

Language based plagiarism detection

Matija Kaniški

Faculty of Organization and Informatics

University of Zagreb

Pavlinska 2, 42000 Varaždin, Croatia

matija.kaniski@foi.hr

Abstract. *Plagiarism is constantly evolving and it occurs in almost all language areas, taking different kind of forms and shapes. To create a sophisticated plagiarism, the linguistic abilities of a particular language are needed. Today's Internet content, adapts more and more to its users and language areas. English is not any more so prevalent on the web because contents in other languages occurs. As plagiarism increasingly evolves development of new solutions that they can compete with is crucial. One of these temporary solutions is the PlagScan plugin for the learning management system Moodle. PlagScan has among other things an exceptional ability to find plagiarism in the Croatian language. So we're not talking about local search and comparison of plagiarism, but rather finding matches of plagiarized content with the original web content in the Croatian language with extremely low cost.*

Keywords. Plagiarism, PlagScan, Detection, Language, Comparison, Turnitin, Crot Pro, Urkunde, VeriCite

1 Introduction

The development of technology makes it easier to access the contents or the appropriation of phraseology and paragraphs by the simple command "copy/paste". Although the development of modern technology solve the problems associated with writing papers and availability of data, the problem of plagiarism and violations of academic honors has increased continually in a progressive matter. It is difficult to determine the exact reasons for plagiarism, but knowing human nature, we can assume that each person tends to achieve success in a lighter and easier way, as in this case, the appropriation of finished content. Thus, frequency of plagiarism is increasing, the development of information and communication technology facilitates unauthorized downloading of contents, but thanks to the same technology, various computer programs and online services are developed which helps to detect plagiarism (Baždarić et. al.,

2009). One of the most popular system for creating online-courses based on socio-constructivist paradigm of learning is the learning management system Moodle (Jadrić, Čukušić & Lenkić, 2013). Lately, the demand for use of e-learning systems like Moodle, Merlin, MuDri, Loom and many others has increased. Such systems represent a content repository for course materials, and student works. Databases with a large number of papers, are very suitable for plagiarism in various forms, where the most commonly used method is "copy/paste". Although there are many alternative software to detect plagiarism, mostly in the form of desktop application, priority of their use is the economical factor. Since systems for e-learning are dominating and intensively used in Universities it is necessary to apply solutions in the form of upgrades that would allow the current system an instant content check and plagiarism search. In this way there is no wastage neither of time nor resources. Desktop applications require a certain PC performance, while upgrades to the system for e-learning is not so demanding in terms of performance for the PC of the user. So far, there exist few software to detect plagiarism in the form of plugins for specific e-learning systems. For example, previous research shows that authors already used and tested plugins Vericite and Crot Pro on the Moodle system. They showed exceptional results by searching the web sources (Hercigonja & Vukovac, 2015). This test was performed as a search for matches with the corresponding web sources. Similar research has been done by Biggam and McCann in 2010 with the aim of raising awareness about plagiarism and improve referencing skills of students (Le Nguyen et. al., 2013). Another study of the same type intends to take place at the University of Rijeka using Turnitin plugin that will automatic scan all thesis's and other papers for plagiarism. All these plugins give high hope for reducing the rate of plagiarism, but the usual problem is the price and possible restrictions for using them. For example, the license per year of the Turnitin software is 12,000.00\$ ("Srednja.hr", 2016). The objective of this case study is to address the strengths and weaknesses of using software for plagiarism detection for different languages as well as their

comparison. The solution that is proposed in this paper, show better results than existing plugins for plagiarism detection in the Croatian language both in this price domain as well as in the domain of the success of finding plagiarism.

The rest of this paper is organized as follows: Section 2 describes the current state of the art in this field, Section 3 presents the result of the performed case study, Section 4 provides comparisons of various tools for plagiarism detection, Section 5 gives an outline of future research steps that needs to be done in order to improve plagiarism detection, and Section 6 concludes the paper.

