

What Persists across Web Crawls?

Genre and News Topic Variation in South Slavic Web Corpora

Nikola Ljubešić^{1,2,3}, Vuk Dinić^{1,4}, Peter Rupnik^{1,4}

¹Jožef Stefan Institute, Jamova 39, Ljubljana, Slovenia,

²Institute of Contemporary History, Privoz 11, Ljubljana, Slovenia,

³Faculty of Computer and Information Science, Večna pot 113, Ljubljana, Slovenia,

⁴Jožef Stefan International Postgraduate School, Jamova 39, Ljubljana, Slovenia,

Correspondence: nikola.ljubestic@ijs.si

Abstract

This work investigates content persistence across two independently constructed versions of the CLASSLA-web corpora for seven South Slavic languages, crawled approximately two years apart, with available document-level genre and topic metadata. Persistence is defined as the presence of a near-duplicate document in both versions of the corpora, which occurs on average in 19% of cases. Genre and news-topic persistence rankings are moderately consistent across languages. Web domain availability after two years is moderately associated with genre persistence ($r = .49$, $R^2 = .24$), but not with topic persistence within news ($r = -.04$). Topic-level differences therefore appear to arise mainly within surviving news websites, where reference-oriented content tends to persist more than event-bound content.

1 Introduction

Large-scale web corpora have become a central resource for natural language processing (NLP), particularly for languages for which manually curated data remain limited (Baroni et al., 2009; Burchell et al., 2025). This is especially relevant for South Slavic languages, where web-derived corpora support both linguistic research and the development of language technologies, including language models, machine translation systems, and downstream NLP applications (Popović, 2020; Ljubešić and Lauc, 2021).

A characteristic feature of web data is the substantial amount of duplicate and near-duplicate content. When the same language-specific web space is crawled repeatedly, some documents are retrieved again, while others disappear, change, or are no longer rediscovered by the crawling process. Successive corpora therefore represent partially overlapping snapshots whose differences reflect both changes on the web and the behaviour of

the corpus-construction pipeline. Although duplicate content is usually treated as noise and removed during preprocessing, overlap between crawls can also serve as a signal of content persistence, which can inform crawling frequency, crawling foci, etc.

In this paper, we investigate such persistence across two independently constructed versions of the CLASSLA-web corpora (Ljubešić and Kuzman, 2024; Kuzman Pungeršek et al., 2026) for seven South Slavic languages, crawled approximately two years apart using comparable crawling and processing procedures. Our aim is not to estimate the persistence of the web as a whole. Rather, taking Version 2 documents as the reference population, we examine which types were also represented by a near-duplicate in Version 1 under a largely unchanged corpus-construction setting.

More specifically, we analyse how document persistence varies across automatically assigned genres (Kuzman et al., 2023) and topics (Kuzman and Ljubešić, 2025), whether these patterns are consistent across languages, and to what extent they can be associated with the continued availability of the source websites. In doing so, we treat cross-crawl duplication not merely as a preprocessing problem, but as evidence about the differential persistence and turnover of web content.

2 Related Work

South Slavic web corpora. Web corpora have long provided an important source of data for South Slavic languages. Early resources such as *s/WaC*, *hrWaC*, *srWaC*, and *bsWaC* followed the Web-as-Corpus methodology (Baroni et al., 2009; Ljubešić and Erjavec, 2011; Ljubešić and Klubička, 2014), while large multilingual collections such as *CC100*, *mC4*, and *HPLT* later expanded the amount of available web data (Conneau et al., 2020; Raffel et al., 2020; Burchell et al., 2025). The *CLASSLA-web* corpora provide a more curated and directly

comparable collection for South Slavic languages, constructed using a shared crawling and processing pipeline (Ljubešić and Kuzman, 2024). Their two independently collected versions therefore offer a suitable setting for studying overlap between crawls.

