Algoritmusok egynyelvű és különböző nyelvek közötti fordítások és plágiumok megtalálására

doktori (Ph.D.) disszertáció

Pataki Máté

MTA SZTAKI

Témavezető:
Prószéky Gábor, az MTA doktora

Pázmány Péter Katolikus Egyetem,
Információs Technológiai Kar,
Multidiszciplináris Műszaki Tudományok Doktori Iskola

Firenze, 2011.

Budapest, 2012.

Tartalomjegyzék

Ábrajegyzék

1. Bevezetés

A plágium nemcsak a felsőoktatásban (Unideb 2010), hanem számos más szakterületen is komoly problémákat okoz (Guttenberg 2011, Schmitt 2012, Ponta 2012). Ahogy terjednek a számítógéppel beadható dolgozatok és a diákok egyre fiatalabb korban ismerkednek meg a számítógéppel, internettel, úgy szivárog be a plagizálás a középiskolákba is. A tudományos életben is sajnos egyre gyakrabban lehet találkozni plagizált cikkekkel, gondolatokkal. A digitális könyvtárak terjedését is lassítják az illegális másolatok, mert a szerzők – nem teljesen alaptalanul – tartanak a bevételkieséstől. A könyvkiadóknál is gyakran azért ragaszkodnak a papír alapú kiadványokhoz, mert ott sokkal könnyebb az illegális másolást normál keretek közé szorítani (Szótár 2005). A cégek honlapján található tartalmakat vagy akár teljes honlapokat is egyre gyakrabban másolják le konkurens cégek (Sváby 2012, Bailey 2012), ahol esetleg a felső vezetés nem is tud erről, csak a honlapszerkesztő gondolta, hogy egyszerűsíti a saját dolgát. A legnagyobb internetes lexikon, a Wikipédia is küzd a plágiumokkal (Wikihu 2011). A Wikipédiára felkerülő anyagok bárki számára ingyenesen elérhetőek és bárki fel is tölthet tartalmat, emiatt viszont rendszeresen ellenőriznie kell az adminisztrátoroknak a tartalmakat, mert nem engedhetik meg, hogy valaki (akár jószándékból), engedély nélküli, jogvédett tartalmat tegyen fel az oldalaikra.

A plágiumkeresés ma már elképzelhetetlen számítógépes segítség nélkül. Senki sem ismerheti az összes, az adott témában megjelenő művet, cikket, diplomát, honlapot. Egy szakdolgozat esetében nem elég érezni, hogy az adott mű plágium, azt be is kell bizonyítani. Ehhez elengedhetetlen egy olyan eszköz, amely hatalmas mennyiségű anyagot rövid idő alatt át tud nézni, és meg tudja nevezni az adott dolgozathoz felhasznált forrásokat és az egyezés mértékét.

A plágiumok elleni védekezés műszaki megoldásait alapvetően két csoportba oszthatjuk, a másolás megakadályozását elősegítő eszközök (másolásvédelem), és a másolás felderítését lehetővé tevő eszközök (plágiumkeresők). Nehéz megóvni digitális tartalmat az illegális másolástól úgy, hogy közben a legális felhasználást ne nehezítse meg a rendszer, sőt egyes esetekben még azt is nehéz megoldani, hogy mindenki hozzáférhessen a tartalomhoz, az általa használt szoftverkörnyezettől függetlenül. A

legtöbb másolásvédelmi rendszer könnyen megkerülhető, így csak névleges védelmet biztosít; más rendszerek sokkal jobban védenek, körülményes a megkerülésük, de csak kiegészítő szoftverekkel, esetenként dedikált hardverrel együtt használhatóak, amit csak akkor fog installálni, megvenni a felhasználó, ha számára igazán értékes a tartalom, amelyet véd. A hátrányos helyzetűek (vakok, gyengénlátók, siketek, elavult gépet használók…) gyakran nem is képesek elérni ezeket a védett tartalmakat, így ezen eljárások bizonyos esetekben még akár jogsértőek is lehetnek (1998. évi XXVI. törvény 6.§).

A plágiumkeresés nem védi meg a tartalmat az illegális másolástól, de ha széles körben használják, követhetővé teszi a mű útját, és megakadályozhatja, hogy valaki a sajátjaként tűntesse fel azt. Ez a védelem kettős: egyrészt másolatot találva a rendszer rögtön meg is nevezi az eredeti forrást és az átfedés mértékét; másrészt, ha az ilyen rendszer létezése széles körben ismert és használata elterjedt, akkor a legtöbben nem fogják felvállalni a plagizálás kockázatát, kitéve magukat a lebukás veszélyének.

1.1. Plágium és plagizálás

> plágium: szellemi tolvajlás, más művének közlése saját név alatt, a mű alapgondolatának vagy részleteinek felhasználása a szerzőre való hivatkozás nélkül. Perbe fogták plágiumért. Bebizonyosodott, hogy novellája az első betűtől az utolsóig plágium. (Magyar Értelmező Szótár)

Két fontos rész van a fenti idézetben, az egyik, hogy a szerzőre való hivatkozás hiánya miatt válik az idézet plágiummá, a másik, hogy elég egy részletet átvenni, azaz nem kell valaki másnak a teljes művét lemásolni és sajátként prezentálni, egy rövid idézet esetében is meg kell jelölni az eredeti szerzőt. Ez utóbbit akkor is meg kell tenni, ha a szerző erre nem tart igényt, és lemondott a műről, már nincsenek jogai rajta, vagy ismeretlen, hiszen például egy diplomadolgozatban, vagy házi feladatban nem az a lényeg, hogy az elkészült munka eredménye ne sértse meg más szerzői jogait (1999. évi LXXVI. törvény), hanem az, hogy a szerző saját, önálló alkotása legyen. (PPKE 2011) Ilyen esetekben teljesen lényegtelen, hogy ki az eredeti szerző, és milyen jogai vannak a művön, egyértelműen meg kell jelölni, hogy mely részek és milyen forrásból lettek átvéve.

A plágium talán a felsőoktatásban okozza a legnagyobb gondot, ezen a területen már a legtöbb feladat, dolgozat illetve diploma digitálisan készül, és a különböző ismerősökön, közösen használt gépeken, szervereken, honlapokon keresztül terjed a diákok között. Már a középiskolákban is ismertek az előre elkészített házi feladatok, olvasónaplók, érettségi tételek, sőt külön honlapok készülnek ezek megosztására, de itt sokkal nehezebb a diákok dolga, hiszen a tanár jobb esetben pontosan ismeri őket, a korábbi teljesítményüket és stílusukat, így egy akárhonnét lemásolt dolgozat esetében igen nagy a lebukás veszélye. Ezzel szemben a felsőoktatásban több ezer diák is felveheti ugyanazt a tárgyat, a beadott munkák kijavítását minden évben változó, akár több tíz fős csoport végzi, ezért a lebukás veszélye is elenyésző.

Amennyiben ezt a gondolatot továbbvisszük, és elképzeljük, hogy adott szakterületen, az országban hány diploma születik, akkor láthatjuk, hogy nincs az a professzor, aki ezeket mind ismerhetné és észrevehetné, ha másolás történt. Anélkül, hogy valakit is megsértenénk, kijelenthetjük, hogy a diplomáknak jelentős része szakmai szempontból sajnos teljesen érdektelen, értéktelen és erről nem a diák tehet. Nincs annyi különböző téma, hogy minden diák valami érdekeset, újat tehessen le az asztalra.

Magyarországon valószínűleg a legnagyobb gondot az egymásról történő másolás okozza, de az angol és német nyelvterületeken – ahol nagyságrendekkel több tartalom található meg az interneten – a legfőbb gondot az internetes oldalakról, például a Wikipédiából másolt szövegek okozzák, és az itthoni trendek alapján hazánk is ebbe az irányba halad.

1.2. Másolásvédelem

Mielőtt rátérnénk a plágiumkereső rendszerekre nézzük meg, milyen előnyökkel rendelkeznek a másolásvédelmi rendszerek.

Mint az a nevében is benne van, megvédi a tartalmakat a másolástól. Nem állíthatjuk, hogy 100%-os védelmet nyújt, de még a gyengébb eljárások esetében is megnehezíti, és körülményessé teszi a másolást.

Nem szorosan másolásvédelmi eljárás, de a Digital Rights Management (DRM), lehetővé teszi, hogy a védelem mellett a mű útját és felhasználását is nyomon kövessék. Ez a kiadóknak pontos információt ad arról, mire is használták fel a művet, és lehetőséget arra, hogy mindenféle kiegészítő szolgáltatásokkal lássák el a

dokumentumokat, például megoldható, hogy a mű nyomtatását az eredeti licensz nem engedélyezi, és amikor ezt mégis megpróbálja a felhasználó, akkor felajánlja, hogy adott összeg befizetésével, egy percen belül már ki is nyomtathatja a művet.

Ha minden mű korlátlanul és ingyen hozzáférhető lenne az interneten, a legtöbben onnét töltenék le, és ezzel a szerzők, kiadók, forgalmazók hatalmas bevételtől esnének el. A másolásvédelemmel megnehezíthető azok dolga, akik le szeretnék másolni, vagy közzé szeretnék tenni a műveket, és ezzel többen „kényszerülnek" megvenni a műveket, azaz legális csatornákon keresztül beszerezni azokat, így a szerzők több bevételhez jutnak.

Az előnyök után most nézzük meg, hogy a másolásvédelmi eljárások használata esetén milyen hátrányokkal kell számolnunk.

Sajnos még a legegyszerűbb másolásvédelmi eljárásról is elmondható, hogy megnehezíti a legális felhasználást is, ha csak a legegyszerűbb, például PDF-fájlokban található védelemre gondolunk, már önmagában az, hogy nem sima szövegként, vagy html-formátumban tesszük közzé a művünket, gondot okozhat egyeseknek. A legtöbb számítógépen alapfelszereltségben nincs pdf olvasására képes program. A mobiltelefonos böngészés is kezd terjedni, ebben az esetben néha még lehetőség sincs ilyen kiegészítő programokat installálni. A PDF fájlok nem tördelhetőek újra a kijelzőnek megfelelő sorhosszal és betűmérettel, így vízszintes és függőleges irányban is görgetni lesz kénytelen a kis képernyőt használó olvasó. A hátrányos helyzetűeknek is gondot okozhat mindenféle kiegészítő programok installálása, ha azokat nem támogatja a böngészésüket segítő alkalmazás (pl.: felolvasóprogram).

Sajnos, nem tudja a másolásvédelem megakadályozni az illegális másolást, és ha éppen azok, akik ennek a dokumentumnak a felhasználói csoportja, könnyedén megkerülik a védelmet, akkor teljesen értelmetlen a használata, csak terhet jelent a szolgáltatónak.

Vannak olyan esetek, amikor egy jogosult személy kénytelen megkerülni a másolásvédelmet. Ilyen lehet például, amikor a valaki a saját dokumentumát pdf-formában menti el, és a program, melyet használ, alapértelmezésben bekapcsolja a másolásvédelmet. Később, ha valamiért nincs már meg az eredeti dokumentum, a felhasználó fel fogja törni ezt a védelmet, hogy hozzájusson a dokumentum tartalmához.

A 1999. évi LXXVI. törvény a szerzői jogról 95/A. paragrafusa kimondja, hogy:

```
a szabad felhasználás kedvezményezettje követelheti, hogy a
jogosult a műszaki intézkedések megkerülésével szemben a 95.
§ alapján biztosított védelem ellenére tegye lehetővé számára
a szabad felhasználást
```

Itt a 95. § a műszaki intézkedések megkerüléséről szól, azaz a másolásvédelem megkerülésének a tiltásáról. Ez a szakasz tehát azt mondja ki, hogy annak ellenére, hogy másolásvédelem van a művön, adott feltételek teljesülése esetén a felhasználók kérhetik a védelem eltávolítását (pl. szabad felhasználás bizonyos eseteiben, fogyatékos személyek jogos igényei esetén).

Nem minden esetben jogszerű a másolásvédelem használata, erre legjobb példa a szoftver, mellyel kapcsolatban az eladó nem akadályozhatja meg, hogy a termékről a vevő biztonsági másolatot készítsen saját céljára. Amennyiben valaki például tanulmányokat árul az interneten, akkor használhat másolásvédelmet, de erre fel kell hívnia a vevő figyelmét, hogy az tisztában legyen vele, hogy vásárlás után mire tudja majd használni a dokumentumot, különösen, ha a másolásvédelem megakadályozza, hogy idézeteket átemeljen a műből a sajátjába, ami legtöbb esetben jogos elvárás.

A korábban említett DRM eljárás felvet pár személyiségi jogi, adatvédelmi problémát, hiszen a legtöbb rendszer esetében az eladó pontosan tudja, hogy ki, mikor, melyik művet nézi meg, nyomtatja ki stb. Nem biztos, hogy minden felhasználó szívesen ad ki magáról ilyen információkat, pláne teljesen idegen cégeknek, ahol nincs is lehetősége befolyásolni azt, hogy ezeket az információkat ki és mire fogja felhasználni.

Főleg tudományos területen az a cél, hogy egy adott kutatás híre minél több másik kutatóhoz eljusson, és minél többen hivatkozzanak az adott cikkre, vagy eredményre. Ebben az esetben a másolásvédelem csak megakadályozza, hogy mindenki hozzáférjen a műhöz, és esetenként még azt is, hogy a webes keresők leindexeljék azt. Utóbbi igen kellemetlen, hiszen annyit jelent, hogy még ha keresi is valaki a cikkünket, akkor se fogja megtalálni például a Google-ben, mert az nem fér hozzá a tartalmához a másolásvédelem miatt.

A teljesség igénye nélkül néhány elterjedtebb másolásvédelmi eljárást érdemes közelebbről is megvizsgálnunk.

A pdf és doc formátumú fájlok esetén az Adobe illetve a Microsoft beépített valamilyen másolásvédelmet. Ezek könnyen használhatóak, és legtöbbször nem is okoznak gondot a másik félnek megnyitáskor, ugyanakkor mind a két megoldás könnyen és automatizálva megkerülhető. Egy ilyen gyenge védelmet egyébként azért is szoktak használni, hogy felhívják a felhasználók figyelmét arra, hogy ezt a dokumentumot nem szabad másolni, így később – mivel a felhasználó szándékosan megkerülte a védelmet – nem hivatkozhat arra, hogy nem tudta milyen feltételekkel használhatja az adott művet.

Léteznek olyan megoldások, amelyek csak az online megjelenítést engedélyezik. A szöveges változatok nem olyan ismertek, de hanganyagok és videók esetében már sokkal elerjedtebbek azok a műsorok, amelyeket nem lehet elmenteni, csak meghallgatni, illetve megnézni. A szöveges változataik is teljesen azonos elven működnek, és legtöbbször valamilyen kis programot kell installálni a gépre a megjelenítéshez. Ezek a megoldások erősen korlátozzák a felhasználást, és ugyan nem olyan egyszerűen, mint az előzőleg említett védelmek, de egy kis utánajárással megkerülhetőek.

Gyakori megoldás, hogy olyan nem szabványos fájlformátumot alkalmaznak a gyártók, amelyet kizárólag az ő megjelenítőjük képes feldolgozni. Hazánkban is egyre népszerűbbek az elektronikus könyvek, de csak lassan terjednek (az OSZK 2010 év végén kapott 10 darab e-könyv olvasót, melyek az érdeklődő olvasók rendelkezésére állnak), külföldön sokkal nagyobb ütemben terjednek (Amazon 2012). A legtöbb ilyen hardver ismeri a legelterjedtebb formátumú szöveges fájlokat, de a hozzá vásárolt könyvek – csak ez által a hardver által támogatott – zárt formátumban vannak. Ennek a megoldásnak a legnagyobb hátránya az, hogy az anyaghoz való hozzáféréshez rendelkeznünk kell ilyen hardverrel. Ha a hardver tönkremegy, elvesztettük a könyvtárunkat is, vagy legalábbis új, kompatibilis hardvert kell vennünk. Esetenként akár észre se vesszük, de csak kölcsönözzük a művet, így azt még csak tovább se adhatjuk. (iTunes 2012)

Gyakran használják azt a védelmet a jogtulajdonosok, hogy korlátozzák a műhöz hozzáférők körét, és ezzel próbálják meg megakadályozni, hogy az kikerüljön illetéktelenek kezébe. Ez természetesen nagyon jó megoldás, ha azok, akiket szeretnénk, hogy hozzáférjenek, nem csak hozzáférnek, de valahogy meg is találják ezeket a műveket. Ezeknek a rendszereknek általában éppen az a hátránya, hogy azok, akik

jogosultak lennének a használatára, nem is tudnak a létezéséről, vagy arról, hogy mihez is férhetnének hozzá. További hátránya, hogy ha ilyen rendszerből dokumentum kiszivárog, akkor attól kezdve nem áll már védelem alatt.

A legbiztonságosabb megoldás a fizikai védelem. Ha senki se fér hozzá a dokumentumhoz, biztos nem fogja senki se lemásolni. Ez a megoldás kicsit túlzottnak tűnik, de sajnos nagyon gyakori. A legszomorúbb példa erre az egyetemi és főiskolai diplomamunkák sorsa, amelyek ugyan elvileg hozzáférhetőek a könyvtárban, ugyanakkor nem lehet bennük keresni, és ezért lehetetlen megtalálni a több ezer diplomadolgozat között a számunkra érdekeseket. Ezek a munkák a plágiumtól való félelem miatt kerültek erre a sorsra, pedig szakmailag éppen az lenne a cél, hogy ezeket a műveket egy digitális könyvtárba rendezzék, és azon keresztül minél többen olvassák. Ideális esetben a diplomázónak át kéne futnia az összes releváns, és az adott témában született korábbi diplomadolgozatot, és például azokhoz kellene hozzáadnia valami újat, azokból kéne meríteni ötleteket, bírálni az ott felvetett gondolatokat, megerősíteni a mérési eredményeket, kiegészíteni új módszerekkel. Ha a diplomák szabadon hozzáférhetőek lennének közös, jól kereshető és használható rendszerben, és az újak is ugyanebbe a rendszerbe kerülnének vissza, akkor a plagizálás könnyen visszaszorítható lenne, ráadásul gyanú esetén a bírálók is könnyedén hozzáférnének az adott művekhez, és kézzel is összehasonlíthatnák, ha gyanúsnak találják valamelyiket. Ezzel el is értünk a plágiumkeresők által nyújtott védelem kérdésköréhez.

1.3. Plágiumkereső rendszerek

A plágiumkereső rendszereknek igen sok fajtája létezik, és legtöbbjük jól használható bizonyos területeken, ugyanakkor jelentős részükre vonatkoznak olyan megkötések, melyek miatt például digitális könyvtárak vagy egyetemi diplomák esetében nem használhatóak. Ebben a fejezetben rövid ismertetés található a fontosabb típusokról, azok előnyeiről és hátrányairól.

Sok rendszer használ vízjelet vagy valamilyen ellenőrzőösszeget a művek eredetiségének, vagy származásának a megállapítására. Az ellenőrzőösszegek jól használhatóak annak az ellenőrzésére, hogy a művet, vagy annak részeit megváltoztatták-e, illetve a mű útja jól nyomon követhető ennek segítségével. A vízjel képek és videók esetében a legelterjedtebb (Picture-shark, WaterMarks), de szöveges

dokumentumok esetében is gyakran használják (Alattar 2004, Kim 2003). Utóbbinál például a szóközök méretének a szemmel észrevehetetlen megváltoztatásával érik el a hatást, és így adott körülmények között még egy fénymásolat esetében is megállapítható, hogy honnét lett átvéve a mű. Mindkét megoldásnál a legnagyobb gondot az jelenti, hogy már egy kisebb változtatás is könnyen a védelem elvesztésével jár, és ha valaki tud arról, hogy a dokumentum ilyen védelem alatt áll, akkor könnyedén és automatizálva eltávolíthatja azt. További hátrány, hogy kisebb idézetek, részletek átvétele esetén nem használható egyik megoldás sem.

