Jak se odhalují plagiáty? Přeházet slova nestačí, překlad ale stroj ošálí

11. července 2018, aktualizováno 18:07

Nástroje na detekci opsaných pasáží mají nemalý problém. Musí porovnat odevzdanou diplomovou práci s miliony stránek už existujících textů. Co nejrychleji a pokud možno spolehlivě. Podívejte se s námi, proč je hledání plagiátů náročný a zajímavý problém kombinující informatiku, lingvistiku a psychologii.

Opsala, nebo neopsala? Zapomněla jenom citaci, nebo úmyslně okopírovala část cizí diplomové práce? Taťána Malá, která byla na necelé dva týdny českou ministryní spravedlnosti, musela na podobné otázky odpovídat poté, co Český rozhlas odhalil, že její magisterská diplomová práce z roku 2011 obsahuje necitované pasáže z cizí diplomky. Později se ukázalo, že i její diplomová práce z roku 2005 obsahuje "jasné plagiátorství".

Malá sice ze své funkce odstoupila, nadále však trvá na tom, že je na své práce pyšná a o plagiáty nejde. „Kdo z vás psal diplomovou práci, víte, jak to funguje," uvedla na tiskové konferenci. „Nelze hodnotit práci z roku 2005 dnešními měřítky." Tehdy podle ní platila jiná pravidla. Později Malá připustila, že možná ne všechno správně odcitovala: „Já jsem vám říkala, že jsem si dělala poznámkový aparát a vykopírovávala jsem si věci do samostatného souboru. A když jsem tu práci zpracovávala, tak jsem možná ten [sporný text] zapomněla odcitovat. To je jediné, co mě napadá. Prostě se mi to stalo, omlouvám se, ale myslím si, že to není důvod k tomu zpochybňovat tu práci jako takovou."

Pomineme teď nesmyslné tvrzení, že v roce 2005 platila „jiná pravidla" - povinnost citovat všechny zdroje ostatně diplomantka stvrdila hned na začátku podpisem u prohlášení o tom, že využila pouze citovanou literaturu. Ve výčtu ovšem neuvedla například diplomovou práci Zity Pavlišové, ze které Malá (tehdy Veličková) opsala - slovo od slova - celé pasáže.

Kauza odhalila řadu nedorozumění, která zjevně panují ohledně diplomových prací obecně a o plagiátorství zejména. V diskuzích jsme narazili na názory, že když práce prošla obhajobou, musí být v pořádku, nebo že opisování cizích textů do diplomek patří. Rozhodli jsme se proto některé tyto omyly uvést na pravou míru a podívat se pod „kapotu" systémům, které mají podobným plagiátorským pokusům zabránit.

Jak lze odhalit plagiát? Oponent i vedoucí jsou na to krátcí

Od samého začátku se Malá hájila tím, že diplomovou práci obhájila za jedna (stupeň A) a tudíž neopisovala. Na tiskové konferenci řekla, že kdyby jí vedoucí nebo oponent řekli, že něco cituje špatně, samozřejmě by to přepracovala. Taková argumentace ignoruje realitu. Vedoucí práce nemůže zkontrolovat každou větu a každý odstavec oproti všem existujícím větám a odstavcům, které kdy byly napsány. Jistě, může si všimnout podezřelých vět, pokusit se je vyhledat a studenta případně upozornit, ale není jeho úkolem odhalit plagiát. Ani oponent při čtení práce nemá obvykle šanci rozpoznat, zda náhodou nějaký odstavec není opsaný odjinud.

Co je plagiátorství a jak správně citovat?

„Za plagiátorství lze považovat úmyslné kopírování cizího textu a jeho vydávání za vlastní, nedbalé nebo nepřesné citování použité literatury, opomenutí citace (byť neúmyslné) některého využitého zdroje," uvádí například Masarykova univerzita v Brně. Ostatní univerzity definují plagiáty podobně.

Správně má autor diplomové práce (nebo jakéhokoli jiného akademického textu) vždy uvést co nejpřesněji, odkud pochází informace, se kterou pracuje. Využívá se k tomu systém citací, který je upraven citační normou. Obecně to znamená, že kratší citované pasáže student oddělí uvozovkami a za ně uvede odkaz, u delších pasáží je vhodné odsadit celý citovaný odstavec.

Samozřejmě, pokud něco takového objeví, musí to nahlásit a následuje obvykle nějaká forma disciplinárního řízení se studentem, který se plagiátu dopustil. Nicméně člověk nemůže opsanou pasáž spolehlivě odhalit jinak než shodou náhod. Zato počítačový systém je na tom o něco lépe. Ale ne o moc.

