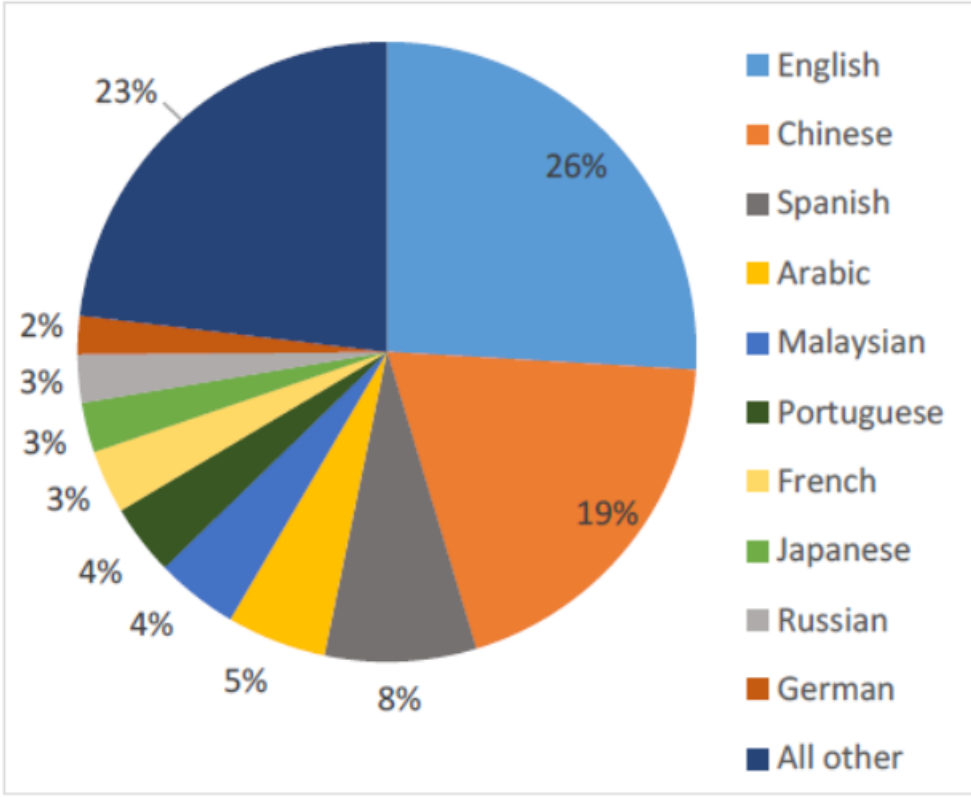
Assessing Multilinguality of Publicly Accessible Websites

Rinalds Vīksna, Inguna Skadiņa, Raivis Skadiņš,
Andrejs Vasiļjevs, Roberts Rozis

MOTIVATION

- Although Internet today is multilingual, language presence on the World Wide Web is very disproportionate.
- “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.” (UNESCO)
- UNESCO encourages its member states to develop comprehensive language-related policies (...) to promote and facilitate linguistic diversity and multilingualism, including on the Internet and in the media.



Source: Internet World Stats, 2020.

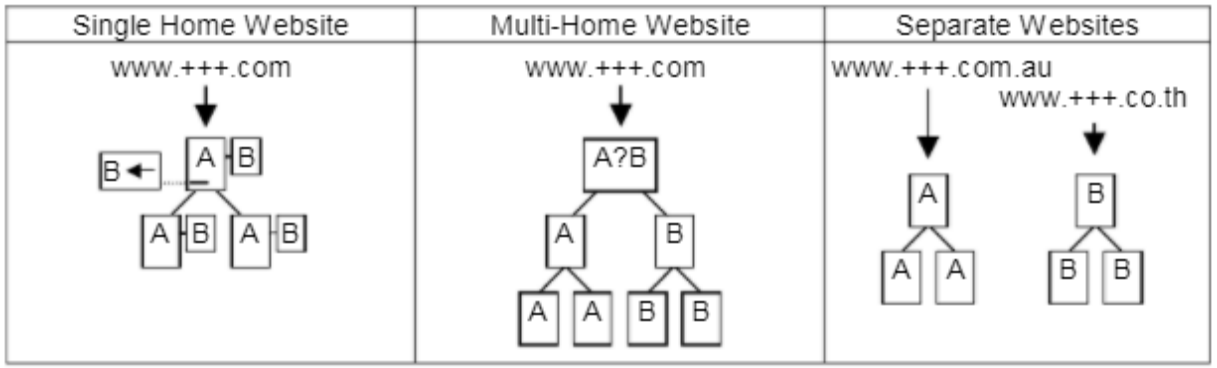
Languages used on the internet by share of internet users in 2020 | European Parliament

MOTIVATION

- Linguistic diversity is a fundamental value of the European Union.
- Making European websites more multilingual is one of the targets of the Connecting Europe Facility Automated Translation (CEF AT).
- CEF AT needs a methodology and a tool to assess the degree of multilingualism of a web site.
- We 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.
- We present a basic scoring tool developed based on open-source software.

MULTILINGUALITY AND MULTILINGUALISM

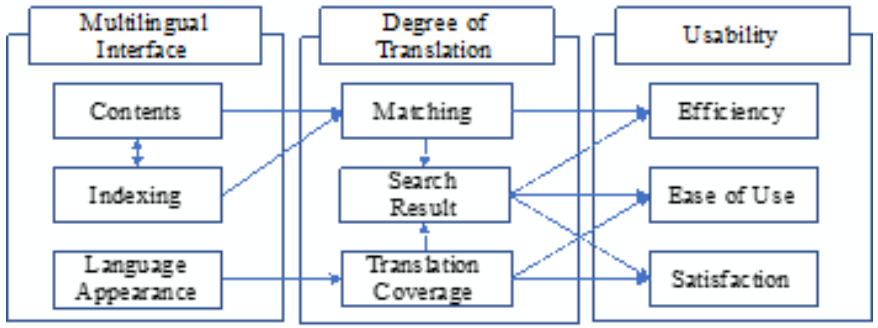
- Multilingual - "(of people or groups) able to use more than two languages for communication, or (of a thing) written or spoken in more than two different languages« (Wikipedia)
- «A “multilingual” web site refers to a web site that uses more than one language." (W3C)



