TAXONOMY OF ACADEMIC PLAGIARISM METHODS

Tedo Vrbanec

MSc, Senior Lecturer, Faculty of Teacher Education, University of Zagreb, Savska cesta 77, 10000 Zagreb, Croatia; e-mail: tedo.vrbanec@ufzg.hr

Ana Meštrović

PhD, Associate Professor, Department of Informatics, University of Rijeka, Radmile Matejčić 2, 51000 Rijeka, Croatia; e-mail: amestrovic@inf.uniri.hr

ABSTRACT

The article gives an overview of the plagiarism domain, with focus on academic plagiarism. The article defines plagiarism, explains the origin of the term, as well as plagiarism related terms. It identifies the extent of the plagiarism domain and then focuses on the plagiarism subdomain of text documents, for which it gives an overview of current classifications and taxonomies and then proposes a more comprehensive classification according to several criteria: their origin and purpose, technical implementation, consequence, complexity of detection and according to the number of linguistic sources. The article suggests the new classification of academic plagiarism, describes sorts and methods of plagiarism, types and categories, approaches and phases of plagiarism detection, the classification of methods and algorithms for plagiarism detection. The title of the article explicitly targets the academic community, but it is sufficiently general and interdisciplinary, so it can be useful for many other professionals like software developers, linguists and librarians.

Key words: *plagiarism methods, plagiarism classification, plagiarism detection, plagiarism cases*

1. INTRODUCTION

The probability that two people with no mutual influence write an identical non-trivial text or produce the identical non-trivial work is very small, and some studies such as Alzahrani et al. demonstrate that it is even impossible (S. M. Alzahrani et al., 2012). The publication of other people's thoughts, words, or deeds, without a clear indication of the source, is called plagiarism. Plagiarism is prohibited in all countries by law (Kumar & Tripathi, 2013).

Plagiarism (Lat. *plagiare* = to steal; Lat. *plagere* = to kidnap; Lat. *plagiarius* = kidnapper, seducer, plunderer) is the partial or complete taking of someone else's intellectual or artistic work, without the clear indication of their authorship. There is a universal consensus on the definition of plagiarism. With very minor variations of the expression, most researchers (M. Y. M. Chong, 2013; Culwin & Lancaster, 2001a, 2001b; Kumar & Tripathi, 2013; Lancaster, 2003; Lukashenko

et al., 2007; Zu Eissen & Stein, 2006) use the definition found in Merriam-Webster's Dictionary, which defines plagiarism as the act of using someone else's words or ideas without giving credit to original author (Merriam-Webster Dictionary, 2016). Encyclopedia Britannica defines plagiarism as the act of taking the writings of another person and passing them off as one's own (Encyclopedia Britannica, 2018). The Cambridge University Press, Oxford Dictionary and University of Oxford define plagiarism as the use of the ideas or works of other people and pretending that they are one's own (Cambridge University Press, 2018; Oxford Dictionary, 2018; University of Oxford & Encyclopedia Britannica, 2018). Meuschke and Gipp define academic plagiarism as the taking of someone else's ideas or expressions without giving recognition to the original authors or sources according to academic principles (Meuschke & Gipp, 2013). Plagiarism.org considers plagiarism and plagiary to be fraud which includes the theft of someone else's work and the subsequent lying about that theft (Plagiarism.org, 2017). Williams considers plagiarism as "a form of cheating and is generally regarded as being morally and ethically unacceptable" (Williams, 2005:3). Probably the most prevalent statement/definition that expresses the core of the problem was given by Bouville: "Plagiarism is a crime against academia" (Bouville, 2008:311–322).

Plagiarism is a problem that has been growing steadily for decades and occurs at all academic levels. Teachers and professors struggle with students who do not know or do not care about norms forbidding and sanctioning plagiarism. Journal editors and reviewers want to preserve and improve the reputation of their journals. Mentors and universities take care of their reputations when accepting theses of their PhD students. These are just a few examples where detection of academic plagiarism is of crucial importance.

Academic plagiarism is the most common object of plagiarism during education and in academic articles. Academic plagiarism refers to plagiarizing (the whole or part of) several types of documents: source code, seminars, critical reviews, professional and scientific articles and non-literal books. Word academic indicates that this type of plagiarism most frequently appears in the academic community. This also means that in the academic context plagiarism is a particularly worrying phenomenon present at all academic levels.

Plagiarism is associated with several *related terms:* forgery, design piracy, brand piracy, replica and copyright infringement. The first three are termed as industrial plagiarism. *Forgery* or imitation is a product that is presented as an original, so the forger makes efforts to convince buyers that they are selling the original product. *Design piracy* is the marketing concept used by manufacturers to capitalize on the customer interest for a product by designing a product that resembles a well-known brand. *Brand piracy* is a situation where a manufacturer cannot protect their name and products in particular country because someone did it before them and with whom it is necessary to reach a financial agreement. *Replica* is a new production of a product from the original manufacturer or the owner of the production and selling rights. *Copyright infringement* is the intensive use of someone's work without permission, with or without the acknowledgment of another author (Aktion Plagiarius, 2018). The economic consequences of industrial plagiarism are severe, and some estimates indicate that 10% of world trade is industrial plagiarism that brings a loss of 200 - 300 billion euros and 200 thousand jobs per year (Aktion Plagiarius, 2018).

Our research was conducted in three distinct phases. The first phase includes a database and journal research papers relevant to the domain of academic plagiarism. The second phase includes analysis of selected research papers and identification of current classifications and taxonomies. The final phase included systematisation of the collected data, classification of academic plagiarism, and related discussion about approaches and phases of plagiarism detection, the classification of methods and algorithms for plagiarism detection.

During the first phase, we performed a search to find research papers addressing the topic of plagiarism in the academic community, plagiarism classification and plagiarism identification methods. Research articles were searched for in relevant databases: Scopus, Web of Science and EBSCO. In order to further extend the search, several selected databases were also included in the search: ScienceDirect database by Elsevier, IEEE Xplore Digital Library and ACM Digital Library. The database search was performed using the following search keywords: "plagiarism", "academic plagiarism", "plagiarism methods", "plagiarism classification", "plagiarism identification". Journals on the other hand were examined by title and abstract, each issue separately.

After the introduction in the first section, the second section describes methods of plagiarizing text. In the third section we give an overview of the current classifications and a new one is proposed, which is implemented according to several criteria. The fourth section is dedicated to the approaches, phases, strategies, methods and algorithms for plagiarism detection. The fifth section contains the final and concluding thoughts.

