

Anti-plagiarism tools for our repositories

Jan Mach

University of Economics, Prague

Charles University in Prague

23. 10. 2013 Seminar on providing access to grey literature

What is plagiarism?

1. clone – presenting another's work, word-for-word, as one's own
2. CTRL-C – presenting another's work as one's own, with minimum changes
3. find/replace – changing key words and phrases but retaining the essential content of the source
4. remix – paraphrasing from several sources into a single text
5. recycle – using an author's previous texts, without citation
6. hybrid – mixing perfectly cited sources with non-cited ones
7. mashup – combining several non-cited sources into a text
8. error 404 – citations to non-existing sources or incorrect information about a source
9. aggregator – correct citation of foreign sources, but practically without any personal input by the author
10. re-tweet – correct citation, but using the original text/structure without significant changes

The Plagiarism Spectrum: Tagging 10 Types of Unoriginal Work

VŠKP testbad

ID souboru1

Nahrát

Uložit

kw: LTE, soubor: korpus/a1.xml

URL dokumentu:

http://cs.wikipedia.org/wiki/LTE

Název:

LTE

Zdroj:

wikipedia.cz

Datum zdroje:

05. 4. 2013 05:02

[dnes](#)

Formát:

html

Datum editace:

18. 4. 2013 02:27

[dnes](#)

Jazyk:

cz

Věta:

Velkou překážkou nasazení LTE v České republice je fakt, že dosud neproběhla aukce kmitočtů pro novou síť.

Odstavec:

LTE (zkratka z anglického 3GPP Long Term Evolution) je technologie určená pro vysokorychlostní Internet v mobilních sítích. Formálně jde o technologii spadající do standardu 3G, přičemž její následovník - LTE Advanced - bude již plnohodnotné 4G řešení. Teoretická rychlost stahování (downlink) je 172,8 Mbps a odesílání (uplink) 57,6 Mbps. V komerčním provozu se nachází např. v severovýchodních zemích či Estonsku, přičemž pokryty jsou zejména hustě osídlené oblasti - v Norsku je to Oslo, ve Švédsku historické centrum Stockholmu apod. Ceny připojení zhruba odpovídají cenám za pevné připojení (ADSL apod.), např. ve Švédsku stál v dubnu 2011 měsíční tarif bez FUP 50 €. V USA je používání LTE předmětem zkoumání úřadů, neboť dle americké armády mohou vysílače firmy LightSquared pracující v pásmu 1559 až 1610 MHz rušit GPS navigaci, která pracuje v sousedícím pásmu 1525 až 1559 MHz.

