IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received May 12, 2022, accepted May 25, 2022, date of publication May 30, 2022, date of current version June 6, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3179375

Comparison of Online Accessibility Evaluation Tools: An Analysis of Tool Effectiveness

RITA ISMAILOVA[1] **AND YAVUZ INAL**[2]
[1]Computer Engineering Department, Kyrgyz-Turkish Manas University, Bishkek 720038, Kyrgyzstan
[2]Department of Design, Norwegian University of Science and Technology, 2821 Gjøvik, Norway

Corresponding author: Yavuz Inal (yavuz.inal@ntnu.no)

ABSTRACT The use of online tools is a common practice for evaluating the accessibility of a website, identifying problems, and providing useful feedback on how to fix detected issues. For ease of accessibility validation, many tools have been developed and implemented successfully over the years. Yet, the results of these tools show differences in terms of coverage, correctness, and reporting-related issues. In this study, we compared online accessibility evaluation tools to understand to what extent they differ in detecting accessibility problems in websites. A total of 41 government websites of different countries were tested for violations of accessibility guidelines using six evaluation tools. We observed that each tool generated different evaluation data for the same websites. As some of the tools are complementary to each other, meaning the highest coverage and completeness can be possible with the right combination of evaluation tools. Therefore, we suggest different tools should be utilized to provide consistency and obtain reliable data from online evaluation tools, thereby improving tool effectiveness.

INDEX TERMS Automated evaluation, government websites, online evaluation tools, WCAG, web accessibility, web performance.

I. INTRODUCTION

The web constitutes an integral part of digital society and offers an enormous level of online information and services for everyone. Web accessibility, however, remains the main concern on many websites across the world. It is necessary to meet an acceptable level of compliance so that everyone regardless of their disabilities can access online services without undue effort [1]. Accessibility evaluation is a widespread practice to check the status of accessibility of a website and develop it to comply with accessibility guidelines [2]. The fastest method to verify compliance is by investigating the code. Evaluation tools, designed specifically for crawling an HTML page, are important for preliminary evaluation [3]. They show problems indicating the lines of HTML code and provide feedback on violated checkpoints.

With tool-based evaluation, it is easier to test a large number of websites in a short amount of time and have an overall insight into the state of accessibility compliance [4]. Over the years, numerous studies have successfully used various evaluation tools to identify accessibility issues in websites [2], [5]–[9].

However, the results of studies derived from different tools show that no tool provides exactly the same outcomes in terms of accessibility violations. Online tools create somewhat different evaluation results at the end of the evaluation process as they have different approaches and evaluation scopes. In line with this argument, previous research pointed out that evaluating the accessibility of a website with only one tool is inadequate to decipher accessibility-related problems (e.g., [10]–[13]). Therefore, most researchers (e.g., [6], [12], [14]) evaluated the accessibility of websites using more than one tool to increase the validity and reliability of their results, as a single tool is not able to identify all possible accessibility problems [10], [15].

For example, [12] evaluated the accessibility of government websites using AChecker and TAW tools. The results showed large differences in the number of errors derived from each tool. TAW identified more accessibility violations when compared to AChecker in most of the evaluated websites. While AChecker reported very few numbers of accessibility violations for a website, TAW identified more than one hundred accessibility problems for the same website.

The associate editor coordinating the review of this manuscript and approving it for publication was Orazio Gambino.

Another study by [8] tested the accessibility of university websites using three evaluation tools, namely TAW, WAVE, and EIII Page Checker. Each tool detected different types of accessibility issues such that WAVE detected contrast errors as the most violated issues in the websites while TAW identified problems related to the compatibility of websites with assistive technologies, different browsers, and other user agents.

Over the years, little research has been conducted to address the importance of tool differences in accessibility evaluation. For example, [5] conducted a case study comparing LIFT Machine with Bobby to explore tool effectiveness in the evaluation of web accessibility. There were differences between the tools in the coverage of accessibility errors and in the reporting of detected issues. [11] compared five accessibility evaluation tools, namely AChecker, TAW, Total Validator, Cynthia Says, and WAVE, in terms of accessibility errors identified and reporting-related issues. The authors observed a significant gap between the results as each tool reported a different number of accessibility errors. The number of violations differed in different success criteria such that TAW and CynthiaSays tools identified the highest number of errors.

In another study, [13] presented two novel frameworks to compare the performance of accessibility evaluation tools in identifying accessibility errors. Two popular tools, WAVE and SiteImprove, were utilized to compare the performance of six evaluated websites. WAVE was not able to identify the components in terms of keyboard navigation while some websites failed to meet such criteria according to SiteImprove. [10] conducted a benchmarking study to explore the effectiveness of six accessibility evaluation tools, namely AChecker, SortSite, Total Validator, TAW, Deque, and AMP in terms of their coverage, correctness, and completeness. The authors reported that online evaluation tools alone were not able to cover most of the accessibility issues.

