Creating register sub-corpora for the Finnish Internet Parsebank

Veronika Laippala[1,2,3], **Juhani Luotolahti**[3],
Aki-Juhani Kyröläinen[2,3], **Tapio Salakoski**[3], **Filip Ginter**[3]
[1] Turku Institute for Advanced Studies, University of Turku, Finland
[2] School of Languages and Translation Studies, University of Turku, Finland
[3] Turku NLP Group, University of Turku, Finland
first.last@utu.fi

Abstract

This paper develops register sub-corpora for the Web-crawled Finnish Internet Parsebank. Currently, all the documents belonging to different registers, such as news and user manuals, have an equal status in this corpus. Detecting the text register would be useful for both NLP and linguistics (Giesbrecht and Evert, 2009) (Webber, 2009) (Sinclair, 1996) (Egbert et al., 2015). We assemble the sub-corpora by first naively deducing four register classes from the Parsebank document URLs and then developing a classifier based on these, to detect registers also for the rest of the documents. The results show that the naive method of deducing the register is efficient and that the classification can be done sufficiently reliably. The analysis of the prediction errors however indicates that texts sharing similar communicative purposes but belonging to different registers, such as news and blogs informing the reader, share similar linguistic characteristics. This attests of the well-known difficulty to define the notion of registers for practical uses. Finally, as a significant improvement to its usability, we release two sets of sub-corpus collections for the Parsebank. The A collection consists of two million documents classified to blogs, forum discussions, encyclopedia articles and news with a naive classification precision of >90%, and the B collection four million documents with a precision of >80%.

1 Introduction

The Internet offers a constantly growing source of information, not only in terms of size, but also in terms of languages and communication settings it includes. As a consequence, *Web corpora*, language resources developed by automatically crawling the Web, offer revolutionary potentials for fields using textual data, such as Natural Language Processing (NLP), linguistics and other humanities (Kilgariff and Grefenstette, 2003).

Despite their potentials, Web corpora are underused. One of the important reasons behind this is the fact that in the existing Web corpora, all of the different documents have an equal status. This complicates their use, as for many applications, knowing the composition of the corpus would be beneficial. In particular, it would be important to know what *registers*, i.e. text varieties such as a user manual or a blog post, the corpus consists of (see Section 2 for a definition). In NLP, detecting the register of a text has been noted to be useful for instance in POS tagging (Giesbrecht and Evert, 2009), discourse parsing (Webber, 2009) and information retrieval (Vidulin et al., 2007). In linguistics, the correct constitution of a corpus and the criteria used to assemble it have been subject to long discussions (Sinclair, 1996), and Egbert & al. (2015) note that without systematic classification, Web corpora cannot be fully benefited from.

In this paper, we explore the development of register sub-corpora for the Finnish Internet Parsebank[1], a Web-crawled corpus of Internet Finnish. We assemble the sub-corpora by first naively deducing four register classes from the Parsebank document URLs and then creating a classifier based on these classes to detect texts representing these registers from the rest of the Parsebank (see Section 4). The register classes we develop are news, blogs, forum discussions and encyclopedia articles. Instead of creating a full-coverage taxonomy of all the registers covered by the Parsebank, in this article our aim is to test this method in

[1] http://bionlp.utu.fi

Proceedings of the 21st Nordic Conference of Computational Linguistics, pages 152–161,
Gothenburg, Sweden, 23-24 May 2017.

the detection of these four registers. If the method works, the number of registers will be extended in future work.

In the register detection and analysis, we compare three methods: the traditional bag-of-words as a baseline, lexical trigrams as proposed by Gries & al. (2011), and Dependency Profiles (DP), co-occurrence patterns of the documents labelled in a specific class, assumed a register, and dependency syntax relations.

In addition to reporting the standard metrics to estimate the classifier performance, we evaluate the created sub-corpora by analysing the mismatches between the naively assumed register classes and the classifier predictions. In addition, we analyse the linguistic register characteristics estimated by the classifier. This validates the quality of the sub-corpora and is informative about the linguistic variation inside the registers (see Section 5).