Jak počítač prohledává stamiliony textových řetězců

S tím, jak se zvyšuje dostupnost textů, které může student během několika sekund zkopírovat do své práce, se rozšířil i nešvar opisování a nepřiznaných citací. Zatímco dříve mohl pedagog mít celkem dobrý přehled o tom, ze kterých knih student čerpal, dnes je ruční kontrola plagiátů stěží představitelná.

Proto začaly po roce 2000 vznikat systémy, které se specializují na odhalování plagiátů. V Česku jsou to především systémy Masarykovy univerzity: Theses.cz a Odevzdej.cz.

K čemu slouží citace?

V akademické práci je obvykle žádoucí vycházet z velkého množství již napsaných prací, článků a knih. Ale je potřeba jasně uvést, z kterých.

Moderní věda je založená na inkrementálním a systematickém rozšiřování našeho poznání a porozumění. To znamená, že žádný vědec neobjevuje věci „na zelené louce", ale vychází z toho, co již objevili a sepsali jiní. Jak řekl Isaac Newton: „Jestliže jsem dohlédl dále, bylo to proto, že jsem stál na ramenech obrů." Rozumějme: nemusel od začátku objevovat všechny matematické a fyzikální poučky, mohl je použít, kombinovat a díky tomu rozšířit.

Citace ve vědecké práci slouží k jasnému označení zdroje informace, se kterou autor pracuje. Díky tomu:

- čtenář ví, kde může hledat více informací
- autor ví, z čeho vychází a na čem staví
- případné chyby lze dohledat ke zdroji
- citovaný autor je oceněn za svůj přínos

Na tom, že někdo v diplomové práci vychází z cizích textů, tedy není vůbec nic špatného, naopak, je to žádoucí, je to podstata moderní vědy. Musí ale bezpodmínečně uvést, odkud přesně čerpá. Nejde jen o to, že by si plagiátor šetřil práci. Jde o to, že by se ztratila informace o původu zdroje.

Do systému Odevzdej.cz (který nyní používá většina univerzit v ČR) mohou svou práci nahrát i studenti před odevzdáním. Systém hledá, zda v textu nejsou textové řetězce shodné nebo téměř shodné s jinou, již nahranou prací.

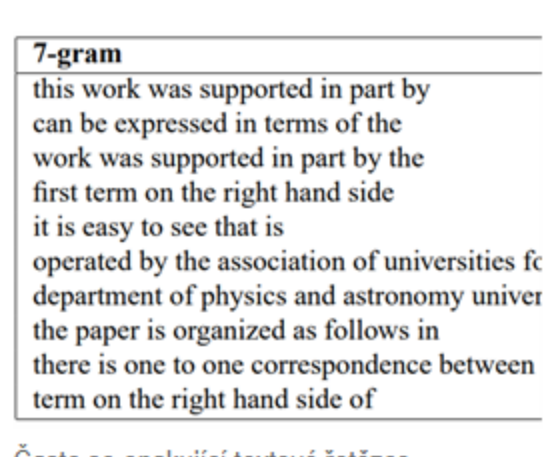
To není až tak primitivní úkol, jak se zdá. Takový nástroj v podstatě musí diplomovou práci (řekněme sto stran textu) rozsekat na dílčí řetězce (třeba 20 písmen) a tyto řetězce srovnat se stejně dlouhými řetězci v milionech stránek ostatních prací. Jinými slovy, šlo by o porovnání 180 tisíců řetězců z kontrolované práce s miliardami řetězců prací již odevzdaných. Protože je potřeba každý úryvek porovnat s každým, znamená to stovky bilionů (stovky milionů milionů) textových operací.

Proto musí autoři takového systému volit různé způsoby, jak hledání optimalizovat (podrobně se tomu věnuje Roman Chýla v článku pro Ikaros). Může například brát jenom celá slova, čímž se počet řetězců sníží. Může také odstranit ze slov diakritiku nebo dokonce samohlásky. Namísto porovnávání řetězce „Opětovným srovnáním došlo k záměně parametrů..." tak počítač srovnává jen „optvnm srvnnm dsl k zmn prmtr", což je kratší.

Existují i další způsoby, jak vyhledávání urychlit. Místo srovnávání textu lze srovnávat jen různé „otisky" textů. Z těch mohou programátoři úmyslně vyřadit některá nedůležitá slova, číslovky nebo zkrátka systém nevytváří otisk každého řetězce. Tím se sice zrychluje vyhledávání, ale stoupá riziko, že nějaká opsaná pasáž systému proklouzne.

Ne každá shoda je plagiát

Tím však problémy jenom začínají. Akademické práce obsahují velké množství okopírovaných textů, frází a dokonce celých pasáží, a přitom se nejedná o plagiát. Jde právě o citace - jasně označené pasáže textu, které jsou převzaté z již publikovaného zdroje, tento zdroj je pak jednoznačně odkázán. Nástroj, který odhaluje plagiáty, tak musí ověřit, zda je shoda „omluvena" citací, nebo zda je to shoda nepovolená, tedy taková, kde autor původ textu zatajil.