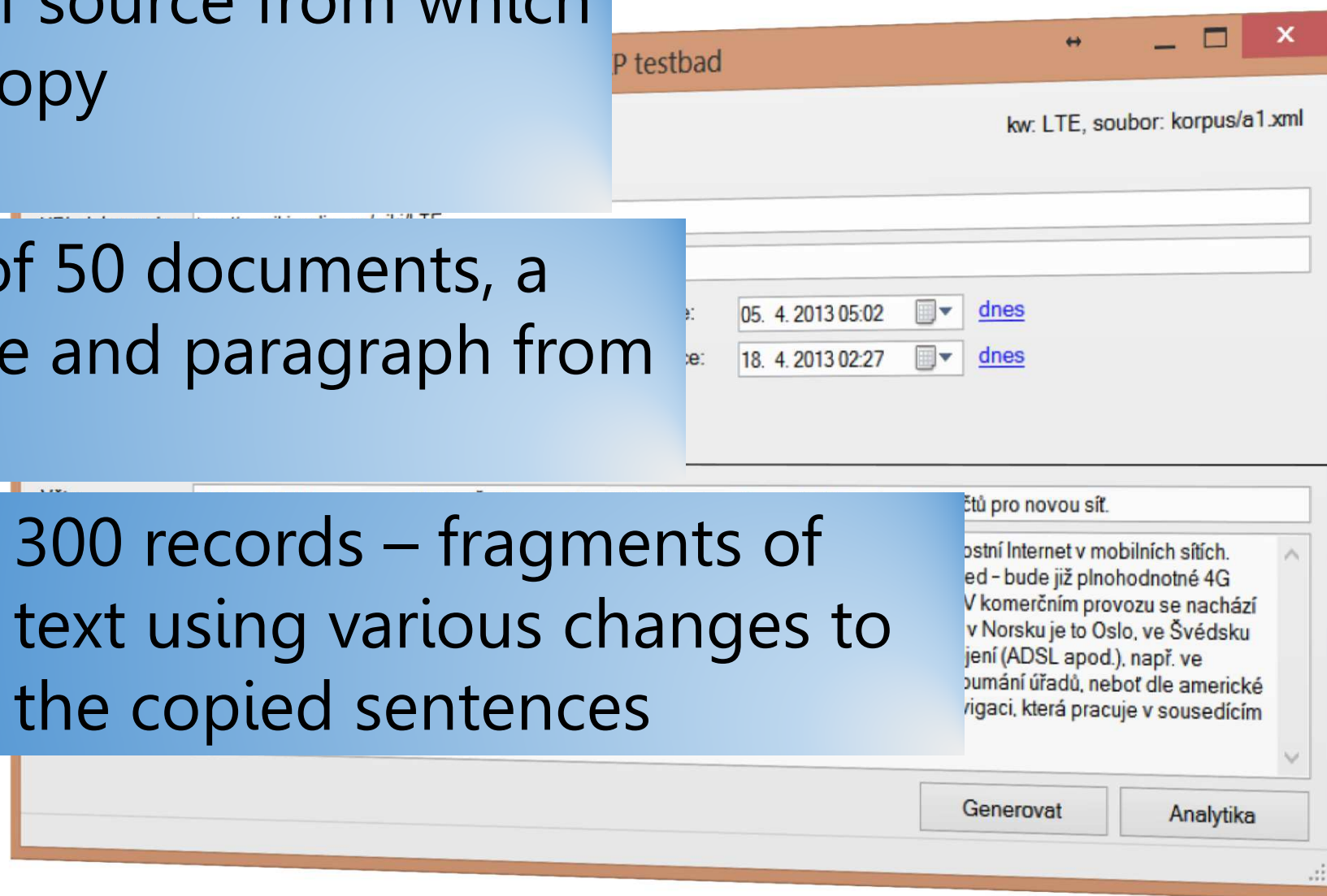
Generovat

Analytika

10 types of source from which students copy

a total of 50 documents, a sentence and paragraph from each

300 records – fragments of text using various changes to the copied sentences














































Transformations used

- a sentence with two words transposed,
- a sentence with diacritics removed,
- a sentence with a single word replaced with another with a similar meaning - paraphrasing a word,
- a sentence with several words replaced with others with similar meanings - paraphrasing a sentence,
- a sentence machine-translated into Czech/English

Hypotheses checked

1. The application is able to detect a single sentence copied from a source document.
2. The application is able to detect a single paragraph copied from a source document. The application is not affected by potential line breaks, indexes etc. in the source or tested document.
3. Successful detection is not impacted if the plagiarist adds/removes a word in the copied sentence.
4. The application can detect Czech texts irrespective of the use of diacritics.
5. Successful detection is not impacted if the plagiarist paraphrases a single word in a sentence.
6. Successful detection is not impacted if the plagiarist paraphrases a whole sentence.
7. Successful detection is not impacted if the plagiarist translates text from/to Czech.
8. The Theses.cz system should achieve the best results in the detection of plagiarism in Czech theses and dissertations.
9. A low percentage of the total number of similarities will be detected from the Anopress source compared to sources freely available on the internet.
10. Better results with EIR and Open Access sources are achieved by foreign tools rather than Czech ones.
11. Very good results for web sources will be achieved by systems using web search services.

Hypotheses checked

Hypotéza	Thesis		Turnitin		Ephorus		GooglePl.		Průměr
1		12%		40%		2%		56%	28%
2		14%		42%		6%		46%	27%
3		100%		100%		0%		0%	50%
4		100%		100%		0%		80%	70%
5		67%		100%		0%		4%	43%
6		0%		88%	na			0%	29%
7		0%		0%		0%		0%	0%
8		10%		50%		10%		30%	25%
9		0%		0%		0%		0%	0%
10		0%		40%		0%		70%	28%
11		20%		50%		0%		80%	38%

TURNITIN

ABOUT THE APPLICATION

- 15 language variants without Czech
- large database of texts
- price based on number of students, hundreds of thousands of crowns
- integration with MOODLE system and others, no API
- GradeMark and PeerMark modules