A szerző azonosítása (authorship attribution) aktívan kutatott számítógépes nyelvészeti terület. (Stamatatos 2008, Juola 2012) Ezzel a megoldással a szöveg nyelvi, nyelvtani elemzésével, a használt szavak alapján próbálják megállapítani, hogy egy adott művet ki írt, vagy két művet ugyanaz a személy írta-e. Irodalmi elemzésekben is használtak már ehhez hasonló eszközöket, egy író különböző korban írt műveinek az elemzésére, vagy adott műben a stílusok változásának a nyomon követésére. (Csernoch 2003) Sajnos ezek az algoritmusok nyelvfüggők, és ahhoz, hogy a rendszer meg tudja állapítani, hogy ki a szerző, rendelkeznie kell már megfelelő mintákkal az adott szerzőtől, ez sok esetben nem biztosítható. A módszer – jelenleg legalábbis – még nem elég megbízható ahhoz, hogy több ezer szerző dokumentumai között megfelelő biztonsággal különbséget tegyen, ugyanakkor egy művön belül ki lehet mutatni vele a stílusváltozásokat. (Juola 2006, 2012)

Léteznek olyan plágiumkereső rendszerek, amelyek nyílt keresőrendszerekre – mint például a Google – épülnek, ilyen rendszer volt például a Plagiarism Search (PSearch). A Copyscape rendszerrel egy honlap taralmát lehet megvédeni a plagizálástól (Copyscape), azaz egy honlapot megadva, ahhoz hasonlóakat, vagy azzal egyezőeket keres a neten. Belső működésére nem térnek ki részletesen az oldalon, annyi azért kiderül, hogy egy metakereső, amely Google-re épül. Hasonló elven működik a Plagiarism Check is, amely egy feltöltött szöveges dokumentumból kiemel egy véletlen mondatot, és arra rákeres a Google segítségével. (PCheck) Az internetről plagizált művek megtalálásában valószínűleg az ilyen, nyílt keresőrendszerre épülő, online szolgáltatás bizonyulhat a leghatékonyabbnak, viszont az interneten közvetlen meg nem található tartalmakban ezek a rendszerek nem képesek keresni. Ma még kevesen teszik fel diplomájukat az internetre, a könyv- és újságkiadók ritkán teszik elérhetővé a teljes

tartalmat a honlapjukon, sőt némely digitális könyvtár is csak regisztráció után érhető el, azaz a kereső nem tudja megtalálni az ott lévő tartalmakat.

Két dokumentum egymással való összehasonlítása a hasonlóságkeresés legegyszerűbb módja. A legismertebb szövegszerkesztő, a Microsoft Word is tartalmazza ezt a funkciót, és a TotalCommander nevű, széles körben használt fájlkezelő program is használható két szöveges formátumú dokumentum összehasonlítására. Kis mennyiségű, azonos nyelven írt, sok közös részt tartalmazó dokumentumok esetén ez az eljárás a leghatékonyabb, és ez adja a legpontosabb eredményt, ugyanakkor nagyobb dokumentumhalmaz elemeinek egymással való összehasonlítása nem oldható meg hatékonyan ezzel a módszerrel. Már 10 dokumentum esetén is 45 összehasonlítási műveletet kell elvégezni, ha párosával össze szeretnénk hasonlítani a műveket. Több ezer dokumentum esetén ez a módszer már egyáltalán nem használható, ugyanakkor, amennyiben egy másik, akár sokkal pontatlanabb módszerrel ki tudja szűrni a felhasználó a nagy adatbázisából azt a húsz-harminc dokumentumot, amelyek egyáltalán szóba jönnek, második lépésben egy ilyen összehasonlító és vizualizáló programot érdemes használnia a hasonlóság mértékének pontosabb megállapítása, és az eredmények megmutatása céljából.

Az előbbitől nagyon eltérő megoldást használ a Glatt Plagiarism Screening Program (GPSP), amely kérdőívet állít elő a műből olyan módon, hogy bizonyos szavakat kitöröl, és utána a szerzőnek ki kell töltenie a hiányzó részeket. A program készítői azzal a jogos feltételezéssel éltek, hogy az eredeti szerző valószínűleg legtöbb helyen ugyanazokat a szavakat használná másodszor is, míg mások nagyobb százalékban illesztenének be eltérő, rokon értelmű szavakat a hiányzók helyére. Ennek a megoldásnak az a hátránya, hogy azzal, hogy kitöltetjük a diákkal a tesztet, már meggyanúsítottuk plagizálással, ráadásul igen sok időt vesz el ez a módszer mind a tanártól, mind a diákoktól. Egyetemi környezetben esetleg használható ez a módszer, amennyiben kevés a diák, de például egy digitális könyvtárban található dokumentumról történő másolást nem fedez fel, ha azt nem diák követi el, hanem például egy tudományos cikk szerzője.

Egy viszonylag új, egyedi keresési eljárás a tudományos művekben, diplomadolgozatokban lévő hivatkozásokat használja fel arra, hogy összehasonlítsa a műveket egymással. Amennyiben két dokumentumban nagyon sok az egyező művekre

való hivatkozás, és a sorrendjük is nagyban megegyezik, akkor azt egyezésnek, plágiumnak veszi. (Gipp 2010) Ennek az eljárásnak nagy előnye, hogy akár fordítási plágiumok esetén is működik, és a hivatkozási lista is elég a plagizálás megállapításához, ami gyakran könnyebben hozzáférhető, mint maga a mű. A hátránya az, hogy kisebb egyezéseket nehezen, vagy egyáltalán nem képes megtalálni, és ha valaki ismeri a rendszer működését, akkor viszonylag kevés munkával megtéveszthetí a keresőt, például eltávolíthatja vagy kicserélheti a hivatkozásokat.

Sok olyan rendszer található az interneten, melyek belső működése teljesen ismeretlen, legtöbbször még olyan alapvető információkra sem derül fény, hogy milyen nyelvű dokumentumokhoz használható a rendszer, nem beszélve arról, hogy milyen algoritmust használ és mennyire megbízható. A Plagiarism Finder (PFind), az EVE Plagiarism Detection System (EVE) és a Turnitin (Turnitin) is mind fizetős rendszerek, de a honlapjukon alig található információ arról, hogy hogyan működnek. Utóbbiról az interneten ilyen semmitmondó információkat találunk:

> Turnitin checks for potential unoriginal content by comparing submitted papers to several databases using a proprietary algorithm. It scans its own databases, and also has licensing agreements with large academic proprietary databases.
>
> http://en.wikipedia.org/wiki/Turnitin
>
> Turnitin uses a matching algorithm that can detect an identical string of words as short as eight words that exist in the Turnitin data-base.
>
> http://www.yorku.ca/acadinte/students/turnitin-students3.htm
>
> This is done by a special algorithm of Turnitin software which uses the following sources for comparison...
>
> http://www.cc.metu.edu.tr/370-turnitin-software
>
> Specifically designed algorithms are used to create a digital fingerprint of any text. ... It's algorithms are designed to detect subtle instances of plagiarism such as: changing word order, adding sentences, or integrating an existing work with his/her work
>
> Sally Neal, Butler University

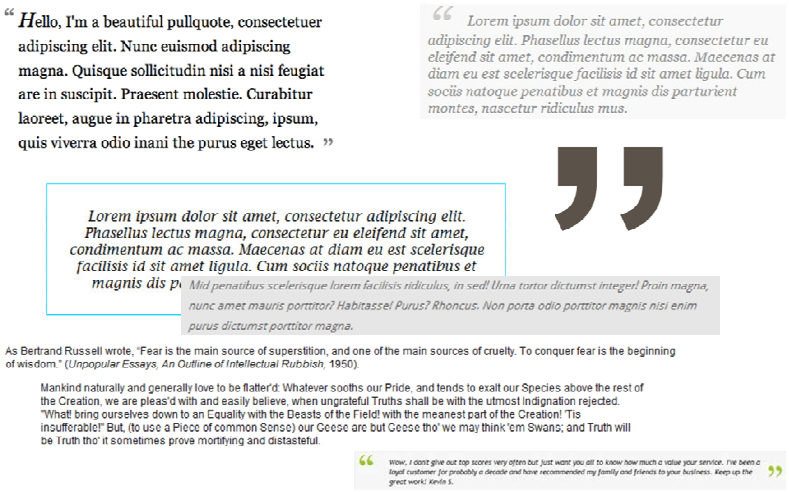
```
Turnitin.com allows the instructor to upload a paper into its
database, where software uses algorithms to create "digital
fingerprints" that has the ability to identify similar
patterns in text.

        https://helpdesk.siu.edu/index.php/CSC/Turnitin-Anti-
                                                  Plagiarism-Suite
```

Így csak következtetni lehet, hogy milyen algoritmuson alapul a működése. Sajnos ezeknél a rendszereknél nem lehet tudni, hogy milyen mértékű másolást talál meg, és a kisebbeknél még az is kérdéses, hogy mennyire lehet megbízni a készítőiben.

A legtöbb és legismertebb plágiumkereső rendszerek ma már szövegdarabolási eljárásokat használnak a plágiumok felderítésére, azaz a szöveget kisebb részekre – jellemzően párszavas darabokra – osztják majd ezek között keresnek hasonlóságot. (KOPI) Ezzel az eljárással részletesebben a következő fejezet foglalkozik.

Érdemes kiemelni – és ez az összes korábban említett szolgáltatásra is igaz –, hogy ez a rendszer nem tudja megállapítani, hogy valami idézet , vagy plágium; nem csak azért, mert olyan sokféle jelölése lehet az idézeteknek (1.1. ábra), hanem azért se, mert intézményenként változik, hogy mekkora idézetek engedélyezettek. (1.2. ábra)



1.1. ábra: Különböző idézet-stílusok

Egy plágiumkereső rendszer csak arra képes, hogy jelezze a felhasználónak, hogy az adott dokumentumban mely más dokumentumból talált meg részeket, mekkora az átfedés vagy a hasonlóság. Annak a megállapítása, hogy ez szabályos módon történt idézés-e, és helyesen meg van-e jelölve a forrás, már a felhasználóra van bízva.

1.2. ábra: A hasonlóság mértékéből még nem lehet következtetni se a mű értékére, se a plagizálás tényére

Egy olyan diplomadolgozat, amelyben egy idézet sincs, amely egyáltalán nem hasonlít egyetlen másik diplomadolgozatra se nagyon valószínűleg ugyanúgy értéktelen, mint az, amelyik csak idézetekből áll, és a szerző semmi sajátot, hasznosat, újat vagy egyedit nem tett hozzá.

1.4. Plágiumkereső mint védelem

A másolásvédelmi eljárásokhoz hasonló módon, most nézzük meg, egy plágiumkereső hogyan védheti meg az oktatási intézmények, könyvkiadók, digitális könyvtárak, konferenciaszervezők, intézmények dokumentumait az illegális másolástól.

Ha valaki másol egy plágiumkereső rendszerbe feltöltött dokumentumról, akkor a plagizálás pillanatok alatt kideríthető. Házi feladatok, diplomadolgozatok, szakmai cikkek esetén ezt a keresést automatikusan el is lehet végezni, és lehet ahhoz kötni a munka elfogadását, hogy a rendszer kiadjon egy igazolást, miszerint nem talált bizonyosnál nagyobb egyezést egyik korábbi munkával sem.

Adott egyetemi dolgozat esetén például nem elég az, ha a tanár érzi azt, hogy a mű, amit a diák beadott, nem az ő munkája, ezt valahogy igazolnia is kell. A plágiumkereső rendszer rögtön megjelöli a forrásokat, így ennek felkutatásával nem kell felesleges időt töltenie az oktatónak, sőt, olyan dokumentumokban is kereshet a rendszer, amelyhez neki nincs is hozzáférése, így meg se találhatná az egyezést.

Az előbbiek miatt a lebukás kockázata jelentősen megnő, és ez nagyon nagy visszatartó erő lehet azoknak, akik meg tudnák oldani a feladatot maguk is, csak egyszerűbb, gyorsabb utat kerestek a munka elvégzéséhez. Sajnos az is előfordul, hogy valaki mással íratja meg a házi feladatát, de ezzel is nagy kockázatot vállal. Számos esetben bukott már le úgy diák, hogy a pénzért vett dolgozatot az eredi szerző több embernek is eladta, illetve csak kisebb módosításokat végzett rajta, esetleg maga is plagizálta. A

plágiumkereső felfedheti ezeket az eseteket még akkor is, ha különböző oktatási intézményekbe került egy-egy példány a műből.

Mivel nem létezik tökéletes védelem, mindig fontos szempont az, hogy a védelem megkerülése nehezebb legyen, vagy több energiába, pénzbe kerüljön, mint annak az értéke, amit véd. Ez a védelem nem kerülhető meg automatikusan, mert legalább minden *n*-edik szót át kell írni a műben ahhoz, hogy ne ismerje fel, természetesen úgy, hogy utána is értelmes maradjon a szöveg, és ne hangozzanak erőltetettnek a mondatok. Ráadásul *n* értéke rendszerről rendszerre változhat, és az is lehet, hogy további finomításokat vezetnek be a rendszer üzemeltetői, azaz el lehet képzelni, hogy a leggyakoribb szavakat (stopword) törlik a dokumentumból darabolás előtt, a szinonimával rendelkezőket pedig a leggyakrabban használt párjukkal helyettesítik, így például hiába írja át a „szorgos” diák a *kocsi*t *autó*ra a *krumpli*t meg *pityóká*ra, a rendszer ugyanúgy meg fogja azt találni.

A legnagyobb előnye a plágiumkeresőnek a másolásvédelemmel szemben talán pont az, hogy a mű szabadon terjeszthetővé válik. Nem kell a védelem kérdésével foglalkozni, mindenki el tudja olvasni, még a speciális hardvert, vagy szoftvert használók is, valamint a webes keresővel is megtalálhatók. Mindennek eredménye, hogy többen olvassák a művet, ismertebb lesz mind a mű, mind a szerzője, illetve kiadója, és természetesen többen hivatkoznak rá, ami tudományos körökben fontos szempont.

A magyar egyetemek és főiskolák – a diákszám csökkenésének és a fejkvóták bevezetésének köszönhetően – elkezdtek versenyezni a diákok kegyeiért. Nem csak az oktatási intézménynek fontos, hogy az egyetem által kibocsátott diplomának mekkora a presztízse, hanem az oda jelentkezőknek is, hogy amikor végeznek, minél jobb esélyeik legyenek a munkaerőpiacon, azaz többen fognak jelentkezni azokba az oktatási intézményekbe, amelyek diplomái többet érnek. A plágiumkereső használatával több módon is növelni lehet az oktatási intézményekben a diplomák és dolgozatok értékét. Az első szempont az lehet, hogy elkerülhetők lesznek az olyan kínos eseteket, amikor utólag, már a diploma kiosztása, vagy a dolgozat értékelése után derül fény egy ilyen esetre. További előnye az ilyen rendszernek, hogy a diákok, éppen a lebukás veszélye miatt, sokkal ritkábban fognak plagizálni, és több energiát fektetnek a diplomába, ezzel annak a színvonala, és a diákok tudása is sokkal jobb lesz. Az jelenti valószínűleg a legnagyobb előnyt, hogy a korábbi évek munkáit ki tudják adni a diákoknak forrásként,

és nem kell tartani a tömeges plagizálástól. Így sokkal nagyobb számban születhetnek olyan diplomák, amelyek hozzátesznek valamit az előző évek munkáihoz, valami újat nyújtanak a szakmának, és nem csak megismétlik, amit már sokan leírtak az előző évben is. Lehet, hogy kicsit utópisztikusan hangzik, de az olyan digitális könyvtár használata, ahol megtalálhatóak a szakdolgozatok, kereshető formában, esetleg tematikusan rendezve, igen egyszerű formája lehet annak, hogy cégek adott területen jártas, új munkaerőre tegyenek szert, hiszen rögtön láthatnák, hogy az adott témában milyen minőségű munkát tett le az illető az asztalra. Ha valaki nagyon jó diplomamunkát írna, az se lenne kizárt, hogy mire kézbe kapja a diplomáját, már két-három állásajánlatot is kap különböző cégektől.

Az előnyei mellett természetesen – mint minden rendszernek – hátrányai, korlátai is vannak a plágiumkereső rendszereknek.

Ahhoz, hogy a védelem érvényesüljön egy nagy rendszert érdemes használnia mindenkinek, vagy pár nagyobbat, mert különben az összes rendszerben keresnie kell a felhasználónak ahhoz, hogy biztos legyen a kezébe került mű egyediségében. Ha meg valaki biztos akar lenni abban, hogy a művét nem másolják, az összes plágiumkeresőbe be kell töltenie, hogy ha éppen ott keresnek a felhasználók, akkor rátaláljanak. Természetesen egyetemi diplomák esetén már az is elég feltétel, hogy az összes, vagy a legtöbb egyetem ugyanazt a rendszert használja.

A másolásvédelem önmagában védi a dokumentumot, ahhoz, hogy egy plágiumkereső rendszer is védje, be kell tölteni a védeni kívánt dokumentumokat a rendszerbe. Ez sok dokumentum esetén, amelyek nincsenek rendezve, illetve rendszerezve, komoly feladat lehet.

1.5. A jövőbeni keresők védelme

Mielőtt a konkrét műszaki megoldásokra térnénk át, érdemes még egy dolgot megjegyezni és áttekinteni a plágiumkeresők fejlődését.

A plágium szó az 1. században keletkezet a *plagiarius* latin szóból, melynek jelentése „emberrabló, gyermekrabló”. Évszázadokig csak kézzel tudtak összehasonlítani műveket egymással, és az olvasó nagy tudásán múlott, hogy megtalálja a megfelelő művet. A 90-es években már bárki összehasonlíthatott dokumentumpárokat egymással, egy számítógép segítségével. A 90-es évek végétől a 2000-es évek elejétől jelentek meg

a dokumentumok adatbázissal való összehasonlítására képes megoldások és az első plágiumkeresők. Magyarországon az első, és mai napig egyetlen, plágiumkereső a KOPI volt, amely 2004-ben indult el. Az ehhez kifejlesztett algoritmust a 2. fejezetben ismertetjük. 2011-ben a KOPI volt az első plágiumkereső a világon, amelyik fordítási plágiumokat is képes volt keresni. Az ehhez elvégzett kutatást a 4. fejezetben mutatjuk be.

A továbbiakban ismertetett algoritmusoknak is vannak gyengeségeik, amelyeket a tudományosság követelményei szerint ismertetni is fogunk, ugyanakkor – ahogy a következőkben is látni lehet– a plágiumkereső algoritmusok folyamatosan fejlődnek, és az, hogy ma egy adott típusú plágiumot nem talál meg a rendszer, nem jelenti azt, hogy azt egy év múlva se fogja. Ma már nagyon jó algoritmusok vannak, amelyek képek eredeti forrását megtalálják (iTrace, GImage), ígéretes eredményeket értek el kutatók belső plágiumkereséshez használható algoritmusok esetében is, amelyek a stílus változásából megállapítják, hogy mely fejezetek nem illenek bele az adott műbe. (Potthast 2011) A szerző azonosítása ugyan jelenleg még csak sok minta alapján működik jól, de a számítógépes nyelvészet rohamos fejlődésével ez az irány is egyre használhatóbb és pontosabb lesz. Az meg, hogy mit hoz a jövő, senki se tudhatja, a legvalószínűbb, hogy pár éven belül szemantikusan elemezni tudja majd a számítógép a műveket, és akkor már gondolatok, ötletek plagizálását is meg tudja találni, akkor is, ha semmi komolyabb szöveges egyezés nincs a két mű között.

Bármit hoz is a jövő, a szemantikus elemzéstől a gondolatolvasásig bármi bizonyulhat egy újabb lépésnek, ezért fontos kiemelni, hogy az, hogy ma nem talál meg egy adott kereső, vagy algoritmus egy művet, az nem jelenti azt, hogy jövőre se fog. Az interneten található, „hogyan játszuk ki a plágiumkeresőt” című Youtube videók és leírások jelentős része már akkor se működött, amikor leírták, a többire pedig a megjelenése után rövid időn belül megoldást találtak az adott rendszer készítői. Minél később bukik le valaki, annál többet veszíthet. Az egyetlen biztos megoldás, hogy valaki ne bukjon le, az az, hogy nem plagizál.