7-gram
this work was supported in part by
can be expressed in terms of the
work was supported in part by the
first term on the right hand side
it is easy to see that is
operated by the association of universities fc
department of physics and astronomy univer
the paper is organized as follows in
there is one to one correspondence between
term on the right hand side of

Často se opakující textové řetězce

Nemluvě o pasážích, které zkrátka patří k akademickému způsobu vyjadřování a vyskytují se tedy ve velkém množství textů. V angličtině jsou to slovní spojení jako „tato práce byla částečně podpořena" nebo „je zřejmé, že jde o", případně „může být vyjádřeno ve formě". Takovéto fráze samy o sobě samozřejmě nemohou být indikátorem toho, že jde o práci opsanou. Systémy proto obvykle hledají více shodných částí, případně shodu neobvykle rozsáhlou, nebo málo používaného řetězce.

„Ačkoli podobnosti může počítač vyhodnotit, o tom, zda se jedná či nejedná o plagiát, nemůže nikdy rozhodovat počítačový systém," upozorňují autoři z Masarykovy univerzity. „Finální verdikt musí vynést člověk. Systém se jen snaží pomoci člověku s tímto rozhodnutím a zvýraznit podezřelé pasáže v dokumentu."

Lidé dokážou strojovou detekci plagiátů ošálit. Zatím

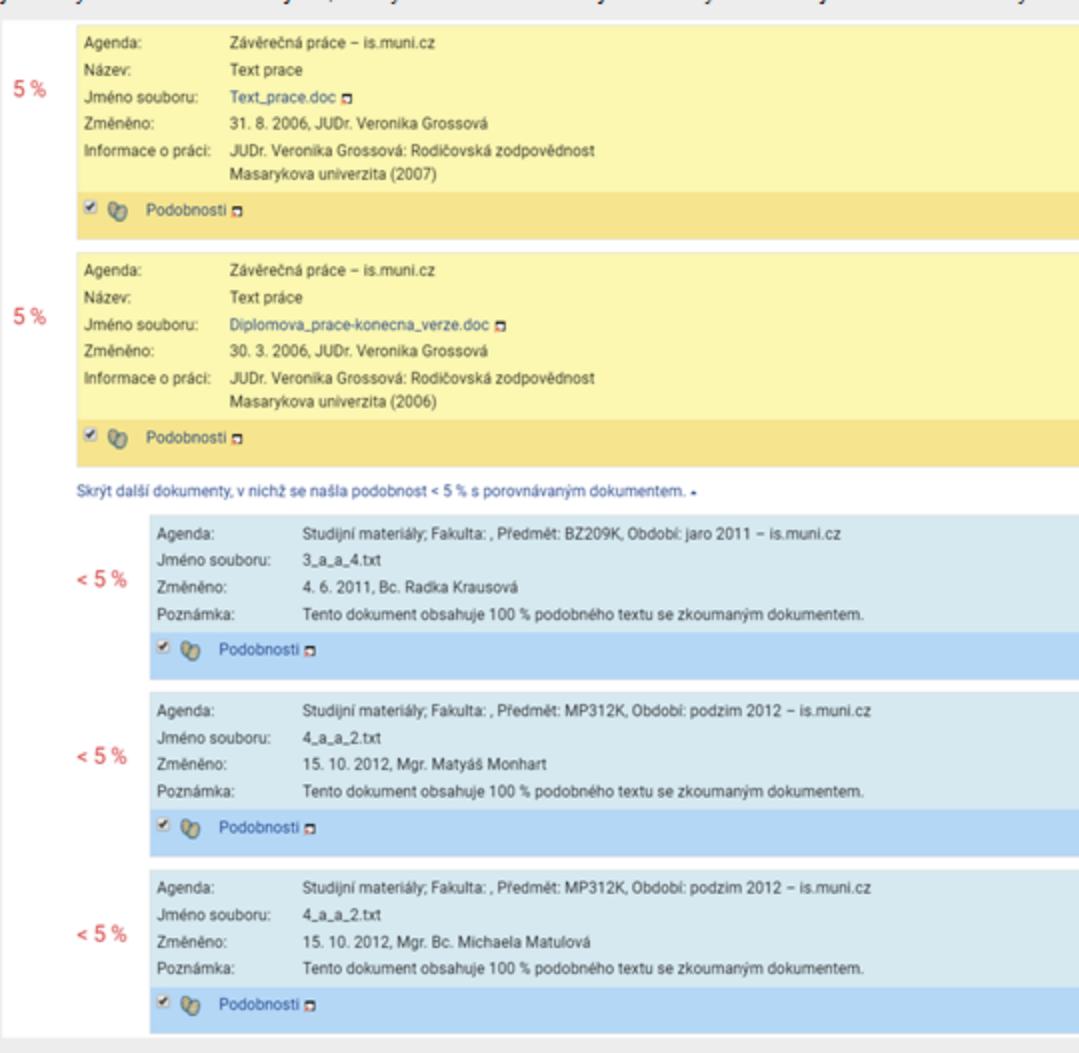
V roce 2005 nebyla zásadně odlišná pravidla pro citace v diplomových pracích. Zásadně se však od té doby proměnilo, jakým způsobem univerzity kontrolují práce. A tím se tedy změnil i přístup studentů. Dalo by se říct, že platila jiná „vnímaná pravidla", protože studenti se nemuseli bát odhalení opisování.

