

Mapping Sámi languages in a Norwegian national web crawl: results from a representativeness study

Magnus Breder Birkenes
National Library of Norway
magnus.birkenes@nb.no

Abstract

The paper reports from a representativeness study covering Sámi languages in a Norwegian national web crawl. We extracted linguistic content from HTML in a considerable portion of the archive and applied language identification on paragraph-level. By aggregating the results on domain-name level, we were able to quantify the amount of Sámi language content in the web archive and identify contributing domains. Further, by comparing the URLs of the national crawl with a Sámi reference corpus, we identified domain names that were lacking in the national crawl and domains that lack depth. The results will be used to increase the coverage of Sámi language content in future web crawls and acquisitions.

1 Introduction

In Norway, Norwegian and Sámi (North Sámi, Lule Sámi and South Sámi) are recognized as "languages of equal worth" under the *Act on the Sámi Parliament and other Sámi legal matters*, Section 1-5).¹ Cultural heritage institutions in Norway collect and make available content in these languages for the general public and research. These collections also include web documents.

The National Library of Norway, a major cultural heritage institution in Norway, has crawled the Norwegian web since the late 1990s. The *Legal Deposit Act*² grants the institution the right to crawl the top-level domain (TLD) *.no* and web pages with relevance for Norwegian culture and society, including indigenous peoples and minorities.

But even if this web crawl (henceforth: "the national crawl") is substantial in size, we know little about the distribution of languages in the resulting corpus, especially the Sámi languages and

other minority languages such as Kven and Romani/Romanes. One could hypothesize that the crawl is less complete for these communities due to various factors such as political priorities, unclear responsibilities, and a lack of knowledge.

The current paper reports on a representativeness study on Sámi languages within the national web crawl. It is carried out as part of the WebData research infrastructure,³ funded by The Research Council of Norway⁴ and hosted at The National Library of Norway. Among the partners is Giellatekno at UiT The Arctic University of Norway which contributes expertise in the field of Sámi language technology,

The study has two goals: Firstly, it aims to provide an overview over web domains with Sámi language content within the national crawl. Secondly, it aims to identify domains and websites that have not been captured in the national crawl and that should be harvested (if still online) or acquired in retrospective through other archivist institutions (for historical material).

We are not aware of previous language-oriented representativeness studies of national crawls, but e.g. Jauhainen et al. (2020) look into building web corpora for minority languages and Dodge et al. (2021) investigate data provenance in large web corpora like CommonCrawl.

2 Data

2.1 Raw WARC files from the national web crawl

A web archive typically consists of so-called WARC files, a standard format for capturing traffic between a crawler and the servers it visits. A WARC file is simply a sequence of client requests and server responses, with some additional meta-data such as the captured URL, the time of record-

¹<https://lovdata.no/dokument/NLE/lov/1987-06-12-56> (unofficial translation by Lovdata)

²<https://lovdata.no/dokument/NL/lov/1989-06-09-32>

³<https://webdata.nb.no>

⁴RCN project number: 350342

ing and the content format (like HTML or PDF) in addition to the content itself.

In this paper, we process WARC files from the national crawl captured between 2005–2009, 2012–2014 and 2019–2025. The missing years were not readily available for processing at the time of the investigation because they had not yet been deduplicated or could not be simply extracted due to storage limitations. Still we consider the selected periods representative enough to tell something substantial about the archive.

Further, using metadata of the WARC files, we restrict the following study to HTML pages, the core object of the web. Other text formats such as PDF and DOC(X) were excluded due to restrictions of time. Especially the former typically needs more processing in order to get high quality text, such as layout analysis and OCR.

The raw data sizes of these HTML records per year are shown in Table 2 of Appendix A. We have processed a total of 1.8 billion HTML responses with 96 TB of payload. There are considerable differences from year to year, partly due to crawling intensity, partly due to data duplication.

2.2 Reference corpus

The most difficult part of the study is to answer what the national crawl does **not** capture. The web is enormous and dynamic, and there exists no such thing as a "complete" crawl. As an approximation, we choose to use the open corpora of the GiellaLT infrastructure⁵ as a point of reference (Moshagen et al. 2023). We use the following Sámi corpora in this study: corpus-sme-orig (North Sámi), corpus-sma-orig (South Sámi) and corpus-smj-orig (Lule Sámi), as of March 2026.⁶ For each corpus, we limit ourselves to HTML documents by only extracting the document URLs referencing actual HTML documents as per the bundled metadata XSL files. We stored the URLs referencing HTML documents in a flat text file.