2. Félig átlapolódó szavas darabolás

A ma használatos plágiumkereső algoritmusok, amely kisebb egyezést is ki tudnak mutatni – azaz nem csak teljes dokumentumokat, több oldalas egyezéseket keresnek – valamilyen daraboló eljáráson nyugszanak.

Ebben a fejezetben azt a kutatást ismertetjük, amelyik a KOPI plágiumkereső egynyelvű algoritmusának a kifejlesztéséhez vezetett (Pataki 2003), amely az ismert szavas darabolásnak egy olyan módosítása, amely lehetővé teszi, hogy a fázis-probléma ellenére használható legyen plágiumkeresésre. Az itt leírtak fontos adalékanyagot adnak majd az automatikus fordítókon alapuló többnyelvű plágiumkereső algoritmusok gyengeségeinek a megértéséhez is. Az itt ismertetett munka részletesebben leírva megtalálható az erről írt diplomamunkámban (Pataki 2002).

2.1. Darabolási eljárások ismertetése

Ahhoz, hogy értékelni tudjuk a daraboló és tömörítő eljárásokat, tudnunk kell, hogy milyen helyet foglalnak el a hasonlóságkeresés folyamatában.

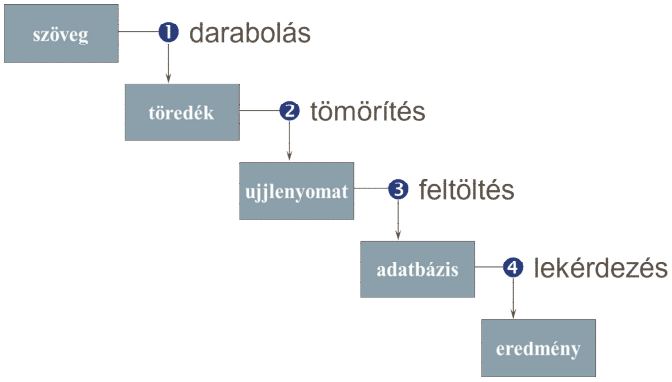
A legelső lépés egy ilyen programban a dokumentumok beszerzése. Mivel ehhez a felhasználáshoz a formázási paraméterekre nincs szükség, ezért a legegyszerűbb egy sima szövegfájl (txt) használata. Minden olyan dokumentum, amelyik nem ilyen formában található, egy ezt megelőző lépésben konvertálásra kerül.

A szövegfájlokat fel kell darabolni kisebb részekre, úgynevezett töredékekre, majd az ezt követő lépésben a töredékek eltárolásra kerülnek egy adatbázisban. Mivel ezek a töredékek sok helyet foglalnának el szöveges formában, ezért nem az eredeti töredék kerül eltárolásra, hanem annak egy úgynevezett „ujjlenyomata”. Ezt egy megfelelő tömörítő eljárással kapjuk az eredeti töredékből (pl.: MD5).

Az adatbázis feltöltése tetszőleges számú lépésben történhet, ehhez minden új dokumentumot fel kell darabolni, majd a töredékek ujjlenyomatát el kell tárolni. A lekérdezést is akármikor elvégezhetjük, akár minden újonnan beérkezett dokumentum eltárolása után is.

Ha később kíváncsiak vagyunk arra, hogy két dokumentum között van-e egyezés, csak le kell kérdeznünk az adatbázisból, hogy hány közös töredéke van ezen két dokumentumnak.

Amennyiben rendelkezésünkre állnak az eredeti dokumentumok, a felhasználó dolgát megkönnyítve, például a hasonlónak ítélt fájlokat egymás mellé téve, vizualizálhatjuk is eredményünket. A 2.1. ábra a teljes folyamatot ábrázolja.



2.1. ábra: Daraboláson alapuló plágiumkeresés lépései

2.1.1. Különböző darabolási eljárások

Az alább felsorolt darabolási eljárásokkal részletesebben Baeza-Yates et al. (1999) és Shivakumar et al. (1995, 1996) foglalkoznak.

A **szavas darabolás** (word chunking) során n darab szó kerül egy töredékbe. A szöveget úgy osztjuk fel, hogy n szavanként új töredéket kezdünk.

```
Eredeti szöveg

Ezen projekt célja, hogy a Monash University-vel együttműködve egy olyan rendszert hozzunk létre, amely hatékony a dokumentum-másolatok felderítésében.

Szavas darabolás (n=5)

ezen projekt célja hogy a monash university vel együttműködve
egy olyan rendszert hozzunk létre amely hatékony a dokumentum
másolatok felderítésében
```

2.2. ábra: Példa szavas darabolásra

Azaz az első szótól az *n*-edikig tart az első töredék, az *(n+1)*-ediktől a *2n*-edikig a második, és így tovább (lásd 2.2. ábra). Ennek az algoritmusnak van egy hátránya, ha szövegek összehasonlítására szeretnénk használni. Ha két szövegben van egyezés, de ezt az egyezést az előtte lévő tartalom miatt nem ugyanott kezdjük darabolni, akkor az eljárás nem fogja megtalálni az egyezést. Például ha egy dokumentum csak abban

különbözik egy másiktól, hogy a címe nem két, hanem három szavas, már nem tudja kimutatni az egyezést. Ezt „fázis-problémának” nevezzük. A fázis probléma kiszűrésének egyik módja az átlapolódó szavas darabolás.

Az **átlapolódó szavas darabolási** eljárás (overlapping word chunking) hasonlít a szavas daraboláshoz, azzal a különbséggel, hogy itt minden szónál kezdődik egy új töredék, amely úgyszintén *n* darab szóból áll. Ezzel az eljárással ki tudjuk szűrni a szöveg esetleges eltolódását. Ebből természetesen az is következik, hogy minden szó *n* darab töredékben lesz benne és (lásd 2.3. ábra), hogy a szavas daraboláshoz képest – ahol csak minden *n*-edik szónál kezdődött el egy darab – itt *n*-szer annyi darab keletkezik.

Eredeti szöveg

Ezen projekt célja, hogy a Monash University-vel együttműködve egy olyan rendszert hozzunk létre, amely hatékony a dokumentum-másolatok felderítésében.

Átlapolódó szavas darabolás (n=5)

ezen projekt célja hogy a
projekt célja hogy a monash
célja hogy a monash university
hogy a monash university vel
a monash university vel együttműködve
monash university vel együttműködve egy
university vel együttműködve egy olyan
stb.

2.3. ábra: Példa átlapolódó szavas darabolásra

Az átlapolódó szavas darabolás *n=1* esetében azonos a szavas darabolás n=1 beállításával, és azt eredményezi, hogy minden szó egy külön töredéket alkot.

Kézenfekvőnek tűnhet, hogy a szövegben lévő mondatok alkossák a töredékeinket, azaz **mondatonkénti darabolás**t használjunk (sentence chunking). Azonban az elsőre magától értetődő megoldás felvet néhány problémát, hiszen a mondathatár megállapítása nem olyan egyértelmű feladat. Ma már nagyon jó statisztikai alapokon nyugvó algoritmusok találhatóak ennek a feladatnak a megoldására, ezek a plágiumkeresés szempontjából már elég pontosak ahhoz, hogy jól használhatóak legyenek.

A mondatonkénti darabolásnak – a teszteltek közül egyedüliként – nincsen paramétere. Kísérletezhetünk azzal a „paraméterezési” lehetőséggel, hogy a vesszőt mondatvégnek számítjuk az egyik esetben (tagmondathatárokat veszünk), és nem vesszük figyelembe a másikban, ez viszont számos egyéb problémát vet fel, hiszen vesszőt számos egyéb okból is használunk, és könnyen módosítható egy szöveg – az értelmének a komolyabb megváltoztatása nélkül – úgy, hogy vesszőket elhagyunk, áthelyezünk, vagy újakat teszünk ki.

A **hash-kódon alapuló darabolás**i eljárás (hashed breakpoint chunking) is paraméterezhető, paraméterét jelöljük *n*-nel. Ez a darabolási eljárás, egy egyszerű és gyors függvényt (hash-függvény) használ annak megállapítására, hogy mely szavak legyenek a töredékek határai. Ehhez minden szóra kiszámítunk egy számértéket, esetünkben a szó betűinek ASCII kódjait összeadjuk. Amennyiben ez a szám maradék nélkül osztható *n*-nel akkor ez a szó töredékhatár. Az eljárásból következik, hogy amennyiben egy szó egyszer töredékhatár, akkor mindig is az lesz. A töredékhatár után álló szó lesz a következő töredék első szava, és az első olyan szó zárja le a töredéket, beleértve a kezdő szót is, amelyik értéke maradék nélkül osztható *n*-nel, azaz töredékhatár.

Eredeti szöveg

Ezen projekt célja, hogy a Monash University-vel együttműködve egy olyan rendszert hozzunk létre, amely hatékony a dokumentum-másolatok felderítésében.

Hash-kódon alapuló darabolás

ezen projekt célja hogy a monash university vel együttműködve egy olyan rendszert hozzunk létre amely hatékony a dokumentum másolatok felderítésében.

2.4. ábra: Példa hash-kódon alapuló darabolásra

Az aláhúzott szavak (2.4. ábra) esetünkben töredékhatárok. Mivel hasonlóság, illetve plágiumkeresés szempontjából nem hordoznak fontos információt a mondatkezdő vagy egyéb nagybetűk, így minden karaktert kisbetűsre változtatunk, és csak ezután számítjuk ki a szavak értékét.

A hash-kódon alapuló darabolásra is, akárcsak az átlapolódó szavas darabolásra, elmondható, hogy *n=1* esetében, azonos a szavas darabolás *n=1* beállításával, és azt eredményezi, hogy minden szó külön töredékbe kerül.

2.1.2. Keletkező töredékek mennyisége

Ahhoz, hogy két dokumentum vagy szakasz hasonlóságát meg tudjuk állapítani, kell, hogy mindkettőből keletkezzenek azonos töredékek, hiszen mi csak ezt az egyezést tudjuk később kimutatni. Minél több azonos töredék van két fájlban, annál nagyobb az esélye, hogy a feldolgozóprogram pozitívnak ítéli meg hasonlóság szempontjából. Ezért amennyiben túl kevés töredékkel dolgozunk, nagy az esélye, hogy átsiklunk bizonyos egyezések felett. Túl sok töredéket sem érdemes használni, mert az adatbázis mérete lesz egyre nagyobb, illetve túl kis töredékek esetén a hamis pozitív találatok aránya is jelentősen megnő. Az adatbázis lekérdezések sebességét is befolyásolja, vagy befolyásolhatja az adatbázis mérete. Ezért alapvető követelmény a darabolási eljárásokkal szemben, hogy lehetőleg minél kevesebb töredéket gyártsanak.

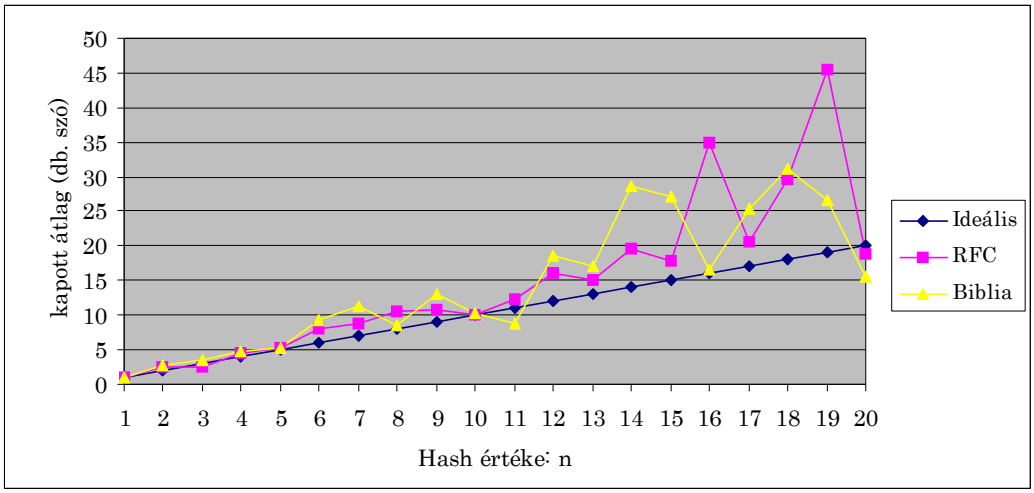
Mivel a tömörítésre használt hash-kódolás miatt az adatbázisba minden töredék azonos hosszúságú számként kerül be, lényegtelen, hogy eredetileg milyen hosszú volt, csak az számít, hogy egy adott dokumentumból melyik eljárás hány töredéket generál. A továbbiakban jelöljük *W*-vel a dokumentumban található szavak számát, valamint *n*-nel a daraboló eljárás paraméterét, *Ch*-val a töredékeket. Most vegyük sorba a különböző eljárásokat.

Az **átlapolódó szavas darabolás**, pontosan *W-(n-1)* töredéket generál, hisz, az utolsó *(n-1)* szót kivéve, minden szónál kezdődik egy töredék. Ez nagyságrendileg ugyanannyi töredéket jelent, mint amennyi a dokumentumban lévő szavak száma, hiszen míg *w* értéke több ezertől több százezerig mozoghat, addig *n* értéke általában nem haladja meg a tizenötöt. Azért, hogy később jobban átlátható legyen az összehasonlítás, a továbbiakban használjuk a *W* közelítő értéket: $|Ch| = W-(n-1) \approx W$

A **mondatonkénti darabolás** esetén elég nehéz még csak közelítő értéket is mondani. Kísérleteink során találkoztunk már olyan dokumentummal is, amiben a mondatok átlagos hossza nem haladta meg a négyet, más dokumentumok esetén viszont a tízes átlag se számít ritkaságnak. Azaz a töredékeink száma *w/4*-től *w/20*-ig akármi lehet. Sőt, szélsőséges esetben még ennél többet, illetve kevesebbet is kaphatunk. Vegyünk

átlagnak 10-et, amely magyar szövegek esetén reális, de természetesen nagyban függ a szövegtől: |Ch| ≈ W/10

Teljesen más a helyzet a **hash-kódon alapuló darabolás** esetében. Az RFC dokumentumokra (RFC), és bibliai idézetekre (Károli 1591, Békés 1951, Református 1993) a 2.5. ábrán látható átlagos töredékhosszokat kapjuk a hash-paraméter függvényében.



2.5. ábra: Töredékek átlagos hossza
hash-kódon alapú darabolás esetében, a paraméter függvényében

Mint az a diagramból is jól kivehető, a felhasznált dokumentumoktól erősen függ, az átlagos érték. Ez érthető is, hiszen elképzelhetőek olyan stílusú vagy nyelvű szövegek, ahol egy adott hash-értéken a leggyakoribb szavak pont töredékhatárt alkotnak. Persze elképzelhető az is, hogy nagyon ritkák az adott paraméteren a töredékhatár szavak, és ha ránézünk az ábrára, láthatjuk, hogy ez a gyakoribb. Természetesen, minél nagyobb a hash paramétere, annál kevesebb szó lesz töredékhatár, és annál nagyobb az esélye, hogy a gyakori szavak nem esnek bele, azaz nagyobb átlagértéket kapunk, mint a hash-értékünk. Ehhez tudni kell még, hogy *n* értéke elvileg akármilyen pozitív szám lehet, a gyakorlatban viszont 2 és 15 között mozog tipikusan. Ezeket mind egybevetve jó közelítést kapunk az egy dokumentumban található töredékek számára a *W/n* értékkel: |Ch| ≈ W/n

Ennek tudatában már látható, hogy a hash-kódon alapuló darabolás az átlapolódó szavashoz viszonyítva jelentősen kevesebb (átlagosan *n*-ed annyi) töredéket generál.

2.1.3. Daraboló-eljárások az irodalomban

Az előző fejezetekben bemutatott darabolási eljárásoknak – és az átlapolódó szavas darabolásnak kiemelten – az ezredforduló környékén még az volt a legnagyobb bajuk, hogy a bemeneti dokumentumhalmaz méretével összevethető méretű adatbázisokat generáltak.

Heintze (1996) az alábbi megoldást javasolta az átlapolódó darabolás optimalizációjára:

> We employ different size fingerprints for storage and search: the fingerprints we store in the database have size 100, but the fingerprints used for searching have size 1000. Importantly, the search fingerprint for a document is a strict superset of the fingerprint used for storage. There are two reasons for this choice. The first is reliability, and is intimately connected with design decisions discussed in Subsection 6.1. The second motivation is security: we want our system to be resilient under attack by would-be plagiarists.

Azaz egy dokumentumból összesen 1000 töredéket tartanak meg. És ezek közül is véletlenszerűen választanak ki 100 darabot, amelyek bekerülnek az index-adatbázisba. Ezzel csökkentik az adatbázis méretét, ugyanakkor biztosítják, hogy egy részletes, dokumentumpárok közötti, összehasonlítás során az egy nagyságrenddel nagyobb, 1000 dokumentum-töredéket tartalmazó, ujjlenyomattal tudjanak dolgozni.

Azért, hogy ne lehessen plágiumok tesztelésére használni az adatbázist a 100 véletlenszerűen kiválasztott töredéket adott időközönként frissítik, és másik 100-ra cserélik le, amelyet a már eltárolt 1000 töredékből választanak ki újra.

> We provide better security by periodically changing the stored fingerprint of a document. The use of two fingerprints provides a particularly convenient way to achieve this: we obtain a new stored fingerprint by simply choosing a different subset of the search fingerprint (since the ratio of sizes involved is 100:1000, this still gives considerable scope for change).

Az 1000 töredék kiválasztása ugyanakkor nem véletlenszerűen történik, hogy ugyanazt a dokumentumot mindig ugyanúgy daraboljanak. Az 1000 töredék kiválasztására példának azt javasolják, hogy az összes töredék közül az 1000 legkisebb hash értékűt tartják meg.

```
One simple strategy is random selection. However this gives poor results. For example suppose that we have a document of length 50,000 (which gives rise to about 50,000 possible substrings of length) and we use 100 substring fingerprints for storage and 1000 substring fingerprints for search. Now consider matching the document against itself. The probability that any particular substring appears in the storage fingerprint is 100/50000 = 1/500. Hence, the expected number of substrings from the search fingerprint that match the storage fingerprint is 1/500x1000 = 2 (i.e. a match ratio of about 2%). The results are of course much worse for documents that are related but not identical. To provide more reliable matches, the selection strategy must select similar substrings from similar documents. One approach is employ a string hash function, and then a fingerprint of size n can be obtained by picking the n substrings with the lowest hash values.
```

Érdemes megjegyezni, hogy ez az algoritmus ugyan biztosítja, hogy ugyanabból a dokumentumból ugyanazokat a töredékeket tartsuk meg két darabolás esetén, ugyanakkor az, hogy két – mondjuk egymással csak néhány százalékban egyező – dokumentumból mindkét esetben az egyező töredékek a kiválasztottak között lesznek, már sokkal kisebb valószínűséggel lesz igaz.

Shivakumar et al. (1995b) elemezték a használatban lévő darabolási eljárásokat, és a mondatonkénti, a hash-kódon alapuló és a szavas darabolást találták csak a gyakorlatban alkalmazhatónak az aktuális irodalmak elemzése után. Az átlapolódó szavas darabolást, főleg Brin et al. (1995) következtetései alapján, a tárolási kapacitás nagysága miatt, már nem is vették be a kutatásukba.

```
Brin et al. (1995) note that using k consecutive words as a chunking unit leads to a “phasing” problem. For instance, inserting a single word in a document shifts all subsequent chunks by one, making them not match with the original document. Brin et al. also consider using overlapping sequences of k words to avoid this problem but find the storage requirement very high. Brin et al. propose using non-overlapping sequences of words with hashed breakpoints as a compromise that avoids the phasing problem while having low storage costs.
```

A mondatonkénti darabolásról megállapították, hogy részleges mondat-egyezést nem talál meg és nehéz a mondatok határát megállapítani.