2 Recent work

The problem of academic plagiarism has been present for centuries. Academic plagiarism is defined as the use of ideas and/or words from sources without giving due acknowledgement as imposed by academic principles. Observations of academic plagiarism reveal a variety of commonly found forms: literal plagiarism, shake and paste plagiarism, paraphrasing, technical disguise, translated plagiarism, idea plagiarism and self-plagiarism (Meuschke, N., & Gipp, B., 2013). Researchers discovered that many of today's the proposed methods for plagiarism detection have a weakness and lacking for detecting some types of plagiarized text. Some of recently proposed plagiarism detection techniques that are widely in use can be classified into character-based methods, structural-based methods, classification and cluster-based methods, syntax-based methods, cross language-based methods, semantic-based methods and citation-based methods (Osman, A. H., Salim, N., & Abuobieda, A., 2012). Most of the work in document plagiarism has been done for academic purpose. Detecting plagiarism is important to judge and mark students' work especially for postgraduates who are strictly prohibited from cheating, rewording, rephrasing, or restating without referencing. In this regard, numerous plagiarism detection systems have been developed. These systems can be classified into two main categories, web-enabled systems and stand-alone systems (Bin-Habtoor, A. S., & Zaher, M. A., 2012). Turnitin is the most well-known commercial plagiarism detection system to which many universities from UK and USA subscribe. It uses an enormous database from the Internet and previous student works to be compared with the query document. (Heckler, N. C., Rice, M., & Hobson Bryan, C., 2013). In the 2004 calendar year, Massey University had 41,436 students enrolled in five Colleges – Business; Creative Arts; Education; Humanities & Social Sciences; and Sciences. The University is spread across three physical campuses. The Turnitin system produces reports which identify the percentage of other text used in an assignment, as

well as a colour grading indicator for assignments which ranges from red (up to 100% copied) through orange, yellow, and green to blue. Preliminary results of this first trial involving 949 assignments over classes controlled by nine lecturers found around 9% of assignments falling in the 'bad' Turnitin yellow to red levels (25% to 100% reported copying) (Goddard, R., & Rudzki, R., 2005). Next we mention cross-language plagiarism. Cross-language plagiarism occurs if a text is translated from a fragment written in a different language and no proper citation is provided. Regardless of the change of language, the contents and, in particular, the ideas remain the same. Whereas different methods for the detection of monolingual plagiarism have been developed, less attention has been paid to the cross- language case. Authors compare two recently proposed cross-language plagiarism detection methods (CL-CNG, based on character n-grams and CL-ASA, based on statistical translation), to a novel approach to this problem, based on machine translation and monolingual similarity analysis (T+MA). They explore the effectiveness of the three approaches for less related languages. CL-CNG shows not be appropriate for this kind of language pairs, whereas T+MA performs better than the previously proposed models (Barrón-Cedeno, et. al.,2010). E-learning is also becoming an increasingly common if not an essential strategy in academic institutions. However, this new teaching mode also brings about new forms of academic misconduct. The most common ones in students' assignments we identified were the following (Orthaber, S.,2009):

- Language, writing style, vocabulary, tone, and grammar were above the students' actual level;
- Pronouns did not correspond to the gender of the text producer;
- Web sites listed in citations were inactive;
- Look of grey letters in the text;
- Strange and poor layout such as more than one fonttype in a single text, references to hyperlinks, paragraphs with background color, strange texts or lines at the top or bottom of the page were indications that the text was downloaded or copy- pasted from the web;
- References to accompanying material that was not included in the text;
- Quotes in the paper did not have citations;
- The length of the paper was considerably longer than instructed;
- Certain information in the copy-pasted text was changed to the extent that the text did not make sense anymore.

The problem described above is not isolated to this particular case. In fact, with the spread and wide use of the internet, internet plagiarism is becoming a ubiquitous issue at several educational institutions.