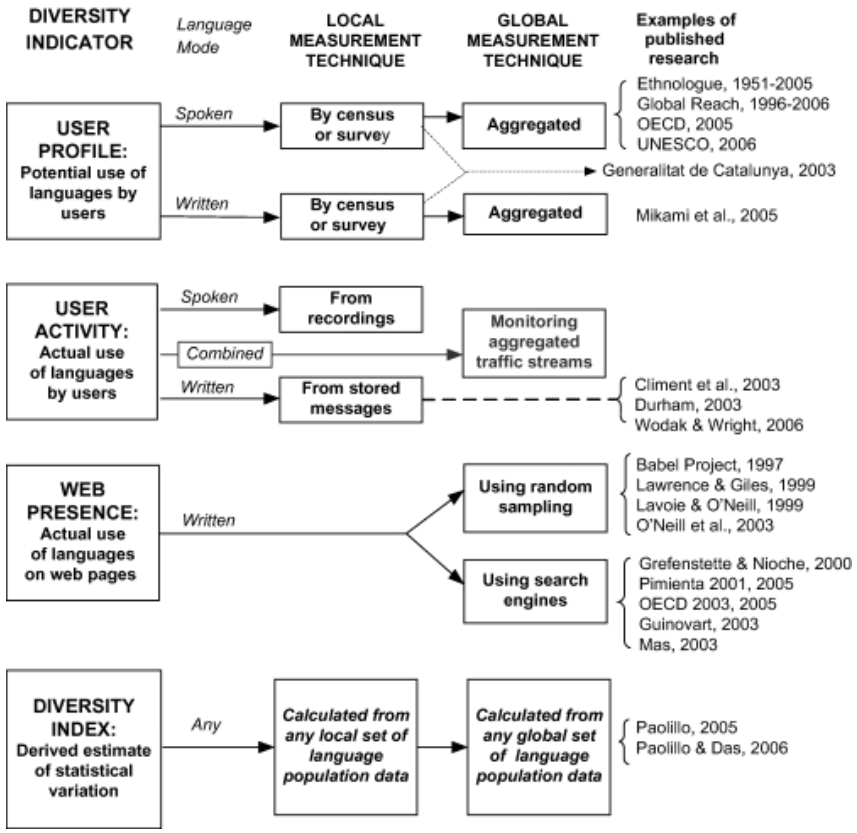
Three Broad Types of Multilingual Websites (Hillier, 2003)

BACKGROUND STUDIES

- Multilingualism on the Internet:
  - **the diversity of languages** as a means of communication on the Internet (analysis of their visibility, accessibility and status)
  - **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
- We found that only a few research papers analyse websites with respect to their multilinguality



Multilingual Evaluation Guidlines model proposed by Lee and Choi (2019)



A taxonomy of different methodologies used to estimate language diversity on the Internet (Gerrand, 2007)

Criteria of Multilingualism Scoring

- **Linguistic quality** criterion is applied to evaluate the linguistic quality of the content in a particular language
- **Technical quality** criterion assesses use of internationalization attributes and other technical aspects
- **Content parallelism** criterion assesses the degree of equivalence of the content in different languages

CRITERIA OF MULTILINGUALISM SCORING

- **Language coverage** represents how many languages are present on a site
- **Language balance** is a measure of evenness/balance of the content coverage in various languages
- **Normalized language balance** represents both how many EU languages are found in a site and how equally the content is distributed between languages
- **Lieberson's diversity index** (LDI) represents how content is distributed in various languages and how many languages are present on a website.

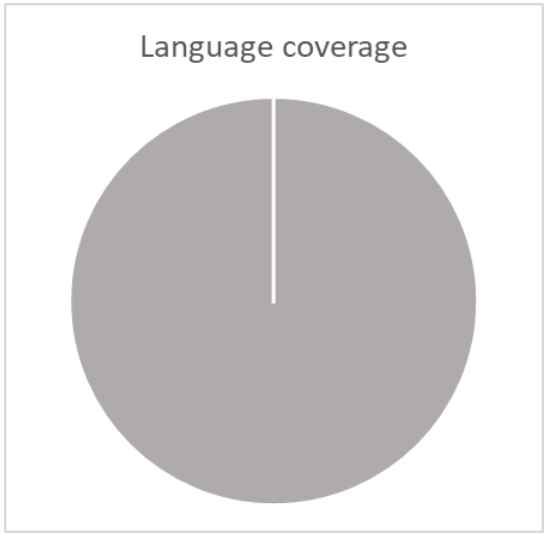
$$LDI = 1 - \sum P_i^2$$

where P_i represents the share of i-th language speakers in a community.

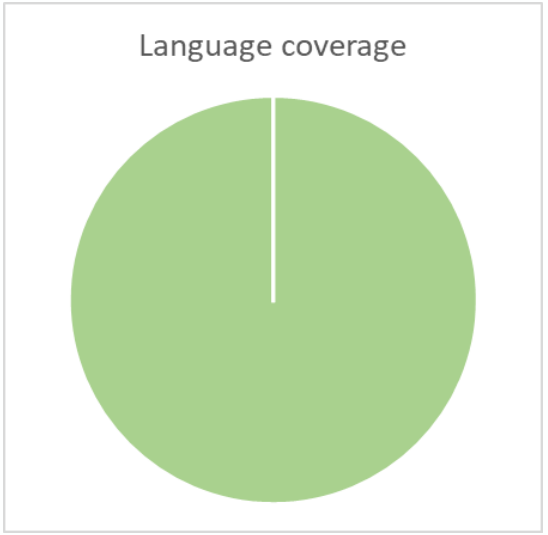
LANGUAGE COVERAGE

24 EU languages (extendable to 26 EEA languages or other set)

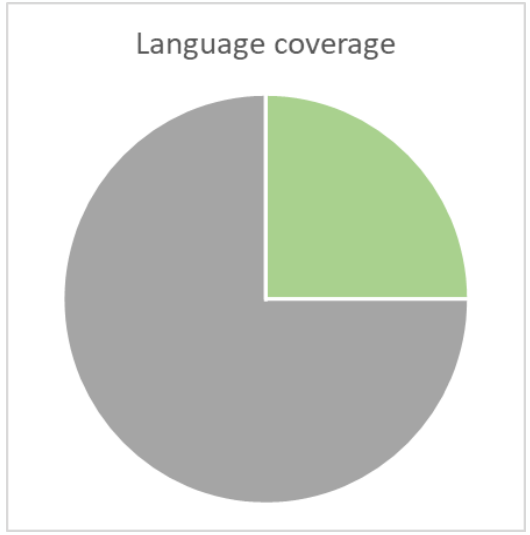
Each of EU languages found adds 1/24 (0,041667)