> If we break up the document into sentences, and store each of the sentences in the repository we will be able to detect document overlaps at the granularity of sentences. A simple metric to quantify the overlap between two documents can be the percentage of shared sentences. This chunking is efficient in disk I/O since only postings of documents that have a common sentence with a query document are retrieved (using our inverted indices). The main disadvantage is that we cannot detect partial sentence overlap using sentences as our chunking primitive. Another important problem is that finding sentence boundaries is a hard problem.

A szavas darabolás alatt az egy szavas darabolást értik, és kizárólag a hasonló szókincsű dokumentumok előszűrésére tartják használhatónak.

> We may also consider each word to be a chunk. Using standard IR techniques we may then denote two documents to have sufficient overlap if they share a "similar" set of words. However, similar documents do not necessarily have significant textual overlap.

A darabolási eljárások közül az átlapolódó szavas darabolás az, amelyik a legkiszámíthatóbb eredményt adja, jól paraméterezhető, nem függ a véletlentől, a mondathatárok megállapításától, ugyanakkor a helyigénye is ennek a legnagyobb. Ezért az átlapolódó szavas darabolás tárhely-optimalizációjára teszek javaslatot a következő fejezetben.

2.1.4. Félig átlapolódó szavas darabolás

A kutatás során elkészítettem egy új darabolási, illetve lekérdezési eljárást is, amelyet félig átlapolódó darabolásnak lehet nevezni. Ennek lényege, hogy az egyik dokumentumot (*q*) átlapolódó szavas darabolással, a másikat (*db*) szavas darabolással dolgozzuk fel, majd ezeket hasonlítjuk össze egymással. Ez a megoldás kiküszöböli a fázisproblémát, hiszen az egyik dokumentumból az összes lehetséges darabot előállítjuk, a másik dokumentumot viszont csak szavas darabolással daraboljuk, így lehetőségünk van vagy az adatbázis méretét (~ $|Ch_{db}|$) csökkenteni *n*-ed részre: $|Ch_{db}| = W/n$, $|Ch_q| \approx W$ vagy a keresési időt (~ $|Ch_q|$), azaz a lekérdezések számát: $|Ch_{db}| \approx W$, $|Ch_q| = W/n$

Ezen darabolási eljárás használata esetén egy szó beszúrása, törlése, illetve átírása mind-mind egy-egy hibát okoznak, azaz egy töredék fog csak módosulni, a többit a

rendszer továbbra is azonosnak fogja értékelni. Ez teljes mértékben egyezik az átlapolódó szavas darabolás esetében tapasztalttal, ahol *n*-szer ennyi töredék van, de ezek a hibák mind *n* darab töredéket érintenek, azaz mindkét eljárás esetében, ahhoz, hogy egy egyezőséget ne találjon meg a rendszer minimum minden *n* szavanként egy különbségnek kell lennie. Most nézzünk a négy leggyakoribb változtatásra egy-egy példát, alapul véve azt, hogy az adatbázist szavas darabolással építettük fel, és a dokumentumot átlapolódó szavas darabolással dolgozzuk fel. A betűk jelöljenek szavakat, a | jel a töredékhatárt, és az aláhúzott töredékek az egyezéseket.

Alapeset:

- Adatbázis: a b c | d e f | g h i | j k l
- Szöveg: a b c | b c d | c d e | d e f | e f g | f g h | g h i | h i j | i j k | j k l
- 4 töredék egyezés

Törlés (f szó):

- Adatbázis: a b c | d e **f** | g h i | j k l
- Szöveg: a b c | b c d | c d e | d e g | e g h | g h i | h i j | i j k | j k l
- 3 töredék egyezés, **1 hiba**

Beszúrás (f után, töredékhatárnál):

- Adatbázis: a b c | d e f | g h i | j k l
- Szöveg: a b c | b c d | c d e | d e f | e f **x** | f **x** g | **x** g h | g h i | h i j | i j k | j k l
- 4 töredék egyezés, **0 hiba**

Beszúrás (f elé, töredékbe):

- Adatbázis: a b c | d e **f** | g h i | j k l
- Szöveg: a b c | b c d | c d e | d e **x** | e **x** f | **x** f g | f g h | g h i | h i j | i j k | j k l
- 3 töredék egyezés, **1 hiba**

Átírás

- Adatbázis: a b c | d e **f** | g h i | j k l
- Szöveg: a b c | b c d | c d e | d e **x** | e **x** g | **x** g h | g h i | h i j | i j k | j k l

- 3 töredék egyezés, **1 hiba** (ha szinonimára írjuk át, akkor **0 hiba**, mert a mai rendszerek azt egy szónak veszik már)

Szócsere (töredéken belül)

- Adatbázis: a b c | d **e f** | g h i | j k l
- Szöveg: a b c | b c d | c d **f** | **d f e** | **f e** g | **e** g h | g h i | h i j | i j k | j k l
- 4 töredék egyezés (a szavak sorrendjét egy töredéken belül a mai rendszerek nem veszik figyelembe) , **0 hiba**

Szócsere (töredékhatáron)

- Adatbázis: a b c | d e **f** | **g** h i | j k l
- Szöveg: a b c | b c d | c d e | d e g | e g f | g f h | f h i | h i j | i j k | j k l
- 2 töredék egyezés, **2 hiba**

Jól látható, hogy egy elemi változtatás legtöbbször 1 hibát okoz, ami el is fogadható, hiszen ez azt jelenti, hogy ahhoz, hogy teljesen eltüntessük a két szöveg közti hasonlóságot ismernünk kell a paramétert, és minden n-edik töredéken változtatnunk kéne. Láthatóan bizonyos műveletek nem is okoznak hibát, és csak egy művelet van, ami két hibát okoz, mégpedig a szócsere, ha pont töredékhatárra esik. Mivel nem lehet tudni, hogy hol van a töredékek határa, és a stopszavak kiszűrése miatt ez nem is egyenletesen *n*-szavanként van, így ez nem igazán használható ki, ráadásul a sokkal gyakoribb eset – hogy nem töredékhatárra esik a csere – nem okoz egyáltalán hibát, így szavak cserélgetésével a gyakorlatban nem lehet a rendszert kijátszani.

Ezt a fentebb ismertetett félig átlapolódó szavas darabolási eljárást használja a KOPI Plágiumkereső rendszer (KOPI), amelyet a volt Informatikai és Hírközlési Minisztérium támogatásával az MTA SZTAKI Elosztott Rendszerek Osztálya (DSD), a Melbourne-i Monash Egyetemmel együtt (Monash), annak eredményeit (MDR) felhasználva végezte.

2.2. Darabolási eljárások összehasonlítása

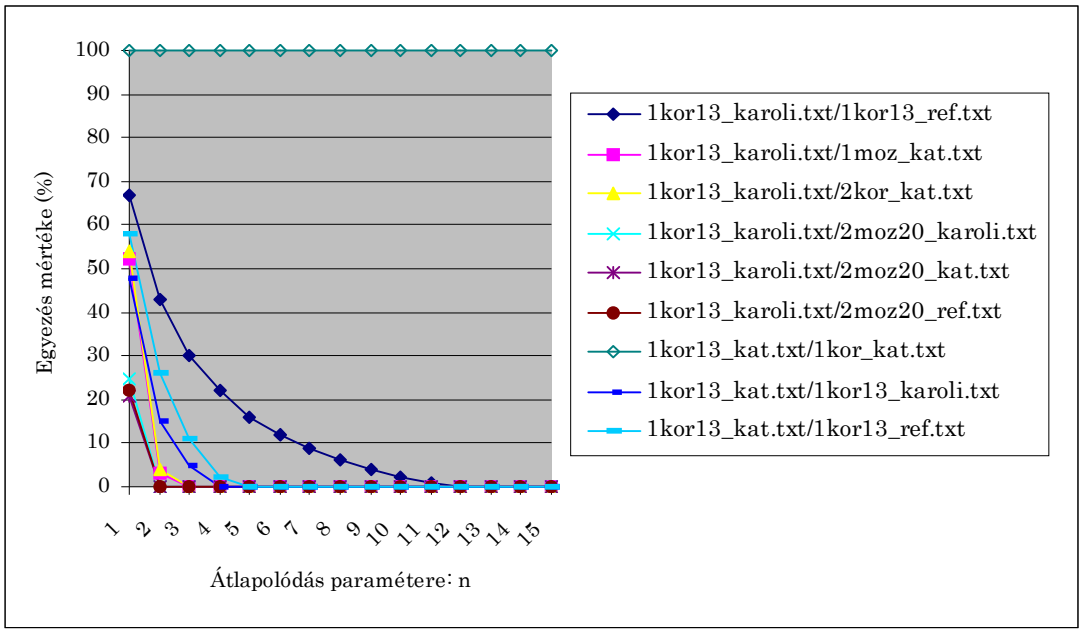
2.2.1. Hasonlóságok kimutatása

Az előző fejezetben már érintőlegesen említettük a paraméterek jelentőségét. Most nézzük meg részletesen, milyen hatással vannak a különböző darabolási eljárások paraméterei a hasonlóság kimutatására.

Azt mindenképpen meg kell említeni, hogy tulajdonképpen lényegtelen, hogy számszerűleg mekkora egyezést mutat ki egy eljárás két dokumentum között, a lényeg, hogy jól elkülöníthetőek legyenek a hasonlóságot felmutató dokumentumpárok a különbözőektől. Egy extrém példával élve, ha egy algoritmus akármely két nem hasonló dokumentum között 20 százalék egyezést talál, az még nem jelenti azt, hogy nem használható, hiszen ha a legkisebb tényleges egyezésre is felmegy ez az érték 40-re, akkor jól elkülöníthető eme két eset. Természetesen az ilyen algoritmus csak a hasonlóság meglétének kimutatására alkalmas, a hasonlóság mértéke nem állapítható meg belőle, de ez a legtöbb alkalmazásban elegendő is, hiszen mint már említettük, ezek az eljárások általában egy gyors előszűrők egy részletesebb, lassabb, de pontosabb összehasonlításhoz, felhasználói vizualizációhoz.

Az **átlapolódó szavas darabolás** paramétere, mint már korábban beláttuk, nincs hatással az adatbázis méretére és a futási időre. Viszont óriási a jelentősége a hasonlóság kimutatásánál, illetve a program aktuális problémára szabásánál.

Tesztekhez a Szent Biblia három magyar fordításával dolgoztunk, ezek: Károli Gáspár fordítása (1591), a katolikus fordítás, Szent István társulat (Békés 1951), és a református fordítás (Református 1993). Ezek tekinthetők úgy, mint egymás átiratai, habár ugyanannak a forrásnak a magyarra fordításai. Ezekből emeltünk ki azonos szakaszokat a tesztekhez, és ezeket egy-egy külön dokumentumként elmentettük. A 2.6. ábra a dokumentumok közötti hasonlóságot ábrázolja az átlapolódó szavas darabolás paraméterének függvényében.



2.6. ábra: Hasonlóság vizsgálata
átlapolódó szavas darabolás esetében, a paraméter függvényében

Mint az a grafikonból is kitűnik, az összes függvény monoton csökkenő, a 100%-os egyezést mutató mintát kivéve, amely végig megtartja az értékét. Ez a monoton csökkenés könnyen belátható. Vegyünk ehhez alapul egy *x* szavas egyezést két dokumentum között. Ebből átlapolódó szavas darabolás esetén *x-(n-1)* töredék képződik. Minél nagyobb *n* értéke, annál kevesebb töredék képződik és annál kisebb lesz az egyezés. Ez a képlet csak addig használható, amíg *n* értéke nem haladja meg *x*-ét, amennyiben meghaladja, az eredmény magától értetődően nulla lesz, azaz nem jelzi ki az egyezést. Ezt a tudást nagyon jól fel lehet használni a paraméter meghatározásához.

Ha két szöveget a stílusa illetve a szóösszetételek, kifejezések szempontjából szeretnénk összehasonlítani akkor egy 2-es, 3-as érték lesz a célravezető. Hiszen az ilyen szerkezetek általában két-három szóból állnak. Ez nagyon jól megfigyelhető a diagramon is. A sötétkékkel jelölt függvény a Szeretet himnuszának (1 kor. 13) a hasonlóságát mutatja a Károli Gáspár, illetve a református fordításban. Ezt a két változatot úgy is szokták nevezni, hogy új illetve régi fordítású református Biblia. Az egyezés mértéke *n*=2 esetében 43%, *n*=3 esetében 30% míg *n*=4 esetében már csak 22%. Ez annyit jelent, hogy a Károli Gáspár fordításban található töredékek ennyi százaléka található meg a református fordításúban is.

A 9.1 mellékletben megtalálható ez a két szövegrészlet valamint a katolikus fordítású is. Jól láthatóak az azonos mondatszerkezetek az első kettőben és az ettől nagyobb mértékben eltérőek a harmadikban.

Amennyiben az ilyen mondatszerkezeteket nem szeretnénk kimutatni nincs más dolgunk, mint, hogy nagyobb értéket adunk a paraméterünknek. Ha egy pár mondatos, folyamatos egyezést is ki szeretnénk mutatni, de nem akarjuk, hogy túl „érzékeny" legyen a keresőprogramunk a mondatszerkezetekre, akkor egy 5-ös 6-os értéket érdemes választani.

Amennyiben csak nagyobb részek egyezésének a kimutatására van szükségünk, jó választás lehet egy 10 feletti paraméter, amely a kisebb hasonlóságokat már nem mutatja ki. Így nem kell ezeknek a későbbi kiszűrésével foglalkozni.

Ezeket a megoldásokat „kombinálni" is lehet, azaz lehet kisebb paramétert választani, ugyanakkor megkövetelni, hogy több azonos töredék legyen adott távolságon belül, ahhoz, hogy egyezésnek vegyük. Ez utóbbi rendszer rugalmasabban paraméterezhető utólag, működés közben, az adatbázis újraépítése nélkül, ugyanakkor ezért a rugalmasságért az adatbázis méretének a növekedésével kell fizetni.

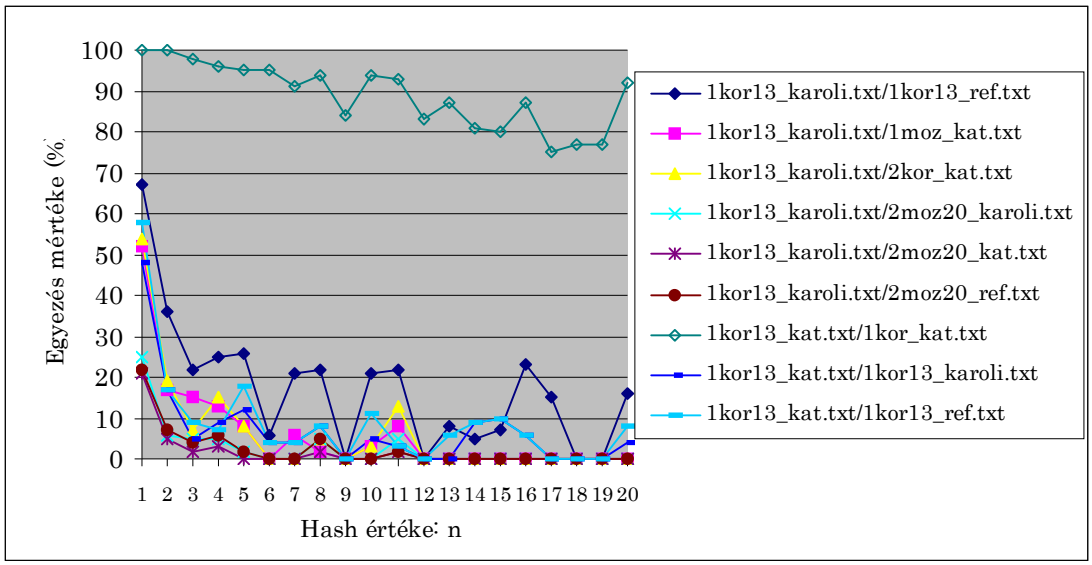
Mint már azt korábban említettük, a **mondatonkénti darabolás**nak nincsen paramétere. Még egy szempontból különleges a mondatonkénti darabolás: nem érzékeny a mondatok felcserélésére. Ez triviális, de nagyon lényeges tulajdonsága. A többi algoritmus nem veszi figyelembe a mondatok határát, ezért érzékenyek a mondatok felcserélésére.

A mondatonkénti darabolás esetén mindenképpen oda kel figyelni arra, hogy a rövidebb mondatok, gyakori kifejezések, idézetek több dokumentumban is megtalálhatóak lehetnek, és ezért könnyen ad egy-egy hamis pozitív találatot. Arra is érdemes odafigyelni, hogy szándékos plagizálásnál könnyebb minden mondaton egy picit változtatni, mint minden *n*-edik szón, így teljes egyezést keresve, szinonima szótár használata nélkül nem feltétlenül alkalmas plágiumkeresésre. Erre a következő fejezetekben külön is ki fogunk térni.

Természetesen el lehet képzelni egy olyan felhasználási területet is, ahol ez a tulajdonsága előnyére válhat. Ha például meg szeretnénk tudni, hogy mely dokumentumban idéznek rendszeresen bibliai verseket, akkor azt legegyszerűbben

mondatonkénti darabolással lehet megvalósítani. Ezeket a hasonlóságokat az átlapolódó szavas darabolással is ki tudnánk mutatni, de nem szabad elfeledkeznünk arról, hogy mennyivel több töredék keletkezne abban az esetben.

A **hash-kódon alapuló darabolás**nál nagyon érdekes diagramot kapunk, ha ábrázoljuk a fájlok között fennálló hasonlóságot a paraméter függvényében (2.7. ábra).



2.7. ábra: Hasonlóság vizsgálata
hash-kódon alapuló darabolás esetében, a paraméter függvényében

Rögtön szembe tűnik, hogy nem monotonok a függvények. Ez igen nagy problémát jelent amennyiben kisebb átfedéseket is ki akarunk mutatni. Például ha éppen olyan paramétert használunk, amelyiknél láthatóan lecsökkennek az értékek, akkor legrosszabb esetben egyáltalán nem kapunk egyezést. Miért van ez a nagy ingadozás? A válasz egyszerű, és tulajdonképpen már a 2.1.2 fejezetben, ahol a keletkező töredékek mennyiségét vizsgáltuk, meg is válaszoltuk. Érdemes összehasonlítani az ott található diagramot ezzel. Azt vehetjük észre, hogy ahol ez a diagram ugrásszerűen leesik, ott a 2.5. ábrán, annak a görbének, amelyiket a bibliai versek feldolgozásából kaptunk, és a töredékek átlagos hosszát ábrázolja, az ideálisnál magasabb értéke van. Ezek azok a helyek, amelyeken az adott stílusú vagy nyelvű szövegnek az adott paraméter mellett kevés töredékhatár szava van.

Az, hogy ez a két dolog így összefügg nagyon jól felhasználható a megfelelő paraméter megtalálásában. Amennyiben rendelkezésünkre állnak olyan stílusú dokumentumok, mint amelyekkel később az adatbázist is fel szeretnénk tölteni, akkor elég csak a

daraboló eljárást lefuttatni ezekre, és megnézni, hogy melyik paraméternél mekkora lett az átlagos töredékhossz. Azok az átlagok, amelyek jóval nagyobbak a paramétereknél, azt jelzik, hogy az adott paraméter nem alkalmas számunkra. A többiből meg az alapján érdemes válogatni, hogy mekkora hasonlóságot szeretnénk kimutatni.

Nagyon kis paraméterek esetén a töredékek száma megközelíti az átlapolódó szavas darabolását, így ezeknél az értékeknél elképzelhető, hogy érdemesebb azt használni, mivel jóval megbízhatóbban ki tudja mutatni a hasonlóságot.

