Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 2108–2116
Marseille, 20-25 June 2022

Assessing the Multilinguality of Publicly Accessible Websites

Rinalds Vīksna[1, 2], Inguna Skadiņa[1, 2], Raivis Skadiņš[1,2], Andrejs Vasiļjevs[1,2], Roberts Rozis[1]
[1]Tilde / Vienības gatve 75a, Riga, Latvia
[2]Faculty of Computing, University of Latvia / Raiņa bulv. 19, Riga, Latvia
{name.surname}@tilde.lv

Abstract

Although information on the Internet can be shared in many languages, the language presence on the World Wide Web is very disproportionate. The problem of multilingualism on the Web, in particular access, availability and quality of information in the world's languages, has been the subject of UNESCO focus for several decades. Making European websites more multilingual is also one of the focal targets of the Connecting Europe Facility Automated Translation (CEF AT) digital service infrastructure. In order to monitor this goal, alongside other possible solutions, CEF AT needs a methodology and easy to use tool to assess the degree of multilingualism of a given website. In this paper we investigate methods and tools that automatically analyse the language diversity of the Web and propose indicators and methodology on how to measure the multilingualism of European websites. We also introduce a prototype tool based on open-source software that helps to assess multilingualism of the Web and can be independently run at set intervals. We also present initial results obtained with our tool that allows us to conclude that multilingualism on the Web is still a problem not only at the world level, but also at the European and regional level.

Keywords: multilingualism, internet, scoring, language equality

1. Introduction

The Internet today is multilingual and diverse, it offers many opportunities for sharing information and knowledge. Although the information on the Internet could be shared in all languages that meet certain technical requirements (e.g., established writing system supported in character encoding standards and keyboard layouts), language presence on the World Wide Web is very disproportionate. For example, according to the W3Techs statistics[1] about 80% of Web content is in 5 languages: 63.6% in English, 6.9% in Russian, 3.9% in Turkish, 3.6% in Spanish and 3.5% in Persian (Farsi), while all other languages account for only 19% of content.

Asymmetries in the volume of online content for different languages can be also observed for Wikipedia[2] with 9 languages representing 47.7% of the content (11.1% in English, 10.5% in Cebuano, 4.7% in Swedish, 4.6% German, 4.1% French, 3.6% Dutch, 3.1% Russian, 3% Spanish and Italian), while the remaining 305 languages account for only about 52% of content. Young (2012) in the study about the multilinguality of Wikipedia articles found that 74% of concepts had articles in only one language and 95% of concepts were in fewer than six languages.[3]

Two recent studies on the multilingual leaders on the Web by CSA Research (Lommel and Sargent, 2019; Sargent and Lommel 2019) examined the language of content by industry. "The Global Website Assessment Index 2019" (Sargent and Lommel 2019) documents languages and social network links on 2,817 of the world's most prominent websites in 37 industry sectors. The study reaffirms the position of English as the global lingua franca, also showing the rapid rise of Chinese and the continued concentration of website localization into core economic languages even as the long tail seen on the most multilingual sites expands. According to CSA "industries that consistently deploy the greatest average number of languages include Automotive, Computer and Electronics, and Consumer Goods. Conversely, the most multilingual brands fall in the social networking and online services categories, even though overall averages in these categories are lower. Leading the way are Google (146 languages), Facebook (141 languages), and Vkontakte (85 languages)." As stated by UNESCO, it is obvious today that "nations, communities and individuals without access to the Internet and its resources will certainly be marginalized with limited access to information and knowledge, which are critical elements of sustainable development."[4] Not only access, but also quality of content, especially at the local level and in local languages is very important. Cultural diversity and multilingualism on the Internet have a key role to play in fostering pluralistic, equitable, open, and inclusive knowledge societies. Therefore, UNESCO encourages its member states to develop comprehensive language-related policies, to allocate resources and use appropriate tools to promote and facilitate linguistic diversity and multilingualism, including on the Internet and in the media.

Linguistic diversity is a fundamental value of the European Union. According to Article 3 of the Treaty on European Union (EU) the Union "shall respect its rich cultural and linguistic diversity". The EU Digital inclusion policy aims to ensure that everybody can contribute to and benefit from the digital world. Therefore, the EU is taking actions to promote multilingualism online. Making European websites more multilingual is one of the targets of the Connecting Europe Facility Automated Translation (CEF AT) digital service infrastructure. In order to monitor this goal, alongside other possible solutions, CEF AT is looking for a methodology and a fully or semi-automatic, easy to use tool to assess the degree of multilingualism of a given website.

In this paper we describe our work to address this need that was carried out in the framework of the ELRC action[5] commissioned by CEF AT.

[1] December, 2021. https://w3techs.com/technologies/overview/content_language

[2] January 10, 2022. https://en.wikipedia.org/wiki/List_of_Wikipedias

[3] http://labs.theguardian.com/digital-language-divide/

[4] https://en.unesco.org/themes/linguistic-diversity-and-multilingualism-internet

[5] https://www.lr-coordination.eu/

Our task was to investigate methods and tools that automatically analyse language diversity on the Web and propose indicators and a methodology on measuring the multilingualism of European websites. In our study we found that only a few research papers analyse websites with respect to their multilinguality (Miraz et al. 2013, Minacapilli 2018, Lee and Choi, 2019). Some research is devoted to the usability and multilingual user experience (Miraz et al.,2013). [6, 7 , 8]

Besides proposed criteria for website multilingualism we present a basic scoring tool developed based on open-source software.[9] The tool can be run on a given list of websites at set intervals to monitor changes in the language coverage over time.

2. Background Studies on Defining Multilingualism

While a multilingual website is usually defined as a website that uses more than one language, the notion of multilingualism on the Internet could refer to several concepts (Leppänen and Peuronen, 2012; Androutsopoulos 2006 and 2007):

- the diversity of languages as a means of communication on the Internet (analysis of their visibility, accessibility and status; Wright 2004; Danet and Herring 2007),
- the practices of multilingual Internet users and the ways in which they draw on and use resources provided by more than one language in their computer-mediated communication (Lee, 2017).