Cross-crawl deduplication during the preparation of South Slavic language-model training data has already suggested substantial variation in such overlap. Deduplication among the earlier WaC corpora removed only single-digit percentages of the data used for pre-training the *BERTi* encoder model, while approximately 15% of additional Common Crawl material overlapped with those corpora (Ljubešić and Lauc, 2021). When newer corpora were used to construct the *XLM-R-BERTi* training dataset and the related continuously pre-trained multilingual encoder model, only 45% of the earlier *BERTi* data remained after deduplication (Ljubešić et al., 2024). Most recently, and most relevant for this work, approximately 20% of documents in the second version of *CLASSLA-web* were matched to documents in the first version (Kuzman Pungershek et al., 2026). These observations show that independently collected web corpora overlap only partially, but do not explain which types of content are most likely to recur.

Web persistence and change. Research on web persistence predates current large-scale NLP corpora. Early longitudinal studies estimated an average web-page lifetime of approximately two years and documented continual change in both pages and websites (Koehler, 1999, 2002). Later work showed that change is highly uneven: frequently visited pages change more often than randomly sampled pages, and content survival follows a characteristic J-shaped decay pattern (Adar et al., 2009). Studies of reference rot and long-term web accessibility similarly demonstrate that substantial portions of previously available web content eventually become unreachable (Klein et al., 2014; Chapekis et al., 2024).

These studies generally examine the continued accessibility or modification of known URLs. Cross-crawl overlap measures a related but different phenomenon. A document is observed as persistent only if sufficiently similar content is retrieved and retained by both crawling and processing pipelines. Failure to observe it again may therefore result from page removal, content change, website unavailability, crawl coverage, or filtering

decisions. In this paper, we use *persistence* in this operational, corpus-level sense and subsequently examine whether differences in cross-crawl overlap can be partly explained by the later availability of source websites.

Duplicate as signal. In contemporary NLP, duplicate web content is normally treated as undesirable redundancy. Deduplication reduces memorisation and training cost without necessarily harming model quality (Lee et al., 2022), and large web datasets may discard substantial portions of their input through exact and near-duplicate removal (Penedo et al., 2023). In contrast, we use near-duplicate matches as evidence of content recurring across independently constructed corpora. Our main contribution is to relate the content recurrence to automatically assigned genre and topic labels across seven South Slavic languages, thereby investigating whether some categories of web content are systematically more persistent than others. These insights can inform the design of future crawls, as well as support retrospective interpretation of existing web corpora, in particular by showing which types of content are more likely to be missed or underrepresented in existing crawls.

3 Data and Research Questions

Our analysis uses the two versions of the *CLASSLA-web* corpora (Ljubešić and Kuzman, 2024; Kuzman Pungershek et al., 2026), which cover seven South Slavic languages. Version 1 was crawled in 2021–2022 and consists of 26 million documents and 11 billion words, and Version 2 in 2024 containing 38 million documents and 17 billion words, both versions using comparable crawling and processing pipelines. Near-duplicates were removed within each version independently.

During the construction of Version 2, documents were also matched across versions using MinHash (Shrivastava and Li, 2014) configured with an LSH similarity threshold of 0.7, 96 MinHash permutations, and four-word shingles. This particular MinHash setup was identified via manual inspection on multiple iterative runs over these particular data collections (Kuzman Pungershek et al., 2026).

We treat a Version 2 document as *persistent* if a near-duplicate was found in Version 1. Thus, persistence is directional: it is the proportion of documents in the later corpus that are also represented in the earlier corpus. Also, this is an operational measure of recurrence across independent crawls

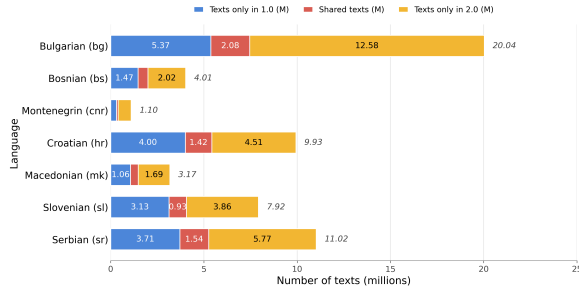


Figure 1: Numbers of documents in CLASSLA-web Versions 1 and 2 and their cross-version overlap by language, adapted from Kuzman Pungeršek et al. (2026).

rather than direct evidence that a particular URL remained unchanged. Across languages, 12–24% of Version 2 documents were matched to Version 1, with an average persistence rate of 19%. Figure 1 summarizes corpus sizes and overlap by language.