With this diversity of tool-based evaluation, the difference in results is inevitable. In the current study, we compared online accessibility evaluation tools to understand to what extent they differ in detecting accessibility issues in websites across WCAG conformance levels. To this end, we tested the compliance of 41 government websites of different countries with WCAG 2.0 using six tools. Unlike the previous studies, we created a larger dataset to determine differences in the evaluation tools.

II. METHODS

A. STUDY DESIGN

Data was collected by testing 41 government websites using six tools, namely AChecker, VaMoLà, AccessMonitor, Examinator, Mauve, Cynthia Says - for a total of 246 tests. However, as errors could be of temporary nature, the test on websites that returned errors was carried out two additional times. For some websites, tests were run more than once, with an interval of 1-2 days between the first and second trials, and with the interval of two weeks between the second and third tests.

Although W3C proposed a new version of WCAG - WCAG 2.1 in 2018, most non-commercial tools test websites for compliance with WCAG 2.0. In addition, the WCAG 2.1 was proposed as an extension to WCAG 2.0, with additional checkpoints on content accessibility on tablets and mobile devices [16]. Therefore, we confined our scope to WCAG 2.0 compliance level to have consistent data and provide inter-reliability in the evaluation results.

B. SELECTION OF TOOLS

Currently, on the World Wide Web Consortium (W3C) page, there are more than 150 web accessibility evaluation tools that are in compliance with WCAG [17]. Of these, 55 are online tools. Although some tools are designed to evaluate a website, in this study, online tools generating reports of evaluation results of single pages were selected. The filtering resulted in 34 tools and from these, non-commercial ones were chosen. Another criterion for selection was the representation of results; thus, filtering was applied to choose tools that represent results on the same page. Finally, the following tools were selected:

AChecker [18]: The tool is based on GDL and uses a probabilistic approach. It evaluates HTML content and reports accessibility issues of three types - known problems, likely problems, and potential problems [19]. The tool can automatically check a single web page only. AChecker also has a module to review conformance with the most recent version of the web content accessibility guidelines, WCAG 2.1.

VaMoLà - Validator: The tool was launched in 2009 in cooperation with AChecker development team [20]. It combines a monitor and a validator of web accessibility. The validator part of the tool evaluates websites for compliance with Italian law requirements, while the monitoring portion ''periodically controls and records the accessibility level of a predefined set of URLs'' [21]. Similar to AChecker, it checks a single page for violations of accessibility guidelines. An online version of the tool was available at the time the data was derived, however, currently only a desktop version of the tool is available.

AccessMonitor [22]: The AccessMonitor is a service by the Digital Experience Team of the Digital Transformation Department, the division of the Portugal Agency for Administrative Modernization. The tool records all scrawled accessibility checkpoints and gives an account of three types - acceptable, to review manually, and not acceptable. It can automatically check a single web page, group of web pages, or websites.

Examinator [23]: The tool is one of the first tools introduced in 2005, which reports accessibility issues based on an overall score from 1 to 10. However, starting from 2015, the tool only allows review of a limited number of pages per session [24] and can automatically check a single web page only.

Mauve [25]: It is a deterministic tool and was first presented by [26]. The tool was developed utilizing XML-based language. It recursively calls for so-called ''checks'' and

''conditions'', and the results of validations at levels are presented in a merged format as an XML file. Recently, a new version of the tool, Mauve++, was proposed by Human Interface in Information System [27]. It can automatically check a single web page, group of web pages, or websites.

Cynthia Says [28]: The tool was developed by the International Centre for Disability Resources on the Internet and Compliance Sheriff. It utilizes a deterministic approach and accessibility reports are organized at five levels: Failed (if a page did not pass the checkpoint and must be fixed), warning (if a page passed the checkpoint but could be improved), passed (if a page passes the checkpoint), visual (if a visual check is required to determine whether this page passed a checkpoint), and N/A in cases when the checkpoint is not relevant for this page. The tool also checks for accessibility violations on a single page. Cynthia Says announced the end of the organization at the end of 2021 [29]. However, test results collected by Cynthia Says were included in this study because we believe that the results of the analysis will help to improve existing accessibility evaluation tools.

C. SELECTED WEBSITES

The sample of the study consists of 41 government websites (Table 1). Mostly, websites have .gov extensions; however, in five of them, URLs are defined by country domain names. In the choice of websites, the corresponding countries' human development index (HDI) was taken into account as it was shown that the difference in WCAG 2.0 violations in websites of countries with different HDI levels was significant [9]. Therefore, in the study, the websites of countries with different HDI were evaluated.