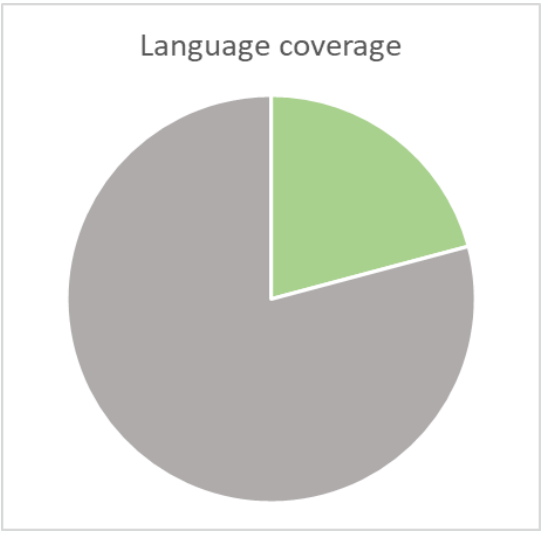
Minimum score 0.



Maximum score 1, all 24 EU languages present



Example website, 6 EU languages present – score 6/24=0,25



Example website, 5 EU languages present and one other language – score 5/24=0,2083

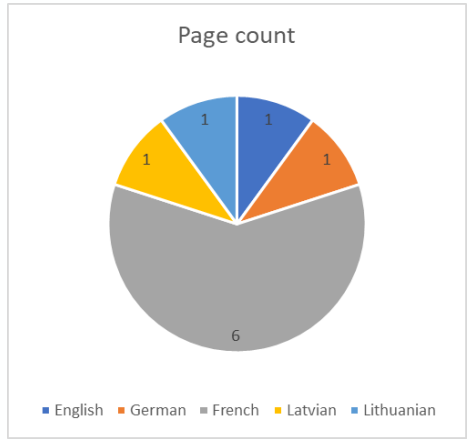
LANGUAGE BALANCE AND LDI

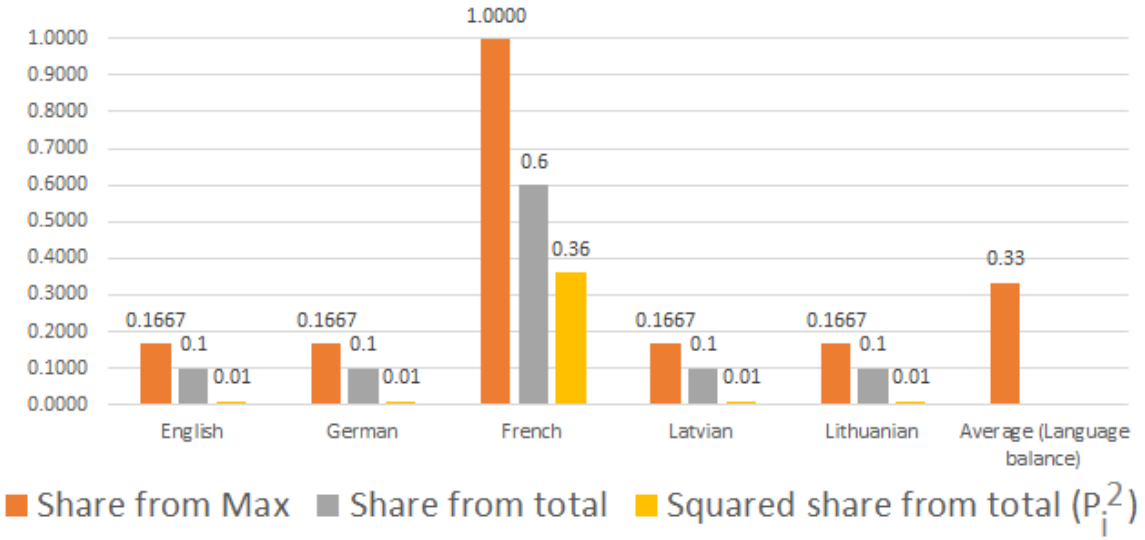
Language balance = (0.1667+0.1667+1.0000+0.1667+0.1667) / 5 = 0,3333

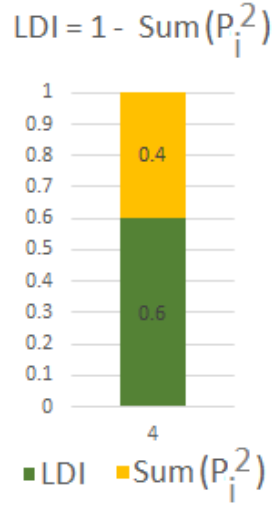
Normalised language balance = 5/24 * 0,3333 = 0.069

LDI = 1 – (0.01+0.01+0.36+0.01+0.01)= 1 - 0.4 = 0.6

Language	Page count	Share from Max	Share from total	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