Finally, we publish four register-specific sub-corpora for the Parsebank that we develop in this paper: blogs, forum discussions, encyclopedia articles and news (see Section 6). We release two sets of sub-corpora: the A collection consists of two million documents with register-specific labels. For these documents, we estimate the register prediction precision to be >90%. The collection B consists of four million documents. For these, the precision is >80%. These sub-corpora allow the users to focus on specific registers, which improves the Parsebank usability significantly (see discussions in (Egbert et al., 2015) and (Asheghi et al., 2016)).

2 Previous studies

Since the 1980s, linguistic variation has been studied in relation to the communicative situation, form and function of the piece of speech or writing under analysis (Biber, 1989; Biber, 1995; Biber et al., 1999; Miller, 1984; Swales, 1990). Depending on the study, these language varieties are usually defined as *registers* or *genres*, the definitions emphasising different aspects of the variation (see discussion in (Asheghi et al., 2016; Egbert et al., 2015). We adopt the term register and define it, following (Biber, 1989; Biber, 1995; Egbert et al., 2015), as a text variety with specific situational characteristics, communicative purpose and lexico-grammatical features.

Studies aiming at automatically identifying registers from the Web face several challenges. Although some studies reach a very high accuracy, their approaches are very difficult to apply in real-world applications. Other studies, adopting a more realistic approach, present a weaker performance. In particular, the challenges are related to the definition of registers in practice: how many of them should there be, and how to reliably identify them? In addition, it is not always clear whether registers have different linguistic properties (Schäfer and Bildhauer, 2016). Based on the situational characteristics of a register, a blog post discussing a news topic and a news article on the same topic should be analysed as different registers. But how does this difference show in the linguistic features of the documents, or does it?

For instance, Sharoff & al. (2010) achieve an accuracy of 97% using character tetragrams and single words with a stop list as classification features, while Lindemann & Littig (2011) report an F-score of 80% for many registers using both structural Web page features and topical characteristics based on the terms used in the documents. They, however, use as corpora only samples of the Web, which can represent only a limited portion of all the registers of the entire Web (Sharoff et al., 2010; Santini and Sharoff, 2009).

Another, more linguistically motivated perspective to study Web registers is adopted by Biber and his colleagues. Using typical end users of the Web to code a large number of nearly random Web documents (48 000) with hierarchical, situational characteristics, they apply a bottom-up method for creating a taxonomy of Web registers (Biber et al., 2015; Egbert et al., 2015). Then, applying a custom built tagger identifying 150+ lexico-grammatical features, they report an overall accuracy of 44.2% for unrestricted Web texts using a taxonomy of 20 registers (Biber and Egbert, 2015). In addition to the relatively weak register identification performance, their approach suffers from a low interannotator agreement for the register classes. Similar problems are also discussed in (Crowston et al., 2011; Essen and Stein, 2004), who note that both experts and end users have troubles identifying registers reliably. This leads to question, whether register identification can at all be possible, if even humans cannot agree on their labelling. This concern is expressed by Schäfer and Bildhauer (2016), who decide to focus on classifying their *COW Corpora*

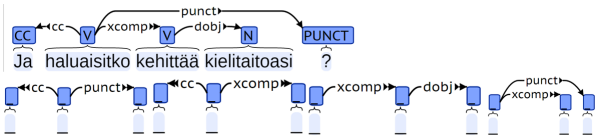

Figure 1: Unlexicalised syntactic biarcs from the sentence *Ja haluaisitko kehittää kielitaitoasi?* 'And would you like to improve your language skills?'

to topic domains, such as *medical* or *science* instead of registers. The recently presented Leeds Web Genre Corpus (Asheghi et al., 2016) shows, however, very reliable interannotator agreement scores. This proves that when the register taxonomy is well developed, the registers can as well be reliably identified.