Dnes už by kopírování celých pasáží textu, jak to předvedla ve svých dvou diplomových pracích paní Malá, studentovi jen tak neprošlo. Její diplomová práce z roku 2011 (ke stažení ve formátu .doc) po nahrání do systému Odevzdej.cz nahlásí opsané pasáže. (Na tyto pasáže jako první z médií upozornil 28. června Český rozhlas, který neuvedl, jakým konkrétním způsobem je odhalil, nezávisle na naší redakci využil stejnou metodu i server Manipulatori.cz. Ještě dříve na podezřelé pasáže v diplomové práci upozornil na svém facebookovém profilu 26. června Libor Nováček, pozn. red.)

Opsané pasáže v diplomové práci

Systém Odevzdej.cz odhalil v diplomové práci Taťány Malé tyto pasáže opsané z jiných zdrojů.

Diplomová práce Taťány Malé (dříve Veličkové) z roku 2011 je ke stažení ve formátu .doc na stránkách její univerzity (Panevropská vysoká škola v Bratislavě). Nahráli jsme ji do systému Odevzdej.cz, který ukázal následující shody s existujícími dokumenty:


5 %
Agenda: Závěrečná práce – is.muni.cz
Název: Text prace
Jméno souboru: Text_prace.doc
Změněno: 31. 8. 2006, JUDr. Veronika Grossová
Informace o práci: JUDr. Veronika Grossová: Rodičovská zodpovědnost Masarykova univerzita (2007)
Podobnosti

5 %
Agenda: Závěrečná práce – is.muni.cz
Název: Text práce
Jméno souboru: Diplomova_prace-konecna_verze.doc
Změněno: 30. 3. 2006, JUDr. Veronika Grossová
Informace o práci: JUDr. Veronika Grossová: Rodičovská zodpovědnost Masarykova univerzita (2006)
Podobnosti

Skrýt další dokumenty, v nichž se našla podobnost < 5 % s porovnávaným dokumentem.

< 5 %
Agenda: Studijní materiály; Fakulta: , Předmět: BZ209K, Období: jaro 2011 – is.muni.cz
Jméno souboru: 3_a_a_4.txt
Změněno: 4. 6. 2011, Bc. Radka Krausová
Poznámka: Tento dokument obsahuje 100 % podobného textu se zkoumaným dokumentem.
Podobnosti

< 5 %
Agenda: Studijní materiály; Fakulta: , Předmět: MP312K, Období: podzim 2012 – is.muni.cz
Jméno souboru: 4_a_a_2.txt
Změněno: 15. 10. 2012, Mgr. Matyáš Monhart
Poznámka: Tento dokument obsahuje 100 % podobného textu se zkoumaným dokumentem.
Podobnosti

< 5 %
Agenda: Studijní materiály; Fakulta: , Předmět: MP312K, Období: podzim 2012 – is.muni.cz
Jméno souboru: 4_a_a_2.txt
Změněno: 15. 10. 2012, Mgr. Bc. Michaela Matulová
Poznámka: Tento dokument obsahuje 100 % podobného textu se zkoumaným dokumentem.
Podobnosti

Výsledek analýzy Odevzdej.cz: práce Taťány Malé (dříve Veličkové) z roku 2011 obsahuje pětiprocentní shodu s diplomovou prací Veroniky Grossové z roku 2006)


< 5 %
Agenda: Studijní materiály; Fakulta: , Předmět: MP312K, Období: podzim 2012 – is.muni.cz
Jméno souboru: 4_a_a_2.txt
Změněno: 15. 10. 2012, Mgr. David Hauser
Poznámka: Tento dokument obsahuje 100 % podobného textu se zkoumaným dokumentem.
Podobnosti