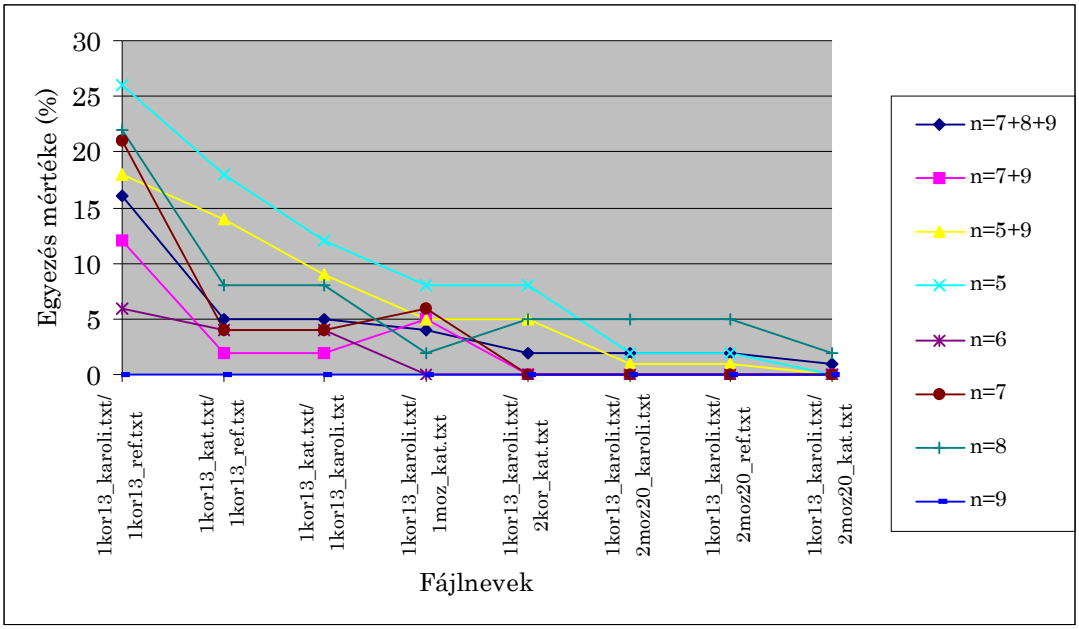
Ha kisebb szakaszok egyezését is ki szeretnénk mutatni, akkor egy 5 körüli érték lehet a legcélravezetőbb. Amennyiben csak hosszabb egyezésekre vagyunk kíváncsiak, ennél nagyobb értékeket kell választanunk. 10 feletti értékeknél viszont már nagyon vigyázni kell arra, hogy egyre nagyobb a valószínűsége annak, hogy az egyező részek pont beleesnek egy-két hosszabb (akár száz szó feletti) töredékbe.

2.2.2. Átlapolódó hash-kódon alapuló darabolás

Elképzelhető, hogy olyan adatbázist kell építenünk, amelybe különböző stílusú illetve nyelvű dokumentumok kerülnek. Például ha egy internetes plágiumkeresőt szeretnénk létrehozni, amely kijelzi az egy oldalnál nagyobb egyezéseket. Az alábbiakban eljátszunk azzal a gondolattal, hogy miként lehetne ezt megoldani hash-kódon alapuló darabolással.

A hash-kódon alapuló darabolással – mint láttuk – az a gond, hogy adott paraméter mellett, nem alkalmas minden szöveg hasonlóságának a kimutatására. Más szóval nem megbízható. Ennek a kiküszöbölésére az alábbi megoldást javasoljuk. Több hash-értéket használva egyszerre, csökkenteni lehet annak a valószínűségét, hogy az adott paraméter nem illeszkedik az adott dokumentumhoz. Ez a megoldás úgy működik, hogy a dokumentumokat kettő, maximum három hash-értékkel is feldaraboljuk, majd felöltjük az adatbázisba. Mindegyik változatot külön-külön. Ezek után, ha egy dokumentumhoz hasonlót keresünk, mindegyikre lefutatjuk a megfelelő lekérdezést, majd a kapott eredményekből átlagot számolunk, vagy a legnagyobb egyezést vesszük alapul. Ez az elgondolás a gyakorlatban is működik, és elég jó eredményeket lehet vele elérni. Ezt a módszert átlapolódó hash-kódon alapuló darabolásnak neveztem el, mert a különböző hash-értékkel kapott darabok átlapolódnak egymással.

A 2.8. ábra ezt a darabolási eljárást hasonlítja össze, a sima hash-kódon alapuló darabolással. Itt mi nem maximumot képeztünk a kapott eredményekből, hanem összeadtuk az egyező töredékek számát, majd ebből számoltuk ki az egyezés mértékét.



2.8. ábra: Átlapolódó hash-kódon alapuló darabolás és a hash-kódon alapuló darabolás összehasonlítása

Az ábrán jól megfigyelhető, hogy míg a 9-es hash-érték sehol se jelzett ki hasonlóságot, addig a mindhárom átlapolódó hash-kódon alapuló darabolási eljárás ki tudott mutatni egyezést, annak ellenére, hogy a 9-es hash-értéket mindhárom esetben használtuk. A maximumszámításnak megvan az a kockázata, hogy túl gyakran jelez egyezést dokumentumok között. Ugyan nem mutattunk külön rá, de ahogy bizonyos paraméterek túl kis, úgy bizonyosak túl nagy hasonlóságot mutatnak, mindkét eset ugyanannyira kerülendő. Ezzel a megoldással az adatbázisunkban kétszer vagy háromszor annyi töredék lesz, mint egy sima hash-kódon alapuló darabolási eljárás esetében. Ez viszont bizonyos körülmények között megengedhető, például ha egy dokumentummásolatokat megbízhatóan detektáló, de nem túl gépigényes rendszert szeretnénk építeni.

Természetesen akármelyik eljárást is alkalmazzuk, a paraméter végleges megválasztása előtt – amennyiben mód nyílik rá – érdemes kisebb mennyiségű példadokumentumra részletes teszteket futtatni. Ha mégis később, használat közben, derül ki, hogy a megválasztott paraméter, vagy eljárás nem alkalmas a feladat megoldására, az egész

adatbázist újra kell építeni. Ehhez az összes dokumentumot újra be kell szerezni, és fel kell dolgozni. Ez esetenként igen komoly feladatot és költséget jelenthet.

Mi itt most a kutatáshoz olyan tesztdokumentumokat használtunk fel, amelyeknél az egyezés jól kimutatható, akkor is, ha az egyezés mértékét százalékban fejezzük ki. Ez viszont nem mindig alkalmazható.

Vegyünk példának egy több száz oldalas könyvet, amelyben van egy több oldalas idézet egy másik, hasonlóan hosszú, dokumentumból. Ezt a hasonlóságot a program kisebbnek fogja megítélni (százalékban kifejezve), mint mondjuk két egyoldalas dokumentumban található közös szakaszt. Ezért a kiértékelésnél azt is figyelembe kell vennünk, hogy hány töredék egyezik meg a két dokumentumban.

A kutatás során a teljes bibliai tesztdokumentum-halmazra kimutattuk a hasonlóságokat, különböző darabolási eljárások és paraméterek mellett. Ezekből vettük a példákat ebben a fejezetben a diagramokhoz. A teljes táblázat megtalálható a 9.2 mellékletben.

2.3. Darabolási eljárások – új eredmények összefoglalása

Ebben a fejezetben bemutattam, két új darabolási eljárást, melyeket egynyelvű plágiumkereséshez lehet használni. A **félig átlapolódó szavas darabolás** a szavas darabolás és az átlapolódó darabolás egyesítése plágiumkeresési célokra, amiről bebizonyítottam, hogy ugyanolyan hatékony a hasonlóságok felismerésében, mint az átlapolódó szavas darabolás, ugyanakkor implementációtól függően vagy n-ed akkora adatbázist igényel, vagy a lekérdezési idő csökken n-ed akkorára (ahol n a szavas darabolás paramétere). Az **átlapolódó hash-kódon alapuló darabolás**ról bebizonyítottam, hogy segítségével kiküszöbölhető a hash-kódon alapuló darabolás szövegfüggősége, és így a gyakorlatban is alkalmassá válik akár ismeretlen szövegek darabolására is.

3. Többnyelvű dokumentum nyelvének megállapítása

A 4. fejezetben a különböző nyelveken írt szövegek összehasonlítására készült algoritmust mutatjuk be, ehhez viszont tudnunk kell, hogy egy adott szöveg milyen nyelven íródott. Mivel a tesztek során gyakran találkoztam vegyes nyelvű dokumentumokkal, így a nyelvdetekcióra használt közismert n-gram algoritmust továbbgondolva egy továbbfejlesztett algoritmust dolgoztam ki. Ez a fejezet ezt a munkát mutatja be.

3.1. Bevezetés

Egy digitális, természetes nyelven íródott dokumentum nyelvének megállapítására számos lehetőség van, és a szakma ezt a problémát nagyrészt megoldottnak tekinti (Cavnar 1994, Řehůřek 2009, Benedetto 2002), ugyanakkor a dokumentum nyelvének megállapítása nem mindig egyértelmű feladat.

A leggyakrabban használt algoritmusok igen jól működnek tesztdokumentumokon vagy jó minőségű, gondosan elkészített gyűjteményeken, ha lehet róluk tudni, hogy egy nyelven íródtak. Nekünk azonban szükségünk volt egy olyan algoritmusra, amely internetről letöltött dokumentumokon is jól – gyorsan és megbízhatóan – működik. A plágiumkereső program interneten talált, megbízhatatlan eredetű, gyakran hibás dokumentumokat dolgoz fel, és ennek során lényeges, hogy a dokumentum nyelvét, illetve főbb nyelveit, megfelelően ismerje fel, azaz többnyelvű dokumentumok esetében is megbízhatóan működjön.

Ennek a feladatnak a megoldásával mások is próbálkoztak, különböző megoldásokat ajánlva a többnyelvű dokumentumok nyelvének, nyelveinek a megállapítására. Martins et al. (2005) azt javasolja weblapok esetében, hogy amennyiben a nyelv megállapítása nem egyértelmű, akkor a teljes lap helyett a legnagyobb összefüggő blokk nyelvét állapítsuk meg, és ezt használjuk a továbbiakban.

```
Finally, multilingual documents are also very common on the
Web. This constitutes a problem if we wish to assign the full
content of a given document to one single language. For
instance homepages for people with foreign names are very
common over the Portuguese Web, at least inside large
institutional sites. To better handle this, we use a simple
heuristic: when a document is neither classified as
```

> Portuguese or English (the two most common languages in our document collection) the algorithm is very likely making a mistake. In these cases, we try to re-apply the n-gram algorithm, weighting the largest continuous text block in the document (blocks are identified in the HTML parsing stage, taking in account the markup information) as three times more important than the normal text. The rationale for this is that the longest block will very likely correspond to a good description of the page, possibly in its main language.

Ez jól működhet abban az esetben, ha csak az első nyelvre vagyunk kíváncsiak, ha viszont a második nyelv is érdekel minket, akkor már ez a megoldás nem alkalmazható.

Prager (1999) a dokumentumokban előforduló betű 2-5 grammok és szavak alapján a nyelveket egy vektortérben ábrázolja és a nyelvek közötti hasonlóságot a vektorok által bezárt szög alapján állapítja meg. Ezek után a többnyelvű dokumentumok nyelveit, illetve azoknak az arányát úgy állapítja meg, hogy előállít egy virtuális, kevert nyelvet a vektorokból, és az ezzel való hasonlóságot nézi meg.

> Now, in the case of bilingual documents, if the component proportions were known, it would be straightforward to create a feature vector from the appropriate weighted mean of the components - this is what we call a virtual mixed language - and measure the cosine of the angle between it and the document. However, we don't know the proportion in advance. In the following, we will denote the two languages being examined as F_i and F_j and we will use f_i and f_j to denote their respective feature vectors. A document D, which is in a mixture of F_i and F_j, is modelled as a vector d which approximates $k = \alpha f_i + (1-\alpha) f_j$ for some mixing weights α and $1-\alpha$ to be determined. k represents the virtual mixed language K (which is, of course, not in $\{F_i\}$). The problem, then, is to find the k (in other words the i, j and α) that minimizes the angle between k and d.

Azt, hogy a két konkrét nyelv (*i* és *j*) melyek legyenek, az egynyelvű keresés alapján állapítja meg, és a legközelebbi *m* nyelv minden lehetséges variációját kipróbálja. Az *m* értékének javasolt 5-ös paraméter esetében például 5x4 azaz 20 nyelvpárt kell végigpróbálni, és ezek után a kapott eredményt összehasonlítani az egynyelvű kereséssel. Amennyiben valamelyik nyelvpár közelebb van a dokumentum vektorához mint a legjobb egynyelvű találat, akkor a dokumentum kétnyelvű, ha nem, akkor egynyelvű.

Xafopoulos et al. (2004) Rejtett Markov Modellt (HMM) használnak a dokumentum nyelvének a megállapítására, és a többnyelvű dokumentumok esetében annyiban változtatták csak meg a modellt, hogy nem a teljes dokumentum esetén dönt a nyelvről, hanem a dokumentum beolvasása folyamán adott karakterenként hoz egy döntést.

```
An application area which is successfully confronted by the
proposed technique, as opposed to other standard techniques
used for TLI, is language tracking. That is, in the presence
of multilingual documents, the identifier manages to pinpoint
the location in the document, where a transition from one
language to another occurs, apart from determining the
identity of the language, thus presenting a flexibility with
respect to transition from one language to another. This is
achieved in a manner similar to the one used in speech
recognition, where the detection of the end of one word and
the beginning of the next one is performed. That is, instead
of selecting a simple DHMM for the description of the entire
document, a sequence of DHMMs is selected. The latter can be
considered a network of DHMMs at a higher level than the
network formed inside each DHMM with respect to its states. A
decision for the language is made every as many characters as
the number of standard states of each DHMM.
```

Ennek a megoldásnak az előnye, hogy azt is megadja, hogy hol vált nyelvet a dokumentum, amennyiben erre szükségünk van. Hátránya, hogy például egy szótárszerűen felépített, gyakori nyelvváltásokat tartalmazó, többnyelvű dokumentumot nem fog tudni helyesen felismerni.

Nyelvfelismerésre a leggyakrabban használt algoritmus az n-gram algoritmus, a fent ismertetett algoritmusok is részben vagy egészében erre épülnek. Ezért döntöttünk úgy, hogy ezt módosítjuk, hogy amennyiben egy dokumentumban nagyobb mennyiségben található más nyelvű szöveg, akkor azt közvetlen jelezze, és így a plágiumkereső rendszer ezt, mint többnyelvű dokumentumot, tudja kezelni.

Az algoritmussal szemben az alábbi elvárásokat támasztottuk:

1. jelezze, ha a dokumentum több nyelven íródott, és nevezze meg a nyelveket,
2. az algoritmus gyors legyen,
3. a szöveget csak egyszer kelljen végigolvasni,
4. ne szótár alapú legyen (kódolási és betanítási problémák miatt).

Az n-gram algoritmust (Cavnar 1994, Dunning 1994) használva csak egyszer kell végigolvasni a dokumentumot és az n-gram statisztikákból meg lehet állapítani, hogy a dokumentum milyen nyelven íródott, valamint – ha vannak megfelelő mintáink – még a kódolását is meg tudja határozni. Ez az algoritmus ugyanaz, mint a korábban ismertetett – plágiumkeresésre alkalmazott – n-gram algoritmus, azzal az különbséggel, hogy itt karakterenként darabolunk, nem szavanként.

A szakirodalomban található n-gram algoritmus viszont nem teljesíti az első feltételt, miszerint a több nyelven íródott dokumentumokat is fel kell ismernie. Ugyan elméletileg elképzelhető lenne, hogy a dokumentumot szakaszokra osztjuk, és szakaszonként állapítjuk meg a dokumentum nyelvét, de ez a megoldás sajnos két esetben is hibás eredményre vezet. Gyakran találkoztunk a plágiumkereső üzemeltetése során olyan dokumentumokkal, amelyek úgy voltak felépítve, mint egy szótár, azaz a két nyelv nem szakaszonként, hanem mondatonként – sőt egyes esetekben szavanként – váltakozott. A másik probléma akkor jelentkezett, amikor a dokumentum – például egy korábbi hibás konverzió miatt – tartalmazott HTML vagy XML elemeket, amelyek miatt rövid dokumentumok esetében hibásan angol nyelvűnek találta az algoritmus azokat.

Ezek kiküszöbölésére kezdtük el továbbfejleszteni az n-gram algoritmust, amely alapból csak arra alkalmas, hogy a dokumentumban leggyakrabban használt nyelvet megállapítsa, de a második leggyakoribb nyelv már nem a második a listában. Ennek oka, hogy a nyelvek hasonlítanak egymásra, és például egy nagyrészt olasz nyelvű dokumentum esetében a spanyol nyelv akkor is nagyobb értéket kap, mint a magyar, ha a dokumentum egy része magyar nyelven íródott.

Az új algoritmusba (Pataki et al. 2011) ezért beépítettünk egy nyelvek közötti hasonlósági metrikát, amelyet a hamis találatok értékének a csökkentésére használunk. A metrika segítségével meg lehet állapítani, hogy a második, harmadik... találatok valódiak-e, vagy csak két nyelv hasonlóságából fakadnak.

A tesztekhez az *Emberi jogok egyetemes nyilatkozatá*nak (UDHR) a különböző nyelvű fordításaiból építettünk nyelvi modelleket, így egészen ritka nyelvek és nyelvjárások is bekerültek a tesztbe.

3.2. Az eredeti n-gram algoritmus

Az n-gram algoritmus működése igen egyszerű, legenerálja egy nyelvnek a leggyakoribb „betű n-gramjait”, azaz a például 1, 2, 3 betű hosszú részeit a szövegnek, majd ezeket az előfordulási gyakoriságuk szerint teszi sorba. A magyar nyelvben ez a 80 leggyakoribb n-gram az általunk használt tesztszövegben (_ a szóköz jele):

1. _	12. o	23. t_	34. _	43. eg	54. _s	65. es	76. tá
2. e	13. á	24. sz	m	44. p	55. al	66. ő	77. c
3. a	14. é	25. el	35. _a	45. _e	56. ta	67. y_	78. re
4. t	15. g	26. ,	_	46. u	57. í	68. z_	79. to
5. s	16. m	27. ,_	36. en	47. le	58. _h	69. tt	80. A
6. l	17. y	28. h	37. ö	48. ó	59. _t	70. ke	
7. n	18. _a	29. k_	38. n_	49. er	60. an	71. _v	
8. k	19. b	30. .	39. _k	50. f	61. ze	72. ás	
9. i	20. d	31. et	40. j	51. ek	62. me	73. ak	
10. r	21. a_	32. gy	41. ._	52. te	63. at	74. _é	
11. z	22. v	33. s_	42. i_	53. és	64. l_	75. ny	

3.1. táblázat: 80 leggyakoribb betű n-gram egy magyar szövegben

Két szöveg összehasonlítása úgy történik, hogy a két n-gram listán összeadjuk az azonos n-gramok helyezéseinek a különbségét, és ez adja a két dokumentum közötti hasonlóság mértékét. Két azonos nyelven írt dokumentum között alig, míg különböző nyelvek között szignifikáns lesz a különbség. Ezért használható ez az algoritmus a dokumentum nyelvének megállapítására.

Példának nézzük meg az angol nyelvű példadokumentumunk első 10 n-gramját, és hasonlítsuk össze a magyarral.

1. _ (1-1)
2. e (2-2)
3. t (3-4)
4. (4-12)
5. n (5-7)
6. i (6-9)
7. a (7-3)

8. s (8-5)
9. r (9-10)
10. h (10-28)

Az eredmény 0+0+1+8+2+3+4+3+1+18 = 40. Ez a különbség egyre nagyobb lesz, ahogy lejjebb megyünk a listában. Mivel nem lehet végtelen hosszú n-gram listát készíteni – és nincs is szükség rá –, mi egy 400-as listával dolgozunk a tesztek során. Az első 400 n-gramot tároljuk el minden nyelvhez, így azokat az n-gramokat, amelyek az egyik listában szerepelnek, de a másikban nem úgy vesszük figyelembe, mintha 400 távolságra lennének.

