Towards better structured and less noisy Web data: Oscar with Register annotations

Veronika Laippala[•] **Anna Salmela**[•] **Samuel Rönnqvist**[•] **Alham Fikri Aji**[○*]
Li-Hsin Chang[•] **Asma Dhifallah**[•] **Larissa Goulart**[‡] **Henna Kortelainen**[•]
Marc Pàmies[⋆] **Deise Prina Dutra**[◇] **Valtteri Skantsi**[•]
Lintang Sutawika[□] **Sampo Pyysalo**[•]

[•]University of Turku [○]Amazon [‡]Montclair State University
[⋆]Barcelona Supercomputing Center [◇]Universidade Federal de Minas Gerais [□]Datasaur.ai
[•]{mavela,annsaln,saanro}@utu.fi

Abstract

Web-crawled datasets are known to be noisy, as they feature a wide range of language use covering both user-generated and professionally edited content as well as noise originating from the crawling process. This article presents one solution to reduce this noise by using automatic register (genre) identification—whether the texts are, e.g., forum discussions, lyrical or how-to pages. We apply the multilingual register identification model by Rönnqvist et al. (2021) and label the widely used Oscar dataset. Additionally, we evaluate the model against eight new languages, showing that the performance is comparable to previous findings on a restricted set of languages. Finally, we present and apply a machine learning method for further cleaning text files originating from Web crawls from remains of boilerplate and other elements not belonging to the main text of the Web page. The register labeled and cleaned dataset covers 351 million documents in 14 languages and is available at `https://huggingface.co/datasets/TurkuNLP/register_oscar`.

1 Introduction

Massive Web-crawled datasets are widely used in Natural Language Processing (NLP), for instance for training language models (Conneau et al., 2020; Raffel et al., 2019; Xue et al., 2020). However, the challenge with these crawled datasets is that they are typically very noisy. First of all, this noise originates from the lack of structure and metadata—the datasets don't include any information on the origin of the documents. This complicates their use, because language on the Web varies extremely, ranging from toxic language, discussion forums and other user-generated content to professionally-like edited texts. Second, the noisiness comes from the crawling process—despite the cleaning efforts, Web-crawled data still contain remains of boilerplate and other elements not belonging to the main text, such as *click here* or *read more*. All these properties affect the automatic processing of text (Maharjan et al., 2018; Barbaresi, 2021; Kilgarriff, 2007).

The automatic identification of Web genres or registers—whether the documents are, e.g., forum discussions, originally spoken, informative or narrative (Biber and Conrad, 2019)— would offer a solution to reduce the noisiness of Web data and to add metadata on the origin of the documents. However, this has been a challenge. There are no gatekeepers ensuring that the users follow any conventions when writing on the Web, and thus, Web language use has been referred to as a jungle (Sharoff, 2010). The available register datasets, almost entirely focusing on English, have been restricted to only selected and well-defined registers, and they do not generalize to the entire Web (Sharoff et al., 2010; Asheghi et al., 2014; Santini, 2008; Madjarov et al., 2019).

Recently, however, the Corpus of Online Registers of English (CORE) (Egbert et al., 2015) sampled from the unrestricted open Web has allowed the modeling of the full range of registers found in Web-crawled datasets. Furthermore, similarly register-annotated datasets in Finnish, Swedish and French (Laippala et al., 2019; Repo et al., 2021) have extended these possibilities to a multilingual setting (Rönnqvist et al., 2021).

In this paper, we benefit from these advances and present Register Oscar, a version of the widely used Oscar dataset (Ortiz Suárez et al., 2019) to which we have automatically created register labels. Furthermore, we introduce and apply a machine learning method for cleaning text files originating from Web crawls—such as the Oscar documents—to filter out noise left after boilerplate removal.

Register Oscar covers 14 languages. To identify the document registers, we use the multilingual

[*]Work done prior to Amazon.