Several methods have been proposed for investigating multilingualism on the Internet by measuring how visible and accessible particular languages are. One method to measure linguistic diversity online is to **survey** what Internet users report on their language choices. This approach was chosen by the research team commissioned by UNESCO (Wright, 2004), who administered the same survey to students of English in ten countries (Tanzania, Indonesia, the United Arab Emirates, Oman, France, Italy, Poland, Macedonia, Japan, and Ukraine).

Another set of studies commissioned by UNESCO (2005) attempted to investigate multilingualism by suggesting the use of **quantitative measures** to study linguistic diversity online. For instance, Paolillo (2005) recommended the use of a linguistic diversity index, a statistical measure which, as part of the measurement of the languages used online, can take into account the variety of languages and the proportion of a particular language group in relation to other language groups of any one country.

The Language Observatory project, founded in 2003, aimed to measure the use of each language on the World Wide Web by **counting the number of pages** on the Web written in each language (Mikami et al, 2005). The proposed tool contained two components: a crawler and a language identification instrument. It needs to be mentioned that the language identification tool was able to identify 184 languages with an average of 94% accuracy.

A widely used index of linguistic diversity is **Lieberson's diversity index (LDI)** (Lieberson, 1981). It is defined as:

$$\mathrm{LDI}=1-\sum P_i^2$$

where P_i represents the share of i-th language speakers in a community. LDI is 0, if community speaks one language, while higher LDI means larger linguistic diversity. Mikami and Kodama (2012) found that "In Europe, the highest LDI belongs to Belgium (0.75). It is followed by Bosnia (0.66), Serbia (0.63), Moldova (0.59), Italy (0.59), Latvia (0.58), Georgia (0.58), Macedonia (0.58), Switzerland (0.58), Albania (0.57), Andorra (0.57), Austria (0.54), Monaco (0.52), and Spain (0.51). These fifteen countries have an LDI over 0.5. Countries with a dominant mother language, such as Germany (0.37), Russia (0.33), the Netherlands (0.29), and France (0.27), generally have lower LDIs. The lowest in Europe is Hungary (0.02)."

Mikami and Kodama (2012) **propose a two-dimensional chart (LL-chart)** with the local language ratio on the horizontal axis and LDI on the vertical axis for measuring language diversity in the cyber world. The motivation for a two-dimensional approach was an observation that languages in cyberspace and languages in the real world have different proportions, as it is demonstrated in Table 1.

Most-used languages in the world		Most-used languages online by user	
English	15%	English	25.4%
Chinese	15%	Chinese	19.3%
Hindi	7.2%	Spanish	8.1%
Spanish	6.9%	Arabic	5.3%
French	3.8%	Portuguese	4.1%

Table 1: Most Used Languages of the World (Babbel Magazine [10]).

To estimate language usage on the Internet Gerrand (2007) proposes a taxonomy that distinguishes among user profile (the number or proportion of active Internet users in each language group), user activity, web presence (the number or proportion of web pages written in each language group), and diversity index as separate indicators of language diversity on the Internet, and further distinguishes between spoken and written languages.

Recently Kelly-Holmes (2019) analysed the evolution of language and technology in relation to multilingualism. She studied the ways in which languages are made available, supported, presented, represented, and managed in digital spaces. According to Kelly-Holmes we are cocooned in linguistic "filter bubbles" (Pariser, 2011), and we are often being steered through the global, multilingual web in a monolingual bubble based on past linguistic behaviour and choices, cocooned from other languages.

Finally, Gazzola et al. (2019) propose new indices to measure linguistic diversity to study the political and economic implications of linguistic diversity in multilingual countries.

[6] https://digital.gov/resources/top-10-best-practices-for-multilingual-websites/

[7] Managing multi-regional and multilingual sites https://developers.google.com/search/docs/advanced/crawling/managing-multi-regional-sites

[8] Chapter 10. Making Your Site Multilingual | Drupal User Guide guide on Drupal.org

[9] Current version of the tool is available at https://github.com/tilde-nlp/Multilingualism-scoring-tool

[10] https://www.babbel.com/en/magazine/internet-language

3. Principles and Criteria of Multilingualism Scoring

3.1 Criteria and Challenges

When assessing the multilingualism of websites, we must distinguish between multiple facets of linguistic and technical criteria. Different criteria have been offered, including usability and multilingual user experience. While some criteria can be detected automatically, others are obvious only to a human eye or to a professional linguist. In our study we identified five main criteria: language coverage, language balance, linguistic quality, technical quality, and content parallelism.

For **language coverage** calculation it is necessary to identify the languages used in the crawled webpages to create a simple summary of the language and page counts. Information could be provided in different granularities, e.g., EU languages, minority languages, etc. Although the task seems rather simple, development of such a tool faces several challenges – multiple languages per page, identification of lesser used languages, etc.

Many websites have unequal coverage of content in different languages with full content in a dominant language and only part of it in other languages. **Language balance** is a measure of evenness/balance of the content coverage in various languages. In the ideal case all the languages on the website would be represented equally. One of challenges for calculating language balance is granularity - the content in different languages can be quantified by webpages, by sentences, by words or by compounding these counts.

The **linguistic quality** criterion is applied to evaluate the linguistic quality of the content in a particular language. Grammatically correct, human-authored or reviewed content should receive the highest score. Low-quality machine-translated content or content with numerous linguistic (e.g., grammatical or spelling) errors should receive lower scores. Here the main challenge is a lack of automatic methods to detect and evaluate human-authored content.[11] Recent survey of manual and automatic methods for translation quality assessment (Lifeng et al., 2021) lists some promising approaches for further investigation.

Technical quality criterion assesses use of internationalization attributes and other technical aspects. This can be done by analyzing the respective HTML code, e.g., encoding, correctness of the language attribute, compliance with w3.org requirements[12], etc.