Ennek megfelelően a két nyelv elméleti minimális távolsága 0, maximális távolsága (r_{max}) pedig – amennyiben a két listának egyetlen közös elemes sincsen – 400^2 azaz 160 000. Ebből a százalékos hasonlóságot a

$$h_{százalékos} = (r_{max} - r)/(r_{max}/100)$$

összefüggéssel kapjuk.

Példának nézzük meg, hogy mekkora hasonlóságot mutatnak különböző nyelvű dokumentumok a mintadokumentumainkhoz képest. Az egyszerűbb olvashatóság érdekében $h_{százalékos}$ értékekkel számolva a különböző nyelvű Szegedről szóló Wikipédia szócikkekre. (SzegedHu, SzegedEn, SzegedDe, SzegedIt, SzegedFr)

A magyar nyelvű szócikk esetén az alábbi eredményt kapjuk, az első 5 találatot kérve: 1. magyar: 35.49, 2. breton: 27.70, 3. szlovák: 27.42, 4. eszperantó: 26.98, 5. közép-fríz: 26.79

Az angol nyelvű szócikk esetén az alábbi eredményt kapjuk: 1. angol: 44.37, 2. skót: 35.67, 3. romans: 35.34, 4. német: 33.74, 5. román: 33.73

A német nyelvű szócikk esetén az alábbi eredményt kapjuk: 1. német: 57.13, 2. holland: 38.15, 3. közép-fríz: 37.71, 4. dán: 37.48, 5. fríz: 36.58

Az olasz nyelvű szócikk esetén az alábbi eredményt kapjuk: 1. olasz: 35.21, 2. román: 33.95, 3. katalán: 33.46, 4. spanyol: 32.18, 5. romans: 31.78

Jól látható az eredményekből, hogy a rokon nyelvek esetében magas hasonlóságot mutat a dokumentum, azaz egy olasz nyelvű dokumentum majdnem ugyanannyi pontot kap az olaszra, mint a spanyolra.

Most nézzük meg, hogy kétnyelvű, 50-50 százalékban kevert dokumentumokra mit kapunk.

Egy magyar-angol nyelvű dokumentum esetén az alábbi eredményt kapjuk: 1. angol: 40.80, 2. magyar: 39.45, 3. skót: 38.41, 4. afrikaans: 34.69, 5. közép-fríz: 34.19

Egy magyar-olasz nyelvű dokumentum esetén az alábbi eredményt kapjuk: 1. olasz: 49.56, 2. romans: 45.25, 3. katalán: 41.60, 4. latin: 41.26, 5. román: 41.18, …, 10. magyar: 38.02

Egy magyar-francia nyelvű dokumentum esetén az alábbi eredményt kapjuk: 1. francia: 38.16, 2. katalán: 36.74, 3. eszperantó: 34.26, 4. spanyol: 34.08, 5. romans: 33.71, …, 7. magyar: 33.2

Egy angol-német nyelvű dokumentum esetén az alábbi eredményt kapjuk: 1. német: 53.47, 2. angol: 44.14, 3. fríz: 40.98, 4. közép-fríz: 40.61, 5. holland: 40.08

Látható, hogy a magyar-olasz ill. magyar-francia kevert szövegben a magyar nyelv bele se került az első 5 találatba. Végül nézzük meg, hogy egy háromnyelvű, harmadolt arányban kevert dokumentumra mit kapunk.

Egy magyar-angol-olasz nyelvű dokumentum esetén az alábbi eredményt kapjuk: 1. angol: 46.55, 2. olasz: 44.55, 3. romans: 43.58, 4. katalán: 42.41, 5. román: 41.11, …, 10. magyar: 38.26

Háromnyelvű szövegben sem kerül be az első öt helyre a magyar nyelv.

3.3. Továbbfejlesztett n-gram algoritmus

Mint láttuk, bizonyos nyelvek hasonlítanak egymásra az n-gram algoritmus szempontjából, így egy többnyelvű dokumentum esetén a második helyen nem minden esetben a dokumentum második nyelvét találjuk, ráadásul az se derül ki, hogy a második nyelv azért került oda, mert valóban szerepel a dokumentumban, vagy azért, mert hasonlít az első nyelvre. Ezért az új algoritmusunkban elkezdtük kiszámolni a nyelvek közötti hasonlóságot, méghozzá a nyelvfelismeréshez használt n-gram minták közötti hasonlóságot. A távolságok tipikus értékeire nézzünk néhány esetet.

A magyar nyelvhez n-gram alapján legközelebb álló nyelvek távolság-értékei: 1. breton: 104 541, 2. közép-fríz: 104 751, 3. svéd: 106 068, 4. eszperantó: 106 469, 5. afrikaans: 106 515

Az angol nyelvhez n-gram alapján legközelebb állók: 1. skót: 85 793, 2. francia: 88 953, 3. katalán: 89 818, 4. latin: 90 276, 5. romans: 92 936

Végül az olasz nyelvhez n-gram alapján legközelebb állók: 1. romans: 79 461, 2. román: 85 232, 3. katalán: 85 621, 4. spanyol: 86 138, 5. latin: 86 247

Számos algoritmussal próbálkoztunk, melyek közül az alább leírt bizonyult a legmegbízhatóbbnak.

Egy *D* dokumentumra kapott százalékos hasonlóság () a százalékos hasonlóság mértékének növekvő sorrendjében legyen: h_1, h_2, h_3 stb., a nyelveket jelölje L_1, L_2, L_3, azaz a h_1 a *D* dokumentum hasonlóságát mutatja az L_1 nyelvű mintánkkal, százalékban. A nyelvek közötti százalékos hasonlóságot pedig jelöljük h^{LiLk}-val. h_i' legyen az új algoritmus által az L_i nyelvre adott érték.

$$h_i' = h_i \quad \text{ha } i=1$$

$$h_i' = h_i - \frac{\sum_{k=1}^{i-1} h_k \times h^{LiLk}}{\sum_{k=1}^{i-1} h_i} \quad \text{ha } i>1$$

Az algoritmus tulajdonképpen minden nyelv valószínűségét csökkenti az előtte megtalált nyelvek valószínűségével, így kompenzálva a nyelvek közötti hasonlóságból adódó torzulást. Nézzük meg, hogy mekkora hasonlóságot mutatnak különböző nyelvű dokumentumok a mintadokumentumainkhoz képest ezzel az új algoritmussal számolva.

Egy magyar nyelvű dokumentum (Szeged Wikipédia szócikke) esetén az alábbi eredményt kapjuk, az első 5 találatot kérve: 1. magyar: 35.49, 2. kínai: 2.09, 3. japán (euc jp): 1.81, 4. koreai: 1.70, 5. japán (shift jis): 1.58

Egy angol nyelvű dokumentum esetén az alábbi eredményt kapjuk: 1. angol: 44.21, 2. nepáli: 3.84, 3. kínai: 2.53, 4. vietnami: 2.08, 5. japán: 1.14

Egy német nyelvű dokumentum esetén az alábbi eredményt kapjuk: 1. német: 57.13, 2. kínai: 2.55, 3. japán (shift jis): 2.19, 4. japán (euc jp): 1.93, 5. nepáli: 1.27

Egy olasz nyelvű dokumentum esetén az alábbi eredményt kapjuk: 1. olasz: 35.21, 2. kínai: 1.07, 3. perzsa: 0.68, 4. japán: 0.57, 5. jiddis: 0.55

Jól látható az eredményekből, hogy a rokon nyelvek esetében a nyelvek hasonlóságából adódó hamis többletpontok kiszűrésre kerültek, azaz egy olasz nyelvű dokumentumnál a spanyol nyelv már meg se jelenik az első öt találatban. Most nézzük meg, hogy a kétnyelvű, 50-50 százalékban kevert dokumentumokra mit kapunk.

Egy magyar-angol nyelvű dokumentum esetében az alábbi eredményt kapjuk: 1. angol: 40.80, 2. magyar: 9.40, 3. thai: 1.54, 4. örmény: 1.39, 5. koreai: 1.37

Egy magyar-olasz nyelvű dokumentum esetében: 1. olasz: 49.56, 2. magyar: 7.44, 3. velszi: 2.31, 4. breton: 1.92, 5. ír: 1.68

Egy magyar-francia nyelvű dokumentum esetében: 1. francia: 38.16, 2. magyar: 2.11, 3. thai: 1.42, 4. koreai: 1.16, 5. kínai: 0.70

Egy angol-német nyelvű dokumentum esetében: 1. német: 53.47, 2. angol: 7.79, 3. velszi: 2.08, 4. fríz: 1.48, 5. nepáli: 1.44

Látható például, hogy a magyar-olasz kevert szövegben a magyar nyelv immár a 2. helyre került, a korábbi – eredeti algoritmus által megadott – 10. helyről.

A kétnyelvű dokumentumok esetében nem mindegy, hogy a nyelvek milyen arányban keverednek, érthető módon egy bizonyos arány felett az egyik nyelv n-gramjai elnyomják a másikét. Ezt egy angol-magyar dokumentum-sorozat segítségével nézzük meg (lásd 3.2. táblázat). Az egyes részek arányát a 9 dokumentum során a 10% angol, 90% magyar összetételről 90% angol és 10% magyar összetételre változtattuk.

A 3.2-es táblázat csak egy példa, de a többi nyelvpárra is hasonló eredményeket kaptunk. Látható, hogy az algoritmus 30% körül kezd el hibázni, azaz akkor találja meg megbízhatóan a második nyelvet, ha az a szöveg több mint 30%-át teszi ki. Hasonló eredményt kapunk egy háromnyelvű, harmadolt arányban kevert, magyar-angol-olasz nyelvű dokumentum esetén is: 1. angol: 46.55, 2. magyar: 7.59, 3. olasz: 6.18, 4. breton: 3.11, 5. skót: 2.85

Láthatjuk, hogy a háromnyelvű szövegben az első három helyen szerepelnek a valós nyelvek, de azért itt el kell mondani, hogy ez csak az egyenlő arányban kevert háromnyelvű dokumentumok esetén működik jól. Ha ez az arány eltolódik, akkor gyorsan kieshet egy-egy nyelv. Tapasztalatunk szerint az új algoritmus három nyelvet

már nem talál meg megbízhatóan, így ilyen dokumentumok tömeges előfordulása esetén más algoritmust ajánlott választani.

10% angol, 90% magyar: 1. magyar: 38.01 2. koreai: 1.53 3. thai: 1.20 4. japán (euc): 1.14 5. japán (shift): 1.09	40% angol, 60% magyar: 1. angol: 37.62 2. magyar: 5.41 3. japán (euc): 1.47 4. thai: 1.46 5. japán (shift): 1.45	70% angol, 30% magyar: 1. angol: 44.92 2. vietnámi: 1.74 3. mingo: 1.67 4. kínai: 1.46 5. armén: 1.36
20% angol, 80% magyar: 1. magyar: 37.93 2. thai: 1.18 3. koreai: 1.17 4. japán: 1.16 5. armén: 1.11	50% angol, 50% magyar: 1. angol: 40.93 2. magyar: 5.30 3. thai: 1.49 4. japán (shift): 1.47 5. japán (euc): 1.37	80% angol, 20% magyar: 1. angol: 46.56 2. vietnámi: 2.07 3. mingo: 2.00 4. japán: 1.47 5. walesi: 1.43
30% angol, 70% magyar: 1. magyar: 37.47 2. angol: 4.91 3. thai: 1.22 4. armén: 1.18 5. japán: 1.16	60% angol, 40% magyar: 1. angol: 41.66 2. magyar: 3.43 3. kínai: 1.50 4. vietnámi: 1.48 5. mingo: 1.45	90% angol, 10% magyar: 1. angol: 48.1 2. vietnámi: 1.51 3. nepáli: 1.40 4. thai: 1.05 5. kínai: 1.05

3.2. táblázat: A felismert nyelvek 9 tesztdokumentum esetén

Ahhoz, hogy megállapítsuk, egy dokumentum egy vagy több nyelven íródott-e, kell választanunk egy olyan értéket, ami felett azt mondjuk, hogy a második nyelv is releváns, azaz a dokumentum többnyelvű. Ezt az értéket a tesztek alapján 4-nek javasolt választani, azaz 4-es érték felett érdemes csak kijelezni a nyelveket. Ez az érték a felhasználási igényeknek megfelelően változtatható. Akkor érdemes valamivel alacsonyabbra állítani, ha mindenképp észre szeretnénk venni, ha a dokumentum kétnyelvű, ha pedig csak igazán nagy idegen nyelvű részek érdekelnek, és nem okoz gondot a hibásan egynyelvűnek talált dokumentum, akkor állíthatjuk akár magasabbra is.

Ezzel a paraméterrel az algoritmust részletesen teszteltük a plágiumkeresőnkbe feltöltött dokumentumokon, és a vele szemben támasztott igényeknek messzemenőkig megfelelőnek találtuk. Ki tudtuk szűrni vele a rosszul konvertált és többnyelvű dokumentumok több mint 90%-át. A tesztek befejezése után az új algoritmust beépítettük a KOPI Plágiumkereső rendszerbe, ahol a korábbi, kevésbé pontos eredményt adó n-gram algoritmust váltotta ki.

3.4. Nyelvfelismerő algoritmus – új eredmények összefoglalása

A jelenleg nyelvfelismerésre széles körben használt n-gram algoritmusnak létrehoztam egy új változatát, amely a korábbi algoritmus végeredményét megtisztítja a nyelvek hasonlóságából adódó hamis pozitív találatoktól. Az új algoritmus a dokumentum szövegében 30%-nál nagyobb arányban jelen lévő nyelveket képes felismerni, akkor is ha ezek a részek nem egyben, összefüggő szövegként, hanem elszórva találhatóak meg. Az új algoritmus alkalmas webes korpuszokban található, a plágiumkeresést már negatívan befolyásoló mennyiségben más nyelvet tartalmazó dokumentumok kiszűrésére.

4. Algoritmus fordítási plágiumok keresésére

4.1. Bevezetés

Európában fontos téma a plágiumkeresés, de még nemzetközi szinten is csak kutatási terület a többnyelvű plágiumkeresés. Weber-Wulff (2010) kétévente teszteli az összes elérhető plágiumkeresőt, 2010-ben 48 plágiumkeresőt tesztelt, és azt állapította meg, hogy:

> The biggest gap in all the plagiarism checkers was the inability to locate translated plagiarism.

Azaz a jelenleg (a kutatás kezdetén) elérhető plágiumkeresők egyáltalán nem foglalkoznak a fordítási plágiumok problémájával. A CLEF 2010 konferencia (PAN 2010) az egyik nagy lépcsőfoka volt a többnyelvű plágiumkeresési algoritmusok fejlődésének, ahol több plágiumkereső algoritmus is versenyzett egymással. Itt az egynyelvű keresésen felül már azonos nyelvcsaládba tartozó (angol, német, spanyol) nyelvek között is keresett plágiumot több versenyző is, és majdnem mind automatikus fordítót használtak a plágiumok megtalálására:

> After analyzing all 17 reports, certain algorithmic patterns became apparent to which many participants followed independently. ... In order to simplify the detection of cross-language plagiarism, non-English documents in D are translated to English using machine translation (services). (Potthast et al. 2010)

Most nézzük meg kicsit részletesebben a fordítási plágiumkeresésre használt algoritmusokat.

Kasprzak et al. (2010) a fordítási plágiumkeresésre azt a megoldást használták, hogy az adatbázis építésénél a forrásdokumentumokat mind lefordították angolra a Google Fordító segítségével, majd ebben a már egynyelvű adatbázisban kerestek.

> With the relatively small document corpus size it was feasible to translate the non-English source documents to English, and to use it as alternative source documents.

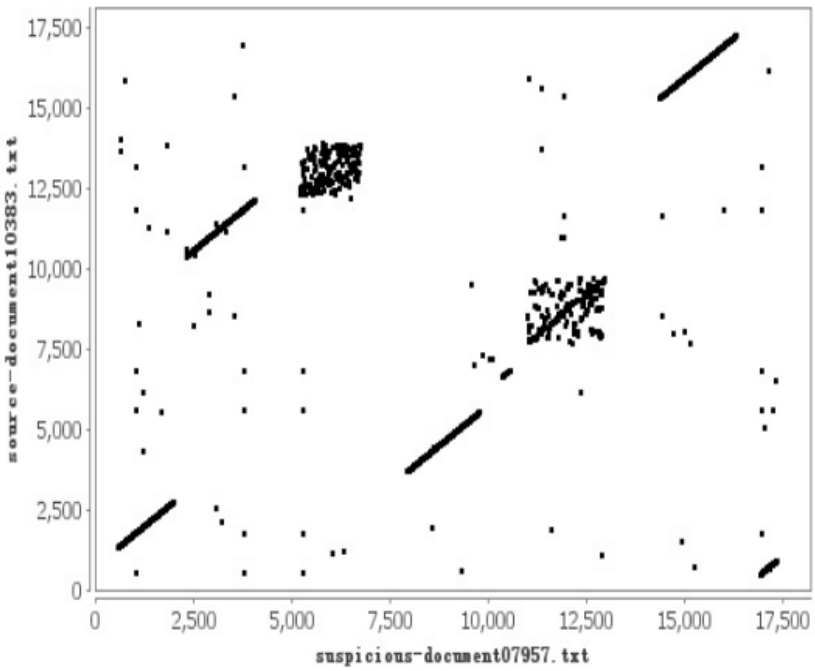
Mint azt ők is elismerik a cikkükben, ez csak azért volt egy kivitelezhető megoldás, mert viszonylag kevés idegen nyelvű dokumentum volt, és előre tudni lehetett, hogy

csak angol nyelvű keresést kell futtatni a dokumentumokon. Egy valós életbeli rendszerben ez nem lenne megoldás, két okból sem: először is nagyon drága lenne az összes lehetséges forrásdokumentum lefordítása, másodszor pedig így egy dokumentumot minden olyan nyelvre le kell fordítani, amilyen nyelven keresni lehet majd benne. Azaz például egy 1 000 000 német dokumentumot tartalmazó webes korpuszban ha szeretnénk angolul, magyarul és németül keresni, akkor 2 000 000 fordításra lesz szükségünk, valamint minden dokumentumot háromszor, három különböző nyelven indexelve fogunk tárolni.

Zou et al. (2010) teljesen más módszert választott a plágiumok keresésére. Először egy gyors algoritmussal kiszűrik azokat a lehetséges dokumentumokat a nagyobb adatbázisból, amelyekben lehet hasonlóság a gyanús dokumentummal.

> The first step is pre-selecting. For each suspicious document, the task is to find out a small list of candidate documents in which the plagiarized content may exist from the source document set quickly. The second step is locating, which compares the suspicious document with each candidate document to get the copy fragments out of the suspicious document. The last step is post-processing, which discards some fragments without plagiarism from the end result.

Majd egy részletes, grafikus összehasonlítást végeznek el, amelynek az alapja egy szó 5-gram algoritmus, és egy futó ablak. A kirajzolt ábráról le lehet olvasni, hogy a két dokumentum hol hasonlít egymásra és milyen hosszan (45 fokos egyenesek) valamint azt is lehet látni, hogy mennyire szó szerinti a másolás, a kisebb különbözőségeket a 45 fokos vonaltól eltérő pontok jelzik (lásd 4.0. ábra). Természetesen ezt algoritmikusan is meg lehet tenni, nem kell a felhasználónak ellenőriznie minden ábrát.



