Comparative Analysis of Plagiarism Detection Systems

Yuliia Shkodkina

PhD, Assistant Professor, Sumy State University, Ukraine

Darius Pakauskas

PhD, Postdoctoral Researcher, Aalto University School of Arts, Design and Architecture, Helsinki, Finland

Abstract

This article compares plagiarism detection systems according to the list of criteria, compiled from the most challenging and important features for users in Ukrainian higher educational institutions. In addition it describes types of text-based plagiarism and provides an overview of the most common and serious forms of plagiarism around the world and in Ukraine in particular.

The authors carried out a comparative analysis of three plagiarism detection systems − Turnitin, Unicheck, eTXT − and highlighted advantages and disadvantages of each one for the use in Ukraine. However, in further studies there might be a need for revising and expanding the list of criteria depending on the subject and the aim of using plagiarism checkers.

Keywords: plagiarism, plagiarism detection system, text-based plagiarism, plagiarism forms.

JEL Classification: I23, I29.

Introduction

The authors are grateful to Unicheck for the materials provided

Plagiarism as the most important and challenging component of quality assurance system in higher education is a key issue to tackle for any reputable university. However, the situation in the universities of some countries, e.g. Ukraine, are rather repressing. Literature shows that high number of Ukrainian students are still engaging in plagiarism, which prevents from new knowledge generation, and thus suppress benefits that the higher education system is capable at providing. University's recognizing this issue and developing anti-plagiarism policy which is supposed to provide the framework for prevention and regulation of plagiarism cases detected. While there are numbers of methods to deal with plagiarisms within higher education, advancements in information and communication technologies, allow us to use plagiarism detection systems, which in simple terms are software solutions to identify plagiarized information, its extent and source.

There are different plagiarism detection systems on the market, and the number is continuously growing. Thus, the comparative analysis of plagiarism detection systems is required, as a means for higher education institutions to choose the system that suits there needs. The aim of this article to propose criteria for plagiarism detection systems choice, and compare most promising solutions currently on the market, based on the criteria.

In this article, firstly, we will define plagiarism, its types, and its detection ways. Further, we will present the extent of plagiarism surrounding Ukrainian universities. Later we will propose criteria for plagiarism detection systems, and finally we will compare the systems.

Theoretical background

The essence and precise definitions of plagiarism in higher education are argued in (Badge et al., 2009; Jensen et al., 2004; Warn, 2006). In (McCabe, 2005; OECD Reviews, 2017; Phillips, 2015; Roig, 2006; Survey Summery, 2013) various forms of plagiarism are discussed. The issue of accuracy and utility of the electronic detection systems were reviewed in the studies of (Badge et al., 2009; Bull et al., 2000; Purdy, 2005; Royce, 2003). Given the growing concern over the quality of education and in particular plagiarism issues in Ukrainian higher educational institutions, this article provides comparative analysis of three plagiarism detection systems and their most important features.

For the purpose of this article plagiarism is considered as "the practice of taking someone else's work or ideas and passing them off as one's own" (English Oxford Dictionaries). The main aim of the article is to compare electronic plagiarism detection systems, which can be used in the sphere of higher education in

Ukraine – as by students, as by Ukrainian higher institutions. Saying 'plagiarism detection system' we imply similarity detection systems, since all detectors check for similarities in texts and return reports on similarities. The matches spotted are the basis for a decision made by an expert if the text being checked is plagiarised or not (Badge et al., 2009; Goddard et al., 2005; Mulcahy et al., 2004). However, we will use hereinafter word 'plagiarism' to refer to namesake detection systems, since it is widely used and gets clear understanding of the essence and the aim of such detectors.

Plagiarism detection

Detection of plagiarism can be done manually. However, this type of detection is heavily time consuming and limited, e.g. teacher could easily identify similar home works among his students' just by reading them, however identifying plagiarised works with the resources available on the internet would require inputting line by line from ones work into search engines, and in case of rogeting – substituting some amount of plagiarised words (Warn, 2006), this approach won't be beneficial. Thus, automatic, software-assisted, plagiarism detection, which allows comparison of vast collection of documents in matter of seconds, is more preferred option.

There are basically two approaches for automatic detection of plagiarism, which have been developed so far and which depend on the object of checks:

1) screening for plagiarizing a text - a whole paper or a paragraph, which means that plagiarism detection system will check text documents;
2) screening for plagiarizing a source code, which means that plagiarism detection system will check computer programs.

Both textual plagiarism and source codes plagiarism occur in academic sphere and are spotted by electronic detection systems. However, source code plagiarism detectors are beyond the scope of this article. Therefore, referring to text-based plagiarism detection tools we omit words 'textual', 'text-based' and alike.

