

Antiplagiátorské nástroje pro naše repozitáře

Jan Mach

Vysoká škola ekonomická v Praze

Univerzita Karlova v Praze

23. 10. 2013 Seminář ke zpřístupňování šedé literatury

Co je plagiát?

1. klonování – vydávání cizí práce, slovo od slova, za vlastní
2. CTRL-C – vydávání cizí práce za vlastní, s minimálním množstvím úprav
3. najít/nahradit – změna klíčových slov a frází bez změny podstaty textu
4. remixování – parafrázování z více zdrojů do jednoho textu
5. recyklování – využití předchozích textů autora, bez autocitace
6. hybridní – mixování velmi dobře citovaných zdrojů s necitovanými
7. míchání zdrojů – kombinace více necitovaných zdrojů do textu
8. chyba 404 – citace neexistujících zdrojů nebo špatné informace o zdroji
9. agregace – korektní citování cizích zdrojů, ale téměř bez vlastního osobního přínosu autorem
10. re-tweet – korektní citování, ale za využití originálního textu/struktury bez podstatnějších úprav

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 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.

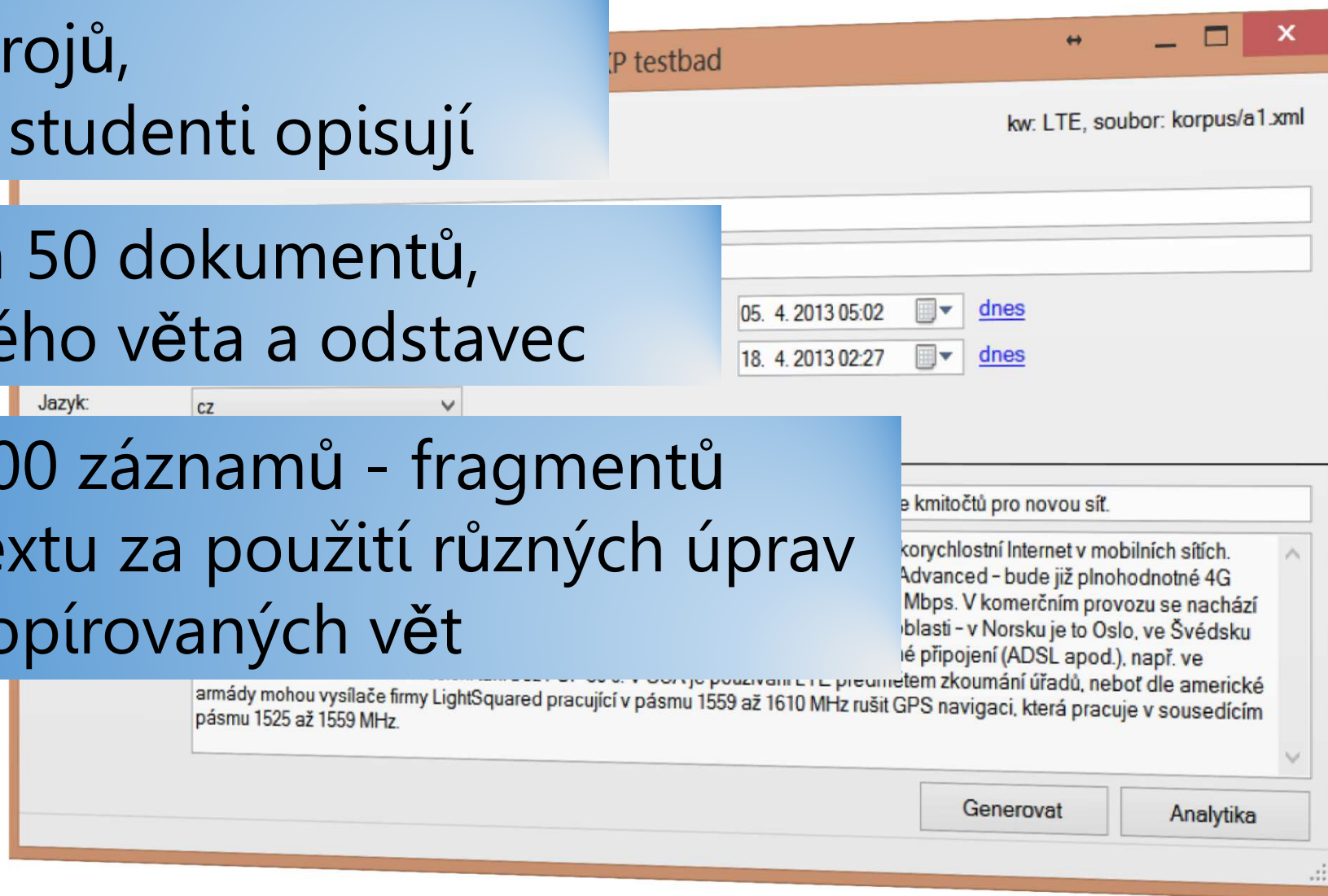
Generovat

Analytika

10 typů zdrojů,
ze kterých studenti opisují

celkem 50 dokumentů,
z každého věta a odstavec

300 záznamů - fragmentů
textu za použití různých úprav
kopírovaných vět



Použité transformace

- věta se dvěma slovy prohozenými,
- věta s odstraněnou diakritikou,
- věta s jedním slovem nahrazeným slovem významově blízkým – parafráze slova,
- věta s více slovy nahrazenými slovy významově blízkými – parafráze věty,
- věta přeložená automaticky do českého/anglického jazyka

Ověřované hypotézy

1. Aplikace umí odhalit jednu větu zkopírovanou ze zdrojového dokumentu.
2. Aplikace umí odhalit jeden odstavec zkopírovaný ze zdrojového dokumentu. Aplikaci nevadí případná zalomení řádků, indexy apod. ve zdrojovém nebo testovaném dokumentu.
3. Pro úspěšnou detekci nevadí, pokud plagiátor přidá/odebere slovo v kopírované větě.
4. Aplikace provádí detekci českých textů nezávisle na diakritice.