EVALUATION OF SIMILARITIES

- processing within 30 s
- configurable size of searched similarities, possibility to exclude citations
- very clear and functional interface with similarities, association of sources

Internet source

Full Source View

cs.wikipedia.org

á pracuje v sousedícím pásmu 1525 až 1559 MHz. Dne 12. června 2012 spustila společnost Telefónica Czech Republic v obci Jesenice u Prahy první komerční provoz sítě čtvrté generace LTE. **Velkou překážkou nasazení LTE v České republice je fakt, že dosud neproběhla aukce kmitočtů pro novou síť.** ČTÚ bude přijímat přihlášky zájemců do 10. září, otevírání obálek s nabídkami se pak bude konat o den později. Samotné přidělení frekvencí proběhne zhruba v lednu roku 2013

1dstavec 1 - bez změn

Velkou překážkou nasazení LTE v České republice je fakt, že dosud neproběhla aukce kmitočtů pro novou síť.

1dstavec 2 - bez změn

LTE (zkratka z anglického 3GPP Long Term Evolution) je technologie určená pro vysokorychlostní Internet v mobilních sítích. Formálně jde o technologii spadající do standardu 3G, přičemž její následovník – LTE Advanced – bude již plnohodnotné 4G řešení. Teoretická rychlost stahování (downlink) je 172,8 Mbps a odesílání (uplink) 57,6 Mbps. V komerčním provozu se nachází např. v severských zemích či Estonsku, přičemž pokryty jsou zejména hustě osídlené oblasti – v Norsku je to Oslo, ve Švédsku historické centrum Stockholmu apod. Ceny připojení zhruba odpovídají cenám za pevné připojení (ADSL apod.), např. ve Švédsku stál v dubnu 2011 měsíční tarif bez FUP 50 €. V USA je používání LTE předmětem zkoumání úřadů, neboť dle americké armády mohou vysílače firmy LightSquared pracující v pásmu 1559 až 1610 MHz rušit GPS navigaci, která pracuje v sousedícím pásmu 1525 až 1559 MHz.

← Match Breakdown

1 cs.wikipedia.org
Internet source

← Match 1 of 20

cs.wikipedia.org
Internet source - 8 urls

- [wiki/Universal_Mobile_Telecom](#)
- [wiki/LTE](#)
- [wiki/WiMAX](#)
- [wiki/Enhanced_Data_Rates](#)
- [wiki/Přepojování_paketů](#)
- [wiki/Paketově_spínaná_dom](#)
- [wiki/General_Packet_Radio](#)
- [wiki/GPRS](#)

blackberry.divoce.cz
Internet source - 4 urls**www.skarpety.slask.pl**
Internet source - 2 urls

Exclude Sources

EPHORUS

ABOUT THE APPLICATION

- the application is used by over 3,000 schools and universities, 4 schools in the CR (Faculty of Business Administration at the University of Economics, Prague)
- interface in Czech
- operator claims a database with billions of web pages, submitted works, journal texts etc.

EVALUATION OF SIMILARITIES

- possible to define a min. similarity percentage
- results sent by e-mail, attachments in PDF
- basic web interface
- no source de-duplication

Soubor 14, zdroj wikipedia.org, kw WIMAX

Odstavec 1 - bez změn

The WiMAX Forum website provides a list of certified devices.

Odstavec 2 - bez změn

dále

odevzdáno:	Nalezeno:
In North America, backhaul for urban operations is typically provided via one or more copper wire line connections, whereas remote cellular operations are sometimes backhauled via satellite. In other regions, urban and rural backhaul is usually provided by microwave links. (The exception to this is where the network is operated by an incumbent with ready access to the copper network.)	In North America, backhaul for urban <i>cellular</i> operations is typically provided via one or more copper wire line <i>T1</i> connections, whereas remote cellular operations are sometimes backhauled via satellite. In <i>most</i> other regions, urban and rural backhaul is usually provided by microwave links. (The exception to this is where the network is operated by an incumbent with ready access to the copper network,

dále

WiMAX

dále

odevzdáno:	Nalezeno:
has more substantial backhaul bandwidth requirements than legacy cellular applications.	has <i>much</i> more substantial backhaul bandwidth requirements than legacy cellular applications.