Textual copying may take different forms. There is a wide range of textual plagiarism types, some of which overlap. The most common types are defined below with the names they are also known as provided in parentheses (Bretag et al., 2009; Phillips, 2015; Survey Summery, 2013).

Secondary source (inaccurate citation) – using a secondary source, while citing only the original sources contained in the secondary one. Secondary source plagiarism leads not only to ignorance of the work done by authors of secondary sources, but it also forms wrong impression as for the scope of the review in the research.

Invalid source (misleading citation, fabrication, falsification, "404 error") – providing inaccurate information in the list of references – an incorrect or non-existent source, which might be as unintentional, demonstrating sloppy research, as intentional action in the pursuit of boosting the list of references.

Duplication (self-plagiarism, auto-plagiarism reuse, recycle) – reusing of the research results from one's own previous papers and presenting them as new ones. Though this type of plagiarism is highly debated and depends on the content copied. Furthermore, it overlaps with repetitive research, which is considered by some authors as a separate type of textual copying and means the repeating of text or data from a similar research with a similar methodology in another research to make it look new one and failing to cite properly.

Paraphrasing (plagiarism, intellectual theft, "find-replace", remix) – rendering someone's text, idea or a piece of research and rewriting it with other words, usually synonyms, without the reference to the original source. It may range from rephrasing to complete rewrite, but it keeps the idea of the original source.

Replication (author submission violation, self-plagiarism) – submitting the same paper to multiple publications, which leads to the publication of the same study more than once.

Misleading attribution (inaccurate authorship) – inaccurate list of authors of the paper, missing authors who denied credit for their contributions made to a study or the opposite - including authors, who did not make any contribution to a study.

Unethical collaboration (inaccurate authorship) – failing to cite in a study the authors who took part in collaborative work on a paper.

Verbatim plagiarism (copy-and-paste, intellectual theft, "ctrl-c") – copying of someone's words, whole paragraphs and works without proper attribution. There are two forms of copy-and-paste plagiarism. The first one is the citing of the source, where the direct text was copying from, but not indicating it as a quote. The other one is the providing of no reference at all and presenting someone's words, paragraphs as their own.

Complete plagiarism (stealing, intellectual theft, "clone") − presenting a paper, a study or other piece of work of another author as one's own and resubmitting it under one's own name.

Also there are several types of textual plagiarism described in (Phillips, 2015), which represent to some extent mix-type plagiarisms mentioned above.

The "Hybrid" refers to the combination of sources cited properly with copied passages from sources − not attributed at all − in one paper. There is a "mashup" type, which represents a combination of copied texts from different sources without proper attribution to them. However, this type may be considered as one of verbatim plagiarism or "copy-and-paste" forms. "Re-tweet" is based too closely on the words and/or sentence structure of the original text, though providing proper citation. If there is a proper citation in a paper, but it lacks originality, it is the "aggregator" plagiarism. The last two types of plagiarism − "retweet" and "aggregator" − we also consider as the variations of "copy-and-paste" forms.

Plagiarisms of these types listed above are happening with different frequency and characterized by different level of seriousness of academic misconduct. Both features depend on the area of application. A survey (Survey Summery, 2013), conducted within different areas of research (science, engineering, medical and social sciences, etc.) with a respondent pool of scientists from 50 different countries, found the five most common and the five most serious types of plagiarism (Figure 1, 2).