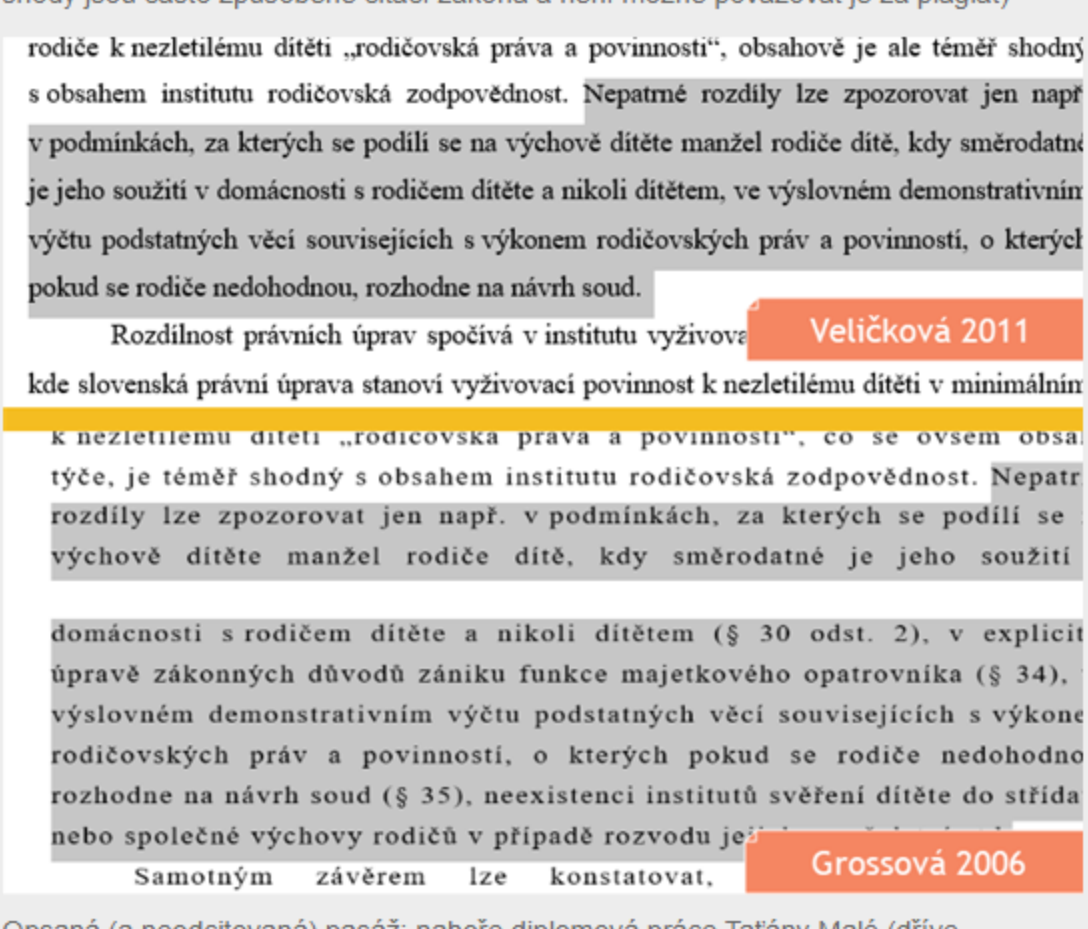
< 5 %
Agenda: Studijní materiály; Fakulta: , Předmět: MP314K, Období: podzim 2012 – is.muni.cz
Jméno souboru: 4_a_a_2.txt
Změněno: 16. 10. 2012, Mgr. Jana Havlová
Poznámka: Tento dokument obsahuje 100 % podobného textu se zkoumaným dokumentem.
Podobnosti

< 5 %
Agenda: Studijní materiály; Fakulta: , Předmět: MP314K, Období: podzim 2012 – is.muni.cz
Jméno souboru: 4_a_a_2.txt
Změněno: 16. 10. 2012, Bc. Vladimír Sokolík
Poznámka: Tento dokument obsahuje 100 % podobného textu se zkoumaným dokumentem.
Podobnosti

< 5 %
Agenda: Studijní materiály; Fakulta: , Předmět: MP312K, Období: podzim 2012 – is.muni.cz
Jméno souboru: 4_a_a_2.txt
Změněno: 16. 10. 2012, Mgr. Bc. Tomáš Straka
Poznámka: Tento dokument obsahuje 100 % podobného textu se zkoumaným dokumentem.
Podobnosti

< 5 %
Agenda: Studijní materiály; Fakulta: , Předmět: MP314K, Období: podzim 2012 – is.muni.cz
Jméno souboru: 4_a_a_2.txt
Změněno: 17. 10. 2012, Mgr. et Mgr. Tereza Hloušková
Poznámka: Tento dokument obsahuje 100 % podobného textu se zkoumaným dokumentem.
Podobnosti

Kromě toho obsahuje řadu kratších shod s různými dalšími studijními materiály (tyto shody jsou často způsobené citací zákona a není možné považovat je za plagiát)


rodiče k nezletilému dítěti „rodičovská práva a povinnosti", obsahově je ale téměř shodný s obsahem institutu rodičovská zodpovědnost. Nepatrné rozdíly lze zpozorovat jen např v podmínkách, za kterých se podílí se na výchově dítěte manžel rodiče dítě, kdy směrodatné je jeho soužití v domácnosti s rodičem dítěte a nikoli dítětem, ve výslovném demonstrativním výčtu podstatných věcí souvisejících s výkonem rodičovských práv a povinností, o kterých pokud se rodiče nedohodnou, rozhodne na návrh soud.