2. PLAGIARISM METHODS

According to Alzahrani et al., unless the original sources are cited correctly, plagiarised parts can arise from paraphrasing, summarising of an original text, combining, reconstruction, generalisation or specification of concepts (S. M. Alzahrani et al., 2012).

Maurer et al. recognise nine methods of plagiarism (Maurer et al., 2006): copy-paste – the verbatim copying of a text; the plagiarism of ideas – the use of similar concepts and thoughts that are not commonly recognised; paraphrasing – grammatical amendments, the use of synonyms, change of word order in a sentence, the use of other words and expressions for the same thoughts; artistic plagiarism – the use of other media for fundamentally the same work; the plagiarism of code – the use of source code, algorithms, classes or functions without licences or references; the lack of links to sources – the existence of quotation marks, but insufficient information about the source, links which are no longer valid; incorrect/imprecise use of quotation marks; disinformation of references – a reference points to a wrong or non-existent source; plagiarism by translation – a translation without reference.

TurnitIn (one of worldwide plagiarism detection software) developers, distinguishes (a) methods of academic plagiarism and (b) the plagiarism of research papers methods (Turnitin Europe, 2016). In the *methods of academic plagiarism,* it lists the following: the submission of someone else's work as one's own, to fulfil a specified teaching obligation; the copying of words or ideas without giving credit to the original author; copying most of the words or ideas that compromises the work; submitting an already submitted work (e.g. from another colleague); not using quotation

marks when quoting; giving incorrect data about sources; the use of someone else's sentences by using substitute words; using someone else's ideas without referencing. The same source (Turnitin Europe, 2016, pt. 1, p. 5) lists the following *methods of research plagiarism:* "claiming authorship on a paper or research that is not one's own; citing sources that were not actually referenced or used; reusing previous research or papers without proper attribution; paraphrasing another's work and presenting it as one's own; repeating data or text from a similar study with similar methodology without proper attribution; submitting a research paper to multiple publications; failing to cite or acknowledge the collaborative nature of a paper or study".

Deep learning is a form of machine learning which solves the problem in unsupervised and simultaneous representative learning by enabling computer-building of complex concepts from simple ones (Goodfellow et al., 2016). It dominates in new research of unsupervised machine learning and has proven to be very effective in solving problems in the field of computer assisted natural language analysis, as it creates very high-quality vector representation of words, so both the syntactic and the semantic similarities of texts can be measured. Since deep learning models generate vector space representation of words and sentences with built-in semantic meanings (Zhang et al., 2018), vectors can be used to generate alternated texts by choosing similar sentences and/or words.

3.   PLAGIARISM CLASSIFICATION

Many authors distinguish a lot of plagiarism types, but there is a great distinction in the depth and width of approach.

3. 1   Related work

One of the first academic plagiarism classifications was made by Martin, who distinguishes verbatim copying, paraphrasing, plagiarism from secondary sources, paper structure plagiarism, plagiarism of ideas and plagiarism of authorship (Martin, 1994).

Park lists five types: collusion (one author claims the credit of a group), commission (the agreed submission of someone else's work), duplication (the same paper in two different contexts), copying/paraphrasing, and submission (someone else's work without the knowledge of the original author) (Park, 2004).

Maurer *et al.* separates plagiarism into categories depending on the intentions of the plagiarists: accidental, unintentional, intentional and self-plagiarism (Maurer et al., 2006).

Schwarzenegger and Wohlers distinguish seven types of plagiarism: complete plagiarism, plagiarism by translation, copy/paste plagiarism, paraphrasing, self-plagiarism, ghostwriter and quoting out of context (Schwarzenegger & Wohlers, 2006).

Roig distinguishes two basic types of academic plagiarism: plagiarism of ideas and plagiarism of text (Roig, 2006). However, the latter is further analysed in great detail as follows: plagiarism verbatim, mosaic (patchwriting and paraphragiarism), inappropriate paraphrasing, paraphrasing

and summarising (of others' work), self-plagiarism, duplicate and redundant publication, data augmentation or fragmentation, inappropriate manipulation of references, citation stuffing, citing sources that were not read or thoroughly understood, reduced recognition of borrowing, selective reporting of literature, selective reporting of methodology, selective reporting of results and ghost authorship. Roig also specifies as many as 27 detailed guidelines to avoid plagiarism.
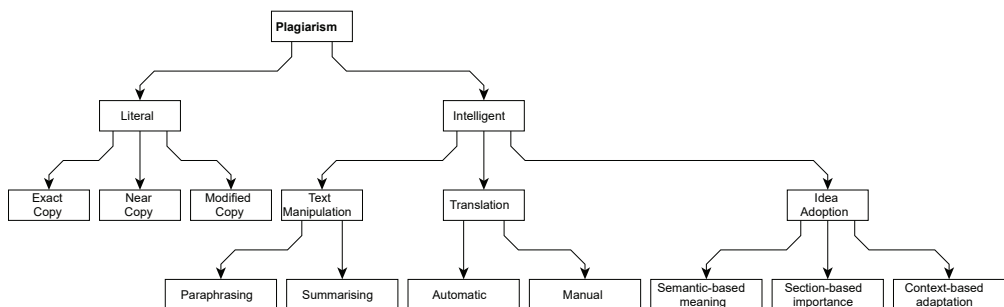
Joy *et al.* consider the plagiarism taxonomy within four mutually complementary aspects: the plagiarism source, the plagiarism method, the plagiarism object and the extrinsic aspect of plagiarism (Joy et al., 2009). The result is a taxonomy with six categories of plagiarism (plagiarism and copying, referencing, deception and inappropriate collaboration, ethics and consequences, source code plagiarism, plagiarism of the source code documentation) and their 23 sub-categories.

Kakkonen and Mozgovoy offered a quite different classification: verbatim copying, plagiarism by paraphrasing, technically disguised plagiarism, deliberate incorrect use of literature and heavy plagiarism, wherein the last category includes a) the use of someone else's ideas, concepts, thoughts; b) translation, c) ghostwriter and d) artistic plagiarism (Kakkonen & Mozgovoy, 2010).

Alzahrani *et al.* propose a plagiarism taxonomy shown in Fig. 1 (Alzahrani et al., 2012). The basis of their taxonomy is the behaviour of the author during plagiarism, in other words the plagiarism method. Based on the plagiarist's behaviour, Alzahrani et al. distinguishe plagiarism between literal and intelligent plagiarism. The literal is simpler and is further divided into three stages of copying. Intelligent plagiarism is a serious academic dishonesty where plagiarists try to hide, obfuscate, and change the original work in various intelligent ways, including text manipulation, translation, and idea adoption.

