ADDED VALUE OF CENTRALISED PLAGIARISM DETECTION SYSTEM ON A NATIONAL LEVEL

Julius Kravjar

Key words: Academic Ethics, Academic Integrity, Higher Education, Metadata, Plagiarism Detection, Theses

Introduction

The use of plagiarism detection (strictly speaking text-matching or detection of text similarities) by means of information and communication technology has become a standard for originality checks. Plagiarism detection software[1] is a good assistant that serves as a support for human decision-making process in plagiarism matters. Software for detection of text similarities does not detect plagiarism, it identifies similarities within the checked document with other sources and these similarities may represent plagiarism. The outputs of text similarity detection do not confirm whether the checked document is original or not, the final decision is made by authorities.

Text similarity detection contributes partially to plagiarism reduction, but text similarity detection alone is an insufficient measure for plagiarism removal or reduction. It is necessary to adopt academic integrity measures, including the academic integrity management system.

According to M. Bek (2018) "The main prerequisite for successful defense against plagiarism has always been and will be the quality work of supervisors with students working on their theses". J. Brandejsová (2018) says: "The essential responsibility lies with the supervisor, who is an expert in the field and is well versed in the relevant literature." Evering and Moorman (2012) consider that the most effective way of dealing with plagiarism is actively addressing issues through instruction and not by means of rules or codes. And they added:

> "The current emphasis on testing and grades has made educators and students alike lose track of the more important goals of schooling, such as lifelong learning and national and global citizenship. Refocusing on higher-order goals can persuade students that plagiarism and other forms of academic dishonesty are not in their long-term best interests."

Upbringing and education towards values in childhood should also continue during school years. Comprehensive curricula at all levels of education with an emphasis on values can significantly contribute to the shaping of the character of pupils and students and also to the culture of academic integrity and, consequently, to plagiarism reduction.

[1] The term "plagiarism detection software" is widely used, but more exact term is "text-matching software" or "text similarities detection software".

DOI: https://doi.org/10.11118/978-80-7509-772-9-0103

> “We must work to ensure that we are putting truth – and integrity – at the forefront of our mission and operations. Academic and research integrity cannot be a side project or an afterthought. Integrity and ethics must be central to everything we do and every decision we make.” (Bertram-Gallant, 2018)

Objectives

The objective is to point out the advantages of uniform metadata collection by a centralised text similarity detection system. In Slovakia, the Centralised Plagiarism Detection System is working closely with the Centralised Repository of Theses and Dissertations – both are in operation since April 2010. All Slovak higher education institutions (HEIs) are required to use this system according to Slovak law. The paper is focused on analytical possibilities of such a system based on uniform collection of theses and metadata.

Methodology

The cooperating systems, Centralised Repository of Theses and Dissertations and Centralised Plagiarism Detection System, are known under a single name SK ANTIPLAG. The Slovak Centre of Scientific and Technical Information (SCSTI) has operated both systems already for ten years. Five types of theses are collected. Today, a rich collection of theses (more than 0.6 million) and metadata are archived and they are used as a base for a wide spectrum of analytical insights useful for HEIs and the Ministry of Education. Several examples of simple and complex insights will be presented.

Uniform collection methodology of theses and metadata (UCM) ensures consistent metadata from all Slovak HEIs using the XML format, which is mandatory for exporting electronic versions of theses and metadata (batch mode) into the SK ANTIPLAG system.

Text similarity detection at higher education institutions

HEIs are free to decide which text similarity detection system will be used in their academic environment. This also applies to Slovak HEIs with one exception: one designated system is used on an obligatory basis according to the amendment to the Higher Education Act (2009). The implementation of SK ANTIPLAG is the first worldwide use of a centralised text similarity detection system, which cooperates with a centralised repository of theses and dissertations (both systems are developed in Slovakia). Before the launch of SK ANTIPLAG, only three HEIs used text similarity detection services. Within a year, all Slovak higher education institutions (public, private, state) started to use the SK ANTIPLAG system and it was a significant step forward. SCSTI is open to share its experience with the use of the SK ANTIPLAG system. The first delegation that wanted to know the Slovak experience with the system was a parliamentary and governmental delegation from Poland – they visited SCSTI already in 2011.