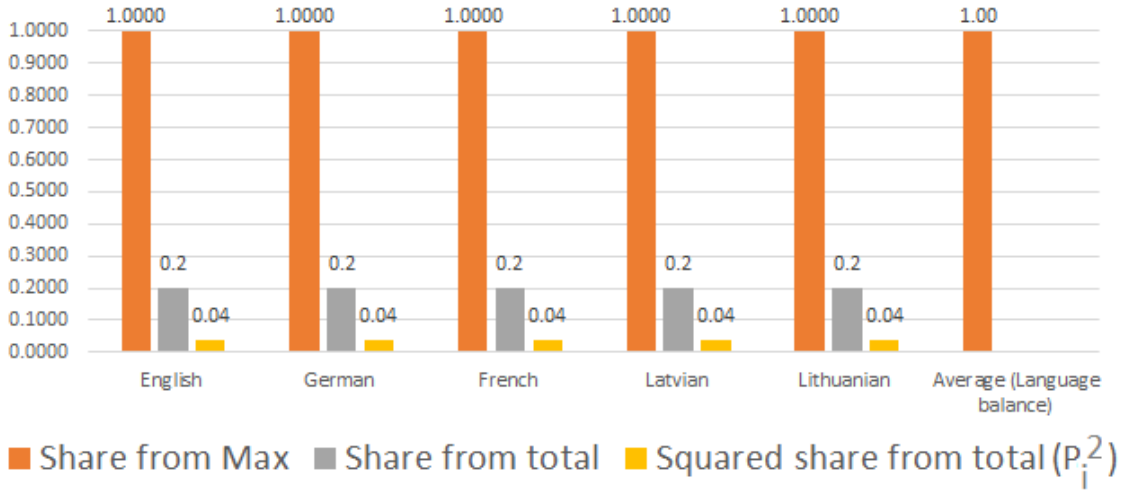
LANGUAGE BALANCE AND LDI

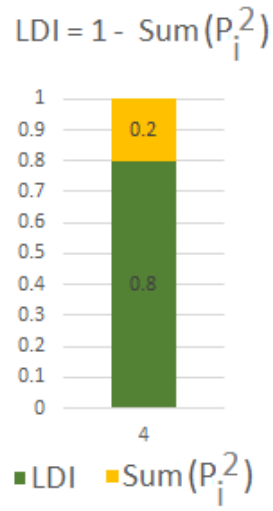
Language balance = (1.0000+1.0000+1.0000+1.0000+1.0000) / 5 = 1.0000

Normalised language balance = 5/24 * 1 = 0.208

LDI = 1 – (0.04+0.04+0.04+0.04+0.04) = 1 - 0.2 = 0.8

Language	Page count	Share from Max	Share from total	Squared share from total (P_i^2)
English	10	1.0000	0.2	0.04
German	10	1.0000	0.2	0.04
French	10	1.0000	0.2	0.04
Latvian	10	1.0000	0.2	0.04
Lithuanian	10	1.0000	0.2	0.04