Rozdílnost právních úprav spočívá v institutu vyživova

kde slovenská právní úprava stanoví vyživovací povinnost k nezletilému dítěti v minimálním

Veličková 2011

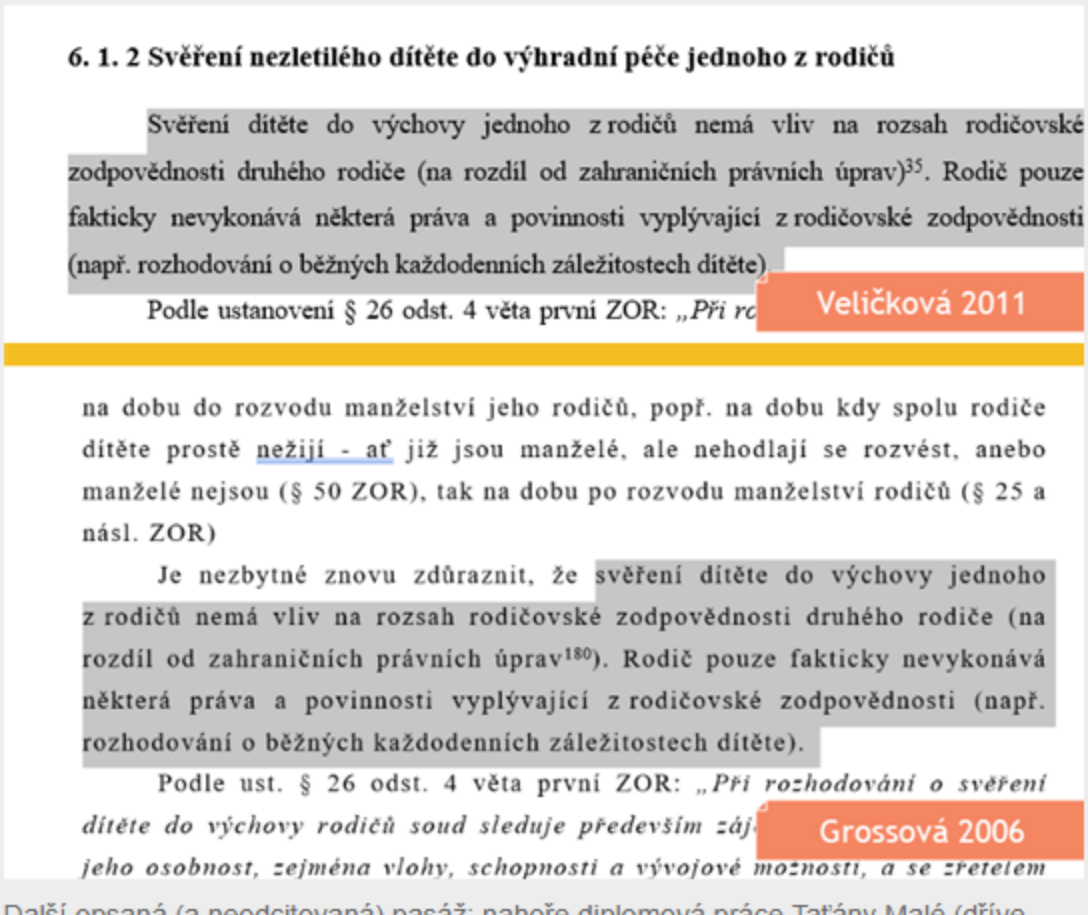
k nezletilému dítěti „rodičovská práva a povinnosti", co se ovšem obsa týče, je téměř shodný s obsahem institutu rodičovská zodpovědnost. Nepatr rozdíly lze zpozorovat jen např. v podmínkách, za kterých se podílí se výchově dítěte manžel rodiče dítě, kdy směrodatné je jeho soužití

domácnosti s rodičem dítěte a nikoli dítětem (§ 30 odst. 2), v explicit úpravě zákonných důvodů zániku funkce majetkového opatrovníka (§ 34), výslovném demonstrativním výčtu podstatných věcí souvisejících s výkone rodičovských práv a povinností, o kterých pokud se rodiče nedohodno rozhodne na návrh soud (§ 35), neexistenci institutů svěření dítěte do střída nebo společné výchovy rodičů v případě rozvodu je

Samotným závěrem lze konstatovat,

Grossová 2006

Opsaná (a neodcitovaná) pasáž: nahoře diplomová práce Taťány Malé (dříve Veličkové), dole diplomová práce Veroniky Grossové. Všimněte si gramatické chyby „se podílí se", která je v obou textech totožná, což ukazuje na kopírování textu metodou kopírovat+vložit (na tuto chybu poprvé upozornil ČRo).