We examine whether persistence varies systematically across the genres and topics automatically assigned to Version 2 documents, and whether the same patterns recur across languages. Topic-level analysis is conducted within the most prominent news genre, for which the topic classifier was developed and evaluated. Finally, we test whether differences in persistence can be partly explained by the later availability of the corresponding web-sites and documents.

4 Results

4.1 Genre Persistence

To contextualize our results on genre persistence, we first estimate the overall genre distribution from equally sized samples (10 thousand documents) of the seven Version 2 corpora. News accounts for 56% of documents, followed by Promotion (10%), Forum and Information/Explanation (8% each), Opinion/Argumentation (6%), Instruction (4%), Mix (3%), Other (2%), and Prose/Lyrical and Legal (1% each). It is important to distinguish between *Other* and *Mix* categories. *Other* is an explicit category used during manual annotation for documents that do not fit any category in the annotation schema, whereas *Mix* is assigned automatically at inference time when no genre reaches the confidence threshold of 0.8 (Kuzman Pungeršek et al., 2026).

Figure 2 presents persistence rates by genre and language. Cell values give the raw persistence percentages, while colour indicates the $\log_2$ ratio between the genre-specific rate and the overall persis-

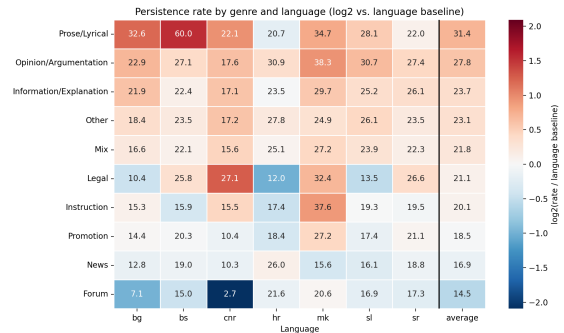


Figure 2: Persistence rates by genre and language. A Version 2 document is considered persistent if a near-duplicate is found in Version 1. *Mix* is a low-confidence fallback class.

tence baseline for that language: red denotes above-baseline persistence, blue below-baseline persistence, and near-white values indicate little difference from the baseline. The results show substantial and moderately consistent differences among genres. Prose/Lyrical content is generally the most persistent, whereas Forum content is among the least persistent. To assess whether the relative ordering of genres is shared across languages, we calculated Spearman rank correlations for all 21 language pairs. The median correlation was $\rho = .53$, indicating moderate cross-language consistency. Pairwise values ranged from $\rho = -.24$ for Croatian and Macedonian to $\rho = .94$ for Bulgarian and Slovenian, with Croatian showing the least consistent pattern overall.

4.2 Topic Persistence

We again start the discussion on persistence of topics with an insight into the most frequent topics in the language-balanced sample of Version 2, most frequent topics being Politics, Arts, Culture, Entertainment and Media, and Economy, Business and Finance (12% each), followed by Sport (11%), Human Interest (8%), and Mix and Lifestyle and Leisure (7% each). The remaining topics individually account for 5% or less of the data. For topic classification, *Mix* is assigned automatically when no single topic reaches the confidence threshold of 0.6; it therefore denotes documents without a confidently identifiable dominant topic rather than a manually defined substantive category (Kuzman Pungeršek et al., 2026).

Because the IPTC topic classifier was developed and evaluated on news data, we report topic persistence within the News genre, which also constitutes

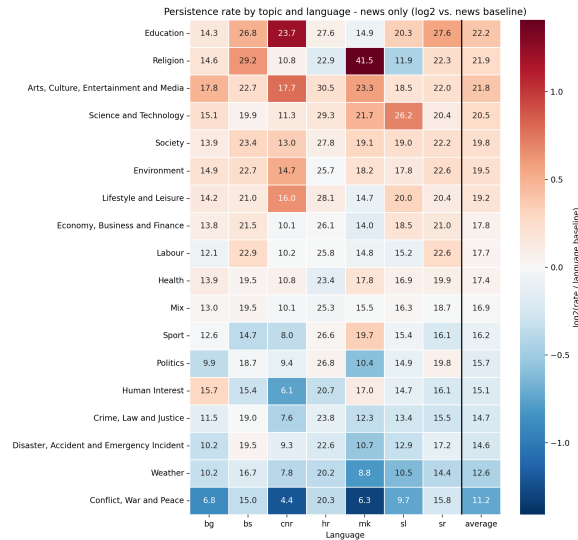


Figure 3: Persistence rates by news topic and language. *Mix* is a low-confidence fallback class.

more than half of the overall content.