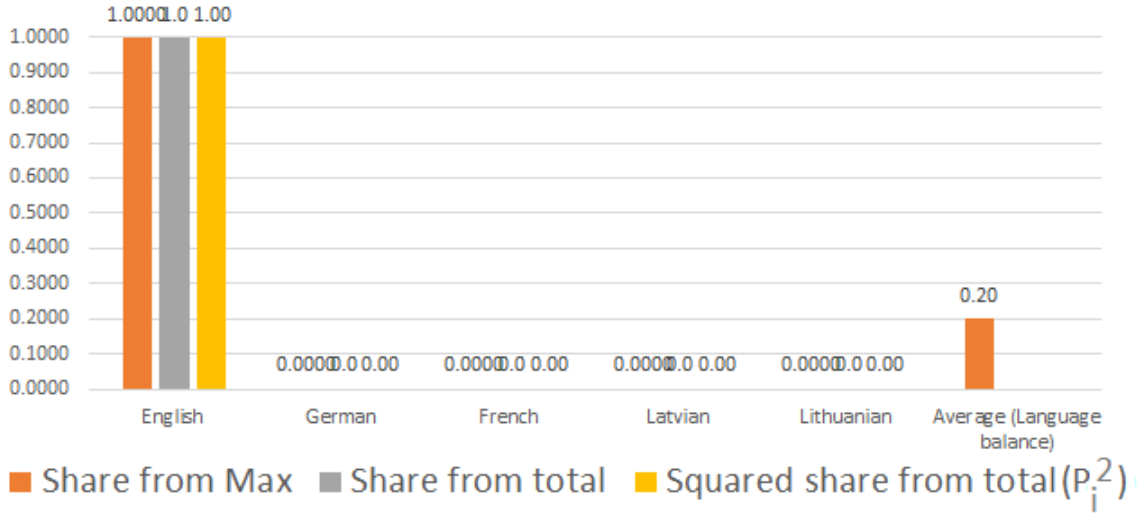
LANGUAGE BALANCE AND LDI

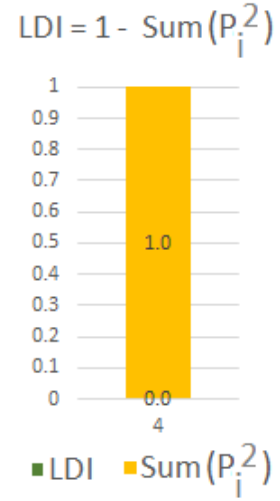
Language balance = (1) / 1 = 1

Normalised language balance = 1/24 * 1 = 0.0416

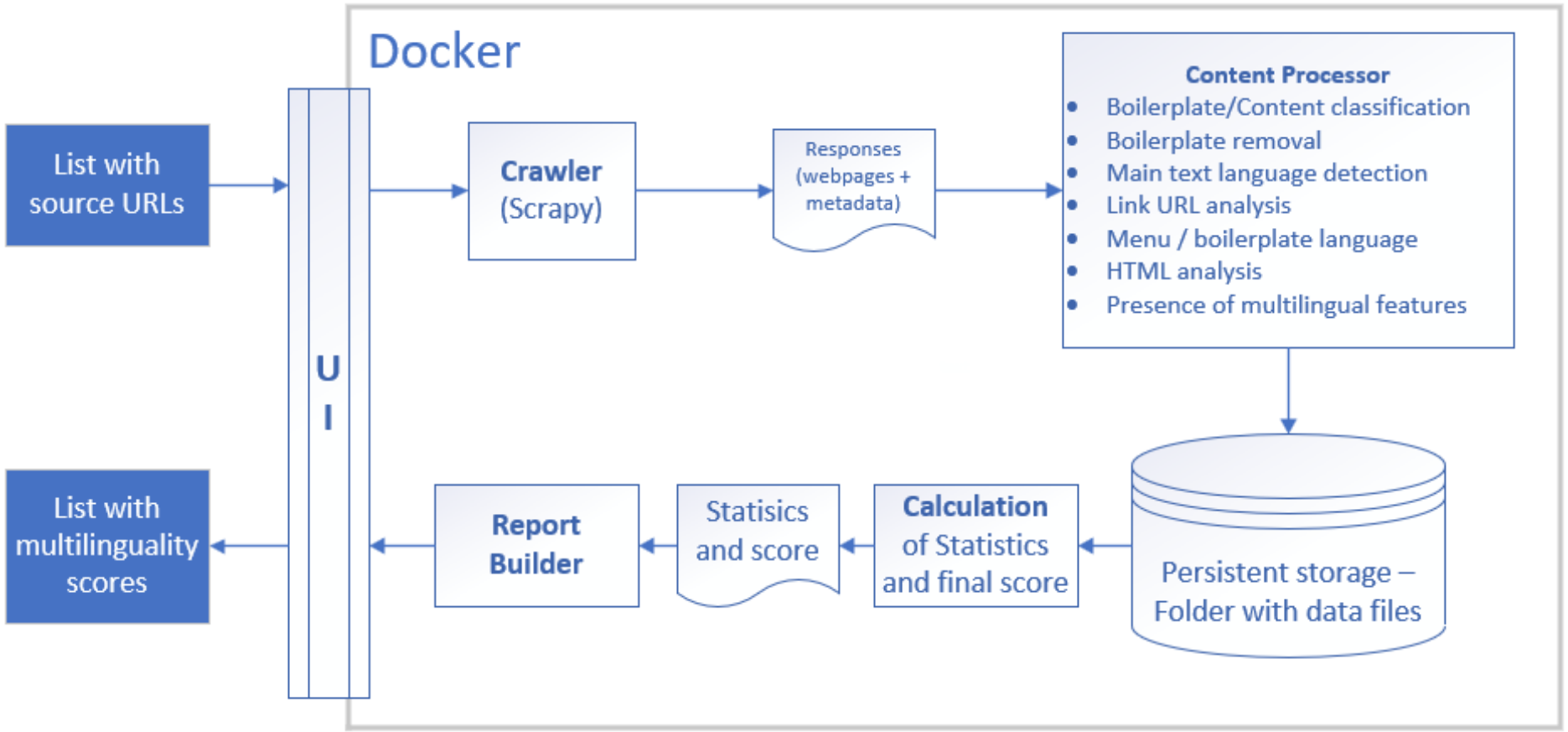
LDI = 1 – (1+0+0+0+0)= 1 - 1 = 0

Language	Page count	Share from Max	Share from total	Squared share from total (P_i^2)
English	10	1	1	1
German	0	0	0	0
French	0	0	0	0
Latvian	0	0	0	0
Lithuanian	0	0	0	0



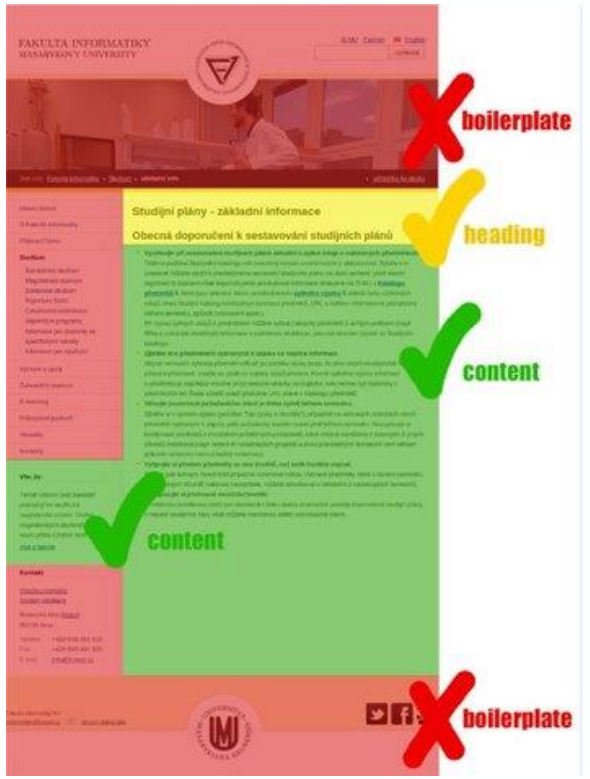


Architecture



COMPONENTS

- Crawler – Scrapy
- Boilerplate removal – jusText
- Language detection – LangDetect
- Scoring – 2 formulas:
  - Normalized language balance = sum(Share1, Share2, …, ShareN)/N
  - Lieberson's diversity index $LDI = 1 - \sum P_i^2$
    - where Pi represents the share of i-th language



THE TOOL

Multilingualism Scoring Tool

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

Crawling data

Urls (each in new line):

OR

Drag and Drop file here
or
Click to select file

Crawling depth: 1

Job name:

Alphanumeric symbols only

Start crawling

Get latest results

© 2021 Tilde

Multilingualism Scoring Tool

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

Stop crawling

Get latest results

Current crawl results

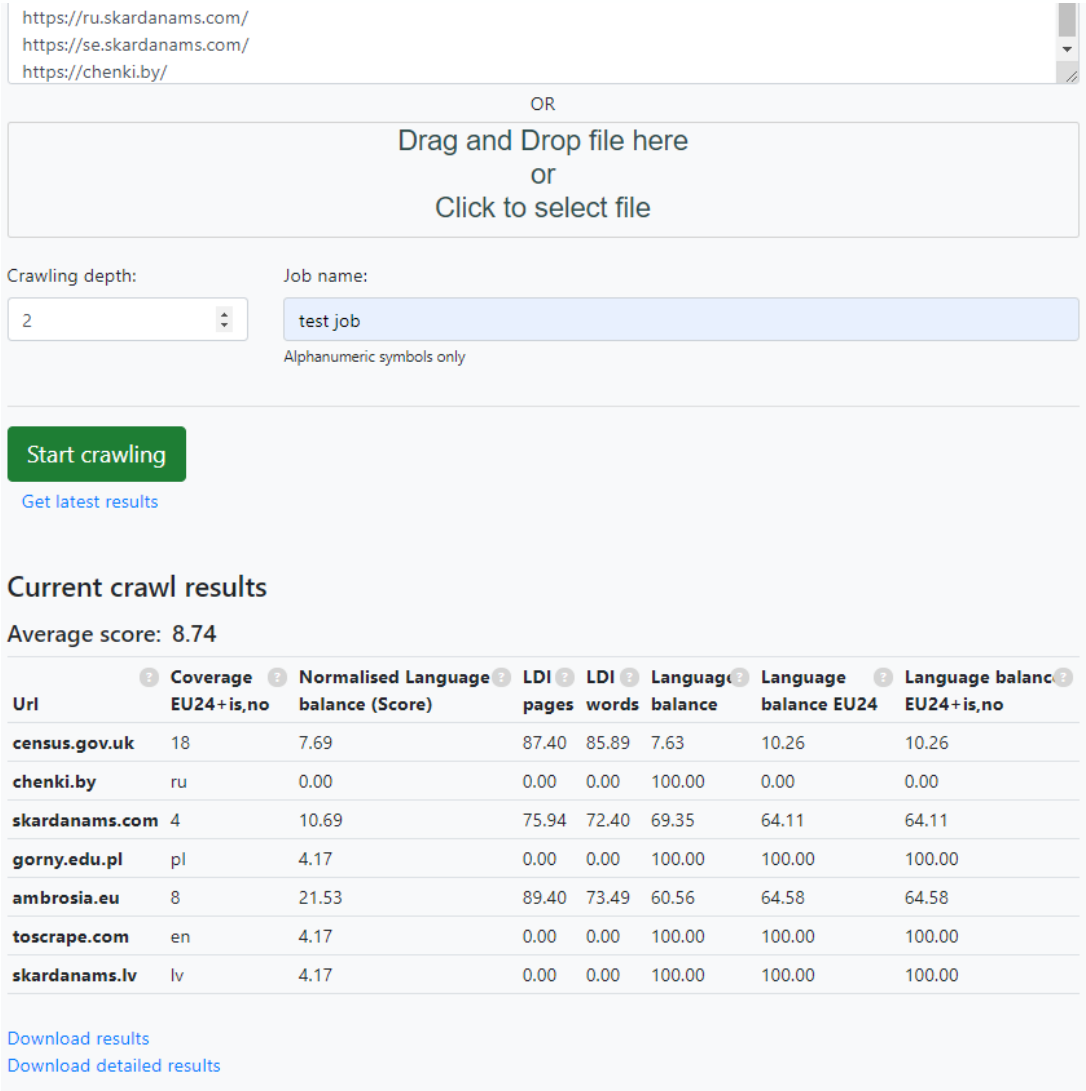
Average score: 7.43

Url	Coverage EU24+is,no	Normalised Language balance (Score)	LDI pages	LDI words	Language balance	Language balance EU24	Language balance 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