dále

odevzdáno:	Nalezeno:
Consequently the use of wireless microwave backhaul is on the rise in North America and existing microwave backhaul links in all regions are being upgraded.[8] Capacities of between 34 Mbit/s and 1	Consequently the use of wireless microwave backhaul is on the rise in North America and existing microwave backhaul links in all regions are being <i>upgraded</i> . [12] Capacities of between 34 <i>Mbps</i> and 1

dále

Gbit/s [9]

nahoru

odevzdáno:	Nalezeno:
are routinely being deployed with latencies in the order	are routinely being deployed with latencies in the order

MUNI SYSTEMS

ABOUT THE APPLICATION

- theses.cz, odevzdej.cz and repozitar.cz
- over 30 public and private schools in the CR and SR
- price per number of students
- extensive database of Czech theses and dissertations, study materials and selected web pages
- API for connection

EVALUATION OF SIMILARITIES

- processing takes some hours
- duplicate documents
- comparing pairs of documents
→ two lists of similarities
- no overall percentage of detected similarities
- similarities displayed only from 5 % of the length of one of the compared document in a pair

Vloženo/změněno: 13. 6. 2013, Ing. Jan Mach
Zkontrolováno: Podobnost tohoto dokumentu byla zkontrolována.
Soubory:  /doc/vse/hpvr9ts/v5_Testovaci_soubor_komplet.docx

[Nápověda k podobnosti souborů](#)

Obsah zkoumaného souboru je z X % podobný souboru níže:



K vloženému souboru nebyl v databázi nalezen žádný podobný dokument.

Obsah souborů níže je z X % podobný zkoumanému souboru:

34 %

Agenda: Zdroj z Internetu:

- <http://cs.wikipedia.org/wiki/Podobnosti>

Změněno: 11. 3. 2013 17:17.40
[Podobnosti](#)

22 %

Agenda: Zdroj z Internetu:

- <http://www.wimax.cz>

Změněno: 9. 1. 2013 21:53.19
[Podobnosti](#)

20 %

Agenda: Zdroj z Internetu:

The first list contains documents with similarity of min. 5 % of the inspected file.
bachelor paper with 40 pages: 2 pages

The second list complements the previous list with other documents, but only with the length of similarity min. 5 % of the found file.

GooglePlagiarism

ABOUT THE APPLICATION

- my own desktop application for personal computers running Windows
- intended for personal analysis of documents by an individual
- searching for whole sentences in the Google search engine

EVALUATION OF SIMILARITIES

- limited number of searches → processing takes several hours
- HTML output without retaining formatting
- highlighted detected sentences and the first corresponding source

forwarded. Odstavec 1 - parafráze slova In the UMTS, both data packets and calls can be conveyed. Soubor 93, zdroj EIZ, kw EDGE Odstavec 1 - bez změn Technology is now a global business, and U.S. companies with the financial talent to manage money worldwide will have an edge on their competition. Odstavec 2 - bez změn The difference in companies is always related to talent. As a headhunter, my primary measure for a great manager is his or her attitude toward hiring talent. I agree absolutely with something Vinod Khosla told me recently. "In most situations," he said. "I would pick a person who has a thirst for great teams, is a great recruiter and a team leader over someone with great"One measure of how important A players are is what happens to companies that lose their talent. Odstavec 1 - MS Translator CZ Technologie je nyní globální podnikání a americké společnosti s finanční talent ke správě peněz po celém světě bude mít výhodu na jejich konkurenci. Odstavec 1 - prohození Technology is a now global business, and U.S. companies with the financial talent to manage money worldwide will have an edge on their competition. Odstavec 1 - parafráze věty Technology is now a global company, and U.S. enterprises with financial talents manage funds global has the edge on the competition. Odstavec 1 - parafráze slova Technology is now a global company, and U.S. companies with the financial talent to manage money worldwide will have an edge on their competition. Soubor 94, zdroj EIZ, kw WIMAX Odstavec 1 - bez změn Usually, everyone present would know who was calling, often everyone would know what the call was about. Odstavec 2 - bez změn Rich Ling and Scott Campbell's excellent edited book surveys the mobile revolution that is a key component in the networked operating system that has It is the rapidly proliferating use of mobile devices to access communication and information at almost all times—and at the same time, to be accessible. As author Ann Light says, "the phone (a) becomes one with our body and (b) follows us about" (p. 195). Yet they are more than technological fads. Odstavec 1 - MS Translator CZ Obvykle všem přítomným věděl, kdo volá, často každý věděl, co bylo o volání. Odstavec 1 - prohození Usually, everyone would present know who was calling, often everyone would know what the call was about. Odstavec 1 - parafráze věty Usually, everyone here will know who called, frequently all will know what the call was about. Odstavec 1 - parafráze slova Usually, everyone here would know who was calling, often everyone would know what the call was about. Soubor 95, zdroj EIZ, kw Wi-Fi Odstavec 1 - bez změn Today, Wi-Fi can apply to products that use any 802.11 standard. Odstavec 2 - bez změn WiFi and VoIP are being widely deployed in enterprises. WiFi is easy and flexible to deploy, and is claimed to be more reliable in terms of coverage while costing less than traditional cellular services. It is also expected that using VoWiFi new converged applications can be developed for mobile workers with new capabilities such as geographic location. Ongoing technical developments, such as dual mode handsets, Session Initiation Protocol (SIP) and softphones are helping in stimulating the further development of. Odstavec 1 - MS Translator CZ Dnes Wi-Fi lze použít na produkty, které používají jakýkoli standard 802.11. Odstavec 1 - prohození Today, Wi-Fi apply can to products that use any 802.11 standard. Odstavec 1 - parafráze věty Today, Wi-Fi can be applied to products using any 802.11 norm. Odstavec 1 - parafráze slova Today, Wi-Fi can be to products that use any 802.11 standard.

