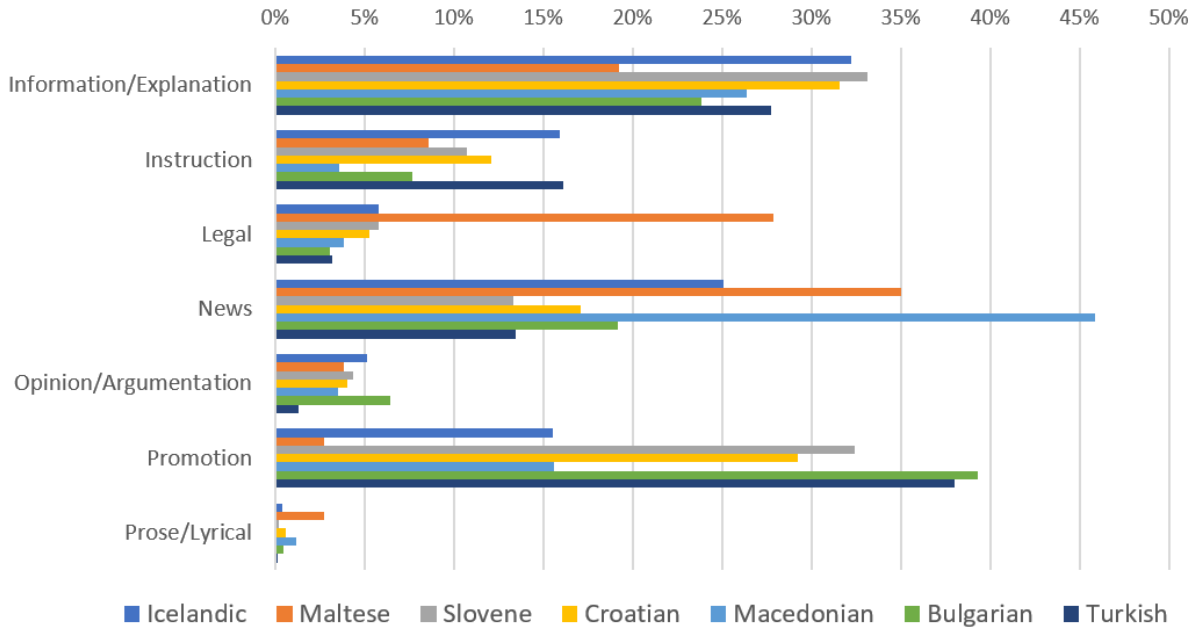

Figure 2: Distribution of genres in Icelandic-, Maltese-, Slovene-, Croatian-, Macedonian-, Bulgarian- and Turkish-English MaCoCu parallel web corpora.

ec.europa.eu, etc., covering 43% of all texts from the corpus. As the translation services in European Union have a preference towards British English (see Forsyth and Cayley (2022)), large amounts of EU texts in the corpus surely impacted the English variety distribution in it. The predominance of British English was also observed in the case of Icelandic, Slovene and Croatian corpora. In contrast, the corpora from the web domains of the countries further to the East, namely Macedonian, Bulgarian and Turkish corpora, show a much bigger influence of American English.

5.2 Genre Distribution

To obtain genre information, we applied the X-GENRE classifier to each text in the English part of the MaCoCu parallel corpora. The analysis of genre distribution, shown in Figure 2, revealed interesting differences between the corpora. The results show that the category *Information/Explanation* is notably present in all corpora, covering 20–30% of all texts. Other two predominant categories are *News* and *Promotion*, mostly covering 15–45% of texts. *News* is especially present in the Macedonian corpus, where it amounts to almost half of all texts, followed by Maltese and Icelandic with 25–35% of texts of this genre. In contrast, *Promotion* represents only up to 15% of texts in these three corpora, while it is much more frequent in Slovene, Croatian, Bulgarian and Turkish corpora, representing 30–40% of texts.

Other genre categories are generally less frequent. *Instruction* represents 5–15% of texts, with the highest frequency in Icelandic and Turkish. *Legal* represents around 5% of corpora. However, legal texts represent 28% of all texts in the Maltese corpus, showing this corpus to be significantly different than the others based on genre distribution as well. *Opinion/Argumentation* is more or less equally represented in all corpora, representing around 5% of texts. This category is the least represented in the Turkish corpus, with only 1% of texts. The least frequent category is *Prose/Lyrical*, representing 0.2–3% of texts, with the largest distribution in the Maltese corpus.