TABLE 1. Websites evaluated in the study.

Country	Government website	Country	Government website
Afghanistan	ceo.gov.af	Malaysia	malaysia.gov.my
Armenia	gov.am	Moldova	gov.md
Azerbaijan	gov.az	Nepal	nepal.gov.np
Bangladesh	bangladesh.gov.bd	Netherlands	rijksoverheid.nl
Burundi	burundi.gov.bi	Norway	regjeringen.no
Cambodia	opmcm.gov.np	Pakistan	pakistan.gov.pk
China	gov.cn	Portugal	portugal.gov.pt
Denmark	stm.dk	Romania	gov.ro
Ethiopia	ethiopia.gov.et	Russia	government.ru
France	gouvernement.fr	Saudi Arabia	saudi.gov.sa
Georgia	gov.ge	South Korea	minwon.go.kr
Germany	bundesregierung.de	Spain	administracion.gob.es
India	india.gov.in	Tanzania	tanzania.go.tz
Indonesia	indonesia.go.id	Thailand	thaigov.go.th
Ireland	gov.ie/ga	Timor-Leste	timor-leste.gov.tl
Israel	gov.il	Turkey	turkiye.gov.tr
Italy	governo.it	Turkmenistan	turkmenistan.gov.tm
Japan	e-gov.go.jp	UK	gov.uk
Kazakhstan	government.kz	Uzbekistan	gov.uz
Kyrgyzstan	gov.kg	Vietnam	chinhphu.vn
Lao	laogov.gov.la		

III. RESULTS

A. VIOLATIONS AT THE CONFORMANCE LEVELS

The number of errors by each tool was counted to verify if different tools detect a different number of violations while testing the same website. First, the number of websites that passed all WCAG 2.0 success criteria were counted for each of the six tools. Results of an overview organized by conformance levels are presented in Table 2. It is important to note that the percentages were calculated based on the number of successful tests when a tool returned a compliance report.

WCAG 2.0 conformance level A consists of guidelines that are essential for websites (must support). Tests, carried out on the selected websites showed that two tools, namely AccessMonitor and Cynthia Says, reported that no website complied with compliance level A guidelines. Mauve found errors in 97.4% of websites (i.e., in 37 out of 38 websites that returned no-error results), followed by AChecker with 80% of websites not complying with level A guidelines. VaMoLà and Examinator tools showed almost similar percentages of websites with violations - 74.4% and 72.7%, respectively.

AccessMonitor reported that no website, selected for the current study, complied with conformance level AA (should support), followed by Cynthia Says, which found errors in 97.6% of websites, and Mauve, which found violations in 81.6% of websites (i.e., in 31 out of 38 successfully tested websites). As mentioned above, the number of successful tests by Examinator was 22, and in 72.7% of these websites, there were violations of conformance level AA checkpoints. VaMoLà and AChecker reported 48.7% and 45% of websites to have conformance level AA errors, respectively.

Almost the same pattern was observed in the number of websites not complying with conformance level AAA (may support), with the difference in the results by Mauve. Again, AccessMonitor and CynthiaSays reported that no website, selected for the study, complied with conformance level AAA. Examinator found errors in 72.7% of tested websites, followed by Mauve (68.4%), VaMoLà (25.6%) and AChecker with 22.5% of websites reported having conformance level AAA errors.

As seen from Table 2, AccessMonitor, Cynthia Says and Achecker showed better results than other tools in terms of successful tests. As for violation detection, results presented by AccessMonitor and Cynthia Says detected errors in all tested websites at all three levels, while Achecker showed lower performance in detecting the violations of level AAA guidelines.

B. VIOLATIONS AT THE GUIDELINES

The numbers of errors were calculated for each accessibility guideline (Table 3). The highest number of violations were

TABLE 2. Violations at the conformance levels by tools.

	A	AA	AAA	Total Tests
AChecker	80% (32)	45% (18)	22.5% (9)	40
VaMoLà	74.4% (29)	48.7% (19)	25.6% (10)	39
Access Monitor	100% (41)	100% (41)	100% (41)	41
Examinator	72.7% (16)	72.7% (16)	72.7% (16)	22
Mauve	97.4% (37)	81.6% (31)	68.4% (26)	38
Cynthia Says	100% (41)	97.6% (40)	100% (41)	41

detected by Mauve (n = 41914), while the lowest number of violations were detected by Examinator (n = 487). More than half of the violations detected by AChecker (80%), VaMoLà (74%), Examinator (62.7%), and AccessMonitor (60%) were at conformance level A. On the other hand, 61% of violations detected by Cynthia Says and 40% of violations detected by Mauve were at level AAA. It is worth mentioning that in the results of deterministic tools (e.g., AChecker and VaMoLà), only confirmed errors were considered.