Download results
Download detailed results

About Multilingualism Scoring Tool

THE TOOL

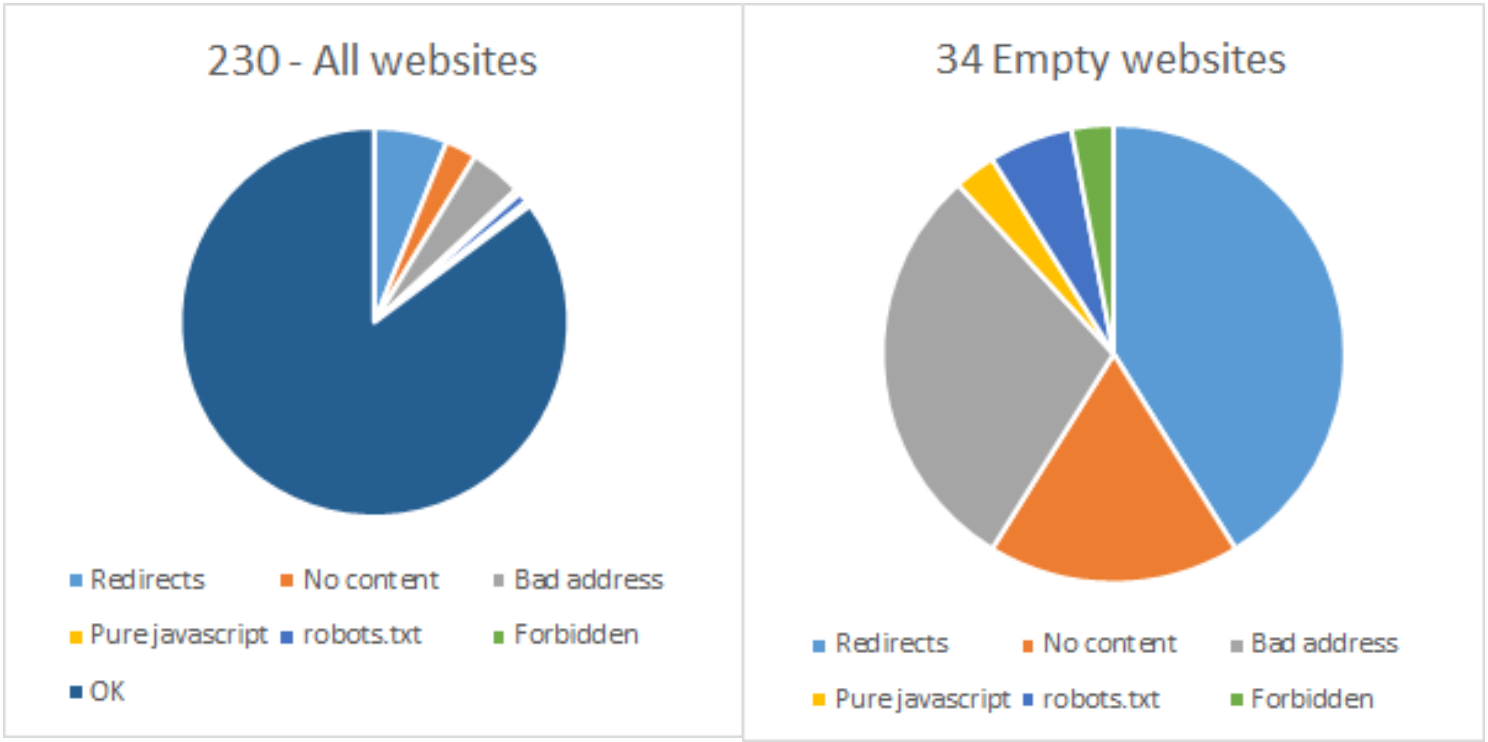


Url	Coverage EU24+is,no	Normalised Language balance (Score)	LDI pages	LDI words	Language balance	Language balance EU24	Language balanc EU24+is,no
census.gov.uk	18	7.69	87.40	85.89	7.63	10.26	10.26
chenki.by	ru	0.00	0.00	0.00	100.00	0.00	0.00
skardanams.com	4	10.69	75.94	72.40	69.35	64.11	64.11
gorny.edu.pl	pl	4.17	0.00	0.00	100.00	100.00	100.00
ambrosia.eu	8	21.53	89.40	73.49	60.56	64.58	64.58
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

Multilingualism Scoring Tool is available for download either as a code and Docker container on the github: https://github.com/tilde-nlp/Multilingualism-scoring-tool

Observations while using tool

- Crawled 230 random websites with depth 1 in two days
- Found 34 empty results:
  - Redirects 14
  - No content 6
  - Bad address 10
  - Javascript 1
  - Restrictive robots.txt 2
  - Forbidden 1
  - OK 196



Observations while using tool

- Number of prepared requests in each depth (before filtering)
  - 'request_depth_count/0': 14,
  - 'request_depth_count/1': 1000,
  - 'request_depth_count/2': 72004,
  - 'request_depth_count/3': 1259091,
- Stopped after ~110K requests
  - 'exception_count': 17784 detailed breakdown:

    'ValueError':2, 'IgnoreRequest':12784, 'CancelledError':1, 'ConnectionRefusedError':4929, 'DNSLookupError':54, 'TimeoutError':11, 'ResponseNeverReceived': 3,
  - Responses 92927:

    200 – 73096; 301 – 7534; 302 – 6338; 307 – 598; 404 – 228; 500 – 77, others <15 each

Observations while using tool

https://president.ee/et/

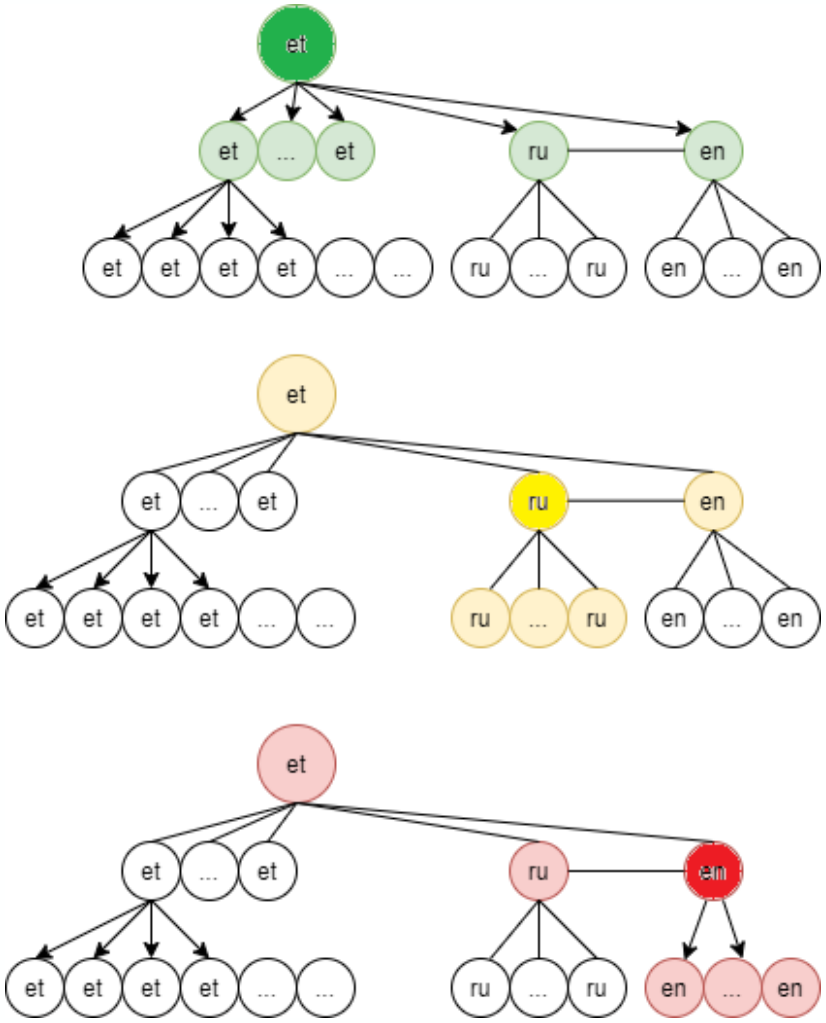
- 1 Hops: Pages: 45; w/o 8, et 35, ru 1, en 1
- 2 Hops: Pages: 2892; w/o 1274, et 1497, ru 39, en 72, fi 2, lv 1, uk 1, lt 1, el 1, ….
- 3 Hops: Pages: 8272; w/o 3670, et 3164, ru 463, en 910, fi 10, lv 4, uk 5, lt 3, ...
- 4 Hops: Pages: 13179; w/o 6566, et 3938, ru 857, en 1694, fi 24, lv 6, uk 8, lt 9, ...