5.3 Genre Distribution in English Varieties

To obtain more information on the interplay of genres and English varieties, we looked at the average distribution of English varieties in each genre across all corpora. The results, shown in Figure 3, reveal that *News* texts and *Legal* texts from the analysed corpora are in average much more frequently written in British English, representing twice as much texts as the texts of these genres written in American English. *News* and *Legal* texts represent 60% of texts in the Maltese corpus, which also provides some explanation on why the Maltese corpus contains so much more British English than the others. In contrast, the category

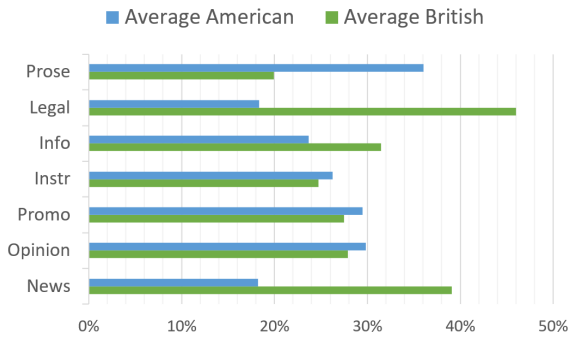

Figure 3: Average distribution of British and American varieties in each genre over all seven corpora. The abbreviated labels represent the following categories: Info – Information/Explanation, Promo – Promotion, Opinion – Opinion/Argumentation, Instr – Instruction, Prose – Prose/Lyrical.

Prose/Lyrical was shown to be more frequently written in American English. An inspection of the domains of the *Prose/Lyrical* texts revealed that in most corpora, a large majority of *Prose/Lyrical* texts come from American religious websites, such as *www.biblegateway.com* and *www.jw.org*, which explains the predominance of American English in this genre. In other genres, namely, *Information/Explanation, Instruction, Promotion* and *Opinion/Argumentation*, the two varieties are more or less equally present.

6 Conclusion

In this paper, we introduce a freely-available English variety classifier for fast and reliable identification of British and American English. The corpus-based approaches to language variety classification can be impacted by topic-related or other biases, occurring due to differences between the corpora on which the model is trained. In contrast, our lexicon-based approach is based on a carefully selected lexicon of words which are confirmed by linguists to be variety-specific, making the results more reliable and explainable. We then show how the classifier can be used to obtain an insight into the characteristics of large parallel corpora, collected with automatic methods. We compare parallel web corpora from European national webs in seven languages. As all languages are paired with English, we obtained meaningful information on the differences between the corpora in terms of English varieties. The results revealed British English is prevalent in Maltese, Icelandic, Slovene and Croatian corpora, while corpora from the Macedonian, Bulgarian and Turkish national webs are more influenced by American English. A stark difference between the use of varieties was observed in the case of the Maltese corpus, where a large majority of texts were written in British English and there were less than 10% of texts in American English. These results reflect the country's historical connection with the United Kingdom, along with a significant presence of EU websites in the corpus, which have a policy of preferring British English. Thus, we show how the classifier can be used for not only comparing corpora, but also obtaining insight into the use of English by native and non-native speaking content writers and translators. By making the classifier freely available, we hope to encourage analyses of the use of English and its varieties among teachers, translators and content creators in the fields of corpus linguistics, translation studies, linguistics and digital humanities.

Furthermore, we extend the comparison by automatically annotating the English texts from the parallel corpora with genre information. The results revealed significant differences between the corpora in terms of genre distribution. Once again, the Maltese corpus was shown to be more different than the others, consisting mostly of *News* and *Legal* texts. *News* is also strongly present in Macedonian and Icelandic corpora, while Slovene, Croatian, Bulgarian and Turkish corpora constitute of large amounts of promotional texts.