3 Finnish Internet Parsebank

Finnish Internet Parsebank (Luotolahti et al., 2015) is a Web-crawled corpus on the Finnish Internet. The corpus is sampled from a Finnish Web-crawl data. The crawl itself is produced using SpiderLing Crawler, which is especially crafted for efficient gathering of unilingual corpora for linguistic purposes. The version we used is composed of 3.7 billion tokens, 6,635,960 documents and has morphological and dependency syntax annotations carried out with a state-of-the-art dependency parser by Bohnet (2010), with a labelled attachment score of 82.1% (Luotolahti et al., 2015). The Parsebank is distributed via a user interface at `bionlp-www.utu.fi/dep_search/` and as a downloadable, sentence-shuffled version at `bionlp.utu.fi`.

4 Detecting registers from the Parsebank

In this Section, we first discuss the development of the naive register corpora from the Parsebank. These will be used as training data for the system identifying the registers from the entire Parsebank. We then motivate our selection of features in the classifier development, and finally, we present the classification results.

4.1 Naive registers as training data

Our naive interpretation of the document registers was based on the presence of lexical cues in the Parsebank document URLs. For the purposes of this article, we used four well-motivated register classes: news, blogs, encyclopedia and forum discussions. These were identified by the presence of the following keywords in the URL: *lehti* 'newspaper', *uutis* 'news', *uutinen* 'news' or *news* for the news class; *wiki* for the encyclopedia class; *blog* for the blog class; and *discussion, forum* or *keskustelu* 'discussion' for the discussion class.

Naive register	Entire PB	Training subset
Blogs	775,885	8,364
Forum discussions	307,797	3,076
Encyclopedia	127,884	1,580
News	771,058	13,197
All	**1,982,624**	**26,217**

Table 1: Total number of documents in the naively assembled register classes and in the subset used in SVM training

In deciding the registers to be searched for, we aimed at a simple, experimental solution that would be informative about the performance of the naive method and offer direct application potentials for the Parsebank to increase its usability. Therefore, instead of creating a full-coverage taxonomy of all registers possibly found online, our aim here was to experiment with a few generally acknowledged, broad-coverage terms for register classes. Once we can in this paper show that the naive method works, the number of the registers will be expanded in future work.

Table 1 presents the proportion of the naively assumed registers in the entire Parsebank and in the subset we use for the classifier training in Section 4.3. The sizes of the retrieved sub-corpora vary significantly. The most frequent, news and blogs, cover more than 10% of the Parsebank documents, respectively, while in particular the encyclopedia corpus remains smaller. Still, the sizes of these classes are relatively large, thanks to the size of the entire Parsebank.

The subset used for the classifier training is created by matching the document URLs of 100,000 first documents from the Parsebank. Of these, 26,216 had a URL that matched one of the keywords defined above. At this stage, all sentence-level duplicates were also removed from the train-

ing data to prevent the classifier from learning elements that are often repeated in Web pages, such as *Lue lisää* 'Read more' but should not be used as classifier features. This proved to be highly needed, as from the 3,421,568 sentences in the 100,000 documents, 497,449 were duplicates.

4.2 Lexical and syntactic approaches to model register variation

The work by Biber and colleagues on the variation of lexico-grammatical features across registers has been very influential in corpus linguistics over the years (Biber, 1995; Biber et al., 1999). Recently, they have extended their work to Web registers and applied the carefully tuned Biber tagger identifying both lexical and grammatical features to explore registers from the Web (Biber et al., 2015; Egbert et al., 2015). Gries & al. (2011) adopt an opposite approach by using simple word-trigrams. Sharoff and colleagues compare a number of different feature sets and conclude that bag-of-words and character n-grams achieve the best results (Sharoff et al., 2010). For detecting mostly thematic domains, Schäfer and Bildhauer (2016) apply lexical information attached to coarse-grained part-of-speech labels.

We compare three methods for detecting the four registers represented by our naively assembled sub-corpora. As a baseline method, we apply the standard bag-of-words, which, despite its simplicity, has often achieved a good performance (Sharoff et al., 2010). Second, we use word-trigrams similar to Gries & al. (2011), and finally, Dependency Profiles (DPs), which are co-occurrences of the register documents with unlexicalised syntactic biarcs, three-token subtrees of dependency syntax analysis with the lexical information deleted (Kanerva et al., 2014) (see Figure 1). As opposed to e.g. keyword analysis (Scott and Tribble, 2006) based on the document words, DPs do not restrict to the lexical or topical aspects of texts, and thus offer linguistically better motivated analysis tools. Many studies on register variation highlight the importance of syntactic and grammatical features (Biber, 1995; Biber et al., 1999; Gries, 2012). Therefore, we hypothesise that DPs would allow to generalise beyond individual topics to differentiate, e.g., between texts representing different registers but discussing similar topics, such as news and forum discussions on sports.