4.0. ábra: Két dokumentum közötti hasonlóság vizualizálva

Az idegen nyelvű dokumentumokat ők is Google Fordító segítségével lefordították, és csak utána dolgozták fel.

```
We used google translation api to translate about two thirds
of the non-English documents into English.
```

Muhr et al. (2010) egy másik megoldást alkalmaztak a plágiumkeresésre, a dokumentumokat átlapolódó, körülbelül 40 szavas, szakaszokra bontották, majd egy információ-visszakereső algoritmus segítségével a gyanús dokumentumok szakaszaihoz próbáltak hasonlóakat találni az adatbázisban lévők között.

```
Source documents are first split into overlap-ping blocks.
Each block is then indexed by a Lucene instance. Suspicious
documents are similarly split into overlapping blocks which
get transformed to boolean Lucene queries. Each query results
in a ranked list of potentially plagiarized source blocks.
```

Megoldásuknak az is az érdekessége, hogy nem egy teljes automatikus fordítót használtak a forrásdokumentumok fordítására, hanem csak szóillesztést.

```
In last year's competition none of the participants tried to
solve the cross-lingual plagiarism subtask. We decided to
give it a try in this year's challenge, building upon
techniques developed in the machine translation community. We
performed word alignment with the BerkleyAligner software
```

> package using the Europarl corpus to provide us with potential translations for each word in German or Spanish source documents. ... For each source document that is not written in English we replaced each word with up to 5 translation candidates. If no translation candidate is available, the word is not replaced. After the words have been replaced the documents are treated similar like the English source documents.

Valamint az is érdekes, hogy az információ-visszakereső miatt nem okozott gondot, ha a célnyelvi szöveg hosszabb, mint a forrásnyelvi, így akár 5 helyettesítő szót is beraktak 1-1 nem angol nyelvű szó helyére. Ennek a megoldásnak az a hátránya – mint az előző kettőnek is – hogy a forrásdokumentumot fordítja le, ráadásul ebben az esetben többszörösére növeli az eltárolandó dokumentum méretét a több lehetséges fordítás párhuzamos használata által.

Egy hasonló algoritmus használatát javasolja Pereira et al. (2010). Ők a dokumentumot bekezdésekre bontják, és ezeket a bekezdéseket töltik fel egy adatbázisba, ahol információ visszakereső algoritmussal a gyanús bekezdésekhez hasonlóakat keresnek. Megoldásuk annyiban különbözik Muhr et al. megoldásától, hogy nem átlapolódó darabokat, hanem bekezdéseket alkalmaznak – azzal a feltételezéssel élve, hogy a fordítás során a bekezdések egyben maradnak – valamint a fordításhoz gépi fordítást alkalmaznak.

> The documents in the collection are translated into a default language so they can be analyzed in a uniform way. The English language was chosen as the default language. A language guesser is used to identify the documents that must be translated and an automatic translation tool is used to translate the documents. ... An information retrieval system is used to retrieve, based on each suspicious document, the documents in the source collection that are candidates of being used as source of plagiarism. Before indexing the source documents, they are divided into several subdocuments, each one containing a single paragraph of the original document. Thus, when submit-ting a query to the system it will only return the relevant subdocuments, not the entire source document. For each passage in the suspicious document, the index is queried and the most relevant subdocuments are returned.

Gottron (2010) is információ visszakereső algoritmust használ, valamint automatikus fordítót, algoritmusa annyiban különbözik Pereira et al. rendszerétől, hogy fix méretű darabokra osztott minden dokumentumot mielőtt feltöltötte az adatbázisba, majd lekérdezésnél ennél rövidebb darabokkal dolgozott.

> Each document d_n was split up into smaller parts d^m_n of a fixed length of $l_d < l$ terms. Each part d^m_n was submitted to the indexing engine as an individual document, but with a reference to the source it was taken from. A similar step was taken for finding candidate documents to compare with a given suspicious document. The suspicious documents I was split into even smaller parts j_i of length $l_q < l_d$. These parts were used as queries to retrieve relevant documents from the index.

Costa-jussà et al. (2010) úgyszintén információ-visszakereső algoritmust használtak a plágiumkereséshez, 50%-ban átlapolódó, 100 szavas töredékekkel. Ezt arra használták, hogy a forrás-dokumentumok közül kiválasszák azokat, amelyeket a következő lépcsőben páronként összehasonlítanak a gyanús dokumentumokkal. Ebben a második lépésben Zou et al. megoldásához hasonló, grafikus összehasonlítást alkalmaznak. A fordítási plágiumok megtalálására azonban szerintük az algoritmusuk nem alkalmas.

> Regarding the first step, it would be interesting to test if retrieving more than one document could improve the recall. Additionally, notice that external plagiarisms were created by using the following offuscations strategies: random text operations, semantic word variation, POS-preserving word shuffling and translation. Our procedure uses bag-of-words, which implies that is not able to detect offuscations based on semantic word variation or translation.

Grozea et al. (2010) megnyerték az első plágiumkeresési versenyt 2009-ben az Encoplot nevű rendszerükkel, amely nagyon pontosan össze tud hasonlítani két dokumentumot, ugyanakkor a rendszerük legnagyobb hátránya, hogy csak párosával képes dokumentumokat összehasonlítani, azaz futási ideje négyzetes a bemeneti dokumentumok számának függvényében.

> The method consists in two stages. In a first stage the values from a string kernel matrix are computed, that give a rough approximation of the similarity between each two documents (a source and a suspicious one). The string kernel used is the normalization of the kernel that counts for each

```
two strings how many character-based N-grams types of a fixed
length N they share. In the second stage the most promising
pairs are examined in detail by creating the encoplot data
for them and the passages in correspondence are extracted
based on some simple heuristics. The encoplot data is a
subset of the dotplot. It consists of a subset of the set of
indexes on which the two documents have the same N-gram,
thus: the first occurrence of an N-gram in one document is
paired with the first occurrence of the same N-gram in the
other document, the second occurrence with the second one and
so on. Computationally, it is worth mentioning that computing
the encoplot data is done in linear time, with the algorithm
we have published in. On the other hand, since our method is
based on pairwise comparisons of documents, its runtime is of
quadratic order in the number of documents.
```

Torrejón et al. (2010) ugyan nem javasoltak megoldást a fordítási plágiumok megkeresésére, de n-gram rendszerük felépítésénél külön megemlítik, hogy az n-gramokon belüli sorrendet érdemes figyelmen kívül hagyni, mert az így kapott rendszer ellenállóbb lesz a szócserékre és a fordításokra.

```
Alphabetic tokens order into every n-gram, processing the
result as canonical representative for the all
possible tokens permutations set. This step
reduces effectivity of words order changes due to sentence
rewriting or translation, improving recall.
```

Érdekes megoldást javasol az n-gram algoritmus optimalizációjára Gupta et al. (2010). A töredékek mennyiségét úgy csökkenti, hogy csak azokat a 9 hosszúságú, nem átlapolódó töredékeket veszi figyelembe, amelyekben minimum egy tulajdonnév található.

```
The suspicious docs are tagged with Lingpipe NE Tagger. Then,
we take non-overlapping n-grams (n=9) which contain at least
one Named Entity (NE). The hypothesis behind is "Information
lies in and around NE".
```

Ez az algoritmus alkalmas lehet az információ-visszakereső algoritmus helyettesítésére, de ahogy a szerzők is javasolják, a megfelelő dokumentumpárok kiválasztása után, egy ennél részletesebb összehasonlításra van szükség.

Egy fuzzy algoritmus használatát javasolja Alzahrani et al. (2010). Algoritmusuk csak dokumentumok páronkénti összehasonlítására alkalmas, így első lépésben n-gram

módszerrel kiszűrik a lehetséges dokumentumpárokat, majd utána egy mondatonkénti összehasonlítást végeznek minden dokumentumpáron. A mondatonkénti fuzzy hasonlóságot pedig úgy számítják ki, hogy a mondat szavait összehasonlítják, 1 pontot adnak ha két szó azonos, 0,5 pontot ha két szó egymás szinonimája, és nullát egyébként. Ezt az értéket elosztják a mondat szavainak számával, és így egy 0 és 1 közötti valószínűséget kapnak a két mondat hasonlóságára.

> Requirements: First, extract a set of features for each $d_q \epsilon D_q$ and $d \epsilon D$. Second, find a list of most promising documents D_x where $D_x c D$ based on shingling and Jaccard similarity coefficient known in IR. Third, perform sentence-wise in-depth analysis using fuzzy semantic-based approach. Last, perform post-processing operations to merge subsequent similar statements into passages or paragraphs. Limitations: Neither intrinsic (i.e. variations in writing styles) nor cross-language plagiarism detection is handled by this algorithm. That is, the languages of both suspicious and candidate documents are considered homogeneous.

Barrón-Cedeño et al. (2008, 2009) kimondottan a fordítási plágiumok megtalálására javasolnak egy algoritmust, amelyik az IBM-1 szóillesztő algoritmusát használja ahhoz, hogy két különböző nyelven íródott szakasz esetén meghatározza a szavak egymásnak való fordításának a *p* valószínűségét, majd ezek szorzatából képezi a teljes valószínűséget. Sajnos a darabolás és a keresés menetét nem írják le, így nem lehet megállapítani, hogy mekkora szakaszokkal, töredékekkel dolgoznak, hogyan keresnek. Ugyanakkor, mint ők is írják, a megoldás ígéretes lehet.

> The results obtained up to now with this method are promising. The application of a statistical machine translation technique, has demonstrated to be a valuable resource for the crosslingual plagiarism analysis.

Az összehasonlító algoritmus egy jó javaslat, a nehézsége csupán az, hogy alkalmazásához rendelkezni kell egy elég nagy, vagy adott szakterületre vonatkozó, párhuzamos korpusszal – amelyből már megfelelő statisztikai következtetéseket lehet levonni – minden olyan nyelvpárra, amelyek között keresni szeretnénk.

A témában az egyik, ha nem a legismertebb blogger, újságíró Jonathan Bailey, aki egy 2011 nyári bejegyzésében az alábbit írta:

> Modern systems, including those that use fingerprinting and string matching, is that they can only detect copied text. Since the definition of plagiarism, often, goes well beyond mere verbatim copying and includes translated works and even just taking the idea, these plagiarism systems are ineffective. (Bailey 2011b)

Azaz a fordítási plágiumkeresés problémája még nincs megoldva, és olyan algoritmusokra van szükség, amelyek a gyakorlatban is használhatóak, implementálhatóak, és alkalmazhatóak nagymennyiségű dolgozat ellenőrzésére és a fordítási plágiumok kiszűrésére.

Ezek a problémák ösztönöztek arra, hogy egy teljesen új, fordításiplágium-keresésre alkalmas algoritmus kutatásába kezdjek, hogy magyar nyelvre, ezen belül is kiemelten az angol-magyar nyelvpárra jól működő megoldást találjak.

4.2. Az algoritmus kialakítása

A CLEF 2010 konferencián versenyző legtöbb a fordítási plágiumok keresésére alkalmas algoritmus, a jelenlegi egynyelvű keresés adaptálása egy adott nyelvpárra. A legjobb plágiumkeresők átlapolódó szavas darabolást (n-gramokat) használnak a szövegek összehasonlítására a plágiumkeresésre. (Potthast 2010) Ez az algoritmus szó szerinti egyezést keres, amelyet számos megoldással igyekeznek javítani, hogy kisebb átírásokat, eltéréseket ne vegyen figyelembe, ezek közül a leggyakrabbak az alábbiak: a) stopszavak szűrése, b) szótövezés, c) bizonyos szavak kicserélése egy szinonimára, d) szavak sorrendezése az n-gramon belül. Ezek a változtatások sokkal nehezebbé teszik a plágiumok elrejtését, és jelentősen megnövelik a lebukás kockázatát, ugyanakkor különböző nyelven írt szövegek között még mindig nem teszik lehetővé az összehasonlítást. Mint láttuk az előző fejezetben, többen próbálkoztak automatikus, gépi fordítók alkalmazásával arra, hogy a két szöveget azonos nyelvre hozzák, ugyanakkor ezek a fordítók eredményei ma még nagyban függenek az adott nyelvpártól, a szöveg témájától, a mondatok összetettségétől. Összefoglalva elmondhatjuk, és ez nem csak a gépi fordítókra igaz – habár azokra kiemelten az – hogy egy fordítás komoly változtatást eredményez a szövegen, hibákat is okozhat, és a szavak mondaton belüli sorrendjén is komolyan változtat, pláne az olyan nem kötött szórendű nyelvek esetében, mint amilyen a magyar.

Egy gépi fordítást alkalmazó plágiumkereső algoritmus tulajdonképpen két – különböző algoritmussal történő – fordítási lépés után hasonlítja össze az eredeti és az új szöveget. Először az, aki plagizálta az eredeti szöveget lefordítja azt, majd ezután a kereső visszafordítja egy gépi fordítóval, majd az ezek után kapott, lefordított-visszafordított szöveget hasonlítja össze az eredeti szöveggel. Mivel a legtöbb mondatnak számos lehetséges fordítása van, így majdnem teljesen biztosak lehetünk benne, hogy komoly különbségek lesznek a mondatok között, nemcsak a szórendben, hanem a használt szavakban, kifejezésekben is.

A 9.3 mellékletben található pár véletlenszerűen kiválasztott Wikipédia oldal, és azok kézi és gépi fordítása (GoogleT). Ezekből jól látszik, hogy a két fordítás mennyire különbözik, ahogy két gépi és két kézi fordítás is valószínűleg igen nagy eltérést mutatna. Most csak mintának álljon itt a legelső és a két utolsó mondat, amelyek igen egyszerűek, nem összetettek, és elvileg könnyen fordíthatóak.

Eredeti angol nyelvű Wikipédia szócikk: Johann Haller

Johann Haller or Jan Haller (1463–1525) is considered one of the first commercial printers in Poland. … Altogether Haller produced 3,530 prints. His masterpieces are illustrated books containing 354 sheets of woodcuts.

Kézi fordítás magyarra: Johann Haller

Johann Haller, vagy Jan Haller, (1463-1525) az, akit az első hivatásos nyomdásznak tartanak Lengyelországban. … Haller összesen 3,530 nyomtatást készített. Mesterművei illusztrálással díszített könyvek melyek 354 oldalnyi fametszetet tartalmaznak.

Gépi fordítás magyarra: Johann Haller

Johann Haller és Jan Haller (1463-1525) tartják az egyik első kereskedelmi nyomtatók Lengyelországban. … Összesen Haller előállított nyomatok 3530. Ő remekművek is illusztrált könyvek tartalmazó 354 darab fametszet.

Magyar kézi fordítás visszafordítása géppel: Johann Haller

Johann Haller, Haller and Jan, (1463-1525), the one of the first professional typography held in Poland. … Haller made a total of 3.530 printing. Masterpieces illusztrálással decorated with books that contain 354 pages of woodcut.

Jól látszik, hogy egy olyan egyszerű mondat esetében is, mint az „Altogether Haller produced 3,530 prints.” Az oda-vissza fordítása egy ettől formailag igen eltérő „Haller made a total of 3.530 printing.” mondatot eredményezett: a mondatban szereplő 4 szóból csak egy szó egyezik, a tulajdonnév, illetve a szám megmaradt.

A gépi fordítók folyamatosan fejlődnek, így hosszú távon el lehet gondolkodni majd a használatukon. Komoly hátrányuk ma még, hogy egy külső programra, vagy szolgáltatásra kell hagyatkozni, és a jó minőségű szolgáltatások mind fizetősek, így komolyabb mennyiségű szöveg rendszeres lefordítása, komoly költségekkel is járna. Ezért is igyekeztem az új algoritmus kialakításakor elkerülni az automatikus fordítóktó-szolgáltatásoktól való függőséget.

A nyelv legkisebb önállóan is értelmes egysége a szó. Ez az, ami egy adott tárgyat, cselekvést, gondolatot jelképez. Nagyobb egységek a kifejezések, frázisok, tagmondatok, amelyek általában egy különálló gondolati egységet reprezentálnak. Egy még magasabb szint a mondat. A mondat legtöbbször egységes, önmagában zárt mondanivalót képvisel. Természetesen egy adott mondat jelentése erősen függhet a környezetétől, ahogy egy adott szó jelentése is, de alapvetően már egy teljes gondolatot, mondanivalót jelképez. A következő formai egység a bekezdés, majd az ennél nagyobb részek, a fejezetek esetleg dokumentumok jönnek.

Most nézzük meg ezeket az egységeket két nyelv között. Az egyszerűség kedvéért, mostantól főleg az angol-magyar nyelvpárt vesszük alapul, hiszen a célunk elsősorban ennek az elemzése, kutatása, de bárki továbbgondolhatja a példákat, felvetéseket, és ha nem is az összes, de a legtöbb nyelv illetve nyelvpár esetében igazak lesznek az alábbi állítások.

Két nyelv között a legkisebb egyezés egy szótő egyezése lehet. Természetesen, ha egy angol szövegben az *eleven* szót olvashatjuk, akkor az annak magyarul nem az *eleven* szó fog megfelelni, hanem a *tizenegy* vagy a *11*, de ennek ellenére beszélhetünk egyezésről. Ugyanakkor érdemes megjegyezni, hogy számos szónak nem lesz megfelelője a másik nyelvben, vagy egyáltalán nem is lesz megfelelője, vagy nem szóként jelentkezik. Most a teljesség igénye nélkül vegyünk sorra pár lehetséges eltérést.

- Összetett szavak: elképzelhető, hogy míg az egyik nyelvben valamit egy szóval, addig a másikban többel fejezzük ki, például *tavaly* és *last year*. Fordítva pedig míg magyarul *szabadlábra helyeznek* valakit, angolul *liberated*.
- Ragozás: a magyar nyelv számos dolgot ragokkal, a szóval egybe írva fejez ki, míg más nyelvek erre elöljárót használnak. Ami magyarul az *álmomban* történt, az angolul *in my dream*.
- Antonima: gyakran egy kifejezést jobb antonimával fordítani, nem önmagával. Míg magyarul valami *nem felel meg* a célnak, addig angolul *inadequate*.
- Műfordítás: műfordítások esetén gyakran találkozunk olyan mondat-, illetve szópárokkal, amelyeket mindenki ért, hogy adott szövegkörnyezetben miért egymás fordításai, ugyanakkor szó szerint nem felelnek meg egymásnak, csak szemantikus szinten A „80 nap alatt a föld körül" magyar fordításában találkozunk a *gentleman* szóval, ahol az angolban a *Mr. Fogg* szerepel.
- Teljes átalakítás: kifejezések és a forrás- valamint célnyelv különbözőségén, illetve kulturális különbözőségéből adódóan. A *Queen's pudding*–ból *rakott palacsinta* lesz, az *egg and spoon races* pedig *ügyességi gyerekjáték*. (Tóth 2005)

Azaz számos eset képzelhető el, amikor egy adott szó nem felel meg egy-egyértelműen a másik nyelv egy szavának, ugyanakkor a szavak jelentős része megtalálható lesz mindkét nyelvben. Ennek ellenére a szavakat jól fel lehet használni arra, hogy fordításokat keressünk, de önmagában két szöveg még nem lesz azonos azért, mert sok közös szavuk van.

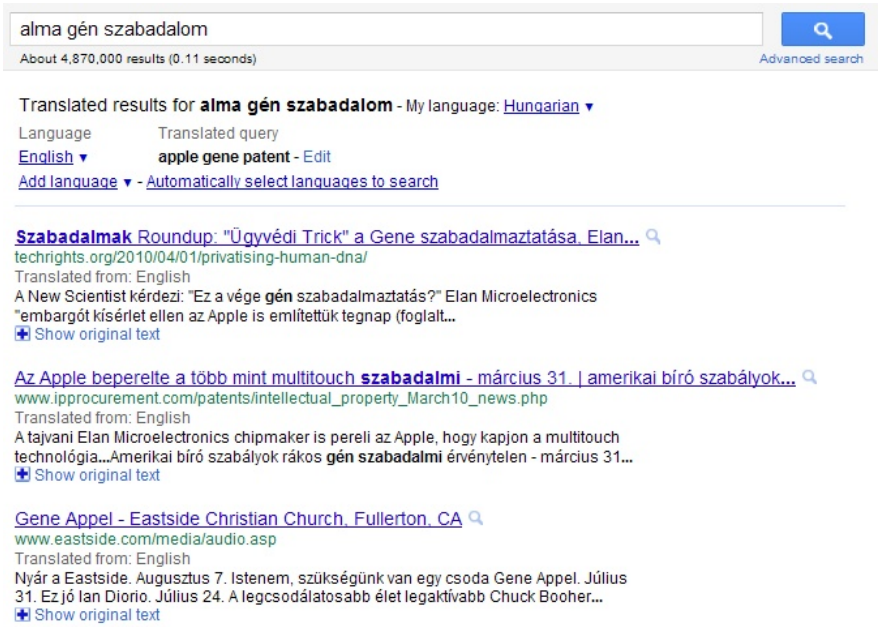
A kifejezések, frázisok, tagmondatok határainak megállapítása időigényes, komoly nyelvi elemzést igényel, így ezzel a lehetőséggel mi most itt nem is foglalkozunk.