6. 1. 2 Svěření nezletilého dítěte do výhradní péče jednoho z rodičů

Svěření dítěte do výchovy jednoho z rodičů nemá vliv na rozsah rodičovské zodpovědnosti druhého rodiče (na rozdíl od zahraničních právních úprav)35. Rodič pouze fakticky nevykonává některá práva a povinnosti vyplývající z rodičovské zodpovědnosti (např. rozhodování o běžných každodenních záležitostech dítěte).

Podle ustanovení § 26 odst. 4 věta první ZOR: „Při ro

Veličková 2011

na dobu do rozvodu manželství jeho rodičů, popř. na dobu kdy spolu rodiče dítěte prostě nežijí - ať již jsou manželé, ale nehodlají se rozvést, anebo manželé nejsou (§ 50 ZOR), tak na dobu po rozvodu manželství rodičů (§ 25 a násl. ZOR)

Je nezbytné znovu zdůraznit, že svěření dítěte do výchovy jednoho z rodičů nemá vliv na rozsah rodičovské zodpovědnosti druhého rodiče (na rozdíl od zahraničních právních úprav180). Rodič pouze fakticky nevykonává některá práva a povinnosti vyplývající z rodičovské zodpovědnosti (např. rozhodování o běžných každodenních záležitostech dítěte).

Podle ust. § 26 odst. 4 věta první ZOR: „Při rozhodování o svěření dítěte do výchovy rodičů soud sleduje především záj

jeho osobnost, zejména vlohy, schopnosti a vývojové možnosti, a se zřetelem

Grossová 2006

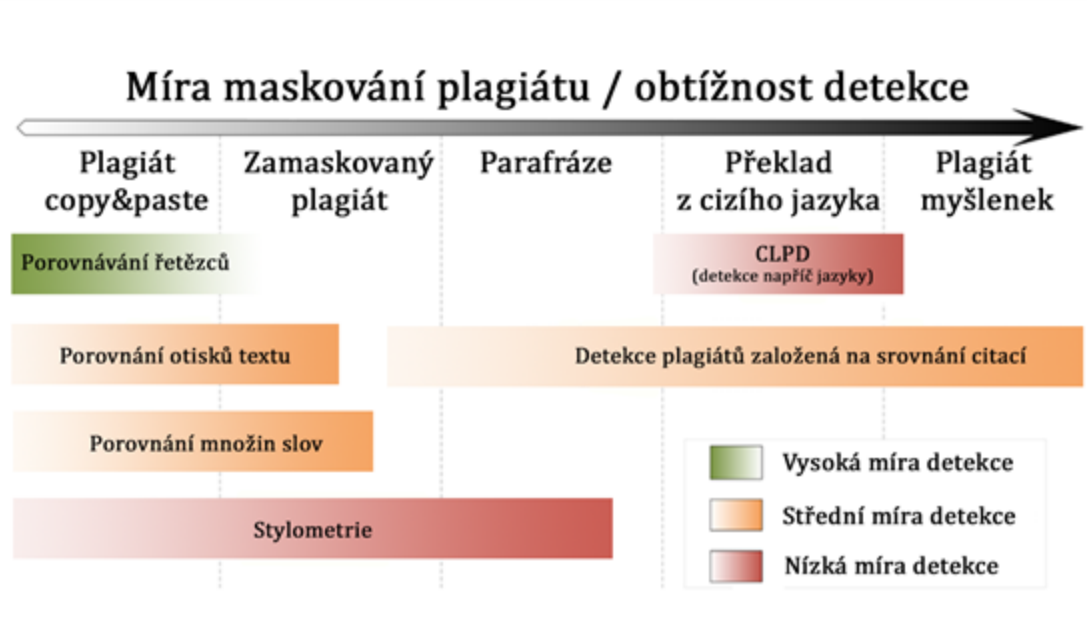
Další opsaná (a neodcitovaná) pasáž: nahoře diplomová práce Taťány Malé (dříve Veličkové), dole diplomová práce Veroniky Grossové

Toto jsou veškeré prameny, které Taťána Malá (dříve Veličková) uvedla ve své diplomové práci z roku 2011. Diplomová práce Veroniky Grossové není zde (ani jinde) zmíněna ani slovem.