An important, but complicated criterion is **content parallelism**. This criterion assesses the degree of equivalence of the content in different languages. One approach would be to link parallel articles in different languages and distinguish between parallel and non-parallel articles or assess content comparability. Documents could be classified and scored as parallel, strongly comparable, weakly comparable or non-comparable (Su and Babych, 2012; Skadiņa et al, 2010) with the help of tools that align multilingual content and assess their level of parallelism (e.g. Pinnis et al, 2012; Su and Babych, 2012).

Another (complementary) approach would be to check for parallelism of navigation. Parallelism of navigation provides the means for switching the language of the website and other multilingual functionality (e.g., supporting identification of translated content by the presence of language id in a link, serving content in separate domain per language using the same or comparable menu structure, and presence of live MT tools/website translators).

3.2 Proposed Evaluation Criteria and Scores

The Multilingualism Scoring Tool calculates multiple metrics. First, we calculate Lieberson's diversity index as it is a widely used index of linguistic diversity. We also introduce scores for language coverage and language balance. We wanted the first version of the tool to be fast, easy to use, and able to handle a large number of websites. Therefore, we do not calculate scores for the linguistic quality, technical quality, and content parallelism in the current version of the tool, as calculation of these metrics requires the use of language-specific tools (e.g., spelling checkers) and much greater computing power. This version of the tool aims to evaluate the overall picture of multilingualism.

Lieberson's diversity index represents how content is distributed in various languages and how many languages are present on a website. When content is in one dominant language, LDI is 0, and as the count of documents in various languages increases, LDI approaches 1. We calculate LDI both on a per-page and per-word basis to detect situations when text in the dominant language is long, while in other languages text may be shorter (e.g., only summaries).

Language coverage represents how many languages are present on a site from a preselected list of languages. By default, we list the 24 EU languages, while also providing a score which includes Icelandic and Norwegian (both Bokmal and Nynorsk).

Language balance is calculated by finding the share of each language against the dominant language, where the dominant language is the one with the largest number of pages, and then calculating an average share. Let n be the number of languages per website, and $pages(i)$ be the number of pages in the language i, then

$$\max_{i\in[1..n]} pages(i)$$

is the number of pages in the dominant language, and the language balance is calculated by the following formula:

$$\frac{1}{n}\sum_{k=1}^{n}\frac{pages(k)}{\max_{i\in[1..n]} pages(i)}$$

For a perfectly balanced website (having an equal number of pages in each language), the language balance will be 1, for a very unbalanced website with a large number of pages in one language and just 1 page per other languages, the language balance approaches $1/n$.

Normalised language balance is the main multilingualism score calculated by the tool. It represents both how many EU languages are found in a site and how equally the content is distributed between languages. This score is obtained by multiplying the language coverage and the language balance. This score may take a value between 0 –

[11] A very simple approach for translation quality would be to apply three categories: machine translation, manual review, professional translation (see https://weglot.com/translation-quality/)

[12] https://www.w3.org/International/articlelist

no content in any of EU languages to 100 – content is present in equal amounts in all 24 EU languages.

The following example illustrates calculation of the proposed multilinguality metrics. Let's consider an example website having a total of 10 pages: one page in English, one in German, six in French, one each in Latvian and Lithuanian (Table 2.)

Language	Page count	Share from Max	Share from total (P_i)	Squared share from total (P_i^2)
English	1	0.1667	0.1	0.01
German	1	0.1667	0.1	0.01
French	6	1.0000	0.6	0.36
Latvian	1	0.1667	0.1	0.01
Lithuanian	1	0.1667	0.1	0.01

Table 2: Example webpage language statistics.

The language coverage is calculated as a fraction of the EU languages present on this website, and, as all five languages are European Union official languages, language coverage is 5/24.

To calculate the share of each language from the dominant language, we divide the number of pages in a given language with the number of pages in language with largest number of pages. In our example French has the largest number of pages – 6, while English is used in one page. As a result, English receives a "share from max" 1/6 = 0,1667 and French itself gets a "share from max" 6/6 =1. We calculate this share for all the languages in the website, and then we calculate the language balance as an average, i.e.,

(0.1667 + 0.1667 + 1 + 0.1667 + 0.1667) / 5 = 0.3333.

When only European languages are used (and the rest ignored) for calculating these scores, we label it as "language balance EU24" to emphasize that some languages are excluded from this score.

To calculate Lieberson's diversity index, we calculate the share of webpages (or words) in each language from the total number of webpages (or words). In an example of Table 2, the total number of pages is 10. Content in English is on one page, so English would have a share of 1/10=0.1 and this share is squared in LDI formula, so P^2 for English is 0.01. For French P=6/10=0.6 and P^2=0.36. P_i^2 values are added together and subtracted from 1 to obtain final LDI score, in our example, LDI = 1 – (0.01+0.01+0.36+0.01+0.01) = 0.6

The final normalised language balance score (taking into account EU24 languages) is calculated as a product of the language balance and language coverage, in our case – 0.3333 * 5/24 = 0.069. Finally, we multiply the language balance, LDI and normalized language balance score by 100 before showing it to the user.

4. Proposed Architecture and Main Tools

4.1 General Overview

The architecture of the proposed Multilingualism Scoring Tool is shown in Figure 1. The crawler uses a list of the URLs given by the user to make the initial requests to the websites and receive response objects. Using the received response objects additional links are extracted for further crawling until the desired depth has been reached.

Received response objects are also sent to the content processor module for text extraction, language detection and further analysis. The extracted text is used for language detection; the language code and metadata about the received page are saved into .TSV files for further analysis.

The report builder module is responsible for providing access to the collected statistics and calculating the final score. The .TSV files produced in the previous step are used to calculate final score and metrics mentioned in Subsection 3.2. The collected statistics and a final score are presented to the user using a web interface (Figure 3) and available to download as .CSV files for further analysis.