Since January 1st, 2019 Poland is the second country in the world that has implemented a centralised text similarity detection system named Jednolity System Antiplagiatowy (JSA), cooperating with the central repository Ogólnopolskie Repozytorium Pisemnych Prac Dyplomowych (ORPPD) – both systems were developed in Poland (jsa.org.pl). In Poland, text similarities detection was widely used already before JSA's operation.

The Slovak system checks the originality of five types of theses: bachelor's, master's, rigorous, doctoral and habilitation theses and the access to the theses is open to the general public at www.crzp.sk (in Slovak language). The Polish system checks the originality of bachelor's, master's and doctoral theses, and the access to these theses is open for thesis supervisors, research promoters and for the teaching staff, but not for public. Metadata related to theses are collected.

In both countries, HEIs do not pay for using centralised text similarity detection, or for licences, implementation, technical support and updates. All collected theses and metadata are stored in one centralised repository. The use of centralised systems is obligatory in Slovakia and in Poland due to amendments to the Higher Education Act.

In Slovenia, all major HEIs use the same system and in the near future it is expected that all HEIs will use this system, which was developed in Slovenia (Ojsteršek, 2018). In Czechia, there is a system used by about 50% of all Czech HEIs; the system was developed in Czechia (www.theses.cz). In Slovenia and Czechia, HEIs use text similarity detection systems on a voluntary basis.

In the literature, there have been several declarations that all HEIs in country use a text similarity detection system. However, a deeper analysis showed that it was not true (Kravjar, 2015).

The role of metadata

Metadata is the key and gate to analyses. If the theses originality check is not accompanied by metadata collection, an opportunity for deeper insights is missed. To name a few types of metadata: author, type of thesis, study field, thesis title, thesis subtitle, unique thesis identifier, language, abstract, key words, number of pages, year, supervisor, opponents, department, faculty, HEI, thesis downloadability, date and time of thesis registration in the central repository, similarity percentage, originality protocol creation date and time, date of thesis publication at www.crzp.sk.

The spectrum of analytical insights will be partially demonstrated on data from the nationwide SK ANTIPLAG system, which is mandatory for all Slovak higher education institutions operating under the Slovak law since April 2010. Many different analytical views on theses and dissertation are available, for example by supervisor, by thesis type, by faculty, by higher education institution, by type of higher education institution, by study field, their combination, etc. There are some analytical views that show a violation of academic integrity by academic staff. One may say that SK ANTIPLAG is not only a detector of text similarities but to some extent a detector of academic misconduct.

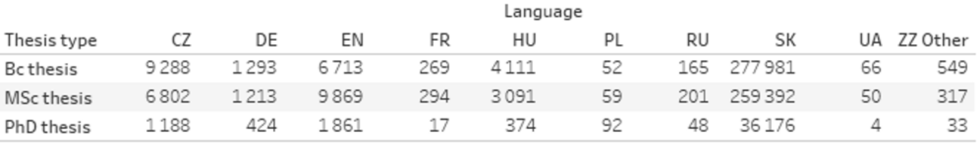
The power of metadata

Uniform metadata collection allows a wide range of insights. The following examples demonstrate their variability. These examples are far from being exhaustive.

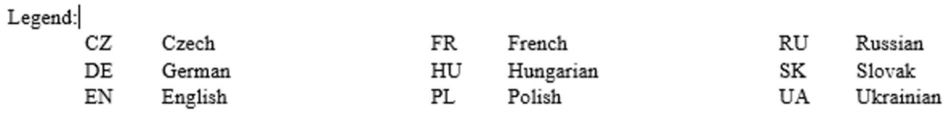
Naturally, Slovak language has the highest share among all languages (Table 1 and Table 2). The selected languages are the languages spoken by our nearest neighbours plus English, French and Russian languages. Language codes are described below the Table 1.