3 Method

3.1 Data extraction

We process the raw WARC files using the Python package `warcio`,⁷ filtering them to only include

⁵<https://giellalt.github.io/AboutGiellaLT.html>

⁶We thereby made sure that no HTML documents had been added after 2025.

⁷<https://github.com/webrecorder/warcio/>

HTML content and responses with HTTP status code 200 (OK), which indicate that the web page existed at crawl time and was served successfully.

For the following study, we are only interested in the linguistic content of web pages. In addition to the markup itself, a typical web page consists of repetitive elements (headers, footers, navigation bars) and other elements that are not part of the main content of the page and are typically labelled as "boilerplate" (Bevendorff et al. 2023, 2594).

In this study, we use the boilerplate removal system `jusText` (Pomikálek 2011). `jusText` uses a combination of rules based on the semantics of HTML elements (e.g. `nav` and `footer` elements hint at repetitive content), link and stop word density to identify (contiguous) blocks of elements containing language. For the best possible extraction, we extended `jusText` with curated stop word lists for Norwegian and Sámi languages.⁸ The output is simply a list of strings, one element per paragraph.

3.2 Language identification

Crucially, we run language detection on all extracted documents on paragraph-level, in order to account for multi-lingual texts. For language detection, we use a Python implementation⁹ of the TextCat algorithm (Cavnar and Trenkle 1994) together with a set of publicly available models (including Norwegian and the Sámi languages) from GiellaLT.¹⁰

The TextCat algorithm computes character n-grams for each text, comparing the particular distribution with a model using out-of-place measurements. Since language detection systems in general struggle with very short texts (e.g. van der Goot 2025), we do not classify documents or paragraphs that are shorter than 10 whitespace separated tokens. All paragraphs above this threshold receive exactly one label.

3.3 Deduplication

In a web crawl, one URL is typically captured multiple times and even though the web is dynamic, often the main content remains unchanged between frequent captures. When we count instances of a

⁸The stop word lists are available here: https://github.com/NationalLibraryOfNorway/maalfriid_toolkit/tree/master/src/maalfriid_toolkit/models

⁹<https://pypi.org/project/gielladetect/>, a packaged version of GiellaLT's `pytextcat` module (<https://github.com/giellalt/CorpusTools>).

¹⁰https://github.com/divvun/CorpusTools/tree/main/corpus_tools/lm

document, we do so only for unique documents within one and the same web domain (e.g. *nrk.no*). We do this by sha1-hashing each document as extracted from justext and picking the first instance. This is a simple and at the same time naïve approach as one single change in the document (e.g. a changed character) leads to a new checksum and hence a new document.

3.4 Data aggregation

We choose to aggregate the language counts per **host**. According to RFC 1738¹¹, a host is the "[t]he fully qualified domain name of a network host", that may include subdomains (e.g. *tv.nrk.no*).

We think that the host is a more suitable identifier to aggregate over than the domain itself: e.g. the domain *kommune.no* is used by all municipalities in Norway, whereas its subdomains point to each individual municipality. If we chose the domain as aggregation level, smaller languages like Sámi would nearly disappear. On the other hand, municipality subdomains like *snasa.kommune.no* or *kaffjord.kommune.no* are expected to have a high volume of Sámi content. For reasons of simplicity, we will refer to both *kommune.no* and *snasa.kommune.no* as domain names.

In order to filter out noise, we only consider content from domain names that have at least 500 tokens in any of the three considered Sámi languages.

3.5 URL set comparison

To assess the overlap between the national crawl and the reference corpus, we compare sets of normalized URLs in the two crawls. Importantly, we only look at the URL strings themselves, and do not investigate what the content looked like at crawl time or the language of the content.

We apply a set of rules to each captured unique URL in the two corpora: We 1) convert the URL to lowercase, 2) remove the protocol (http/https), 3) merge subdomains that begin with *www* with the main domain, since these two are often used interchangeably, 4) strip optional trailing slashes, 5) sort the URL parameters alphabetically, and finally 6) cut the string at the hash, effectively removing internal links and JavaScript directives.