https://president.ee/ru/index.html

- 1 Hops: Pages: 37; w/o 7, ru 27, en 1, et 2
- 2 Hops: Pages: 673; w/o 158, ru 434, en 38, et 42, uk 1
- 3 Hops: Pages: 5373; w/o 2147, et 1540, ru 824, en 812, uk 5, fi 5, lv 3, lt 2, ...
- 4 Hops: Pages: 12031; w/o 5980, et 3367, ru 898, ne 1670, uk 8, fi 19, lv 6, ...

https://president.ee/en/index.html

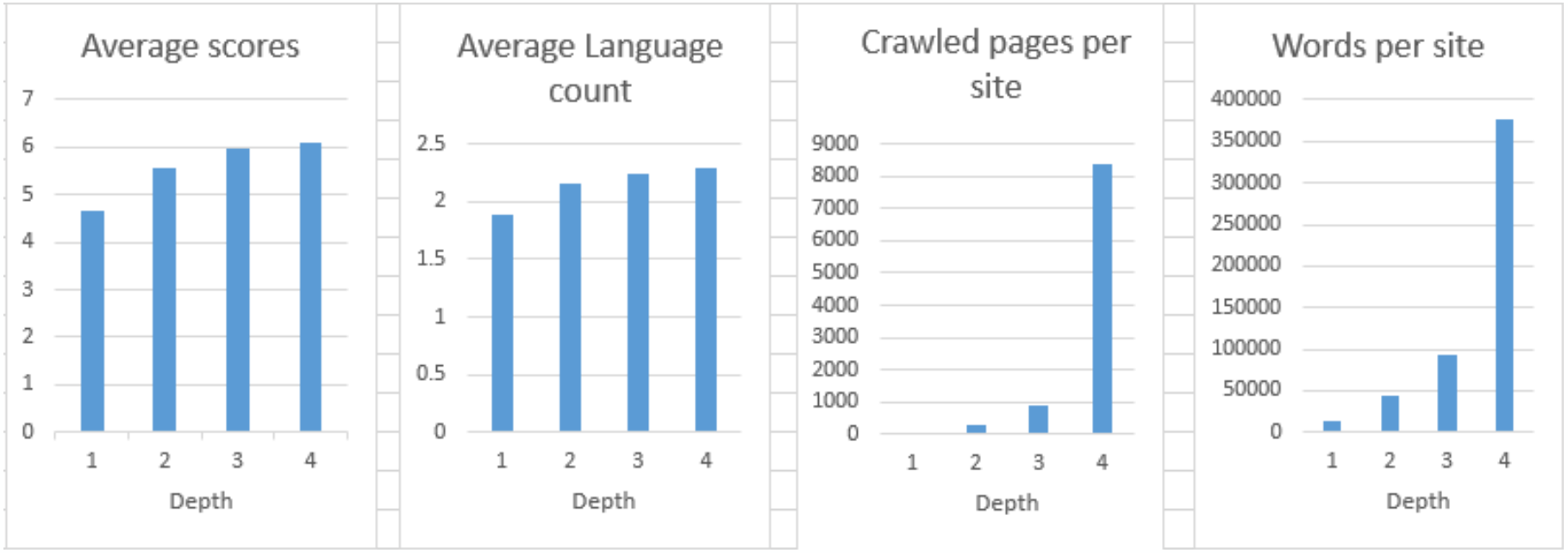
- 1 Hops: Pages: 39; w/o 6, ru 1, en 31, et 1
- 2 Hops: Pages: 1133; w/o 293, ru 28, en 733, et 43, uk 2, fi 3, de 6, sv 2, lv 2, ...
- 3 Hops: Pages: 6162; w/o 2445, et 1625, ru 456, en 1537, fi 14, de 15, sv 4, lv 5, ...
- 4 Hops: Pages: 12471; w/o 6149, et 3421, en 1881, ru 879, fi 28, de 21, sv 5, lv 6, ...



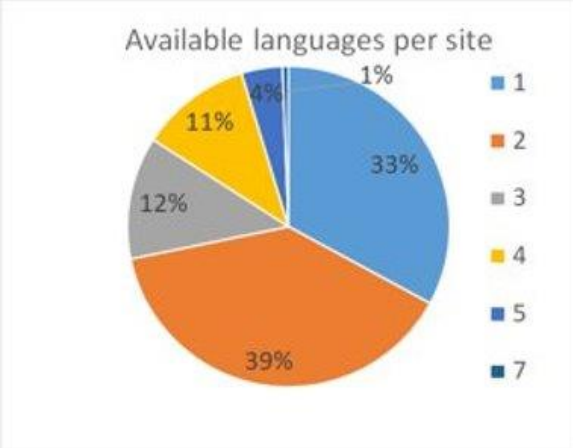
SOME RESULTS AND OBSERVATIONS

URLs	Crawling depth	Crawl time (h)	Memory usage max (MB)	Average score (normalised lang. balance)	Average coverage of EU languages	Average number of pages/site	Average number of words/site
198	1	0:50	160	4.65	1.89	33	14516
198	2	12:54	1030	5.57	2.15	262	44592
198	3	48	1238	5.97	2.25	885	93242
198	4	>250	6672	6.08	2.29	8374	376787
600	2	52	1023	5.56	2.02	227	53663

RESULTS ON TWO TEST LISTS



Depth of crawling 3



CONCLUSION

- To measure multilinguality, we created an open-source tool for scoring multilinguality that calculates several scores to measure multilingualism over the Web: Lieberson's diversity index, language coverage, language balance and normalised language balance.
- European websites currently are not very multilingual – on average content is presented only in 2-3 languages
- Our next steps include assessment of more complicated multilingualism criteria, such as linguistic quality, technical quality and content parallelism, and implement them into next versions of the tool.

THANK YOU FOR YOUR ATTENTION!

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.

European Language Resource Coordination
Connecting Europe Facility

NATIONAL DEVELOPMENT PLAN 2020

EUROPEAN UNION
European Regional Development Fund

INVESTING IN YOUR FUTURE