In two studies of plagiarism detection supporting tools published by very similar group of authors (Foltýnek et al., 2020; Weber-Wulff et al., 2013), they used to distinguish three types of plagiarism: verbatim copying, paraphrasing, and applying technical tricks. The authors focused more on plagiarism techniques.
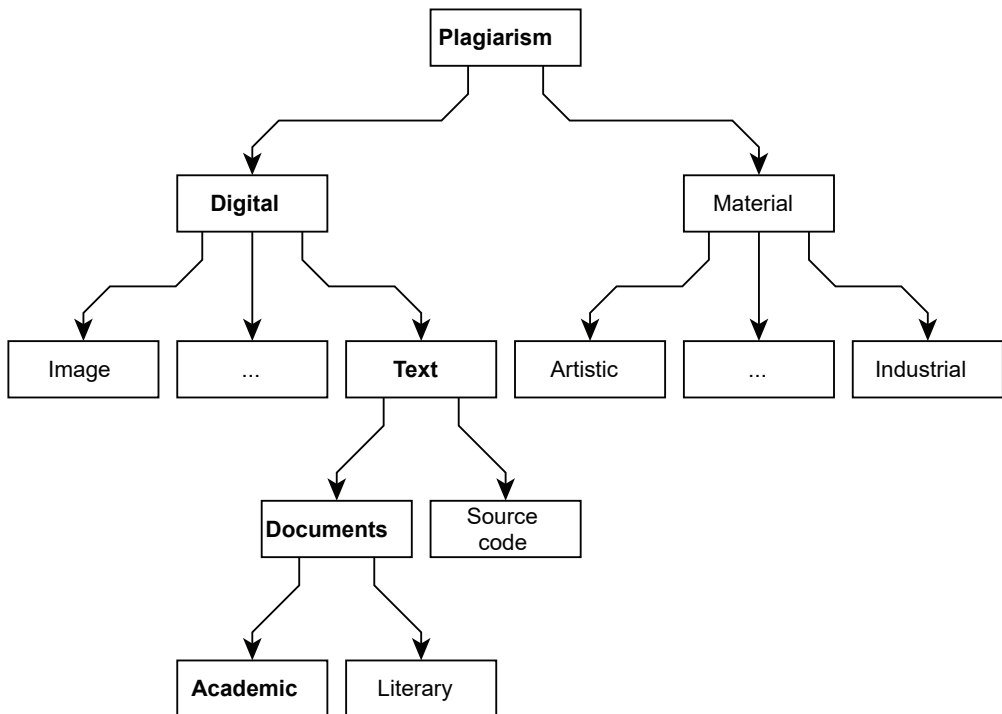
Graph 1. Taxonomy of Plagiarism

Source: Alzahrani et al. (2012)

3. 2   Proposed classification

In this section, we propose a new, comprehensive classification of plagiarism. In the widest sense, plagiarism can be categorized into types according to several criteria. According to the source and intention (Fig. 2), the plagiarism objects can be material (industrial, artistic) and non-material. Those non-material could be originally in the digital form (texts, source code and alike) or can be digitalised (artistic paintings, songs and alike). Text plagiarism can be divided into academic and literary (Meuschke & Gipp, 2013). Literary plagiarism causes artistic and direct financial loss to the original author while academic plagiarism can cause academic and indirect financial loss. The systematic verification of digital text documents, i.e. their originality, and generally, the systematic struggle against plagiarism is dominantly carried out in and by academic circles (Vrbanec & Meštrović, 2017).

According to the criteria of the technical implementation of plagiarism, academic plagiarism can be divided into the following types (Beames, 2012; Juričić, 2012): *clone or complete plagiarism* – the insinuation of someone else's document as being one's own; *translation* – the translation of someone else's document from another language without quoting the authorship and the author's permission; *copy* – a document which contains a significant portion of text from one source, without significant changes; *substitute* – the keywords and expressions in a document have been changed, however the document has retained the initial meaning and content of the original document; *remix* – a document in which other documents have been paraphrased and put together in a way that they act as a conceived whole; *self-plagiarism* – the use of one's earlier documents without appropriate references; *hybrid* – a document in which correctly quoted parts and those copied are combined; *mashup* – the inconsistent mixture of documents of various sources without correct citation; *waste* – a document which includes citations from non-existent or incorrect sources; *aggregator* – a document in which the sources are correctly cited but contains no originality; *repetition* – a document which includes the corresponding citations but sticks too much to the text or structure of the source documents; *ghost-product* – a document which is the result of the service (most often paid) of some other author than the one who signed it.
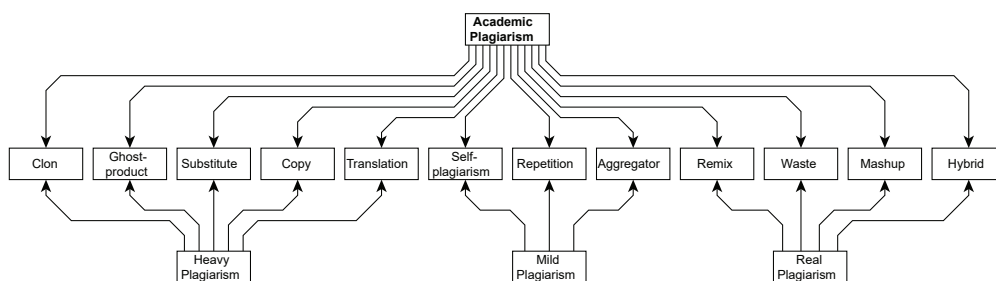
Graph 2. Objects of Plagiarism

Source: Authors

According to the potential severity of the consequences, the previous classification could be reduced to three categories: heavy, real and mild plagiarism. *Heavy plagiarism* includes clone, translation, copy, substitute, and ghost-product. Here both the intentions and potential damage from plagiarism are the greatest, and the plagiarist is the most ruthless or very naïve. *Real plagiarism* includes remix, hybrid, mashup, and waste. In the academic community such plagiarism is common, especially in meeting student obligations. It is difficult to distinguish the intention, ignorance, or naiveté of the author of plagiarism, and their detection is also difficult. *Mild plagiarism* includes self-plagiarism, aggregator and repetition. From a moral, ethical and legal standpoint, this category of plagiarism is the most benign, but it is certainly not allowed or justified.