The ordering of news topics is moderately to strongly consistent across languages. The median pairwise Spearman correlation is $\rho = .55$, and all 21 language-pair correlations are positive, ranging from $\rho = .43$ to $\rho = .92$. Topic rankings are therefore somewhat more uniformly shared across languages than genre rankings.

The most persistent topics tend to contain reusable or reference-oriented content, including Education, Religion, Arts and Culture, and Science and Technology. The least persistent topics are more event-bound, particularly Conflict, Weather, Disasters, and Crime.

4.3 Persistence and Web Domain Availability

A document may fail to recur across crawls because its source website has disappeared, but also because the page has changed, moved, been removed, or was not rediscovered by the crawler. To assess the contribution of website survival, we sampled 100 Version 2 documents for each genre and each news topic in every language and tested the availability of their source web domains in July 2026, approximately two years after the Version 2 crawl, obtaining thereby genre-specific and topic-specific website availability rates, structurally comparable to the persistence rates presented previously. Average website availability remained between roughly 70% and 90%, substantially above the 12–24% cross-crawl persistence rates.

To estimate the relationship between persistence

and web domain availability, we calculated Pearson correlations between persistence rates and availability rates. We additionally fitted a simple linear regression of persistence on domain availability to obtain R^2 . All language-category combinations received equal weight. The complete availability rates are reported in Appendix A.

Across the 70 genre-language combinations, web domain availability was moderately associated with persistence ($r = .49$, $p < .001$, $R^2 = .24$). Genres occurring on more durable websites therefore also tended to recur more often across crawls. In contrast, no association was observed across news-topic-language combinations ($r = -.04$, $R^2 < .01$). Website survival may thus explain part of the variation among genres, whereas topic-level differences within news appear to arise mainly from processes within surviving websites. Our working hypothesis still remains that reference-oriented content (news on education, arts, religion, science) has a much slower turnover than event-bound content (news on wars, crime, disasters and weather).

5 Conclusion

We examined content persistence across two independently constructed versions of the CLASSLA-web corpora for seven South Slavic languages, collected approximately two years apart. On average, 19% of Version 2 documents had a near-duplicate in Version 1, indicating substantial turnover between otherwise comparable crawls.

Persistence varied systematically across both genre and topic. Genre rankings were moderately consistent across languages, with Prose/Lyrical content generally among the most persistent and Forum content among the least. Topic-level patterns within news were more uniform: reusable and reference-oriented topics tended to persist more, whereas event-bound topics such as Conflict, Weather, Disasters, and Crime were less persistent.

Website survival explained only part of these differences. Web domain availability was moderately associated with genre persistence, but unrelated to persistence across news topics. Much of the observed news turnover therefore appears to arise within surviving websites.

Future work should exploit the released data¹ primarily to disentangle the mechanisms underlying cross-crawl persistence.

¹The derived TSV-formatted dataset is available at <http://hdl.handle.net/11356/1972>.

Limitations

Our measure of persistence captures recurrence across two independently constructed corpora rather than direct survival of individual web pages. Documents may fail to match because they were created only after the first crawl, were removed or substantially changed, were not rediscovered by the crawler, were filtered during processing, or fell below the near-duplicate threshold. The latter may also affect genres differently depending on document length and structure. Accordingly, the observed rates should be interpreted as cross-crawl recurrence rather than direct estimates of content lifetime or turnover.

Genre and topic labels were assigned automatically and therefore contain classification errors. Topic analysis is restricted to news because the classifier was primarily evaluated in that genre.

The availability analysis is based on samples of 100 documents per category and language and measures domain reachability only at a single follow-up point in July 2026, rather than during the interval between the two crawls. It therefore provides only an indirect proxy for website survival, particularly for news topics, which largely share domains and show limited variation in availability. Finally, the study covers two crawl waves, seven closely related languages, and one corpus-construction and near-duplicate matching setup, limiting generalizability to other web collections and time intervals.