Seznam významných URL

7 vět, 705 znaků: [LTE – Wikipedie](#)

2 vět, 215 znaků: [Pro spotřebitele - Slovníček | APMS](#)

3 vět, 346 znaků: [Digitálník.cz - o2 spustí lte a zrychlí 3g](#)

3 vět, 346 znaků: [3AM's geoBlog](#)

2 vět, 220 znaků: [LTE – MobilMania.cz](#)

6 vět, 711 znaků: [Přepojování paketů – Wikipedie](#)

6 vět, 711 znaků: [Universal Mobile Telecommunications System – Wikipedie](#)

Without retention of size and line breaks, navigation in a text during checks is much more difficult.

Evaluation of system control and function

Evaluation	Thesis	Turnitin	Ephorus	GooglePl.
processing time	✗	✓	⚠	✗
clarity of results	✗	✓	⚠	⚠
display of overall similarity	✗	✓	⚠	⚠
minimum similarity	✗	✓	✓	⚠
price	✓	✗	⚠	✓
integration with school IR's	✓	✗	⚠	✗
de-duplications of sources	✗	✓	✗	✓

The Thesis.cz system stands out thanks to its low price and possibility of repository integration.

The Turnitin application excels through its user interface and available functions, but is expensive and not easy to integrate.

The Ephorus system would be a good compromise between Thesis and Turnitin, yet ...

Number of documents detected according to source

Category	Corpus	Thesis	Turnitin	Ephorus	GooglePl.	average
wikipedia.cz	5	3	5	2	5	3,75
wikipedia.org (en)	5	1	3	2	5	2,75
ETDs (cz)	5	1	2	1	1	1,25
ETDs (en)	5	0	3	0	2	1,25
NDLTD	5	0	0	0	1	0,25
Anopress	5	0	0	0	0	0
Arxive.org	5	0	1	0	3	1
Google.cz (cz)	5	2	3	0	5	2,5
Google.com (en)	5	0	2	0	3	1,25
el. inf. Resources	5	0	3	0	4	1,75
total	50	7	22	5	29	15,75

Category	Corpus	Thesis	Turnitin	Ephorus	GooglePl.	average
wikipedia.cz	100%	60%	100%	40%	100%	75%
wikipedia.org (en)	100%	20%	60%	40%	100%	55%
ETDs (cz)	100%	20%	40%	20%	20%	25%
ETDs (en)	100%	0%	60%	0%	40%	25%
NDLTD	100%	0%	0%	0%	20%	5%
Anopress	100%	0%	0%	0%	0%	0%
Arxive.org	100%	0%	20%	0%	60%	20%
Google.cz (cz)	100%	40%	60%	0%	100%	50%
Google.com (en)	100%	0%	40%	0%	60%	25%
el. inf. Resources	100%	0%	60%	0%	80%	35%
average	100%	14%	44%	10%	58%	32%

Low number of documents found using the Ephorus system.