Since the analyzed parts of the national crawl consist of 1.8 billion responses (cf. Table 2 of Appendix A), we first reduce the set of URLs from

lang	hosts	tokens	share
nob	327,123	29,951,618,479	68.16%
eng	131,526	10,611,419,013	24.15%
nno	45,726	2,019,595,019	4.60%
other	48,439	1,317,242,029	3.00%
sme	725	39,130,826	0.09%
sma	145	3,369,714	0.01%
smj	154	2,224,880	0.01%

Table 1: Overview of languages in the national crawl

the national crawl to the domain names captured in the reference corpus before applying the above set of normalization rules. This step reduces computational costs and does not affect URL-level overlap because URLs from domains absent from the reference corpus cannot contribute to the overlap.

4 Results

4.1 Overview of Sámi language content

Table 1 illustrates the extent of Sámi languages in the national crawl. As a point of comparison, we have included the two official Norwegian languages (Bokmål and Nynorsk) as well as English. Other languages were collapsed into a label *other*.¹²

In comparison to the other languages, the number of identified hosts containing Sámi content is rather low. Of a total of 43.9 billion tokens, the Sámi languages account for 44.7 million tokens or about 0.11 % of the archive. Still, the raw token counts for North, Lule and South Sámi align quite well with the reported corpus size for these languages, which was 38.18 million tokens in 2024 (cf. Wiecheteck et al. 2024, 15926)

Table 3, 4 and 5 in Appendix B show the top 20 hosts in terms of token volume for each of the three languages. When looking at the individual hosts, two numbers are particularly interesting: the absolute number of tokens in a Sámi language and the relative amount of tokens in Sámi on that particular host.

Exploring tables across the identified 831 unique hosts with Sámi content can be challenging qualitatively. We therefore developed a web application for the project that visualizes these two dimensions and lets the researchers look at individual hosts and their contents qualitatively using an interactive scatter plot. A screenshot of the application is shown in Figure 1.

¹¹<https://datatracker.ietf.org/doc/html/rfc1738>

¹²We recognized content in 46 languages, but note that our language detection system only supports 51 languages.

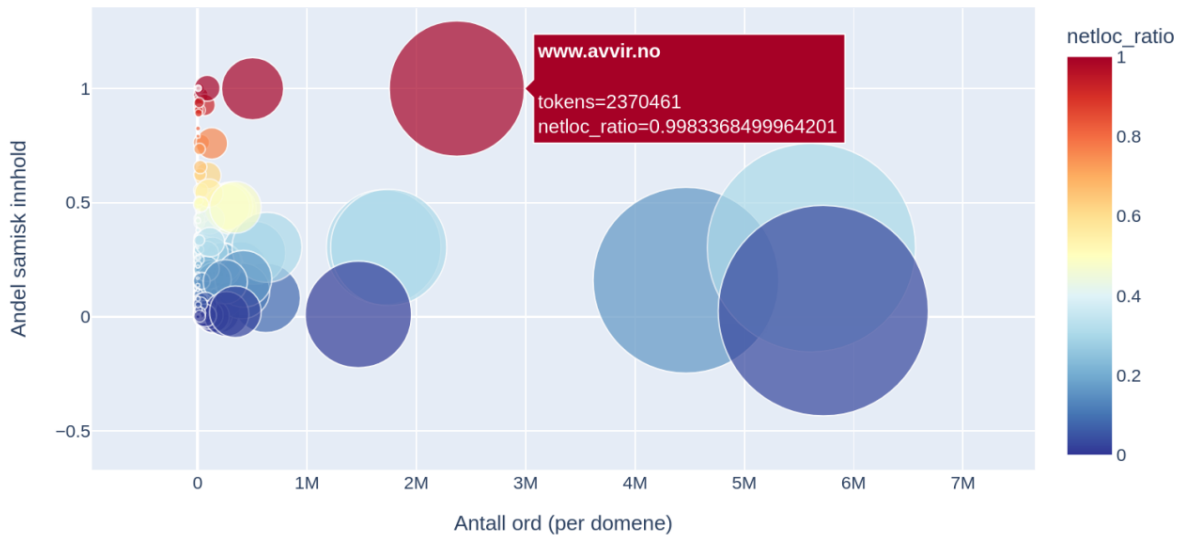


Figure 1: Screenshot of the web application used for qualitative inspection of relevant domain names and contents