Podobných pasáží je v dokumentu více (viz analýza ČRo). Lze tedy bez pochybností říci, že Taťána Malá jednoznačně využila celé pasáže textu z jiné diplomové práce (Grossová, 2006, ke stažení zde), aniž tyto pasáže oddělila uvozovkami či je jakkoli jinak označila. Nejen to, diplomovou práci Grossové dokonce Malá v celé své práci ani jednou nezmínila.

Navíc jsou v obou textech shodné gramatické chyby. Je prakticky vyloučeno, že k něčemu takovému mohlo dojít nedopatřením nebo náhodou, nejjednodušším vysvětlením je úmyslné zkopírování a nepřiznání zdroje - tedy plagiát.

Doslovné opisování už dnes studentům tak snadno neprojde. To ale neznamená, že mají plagiátoři konečnou. Pokud diplomanti chtějí opisovat dnes, samozřejmě si cesty najdou. Jen musí být kreativnější než programátoři, kteří programují nástroje na detekci plagiátů...



Metody odhalování plagiátů a jejich úspěšnost proti různým typům opisování

Metoda porovnávání textových řetězců, respektive jejich otisků, bezpečně odhalí pouze zkopírování celých pasáží (tedy ono pověstné Ctrl+C, Ctrl+V). Student by však mohl jen přeházet několik slov nebo nahradit slova nějakými synonymy (maskování plagiátu) a počítačový systém fungující na principu srovnávání textových řetězců by takový plakát nejspíše neodhalil.

Programátoři proto vyvinuli metodu, ve které porovnávají slovní zásobu jednotlivých částí dokumentů. Pokud se styl textu (například používaná slovní zásoba, složitost vět a další parametry) zásadně odlišují v rámci dokumentu, nebo naopak podezřele podobají části cizího dokumentu, může takový systém označit shodu. Tato metoda však není spolehlivá, protože je téměř nemožné vyvážit parametry takovým způsobem, aby neoznačovala jako podezřelé pasáže, u kterých to není na místě.

I takové detekci se student může vyhnout tím, že přeloží „ukradenou" část z jiného jazyka. Protože téměř jakákoli věta může být přeložena více než jedním způsobem, je téměř nemožné takové přeložené pasáže zachytit. Alespoň nyní. Do budoucna si ovšem lze představit, že systémy založené například na překladatelských službách typu Google Translate by mohly odhalit i přeložené pasáže.

Když autor okopíruje „jen" myšlenku

Okopírovat cizí text bez uvedení zdroje je očividný plagiát. Z etického hlediska se stejného prohřešku dopouští i ten, kdo z cizího textu vezme pouze myšlenky, aniž by tento text citoval jako zdroj inspirace. Právě tento jev se snaží odhalit takzvaná CBPD (Citation-based Plagiarism Detection, tedy detekce plagiátů založená na srovnávání citací). Jde o relativně novou metodu, která je navržena přímo pro odhalování plagiátů v akademických textech.

Řekněme, že student píše diplomovou práci na téma, které je v Česku relativně nové, zatímco v USA a v Japonsku už nějaký pátek frčí. Existuje tedy řada prací, které například v teoretické části pěkně shrnují současné vědecké poznání. Student by mohl tuto teoretickou část „obšlehnout", aniž by nutně šlo o kopírování nebo překlad. Prostě by jen napsal vlastními slovy totéž, co už někdo napsal.

Protože by odkazoval na stejné dokumenty, systém založený na detekci shodných citací by takovouto pasáž mohl označit za podezřelou. Protože zároveň zvýrazní „citačně příbuzné" pasáže v jiných textech, je pak na posouzení pověřené osoby, aby zhodnotila, zda jde o shodu náhodnou, nebo zda si student ušetřil práci a „opsal myšlenku" bez uvedení zdroje inspirace.

Rozpoznat plagiát nedokáže stoprocentně žádný stroj (a už vůbec ne člověk). Pokud se student rozhodne podvádět a neuvést zdroj svých informací, má velkou šanci, že mu to v době odevzdání práce projde. Zvláště když si sám může předem zdarma vyzkoušet, co systém na detekci plagiátů odhalí a co už ne.

Vzhledem k tomu, že se nástroje na detekci plagiátů neustále vylepšují, je celkem pravděpodobné, že systém může odhalit opsanou pasáž zpětně. Prakticky všechny absolventské práce se archivují v digitální podobě, takže s každým rokem se zvyšuje pravděpodobnost, že nějaký nový algoritmus plagiát zaznamená.

Jedinou obranou proti detekci plagiátů tak zřejmě zůstává nudná poctivost. Ve chvíli, kdy opsanou pasáž přiznáte a uvedete její zdroj, se z ní stává legitimní citace. Jen pak asi nemůžete tak snadno použít celé stránky cizích textů, protože to ihned působí maličko lajdácky.