Graph 3. Types of Academic Plagiarism

Source: Authors

As we can see in Fig. 3, these two classifications are not mutually completely independent. That is easier to realise if we introduce one more pragmatic criterion of classification: the potentiality of automatic detection, i.e. the plagiarism detection complexity criterion. So, within the classification by type, we can divide all types of plagiarism into those which are easy or difficult to detect, which result in the matrix shown in Table 1 and illustrated in Fig. 3. Easy or simple detection means that it is possible to detect them with automated software systems for plagiarism detection, whilst difficult or complex detection means that the plagiarism analysis of a human expert is needed.

Table 1. Category / complexity of detection matrix

Category	The complexity of detection	
	Simple	*Complex*
Heavy plagiarism	Clone, Copy, Substitute	Translation, Ghost-product
Real plagiarism	–	Remix, Hybrid, Mashup, Waste
Mild plagiarism	Self-plagiarism, Repetition	Aggregator

Source: Authors

According to linguistic origin, we can classify plagiarism as monolingual and plagiarism by translation. Plagiarism by translation can arise by the translation of documents from one or more languages, and the precondition for their automatic detection is usage of automatic translation software as a module for plagiarism detection software but this is barely satisfactory even for the main world languages and not for the rest of them.

4. PLAGIARISM DETECTION

In 1927 Charles Bird "*first researched the application of statistical methods in the detection of plagiarism with multiple choice answers*" (Chong, 2013). In the 1960s, the first methods were developed aiming to detect plagiarism of texts with multiple choice answers, whereas the first systems for plagiarism detection of written texts were developed in the 1970s for source code, and in the 1990s for natural languages (Chong, 2013).

The first papers about the plagiarism of texts and source code date from the 1970s (Alzahrani et al., 2012). They predominately dealt with the plagiarism detection in source code written in Pascal and C programming languages.

Twenty years later, papers appeared in which statistical computer methods of the detection of copying were presented in natural languages. In the 1990s scientists began to publish papers about the academic plagiarism and so Samuelson polemicized about the ethics and infringement of authors' rights in the case of self-plagiarism (Samuelson, 1994).

At the turn of the millennium, authors mainly dealt with the problems of detecting plagiarism in closed systems within academic institutions and web plagiarism. Contemporary researchers are trying to (1) improve the existing systems to make them more efficient and effective, (2) use the semantic and stylistic similarities of documents and (3) find methods of extracting knowledge from them. In order to reduce the complexity of determining the measure of the semantic similarities of documents and to make the organisation of the information and knowledge contained within them easier, some contemporary researchers use ontologies (Harispe et al., 2014; Leroy & Rindflesch, 2005; Patwardhan et al., 2003), particularly when it concerns the problem of word-sense disambiguation (WSD).

4. 1 Classifications

Lancaster and Culwin classified the approaches to plagiarism detection according to five criteria (Lancaster & Culwin, 2005):

Traditional classification

In traditional classification documents are calculated either by attributes (Attribute Counting Systems) or by structures (Structure Metric Systems). Lancaster and Culwin considers such classification incomplete, because some systems have an approach which does not belong to either of the two classes.

Classification according to the type of corpus which is processed

According to the *types of documents* which are being processed, the corpus of documents can be formed by source code, text documents or both. According to the *source of documents*, the corpora can be internal (documents available to an organisation), external (all Internet sources) or mixed. According to the *method of work*, access can be with or without tokenisation.

Classification according to the availability of the plagiarism detection system

According to the *setting*, they can be local or on the web. According to the *openness*, they can be public or private.

Classification according to the number of documents which are simultaneously processed

A metric is used which can be singular, paired, and corporal (n - dimensional; n = number of documents in a corpus).

Classification according to the complexity of the metrics used

Metrics can be superficial or structural.

According to Maurer *et al.*, the strategy of plagiarism detection should be carried out in three phases (authors call them methods) (Maurer et al., 2006):

1. The use of local documents' repository, i.e. the comparison of a verified document word-by-word with potential sources of plagiarism.

2. The comparison of a verified document with all available web sources in a way that the characteristic parts or sentences are compared, not the whole documents.

3. The use of stylometry i.e. an algorithm for a linguistic analysis that compares the style of sequential sections of an observed document and draws attention to the inconsistency and change of style, which indicates the increased probability of plagiarism.

Culwin and Lancaster identify a four-phase model for the plagiarism detection (Culwin & Lancaster, 2001a): (1) a collection phase in which documents fill the repository of all relevant documents, (2) a detection phase in which a software system recognises the suspicious pairs of documents, (3) a confirmation phase in which a human expert confirms or rejects doubt about plagiarism and (4) an investigation phase in which a human expert confirms the plagiarism and determines sanctions for the plagiarists.

Williams states three strategies, which we could call an evolutionary approach in the anti-plagiarism efforts (Williams, 2005:5): "Various strategies can be employed by academics to police plagiarism, ranging from simple Web search techniques used by individual lecturers, to the employment of easy-to-use freeware capable of tracking plagiarism between cohorts of students, as well as to quite elaborate systemic approaches involving the engagement of commercial plagiarism detection agencies."

4. 2 Methods

In the processes of plagiarism detection, the plagiarism detection methods and algorithms are key elements. An ideal algorithm for the plagiarism detection should be able to determine (Kakkonen & Mozgovoy, 2010):

1. Verbatim copying of initially digital documents and digitalised analogue sources.

2. Paraphrasing in the forms of the addition or removal of words or letters, the addition of intentional spelling and grammatical errors, substitution of words with synonyms, changing of word order in sentences or expressions, and changes in grammar or style.

3. The detection of technical tricks which attempt to exploit the weaknesses of existing automatic systems for the plagiarism detection, such as the use of fonts which are similar in appearance, but are different by code, the use of white letters in place of spaces to confuse plagiarism detection software, and the use of images of text instead of text, etc.

4. Intentionally incorrect referencing in a form of wrong or inaccurate marking of quotation marks, deliberately inaccurate or non-existing references, and use of outdated links to sources.

5. Heavy plagiarism is the plagiarism of ideas (similar concepts or thinking beyond that generally known, without correct referencing), plagiarism of translated text (translation without the acknowledgment of the original author), the use of the text of a ghost-writer, and artistic plagiarism (someone else's work in another medium).

With such an ideal algorithm, we are getting closer to the development of existing and new methods, algorithms and methodologies.

Today we can classify the developed methods into two classes: *external* (extrinsic) and *internal* (intrinsic), whether the evidence for plagiarism is sought by comparing potential plagiarism with a potential original or whether it is sought within the document itself (Chong et al., 2010).