5. Pro úspěšnou detekci nevadí, pokud plagiátor parafrázuje jedno slovo ve větě.
6. Pro úspěšnou detekci nevadí, pokud plagiátor parafrázuje celou větu.
7. Pro úspěšnou detekci nevadí, pokud plagiátor přeloží text z/do českého jazyka.
8. Systém Theses.cz by měl v detekci plagiátorství u českých VŠKP dosahovat nejlepších výsledků.
9. U zdroje Anopress bude nalezeno nízké procento z celkového počtu podobností oproti zdrojům volně dostupným na Internetu.
10. Lepších výsledků u EIZ a Open Access zdrojů dosáhnou nástroje zahraniční oproti českým.
11. Velmi dobrých výsledků u webových zdrojů budou dosahovat systémy využívající služeb webových vyhledávačů.

Ověřované hypotézy

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

O APLIKACI

- 15 jazykových mutací bez češtiny
- velká databáze textů
- cena dle počtu studentů, v řádu stovek tisíc Kč
- integrace se systémy MOODLE apod., bez API
- moduly GradeMark a PeerMark

VYHODNOCENÍ PODOBNOSTÍ

- zpracování do 30 s
- konfigurovatelná velikost hledaných podobností, možnost vyloučení citací
- velmi přehledné a funkční rozhraní s podobnostmi, sdružování zdrojů

Originality

GradeMark

PeerMark

Testovací korpus CZ

BY JAN MACH

turnitin

20%

SIMILAR

--

OUT OF 100

Match Breakdown

1 cs.wikipedia.org
Internet source 9%

Match 1 of 20

cs.wikipedia.org
Internet source - 8 urls 9%

- wiki/Universal_Mobile_Telecom... 3%
- wiki/LTE 2%
- wiki/WiMAX 2%
- wiki/Enhanced_Data_Rates... 2%
- wiki/Přepojování_paketů 1%
- wiki/Paketově_splínaná_dom... 1%
- wiki/General_Packet_Radio... 1%
- wiki/GPRS 1%

blackberry.divoce.cz
Internet source - 4 urls 2%

www.skarpety.slask.pl
Internet source - 2 urls 2%

Exclude Sources

View/edit excluded 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 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.

EPHORUS

O APLIKACI

- aplikaci používá přes 3 000 škol a univerzit, v ČR 4 školy (FPH na VŠE)
- rozhraní v češtině
- dle provozovatele databáze s miliardami webových stránek, odevzdanými pracemi, texty z časopisů aj.