4.3 Classifier development and testing

To predict the registers for all the Parsebank documents, we trained a linear SVM. In the training and testing, we used a subset of the Parsebank described in Table 1. As features, we used the sets described in the previous Section. Specifically, the four register classes were modelled as a function of the feature sets, i.e. a co-occurence vector of the used features across the documents. These vectors were L2-normalised and then used to model the register class of a given document by fitting a linear SVM to the data, as implemented in the Scikit package[2] in Python. To validate the performance of the fitted model, we implemented a 10-fold cross-validation procedure with stratified random subsampling to keep the proportion of the register classes approximately equal between the training and test sets.

The results of the SVM performance are described in Table 2. First, the best results are achieved with the bag-of-words and lexical n-gram approaches with an F-score of 80 % and 81%. This already confirms that the registers can be identified and that our naive method of assuming the registers based on the URLs is justified.

Second, although the DPs consisting of syntactic biarcs would allow for detailed linguistic examinations of the registers, and even if they follow the influential work by Biber and colleagues on modelling registers, their classification performance is clearly lower than those of the lexical approaches. The average F-score for the biarcs is only 72%. Interestingly, combining biarcs and the bag-of-words results in a very similar F-score of 79% and does not improve the classifier performance at all. In other words, three of the four feature sets attest very similar performances. This can suggest that the remaining 20% of the data may be somehow unreachable with these feature sets and requires further data examination, which we will present in Section 5.1 and 5.2.

Third, it is noteworthy that the classifier performance varies clearly across the registers. News and blogs receive the best detection rates, rising to the very reliable 86% and 79% F-score, respectively, while the discussion and encyclopedia article detection rates are clearly lower, 70% and 73%. Naturally, the higher frequency of blogs and news in the training and test set explains some of these differences. Still, these differences merit fur-

[2] `http://scikit-kearn.org/stable/`

	Blog	Discussion	Encyclopedia	News	Avg.
Bag-of-words					
Precision	0.80	0.66	0.69	0.86	0.81
Recall	0.76	0.73	0.74	0.86	0.80
F-score	0.78	0.69	0.71	0.86	0.80
Biarcs					
Precision	0.70	0.50	0.43	0.83	0.73
Recall	0.66	0.47	0.60	0.83	0.72
F-score	0.68	0.48	0.51	0.82	0.72
Bag-of-words + biarcs					
Precision	0.78	0.63	0.65	0.86	0.80
Recall	0.75	0.69	0.73	0.85	0.79
F-score	0.76	0.66	0.69	0.85	0.79
Uni- bi - trigrams					
Precision	0.81	0.68	0.71	0.86	0.81
Recall	0.77	0.73	0.75	0.86	0.81
F-score	0.79	0.70	0.73	0.86	0.81

Table 2: The SVM results achieved with the four feature sets.

ther analyses in future work.

Finally, the variation between the precision and recall rates across the registers requires closer examination. While the precision and recall are very similar for the news class, for the blogs the precision is higher than the recall. This suggests that some features, words in this case, are very reliable indicators of the blog register, but that they do are not present in all the class documents. For the discussion and encyclopedia classes, the recall is higher than the precision, indicating that such reliable indicators are less frequent.

5 Validating the classifier quality

The classifier performance seems sufficiently reliable to be applied for identifying the registers from the Parsebank. Before classifying the entire corpus, we will, however, in this Section seek answers to questions raised by the SVM results. First, we analyse the classifier decisions to find possible explanations for the 20% of the data that the SVM does not detect. This will also ensure the validity our naive method for assuming the register classes. Second, we study the most important register class features, words in our case, estimated by the SVM. These can explain the variation between the precision and recall across the registers revealed, and also further clarify the classifier's choices and the classification quality.