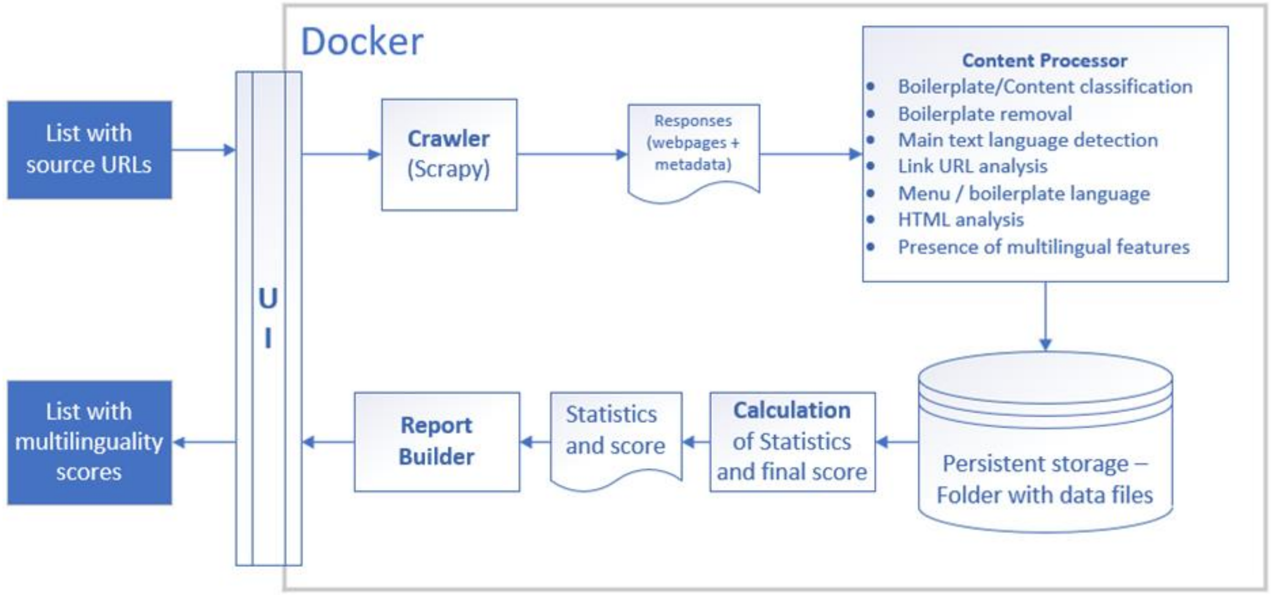

Docker
List with source URLs
U I
Crawler (Scrapy)
Responses (webpages + metadata)
Content Processor
Boilerplate/Content classification
Boilerplate removal
Main text language detection
Link URL analysis
Menu / boilerplate language
HTML analysis
Presence of multilingual features
List with multilinguality scores
Report Builder
Statistics and score
Calculation of Statistics and final score
Persistent storage – Folder with data files

Figure 1: Proposed architecture of the Multilingualism Scoring Tool.

4.2 Implementation

The proposed architecture has been implemented using the Scrapy crawler 2.4. Scrapy is written in Python, has extensive documentation[13], it can be easily extended using customized middleware, and can process multiple sites simultaneously. The Scrapy crawler is modified using a custom spider class to produce HTTP responses, which are fed to the content processor module.

The Content Processor module parses the responses from Scrapy to extract the necessary data – text content, HTML structure, and other features present on the webpage. To extract textual content, we use the jusText library[14] (Pomikálek, 2011), which finds and removes boilerplate content, such as navigation links, headers, and footers from HTML pages. It is designed to preserve mainly text containing full sentences and it is therefore well suited to extract plain text from the given webpage. The boilerplate content is not analysed and is discarded.

The extracted text is used to determine the main language of the webpage, using the langdetect library[15], which supports all official European languages. The detected language id together with metadata (URL, relative depth of the webpage, time of crawl, word count of text) are saved to a file in the analysis directory as a new row. File names correspond to the domains crawled, with "." replaced by underscores. In the current version, a folder with files with filenames derived from the domain name is used as a database for simplicity.

Crawling and analysing each domain is independent of other domains and can be done in parallel. Currently, the tool uses the Scrapy crawler set to crawl and analyse 16 parallel domains, however, this number could be set higher depending on available hardware.

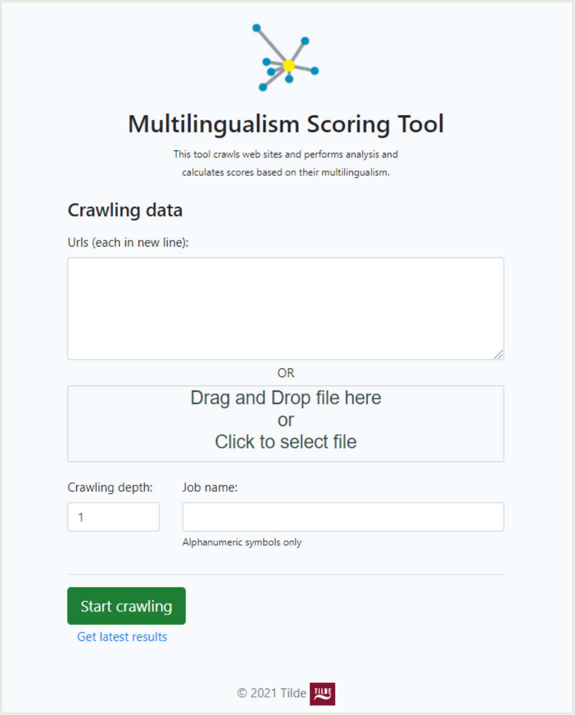
At any point during or after a crawl, the Reporter class may be called to analyse the data collected. The Reporter instance iterates through all the requested domain names and calculates aggregated statistics. To calculate the statistics of a website, the corresponding file is read line by line and statistics (language, word count, etc.) about the webpages are aggregated. Using statistics about pages in a website, different metrics are calculated: language balance, Lieberson's diversity index, and normalised scores. We also keep count of webpages without considerable textual content.