With the two classification approaches, we obtained valuable information on the characteristics of the datasets. As such datasets are often used for creation of machine translation systems, various NLP tools, as well as linguistic studies, it is crucial that the users are provided with the information on what types of texts and language varieties the datasets consist of. The MaCoCu project will provide this information for all their datasets, covering 13 European under-resourced languages: Albanian, Bosnian, Bulgarian, Catalan, Croatian, Icelandic, Macedonian, Maltese, Montenegrin, Slovene, Serbian, Turkish and Ukrainian. The datasets will be made freely available by June 2023. As the initial analysis of the English variety and genre distribution in corpora, presented in this paper, revealed that this information highlights important differences between the corpora, in the future, we plan to extend the analysis to all 13 newly available MaCoCu corpora. Furthermore, one important downstream task that we did not tackle in this work

is the inspection of the impact of the variation in variety and genre on machine translation and other systems based on these and other datasets, which we also plan to analyse in future studies.

Limitations

In this paper, we describe how we devised a lexicon-based classifier for American and British English. We argue for the lexicon-based approach as a better alternative to the corpus-based approach, as it is rule-based and explainable. However, we are aware that a lexicon-based approach is less feasible or impossible for classification of varieties of other languages or identification between languages. While the corpora-based approaches can be performed on all languages where at least one corpus of appropriate size exists, this approach requires an availability of a lexicon or at least linguistic rules on which a new lexicon needs to be based.

Secondly, by using the lexicon-based approach, we prefer reliability over coverage. If no variety-specific word is present in the analysed text, the text is left unlabeled. This was the case for 30 to 50% of texts in our analysed corpora. Furthermore, our lexicon is based on words only, and does not take account of variety-specific multiword expressions. Consequently, one should be aware that the findings reflect only the characteristics of the texts that were long enough and had any variety-specific word. Furthermore, while the corpora were collected by crawling the national web domains, there might exist texts on the web that were deliberately or not left out of the final datasets. This means that the nature of these corpora does not necessarily reflect the English variety distribution of all texts found on a national web.

Thirdly, in this research, we limit ourselves to the two most recognized varieties of the Standard English. We are aware that numerous other varieties from throughout the world exist. As this analysis has been done on texts from non-native English-speaking European countries, we consider that focusing on the two varieties which are often considered to be the main varieties is appropriate, albeit simplistic. However, we are aware that some of British or American-specific words might overlap with words that are also typical for other English varieties, such as Australian, Canadian, Irish, etc., and could for instance classify Irish English as British. We are aware that our pragmatic approach could be regarded as discriminatory towards other English varieties. While our classifier can be used on any English text, we should be aware that it solely provides information on the frequency of words, defined to be British or American. We leave discussions whether these texts are by that truly British, or whether we are talking about European English with British influence to the linguists, as we are aware that defining how many English varieties are there and what are their key differences is outside of our expertise.

Finally, in contrast to the English variety classifier which can be used only for English, the genre classifier is multilingual and covers all of the languages, included in the XLM-RoBERTa language model (Conneau et al., 2020). On the other hand, while the English variety classifier does not require massive computational resources, genre identification requires the use of a GPU. We are aware that not everyone is privileged to have access to such computational resources to be able to reproduce our research.

Ethics Statement

We are aware that collecting texts from the web can raise questions of respecting the intellectual property and privacy rights of the original authors of the texts. The web corpora, analysed in this paper, have been collected by crawling the national top-level domains. To assure that no sensitive data would be included, only texts that have been freely accessible were included in the corpora. We are aware that the datasets might still include some texts that the authors do not consent to be included. To mitigate this, the datasets are published with a notice, which informs the authors of the text that the texts can be taken out of the corpora upon their request. Secondly, for privacy issues, the sentences in the published corpora that contain personal information are flagged, so that the corpora users can leave them out of their research if the nature of their study would reveal this information. In our paper, we look into and report on the overall characteristics of the texts and do not examine texts more closely or produce systems which could abuse personal information or intellectual property rights. That is why anonymisation or additional filtering was not necessary.

Secondly, as mentioned in Limitations, our English variety classifier labels a text to be British or English based on the counts of variety-specific words. While it is a useful tool for quick inspec-

tion of the differences in English between various corpora, it is meant to be used on English texts, produced by non-native English speakers. As the British and American-specific words it detects could overlap with other English varieties, such as Irish, Australian, Canadian, Indian etc., one should not use it with the intention of belittling other varieties or proving that the entire world uses only the two mentioned varieties.