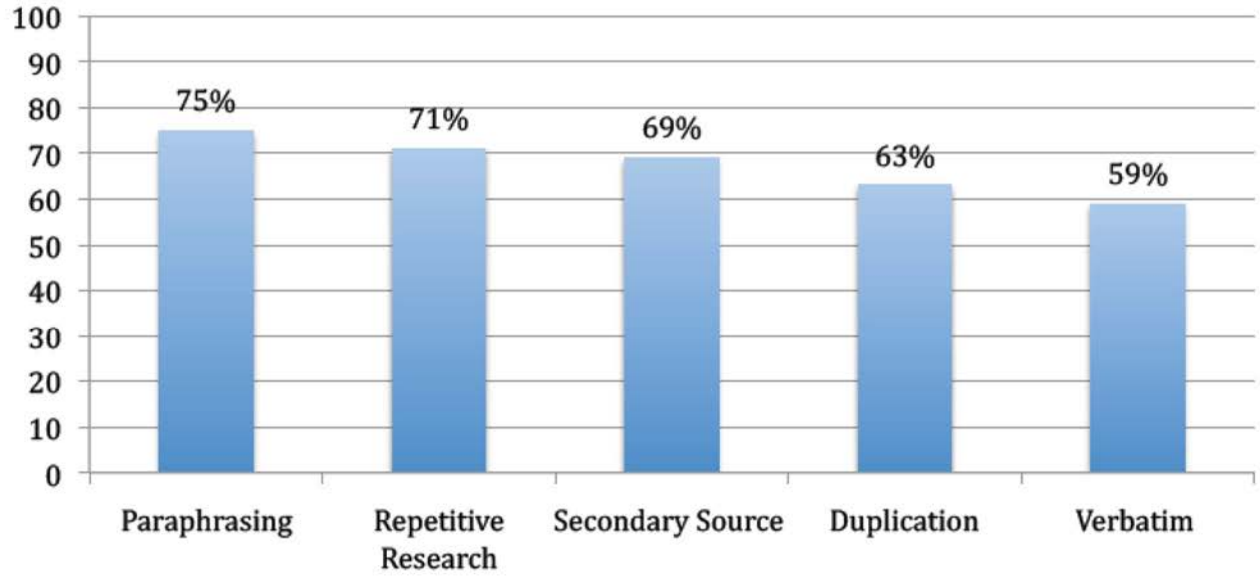

Figure 1. Most common forms of plagiarism in research around the world

Source: [18].

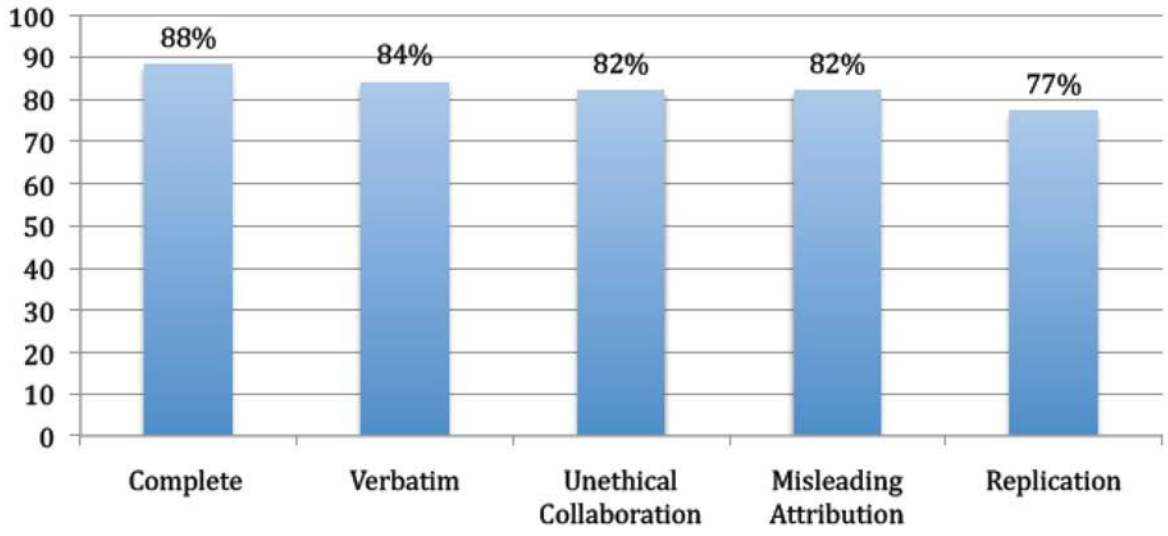

Figure 2. Most serious forms of plagiarism in research around the world

Source: [18].

The findings of another survey (Phillips, 2015), which surveyed higher and secondary educators − secondary instructors, educators at undergraduate as well as graduate schools − from around the world, showed most frequent and most problematic types of plagiarisms among students (Figure 3).