The tools mostly detected violations of the same checkpoints, however, the numbers of violations differed. That is, violations were detected in 21 out of 30 checkpoints (70%) of the conformance level A, 15 out of 20 (75%) of level AA, and in 14 out of 28 (50%) of level AAA. The following comparisons were made within violations detected by six tools on a set of given sample websites only. For instance, the distribution of violated checkpoints by tools showed that Mauve detected violations in 18 checkpoints at level A and 12 at level AA. Examinator detected 17 checkpoints at level A and 9 at level AA.

The difference in the number of violations was high. For example, at level A, Mauve identified 14927 violations, while Examinator detected 291 violations. Most of the violations detected by Mauve were that of checkpoints 1.1.1 (non-text content), 2.4.4 (link purpose in context), 3.1.1 (language of page), and 4.1.2 (name, role, value). Although the difference in the violation of checkpoint 1.1.1 was high, two other tools, Achecker and Cynthia Says, detected 1497 and 1878 violations, respectively. However, in the three remaining checkpoints, results by Mauve were much higher. The same pattern was observed in conformance levels AA and AAA.

AChecker and VaMoLà failed to detect the absence of alternatives for time-based media (guideline 1.2) or time arrangements in auto-updating and flash objects (guideline 2.2 on providing users enough time to read and use content and guideline 2.3 on physical reactions). Violations of guidelines 1.3 (on making content adaptable) and 1.4 (on making content distinguishable) at conformance level AA were

TABLE 3. Violations at the checkpoints by tools.

	AChecker	VaMoLà	Access Monitor	Examinator	Mauve	CynthiaSays
1.1.1	1497	858	121	51	2367	1878
1.2.1	-	-	1	-	12	4
1.2.2	-	-	-	-	4	4
1.3.1	67	36	226	88	479	260
1.4.1	1	-	-	-	150	-
2.1.1	58	58	4	3	264	-
2.1.2	-	-	-	-	44	-
2.2.1	-	1	2	2	37	-
2.2.2	3	4	2	8	-	-
2.3.1	-	-	-	3	-	-
2.4.1	-	-	86	28	8	45
2.4.2	1	1	41	16	2	-
2.4.3	-	-	-	4	-	-
2.4.4	26	259	51	13	4940	744
3.1.1	17	10	45	17	1113	2
3.2.1	-	2	1	1	4	-
3.2.2	3	-	29	10	8	27
3.3.1	-	2	-	2	7	-
3.3.2	34	23	34	14	51	47
4.1.1	10	9	16	5	61	-
4.1.2	-	-	76	26	5376	99
Total A	**1717**	**1263**	**735**	**291**	**14927**	**3110**
1.2.4	1	-	-	2	-	-
1.2.5	-	-	-	4	2	-
1.3.4	-	-	-	-	12	-
1.3.5	-	-	-	-	162	-
1.4.3	-	-	23	13	1022	3916
1.4.4	220	246	62	25	4162	48
1.4.5	-	-	61	19	516	-
1.4.10	-	-	-	-	1409	-
1.4.11	-	-	-	-	1224	-
1.4.12	-	-	-	-	117	-
2.4.5	-	-	16	5	-	-
2.4.6	24	21	1	-	3	5
2.4.7	-	-	1	1	46	-
3.2.4	1	-	-	-	-	-
4.1.3	-	-	-	-	1473	-
Total AA	**246**	**267**	**164**	**69**	**10148**	**3969**
1.2.8	-	-	2	2	-	5
1.2.9	-	-	-	-	96	-
1.4.6	177	176	20	9	116	5311
1.4.7	-	-	-	-	3676	-
1.4.8	-	-	53	26	12950	2672
1.4.9	-	-	62	24	-	-
2.1.3	-	-	3	2	-	-
2.2.4	-	-	2	2	-	-
2.4.10	-	-	69	26	-	59
2.4.8	-	-	11	4	-	4
2.4.9	-	-	51	31	-	1025
3.1.4	-	-	-	-	-	2
3.2.5	-	-	1	1	1	1984
3.3.5	-	-	-	-	-	25
Total AAA	**177**	**176**	**274**	**127**	**16839**	**11087**

missing. On conformance level AAA, violations related to operability and understandability principles were not detected. An almost similar pattern was observed in the results by AccessMonitor and Examinator. Cynthia Says mostly missed violations on the operability principle (guidelines 2.1, 2.2, and 2.3). Mauve detected more errors than other tools, both as regards the number of violated guidelines (except level AAA) and the number of errors at each guideline.