Lukashenko *et al.* distinguish two classes of methods: *methods for prevention* which are time-consuming but have long-term effects and *methods for detection* which are short-term and have rapid effects (Lukashenko et al., 2007). According to same authors, (p. 1), methods of prevention are "precautions with which the goal is to prevent the development of illness." They do not act as rapidly as the methods of plagiarism detection; however, their effect is long-term, and therefore very desirable. Williams supports the attitude that the main course of prevention is the assigning of innovative and interesting tasks and that in addition to prevention there must also be deterrence, which discourages attempts at plagiarism due to unprofitability and the potential penalties (Williams, 2005).

Methods of prevention include the propagation of a policy of honesty and integrity which strives to influence the awareness of the whole of society, more precisely, conscientiousness, morality, ethics, attitude and so on..., the education of all the people or stakeholders of the system. Considering that it is difficult to influence the whole society without a great political agenda, it is necessary to influence the very important organised sections: the science, higher and secondary education so that they systematically promote the values of so-called academic integrity. An adequate system of penalisation i.e. the adoption of regulations and penalties for their violation on a social or systemic level must follow methods of detection. These two methods act as prevention and treatment. According to Turnitin Europe, a method of the prevention of academic plagiarism should include (Turnitin Europe, 2016) the education of students by professors; the adoption, open disclosure, and promotion of a policy of academic integrity; a developed system of penalisation proportional to the degree of plagiarism, the consequences and the intention of plagiarists; systematic raise of awareness among the students through discussions and within the syllabus of individual courses; teachers should help students with examples of proper referencing and should have plagiarism in mind during the creation of tasks; the use of plagiarism detection software, with the free usage for students so that they themselves would be able to practice.

Statistical methods do not strive to "understand" the document. These methods do not always strictly extract statistical values from the documents. In addition to the frequency of words, they also calculated their weighted values. In the statistical values, some authors include various

measures of distance (Li et al., 2004): the Hamming distance, the Euclidean distance, the Lempel-Ziv distance, compression distance, information distance and normalised information distance. According to our experience, K-character statistics is effective too; for example, 2-character reliably identifies the language in which the document is written, 3-character classifies the document. Statistical methods are often the components of other methods.

Methods of the copying detection include algorithms that can be divided into four subcategories (Aho, 2014; Michailidis & Margaritis, 2001; Stein, 2007; Stein & Zu Eissen, 2006; Stephen, 1992).

- Classical algorithms or algorithms for the comparison of character strings are numerous e.g. Brute-Force (Naive), Knuth-Morris-Pratt, Boyer-Moore, Boyer-Moore-Smith, Boyer-Moore-Horspool, Boyer-Moore-Horspool-Raita, Simon, Colussi, Galil, Apostolico-Giancarlo, Turbo-BM, Reverse Colussi, Sunday algorithms (Quick Search, Optimal Mismatch, Maximal Shift) and Ratcliff/Obershelp. Some of the algorithms can search text similarity of several sources, e.g. Commentz-Walter, Hume, Baeza-Yates.

- Suffix automation algorithms are Reverse Factor, Turbo Reverse Factor, Suffix Tree and the Aho-Corasick algorithm.

- Bit-parallelism algorithms are the Shift-Or algorithm, Shift-And and BNDM.

- Examples of algorithms and methods of using summaries are the algorithms Harrison, Karp-Rabin, Running Karp-Rabin Greedy String Tiling, Las Vegas, Monte Carlo, winnowing (Schleimer et al., 2003), Wu-Manber's algorithm for multiple samples (Wu & Manber, 1994), the method of chunking (Stein & Zu Eissen, 2006) etc. They use a cryptographic hash function such as MD5 to obtain summaries from small or large parts of a text. The sensitivity of the algorithm determines the size of the text. Even the slightest alteration of the text changes the summary. Similarity matrices are created from the summaries of the two documents which are compared (Stein, 2007). These methods are demanding according to the necessary computing resources (Stein & Zu Eissen, 2006).

The methods of detecting paraphrasing and semantic similarities are two groups of related methods, and they are here together because detecting paraphrasing has the consequence of detecting semantic similarity, while detection of semantic similarity reveals paraphrasing. The detection of semantic similarities is a threefold problem (Chong, 2013; Ram et al., 2014): the detection of lexical changes, changes of the text structure and the most complex of them – the detection of paraphrasing. Examples of these methods are: Natural Language Processing – NLP methods (Chong et al., 2010; Chong, 2013), Morphological Analysis (Marsi & Krahmer, 2010), Syntactic Parsing – a method of comparing the meta-information of documents, methods of the automatic extraction of summaries (Aliguliyev, 2009; Das & Martins, 2007; Spärck Jones, 2007), Keyword Similarity (Stein & Zu Eissen, 2006), a method of tokenisation (Lujo, 2010), and Deep Learning methods (Le & Mikolov, 2014; Mikolov et al., 2013; Pennington et al., 2014) that create a vector space of words, sentences, or phrases with embedded semantic meaning. Alzahrani and Salim use fuzzy semantic similarity (Alzahrani & Salim, 2010), whereas Hsiao *et al.* use "fuzzy strengths as a function of the semantic proximity between two objects" (Hsiao et al., 2014, p. 2), given that plagiarism is not

always completely obvious to determine (Chong, 2013, p. 1). The semantic similarity of a text is defined according to the following criteria: a different vocabulary, changes of vocabulary within the same text, incoherence of text, identicalness of punctuation, amount of similarity among the texts, same spelling mistakes, equal statistical distribution of words, same syntax, equally long sentences, same sequence of themes, consistent use of the same phrases and expressions, frequency of words, preferences in using short or long sentences, readability of text, references which are missing in the list of literature (Clough, 2000). In the field of the semantic similarities of texts (Harispe et al., 2014; Marsi & Krahmer, 2010; Zervanou et al., 2014), artificial intelligence methods are being intensively developed, natural language processing methods, data mining, methods of stylometric analysis of text, methods of extraction and presentation of knowledge and meaning from documents (Jakupović et al., 2013; Rauker Koch et al., 2014; Pavlić, Jakupović, et al., 2013; Pavlić, Meštrović, et al., 2013; Rajagopal et al., 2013), data and natural languages (graphical methods of the presentation of knowledge such as BG (Basic Conceptual Graphs) and NOK (Nodes of Knowledge), data models, semantic networks, neural networks, MultiNets method, HSF method for the representation of examples in natural languages). Stylometric methods (Zu Eissen & Stein, 2006) have become so reliable that the legislations of the USA, UK and Australia acknowledge the analyses carried out (Brennan & Greenstadt, 2009).