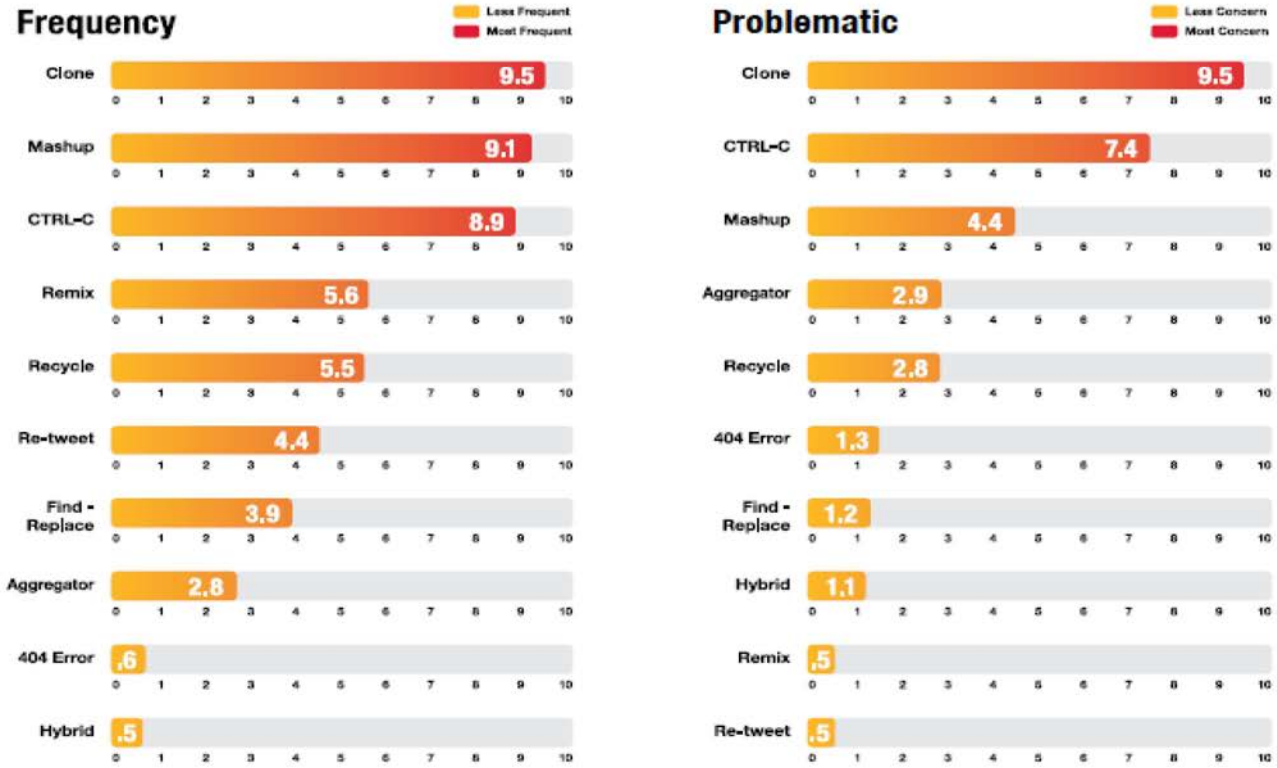

Figure 3. Frequency and seriousness scores of different types of plagiarisms among students around the world

Source: [13].

When it comes to students and scientists in Ukraine plagiarism is believed to be not only widespread but also a common practice in academic sphere. According to the analysis (IED, 2015) more than 90% of students plagiarize in different forms and "on average, no less than 50% of dissertations do not meet minimum standards of academic quality, or are plagiarized, or both". The same study provided a survey on the frequency of different types of plagiarism in Ukraine (Figure 4).

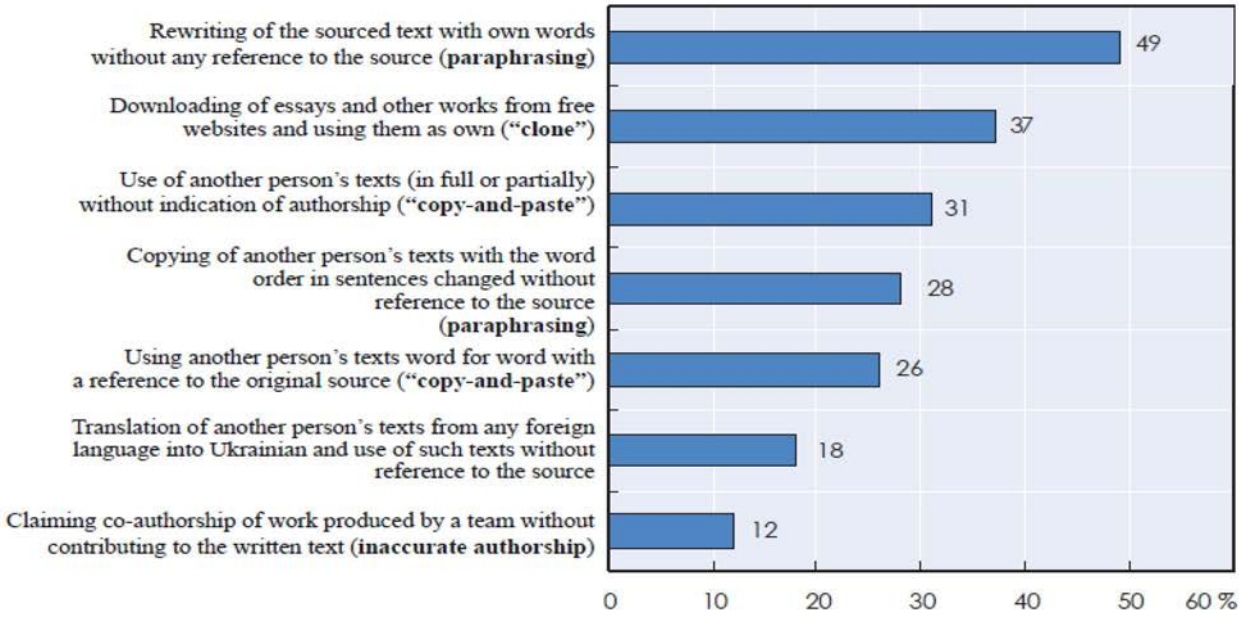

Figure 4. Frequency of different types of plagiarism among students and in research in Ukraine

Source: [8].