5.1 Mismatches between the SVM predictions and the naively assumed registers

Mismatches between the SVM predictions and the naively assumed register labels are informative both about the SVM performance and about the coherence of the naive corpora: a mismatch can occur either because the classifier makes a mistake or because the document, in fact, does not represent the register its URL implies. This can also explain why the classifier results achieved with different feature sets were very similar.

We went manually through 60 wrongly classified Parsebank documents that did not belong to the subset on which the SVM was trained on. Although the number of documents was not high, the analysis revealed clear tendencies on the classification mismatches and the composition of the naively presumed registers.

Above all, the analysis proved the efficiency of our naive method of assuming the register. The blog, encyclopedia and discussion classes included only one document, respectively, where the URL did not refer to the document register. The news class included more variation, as in particular the documents with *lehti* 'magazine' in the URL included also other registers than actual news. Of the 15 analysed documents naively presumed news, nine were actual news, four columns or editorials and two discussions.

For the mismatches between the naive register labels and the SVM predictions, our analysis showed that many could be explained with significant linguistic variation within the registers, both in terms of the communicative aim of the document and its style. For instance, some of the blogs we analysed followed a very informal, narrative model, while others aimed at informing the reader on a current topic, and yet others resembled advertisements with an intention of promoting or sell-

ing. Also the distinction between news and encyclopedia articles that both can focus on informing the reader was for some documents vague in terms of linguistic features. Similarly, some shorter blog posts and forum discussion posts appeared very similar.

Similar communicative goals thus seem to result in similar linguistic text characteristics across registers. In addition to explaining the mismatches between the SVM predictions and the naively assumed registers, this linguistic variation inside registers clarifies the SVM results and the fact that the performances of three of the four feature sets were very similar. On one hand, this could also suggest that the registers should be defined differently than we have currently done, so that they would better correspond to the communicative aims and linguistic characteristics of the texts. For instance, the register taxonomy proposed by Biber and colleagues (Biber et al., 2015; Egbert et al., 2015) with registers such as *opinion* or *informational persuasion* follows better these communicative goals and could, perhaps, result in better classification results. On the other hand, such denominations are not commonly known and the registers can be difficult to identify, as noted by Asheghi & al. (2016). Also, they can result in very similar texts falling to different registers. For instance in the taxonomy presented by Biber & al, *personal blogs, travel blogs* and *opinion blogs* are all placed in different registers.

5.2 The most important register features

To obtain a better understanding of the classifier's decisions and the features it bases its decisions on, we analysed the most important words of each register class as estimated by the SVM classifier based on the training corpus. These words can be seen as the *keywords* of these classes. In corpus linguistics, keyword analysis (Scott and Tribble, 2006) is a standard corpus analysis method. These words are said to be informative about the corpus topic and style. (See, however, also Guyon and Elisseeff (2003) and Carpena & al. (2009).) To this end, we created a frequency list of the 20 words which were estimated as the most important in each register class across the ten validation rounds. Table 3 presents, for each class, five of the ten most frequent words on this list that we consider the most revealing.

The most important words for each register class listed in Table 3 reveal clear tendencies that the classifier seems to follow. Despite the extraction of sentence-level duplicates presented in Section 3, the blog and forum discussion classes include words coming from templates and other automatically inserted phrases, such as *Thursday*, *anonymous* and the English words. Although these do not reveal any linguistic characteristics of the registers, they thus allow the classifier to identify the classes, and also explain the higher precision than recall reported in Section 2. Interestingly, Asheghi & al. (2016) report similar keywords for both blogs and discussions in English, which demonstrates the similarity of these registers across languages. In our data, the encyclopedia and news classes include words reflecting topics, such as *20-tuumaiset* '20-inch', and for instance verbs denoting typical actions in the registers, such as *kommentoi* 'comments'. These are more informative also on the linguistic characteristics of the registers and their communicative purposes.