3 PlagScan plagiarism detection software

PlagScan plugin is specialized for detection of “copy/paste” manifestation of plagiarism (see Figure 1). PlagScan works by searching for matching content from web sources. Price for the licensed product is 19.99\$ on a monthly basis.

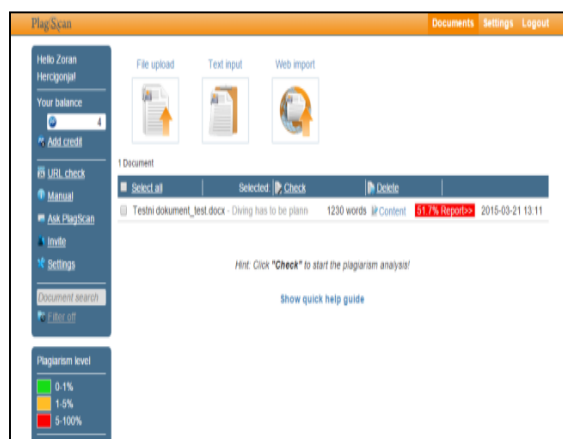


Figure 1. PlagScan interface

The trial version of the product includes a 30 day try out period with additional restrictions. The trial version requires a certain number of words, maximum 2,000 words per document (“Moodle plugins Library: PlagScan Plagiarism”, 2016). The overall quality of plagiarism search patterns, determining the level of copied content as well as the thoroughness of the report, is equal to the quality of the licensed version of PlagScan (see Figure 2).

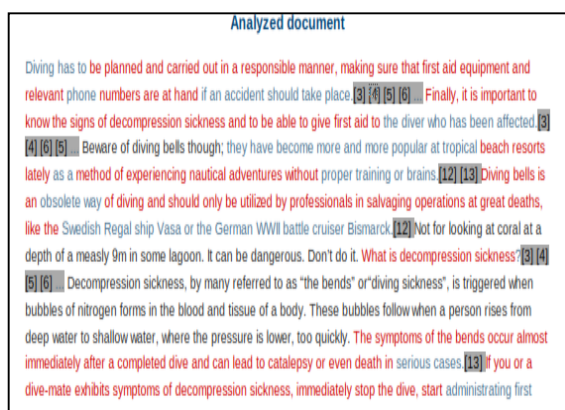


Figure 2. Document analysis

PlagScan supports multiple Moodle versions: 2.0, 2.1, 2.2, 2.3, 2.4, 2.5, 2.6, 2.7, and 2.8. It also supports several different types of document formats: pdf, doc and docx (“PlagScan Plagiarism Checker”, 2016).

3.1. Materials and methods

For the purpose of testing and quality control of this plugin, two tests were performed. Both tests were searching correspondence between the tested content and the web sources. The aim was to compare the quality of the plugin by conducting the process of plagiarism detection for the content in English and Croatian language. The duration time of the process of detection and the percentage of detected plagiarism was measured. Two test documents on the same topic were made: one in Croatian language and the other one in English. Each document contained the same number of words: 873 words in total. The number of words used is reduced to 873 words because of the trial restrictions for PlagScan. Contents are taken from the same source, Wikipedia on the topic “History of the steam engine.” Wikipedia was used because the frequency of visiting and the supports for multilingual presentation of content. It is well known that Wikipedia published content in a various number of languages. That was the reason for using such contents. The contents are taken directly with the method “copy/paste”.

3.2. Results

In both conducted test the goal was to determine the quality level of the plugin. Therefore, the following quality indicators were specified:

1. The duration time of the process of detection,
2. Percentage of copied content and
3. The number of detected web sources.

3.3.1. The test document in English

The duration time of the process of detection was 6.4 seconds. The process of detection found concurrency of 94.0% within the content (see Figure 3).