The tool is designed to allow score calculation considering only languages belonging to one of three groups: official EU languages, European Economic Area languages or all languages detected. Default values may be changed by editing the configuration file "report_settings.ini". Language groups could be considered when calculating the normalised language balance – the final score, as well as language balance EU24, language balance EU24 plus Icelandic and Norwegian languages, coverage EU24 and other metrics. The tool uses ISO 639-1 codes for language identification and export to .CSV format.

4.3 Interface

For ease of use, the tool has a simple web interface (Figure 2), where users can enter a list of websites to analyse and see the results as they are crawled (Figure 3).

The main application is started using the module app.py, which starts the tornado[16] web server used for the UI. This module creates an application which calls the class ScoringTool that manages the crawling, calculation, and presentation of the aggregated statistics. A Docker container is used for portability and easy deployment of the tool.



Multilingualism Scoring Tool

This tool crawls web sites and performs analysis and calculates scores based on their multilingualism.

Crawling depth: 1

Figure 2. Multilingualism scoring tool UI (before crawl).

To start the tool, a list of seed URLs is required. Seed URLs are read line-by-line; therefore, they must be separated by a newline (i.e., each link should be on its own line). The best practice is to enter multiple seed URLs (one for each language) for each domain to get the best results using limited crawling depth. URLs should also be full, containing a protocol (http:// or https://). We found that subdomains are often used to provide translated versions of pages, therefore subdomains are considered part of the domain. On the other hand, websites ending with different suffixes (e.g., example.eu, example.de) are considered two different domains. While in some cases this approach is also used to switch languages, more often websites structured this way are intended to be multi-regional.

Crawling depth is pre-set as 1, this setting determines how many times the crawler will extract and follow links. Recommended values are in the range 2-4, with higher values giving more precise results but longer crawl times, and lower values giving faster crawl time but less precise results.

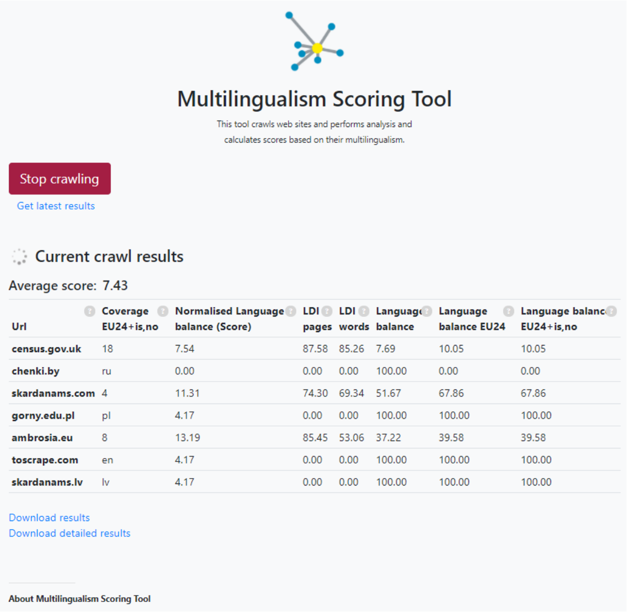
The job name field is used to specify the current job name. Crawl results will be named using this job name, saved for future reference, and prepared for download. The job name may consist of letters, numbers, and space symbols only.

[13] https://docs.scrapy.org/en/latest/index.html

[14] https://pypi.org/project/jusText/

[15] https://github.com/Mimino666/langdetect

[16] https://github.com/tornadoweb/tornado/



Multilingualism Scoring Tool

This tool crawls web sites and performs analysis and calculates scores based on their multilingualism.

Current crawl results

Average score: 7.43

Url	Coverage EU24+is,no	Normalised Language balance (Score)	LDI pages	LDI words	Languag balance	Language balance EU24	Language balanc EU24+is,no
census.gov.uk	18	7.54	87.58	85.26	7.69	10.05	10.05
chenki.by	ru	0.00	0.00	0.00	100.00	0.00	0.00
skardanams.com	4	11.31	74.30	69.34	51.67	67.86	67.86
gorny.edu.pl	pl	4.17	0.00	0.00	100.00	100.00	100.00
ambrosia.eu	8	13.19	85.45	53.06	37.22	39.58	39.58
toscrape.com	en	4.17	0.00	0.00	100.00	100.00	100.00
skardanams.lv	lv	4.17	0.00	0.00	100.00	100.00	100.00

Figure 3: Multilingualism Scoring Tool UI (during crawl).

4.4 Memory Requirements

Memory requirements of the tool are dependent on the size of a crawl. Observed RAM usage (using "sudo docker container stats" command) was below 1 GB during our smaller crawls, while on deep crawl (200 sites, 4 hops, more than 1.3 M pages downloaded) it reached 6.6 GB and could get higher on larger crawls. Downloaded pages are discarded after analysis, therefore, disk space required for metadata storage is on average 100 bytes per page crawled.

5. Some Results and Observations

The tool was evaluated by crawling the European Commission's website[17] and two lists of randomly selected URLs of European SMEs. The European Commission's website was crawled using depth 2 and as starting URLs using landing pages for all European languages to avoid favouring any one language. The list of 608 URLs was crawled in a depth of 2 only. The goal of this evaluation was to test the overall performance of the tool, usability of selected criteria, as well as to obtain some initial information regarding the multilinguality of European websites.

Results of both crawls are summarized in Table 4 and Figure 6. As expected, the deeper the crawl, the longer time it takes, the more pages are crawled, content in more languages is discovered, and the normalized language balance score increases. While crawling one of the sites with depth 4 our crawler hit a crawler trap, where the site generates a large number of unique links, causing our crawler to crawl in total 1,235,529 links from this site. This explains the large average number of words/pages per site observed in Table 4. As most of the auto-generated links point to the same content, such spider traps could be detected in future versions of the tool by identifying duplicates on the page's content level, using a hash function.