Table 1

Thesis type by language in absolute numbers for the period 2010–2019



					Language					
Thesis type	CZ	DE	EN	FR	HU	PL	RU	SK	UA	ZZ Other
Bc thesis	9 288	1 293	6 713	269	4 111	52	165	277 981	66	549
MSc thesis	6 802	1 213	9 869	294	3 091	59	201	259 392	50	317
PhD thesis	1 188	424	1 861	17	374	92	48	36 176	4	33

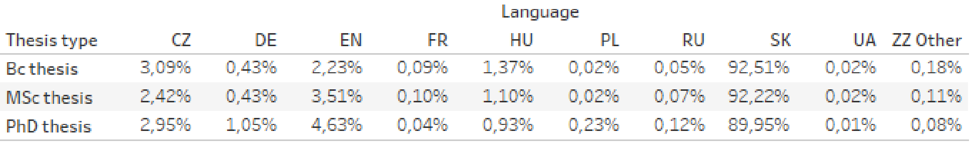


Legend:

CZ	Czech	FR	French	RU	Russian
DE	German	HU	Hungarian	SK	Slovak
EN	English	PL	Polish	UA	Ukrainian

Table 2

Thesis type by language (%) for the period 2010–2019

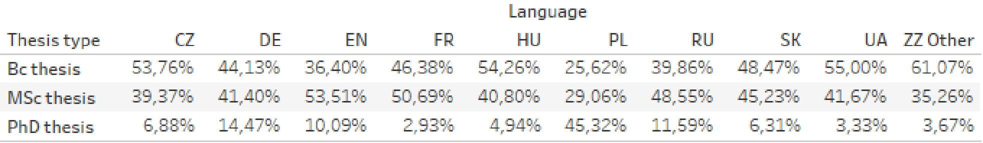


					Language					
Thesis type	CZ	DE	EN	FR	HU	PL	RU	SK	UA	ZZ Other
Bc thesis	3,09%	0,43%	2,23%	0,09%	1,37%	0,02%	0,05%	92,51%	0,02%	0,18%
MSc thesis	2,42%	0,43%	3,51%	0,10%	1,10%	0,02%	0,07%	92,22%	0,02%	0,11%
PhD thesis	2,95%	1,05%	4,63%	0,04%	0,93%	0,23%	0,12%	89,95%	0,01%	0,08%

Bachelor and master theses in Slovak language have approximately the same share (Table 3). In the similar situation are bachelor and master theses written in German, French and Polish languages.

Table 3

The share of thesis types in each language (%) for the period 2010–2019

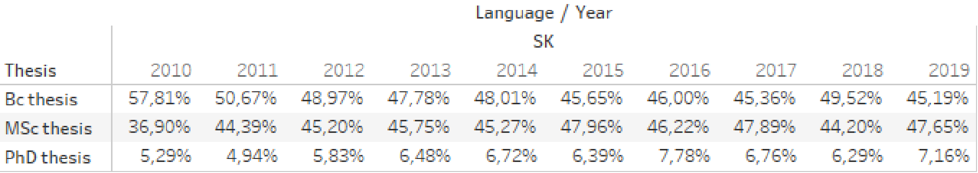


					Language					
Thesis type	CZ	DE	EN	FR	HU	PL	RU	SK	UA	ZZ Other
Bc thesis	53,76%	44,13%	36,40%	46,38%	54,26%	25,62%	39,86%	48,47%	55,00%	61,07%
MSc thesis	39,37%	41,40%	53,51%	50,69%	40,80%	29,06%	48,55%	45,23%	41,67%	35,26%
PhD thesis	6,88%	14,47%	10,09%	2,93%	4,94%	45,32%	11,59%	6,31%	3,33%	3,67%

Except for the years 2010 and 2011, the difference in the share of Slovak bachelor and master theses is really small and the share of master theses starts to surpass bachelor ones (Table 4).