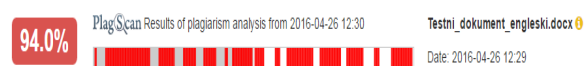


Figure 3. Percentage of plagiarism

High level of copied content is announced and marked in red color. The process of detection revealed two levels of plagiarism: complete and partial plagiarism. Part of the content is marked in red color and described as exact match or a complete coincidence while the other part of the content is marked in green color described as possibly altered text through modification of the text. This can be interpreted as text recognition of partial plagiarism that was not fully taken up by the method “copy/paste” but rather was slightly modified by using own words of the author (see Figure 4).

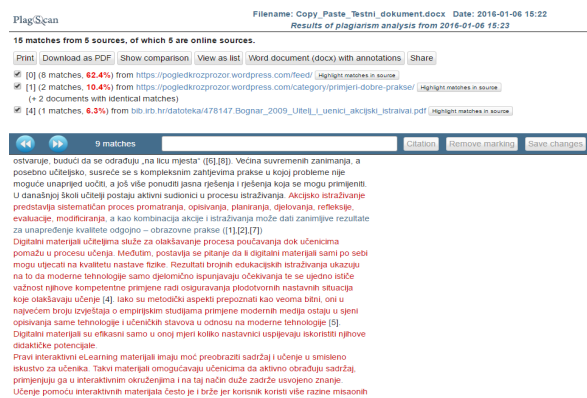


Figure 4. Detailed analysis

Within the text, PlagScan made additional analysis. PlagScan marked the content in red, blue and green color. In the text he was able to recognize pieces of content that have been cited. This category is called marked as quotation. Number of located Internet sources was 29.

3.3.2. The test document in Croatian language

The duration time of the process of detection was 2 seconds. The process of detection found concurrency of 94.3% within the content (see Figure 5).

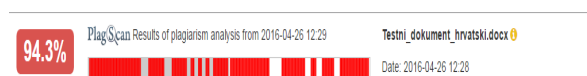


Figure 5. Percentage of plagiarism

Number of located Internet resources was 11. PlagScan as in the first test made within the content an additional selection. Part of the content is marked in red and described as exact match or a complete coincidence while the other part of the content is marked in green color described as possibly altered text that is the result of modification of the text (see Figure 6).

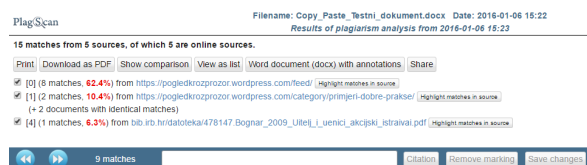


Figure 6. Detailed analysis

In the text PlagScan did not identify parts such as the cited content, although such content in the test document was present. So, he did not recognize the category marked as quotation. This is the primary disadvantage in comparison with the previous test over English content (see Table 1).

Table 1. Comparison of results






Tests/Features	1 st test	2 nd test
The duration time of the process of detection	6.4 seconds	2 seconds
Percentage of plagiarized content	94.00%	94.03%
Number of located Internet sources	29	11

The performed tests showed excellent results. Percentages of the plagiarized content reached almost 100%. Judging by the number of detected web sources, the first test with the content in English, took three times longer than the second test with the content in Croatian language. Although both test documents contained the same number of words for the analysis (total 873) the cause of a prolonged duration for the first test is the existence of multiple web sources with the same content (total 29). It is evident that the test content in Croatian language resulted in identification of only 11 sources. Test content in Croatian language showed a higher percentage of copied content in relation to the test content in English language. The difference in percentage between both test is very small, only 0.3%. In the test with the content in Croatian language the category marked as quotation was not recognized, which resulted in an increase in the percentage of plagiarized content of 0.3%. For example, this category has been successfully recognized and registered with the test content in English. Therefore, these 0.3% were not recorded in the category of complete plagiarism. In both tests PlagScan proved its quality in finding plagiarism at both speaking regions.