C. VIOLATIONS AT THE PRINCIPLES

The difference in violation detection by different tools in individual websites was examined, and for this exercise, ten websites with the highest and lowest occurrences were selected. The difference in violation of checkpoint 1.1.1 (non-text content) was the most diverse; the difference was mostly observed on the Azerbaijan government website. AChecker and VaMoLà reported 275 violations followed by Mauve (n = 563), AccessMonitor (n = 3) and Cynthia Says (n = 4). Examinator returned an error message. The difference in the tool results was 560 detections. The lowest number of differences were found on the website of the Norwegian government: VaMoLà and Cynthia Says reported no violations. AChecker reported 1, AccessMonitor 2, and Mauve 4 violations. Examinator again returned an error message. The difference in the tool results was 3 detections.

Another pattern detected was that Mauve performed better in identifying the violations of checkpoints 2.4.4 (link purpose), 2.4.5, (multiple ways), 4.1.2 (name, role, value), 1.4.3 (Minimum contrast), and 1.4.8 (visual presentation) than other tools. The maximum difference in detected violations was up to 1279 occurrences on the Chinese government website. Cynthia Says found the highest number of violations of checkpoints 1.3.5 (identify input purpose), 1.4.6 (contrast - enhanced), and 1.4.12 (text spacing) compared to other tools. All these checkpoints are related to the perceivable principle of accessibility. Thus, as a next step, the differences were visualized by WCAG principles based on the data of ten websites with high occurrences.

Results showed that for the perceivable principle, differences were mostly observed in the number of detections by Mauve and Cynthia Says (Figure 1). This means that these tools can better recognize the absence of text alternatives and media equivalents for time-dependent presentations. In 9 cases out of 10, Mauve detected more violations than other tools. For example, the difference in detection of these kinds of violations was up to 2189 detections in the case of the Azerbaijani government website.

The same pattern was observed in the detections of violations in the operable principle. Mauve detected more violations than other tools. Although there was a difference (n = 582) in the detection of violations in the case of the Chinese government website (i.e., Mauve detected 1011 violations and Cynthia Says detected 429 violations) - on average, results by Cynthia Says were close to those by other tools.

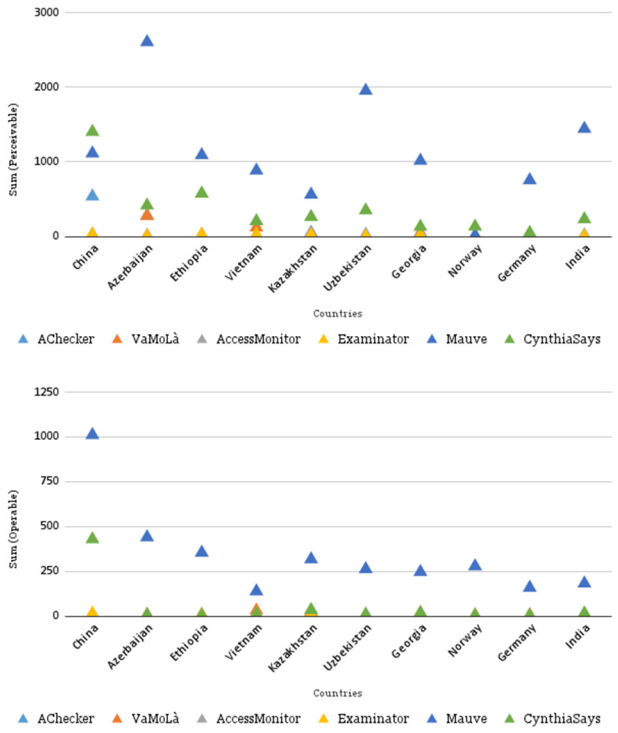

FIGURE 1. Highest differences in perceivable (upper) and operable (lower) principles violation detections by tools.