There are still no satisfactory solutions for finding obfuscated plagiarism. Promising research directions most often are quite demanding in terms of the required computing resources. Nowadays, these are Machine Learning, Deep Learning, and usage of high dimensional vector space. Therefore, when they can, researchers use *heuristics* as well (Dolan et al., 2004; El Bachir Menai & Bagais, 2011; Ganitkevitch et al., 2013; HaCohen-Kerner & Tayeb, 2017).

The methods and algorithms classification which can be used for the plagiarism detection is not unambiguous, because some of them use elements which could belong to several classes.

5. CONCLUSION

Plagiarism is a very dangerous and persistent phenomenon that is presented from different perspectives: history, development, theoretical classification, ways of creation and discovery, with an emphasis on the academic type of plagiarism. This phenomenon must be brought under control, and this can be done by automatically detecting it with the proper software. Discovering academic plagiarism is a task whose complexity varies from trivial to extremely complex, especially since there are many types and methods of their creation as well as combinations thereof. Today, increasingly reliable and efficient plagiarism detection methods, methods and algorithms are being developed. There is also software to detect it, but it is often powerless to detect complex types of plagiarism. In addition, this existing software often does not guarantee confidentiality or has restrictions on the number of documents submitted or the number of words in the document to be checked and they carry high costs. The most plagiarism detection software, once open and free, now are commercial or abandoned. And despite all of development, we still does not have the ability to effectively and reliably detect plagiarism. Yet, the plagiarism detection problem slowly converges into the category of solvable, computer-supported problems. There is a lot of potential for developing effective plagiarism detection software, based on open access to scientific databases

and deep learning models. In this sense it is very important for the effective deal with the academic plagiarism, international institutional support (political and financial) in the creation of open access databases into which scholars, researchers and scientists could freely upload their papers after primary being published. All scientific papers in those databases should be freely accessible to the overall interested public. These databases could be (a) the foundation for dissemination of knowledge and new scientific insights and discoveries, and (b) the source of papers (reference corpus) for (today's and future) plagiarism detection software. That should be strategic priority of academia and it should develop as freely and publicly available service.

REFERENCES

Aho, A. V. (2014) Algorithms for finding patterns in strings. Algorithms and Complexity, 1, 255. http://theory.snu.ac.kr/mediawiki/images/a/a2/AHO.pdf

Aktion Plagiarius. (2018) Innovation vs. Imitation. https://www.plagiarius.com/index.php?ID=39

Aliguliyev, R. M. (2009) A new sentence similarity measure and sentence based extractive technique for automatic text summarization. Expert Systems with Applications, 36(4), 7764–7772. https://doi.org/10.1016/j.eswa.2008.11.022

Alzahrani, S. M., Salim, N., Abraham, A. (2012) Understanding Plagiarism Linguistic Patterns, Textual Features, and Detection Methods. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 42(2), 133–149. https://doi.org/10.1109/TSMCC.2011.2134847

Alzahrani, S., Salim, N. (2010) Fuzzy semantic-based string similarity for extrinsic plagiarism detection: Lab report for PAN at CLEF 2010. CEUR Workshop Proceedings, 1176. http://ims-sites.dei.unipd.it/documents/71612/86374/CLEF2010wn-PAN-AlzahraniEt2010.pdf

Beames, S. (2012) White Paper - The Plagiarism Spectrum: Instructor Insights into the Ten Types of Plagiarism. 18. http://www.ed.ac.uk/files/atoms/files/10-types-of-plagiarism.pdf

Bouville, M. (2008) Plagiarism: Words and Ideas. Science and Engineering Ethics, 14(3), 311–322. https://doi.org/10.1007/s11948-008-9057-6

Brennan, M. R., & Greenstadt, R. (2009) Practical Attacks Against Authorship Recognition Techniques. http://www.cs.drexel.edu/~mb553/stuff/brennan_iaai09.pdf

Cambridge University Press. (2018) Meaning of "plagiarize" in the English Dictionary. http://dictionary.cambridge.org/dictionary/english/plagiarize?q=plagiarism

Chong, M., Specia, L., Mitkov, R. (2010) Using natural language processing for automatic detection of plagiarism. Proceedings of the 4th International Plagiarism Conference (IPC 2010), Newcastle, UK. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.458.9440&rep=rep1&type=pdf

Chong, M. Y. M. (2013) A study on plagiarism detection and plagiarism direction identification using natural language processing techniques [University of Wolverhampton]. http://wlv.openrepository.com/wlv/handle/2436/298219

Clough, P. (2000) Plagiarism in natural and programming languages: an overview of current tools and technologies. Research Memoranda: CS-00-05, Department of Computer Science, University of Sheffield, UK, 1–31.

Culwin, F, & Lancaster, T. (2001a) Plagiarism, prevention, deterrence & detection. Available for ILT Members From.

Culwin, F., Lancaster, T. (2001b) Plagiarism issues for higher education. VINE, 31(2), 36–41. https://doi.org/10.1108/03055720010804005

Das, D., Martins, A. F. A. F. (2007) A survey on automatic text summarization. Literature Survey for the Language and Statistics II Course at CMU, 4, 192–195. http://stuyresearch.googlecode.com/hg-history/132ed87460529c48ae57bc388ef1083ba07791a5/blake/resources/das-martins.07.pdf

Dolan, B., Quirk, C., Brockett, C. (2004) Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. Proceedings of the 20th International Conference on Computational Linguistics, 350-es. https://doi.org/10.3115/1220355.1220406

El Bachir Menai, M., Bagais, M. (2011) APlag: A plagiarism checker for Arabic texts. ICCSE 2011 - 6th International Conference on Computer Science and Education, Final Program and Proceedings, 1379–1383. https://doi.org/10.1109/ICCSE.2011.6028888

Encyclopedia Britannica. (2018) Plagiarism. http://www.britannica.com/topic/plagiarism

Foltýnek, T., Dlabolová, D., Anohina-Naumeca, A., Razı, S., Kravjar, J., Kamzola, L., Guerrero-Dib, J., Çelik, Ö., Weber-Wulff, D. (2020). Testing of Support Tools for Plagiarism Detection. ArXiv Preprint ArXiv:2002.04279. http://arxiv.org/abs/2002.04279

Ganitkevitch, J., Durme, B. Van, Callison-Burch, C. (2013) PPDB: The Paraphrase Database. Proceedings of NAACL-HLT 2013, 758–764. http://www.aclweb.org/anthology/N13-1092

Goodfellow, I., Bengio, Y., Courville, A. (2016) Deep learning. In Healthcare Informatics Research (Vol. 22, Issue 4). MIT Press. https://doi.org/10.1038/nature14539

HaCohen-Kerner, Y., Tayeb, A. (2017) Rapid detection of similar peer-reviewed scientific papers via constant number of randomized fingerprints. Information Processing & Management, 53(1), 70–86. https://doi.org/10.1016/j.ipm.2016.06.007

Harispe, S., Sánchez, D., Ranwez, S., Janaqi, S., Montmain, J. (2014) A framework for unifying ontology-based semantic similarity measures: A study in the biomedical domain. Journal of Biomedical Informatics, 48, 38–53. https://doi.org/10.1016/j.jbi.2013.11.006

Hsiao, D. K., Neuhold, E. J., Sacks-Davis, R. (2014) So far (schematically) yet so near (semantically). Interoperable Database Systems (DS-5): Proceedings of the IFIP WG2. 6 Database Semantics Conference on Interoperable Database Systems (DS-5) Lorne, Victoria, Australia, 16-20 November, 1992, 25, 283.