Acknowledgements

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), the research project "Basic Research for the Development of Spoken Language Resources and Speech Technologies for the Slovenian Language" (J7-4642), and the research programme "Language resources and technologies for Slovene" (P6-0411).

References

Kevin Atkinson and Benjamin Titze. 2020. Variant Conversion (VarCon). http://wordlist.aspell.net/varcon/.

ES Atwell, Junaid Arshad, Chien-Ming Lai, Lan Nim, N Rezapour Ashregi, Josiah Wang, and Justin Washtell. 2007. Which English dominates the world wide web, British or American? In *Proceedings of CL'2007 Corpus Linguistics Conference*. UCREL, Lancaster University.

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. 2022a. Bulgarian-English parallel corpus MaCoCu-bg-en 1.0. Slovenian language resource repository CLARIN.SI.

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. 2022b. Croatian-English parallel corpus MaCoCu-hr-en 1.0. Slovenian language resource repository CLARIN.SI.

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. 2022c. Icelandic-English parallel corpus MaCoCu-is-en 1.0. Slovenian language resource repository CLARIN.SI.

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. 2022d. Macedonian-English parallel corpus MaCoCu-mk-en 1.0. Slovenian language resource repository CLARIN.SI.

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. 2022e. Maltese-English parallel corpus MaCoCu-mt-en 1.0. Slovenian language resource repository CLARIN.SI.

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. 2022f. Slovene-English parallel corpus MaCoCu-sl-en 1.0. Slovenian language resource repository CLARIN.SI.

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. 2022g. Turkish-English parallel corpus MaCoCu-tr-en 1.0. Slovenian language resource repository CLARIN.SI.

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, et al. 2022h. Macocu: Massive collection and curation of monolingual and bilingual data: focus on under-resourced languages. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 301–302.

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. *Language resources and evaluation*, 43(3):209–226.

Salvino Busuttil and Lino Briguglio. 2023. Malta. https://www.britannica.com/place/Malta. Encyclopaedia Britannica.

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.

BNC Consortium et al. 2007. British national corpus. *Oxford Text Archive Core Collection*.

Paul Cook and Graeme Hirst. 2012. Do Web Corpora from Top-Level Domains Represent National Varieties of English? In *Proceedings of the 11th International Conference on Textual Data Statistical Analysis*, pages 281–293.

Mark Davies. 2013. Corpus of Global Web-Based English. https://www.english-corpora.org/glowbe/.

Mark Davies. 2016. Corpus of News on the Web (NOW). https://www.english-corpora.org/now/.

Jonathan Dunn. 2019. Modeling Global Syntactic Variation in English Using Dialect Classification. *NAACL HLT 2019*, 660:42.

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.

Julia Forsberg, Susanne Mohr, and Sandra Jansen. 2019. "The goal is to enable students to communicate": Communicative competence and target varieties in TEFL practices in Sweden and Germany. *European Journal of Applied Linguistics*, 7(1):31–60.

Angus Forsyth and Mireille Cayley, editors. 2022. *English Style Guide: A handbook for authors and translators in the European Commission*. European Commission.

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.

Albert Sydney Hornby. 1995. *Oxford Advanced Learner's Dictionary of Current English*. Oxford, England: Oxford University Press.

Adam Kilgarriff and Adam Kilgarri. 2001. Comparing corpora. In *International Journal of Corpus Linguistics*. Citeseer.

Taja Kuzman, Nikola Ljubešić, and Senja Pollak. 2022a. Assessing Comparability of Genre Datasets via Cross-Lingual and Cross-Dataset Experiments. In *Jezikovne tehnologije in digitalna humanistika: zbornik konference*, Jezikovne tehnologije in digitalna humanistika: zbornik konference, page 100–107. Institute of Contemporary History.

Taja Kuzman, Peter Rupnik, and Nikola Ljubešić. 2022b. 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, Anna Salmela, Samuel Rönnqvist, Alham Fikri Aji, Li-Hsin Chang, Asma Dhifallah, Larissa Goulart, Henna Kortelainen, Marc Pàmies, Deise Prina Dutra, et al. 2022. Towards better structured and less noisy web data: Oscar with register annotations. In *Proceedings of the Eighth Workshop on Noisy User-generated Text (W-NUT 2022)*, pages 215–221.