Documents from Anopress were not detected by any system.

Most documents were detected by the Turnitin and GooglePlagiarism systems.

Number of documents detected according to document language

Language	Corpus	Thesis	Turnitin	Ephorus	GooglePl.	average
Czech	19	6	10	3	11	7,5
English	30	1	12	2	18	8,25
Slovak	1	0	0	0	0	0
total	50	7	22	5	29	15,75

Language	Corpus	Thesis	Turnitin	Ephorus	GooglePl.	average
Czech	100%	32%	53%	16%	58%	39%
English	100%	3%	40%	7%	60%	28%
Slovak	100%	0%	0%	0%	0%	0%

The Theses.cz system detected an average number of Czech documents, but posted the worst results for English documents.

Still, however, more than Ephorus overall. A reduction in the 5% limit would significantly enhance the success rate of Theses.cz!

Number of records detected according to type of change

– suspicion of plagiarism

Transformation	Corpus	Thesis	Turnitin	Ephorus	GooglePl.	average
one sentence	50	6	20	1	28	13,75
one paragraph	50	7	21	3	23	13,5
swapping words	50	6	20	1	0	6,75
no diacritics	19	5	9	1	8	5,75
paraphrased sentence	31	0	10	0	0	2,5
paraphrased word	50	4	20	1	1	6,5
translation	50	0	0	1	0	0,25
total	300	28	100	8	60	49,00

Transformation	Corpus	Thesis	Turnitin	Ephorus	GooglePl.	average
one sentence	100%	12%	40%	2%	56%	28%
one paragraph	100%	14%	42%	6%	46%	27%
swapping words	100%	12%	40%	2%	0%	14%
no diacritics	100%	26%	47%	5%	42%	30%
paraphrased sentence	100%	0%	32%	0%	0%	8%
paraphrased word	100%	8%	40%	2%	2%	13%
translation	100%	0%	0%	2%	0%	1%
average	100%	10%	35%	3%	21%	17%

















































Searching for whole sentences in the GooglePlagiarism application does not detect text changes.

The Ephorus system detected only 8 similar passages in the text, however these were mainly transcriptions of abbreviations.

Number of detected records according to type of change

– **proof** of plagiarism

Transformation	Corpus	Thesis	Turnitin	Ephorus	GooglePl.	average
one sentence	50	5	8	0	25	9,5
one paragraph	50	6	10	1	9	6,5
swapping words	50	1	7	0	0	2
no diacritics	19	4	6	0	7	4,25
paraphrased sentence	31	0	2	0	0	0,5
paraphrased word	50	3	8	0	1	3
translation	50	0	0	0	0	0
total	300	19	41	1	42	25,75

Transformation	Corpus	Thesis	Turnitin	Ephorus	GooglePl.	average
one sentence	100% 	10% 	16% 	0% 	50% 	19% 
one paragraph	100% 	12% 	20% 	2% 	18% 	13% 
swapping words	100% 	2% 	14% 	0% 	0% 	4% 
no diacritics	100% 	21% 	32% 	0% 	37% 	22% 
paraphrased sentence	100% 	0% 	6% 	0% 	0% 	2% 
paraphrased word	100% 	6% 	16% 	0% 	2% 	6% 
translation	100% 	0% 	0% 	0% 	0% 	0% 
average	100% 	7% 	15% 	0% 	15% 	9% 

The Ephorus system actually detected only one document clearly showing plagiarism.

As yet none of the systems is able to search for a translated text.

GooglePlagiarism best detects sentences without changes, while Turnitin best detects sentences with changes.

Final summary

The Turnitin application achieves very good results, but is very expensive.

The Ephorus application is inadequate at detecting duplicates in the text corpus.

The Theses.cz application is a good compromise between price and capability. Removing the 5% limit on similarity detection would help.

Searching for sources online in GooglePlagiarism is very effective at detecting copied texts.

You can find detailed test results in the proceedings of the
Seminar on providing access to grey literature 2013

<http://nrgl.techlib.cz/index.php/Workshop>

Jan Mach
machj@vse.cz