Jakupović, A., Pavlić, M., Meštrović, A., Jovanović, V. (2013) Comparison of the Nodes of Knowledge method with other graphical methods for knowledge representation. Information & Communication Technology Electronics & Microelectronics (MIPRO), 2013 36th International Convention On, 1004–1008.

Joy, M., Cosma, G., Sinclair, J., Yau, J. (2009) A taxonomy of plagiarism in computer science. Proceedings of EDULEARN09 Conference: 6th - 8th July 2009. http://eprints.dcs.warwick.ac.uk/84/1/joy_cosma_sinclair_yau_edulearn_09.pdf

Juričić, V. (2012). Detekcija plagijata u višejezičnom okruženju. University of Zagreb.

Kakkonen, T., Mozgovoy, M. (2010) Hermetic and Web Plagiarism Detection Systems for Student Essays - An Evaluation of the State-of-the-Art. Journal of Educational Computing Research, 42(2), 135–159. https://doi.org/10.2190/EC.42.2.a

Kumar, R., Tripathi, R. C. (2013) An Analysis of Automated Detection Techniques for Textual Similarity in Research Documents. International Journal of Advanced Science and Technology, 56, 99–110. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.365.5776&rep=rep1&type=pdf

Lancaster, T. (2003). Effective and efficient plagiarism detection. https://www.researchgate.net/profile/Thomas_Lancaster/publication/228729388_Effective_and_efficient_plagiarism_detection/links/0fcfd50f68dcf52345000000.pdf

Lancaster, T., Culwin, F. (2005) Classifications of plagiarism detection engines. Innovation in Teaching and Learning in Information and Computer Sciences, 4(2). https://doi.org/10.11120/ital.2005.04020006

Le, Q. V., Mikolov, T. (2014) Distributed Representations of Sentences and Documents. International Conference on Machine Learning, 14, 1188–1196. http://www.jmlr.org/proceedings/papers/v32/le14.pdf

Leroy, G., Rindflesch, T. C. (2005) Effects of information and machine learning algorithms on word sense disambiguation with small datasets. International Journal of Medical Informatics, 74(7–8), 573–585. https://doi.org/10.1016/j.ijmedinf.2005.03.013

Li, M., Chen, X., Li, X., Ma, B., Vitanyi, P. M. B. (2004) The Similarity Metric. IEEE Transactions on Information Theory, 50(12), 3250–3264. https://doi.org/10.1109/TIT.2004.838101

Lujo, R. (2010) Lociranje sličnih logičkih cjelina u tekstualnim dokumentima na hrvatskome jeziku [Fakultet elektrotehnike i računarstva - Zagreb]. https://bitbucket.org/trebor74hr/text-hr/src

Lukashenko, R., Graudina, V., Grundspenkis, J. (2007) Computer-based plagiarism detection methods and tools: an overview. 40. http://dl.acm.org/citation.cfm?id=1330642

Marsi, E., Krahmer, E. (2010) Automatic analysis of semantic similarity in comparable text through syntactic tree matching. Proceedings of the 23rd International Conference on Computational Linguistics, 752–760. http://dl.acm.org/citation.cfm?id=1873866

Martin, B. (1994) Plagiarism: a misplaced emphasis. Journal of Information Ethics, 3(2), 36–47. http://www.bmartin.cc/pubs/94jie.pdf

Maurer, H. A., Kappe, F., Zaka, B. (2006) Plagiarism - A Survey. Journal of Universal Computer Science, 12(8), 1050–1084. http://jucs.org/jucs_12_8/plagiarism_a_survey/jucs_12_08_1050_1084_maurer.pdf

Merriam-Webster Dictionary. (2016) Simple Definition of Plagiarism. http://www.merriam-webster.com/dictionary/plagiarism

Meuschke, N., Gipp, B. (2013) State-of-the-art in detecting academic plagiarism. International Journal for Educational Integrity, 9(1), 50–71. http://gipp.com/wp-content/papercite-data/pdf/meuschke13.pdf%5Cnhttp://www.ojs.unisa.edu.au/index.php/IJEI/article/view/847/

Michailidis, P. D., Margaritis, K. G. (2001) On-line string matching algorithms: Survey and experimental results. International Journal of Computer Mathematics, 76(4), 411–434. https://www.researchgate.net/profile/Konstantinos_G_Margaritis/publication/234025323_On-line_string_matching_algorithms_Survey_and_experimental_results/links/0fcfd50dae3bb4ef83000000.pdf

Mikolov, T., Chen, K., Corrado, G., Dean, J. (2013) Efficient estimation of word representations in vector space. ArXiv Preprint ArXiv:1301.3781. https://arxiv.org/abs/1301.3781

Oxford Dictionary. (2018) Definition of plagiarism in English. http://www.oxforddictionaries.com/definition/english/plagiarism

Park, C. (2004) Rebels without a clause: towards an institutional framework for dealing with plagiarism by students. Journal of Further and Higher Education, 28(3), 291–306. https://doi.org/10.1080/0309877042000241760

Patwardhan, S., Banerjee, S., Pedersen, T. (2003) Using measures of semantic relatedness for word sense disambiguation. In Computational linguistics and intelligent text processing (pp. 241–257). Springer. http://www.d.umn.edu/~tpederse/Pubs/cicling2003-3.pdf

Pavlić, M., Jakupović, A., Meštrović, A. (2013) Nodes of Knowledge Method for Knowledge Representation. Informatologia, 46(3), 206.