4 PlagScan versus various plagiarism detection software

For the purpose of testing and comparing the quality of PlagScan versus other plagiarism software a new test document was created. The prepared test document contained common knowledge on the topic of diving. Sentences and fragments of the test document were chosen randomly and literally taken from Web sites, international blogs and chat groups by the method of “copy/paste”. Due the restrictions of API call’s the sample document contained exactly 1,410 words and fifty different sources in English.

Table 2. PlagScan versus various plagiarism detection software

	PlagScan	VeriCite	URKUND	Turnitin	Crot Pro
Accessories / Features					
The time required to detect plagiarism	5 minutes and 14 seconds	3 seconds	5 seconds	8 minutes and 34 seconds	24 hours
The percentage of plagiarized content	51.7%.	95%	91%	64%	26%
The number of detected web source from a total of 50	27	43	35	10	9
Sensitivity for pattern recognition	High-level words in a row	High-level line	High-level line	Low-level section	Low-level section

According to the table 2 the best results in all of these features shows VeriCite (“VeriCite”, 2016). The time required for the process of detecting plagiarism is extremely short. Also the obtained output results for recognized plagiarized content are very convenient. From 95% VeriCite identified 43 sites from where the content was taken. Urkunde (“Urkunde”, 2016) literally compete with VeriCite if we look at the time required to detect plagiarism that took just about few seconds. Also the percentage of determined plagiarized content, the number of detected web sources and the sensitivity for pattern recognition at the level of the word are very similar. On the other side PlagScan (“PlagScan”, 2016), a commercial plugin, who is not distinguished by the best results, direct compete and exceeds the listed plugins at the level of pattern sensitivity. PlagScan recognizes and searches web sources using individual words in the pattern. This high level of sensitivity contributes to its quality. Turnitin (“Turnitin”, 2016) and Crot Pro (“Crot Pro”, 2016) have common features regarding the number of detected web sources and the low sensitivity for pattern recognition. The last category contains plugins which are very poor quality, one of them is the free plugin Crot Pro. The very process of analysis and detection take too long considering meager results of the report. Compared with commercial plugins, Crot Pro is very poor regarding all features listed in the table. From all of the tested plugins listed in the table, taking into account the basic functionality and performance results to detect plagiarism, commercial plugins VeriCite and PlagScan have proven to be better and are recommended to use. Of course also the plugin Urkunde is a direct rival for the plugin VeriCite, both in functionality and in the results of detecting plagiarism. If we compare the price of the license for

the two plugins, it can be said that the one-year license for Urkunde is too high in relation to its functionality, detailed reports and the use of available repositories. For example, VeriCite with almost identical features and functionality achieved about the same results with half the amount of the one-year license that Urkunde demands. For Urkunde we are annually required to pay 2058.28\$, while VeriCite’s monthly subscription requires only 93.00\$ what would calculated on an annual basis be 1116.00\$. Urkunde is overpaid for its functionality. Besides that VeriCite and its functionality includes the ability to detect plagiarism on a local basis of documents that are submitted under the same Assignment. Furthermore, the monthly subscription of 19.99\$ for the plugin PlagScan is completely justified. PlagScan has a very high pattern sensitivity at the word level, and generates extremely detailed reports with lots of information. Such functionality and features correspond to the proposed amount.

5 Discussion

Under consideration for further research, it is necessary to determine the quality of the software through the search for other forms of plagiarism, such as “Shake&Paste”. “Shake&Paste” plagiarism implies that the plagiarist has managed to fit copied contents from a number of different authors in the same section. In this section the plagiarist can now combine multiple taken contents from different authors or dismiss repetitive portions or connect parts of taken sentences. This kind of plagiarism is very hard to detect. Former software in this area are not successful

for detecting “Shake&Paste” plagiarism. They must be further developed. Another important question for future research is to improve the software so that it can fully distinguish clearly marked quotes from copied contents which was the case with the performed analysis of this work.