Therefore, there are types of plagiarism − as paraphrasing, complete plagiarism ("clone") and "copy-and-paste" with its variations − which are in top most common types around the world, including Ukraine, among students as well as researchers. As for the latter, self-plagiarism with its variations is also on the list. Apparently, the plagiarism detection tools are supposed to be able to detect the most common types of plagiarism. That is why these types of plagiarism should be added to the list of features for comparison of plagiarism detectors, let alone the most serious types.

Criteria for plagiarism detection systems

There is the list of the features of plagiarism detection systems, which are considered to be the most important ones for users in Ukrainian universities (Figure 5).

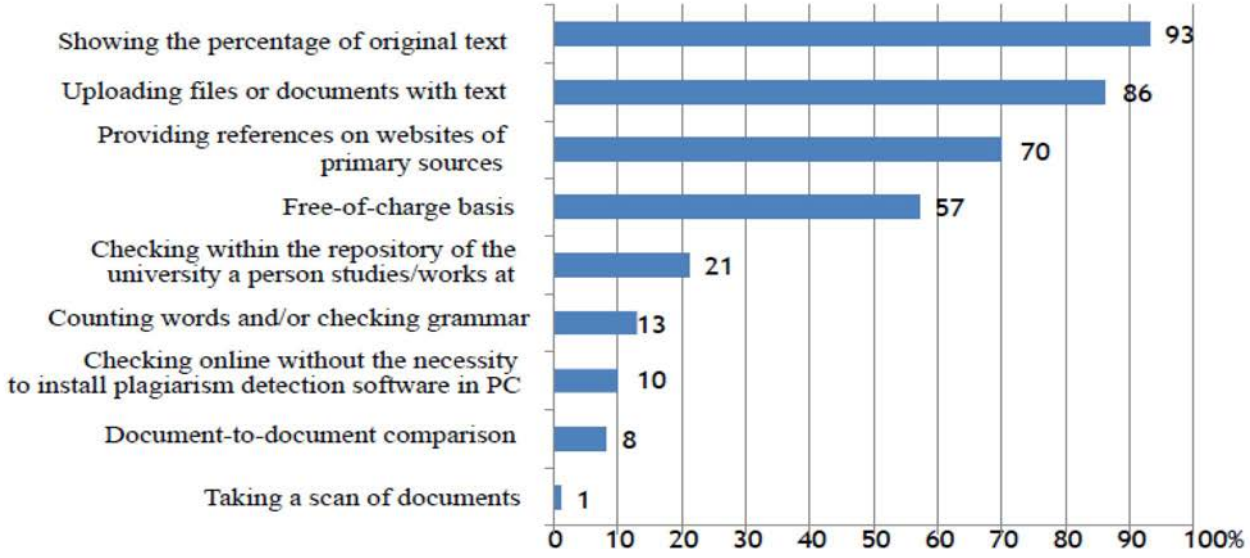

Figure 5. Rating of most important features of anti-plagiarism softwares according to Ukrainian higher educational institutions

Source: [1].

We suggest 4 criteria, where each of them has a set of features, for plagiarism detection systems comparison and choice – (1) affordability, (2) material support, (3) functionality, and (4) showcasing.

Affordability describes how a system is coping with organizational restrictions in a form of policies and available resources – are there monetary resources available to purchase a system and use it for a needed group of people. It can also relate to accessibility, whether system can be implemented and used within organizational infrastructure – for instance third party software install within computers on university campus, or country based restriction for the internet access.

Material support aims at examining what types of plagiarized information is needed to be deal with. It can be a text, a picture, an audio format, or links to different resources.

Functionality deals with the plagiarism detection itself. It describes how a system helps in detecting plagiarized works. Whether system allows to determine to which extent works are similar, or in which extent work is original. Does it allow to compare to external resources and databases, or within uploaded works itself.

Showcasing is a system's ability to establish the discussion between the parties that involved in resolving the plagiarism case. For instance, including a student into a system and showcasing plagiarism results on his/her work. Table 1 shows criteria and features for each of it.