Proceedings of the 2022 COLING Workshop: The 8th Workshop on Noisy User-generated Text (W-NUT 2022), pages 215–221

register model by Rönnqvist et al. (2021) based on four languages. To evaluate the model on the wider set of languages included in Register Oscar, we present new CORE-style annotated evaluation datasets in eight languages: Arabic, Catalan, Chinese, Hindi, Indonesian, Portuguese, Spanish and Urdu. We find that the zero-shot performance of the model on these culturally and linguistically different languages is 0.70 F1-score, similar to the previously reported zero-shot results.

In sum, our main contributions are:

- We provide automatic register annotations for 351M documents in Oscar in 14 languages, using the register model by Rönnqvist et al. (2021).
- We present new manually annotated register corpora for eight languages and evaluate the register identification model on these.
- We introduce and apply a new machine learning method for cleaning text files from Web crawls.

The register annotations for Oscar are available at `https://huggingface.co/datasets/TurkuNLP/register_oscar`, and the new manually annotated register corpora and the text quality annotations used to train the cleaning system at `https://github.com/TurkuNLP/multilingual-register-labeling`.

2 Data

Oscar (Ortiz Suárez et al., 2019) is our main source of data. We use the version available at `https://huggingface.co/datasets/oscar` in the following 14 languages: Arabic, Basque, Bengali, Catalan, Chinese, English, French, Hindi, Indonesian, Portuguese, Spanish, Swahili, Urdu and Vietnamese. Following the Big Science project[1], the languages were selected so that they represent a variety of language families and geographical locations and include also low-resource languages.

The new manually annotated multilingual register corpora cover eight languages created as a part of the current study, the main objective being to allow for a more extensive evaluation of the register identification model on the Oscar languages. The newly annotated datasets include culturally and linguistically varied languages. As register is deeply associated with the situational context of the text (Biber, 1988) and, e.g., blogs can have very different characteristics in different cultures, this offers a unique chance to evaluate the robustness of the register model.

Narrative NA
News report / news blog, narrative blog
Opinion OP
Review, opinion blog, advice
Informational Description IN
Description of a thing or a person, research article
Interactive Discussion ID
How-to HI
How-to / instruction, recipe
Informational Persuasion IP
Description with intent to sell
Lyrical LY
Spoken SP
Machine Translated MT

Table 1: Main registers and examples of sub-registers.

The documents for the annotation were randomly sampled from a recent Common Crawl[2] dataset. The annotation was done using a custom annotation tool. Most of the annotators have a background in linguistics or NLP. The annotators were given a detailed tutorial to the register scheme, see `https://turkunlp.org/register-annotation-docs/`.

The annotations of the new datasets follow the hierarchical CORE register scheme consisting of eight main registers, tens of subregisters, and the category Machine Translated, see Table 1. To cover all the documents found in the online jungle, the scheme has been created in a data-driven manner and allows for the annotation of *hybrid* documents simultaneously assigned to several registers (Biber et al., 2020; Egbert et al., 2015). For instance, a lifestyle blog telling about the writer's day and promoting a product would be annotated as both *Narrative* and *Informational Persuasion*.

The newly annotated register corpora are described in Table 2. Their sizes vary, Indonesian being the largest and Arabic the smallest language. Overall, the sizes are relatively small. Therefore, we focus here on the main register level. The register distributions are also very uneven. This was expected, as similar distributions have been found for the four original languages (Laippala et al., 2019; Repo et al., 2021).

The text quality annotations are used to train the

[1] `https://bigscience.huggingface.co/`

[2] `https://commoncrawl.org/`

	HI	ID	IN	IP	LY	NA	OP	SP	HYB	MT	No label	Total
ar	2	3	12	7		32	10		23	3		92
ca	2	2	41	11	2	34	10	2	2	3	2	111
es	6	3	25	27		31	4		3	1		100
hi	3	1	26	12	10	82	6	2	13	5	1	161
id	34	5	153	131	10	239	79	2	504	29	4	1190
pt	24	6	47	101	3	97	23		31		2	334
ur	1	1	13	9	2	94	22		17	1		160
zh	8	5	58	104	1	84	27	1	24	5		317

Table 2: New multilingual register corpora. Hybrids (HYB) are presented as one class. No label refers to documents for which the annotators could not find a suitable register.