VYHODNOCENÍ PODOBNOSTÍ

- možno definovat min. procento podobností
- odesílání výsledků e-mailem, přílohy v PDF
- základní rozhraní na webu
- chybí deduplikace zdrojů

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

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

nahoru

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

SYSTÉMY MUNI

O APLIKACI

- theses.cz, odevzdej.cz a repozitar.cz
- přes 30 veřejných i soukromých škol z ČR a SK
- cena dle počtu studentů
- rozsáhlá databáze českých VŠ závěrečných prací, studijních materiálů a vybraných webových stránek
- API pro napojení

VYHODNOCENÍ PODOBNOSTÍ

- zpracování v řádu hodin
- duplicitní dokumenty
- porovnávání dvojic dokumentů
→ dva seznamy podobností
- chybí celkové procento nalezených podobností
- zobrazeny podobnosti pouze od 5 % délky jednoho z porovnávané dvojice dokumentů

Vloženo/změněno: 13. 6. 2013, Ing. Jan Mach
Zkontrolováno: Podobnost tohoto dokumentu byla zkontrolována.
Soubory:  [/doc/vse/hpvr9ts/v5_Testovací_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

Druhý seznam doplňuje předchozí o dokumenty další, ale pouze s délkou podobnosti min. 5 % nalezeného souboru.

22 %

Agenda: Zdroj z Internetu:

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

Změněno: 9. 1. 2013 21:53.19

20 %

Agenda: Zdroj z Internetu:

GooglePlagiarism

O APLIKACI

- vlastní desktopová aplikace pro osobní počítače s Windows
- určena pro osobní analýzy dokumentů jednotlivcem
- hledání celých vět ve vyhledávači Google

VYHODNOCENÍ PODOBNOSTÍ

- limitovaný počet vyhledávání → doba zpracování v řádu hodin
- výstupem HTML bez zachování formátování
- vyznačené nalezené věty a první odpovídající zdroj

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 now a 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 de 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](#)

Pokud není zachována velikost a zalomení řádků, je velmi snížena orientace v textu při kontrole.

Hodnocení ovládání a funkcí systémů

Hodnocení	Thesis	Turnitin	Ephorus	GooglePl.
doba zpracování	✗	✓	⚠	✗
přehlednost výsledků	✗	✓	⚠	⚠
zobrazení celkové podobnosti	✗	✓	⚠	⚠
minimální podobnost	✗	✓	✓	⚠
cena	✓	✗	⚠	✓
integrace s IS školy	✓	✗	⚠	✗
deduplikace zdrojů	✗	✓	✗	✓

System Thesis.cz vyniká příznivou cenou a možností integrace s repozitáři.

Aplikace Turnitin exceluje uživatelským rozhraním a dostupnými funkcemi, ale je drahá a nelze ji snadno zaintegrovat.

System Ephorus by byl dobrým kompromisem mezi Thesis a Turnitin, ale ...

Počet nalezených dokumentů podle zdroje

Kategorie	Korpus	Thesis	Turnitin	Ephorus	GooglePl.	Průměr
wikipedia.cz	5	3	5	2	5	3,75
wikipedia.org (en)	5	1	3	2	5	2,75
VŠKP (cz)	5	1	2	1	1	1,25
VŠKP (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
EIZ	5	0	3	0	4	1,75
Celkem	50	7	22	5	29	15,75

Kategorie	Korpus	Thesis	Turnitin	Ephorus	GooglePl.	Průměr
wikipedia.cz	100%	60%	100%	40%	100%	75%
wikipedia.org (en)	100%	20%	60%	40%	100%	55%
VŠKP (cz)	100%	20%	40%	20%	20%	25%
VŠKP (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%
EIZ	100%	0%	60%	0%	80%	35%
Průměr	100%	14%	44%	10%	58%	32%

Nízký počet dokumentů nalezených systémem Ephorus.

Dokumenty z Anopressu nebyly nalezeny žádným systémem.

Nejvíce dokumentů našly systémy Turnitin a GooglePlagiarism.