Table 1. Groups of features for plagiarism detection systems comparison and choice

Affordability		
Free-of-charge-basis	Checking without registration (demo-version/test-version)	Checking online without the necessity of installing to PC
Uploading files from cloud services	Average speed of checking	Accounts with storage
Material support		
Uploading files or documents with text	Text-format support	Image-format support

Table 1 (cont.). Groups of features for plagiarism detection systems comparison and choice

Functionality		
Showing the percentage of original text (originality score)	Showing the percentage of similarities (similarity score)	List of original sources
List of links to websites of primary sources	Option of downloading a similarity report	Internet checking
Checking within open access sources	Databases checking	Checking within the repository of the university a person studies/works
Document-to-document comparison	Citation and reference recognition	Checking for paraphrasing
Checking for self-plagiarism	Language support	Checking for translated plagiarism
Grammar check	Multiple document support	Detection of hidden symbols, replaced letters from other alphabets
Showcasing		
Commenting (student-teacher feedback)		

Plagiarism detection systems

There are several plagiarism detection systems available in Ukraine, which can be divided in two groups. The first one includes commercial or fee-based official plagiarism detectors, such as Unicheck, Strikeplagiarism, Anti-Plagiarism. The second group comprises of free-of-charge plagiarism detection systems, such as eTXT, Advego Plagiatus, Anti-Plagiarism etc. Although the free-of-charge status of systems in the second group is quite debatable, which is demonstrated further on the example of one of these softwares.

For the purpose of this article three plagiarism detection systems were chosen for comparison. They are Unicheck, eTXT and Turnitin (ETXT; Turnitin; Unicheck).

Unicheck (former Unplag)— is a relatively new online plagiarism detection software, which was launched in 2014 and can be used by individuals, separate departments or the whole institution. It makes checks of uploaded documents against the Internet − web pages indexed by Google and Yahoo − as well as against Open Access Sources and personal library. It supports almost all text file formats. Unicheck was chosen as an example of plagiarism detection systems of the first group, because others fail to have features considered important for this comparison. For instance, Anti-Plagiarism does not have any free test version to try it; Strikeplagiarism does not allow to check several documents at a time, has some difficulties uploading large files and with PDF files.

The other anti-plagiarism system, which was chosen to represent the second group, is eTXT. It is one of the most popular free-of-charge anti-plagiarism programmes in Ukraine, Russia, Belarus, Kazakhstan. It has quite similar features to other plagiarism detection systems of the same group, but a bit more. For instance, Advego does not have the option of multiple document checking.

Turnitin was chosen as the third one for comparison to represent a plagiarism detection system, which is one of the global leading anti-plagiarism software with rather long history of dealing with plagiarism.

Comparative analysis of plagiarism detection systems

Table 2. Comparison of plagiarism detection systems

Features	Turnitin	Unicheck	eTXT
Free-of-charge basis	No	No	Yes
Checking without registration (demo-version/test-version)	No	Yes in demo-version	No
Checking online without the necessity of installing to PC	Yes	Yes	No
Uploading files or documents with text	Yes	Yes	Yes
Text-format support	Any	Most of text formats	Most of text formats
Image-format support	Yes	No	Yes
Uploading files from cloud services	Yes	Yes	Yes
Showing the percentage of original text (originality score)	No	Yes	Yes
Showing the percentage of similarities (similarity score)	Yes	Yes	Yes

Table 2 (cont.). Comparison of plagiarism detection systems

Features	Turnitin	Unicheck	eTXT
List of original sources	Yes	Yes	Yes
List of links to websites of primary sources	Yes	Yes	Yes
Option of downloading a similarity report	Yes	Yes	Yes
Internet checking	Yes	Yes	Yes
Checking within open access sources	Yes	Yes	Yes
Databases checking	Yes	No	No
Checking within the repository of the university a person studies/works	No	Yes	No
Document-to-document comparison	No	Yes	No
Citation and reference recognition	Yes	Yes	Yes/No
Checking for paraphrasing	Yes	Yes	Yes
Checking for self-plagiarism	Yes	Yes	No
Language support: Ukrainian	No	Yes	Yes
Checking for translated plagiarism	Yes	No	Yes
Average speed of checking	10 seconds per page	4-10 seconds per one page (275 words)	Depends
Accounts with storage	Yes	Yes	No
Grammar check	Yes	No	Yes
Multiple document support	Yes	Yes	Yes
Commenting (student-teacher feedback)	Yes	Yes	No
Detection of hidden symbols, replaced letters from other alphabets	No (for Ukrainian)	Yes	Yes

Both Turnitin and Unicheck do not have free versions. However, Unicheck provides a demo-version to check a text up to 500 words and three checks. Whereas Turnitin is said to offer demo check as well, but as a matter of actual practice it rather has a test version of interface of the software than a testing of the software itself. In order to check a text online on eTXT a user has to register and to have a rating, which allows to make a few checks with maximum 5000 characters each. The other option is to download an application of eTXT and install it. However, after installation in order to check long texts this system requests CAPTCHAs all the time, which makes impossible to leave documents for checking alone, because the application freezes up in that case. To avoid this disadvantage is possible, but only in a paid-based version of the app, which in its turn equates it with Unicheck and Turnitin in terms of free-of-charge basis. Furthermore, in the light of recent cyber attacks in Ukraine online plagiarism checking is safer, since it does not require installation of applications to personal computers.