6 Finnish Internet Parsebank with register sub-corpora

The classifier performance results reported in Section 4.3 and the analysis described in Section 5 proved that the developed classifier is sufficiently reliable to improve the usability of the Parsebank. In this Section, we apply the model to classify the entire Parsebank.

6.1 Detecting five register classes with a four-class SVM

We classified all the Parsebank documents with the bag-of-words feature set and parameters reported in Section 4.3. The SVM was developed to detect the four classes for which we had the training data thanks to the naive labels present in the document URLs. The addition of a negative class to the training data, with none of the labels in the URLs, would have increased significantly its noisiness, as these documents could still, despite the absence of the naive label, belong to one of the positive classes. Therefore, we needed to take some additional steps in the Parsebank classification, as the final classification should still include a fifth, negative class.

First, we ran the four-class classifier on all the Parsebank documents. In addition to the register labels, we also collected the scores for each regis-

Blogs	Forum discussions	Encyclopedia	News
kirjoitettu 'written'	*keskustelualue* 'discussion area'	*20-tuumaiset* '20-inch'	*kertoo* 'tells'
ihana 'wonderful'	*wrote*	*opiskelu* 'studying'	*aikoo* 'will'
kl. 'o'clock'	*nimetön* 'anonymous'	*wikiin* 'to the wiki'	*tutkijat* 'researchers'
torstai 'Thursday'	*ketjussa* 'in the thread'	*perustettiin* 'was founded'	*huomauttaa* 'notes'
archives	*forum*	*liitetään* 'is attached'	*kommentoi* 'comments'

Table 3: The most important features for each register class as estimated by the classifier. The original words are italicised and the translations inside quotations. Note that some words are originally in English.

Register	Proportion
Narrative	31.2%
Informational description	24,5%
Opinion	11.2%
Interactive discussion	6.4%
Hybrid	29.2%

Table 4: Register frequencies in the English Web, as reported by Biber & al. (2015)

ter, as assigned by the SVM, and sorted the documents based on these scores. Then, we counted a naive precision rate for the predictions by counting the proportion of the correct SVM predictions that matched the naive register label gotten from the URL. This gave us a sorted list of the Parsebank documents, where, in addition to the scores assigned by the classifier, we also have an estimate of the prediction precisions. From this sorted list, we could then take the best ranking ones that are the most likely to be correctly classified.

These estimated precisions for the documents descend from 1 for the most reliably classified documents to 0.74 for the least reliable ones. The question is, where to set the threshold to distinguish the documents that we consider as correctly predicted and those that we do not. As we do not know the distribution of registers in the Finnish Web, this is difficult to approximate. The study by Biber & al. (2015) on a large sample of Web documents reports the most frequent registers in English. These are described in Table 4. Our news and encyclopedia registers would most likely belong to the *informational* category, blogs to the *narrative* and Forum discussions naturally to the *interactive discussion* category. Very likely many could also be classified as *hybrid*. Based on these, we can estimate that the registers we have can cover a large proportion of the Finnish Web and of the Parsebank, in particular if we consider them as relatively general categories that can include a number of subclasses, similar to (Biber et al., 2015).

6.2 Collections A and B

To improve the Parsebank usability to the maximum, we decided to release two sets of sub-corpora: the A collection includes all the Parsebank documents with best-ranking scores assigned by the SVM, where the naive match precision threshold was set to 90%, and the B corpora where the threshold was set to 80%[3] This allows the users to choose the precision with which the register labels are correct.

The sizes of the sub-corpus collections are presented in Tables 5 and 6. The A collection consists of altogether 2 million documents classified to four registers. Of these, the URLs of nearly 800,000 documents match the SVM prediction, and more than a million do not have a naive label deduced from the URL. The news sub-corpus is clearly the largest covering nearly 50% of the total, blogs including 0.5 million documents.