5.1 Observations from the ec.europa.eu Crawl

The European Commission's website is quite multilingual and contains fully functional landing pages (e.g., https://ec.europa.eu/info/index_es) for all European official languages. The goal of this evaluation was to find out if the tool can extract content in all European languages, as well as evaluate the scoring. The statistics of detected languages and webpages found are shown in Table 3. The normalised language balance score is calculated as 18.91, and Lieberson's diversity index is 92.48.

All EU official languages have about the same number of pages, except for English, for which the tool found significantly more content. The tool found and followed links to more technical parts of this website, where content is mostly in English (e.g., https://wikis.ec.europa.eu/). This extra content in English significantly lowers the normalised language balance score, while having little effect on the Lieberson's diversity index.

lang.	pages	lang.	pages	lang.	pages
bg	352	fr	347	nl	305
en	2031	ga	286	pl	301
es	361	hr	270	pt	307
cs	312	it	317	ro	359
da	303	lv	287	sk	284
de	370	lt	291	sl	263
et	296	hu	366	fi	276
el	374	mt	311	sv	249
		ca	4	no	2

Table 3: Languages and page count detected on europa.eu.

5.2 Observations from a Deep Crawl

For the first round of evaluation, we used a list of 198 URLs. The list was created randomly and included 72 .com websites, 13 .eu websites, 55 .it websites, 6 for .de, .es, .pt, 5 for .fr, .be, 4 for .pl, 3 for .lt, 2 for .bg, .dk, .et, .fi, .lv and one for .lu, .org, .wine and .srl.

The list of 198 URLs was crawled at depths of 1, 2, 3 and 4.

While performing the crawl containing 198 URLs, we analysed how many, and which sites were not crawled (Figure 7). Out of 198 URLs, 5 URLs were duplicates and were merged into a single domain by the tool. The remaining 193 URLs were crawled, and 170 domains had at least one HTML page with detected text content and were scored. At a crawling depth of 3, out of those 170 websites more than half had only one (54 websites) or two languages (65 websites), while 23 had three languages, 16 websites had content in 4 languages, 8 websites had content in 5 languages, and two websites had content in 6 or 7 languages (Figure 4). From websites that presented information at least in 5 languages (12 in total), 7 were .com sites, one .es and one .lv, while three for .it. All these websites had information in English, 9 had information in Spanish and German, 8 in Italian and French. From 54 websites that were identified as monolingual, most were in Spanish (9), French (9), Italian (11) or English (9).

[17] https://ec.europa.eu/

URLs	Crawling depth	Crawl time (hours)	Memory usage (MB)	Average number of languages (24 EU languages)	Language balance (all languages)	Normalised Language balance	LDI (pages)	Average number of pages/site	Average number of words/site
198	1	0:50	160	1.89	65.82	4.65	11.47	33	14516
198	2	12:54	1030	2.15	64.69	5.57	21.52	262	44592
198	3	48	1238	2.25	65.64	5.97	24.17	885	93242
198	4	>250	6672	2.29	66.59	6.08	24.83	8374	376787
608	2	52	1023	2.02	63.81	5.56	19.71	227	53663

Table 4: Results of 198 and 608 webpage crawl run.

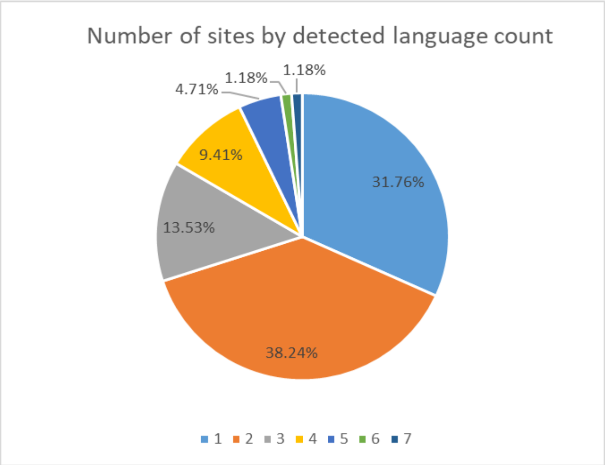


Figure 4: Breakdown of 170 successfully crawled sites by detected language count on sites.

23 domains were crawled, but the tool did not find any textual content. The breakdown of causes follows: 2 URLs contained a semicolon ";", which caused these links to be dropped by the tool; 7 URLs actually contained very little content (or had placeholders, like "In construction"); 8 URLs had redirects to other domains; Restrictive robots.txt, content served using JavaScript only, connection problems, and domain in capital letters were found as causes failure to extract text page each on single URL. On two sites Scrapy was unable to extract any links for unknown reasons (the only commonality was that both URLs used WordPress). Some of the failures (Figure 7, right) to score a domain are easily corrected by providing an up to date or correct URL (in case of redirect, and wrongly uppercased URL), as well as by making sure that input URLs are entered without quotes, semicolons, and other invalid characters. Websites that use JavaScript to display all content seem to be rare, in this crawl only one such website was encountered, and it had very little text. In other cases (no text, connection problems, robots.txt) there is little we could do to be able to crawl such sites. Websites that use different domains for translated or localized versions of the site are scored as separate sites and thus often monolingual. One such example is ekoseses.lt Lithuanian site (scored as Lithuanian monolingual) and ecosisters.eu site in English (not scored in this run).

5.3 Observations from a Wide Crawl

During a second evaluation round, the list of 608 URLs was crawled at a depth of 2. This crawl finished in 52 hours. Out of the given 608 links, 581 sites were extracted after deduplication, and of those, 87 sites were found to have no text. Selected URLs contained 218 .com sites, 160 .it, 29 .eu, 25 .pl, 20.es, 16 .pt, 19 .ro, 13 .de, 12 .fr, 10.be, and 5 .org and .bg, and many pages with 1-2 representatives of domain.