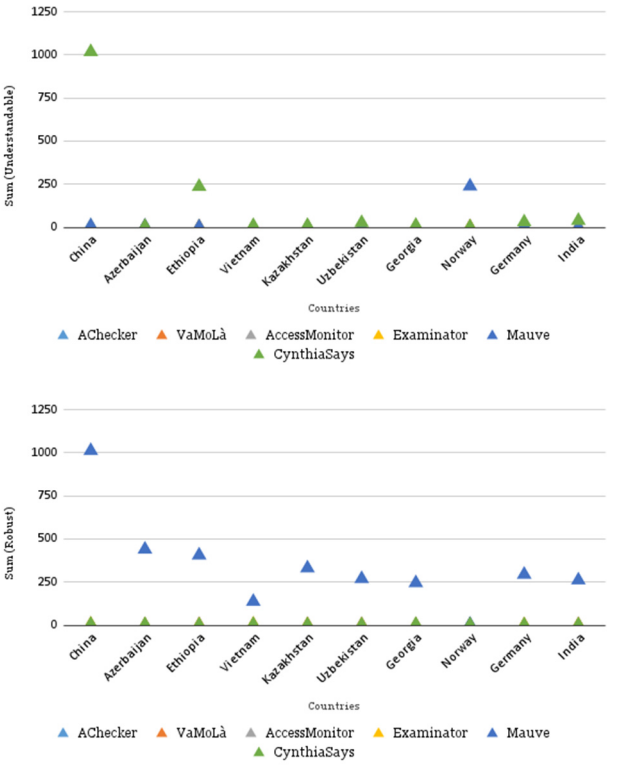

FIGURE 2. Highest differences in understandable (upper) and robust (lower) principles violation detections by tools.

However, with regards to the violations of the understandable principle, Cynthia Says demonstrated better

performance by detecting more errors than other tools in 9 websites out of 10 (Figure 2). For example, in the case of the Chinese government website, this tool reported 1018 violations, while AChecker detected 8 violations followed by Mauve (n = 7), AccessMonitor (n = 3), and Examinator (n = 3).

Analysis of the violations detected in the robust principle showed that Mauve found more errors than other tools. Among 10 websites, where there was the highest difference in the number of detected violations, the difference in the analysis of the Chinese government website was the highest - Mauve detected 1012 violations followed by AccessMonitor (n = 4), Examinator (n = 4), and AChecker (n = 1). VaMoLà reported no violations (Figure 2). The results of the current analysis showed that when considering the number of violation detection by accessibility principles, mostly the dissimilarity was observed between the results derived from Mauve and those of other tools.

IV. DISCUSSION AND CONCLUSION

This study aimed to understand to what extent online accessibility evaluation tools differ in detecting accessibility problems in websites. Data from 41 government websites showed that there were differences in the evaluation results. According to AccessMonitor, no websites complied with all conformance levels of WCAG 2.0 followed by Cynthia Says, which detected that no websites complied with conformance levels A and AAA, and all websites except one failed to meet level AA conformance.

The smallest difference was between the outputs of AChecker and VaMoLà which was not unexpected, as VaMoLà was developed based on AChecker. The number of violations found by Mauve was significantly higher than that of other tools; this was observed at all conformance levels. In addition, the analysis carried out on the differences in different accessibility principles showed that Mauve detected more violations. This result may be due to the fact that it is a deterministic tool and runs recursively. Mauve performed better in checkpoints 2.2.2 (pause, stop, hide), 2.3.1 (three flashes or below threshold), and 2.4.3 (focus order) than other tools. Examinator found more issues with violations in checkpoints 1.4.9 (images of text, no exception), 2.4.10 (section headings), and VaMoLà in checkpoint 3.2.2 (on input).

The picture was completely different when examining the number of websites on which violations were detected rather than on the number of detected violations. From this perspective, Access Monitor showed higher performance. The tool detected some violations in websites where other tools failed. For example, all tools except for Access Monitor failed to detect non-compliance with level A checkpoints 1.3.1 (info and relationships), 2.4.1 (bypass blocks), 2.4.2 (page titles), and 3.1.1 (the language of pages). In addition, at level AA, the violations of checkpoint 1.4.5 (images of text) were detected only by this tool. Consequently, it prompted the low - or absence of - correlation in the number of violations by different tools, both when compared to conformance levels and accessibility principles.

We observed that online evaluation tools had different coverage thus they produced different outcomes in terms of accessibility issues. Some tools were good at detecting one type of error, while others were good at detecting different types of errors. Therefore, it may be worth looking at the violations at which tools showed similar performance. In the following WCAG checkpoints, none of the tools detected violations: Checkpoints 1.2.3 (Audio Description or Media Alternative) and 1.4.2 (Audio Control Level) at level A and 3.1.2 (Language of Parts), 3.2.3 (Consistent Navigation Level), 3.3.3 (Error Suggestion), and 3.3.4 (Error Prevention) at level AA. These checkpoints were not violated in the evaluated websites due to the fact that they did not contain a pre-recorded video; thus, no content covered by checkpoints 1.2.3, 1.2.6, and 1.2.7 was presented on the websites. The same can be said about outdated technologies such as flash (e.g., checkpoint 2.3.2) or user-controllable data insertion fields (covered by checkpoints 2.5.2, 2.5.3, 2.5.3, and 2.5.4). However, checkpoints on making text content readable and understandable (3.1.2, 3.1.3, 3.1.4, 3.1.5) require more advanced technology as they are related to the content language and text understanding. Detection of these violations can be programmatically challenging and, thus, developers should pay attention to these checkpoints during the web development process.