In the B collection, the total number of documents rises to four million, which presents nearly 60% of the Parsebank. Similarly to the A collection, News and Blogs are the largest register-specific classes. In this version, the number of documents with mismatches between the classifier predictions and the naively assumed registers is evidently higher than in the A, and also the number of documents without any naive label is higher. This naturally implies a lower register prediction quality. Despite this, the B collection offers novel possibilities for researchers. It is a very large corpus, where the registers should be seen as upper-level, coarse-grained classes. In addition to offering register-specific documents, this collection can be seen as a less noisy version of the Parsebank, which is useful also when the actual registers are not central.

[3]These corpora will be put publicly available on the acceptance of this article.

Register	Register total	URL match	W/o naive label	Mismatch
Blogs	521,777	274,447	224,630	22,700
Forum discussions	326,561	124,971	168,416	33,174
Encyclopedia	204,019	59,596	132,670	11,753
News	966,938	325,138	609,500	32,300
A collection total	**2,019,295**	**784,152**	**1,135,216**	**99,927**

Table 5: Sizes of the register classes in the A collection

Register	Register total	URL match	W/o naive label	Mismatch
Blogs	1,122,451	450,202	604,877	67,372
Forum discussions	673,678	189,441	394,501	89,736
Encyclopedia	483,425	74,679	376,586	32,160
News	2,425,261	542,970	1,735,926	146,365
B collection total	**4,704,815**	**1,257,292**	**3,111,890**	**335,633**

Table 6: Sizes of the register-specific classes in the B collection

7 Conclusion

The aim of this article was to explore the development of register sub-corpora for the Finnish Internet Parsebank by training a classifier based on documents for which we had naively deduced the register by the presence of keyword matches in the document URL. We also experimented with several feature sets in detecting these registers and evaluated the validity of the created sub-corpora by analysing their linguistic characteristics and classifier prediction mistakes.

First of all, the results showed that our naive method of assuming the document registers is valid. Only the news class proved to include some documents belonging to other, although related, registers. Of the four feature sets we experimented on, the best classification performance was achieved with the bag-of-words and lexical trigram sets. The average F-score of 81% proved that the registers can be relatively reliably identified. In addition, the analysis of the classifier mistakes showed that texts with similar communicative purposes, such as news articles and blog posts that both aim at informing the reader, share linguistic characteristics. This complicates their identification, and attests of the challenges related to defining registers in practice, as already discussed in previous studies.

After validating the classifier performance and the quality of the naively assembled sub-corpora, we classified the entire Parsebank using the four-class model developed with the naive registers. To create a fifth, negative class for the documents not belonging to any of the four known registers, we sorted the documents based on the scores estimated by the SVM and counted a naive classification precision based on the proportion of the documents with matching naive register labels deduced from the URL and classifier predictions. This allowed us to establish a precision threshold, above which we can assume the document labels to be sufficiently reliably predicted. To improve the Parsebank usability, we release to sets of sub-corpora: the A collection includes two million documents classified to four register-specific corpora with a precision above 90%, and the B collection four million documents with a precision above 80%.

Naturally, this first sub-corpus release leaves many perspectives and needs for future work. More precisely and reliably defined register classes would further increase the usability of the sub-corpora. Also the number of available registers should be increased, as the *none* class currently includes still many registers. The naming of the registers and their inner variation would also merit further analyses to decide how to deal with linguistically similar texts that at least in our current system belong to different registers, such as different texts aiming at informing the reader.

Acknowledgements

This work has been funded by the Kone Foundation. Computational resources were provided by CSC - It center for science.

References

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

Douglas Biber and Jesse Egbert. 2015. Using grammatical features for automatic regisgter identification in an unrestricted corpus of documents from the Open Web. *Journal of Research Design and Statistics in Linguistics and Communication Science*, 2(1).

Douglas Biber, S. Johansson, G. Leech, Susan Conrad, and E. Finegan. 1999. *The Longman Grammar of Spoken and Written English*. Longman, London.

Douglas Biber, Jesse Egbert, and Mark Davies. 2015. Exloring the composition of the searchable web: a corpus-based taxonomy of web registers. *Corpora*, 10(1):11–45.

Douglas Biber. 1989. *Variation across speech and writing*. Cambridge University Press, Cambridge.

Douglas Biber. 1995. *Dimensions of Register Variation: A Cross-linguistic Comparison*. Cambridge University Press, Cambridge.