At this point, it is worth mentioning that Unicheck provides trial version of its software for universities, though with limited number of pages for a free check. Unicheck also started to give the free check for universities' scientific journals in Ukraine, which makes this plagiarism detector definitely the promising one for the use in higher educational institutions of Ukraine.

As for file-format support, only Unicheck does not support image-formats. Since we are considering systems of text-based plagiarism detection, not supporting image-formats does not count as a crucial disadvantage.

As it was mentioned above in Figure 5, showing the percentage of original text is important feature for users. This can be explained by the regulatory documents in most of higher educational institutions in Ukraine, which contain a "barrier" − minimum originality score − for the text to be approved by educators. In this respect, this feature makes Unicheck and eTXT more suitable for the use in Ukraine than Turnitin.

As for database scope, Turnitin definitely has a huge one, which includes the Internet sources, scientific journals (the largest number among plagiarism detection software) and its own database formed from papers submitted by users. Though the latter might be considered as a disadvantage of this software, since it does not have an option to delete papers submitted for check in its database. As for database scope of Unicheck, it is comprised of two sources, which are the Internet and personal database - Personal Library. What is good about the latter that it has an option of choosing access rights and papers submitted earlier can be deleted.

Furthermore, Unicheck allows to conduct document-to-document comparison unlike other systems.

What stands out Turnitin is its feature to detect paraphrasing, which is available in other plagiarism detection softwares to some extent. All the plagiarism detection systems can detect low-level paraphrasing. Though, when a text is paraphrased thoroughly, it is difficult almost for any programme to spot plagiarism.

Both Turnitin and Unicheck make checks against user's documents uploaded earlier, which can prevent the one from self-plagiarism.

As for another type of plagiarism, which is not easy to be spotted − translated plagiarism or multi-language plagiarism, Turnitin seems to have this feature as well. However, in reality, after two and more translations from one language to another it fails to detect it. The same happens when it comes to Russian-Ukrainian and inverse translations. It is the case for both Turnitin and Unicheck, while for eTXT it is not a big problem and in most cases the latter is able to detect translated plagiarism within these two languages. Though when more rounds of translation are done and more languages are included it fails to detect it. Apparently, in terms of this kind of plagiarism no kind of plagiarism detection system can catch it. It is the case when only manual checking might be helpful.

Grammar check is not available in Unicheck system. Instead it can detect letters replaced with ones from another language, spotting more than 90% of copied and amended text. Turnitin is supposed to detect letters from other alphabets as well, but it does not recognize special symbols from other languages and exclude words containing them, which may decrease the amount of matches and hence to increase the originality score. As for eTXT it also can detect hidden symbols and letters from other alphabets, considering them as mistakes and not influencing similarity score.

As for the option of commenting, eTXT does not have it, which makes it less appropriate and useful for using in universities during the educational process.

As for citation and list of references recognition, Turnitin and Unicheck deal with both, whereas eTXT has an option to exclude quotations from the screening, but nothing concerning lists of references.

One peculiar fact was found about Unicheck system. The efficiency of check appeared to depend on the format of a file uploaded for screening as well as on the language the checking text was written in. The best results were demonstrated when the file was in DOC format and Unicheck detected 100% similarity. When the same text was uploaded in PDF format, the system detected less similarities and scored around 80%, considering the rest as original text though it was entirely duplicated one. The least similarities were found when the same text was uploaded in Pages format and the originality score showed more than 90%. These findings related to the text in Ukrainian. However, when the text was in English the system detected 100% similarities regardless the format of document uploaded for check.

Last but not least, due to the introduction of the presidential decree banning Russian websites eTXT requires proxy servers to be used alongside. The point is that it utilizes Russian search engines − Yandex, Rambler etc.

Discussion and conclusions

Although, the wide access to the Internet allows easier to plagiarize ones work, it can also allow to counter-attack plagiarism, and due to easy accessible information to identify and locate plagiarized sources. This feature is one of many advantages that plagiarism detection system can provide for a higher education institution. In this article we compared plagiarism detection systems, by proposed criteria and set of features, in the context of higher education in Ukraine.