Since there has been an increasing demand to access online information and services in recent years [30], the accessible web has gained outstanding importance to the public for effective digitalization. Poor accessibility of websites leads to the exclusion of some groups of people including people with disabilities from digital society. To remedy this, web accessibility practices need to be incorporated into the web development process effectively so that developers can ensure that content is accessible for everyone regardless of their disabilities [31]–[34].

Results of our study corroborate the findings of [10], [35], who found that online evaluation tools cannot detect all accessibility errors of a website and there is always a need for manual testing. However, any evaluated website can perform a large number of repetitive tests, and thus, tool-based evaluation helps developers evaluate the accessibility of their websites, identify problems and obtain useful feedback on how to solve them.

Using one tool cannot always help developers to find all violations of WCAG. We observed that some of the tools are complementary to each other, meaning the highest coverage and completeness can be possible with the right combination of online evaluation tools. Using different tools can help maximize the coverage of accessibility success criteria. We, therefore, suggest that different tools should be utilized to provide consistency and obtain reliable data from online evaluation tools, thereby improving tool effectiveness. We hope the present study helps developers choose the best combination of evaluation tools to address accessibility guidelines.

Results presented in the study draw the accessibility issues on home pages of selected websites only, which can be considered as a limitation of the study. In addition, in the present study, the focus was on the number of violations detected by each tool, and no insight into each detection was analyzed. As a further study, the cases of detection are planned to be analyzed in detail to understand the reasons for some tools to detect them.

REFERENCES

[1] R. Ismailova and Y. Inal, "Web site accessibility and quality in use: A comparative study of government web sites in Kyrgyzstan, Azerbaijan, Kazakhstan and Turkey," *Universal Access Inf. Soc.*, vol. 16, no. 4, pp. 987–996, Nov. 2017.

[2] A. P. Freire, T. J. Bittar, and R. P. M. Fortes, "An approach based on metrics for monitoring web accessibility in Brazilian municipalities web sites," in *Proc. ACM Symp. Appl. Comput.*, 2008, pp. 2421–2425.

[3] C. Benavídez, J. L. Fuertes, E. Gutiérrez, and L. Martínez, "Semi-automatic evaluation of web accessibility with HERA 2.0," in *Proc. Int. Conf. Comput. Handicapped Persons*. Berlin, Germany: Springer, 2006, pp. 199–206.

[4] N. E. Youngblood and S. A. Youngblood, "User experience and accessibility: An analysis of county web portals," *J. Usability Stud.*, vol. 9, no. 1, pp. 25–41, 2013.

[5] G. Brajnik, "Comparing accessibility evaluation tools: A method for tool effectiveness," *Universal Access Inf. Soc.*, vol. 3, nos. 3–4, pp. 252–263, Oct. 2004.

[6] O. Gambino, R. Pirrone, and F. D. Giorgio, "Accessibility of the Italian institutional web pages: A survey on the compliance of the Italian public administration web pages to the Stanca Act and its 22 technical requirements for web accessibility," *Universal Access Inf. Soc.*, vol. 15, no. 2, pp. 305–312, Jun. 2016.

[7] R. Ismailova and Y. Inal, "Accessibility evaluation of top university websites: A comparative study of Kyrgyzstan, Azerbaijan, Kazakhstan and Turkey," *Universal Access Inf. Soc.*, vol. 17, no. 2, pp. 437–445, Jun.2018.

[8] S. Alim, "Web accessibility of the top research-intensive universities in the U.K.," *SAGE Open*, vol. 11, no. 4, 2021, Art. no. 21582440211056614.

[9] Y. Inal and R. Ismailova, "Effect of human development level of countries on the web accessibility and quality in use of their municipality websites," *J. Ambient Intell. Hum. Comput.*, vol. 11, no. 4, pp. 1657–1667, Apr. 2020.

[10] M. Vigo, J. Brown, and V. Conway, "Benchmarking web accessibility evaluation tools: Measuring the harm of sole reliance on automated tests," in *Proc. 10th Int. Cross-Disciplinary Conf. Web Accessibility*, 2013, pp. 1–10.

[11] M. Pădure and C. Pribeanu, "Exploring the differences between five accessibility evaluation tools," in *Proc. Int. Conf. Hum.-Comput. Interact. (RoCHI)*, 2019, pp. 87–90.

[12] N. A. Karaim and Y. Inal, "Usability and accessibility evaluation of Libyan government websites," *Universal Access Inf. Soc.*, vol. 18, no. 1, pp. 207–216, Mar. 2019.

[13] A. Alsaeedi, "Comparing web accessibility evaluation tools and evaluating the accessibility of webpages: Proposed frameworks," *Information*, vol. 11, no. 1, p. 40, Jan. 2020.