Acknowledgements

This paper was produced as part of the project “Large Language Models for Digital Humanities” (Grant GC-0002), the research programme “Language Resources and Technologies for Slovene” (Grant P6-0411), and the Research Infrastructure DARIAH-SI (I0-E007), all funded by the ARIS Slovenian Research and Innovation Agency, as well as a part of the Datavysts MSCA network, funded by Horizon Europe (grant number 101169037) and UK Research and Innovation (UKRI) under the UK government’s Horizon Europe funding Guarantee (grant number EP/Z536581/1).

This research was also supported by LLMs4EU, co-funded by the Digital Europe Programme under GA 101198470. This project is co-funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Commission. Neither the Euro-

pean Union nor the granting authority can be held responsible for them.

References

- Eytan Adar, Jaime Teevan, Susan T. Dumais, and Jonathan L. Elsas. 2009. [The web changes everything: Understanding the dynamics of web content](#). In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, pages 282–291, Barcelona Spain. ACM.
- 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.
- Laurie Burchell, Ona De Gibert Bonet, Nikolay Arefyev, Mikko Aulamo, Marta Bañón, Pinzhen Chen, Mariia Fedorova, Liane Guillou, Barry Haddow, Jan Hajič, Jindřich Helcl, Erik Henriksson, Mateusz Klimaszewski, Ville Komulainen, Andrey Kutuzov, Joonas Kytöniemi, Veronika Laippala, Petter Mæhlum, Bhavitvya Malik, and 16 others. 2025. [An expanded massive multilingual dataset for high-performance language technologies \(HPLT\)](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 17452–17485, Vienna, Austria. Association for Computational Linguistics.
- Athena Chapekis, Samuel Bestvater, Emma Remy, and Gonzalo Rivero. 2024. [When online content disappears](#).
- 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.
- Martin Klein, Herbert Van De Sompel, Robert Sander-son, Harihar Shankar, Lyudmila Balakireva, Ke Zhou, and Richard Tobin. 2014. [Scholarly Context Not Found: One in Five Articles Suffers from Reference Rot](#). *PLoS ONE*, 9(12):e115253.
- Wallace Koehler. 1999. An analysis of web page and web site constancy and permanence. *Journal of the American Society for Information Science*, 50(2):162–180.
- Wallace Koehler. 2002. [Web page change and persistence—a four-year longitudinal study](#). *Journal of the American Society for Information Science and Technology*, 53(2):162–171.

- Taja Kuzman and Nikola Ljubešić. 2025. [LLM teacher-student framework for text classification with no manually annotated data: A case study in IPTC news topic classification](#). *IEEE Access*, pages 1–1.
- Taja Kuzman, Igor Mozetič, and Nikola Ljubešić. 2023. Automatic genre identification for robust enrichment of massive text collections: Investigation of classification methods in the era of large language models. *Machine Learning and Knowledge Extraction*, 5(3):1149–1175.
- Taja Kuzman, Pungeršek, Peter Rupnik, Vít Suchomel, and Nikola Ljubešić. 2026. The growing gains and pains of iterative web corpora crawling: Insights from South Slavic CLASSLA-web 2.0 corpora. *arXiv preprint arXiv:2601.11170*.
- Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. 2022. [Deduplicating Training Data Makes Language Models Better](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8424–8445, Dublin, Ireland. Association for Computational Linguistics.
- Nikola Ljubešić and Tomaž Erjavec. 2011. hrWaC and slWaC: Compiling web corpora for croatian and slovene. In *International Conference on Text, Speech and Dialogue*, pages 395–402. Springer.
- Nikola Ljubešić and Filip Klubička. 2014. {bs, hr, sr} wac-web corpora of bosnian, croatian and serbian. In *Proceedings of the 9th web as corpus workshop (WaC-9)*, pages 29–35.
- Nikola Ljubešić and Taja Kuzman. 2024. [CLASSLA-web: Comparable web corpora of South Slavic languages enriched with linguistic and genre annotation](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3271–3282, Torino, Italia. ELRA and ICCL.
- Nikola Ljubešić and Davor Lauc. 2021. [BERTiC - the transformer language model for Bosnian, Croatian, Montenegrin and Serbian](#). In *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*, pages 37–42, Kiyv, Ukraine. Association for Computational Linguistics.
- Nikola Ljubešić, Vít Suchomel, Peter Rupnik, Taja Kuzman, and Rik van Noord. 2024. [Language models on a diet: Cost-efficient development of encoders for closely-related languages via additional pretraining](#). In *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages @ LREC-COLING 2024*, pages 189–203, Torino, Italia. ELRA and ICCL.
- Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Hamza Alobeidli, Alessandro Cappelli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The refinedweb dataset for falcon llm: Outperforming curated corpora with web data only. In *Advances in Neural Information Processing Systems*, volume 36.
- Maja Popović. 2020. Neural machine translation for translating into croatian and serbian. In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 102–113. International Committee on Computational Linguistics.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. [Exploring the limits of transfer learning with a unified text-to-text transformer](#). *Journal of Machine Learning Research*, 21(140):1–67.
- Anshumali Shrivastava and Ping Li. 2014. In defense of minhash over simhash. In *Artificial intelligence and statistics*, pages 886–894. PMLR.