Language	Texts	Accepted lines	Rejected lines	Lines total
English	104	3 360	2 812	6 172
Finnish	89	1 797	2 480	4 277
French	1 807	57 345	26 171	83 516
German	112	2 529	925	3 454
Spanish	70	1 536	1 483	3 019
Swedish	2 114	47 302	51 099	98 401
Total	4 296	113 869	84 970	198 839

Table 3: Text quality annotations.

model behind the cleaning pipeline. The method is trained on documents annotated line-by-line as *accept* or *reject* according to if the line was part of the main text or not. The statistics of this dataset are described in Table 3. The documents were retrieved from Common Crawl using the same pipeline as the register annotated documents, and they were pre-processed for boilerplate removal using Trafilatura version 0.3.

3 Methods

The register labeling of the Oscar documents is done using the *master multilingual* model by Rönnqvist et al. (2021). The model is based on a fine-tuned XLM-R (Conneau et al., 2020) using French, Finnish and English data, and is available at `https://github.com/TurkuNLP/multilingual-register-labeling`. To account for hybrid documents (see Section 2), the model is multi-label allowing to predict several registers for one document.

The register model has been reported to achieve an F1-score of 0.77 on a multilingual dataset. Furthermore, it outperforms also monolingual language-specific neural classifiers in these languages (Rönnqvist et al., 2021), and provides much higher performance than earlier systems based on SVMs or statistical techniques that additionally would not allow for the modeling of languages without training data (Laippala et al., 2021; Biber and Egbert, 2016). Therefore, the use of the XLM-R is motivated in the current study despite the computational costs. Finally, we also evaluate the performance of the XLM-R-based register identification model on the new multilingual register corpora.

The cleaning of the Oscar documents from remains of boilerplate and elements not belonging to the main text works on text files and is based on machine learning, unlike boilerplate removal that is typically rule-based and takes html as input.

The pipeline consists of three steps. First, the data is run through a heuristic filtering script with language detection using langdetect to filter out e.g., documents that are less than 75 words long or have a high ratio of digit (> 0.075) or foreign characters (>0.02).

Second, an XLM-R (Conneau et al., 2020) classifier is trained to predict whether a document is machine generated or not, using data from our ongoing register annotation projects where Machine translation and generation is one of the register categories. We optimize learning rate using a grid of rates between 1e-7..9e-5.

Third, we filter out lines, defined as sequences of characters separated by a line break, that do not belong to the main text of the document. This step uses the text quality annotations described in Section 2 and includes two XLM-R models: a bag-of-lines classifier to predict whether a line is main text content or not, and another one with an extra Long Short-Term Memory (LSTM) layer to predict the line quality based on sequences of embeddings retrieved from the first model. We optimize the learning rate within the range of 1e-7..9e-5, and compare the performances of the first model to the entire architecture.

4 Evaluation

4.1 Register model performance on the new languages

Figure 1 presents the performance of the register identification model on the new multilingual register corpora and on English and French already used in the original model development (Rönnqvist et al., 2021). The model performance varies between 0.58 and 0.82 for the new languages, the lowest being for Indonesian and the highest for Urdu. Overall, the total average F1-score on all the evaluation datasets is 0.70.

Model	Accuracy	sd	t-value
Bag-of-lines XLM-R	0.84	0.011	
Sequential XLM-R	0.88	0.002	t(2) = 45

Table 4: Performances of the line-wise cleaning models.

The performance of the model on the new set of languages is somewhat lower than the original performance reported by Rönnqvist et al. (2021), 0.77. However, importantly, the original setting was multilingual with the same languages included in the training and testing, whereas ours is zero-shot. This explains the decrease—similarly, Rönnqvist et al. (2021) report an F1-score of 0.71% on a zero-shot experiment.

4.2 Cleaning pipeline

The classifier predicting whether entire documents are machine generated or not achieved a mean F1-score of 0.98, averaged over three instances (*SD* 0.001).