Table 4

The share of thesis types in Slovak language per year (%)

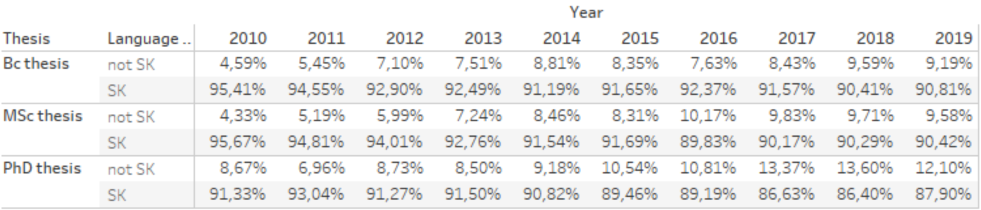


					Language / Year					
					SK					
Thesis	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019
Bc thesis	57,81%	50,67%	48,97%	47,78%	48,01%	45,65%	46,00%	45,36%	49,52%	45,19%
MSc thesis	36,90%	44,39%	45,20%	45,75%	45,27%	47,96%	46,22%	47,89%	44,20%	47,65%
PhD thesis	5,29%	4,94%	5,83%	6,48%	6,72%	6,39%	7,78%	6,76%	6,29%	7,16%

There is a mild growth in theses written in languages other than Slovak (Table 5). Between 2010 and 2019, the share of bachelor and master theses written in the other than Slovak language doubled. For PhD theses, the growth was about 50%.

Table 5

The share of Slovak language and other languages group per thesis type and per year (%)

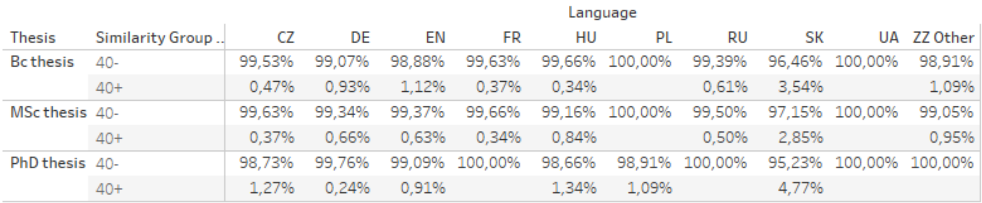


		Year									
Thesis	Language ..	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019
Bc thesis	not SK	4,59%	5,45%	7,10%	7,51%	8,81%	8,35%	7,63%	8,43%	9,59%	9,19%
	SK	95,41%	94,55%	92,90%	92,49%	91,19%	91,65%	92,37%	91,57%	90,41%	90,81%
MSc thesis	not SK	4,33%	5,19%	5,99%	7,24%	8,46%	8,31%	10,17%	9,83%	9,71%	9,58%
	SK	95,67%	94,81%	94,01%	92,76%	91,54%	91,69%	89,83%	90,17%	90,29%	90,42%
PhD thesis	not SK	8,67%	6,96%	8,73%	8,50%	9,18%	10,54%	10,81%	13,37%	13,60%	12,10%
	SK	91,33%	93,04%	91,27%	91,50%	90,82%	89,46%	89,19%	86,63%	86,40%	87,90%

In Table 6 the symbol 40+ means the share of theses with similarity greater than 40% and the symbol 40 – means the share of theses with similarity less than 40%. SK ANTIPLAG has a rich comparative corpus of Slovak documents, which is not the case for other languages. Therefore, the share of 40+ theses in Slovak language is the highest.

Table 6

Language by thesis types and similarity group (%)