Similarly to the previous experiment, most of the sites were monolingual (185 sites) or bilingual (192), while the content of two sites was in 11 languages (both .com sites from the manufacturing industry), one in 8 languages (.it domain, winery), two in 7 languages and four in 6 languages (Figure 5). English and German were present in all 9 sites, while French on 7 sites, Spanish - 6, Italian – 5.

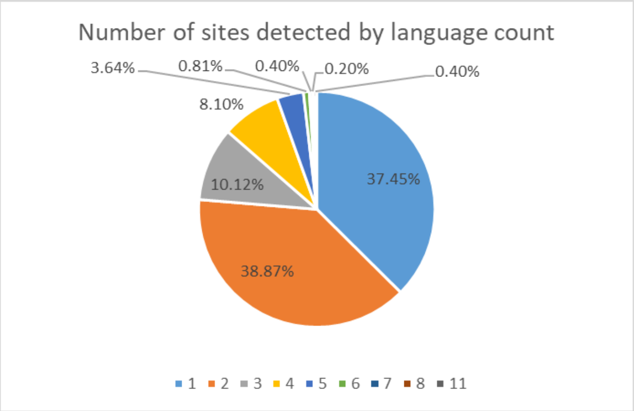


Figure *5:* Breakdown of 494 successfully crawled sites by detected language count on sites.

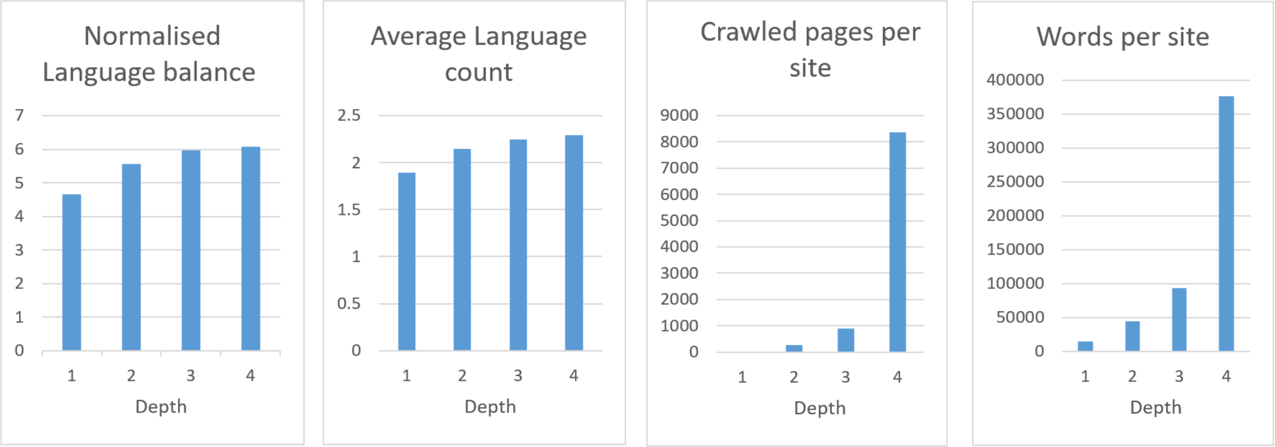


Figure 6: Results of 198 site crawl on various depth settings.

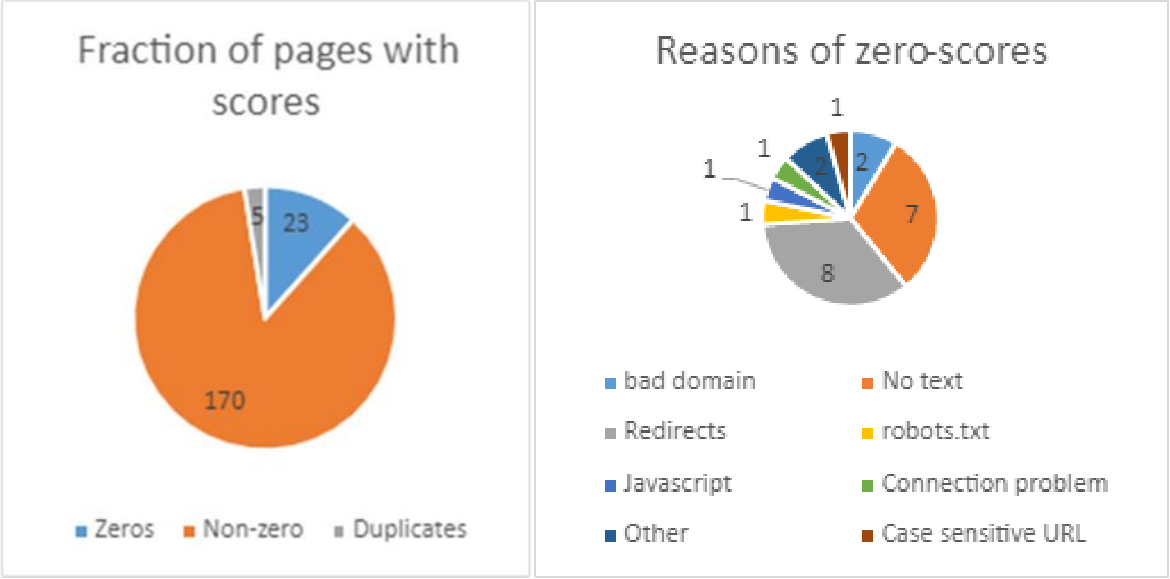


Figure 7: Failure analysis.

6. Conclusion

In this paper we discussed different methods and scores that can be applied to measure language diversity on the Internet. Our primary interest was to track and assess the multilinguality of EU websites.

To measure multilinguality, we created an open-source tool for scoring multilinguality[18] (Vīksna, 2021). The tool calculates several scores to measure multilingualism over the Web: Lieberson’s diversity index, language coverage, language balance and normalised language balance.

The tool can be used to track the multilinguality of a list of representative websites. It can crawl a large number of websites and provide some initial results almost immediately, updating scores as more pages are crawled. The time necessary for the complete crawling increases quickly as we try to crawl a website exhaustively.