Počet nalezených dokumentů podle jazyka dokumentu

Jazyk	Korpus	Thesis	Turnitin	Ephorus	GooglePl.	Průměr
česky	19	6	10	3	11	7,5
anglicky	30	1	12	2	18	8,25
slovensky	1	0	0	0	0	0
Celkem	50	7	22	5	29	15,75

Jazyk	Korpus	Thesis	Turnitin	Ephorus	GooglePl.	Průměr
česky	100%	32%	53%	16%	58%	39%
anglicky	100%	3%	40%	7%	60%	28%
slovensky	100%	0%	0%	0%	0%	0%

System Theses.cz našel průměrný počet českých dokumentů, u anglických dokumentů dosáhl výsledků nejhorších.

Stále však v celku více než Ephorus. Snížení 5% hranice by úspěšnosti Theses.cz výrazně pomohlo!

Počet nalezených záznamů podle typu úprav

– podezření na plagiátorství

Úprava	Korpus	Thesis	Turnitin	Ephorus	GooglePl.	Průměr
jedna věta	50	6	20	1	28	13,75
jeden odstavec	50	7	21	3	23	13,5
prohození slova	50	6	20	1	0	6,75
bez diakritiky	19	5	9	1	8	5,75
parafráze věty	31	0	10	0	0	2,5
parafráze slova	50	4	20	1	1	6,5
překlad	50	0	0	1	0	0,25
Celkem	300	28	100	8	60	49,00

Úprava	Korpus	Thesis	Turnitin	Ephorus	GooglePl.	Průměr
jedna věta	100%	12%	40%	2%	56%	28%
jeden odstavec	100%	14%	42%	6%	46%	27%
prohození slova	100%	12%	40%	2%	0%	14%
bez diakritiky	100%	26%	47%	5%	42%	30%
parafráze věty	100%	0%	32%	0%	0%	8%
parafráze slova	100%	8%	40%	2%	2%	13%
překlad	100%	0%	0%	2%	0%	1%
Průměr	100%	10%	35%	3%	21%	17%

Hledání celých vět v aplikaci GooglePlagiarism nedetekuje úpravy textu.

Systém Ephorus detekoval pouze 8 zkopírovaných pasáží textu, především ale jen typu rozepsání zkratky.

Počet nalezených záznamů podle typu úprav

– důkazy plagiátorství

Úprava	Korpus	Thesis	Turnitin	Ephorus	GooglePl.	Průměr
jedna věta	50	5	8	0	25	9,5
jeden odstavec	50	6	10	1	9	6,5
prohození slova	50	1	7	0	0	2
bez diakritiky	19	4	6	0	7	4,25
parafráze věty	31	0	2	0	0	0,5
parafráze slova	50	3	8	0	1	3
překlad	50	0	0	0	0	0
Celkem	300	19	41	1	42	25,75

Úprava	Korpus	Thesis	Turnitin	Ephorus	GooglePl.	Průměr
jedna věta	100%	10%	16%	0%	50%	19%
jeden odstavec	100%	12%	20%	2%	18%	13%
prohození slova	100%	2%	14%	0%	0%	4%
bez diakritiky	100%	21%	32%	0%	37%	22%
parafráze věty	100%	0%	6%	0%	0%	2%
parafráze slova	100%	6%	16%	0%	2%	6%
překlad	100%	0%	0%	0%	0%	0%
Průměr	100%	7%	15%	0%	15%	9%

Systém Ephorus
ve skutečnosti jen jednou
nalezl dokument jasně
prokazující plagiátorství.

Zatím žádný ze systémů
neumí vyhledat přeložený
text.

Věty bez úprav vyhledává
nejlépe GooglePlagiarism,
věty s úpravami nejlépe
Turnitin.

Závěrečné shrnutí

Aplikace Turnitin dosahuje velmi dobrých výsledků, ale je velmi drahá.

Aplikace Ephorus vyhledávala duplicity v testovacím korpusu nedostatečně.

Aplikace Theses.cz je dobrým kompromisem mezi cenou a schopnostmi. Zlepšením by bylo odstranění 5% limitu na délku podobnosti.

Vyhledáváním zdrojů online v GooglePlagiarism je velmi účinné na detekci zkopírovaných textů.

**Detailní výsledky testu naleznete ve sborníku
Semináře ke zpřístupňování šedé literatury 2013**

<http://nysl.techlib.cz/sbornik/>

Jan Mach
machj@vse.cz