The x-axis shows the number of tokens, the y-axis shows the relative amounts of Sámi tokens to other languages. The distributions of the two variables are heavily skewed: most hosts have a small amount of Sámi tokens, the median being at 2,820 tokens, but there is still a long tail of hosts with a very high amount of tokens (up to 6.7 million). Equally, the ratios concentrate around 0 %, with a median at 1 %.

There are two interesting and important outliers: The online newspaper Ávvir (*avvir.no*) publishes content nearly exclusively in North Sámi, at the same time having a large volume of tokens (2.3 million). The Norwegian Government (*regjeringen.no*), on the other hand, has a much lower proportion of Sámi languages (1.1 %), but due to the size of the web page, the volume (1.6 million tokens) is still substantial.

4.2 Comparison with a reference corpus

4.2.1 Overlap in domain names

In part two of the study, we use set comparison to analyze the two URL sets described in section 3.5. First we extract the domain names from the normalized URLs. Out of the 112 unique domain names in the reference corpus, the national crawl captured content for 80 of them (71.4 %). Of these 80 domain names, 50 are from the *.no* TLD. The remaining domain names are found among *.se* (14), *.fi* (6) and international TLDs like *.org* and *.com* (10 in total).

On the other hand, 32 domain names of the reference corpus were **not** captured by the national

crawl. Only three of these domain names belong to the *.no* TLD, which is consistent with the national crawl setup: *.no* domains are typically accepted automatically, whereas other TLDs are subject to an individual assessment. 24 of the missing domain names are in neighboring Sweden (13) and Finland (11), whereas the remaining five are on other international TLDs. Most missed domain names belong to state and municipality-level institutions.

4.2.2 Crawl depth

We also investigate the depth of the national crawl by looking at the union of the two URL sets described in section 3.5. Selecting URLs from the 112 domain names of the reference corpus, the reference corpus contains 13,722 unique URLs and the national crawl 6,612,397 unique URLs after normalization. Importantly, both URL sets reference HTML documents only.

The result is that the national crawl only captures 19.7 % of URLs of the reference corpus. This tells us something important: Whereas the domain-level overlap was substantial, the URL-level intersection proved more limited. Even when reducing the URLs of reference corpus only to the overlapping domain names, the coverage increases only slightly, to 22.2 %.

Of the 112 domain names of the reference corpus, the national crawl has URL-level overlaps only with 35 of these. And only on 18 of these 35 domains have more than 50 % of the URLs of the reference corpus been captured. For the remaining domain names, the national crawl either captures nothing (32) or other URLs (45). Even within the

national administration, the coverage seems low, with only 11 % URL overlap on the website of the Norwegian government (*regjeringen.no*).

5 Discussion and outlook

Part 1 of the study identified a sizable amount of Sámi content (measured in tokens) in the national crawl. When compared with a reference corpus in part 2, we notice that even if a substantial portion of the relevant domain names has been visited by the crawler, the crawl seems to lack depth, as only about 1/5 of the URLs of the reference corpus has been captured by the crawler.

We want to stress that the numbers need to be interpreted cautiously. For certain domains like *nrk.no* and *bibel.no* we have identified a larger share of near-duplicates, which artificially increase the token counts. A future investigation should apply methods such as SimHash or MinHash to reduce the amount of such near-duplicates which are characteristic in web archives. The same cautiousness applies to the second part of the study: Comparison of URLs across harvests at different times is problematic since the web is highly dynamic and URLs are subject to change.

The results of this study are important for the ongoing infrastructure project where the national crawl will be made available for research purposes: The resulting list of domain names can, after manual inspection, be used as seeds for the crawler. The identified lack of depth in existing domain names indicates that a more focused crawl should be conducted for the Sámi language domains in order to capture as much content as possible. For historical material, the data would need to be acquired through other web archives, e.g. from those maintained in the Nordic countries.