6 Conclusion

Plagiarism as a tribute now has a serious and highly sophisticated enemy, multilingual software for plagiarism prevention, not only in world languages, but also in Croatian language as well. Such solutions could lead to a reduction in the number of plagiarism for the overall field of education. PlagScan showed a very high quality through the use of its functionality that can truly be of great help to education. He could be the foundation for the development of similar software. But the development of highly sophisticated software to detect plagiarism, raises the question of payment for the license on the one side and the return of investment on the other side. If the user is a teacher and he needs to detect plagiarism for student assignments several times a year, then the full potential of this plugin is unused. The license cost for the plugin of several thousand dollars per year does not return the investment. On the other side a multilingual software to detect plagiarism would be a good solution in terms of cost-effectiveness license payment.

References

- Barrón-Cedeno, A., Rosso, P., Agirre, E., & Labaka, G. (2010, August). Plagiarism detection across distant language pairs. In *Proceedings of the 23rd International Conference on Computational Linguistics* (pp. 37-45). Association for Computational Linguistics.
- Baždarić, K., Pupovac, V., Bilić-Zulle, L., & Petrovečki, M. (2009). Plagiranje kao povreda znanstvene i akademske čestitosti. *Medicina Fluminensis*, 45(2), 108-117.
- Bin-Habtoor, A. S., & Zaher, M. A. (2012). A Survey on Plagiarism Detection Systems. *International Journal of Computer Theory and Engineering*, 4(2), 185.
- Crot Pro. (2016). Retrieved from <http://www.crotsoftware.com>
- Heckler, N. C., Rice, M., & Hobson Bryan, C. (2013). Turnitin systems: A deterrent to plagiarism in college classrooms. *Journal of Research on Technology in Education*, 45(3), 229-248.
- 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), 7.
- Hercigonja, Z., & Vukovac, D. P. (2015, January). Plagiarism Detection with Moodle Plugins. In 17. CARNetova korisnička konferencija-CUC 2015.
- Jadrić, M., Čukušić, M., & Lenkić, M. (2013). „E-učenje: Moodle u praksi“. *Ekonomski fakultet u Splitu*.
- Le Nguyen, T. T., Carbone, A., Sheard, J., & Schuhmacher, M. (2013, January). Integrating source code plagiarism into a virtual learning environment: benefits for students and staff. In *Proceedings of the Fifteenth Australasian Computing Education Conference-Volume 136* (pp. 155-164). Australian Computer Society, Inc..
- Meuschke, N., & Gipp, B. (2013). State-of-the-art in detecting academic plagiarism. *International Journal for Educational Integrity*, 9(1).
- Moodle plugins Library: PlagScan Plagiarism. (2016). Retrieved from https://moodle.org/plugins/view/plagiarism_plagscan
- Orthaber, S. (2009). Detecting and preventing internet-plagiarism in a foreign language e-learning course. *International Journal of Advanced Corporate Learning*, 2(2), 20-4.
- Osman, A. H., Salim, N., & Abuobieda, A. (2012). Survey of text plagiarism detection. *Computer Engineering and Applications Journal (ComEngApp)*, 1(1), 37-45.
- PlagScan Plagiarism Checker. (2016). Retrieved from https://api.plagscan.com/PlagScan_Moodle_Manual-Admin_EN.pdf
- PlagScan. (2016). Retrieved from <http://www.plagscan.com>
- Srednja.hr. (2016). Retrieved from <http://m.srednja.hr/Studenti/Vijesti/U-Hrvatsku-stigao-lovac-na-plagijate-Nema-vise-prepisivanja-seminarskih-diplomskih-i-doktorskih-radova>
- Turnitin. (2016). Retrieved from <http://turnitin.com>
- Urkunde. (2016). Retrieved from <http://www.orkunde.com/en>
- VeriCite. (2016). Retrieved from <https://www.vericite.com>