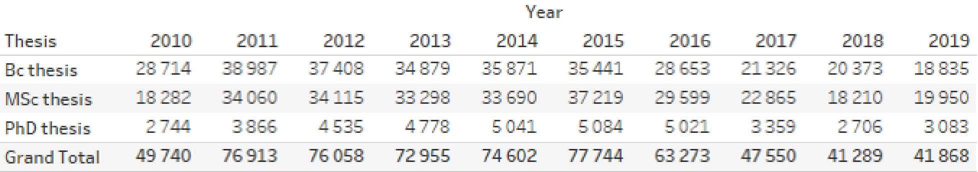


		Language									
Thesis	Similarity Group ..	CZ	DE	EN	FR	HU	PL	RU	SK	UA	ZZ Other
Bc thesis	40-	99,53%	99,07%	98,88%	99,63%	99,66%	100,00%	99,39%	96,46%	100,00%	98,91%
	40+	0,47%	0,93%	1,12%	0,37%	0,34%		0,61%	3,54%		1,09%
MSc thesis	40-	99,63%	99,34%	99,37%	99,66%	99,16%	100,00%	99,50%	97,15%	100,00%	99,05%
	40+	0,37%	0,66%	0,63%	0,34%	0,84%		0,50%	2,85%		0,95%
PhD thesis	40-	98,73%	99,76%	99,09%	100,00%	98,66%	98,91%	100,00%	95,23%	100,00%	100,00%
	40+	1,27%	0,24%	0,91%		1,34%	1,09%		4,77%		

The decrease in the number of thesis types is most likely caused by two main factors: the demographic development and the growing number of students studying abroad (Table 7). The centralised system started to work at the end of April 2010 and that is the reason why the numbers for 2010 are lower.

Table 7

Number of thesis types per year



	Year									
Thesis	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019
Bc thesis	28 714	38 987	37 408	34 879	35 871	35 441	28 653	21 326	20 373	18 835
MSc thesis	18 282	34 060	34 115	33 298	33 690	37 219	29 599	22 865	18 210	19 950
PhD thesis	2 744	3 866	4 535	4 778	5 041	5 084	5 021	3 359	2 706	3 083
Grand Total	49 740	76 913	76 058	72 955	74 602	77 744	63 273	47 550	41 289	41 868

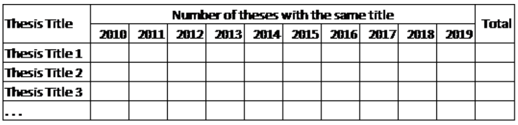
Metadata are able to detect an academic integrity breach in some cases. Repeating theses titles, high number of theses per supervisor, theses with higher similarity percentage are indicators of a potential integrity breach.

Repeating or similar titles of theses may imply academic misconduct or plagiarism among students. Such titles of theses could mean a failure of the HEIs, because the titles of theses from all Slovak HEIs are very simple accessible at the www.crzp.sk portal. There are different ways of repeating theses insights. All further examples are in the structural form because of sensitive data.

Number of theses with the same title (Table 8) can be seen from the level of faculty, HEI, HEI group, all HEIs, study field, group of study fields, scientific area, thesis type, supervisors, similarity percentage and combinations of them. However, it should be taken into account that the same thesis title does not always mean the same assignment.

Table 8

Number of theses with the same title

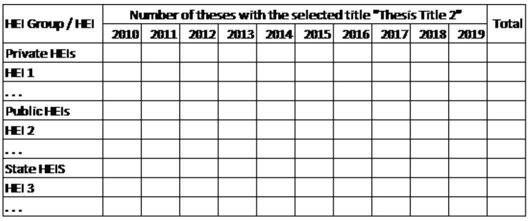


Thesis Title	Number of theses with the same title										Total
	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	
Thesis Title 1											
Thesis Title 2											
Thesis Title 3											
...											

Table 9 gives us deeper insight how the selected repeated thesis title is distributed among HEI groups and HEIs.

Table 9

Number of selected repeating thesis title according to HEI groups and HEIs

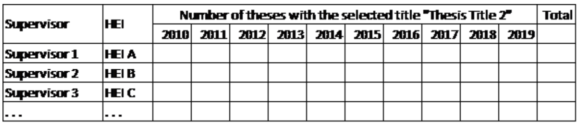


HEI Group / HEI	Number of theses with the selected title "Thesis Title 2"										Total
	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	
Private HEIs											
HEI 1											
...											
Public HEIs											
HEI 2											
...											
State HEIS											
HEI 3											
...											

Table 10 shows the selected repeated thesis title distribution by HEIs and supervisors.