Ultimately, the goal of this endeavor is to increase the coverage of Sámi content in the web archive. This can both stimulate research on Sámi language and culture and the development of language technology for Sámi languages. We also hope that this study can serve as inspiration for similar initiatives at other cultural heritage institutions.

Ethical considerations

We have chosen not to publish the complete lists of hosts and URLs for reasons of privacy. The domain lists in the appendix are therefore limited to the top 20 hosts per language. Certain host names,

such as those used in blogs, contain potentially identifiable personal information. The complete lists are available upon request with a valid research purpose.

Limitations

The paper considers only a part of the national crawl, with several years missing, and only HTML pages with text extracted by *jusText*. As noted in the discussion, the raw token numbers need to be interpreted carefully, since we did not apply near-duplicate detection. Further, and very importantly, the data have not yet been investigated manually by speakers with proficiency in Sámi languages, which will be essential in the process of increasing the coverage of Sámi languages in the web archive. The language detection system, while used extensively by *GiellaLT* and The National Library of Norway, has to our knowledge not been externally audited or evaluated. Finally, the study relies on URL set comparisons and not content comparison when comparing with a reference corpus due mainly to differences in harvesting.

Acknowledgments

We are grateful to Trond Trosterud for pointing us to the *GiellaLT* corpora and for comments on the token counts and domain distribution presented in this paper.

References

- Janek Bevendorff, Sanket Gupta, Johannes Kiesel, and Benno Stein. 2023. [An empirical comparison of web content extraction algorithms](#). In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '23*, page 2594–2603, New York, NY, USA. Association for Computing Machinery.
- William B. Cavnar and John M. Trenkle. 1994. N-gram-based text categorization. In *Proceedings of third annual symposium on document analysis and information retrieval, Las Vegas, April 1994*, page 161–175.
- Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. 2021. [Documenting large webtext corpora: A case study on the colossal clean crawled corpus](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1286–1305, Online and Punta Cana, Dominican Republic.

Heidi Jauhiainen, Tommi Jauhiainen, and Krister Lindén. 2020. [Building web corpora for minority languages](#). In *Proceedings of the 12th Web as Corpus Workshop*, pages 23–32, Marseille, France. European Language Resources Association.

Sjur Nøsterbø Moshagen, Flammie Pirinen, Lene Antonsen, Børre Gaup, {Inga Lill Sigga} Mikkselsen, Trond Trosterud, Linda Wiechetek, and Katri Hiovain. 2023. *The GiellaLT infrastructure: A multilingual infrastructure for rule-based NLP*, pages 70–94. NEALT Monograph Series. University of Tartu, Estonia.

Jan Pomikálek. 2011. *Removing Boilerplate and Duplicate Content from Web Corpora*. Ph.D. thesis. Masaryk University Brno.

Rob van der Goot. 2025. [Identifying open challenges in language identification](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18207–18227, Vienna, Austria.

Linda Wiechetek, Flammie Pirinen, Maja Lisa Kappfjell, Trond Trosterud, Børre Gaup, and Sjur Nøsterbø Moshagen. 2024. [The ethical question – use of indigenous corpora for large language models](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15922–15931, Torino, Italia. ELRA and ICCL.

A Raw data analyzed per year

Table 2 shows the processed WARC data in the national crawl, i.e. HTML responses with status code 200. The table shows the number of such responses per year and the aggregated content length, i.e. the aggregated size of the response payloads.

year	responses	size (GB)
2005	199,444,309	5,914
2006	213,317,637	6,921
2007	194,967,792	6,794
2008	449,943,668	16,995
2009	156,183,539	4,273
2012	9,085,080	496
2013	50,400,889	2,748
2014	50,344,623	2,610
2019	6,122,214	465
2020	28,077,300	2,182
2021	17,854,548	1,540
2022	9,033,690	961
2023	124,720,421	11,910
2024	170,892,021	17,392
2025	130,324,074	14,933

Table 2: Processed HTML responses and size of HTML payloads per year in the national crawl

B Top 20 domains per Sámi language by token counts

Tables 3, 4 and 5 show the top 20 hosts in terms of token counts for North, Lule and South Sámi respectively. The data have been deduplicated at both the content and domain levels as described in Section 3.3. *tokens* refers to the aggregated token count of paragraphs classified as the given language, *paras* is the number of such paragraphs, *docs* is the number of documents with at least one paragraph in the respective language. *ratio* is the proportion of the respective Sámi-language tokens to all other tokens on that host.