Bernd Bohnet. 2010. Very high accuracy and fast dependency parsing is not a contradiction. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, pages 89–97, Stroudsburg, PA, USA. Association for Computational Linguistics.

Pedro. Carpena, Pedro. Bernaola-Galván, Michael Hackenberg, Ana. V. Coronado, and Jose L. Oliver. 2009. Level statistics of words: Finding keywords in literary texts and symbolic sequences. *Physical Review E (Statistical, Nonlinear, and Soft Matter Physics)*, 79(3):035102.

Kevin Crowston, Barbara Kwaśnik, and Joseph Rubleske, 2011. *Genres on the Web: Computational Models and Empirical Studies*, chapter Problems in the Use-Centered Development of a Taxonomy of Web Genres, pages 69–84. Springer Netherlands, Dordrecht.

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.

S. Meyer Zu Essen and Barbara Stein. 2004. Genre classification of web pages: User study and feasibility analysis. *Proceedings of the 27th Annual German Conference on Artificial Intelligence*, pages 256–259.

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 *Web as Corpus Workshop (WAC5)*, pages 27–36.

Stephan Gries, John Newman, and Cyrus Shaoul. 2011. N-grams and the clustering of registers. *Empirical Language Research*.

Stephan Gries, 2012. *Methodological and analytic frontiers in lexical research*, chapter Behavioral Profiles: a fine-grained and quantitative approach in corpus-based lexical semantics. John Benjamins, Amsterdam and Philadelphia.

Isabelle M. Guyon and Andre Elisseeff. 2003. An introduction to variable and feature selection. *The journal of machine learning research*, 3:1157–1182.

Jenna Kanerva, Matti Luotolahti, Veronika Laippala, and Filip Ginter. 2014. Syntactic n-gram collection from a large-scale corpus of Internet Finnish. In *Proceedings of the Sixth International Conference Baltic HLT 2014*, pages 184–191. IOS Press.

Adam Kilgariff and Gregory Grefenstette. 2003. Introduction to the special issue on Web as Corpus. *Computational Linguistics*, 29(3).

Christoph Lindemann and Lars Littig, 2011. *Genres on the Web: Computational Models and Empirical Studies*, chapter Classification of Web Sites at Super-genre Level, pages 211–235. Springer Netherlands, Dordrecht.

Juhani Luotolahti, Jenna Kanerva, Veronika Laippala, Sampo Pyysalo, and Filip Ginter. 2015. Towards universal web parsebanks. In *Proceedings of the International Conference on Dependency Linguistics (Depling'15)*, pages 211–220. Uppsala University.

C.R. Miller. 1984. Genre as social action. *Quaterly journal of speech*, 70(2):151–167.

Marina Santini and Serge Sharoff. 2009. Web genre benchmark under construction. *JLCL*, 24(1):129–145.

Roland Schäfer and Felix Bildhauer, 2016. *Proceedings of the 10th Web as Corpus Workshop*, chapter Automatic Classification by Topic Domain for Meta Data Generation, Web Corpus Evaluation, and Corpus Comparison, pages 1–6. Association for Computational Linguistics.

Mike Scott and Chistopher Tribble. 2006. *Textual Patterns: keyword and corpus analysis in language education*. Benjamins, Amsterdam.

Serge Sharoff, Zhili Wu, and Katja Markert. 2010. The web library of Babel: evaluating genre collections.

John Sinclair. 1996. Preliminary recommendations on corpus typology.

John Swales. 1990. *Genre analysis: English in academic and research settings*. Cambridge University Press, Cambridge.

Vedrana Vidulin, Mitja Lustrek, and Matjax Gams. 2007. Using genres to improve search engines. In *Workshop "Towards genre-enabled Search Engines: The impact of NLP" at RANLP*, pages 45–51.

Bonnie Webber. 2009. Genre distinctions for discourse in the Penn treebank. In *Proceedings of the joint conference of the 47th annual meeting of the ACL and the 4th international joint conference on natural language processing of the AFNLP.*, pages 674–682. Association for Computational Linguistics.