Table 10

Ranking of supervisors according to the total number of selected repeating thesis title

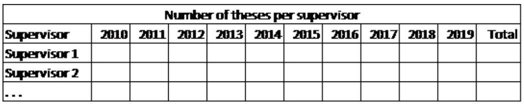


Supervisor	HEI	Number of theses with the selected title "Thesis Title 2"										Total
		2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	
Supervisor 1	HEI A											
Supervisor 2	HEI B											
Supervisor 3	HEI C											
...	...											

The number of theses per supervisor should not be excessively high. High numbers of theses per supervisor decrease the time during which supervisor can devote himself to a student and to his/her thesis. And that represents academic misconduct by the supervisor and the HEI, too. The insight like Table 11 was used by MinEdu to prepare measures to reduce the number of thesis per some supervisors. Many other variants of this table are possible.

Table 11

Number of thesis per supervisor

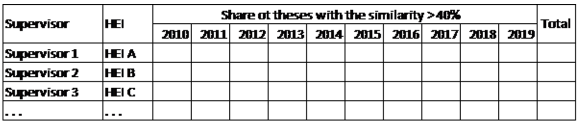


Number of theses per supervisor											
Supervisor	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	Total
Supervisor 1											
Supervisor 2											
...											

Higher similarity percentage of a thesis can be an indicator that there was a failure in the supervisor-student relationship. Table 12 offers a ranking of supervisors that have higher share of theses with similarity greater than 40%.

Table 12

Ranking of supervisors with a higher share of theses with similarity 40%

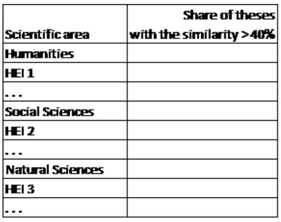


Supervisor	HEI	Share ot theses with the similarity >40%										Total
		2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	
Supervisor 1	HEI A											
Supervisor 2	HEI B											
Supervisor 3	HEI C											
...	...											

The share of theses with similarity greater than 40% by scientific areas is displayed in Table 13.

Table 13

Share of theses by scientific areas and HEIs with similarity 40%

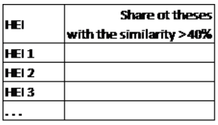


Scientific area	Share of theses with the similarity >40%
Humanities	
HEI 1	
...	
Social Sciences	
HEI 2	
...	
Natural Sciences	
HEI 3	
...	

A ranking of HEIs by a share of theses with similarity greater than 40% is given in Table 14.

Table 14

Share of HEI theses with similarity 40%

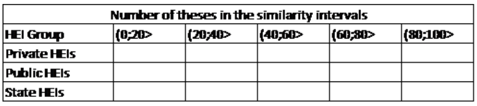


HEI	Share ot theses with the similarity >40%
HEI 1	
HEI 2	
HEI 3	
...	

Tables 15, 16, 17, 18 offer insights for the number of theses in five similarity intervals across HEI groups, thesis types and scientific areas. Similarity intervals can be defined by the user.

Table 15

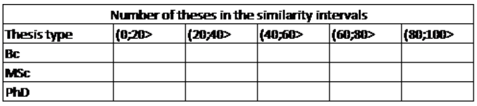
Number of theses in the selected similarity intervals by HEI group



Number of theses in the similarity intervals					
HEI Group	(0;20>	(20;40>	(40;60>	(60;80>	(80;100>
Private HEIs					
Public HEIs					
State HEIs					

Table 16

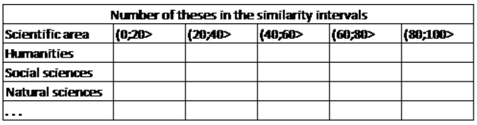
Number of theses in the selected similarity intervals by thesis type



Number of theses in the similarity intervals					
Thesis type	(0;20>	(20;40>	(40;60>	(60;80>	(80;100>
Bc					
MSc					
PhD					

Table 17

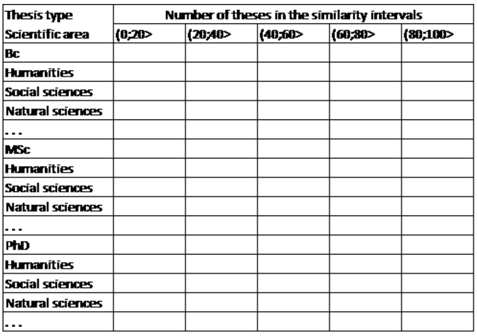
Number of theses in the selected similarity intervals by scientific area



Number of theses in the similarity intervals					
Scientific area	(0;20>	(20;40>	(40;60>	(60;80>	(80;100>
Humanities					
Social sciences					
Natural sciences					
...					