host	tokens	docs	paras	ratio
www.nrk.no	5,722,732	36,028	222,976	2.70%
www.gavpi.org	5,614,019	34,292	189,227	30.24%
bibel.no	4,465,546	18,395	37,212	15.99%
www.avvir.no	2,370,461	11,751	70,926	99.83%
samisketallforteller.no	1,747,008	1,562	22,570	30.35%
samilogutmuitalit.no	1,701,250	1,513	21,899	30.48%
www.regjeringen.no	1,469,519	6,304	45,292	1.03%
www.nsr.no	634,133	3,226	16,181	30.45%
samiskbibliotektjeneste.tromsfylke.no	623,469	5,284	16,402	8.26%
skuvla.info	516,893	339	8,666	28.03%
www.minaigi.no	502,330	1,344	3,815	99.78%
www.samediggi.fi	418,638	1,128	10,759	16.63%
www.kafjord.kommune.no	403,705	3,303	11,526	11.39%
sametinget.no	388,392	1,729	11,034	20.90%
www.samediggi.no	345,324	1,542	6,003	47.93%
www.unn.no	344,787	8,208	15,003	2.19%
samas.no	312,118	1,706	9,040	46.84%
samiallaskuvla.no	289,667	1,740	8,233	48.53%
allaskuvla.no	287,486	1,727	8,172	48.21%
samiskhogskole.no	286,837	1,721	8,165	48.23%

Table 3: Top 20 hosts for North Sámi

host	tokens	docs	paras	ratio
www.nrk.no	587,232	3,528	24,741	0.28%
www.nord-salten.no	218,426	1,083	4,720	2.03%
www.gavpi.org	207,666	4,376	6,857	1.12%
www.udir.no	132,606	450	3,170	0.61%
www.infonuorra.no	106,087	507	2,865	4.09%
samiskbibliotektjeneste.tromsfylke.no	83,875	1,212	2,438	1.11%
issuu.com	80,003	978	1,737	0.00%
nordsalten.no	63,317	257	855	3.34%
www.regjeringen.no	54,628	271	1,579	0.04%
sametinget.no	40,592	274	1,044	2.18%
www.samedigge.no	39,366	173	1,051	6.71%
giellatekno.uit.no	31,979	11	1,517	4.83%
www.calliidlagadus.org	29,160	192	192	55.80%
www.statped.no	24,749	476	810	0.19%
nsr.no	24,722	58	736	3.07%
arran.no	23,215	125	679	18.94%
samiskeveivisere.no	19,944	76	587	3.88%
forum.infonuorra.no	19,094	202	248	2.42%
www.hamaroy.kommune.no	19,062	194	584	1.42%
sapmi.arbeiderpartiet.no	16,241	9	634	12.42%

Table 4: Top 20 hosts for Lule Sámi

host	tokens	docs	paras	ratio
www.nrk.no	407,652	2,817	17,338	0.19%
www.gavpi.org	339,209	6,848	12,357	1.83%
samisketallforteller.no	311,033	807	5,188	5.40%
samilogutmuitalit.no	299,630	786	4,970	5.37%
bibel.no	289,371	1,106	1,194	1.04%
www.infonuorra.no	281,593	1,377	9,034	10.85%
www.udir.no	208,112	554	5,286	0.96%
samiskbibliotektjeneste.tromsfylke.no	102,095	1,595	3,636	1.35%
www.regjeringen.no	92,655	657	2,594	0.07%
www.raadstua.no	84,800	1	684	50.46%
www.frisktigers.no	84,450	1,110	1,114	0.02%
sametinget.no	81,606	500	2,217	4.39%
issuu.com	78,694	593	1,183	0.00%
www.saemiedigkie.no	76,659	301	2,096	41.14%
www.statped.no	29,263	566	1,044	0.22%
www.statsforvalteren.no	26,124	206	798	0.11%
nsr.no	24,005	75	670	2.98%
ovttas.no	23,335	558	667	13.96%
www.kirken.no	21,615	126	632	0.11%
www.kyrkja.no	21,146	124	617	0.13%

Table 5: Top 20 hosts for South Sámi