The performances of the bag-of-lines classifier and the sequence-to-sequence architecture identifying the text quality based on the line-wise annotations are described in Table 4. The results are means over three runs. We can see that while both methods achieve competitive results, the sequence-to-sequence model outperforms the classifier approach by four percentage points. This was to be expected considering that lines featuring actual text and noise are not evenly distributed in a document—instead, there may be long passages of actual text, and then again several lines of noise. The sequence-to-sequence approach can take advantage of this ordering, resulting in a higher performance.

	Texts	Main content lines	Noise lines	Words	Cleaned texts
ar	9.01M	43.5M	901k	2.65B	**3.36M**
bn	1.11M	7.19M	332k	358M	**1.1M**
ca	2.46M	9.43M	235k	556M	**1.22M**
en	304M	2.99B	103M	169B	**214M**
es	56.3M	393M	8.76M	21.3B	**34.6M**
eu	257k	835k	12.6k	37.1M	**112k**
fr	59.4M	360M	9.96M	18.6B	**34.2M**
hi	1.91M	9.03M	370k	630M	**1.13M**
id	9.95M	43.4M	590k	2.1B	**6.23M**
pt	26.9M	162M	2.79M	8.49B	**15.9M**
sw	24.8k	38.7k	1.19k	1.37M	**24.7k**
ur	429k	1.86M	51.9k	162M	**260k**
vi	9.9M	76.5M	2.61M	4.86B	**7.09M**
zh	41.7M	186M	5.99M	24.7B	**31.2M**
Total	524M	4.28B	136M	253B	**351M**

Table 5: Data sizes before and after the cleaning.

4.3 Register Oscar in numbers

Table 5 describes the Oscar dataset we use and the effect of the cleaning pipeline to its size. The word counts represent space-separated tokens except for Arabic and Chinese, where the texts were tokenized with UDPipe (Straka and Straková, 2017). Overall, the filtering reduced the dataset sizes relatively aggressively to ~30-40% of the original. However, for most of the languages, the sizes are still giant—English, French, Spanish, Portuguese and Chinese cover tens of millions of documents, and Arabic, Bengali, Catalan, Hindi, Indonesian and Vietnamese 1-10 million documents. Basque, Swahili and Urdu have only 20,000-260,000 cleaned documents, but their sizes were small already in the uncleaned version. Finally, Figure 2 in Appendix presents the register distributions for each language in the cleaned dataset. For most languages, the distributions follow the training data—Narrative and Informational Description are the most frequent, while Spoken and Lyrical feature a much smaller proportion of the data. E.g., English Lyrical covers 164,105 documents. For some of the lower-resource languages—Bengali, Hindi, Swahili and Urdu—the vast majority of the documents are predicted as Narrative. This can be related to many aspects of the data collection and processing, and will be examined in future work.

5 Conclusions

In this paper, we have presented automatically produced register annotations for the widely used Oscar dataset in 14 languages, and we have evaluated the register identification model against new datasets covering eight languages not included in the original model development. Furthermore, we have described a machine-learning method for cleaning text data originating from Web crawls, and we have applied the method to further clean the documents in the entire dataset.

The evaluation showed that the performance of the register model is comparable to previously reported zero-shot results, although the newly annotated datasets feature linguistically and culturally diverse languages. This suggests that multilingual register identification can be used to provide structure and improve the usability of Web-crawled data, where the content ranges from noisy user-generated text to professionally edited documents. The register annotations automatically produced in this study cover altogether eight