The tool was applied to score a broad range of European websites. The scoring results show that today European SME websites on average provide content in only 2-3 languages. The normalised language balance score is in the range of 4.65-6.08 showing that most of the content is provided in only one or two languages per site.

Our plan for the future is to continue investigating possible ways to automatically assess more complicated multilingualism criteria, such as linguistic quality, technical quality and content parallelism, and implement them into next versions of the tool.

7. Acknowledgements

This work has been funded by the CEF Automated Translation programme project SMART 2019/1083 and the European Regional Development Fund research project “AI Assistant for Multilingual Meeting Management" No. 1.1.1.1/19/A/082.

8. Bibliographical References

Androutsopoulos, J. (2006). Introduction: sociolinguistics and computer-mediated communication. *Journal of Sociolinguistics*, 10(4): 419-38.

[18] https://github.com/tilde-nlp/Multilingualism-scoring-tool

Androutsopoulos, J. (2007). Language choice and code switching in German-based diasporic web forums. In B. Danet and S. C. Herring (eds) *The Multilingual Internet: Language, Culture, and Communication Online*, New York: Oxford University Press.

Danet, B. and Herring, S. C. (eds) (2007). The Multilingual Internet: Language, Culture, and Communication Online, New York: Oxford University Press.

Gazzola, M., Templin, T. and McEntee-Atalianis, L. J. (2019). Measuring Diversity in Multilingual Communication. SpringerLink

Gerrand, P. (2007). Estimating Linguistic Diversity on the Internet: A Taxonomy to Avoid Pitfalls and Paradoxes. *Journal of Computer-Mediated Communication*, 12 (4): 1298–1321.

Lifeng, H., Smeaton, A. and Jones, G. (2021). Translation Quality Assessment: A Brief Survey on Manual and Automatic Methods. In: *Proceedings for the First Workshop on Modelling Translation: Translatology in the Digital Age*, 15–33. Association for Computational Linguistics.

Kelly-Holmes, H. (2019). Multilingualism and Technology: A Review of Developments in Digital Communication from Monolingualism to Idiolingualism. *Annual Review of Applied Linguistics*, 39: 24 – 39. Cambridge University Press.

Lee, C. (2017). Multilingualism online. New York: Routledge.

Lee, T. H. and Choi. (2019). Multilingual access support evaluation guideline in the website of public library. iConference IDEALS @ Illinois: Multilingual access support evaluation guideline in the website of public library.

Leppänen, S. and Peuronen, S. (2012). Multilingualism on the internet. In Martin-Jones, M., Blackledge A., and Creese Angela. *The Routledge Handbook of Multilingualism.* London and New York: Routledge.

Lieberson S. (1981). Language Diversity and Language Contact: Essays by Stanley Lieberson. Stanford, California: Stanford University Press.

Lommel, A. and Sargent, B. B. (2019). Top Target Languages by Vertical Sector: 2019. CSA Research.

Mikami, Y. and Kodama, S. (2012). Measuring linguistic diversity on the Web. In: *Net.lang: towards the multilingual cyberspace.* Unseco.

Mikami, Y., Vigna, S., Zavarsky P., Rozan M., Suzuki, I., Takahashi, M., Maki, T., Ayob, I. N., Boldi, P. and Santini, M. (2005). The language observatory project (LOP). In*: Special interest tracks and posters of the 14th international conference on World Wide Web - WWW 05.*

Minacapilli, C. A. (2018). A Heuristic Evaluation of Multilingual Lombardy: Museums' Web Sites. Université de Genève.

Miraz, M. H., Ali, M. and Excell, P. (2013). Multilingual Website usability analysis based on an international user survey. In: *The proceedings of the fifth international conference on Internet Technologies and Applications* (ITA 13), 236-244.

Paolillo, J. (2005). Language diversity on the Internet: examining linguistic bias. In: UNESCO Institute for Statistics (ed.) *Measuring Linguistic Diversity on the Internet,* Paris: UNESCO.

Pariser, E. (2011). The filter bubble: What the Internet is hiding from you. Penguin UK.

Pinnis, M., Ion, R., Ştefănescu, D., Su, F., Skadiņa, I., Vasiļjevs, A., Babych, B. (2012). ACCURAT toolkit for Multi-Level Alignment and Information Extraction from Comparable Corpora. In: *Proceedings of System Demonstrations Track of ACL 2012*, 91-96, Jeju Island, Republic of Korea.

Pomikalek, J. (2011). Removing boilerplate and duplicate content from web corpora. Ph.D. dissertation, Masarykova univerzita, Fakulta informatiky, https://is.muni.cz/th/45523/fi_d/phdthesis.pdf.

Sargent, B.B. and Lommel, A. 2019. Global Website Assessment Index 2019. CSA Research.

Skadiņa, I., Aker, A., Giouli, V., Tufis, D., Gaizauskas, R., Mieriņa, M., Mastropavlos, N. (2010). A Collection of Comparable Corpora for Under-resourced Languages. In: *Proceedings of the Fourth International Conference Baltic HLT 2010*, Frontiers in Artificial Intelligence and Applications, Vol. 219, 161-168. IOS Press.

Su, F., Babych, B. (2012). Development and Application of a Cross-language Document Comparability Metric. In: *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012),* 3956-3962.

UNESCO Institute for Statistics (ed.) (2005) Measuring Linguistic Diversity on the Internet, Paris: UNESCO.

Wright, S. (2004). Introduction [to special issue on Multilingualism on the internet. In: *International Journal on Multicultural Societies* 6(1): 5-13.

Young, H. (2012). The digital language divide.How does the language you speak shape your experience of the internet? *Guardian*.

9. Language Resource References

Pomikálek, J. (2011). jusText. Distributed via LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University, http://hdl.handle.net/11858/00-097C-0000-000D-F696-9.

Vīksna, R. (2021). Multilingualism Scoring Tool. ELRC project, distributed via tilde-nlp github repository: GitHub - tilde-nlp/multilingualism-scoring-tool