Marco Lui and Paul Cook. 2013. Classifying English documents by national dialect. In *Proceedings of the Australasian Language Technology Association Workshop 2013 (ALTA 2013)*, pages 5–15.

Marco Lui, Ned Letcher, Oliver Adams, Long Duong, Paul Cook, and Timothy Baldwin. 2014. Exploring methods and resources for discriminating similar languages. In *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects*, pages 129–138.

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

Randolph Quirk. 2014. *Grammatical and lexical variance in English*. Routledge.

Francisco Rangel, Paolo Rosso, Martin Potthast, and Benno Stein. 2017. Overview of the 5th author profiling task at PAN 2017: Gender and language variety identification in Twitter. *Working notes papers of the CLEF*, 48.

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.

Dmitri Roussinov, Kevin Crowston, Mike Nilan, Barbara Kwasnik, Jin Cai, and Xiaoyong Liu. 2001. Genre based navigation on the web. In *Proceedings of the 34th annual Hawaii international conference on system sciences*, pages 10–pp. IEEE.

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

Serge Sharoff. 2021. Genre annotation for the web: text-external and text-internal perspectives. *Register studies*.

Serge Sharoff, Zhili Wu, and Katja Markert. 2010. The Web Library of Babel: evaluating genre collections. In *LREC*. Citeseer.

Vasiliki Simaki, Panagiotis Simakis, Carita Paradis, and Andreas Kerren. 2017. Identifying the authors' national variety of English in social media text. Association for Computational Linguistics.

Jade Goldstein Stewart and J Callan. 2009. *Genre oriented summarization*. Ph.D. thesis, Carnegie Mellon University, Language Technologies Institute, School of Computer Science.

Jörg Tiedemann and Nikola Ljubešić. 2012. Efficient discrimination between closely related languages. In *Proceedings of COLING 2012*, pages 2619–2634.

Muhammad Romi Ario Utomo and Yuliant Sibaroni. 2019. Text classification of British English and American English using support vector machine. In *2019 7th International Conference on Information and Communication Technology (ICoICT)*, pages 1–6. IEEE.

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)*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

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.

Marcos Zampieri, Preslav Nakov, and Yves Scherrer. 2020. Natural language processing for similar languages, varieties, and dialects: A survey. *Natural Language Engineering*, 26(6):595–612.

Marcos Zampieri, Liling Tan, Nikola Ljubešić, and Jörg Tiedemann. 2014. A report on the DSL shared task 2014. In *Proceedings of the first workshop on applying NLP tools to similar languages, varieties and dialects*, pages 58–67.

A Appendix

A.1 Genre Categories

Label	Description	Examples
Information/Explanation	An objective text that describes or presents an event, a person, a thing, a concept etc. Its main purpose is to inform the reader about something.	research article, encyclopedia article, product specification, course materials, biographical story/history.
Instruction	An objective text which instructs the readers on how to do something.	how-to texts, recipes, technical support
Legal	An objective formal text that contains legal terms and is clearly structured.	small print, software license, terms and conditions, contracts, law, copyright notices
News	An objective or subjective text which reports on an event recent at the time of writing or coming in the near future.	news report, sports report, police report, announcement
Opinion/Argumentation	A subjective text in which the authors convey their opinion or narrate their experience. It includes promotion of an ideology and other non-commercial causes.	review, blog, editorial, letter to editor, persuasive article or essay, political propaganda
Promotion	A subjective text intended to sell or promote an event, product, or service. It addresses the readers, often trying to convince them to participate in something or buy something.	advertisement, e-shops, promotion of an accommodation, promotion of company's services, invitation to an event
Prose/Lyrical	A literary text that consists of paragraphs or verses. A literary text is deemed to have no other practical purpose than to give pleasure to the reader. Often the author pays attention to the aesthetic appearance of the text. It can be considered as art.	lyrics, poem, prayer, joke, novel, short story

Table 3: Descriptions of genre labels, with examples.