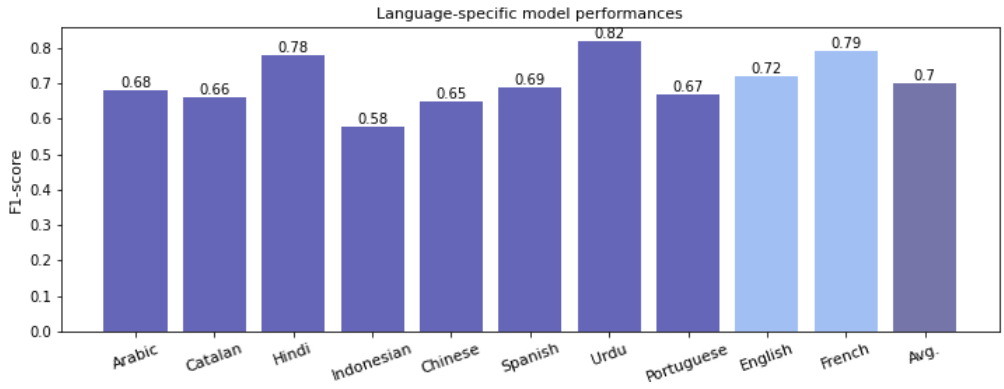

Figure 1: Language-specific performances of the register model.

registers and 351 million documents, available at `https://huggingface.co/datasets/TurkuNLP/register_oscar`. The new manually annotated register datasets and the text quality annotations used to develop the cleaning pipeline can be found at `https://github.com/TurkuNLP/multilingual-register-labeling`.

Acknowledgements

We thank the Emil Aaltonen Foundation and Academy of Finland for financial support and the Big Science project for collaboration. We also wish to acknowledge CSC – IT Center for Science, Finland for computational resources.

References

Rezapour Noushin Asheghi, Katja Markert, and Serge Sharoff. 2014. Semi-supervised graph-based genre classification for web pages. In *Proceedings of TextGraphs-9*, pages 39–47.

Adrien Barbaresi. 2021. Trafilatura: A web scraping library and command-line tool for text discovery and extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 122–131, Online. Association for Computational Linguistics.

Douglas Biber. 1988. *Variation across speech and writing*. Cambridge University Press, Cambridge.

Douglas Biber and Susan Conrad. 2019. *Register, Genre, and Style*. Cambridge University Press, Cambridge.

Douglas Biber and Jesse Egbert. 2016. Using grammatical features for automatic register identification in an unrestricted corpus of documents from the open web. *Journal of Research Design and Statistics in Linguistics and Communication Science*, 2:3–36.

Douglas Biber, Jesse Egbert, and Daniel Keller. 2020. Reconceptualizing register in a continuous situational space. *Corpus Linguistics and Linguistic Theory*, Ahead of print.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard 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, Online. Association for Computational Linguistics.

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*, 66:1817–1831.

Adam Kilgarriff. 2007. Last Words: Googleology is Bad Science. *Computational Linguistics, Volume 33, Number 1, March 2007*.

Veronika Laippala, Jesse Egbert, Douglas Biber, and Aki-Juhani Kyröläinen. 2021. Exploring the role of lexis and grammar for the stable identification of register in an unrestricted corpus of web documents. *Language resources and evaluation*.

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, Turku, Finland. Linköping University Electronic Press.

Gjorgji Madjarov, Vedrana Vidulin, Ivica Dimitrovski, and Dragi Kocev. 2019. Web genre classification with methods for structured output prediction. *Information Sciences*, 503:551 – 573.

Suraj Maharjan, Manuel Montes, Fabio A. onzález, and Thamar Solorio. 2018. A genre-aware attention model to improve the likability prediction of books. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages

3381–3391. Association for Computational Linguistics.

Pedro Javier Ortiz Suárez, Benoit Sagot, and Laurent Romary. 2019. Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019. Cardiff, 22nd July 2019, pages 9 – 16, Mannheim. Leibniz-Institut für Deutsche Sprache.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *CoRR*, abs/1910.10683.

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 *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 183–191, Online. Association for Computational Linguistics.

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, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

Marina Santini. 2008. Zero, single, or multi? genre of web pages through the users' perspective. *Information Processing & Management*, 44(2):702–737.

Serge Sharoff. 2010. In the garden and in the jungle comparing genres in the bnc and internet.

Serge Sharoff, Zhili Wu, and Katja Markert. 2010. The web library of babel: evaluating genre collections. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Milan Straka and Jana Straková. 2017. Tokenizing, pos tagging, lemmatizing and parsing ud 2.0 with udpipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99, Vancouver, Canada. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. *CoRR*, abs/2010.11934.