A Domain availability by language and genre/topic.

In Table 1 we present full availability rates (percentage of websites that were still available two years after the Version 2 crawl, as described in Section 4.3) for individual genres and languages. For documents with the genre label News, we further present their availability rates by their topic and language in Table 2.

Genre	bg	bs	cnr	hr	mk	sl	sr	Average
Prose/Lyrical	0.93	0.97	0.93	0.92	0.96	0.86	0.93	0.93
Other	0.89	0.87	0.86	0.90	0.93	0.87	0.89	0.89
Instruction	0.89	0.91	0.88	0.76	0.97	0.86	0.84	0.87
Information/Explanation	0.87	0.92	0.82	0.80	0.92	0.88	0.89	0.87
Mix	0.81	0.85	0.91	0.83	0.96	0.89	0.85	0.87
Opinion/Argumentation	0.85	0.75	0.91	0.80	0.98	0.94	0.82	0.86
News	0.82	0.87	0.85	0.85	0.81	0.96	0.87	0.86
Legal	0.95	0.85	0.80	0.42	0.96	0.83	0.91	0.82
Promotion	0.92	0.89	0.68	0.72	0.93	0.65	0.86	0.81
Forum	0.73	0.83	0.21	0.91	0.99	0.92	0.87	0.78

Table 1: Web domain availability rate by genre and language.

Topic	bg	bs	cnr	hr	mk	sl	sr	Average
Weather	0.82	0.90	0.94	0.93	0.93	0.92	0.92	0.91
Disaster, Accident and Emergency Incident	0.83	0.92	0.91	0.92	0.87	0.96	0.92	0.90
Labour	0.88	0.88	0.91	0.92	0.90	0.91	0.92	0.90
Science and Technology	0.93	0.86	0.91	0.91	0.93	0.91	0.85	0.90
Conflict, War and Peace	0.84	0.87	0.91	0.89	0.93	0.94	0.85	0.89
Environment	0.77	0.88	0.92	0.88	0.87	0.97	0.92	0.89
Mix	0.80	0.85	0.94	0.91	0.87	0.95	0.87	0.88
Lifestyle and Leisure	0.81	0.90	0.92	0.88	0.86	0.94	0.88	0.88
Arts, Culture, Entertainment and Media	0.88	0.85	0.91	0.88	0.83	0.96	0.85	0.88
Crime, Law and Justice	0.84	0.84	0.91	0.88	0.87	0.95	0.90	0.88
Human Interest	0.85	0.82	0.92	0.87	0.81	0.97	0.90	0.88
Health	0.71	0.87	0.95	0.91	0.88	0.94	0.88	0.88
Religion	0.81	0.95	0.91	0.85	0.78	0.93	0.88	0.87
Politics	0.85	0.84	0.92	0.87	0.88	0.88	0.81	0.86
Education	0.83	0.89	0.94	0.87	0.82	0.73	0.91	0.86
Society	0.75	0.83	0.93	0.90	0.80	0.92	0.88	0.86
Sport	0.90	0.92	0.88	0.72	0.72	0.94	0.91	0.86
Economy, Business and Finance	0.84	0.80	0.89	0.86	0.77	0.95	0.83	0.85

Table 2: Web domain availability by news topic and language.