While we can conclude that all compared programs have advantages and disadvantages, each institution needs to discuss and decided importance of criteria and features aggregating them, for a plagiarism detection system's selection. First of all it depends on the purpose of the one, conducting checks. If it is an individual, eTXT and other similar plagiarism detectors are more appropriate for personal use. When it comes to institutions' use, such plagiarism checkers as Turnitin or Unicheck are more suitable. Since at the present for Ukrainian universities there are few options available, Unicheck appears to be one of the most appropriate and efficient, meeting most of our criteria in the comparison.

However, any plagiarism detection system is a tool for spotting similarities and any decision as to whether they are plagiarism or not should be taken by a person, who has relevant expertise and is able to make a qualitative judgement. Needless to say, any university should have quality assurance system and anti-plagiarism policy for prevention as well as regulation of plagiarism cases.

References

1. Analitychna dovidka za rezultatamy doslidzhennia praktyk akademichnoi dobrochesnosti u vyshchykh navchalnykh zakladakh Ukrainy. Ministerstvo osvity i nauky Ukrainy. Instytut osvitnoi analityky (2016). Kyiv.
2. Badgc, J., & Scott, J. Dcaling with plagiarism in thc digital agc (2009). Rctricvcd from http://evidencenet.pbworks.com/f/Badge_Scott_plagiarism.pdf.
3. Bretag, T. & Mahmud, S. (2009). A model for determining student plagiarism: Electronic detection and academic judgement. *Journal of University Teaching & Learning Practice*, *6*(1). Available at http://ro.uow.edu.au/jutlp/vol6/iss1/6.
4. Bull, J., Colins, C., Coughlin, E. & Sharp, D. (2000). Technical review of plagiarism detection software report. Retrieved from http://www.jisc.ac.uk/media/documents/programmes/plagiarism/southbank.pdf. Accessed 31 October, 2009.
5. English Oxford Dictionaries. Available at https://en.oxforddictionaries.com/definition/plagiarism.
6. ETXT website. Available at https://www.etxt.ru.
7. Goddard, R., & Rudzki, R. (2005). Using an electronic text-matching tool (turnitin) to detect plagiarism in a new Zealand university. *Journal of University Teaching & Learning Practice*, *2*(3).
8. IED (2015). Academic Culture of Ukrainian Student Community: Primary Factors of Formation and Development, Kyiv. Available at http:// iro.org.ua/uploads/96491691.pdf.
9. Jensen, J., & De Castell, S. (2004). 'Turn it in': Technological challenges to academic ethics. Education, Communication & Information, *4*(2), 311-330.
10. McCabe, D. (2005). Cheating among college and university students: A North American perspective. *International Journal for Educational Integrity*, 1(1). Available at https://www.ojs.unisa.edu.au/index.php/IJEI/article/view/14/9.
11. Mulcahy, S., & Goodacre, C. (2004). Opening pandora's box of academic integrity: Sing plagiarism detection software. *Beyond the Comfort Zone: Proceeedings of the 21st ASCILITE Conference,* 688-696.
12. OECD Reviews of Integrity in Education: Ukraine 2017. Academic dishonesty − cheating and plagiarism in Ukrainian higher education. Available at http://dx.doi.org/10.1787/9789264270664-13-en.
13. Phillips, V. (2015). Turnitin Report. Available at https://www.geteducated.com/elearning-education-blog/10-types-of-plagiarism-and-academic-cheating/.
14. Purdy, J. (2005). Calling off the hounds: Technology and the visibility of plagiarism. *Pedagogy*, *5*(2), 275-296.
15. Roig, M. (2006). Avoiding plagiarism, self-plagiarism, and other questionable writing practices: A guide to the ethical writing. Retrieved from http://facpub.stjohns.edu/~roigm/plagiarism/Index.html. Accessed 29 January 2008.
16. Royce, J. (2003). Trust or trussed? Has turnitin.com got it all wrapped up? *Teacher Librarian*, *30*(4).
17. Stein, Benno; Koppel, Moshe; Stamatatos, Efstathios (2007). Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection PAN'07. *SIGIR Forum*, *41*(2).
18. Survey Summery (2013). Research Ethics: Decoding Plagiarism and Attribution in Research. Researcher Insights into the Types of Plagiarism & Attribution Issues. Available at https://www.ithenticate.com/resources/infographics/types-of-plagiarism-research.
19. Turnitin website. Available at http://www.turnitin.com.
20. Unicheck website. Available at https://unicheck.com.
21. Warn, J. (2006). Plagiarism software: No magic bullet! *Higher Education Research and Development*, 25, 195-208.