Pavlić, M., Meštrović, A., Jakupović, A. (2013) Graph-Based Formalisms for Knowledge Representation. Proceedings of the 17th World Multi-Conference on Systemics Cybernetics and Informatics (WMSCI 2013), 2, 200–204. http://www.iiis.org/CDs2013/CD2013SCI/SCI_2013/PapersPdf/GA426XX.pdf

Pennington, J., Socher, R., Manning, C. (2014) Glove: Global Vectors for Word Representation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 1532–1543. https://doi.org/10.3115/v1/D14-1162

Plagiarism.org. (2017) What is Plagiarism? https://www.plagiarism.org/article/what-is-plagiarism

Rajagopal, D., Cambria, E., Olsher, D., Kwok, K. (2013) A graph-based approach to commonsense concept extraction and semantic similarity detection. Proceedings of the 22nd International Conference on World Wide Web Companion, 565–570. http://dl.acm.org/citation.cfm?id=2487995

Ram, R. V. S. V. S., Stamatatos, E., Devi, S. L. L. (2014) Identification of Plagiarism Using Syntactic and Semantic Filters. In Computational Linguistics and Intelligent Text Processing (pp. 495–506). Springer.

Rauker Koch, M. R. R., Pavlić, M., Jakupović, A. (2014) Application of the NOK method in sentence modelling. Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2014 37th International Convention On, 1176–1181. http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6859746

Roig, M. (2006) Avoiding plagiarism, self-plagiarism, and other questionable writing practices: A guide to ethical writing. St John's University. https://graduateschool.vt.edu/sites/default/files/u12/academics/plagiarism.pdf

Samuelson, P. (1994) Self-plagiarism or fair use. Communications of the ACM, 37(8), 21–25. http://people.ischool.berkeley.edu/~pam/papers/SelfPlagiarism.pdf

Schleimer, S., Wilkerson, D. S. S., Aiken, A. (2003) Winnowing: local algorithms for document fingerprinting. Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data, 76–85. http://dl.acm.org/citation.cfm?id=872770

Schwarzenegger, C., Wohlers, W. (2006) Quellen zitieren, nicht plagiieren. Universität Zürich Unijournal 4/06, 16. http://www.kommunikation.uzh.ch/dam/jcr:00000000-086d-f41b-0000-00006b8d9335/unijournal-2006-4.pdf

Spärck Jones, K. (2007) Automatic summarising: The state of the art. Information Processing & Management, 43(6), 1449–1481. https://doi.org/10.1016/j.ipm.2007.03.009

Stein, B. (2007). Principles of hash-based text retrieval. Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 527–534. http://dl.acm.org/citation.cfm?id=1277832

Stein, B., Zu Eissen, S. M. M. (2006) Near similarity search and plagiarism analysis. From Data and Information Analysis to Knowledge Engineering, 430–437. https://www.researchgate.net/profile/Benno_Stein/publication/221649255_Near_Similarity_Search_and_Plagiarism_Analysis/links/004635141d122596aa000000.pdf

Stephen, G. A. (1992) String search. University College of North Wales. http://www.quretec.com/u/vilo/edu/2002-03/Tekstialgoritmid_I/Books/TR92gas01.us.ps

Turnitin Europe. (2016) Plagiarism in a Digital World series. Turnitin. http://go.turnitin.com/en/whitepaperLP1

University of Oxford & Encyclopedia Britannica. (2018) Plagiarism. University of Oxford. http://www.britannica.com/topic/plagiarism

Vrbanec, T., Meštrović, A. (2017) The struggle with academic plagiarism: Approaches based on semantic similarity. 2017 40th International Convention on Information and Communication Technology, Electronics and Microelectronics, MIPRO 2017 - Proceedings. https://doi.org/10.23919/MIPRO.2017.7973544

Weber-Wulff, D., Jannis, C. M., Elin Zincke, T. (2013) Plagiarism Detection Software Test 2013. https://plagiat.htw-berlin.de/software-en/test2013/report-2013/

Williams, J. B. (2005) Plagiarism: Deterrence, Detection and Prevention (p. 20). Universitas 21 Global. https://www.economicsnetwork.ac.uk/handbook/printable/plagiarism.pdf

Wu, S., Manber, U. (1994) A fast algorithm for multi-pattern searching. 11. http://webglimpse.net/pubs/TR94-17.pdf

Zervanou, K., Iosif, E., Potamianos, A. (2014) Word Semantic Similarity for Morphologically Rich Languages. LREC, 1642–1648. http://www.lrec-conf.org/proceedings/lrec2014/pdf/973_Paper.pdf

Zhang, O. R., Cohen, T., McGill, S. (2018) Did Gaius Julius Caesar Write De Bello Hispaniensi? A Computational Study of Latin Classics. Human IT, 14(1), 28. http://search.ebscohost.com/login.aspx?authtype=shib&custid=s4753785&groupid=knjiznica&profile=eds

Zu Eissen, S. M. M., Stein, B. (2006) Intrinsic plagiarism detection. In Advances in Information Retrieval (pp. 565–569). Springer. http://link.springer.com/10.1007%2F11735106_66

TAKSONOMIJA METODA AKADEMSKOG PLAGIRANJA

Tedo Vrbanec

Mr. sc., viši predavač, Učiteljski fakultet, Sveučilište u Zagrebu Savska cesta 77, 10 000 Zagreb, Hrvatska;
e-mail: tedo.vrbanec@ufzg.hr

Ana Meštrović

Dr. sc., izvanredna profesorica, Odjel za informatiku, sveučilište u Rijeci, radmile Matejčić 2, 51 000 Rijeka, Hrvatska; *e-mail:* amestrovic@inf.uniri.hr

SAŽETAK

Rad daje pregled domene plagiranja tekstnih dokumenata. Opisuje porijeklo pojma plagijata, daje prikaz definicija te objašnjava plagijatu srodne pojmove. Ukazuje na širinu domene plagiranja, a za tekstne dokumente daje pregled dosadašnjih taksonomija i predlaže sveobuhvatniju taksonomiju prema više kriterija: porijeklu i namjeni, tehničkoj provedbi plagiranja, posljedicama plagiranja, složenosti otkrivanja i (više)jezičnom porijeklu. Rad predlaže novu klasifikaciju akademskog plagiranja, prikazuje vrste i metode plagiranja, tipove i kategorije plagijata, pristupe i faze otkrivanja plagiranja. Potom opisuje klasifikaciju metoda i algoritama otkrivanja plagijata. Iako cilja na akademskog čitatelja, može biti od koristi u interdisciplinarnim područjima te razvijateljima softvera, lingvistima i knjižničarima.

***Ključne riječi**: metode plagiranja, klsifikacija plagijata, detekcija plagijata, slučajevi plagiranja*