[14] N. Sabev, G. N. Georgieva-Tsaneva, and G. Bogdanova, "Research, analysis, and evaluation of web accessibility for a selected group of public websites in Bulgaria," *J. Accessibility Des. All*, vol. 10, no. 1, pp. 124–160, 2020.

[15] B. Giovanna, M. Manca, F. Paternò, and F. Pulina, "Flexible automatic support for web accessibility validation," in *Proc. ACM Hum.-Comput. Interact.*, vol. 4, Jun. 2020, Art. no. 83.

[16] World Wide Web Consortium (W3C). (2018). *Web Content Accessibility Guidelines (WCAG) 2.1*. [Online]. Available: https://www.w3.org/TR/WCAG21/#comparison-with-wcag-2-0

[17] World Wide Web Consortium (W3C). (2016). *Web Accessibility Evaluation Tools List*. [Online]. Available: https://www.w3.org/WAI/ER/tools/?q=wcag-20-w3c-web-content-accessibility-guidelines-20

[18] *AChecker*. [Online]. Available: https://achecker.achecks.ca/checker/index.php

[19] G. Gay and C. Q. Li, "AChecker: Open, interactive, customizable, web accessibility checking," in *Proc. Int. Cross Disciplinary Conf. Web Accessibility (W4A)*, 2010, pp. 1–2.

[20] S. Mirri, L. A. Muratori, M. Roccetti, and P. Salomoni, "Metrics for accessibility on the VaMoLà project," in *Proc. Int. Cross-Disciplinary Conf. Web Accessibililty (W4A)*, 2009, pp. 142–145.

[21] M. Battistelli, S. Mirri, L. A. Muratori, P. Salomoni, and S. Spagnoli, "Making the tree fall sound: Reporting web accessibility with the VaMoLà monitor," in *Proc. 5th Int. Conf. Methodol., Technol. Tools Enabling e-Government*, 2011, pp. 1–12.

[22] *AccessMonitor*. [Online]. Available: https://accessmonitor.acessibilidade.gov.pt

[23] *Examinator*. [Online]. Available: http://examinator.net

[24] C. Benavidez. (2015). *Evaluación de la Accesibilidad Web. Examinator*. Accessed: Oct. 5, 2021. [Online]. Available: http://examinator.net/

[25] *Mauve*. [Online]. Available: http://mauve.isti.cnr.it

[26] A. G. Schiavone and F. Paternò, "An extensible environment for guideline-based accessibility evaluation of dynamic web applications," *Universal Access Inf. Soc.*, vol. 14, no. 1, pp. 111–132, Mar. 2015.

[27] Human Interface in Information System. (2020). *Multiguideline Accessibility and Usability Validation Environment (MAUVE + +)*. [Online]. Available: https://mauve.isti.cnr.it/

[28] *Cynthia Says*. [Online]. Available: http://www.cynthiasays.com

[29] B. Henry. (2021). *CynthiaSays.com Accessibility Website Scan Announcement*. [Online]. Available: https://www.tpgi.com/cynthiasays-com-accessibility-website-scan-announcement/

[30] Y. Inal, F. Guribye, D. Rajanen, M. Rajanen, and M. Rost, "Perspectives and practices of digital accessibility: A survey of user experience professionals in Nordic countries," in *Proc. 11th Nordic Conf. Hum.-Comput. Interaction, Shaping Experiences, Shaping Soc.*, Oct. 2020, pp. 1–11.

[31] Y. Inal and R. Ismailova, "How do computer engineering students construe usability and accessibility? A comparative study between Turkey and Kyrgyzstan," *Tehnički vjesnik*, vol. 25, no. 5, pp. 1339–1347, 2018, doi: 10.17559/TV-20170205083820.

[32] Y. Inal, K. Rızvanoğlu, and Y. Yesilada, "Web accessibility in Turkey: Awareness, understanding and practices of user experience professionals," *Universal Access Inf. Soc.*, vol. 18, no. 2, pp. 387–398, Jun. 2019.

[33] J. Lazar, A. Dudley-Sponaugle, and K.-D. Greenidge, "Improving web accessibility: A study of webmaster perceptions," *Comput. Hum. Behav.*, vol. 20, no. 2, pp. 269–288, Mar. 2004.

[34] G. Farrelly, "Practitioner barriers to diffusion and implementation of web accessibility," *Technol. Disability*, vol. 23, no. 4, pp. 223–232, Dec. 2011.

[35] A. Leitner, I. Ciupa, B. Meyer, and M. Howard, "Reconciling manual and automated testing: The AutoTest experience," in *Proc. 40th Annu. Hawaii Int. Conf. Syst. Sci. (HICSS)*, Jan. 2007, p. 261.