Table 18

Number of theses in the selected similarity intervals by thesis type and scientific area



Thesis type	Number of theses in the similarity intervals				
Scientific area	(0;20>	(20;40>	(40;60>	(60;80>	(80;100>
Bc					
Humanities					
Social sciences					
Natural sciences					
...					
MSc					
Humanities					
Social sciences					
Natural sciences					
...					
PhD					
Humanities					
Social sciences					
Natural sciences					
...					

Conclusion

If text similarity detection systems collect metadata in a uniform way, then their indisputable advantage are analytical insights. The absence of metadata collection means the absence of analytical insights. Metadata collection means more work that is rewarded by a range of analytical possibilities. More work means automated collection of metadata from academic information systems.

Centralised systems are relatively new and their comparative corpora are not as rich as those of the systems existing for twenty or more years. SK ANTIPLAG's ability to detect text similarities is very good in the local language thanks to rich comparative corpus of Slovak documents, but is weaker in other languages.

SK ANTIPLAG collects theses and metadata according to a uniform collection methodology and provides analytical insights that have common and comparable data base. This feature is out of reach for an academic environment where text similarity detection systems collect theses only.

The systems for the detection of text similarities are not a panacea, they have inherent limitations. One of them is the comparative corpus, which is the base for the comparison. No comparative corpus is all-embracing. These systems are only an element of the whole mosaic that helps to reduce plagiarism and to increase the level of academic integrity.

Acknowledgement

I would like to thank Mr. Matúš Benkovič (certified translator) for his help with the proofreading.

References

Bek, M. (2018). In: Masarykova univerzita řeší, jak zdokonalit boj s plagiáty, Zprávy z MUNI, `https://www.em.muni.cz/udalosti/10816-masarykova-univerzita-resi-jak-zdokonalit-boj-s-plagiaty`, Accessed: Nov 5, 2019

Bertram Gallant, T. (2018). Fake News, Truth & the Higher Education Imperative, Accessed: Apr 7, 2020

Brandejsová, J. (2018). In: Masarykova univerzita řeší, jak zdokonalit boj s plagiáty, Zprávy z MUNI, `https://www.em.muni.cz/udalosti/10816-masarykova-univerzita-resi-jak-zdokonalit-boj-s-plagiaty`, Accessed: Nov 5, 2019

Evering, L. C., & Moorman, G. (2012). Rethinking Plagiarism in the Digital Age. Journal of Adolescent & Adult Literacy, pp. 35–44

Higher Education Act Amendment, (2009).

Koz⍰owski M. (2019). Personal communication

Kravjar, J. (2015). SK ANTIPLAG is bearing fruit, Plagiarism across Europe and Beyond 2015, pp. 147–163, `https://plagiarism.pefka.mendelu.cz/files/proceedings_15.pdf`

Kravjar, J. (2018). Integrity is not component of ethics, integrity is much more, Apr 7, 2020
M. Ojsteršek, personal communication

Unknown (2019). Centrálny register záverečných a kvalifikačných prác, www.crzp.sk, Acessed Nov 7, 2019

Unknown (2019). Jednolity System Antiplagiatowy, `https://jsa.opi.org.pl`, Accessed: Nov 19, 2019

Unknown (2019). Theses.cz – Vysokoškolské kvalifikační práce, www.theses.cz, Accessed: Nov 11, 2019

Author

Julius Kravjar, Slovak Centre of Scientific and Technical Information, Lamacska cesta 8/A, 811 04 Bratislava, Slovakia e-mail: `julius.kravjar@gmail.com`