A Appendix

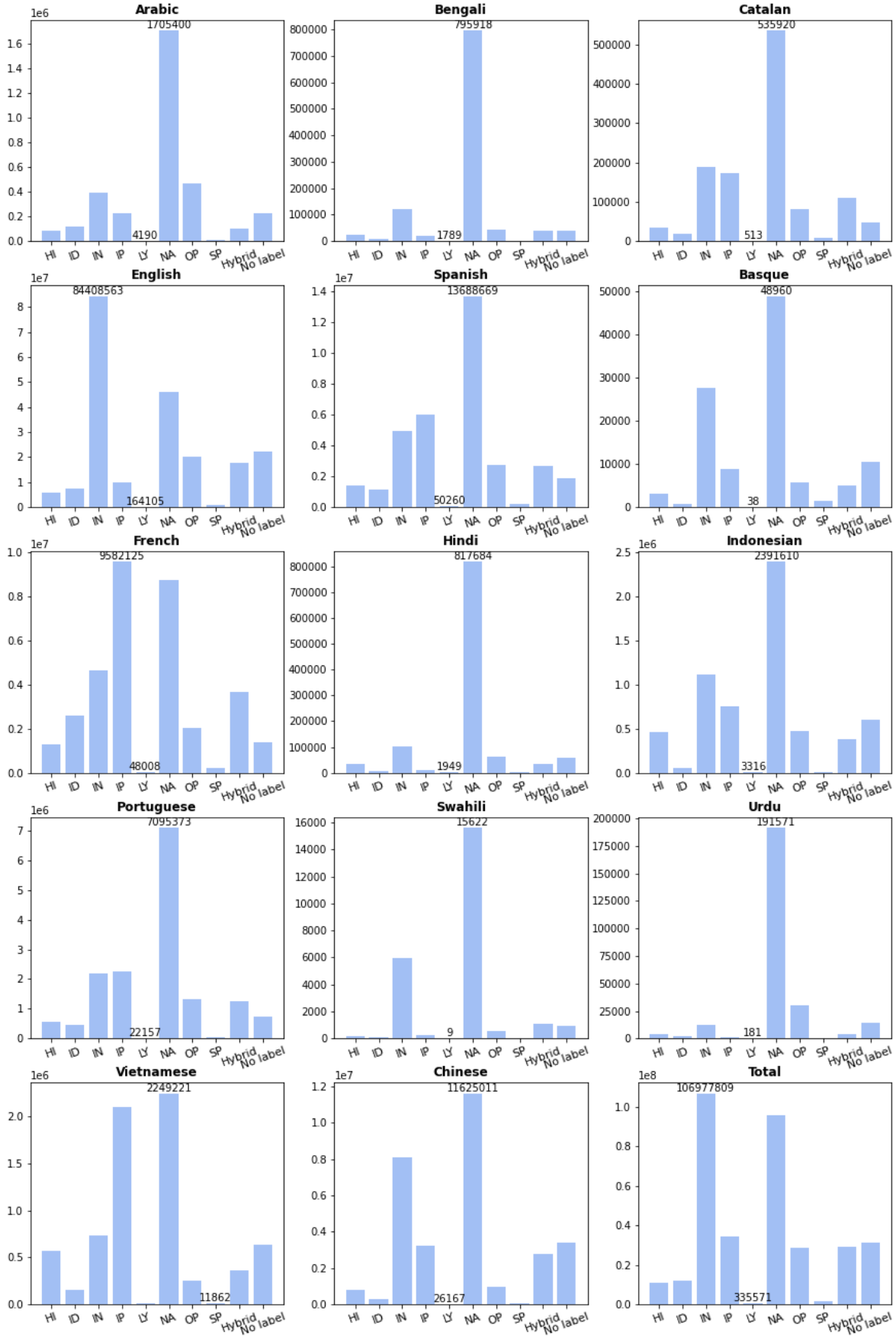

Figure 2: Register distributions per language in the cleaned dataset. The sizes of the largest and the smallest class for each language are indicated. Please note the varying scales of the figures.