Két nyelven is meglévő szövegek mondatszintű párhuzamosításra (sentence alignment) számos szoftver létezik, ezek azt használják ki, hogy egy szöveg és a fordítása nagyon jól megfeleltethető egymásnak a mondatok szintjén.

Az ennél magasabb szintek, bekezdések, fejezetek határai már nem olyan könnyen, egyértelműen határozhatóak meg, illetve szándékos plagizálás esetén kevés energiával

eltüntethetőek, megváltoztathatóak, így ezek egyezésének a vizsgálatára úgyszintén nem térünk most ki.

Mint láttuk, fordítások esetében a plágiumkereső szempontjából legegyszerűbben használható szint a szavak, vagy a mondatok szintje. A szavak esetében viszont lényeges a pozíciójuk, a többi szóhoz viszonyítva, a szövegkörnyezet, hiszen bármely két azonos nyelven íródott szövegben vannak azonos szavak, még akár ezek mértéke magas is lehet, akkor se biztos, hogy a két szövegnek ugyanaz a jelentése. Mint azt a webes keresők esetében látjuk – ahol adott szót tartalmazó szövegekre keresünk – nagyon nagy az olyan találatok száma, amelyek ugyan megfelelnek a keresőkérdésnek, de semmi közük sincs ahhoz, amit kerestünk. Ez két különböző nyelv esetében még inkább így lesz, hiszen egy adott szónak a másik nyelvben számos másik felel, vagy felelhet meg, így még ez is egy komoly bizonytalanságot eredményez.



4.1. ábra: Az „alma gén szabadalom” szavakra való keresés a Google fordító keresőjében

A Google-nek van egy igen kényelmes szolgáltatása, azok számára, akik nem beszélnek egy adott idegen nyelvet, de szeretnének az adott nyelvű oldalakon keresni. A keresőkérdéseket feltehetjük a Google-nek az anyanyelvünkön, majd a keresés az adott idegen nyelven történik, és az eredményt meg visszafordítja a Google a mi nyelvünkre. (GTSearch)

4.2. ábra: Az „alma gén szabadalom” szavakra való keresés a Webfordítás keresőjében

Hasonló szolgáltatást nyújt a Morphologic által fejlesztett webforditas.hu is. Ha az *alma* és a *szabadalom* szavakra rákeresünk, akkor valószínűleg a génmódosított szabadalmaztatott almákról szeretnénk valami információt kapni, ugyanakkor mindkét keresőben az Apple cég szabadalmai fognak előjönni. A fordítók nem tesznek különbséget az almáról és az Apple cégről szóló oldalak között. Magyarul ez a két szó még csak nem is hasonlít egymásra, ugyanakkor angolul csak a nagy kezdőbetű választja el őket egymástól, amit nem vesz figyelembe semelyik kereső se. Ha pontosítjuk a keresést a *gén* szóval, akkor már jobb találatokat kapunk, de még így is vezető helyet foglal el az Apple cég mindkét keresőben (4.1. és 4.2. ábra).

Természetesen ez nem azt jelenti, hogy a szavak nem használhatók két szöveg közti egyezés magtalálására, de önmagában nem elég, hiszen, ha valaki lefordít egy egyoldalas szöveget angolról, és beteszi a 120 oldalas magyar diplomájába, akkor ennek a megtalálása pusztán a szavak használatával lehetetlen. Mindenképpen definiálnunk kell egy szövegkörnyezetet, ahol a szavakat keressük. Ezért jobb kiindulási pontnak tűnt a mondat alapú keresés, ahol a szavaknak van szövegkörnyezetük, ráadásul a mondat már elég egyedi ahhoz, hogy két adott dokumentumban – még ha azonos témában íródtak is – kis eséllyel lesz két azonos mondat (az 1-3 szavas mondatokat nem számítva). Ez első hallásra nehezen hihetőnek tűnik, de ha belegondolunk, hogy a legtöbb nyelvnek több százezer szava van (Oxford), akkor a nyelvtani szabályokat most

figyelmen kívül hagyva, százezer szóval számolva az adott nyelven egy *n* szóból álló mondat (S_n) összes lehetséges változata: $|S_n| = (10^5)^n$

Ez egy rövidebb 5 szavas mondat esetében is már: $|S_{10}| = 10^{25}$

Ha hozzávesszük, hogy például a magyar nyelvben a legtöbb szónak számos toldalékolt alakja van, akkor ez a szám még jelentősen növekedne, de még az angol esetében is a többesszám és egyéb alakok miatt az alapszókincs többszöröse a ténylegesen előforduló szóalakok száma. Ezért tekinthetünk úgy egy mondatra, mint egyedi alkotásra. Sokak szerint ezért már egyetlen mondatnál kezdődik a plagizálás, azaz egy (tartalmas) mondat már rendelkezik annyi egyedi tulajdonsággal, hogy lemásolása esetén lehet plagizálásról beszélni. Fry (1989), angol színész, komikus és szövegíró, így fogalmazza meg a mondatok sokszínűségét és egyediségét egy vígjátékában:

> Imagine a piano keyboard, eh, 88 keys, only 88 and yet, and yet, hundreds of new melodies, new tunes, new harmonies are being composed upon hundreds of different keyboards every day in Dorset alone. Our language, tiger, our language: **hundreds of thousands of available words, frillions of legitimate new ideas,** so that I can say the following sentence and be **utterly sure that nobody has ever said it before** in the history of human communication: "Hold the newsreader's nose squarely, waiter, or friendly milk will countermand my trousers." Perfectly ordinary words, but never before put in that precise order. A unique child delivered of a unique mother.

Érdemes megnézni a Wikipédia ide vonatkozó oldalán található összefoglaló táblázatot, amelyből itt csak egy kivonatot mutatunk be. (Szókincs)

Dokumentum, bemeneti adat, szövegkörnyezet	Szavak száma	$\|S_{10}\|$
Egy szöveg leggyakoribb szavai közül ennyi adja ki annak 25%-át.	15	5,8E+11
Egy szöveg leggyakoribb szavai közül ennyi adja ki annak 60%-át.	100	1,0E+20
Kb. egy 2 éves gyerek szókincse	300	5,9E+24
Az Ogden-féle egyszerű angol nyelv (Basic English) szókincse	850	2,0E+29
Ennyi szót használnak az első osztályosok olvasástanításában.	1000	1,0E+30
Kb. egy 6 éves gyerek szókincse	2500	9,5E+33
Arany János Toldi c. művében felhasznált szókincse	3000	5,9E+34

Az átlagember aktív szókincse (élő-aktív és szunnyadó-aktív)	3 000-5 000	5,9E+34
Középfokú nyelvtudásnak megfelelő szókincs	3 500-3 900	2,8E+35
Kb. egy 11 éves gyerek szókincse	5 000	9,8E+36
Az átlagember passzív szókincse	5 000-10 000	5,6E+38
Ennyi szóval a Shreket 95%-ban megértjük.	6 000	6,0E+37
Ennyi szó szükséges a 20. századi angol próza megértéséhez.	8-9 000	1,1E+39
Ennyi szóval a tankönyveket 95%-ban megértjük.	10-12 000	1,0E+40
Egy kétnyelvű kisszótár terjedelme (címszavak)	10-30 000	1,0E+43
Shakespeare (műveiben felhasznált) szókincsét ennyire becsülik	18-25 000	1,7E+43
Petőfi Sándor verseiből kimutatható szókincse	22 719	3,7E+43
Egy átlag értelmiségi egyévi beszédét gondolatban rögzítve kb. ennyiféle szó fordulna elő.	25-30 000	3,0E+44
Igen művelt embereknél a passzív szókincs nagysága	50-60 000	2,5E+47
Kb. ennyi mai magyar szót tartanak számon.	60-100 000	1,1E+49
Egy kétnyelvű nagyszótár terjedelme (címszavak)	120 000	6,2E+50
A 20 kötetes Oxford English Dictionary 2. (nyomtatott) kiadásának (1989) terjedelme (címszavak)	291 500	4,4E+54
A 33 kötetes Deutsches Wörterbuch terjedelme (1960-as kiadás, címszavak)	350 000	2,8E+55

4.3. táblázat: Szókincsméretek összehasonlító listája

Jól látható a táblázatból, hogy már egy kétéves gyerek is több száz szót ismer, és ha csak rövidebb, pár szó hosszú, mondatokat vesszük, akkor is több százezer mondatot tud elméletileg összetenni.

Mahmoud (2006), arab nyelvész írta, hogy az idegen-nyelv oktatásban fontos eszköznek tartja a visszafordítást, azaz a célnyelvről (L2) a tanulók anyanyelvére (L1) való fordítás után a hibás mondatok visszafordítását, hogy a tanulók láthassák, hogy mit rontottak el.

„L2 learners customarily rely on their L1, especially in acquisition-poor environments where exposure to the L2 is confined to a few hours per week of formal classroom instruction. For many teachers and students, the use of L1 is a learning and communication strategy that can be used in the classroom for various purposes, such as to explain difficult concepts. I support L2 to L1 written translation in particular not only because of its benefits as a post-reading task, but also because the incorrect translations can be

> back-translated into the L2 and used as a source of information to be fed into other testing techniques. Translation is also useful because it draws the students' attention to the entire reading passage at the word, sentence, and text level. Because translation does not require students to respond in the L2, it focuses on comprehension, the skill it purports to develop." [Mahmoud2006]

Számunkra azért lényeges ez a feladat, mert jól mutatja nyelvoktatásban azt az igyekezetet – és ez igaz a hazai nyelvoktatásra is – hogy a célnyelvi szöveg visszafordításával, egy azonos jelentésű mondatot kell kapnunk (csak a jelentése lesz azonos, a mondat maga nem biztos). Azaz a fordítás egy szimmetrikus művelet, ha egy adott mondatot lefordítunk egy célnyelvre, akkor annak egy lehetséges visszafordítása lesz önmaga. Ez triviálisnak tűnik, de egy lényeges és szükséges megállapítás a továbbiak megértéséhez.

Érdemes az előzőekhez még hozzátenni, hogy legtöbb esetben nincs a mondatoknak, szövegeknek egy jó, helyes fordítása, a fordító a körülmények figyelembe vételével – többek között a szövegkörnyezet és az olvasóközönség – választ számos lehetséges jó megoldás közül. Fischer (2008) ezt így fogalmazza meg:

> A nyelvészeti fordítástudomány eredményei – amelynek fontos területe az ekvivalencia kutatása – eloszlathatják azt a téves elképzelést, mely szerint a fordítás automatikus és teljes megfeleltetést (ekvivalenciát) feltételez a két nyelv között. A kutatók különböző megközelítései és a számtalan ekvivalencia-elmélet éppen arra világítanak rá, hogy az ekvivalencia több szinten, több szempont szerint értelmezhető Ezek ismerete tehát éppen abban erősítheti meg a tanulót, hogy nincs egyetlen helyes (ekvivalens) válasz.

Összefoglalva az előzőeket, láthatólag a mondat megfelelő egység ahhoz, hogy plágiumot keressünk, illetve szövegek közötti egyezéseket. Az alábbi előnyei vannak:

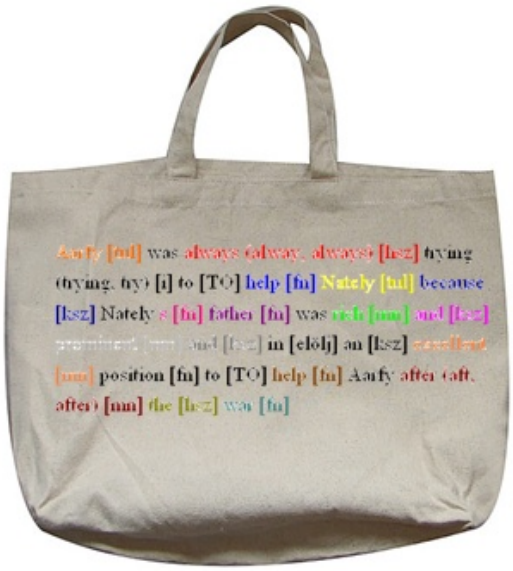
- egyetlen gondolati egységet képvisel
- határai nagy pontossággal meghatározhatóak
- elég egyedi ahhoz, hogy két szöveg között több mondat egyezésekor már valami közös forrást feltételezzünk

- fordítások között is megfeleltethetőek a mondatok egymásnak

Azok után, hogy beláttuk, hogy érdemes a mondatok közötti hasonlóságot vizsgálnunk, ahhoz, hogy a fordítási plágiumot megtaláljuk, meg kell meghatároznunk egy metrikát, amely különböző nyelven íródott mondatok közötti hasonlóság mértékét határozza meg.

4.2.1. Fordítások összehasonlítása – hasonlósági metrika

Mint korábban említettük egy angol és egy magyar nyelvű mondat szavai, ha nem is teljes mértékben, de megfeleltethetők egymásnak. A két nyelv nyelvtanának különbségéből és a magyar nyelv kötetlen szórendjéből adódóan a szavak sorrendje teljesen lényegtelen ebben a megfeleltetésben, azaz az angol nyelvű mondat első, második, harmadik… szava bárhol lehet a magyar mondatban, és fordítva. Célunk tulajdonképpen egy olyan algoritmus előállítása, amely minden mondathoz előállítja az összes lehetséges fordítást, és ezek között keres hasonlóságot.



4.4. ábra: Szózsák :-)

A sorrendet figyelembe nem vevő, egy szöveg szavait reprezentáló modell (Harris 1954) a szózsák (bag of words), ez egy adott szöveg összes szavát tartalmazó, de a sorrendet nem figyelembe vevő halmaz, amelyet számos helyen használnak a szakirodalomban, például dokumentumok csoportosítására, SPAM szűrésre, de még érzelmek felismerésére is (Miháltz 2010). Mi most sokkal kisebb egységben fogjuk a szózsákot alkalmazni, a mondatok szintjén.

Egy n szóból álló mondatot (S_n) képviseljenek a benne lévő szavak (w_1, w_2, ..., w_n).

$$w_i^x \in S_n^x \text{ ahol } w_i^x \text{ az } x. \text{ mondat } i. \text{ szava}$$

Természetesen ez egyszerűsítés, hiszen elméletileg ugyanazokból a szavakból más mondatokat is össze lehet rakni, de azért elég egyértelműen visszaállítható egy mondat értelme a szavak ismeretében, így túl sok hibát ez az átalakítás nem fog eredményezni.

$$S_n^x = \left\{w_1^x, w_2^x, ..., w_n^x\right\}$$

Most definiáljuk két mondat hasonlóságának a mértékét (*Sim*) a bennük levő közös szavak számával.

$$\mathrm{Sim}\left(S_n^x, S_m^y\right) = \left|S_n^x \cap S_m^y\right|$$

Ez már egy jó megközelítés, de számos dolgot nem vesz figyelembe. Például egy hosszú és egy rövid mondat hasonlósága így maximum akkora lehet, amekkora a rövid mondat hossza, ami helyes is; ugyanakkor ha például a hosszú mondatban megtalálható a rövid mondat összes szava, akkor ez a két mondat ugyanannyira hasonló lesz, mintha a rövid mondatot önmagával hasonlítottuk volna össze. Ez nem jívánatos, ezért nem csak a közös szavakat, hanem a hiányzó szavakat is figyelembe kell venni. Ezeket érdemes súlyozni is: legyen a megtalált szavak súlya α a nem megtaláltaké β.

$$\mathrm{Sim}\left(S_n^x, S_m^y\right) = \alpha \cdot \left|S_n^x \cap S_m^y\right| - \beta \cdot \left|S_n^x \setminus S_m^y\right|$$

Ha például α értékét 3-nak, β értékét 1-nek vesszük, akkor az azt jelenti, hogy minden olyan szót, amelyik megtalálható a másik mondatban, háromszoros súllyal veszünk figyelembe, a hiányzó szavakhoz képest.

Ez a képlet már jobban megfelel, de nem szimmetrikus $S^x \setminus S^y$ miatt, azaz: $Sim(S^x,S^y) \neq Sim(S^y,S^x)$. Mint korábban láttuk, annak az esélye, hogy S^x mondat S^y-nak a fordítása elvileg ugyanannyi kellene, hogy legyen, mint annak, hogy S^y mondat S^x-nek a fordítása, mert a fordítás szimmetrikus művelet. Ezt úgy lehet kiküszöbölni, hogy kiszámoljuk mindkét értéket, majd például vesszük az összegét. Ugyanakkor azért vezettük be az egyenlet második tagját ($S^x \setminus S^y$), mert azok a szavak, amelyek csak az egyik mondatban találhatóak meg, csökkentik annak a valószínűségét, hogy a két mondat egymás fordítása. Ha kisebb annak ez esélye, hogy S^x fordítása S^y-nak, mint a fordítottja – azaz $Sim(S^x,S^y) < Sim(S^y,S^x)$ –, akkor az azt jelenti, hogy S^x hosszabb, azaz több olyan szó van benne, aminek nincs fordítása a másik mondatban. Ez lényeges, és

hiába kapunk $Sim(S^y,S^x)$-re nagyon magas értéket, ha $Sim(S^x,S^y)$ alacsony, akkor majdnem biztos, hogy a két mondat nem fordítása egymásnak, esetleg az egyik a másik része. Ezért a továbbiakban úgy számoljuk ki $Sim(S^x,S^y)$ értékét, hogy a korábban definiált értékek közül az alacsonyabbat vesszük. Ezzel az új képlet:

$$\mathrm{Sim}\left(S_n^x, S_m^y\right) = \min\left(\alpha \cdot \left|S_n^x \cap S_m^y\right| - \beta \cdot \left|S_n^x \setminus S_m^y\right|, \alpha \cdot \left|S_m^y \cap S_n^x\right| - \beta \cdot \left|S_m^y \setminus S_n^x\right|\right)$$

Ez a definíció már eleget tesz a szimmetria követelményének, azaz most már

$$\mathrm{Sim}\left(S_n^x, S_m^y\right) = \mathrm{Sim}\left(S_m^y, S_n^x\right)$$

A továbbiakban még néhány lényeges dolgot figyelembe kell vennünk ahhoz, hogy a szózsák algoritmus fordítások esetében is jól működjön. Mivel S^x és S^y nyelve nem azonos, ezért definiálnunk kell, hogy mit jelent két szó egyenlősége/azonossága, és mit jelent a különbözősége. Azaz mikor mondjuk, hogy $w_i \equiv w_j$ és mikor mondjuk, hogy $w_i \not\equiv w_j$; ahhoz, hogy ezt meghatározzuk, definiálnunk kell még egy műveletet, a fordítás műveletét, azaz egy fordításfüggvényt. Az alábbiakban W szavak halmazát jelenti, melyeket itt egy adott szó, w fordításaként kapunk: például x nyelvű mondat egy szavának, w^x-nek a lehetséges fordításait y nyelvre W^y jelöli.

$$\mathrm{trans}\left(w_i^x\right) = W^y \text{ ahol } w_j^y \in W^y$$

$$\mathrm{trans}\left(w_j^y\right) = W^x \text{ ahol } w_i^x \in W^x$$

Mivel a fordítás szimmetrikus művelet ezért

$$w_i^x \in \mathrm{trans}\left(w_j^y\right) \Rightarrow w_j^y \in \mathrm{trans}\left(w_i^x\right)$$

Ezek alapján

$$w_j^y \in \mathrm{trans}\left(w_i^x\right) \Rightarrow w_i^x \equiv w_j^y$$

$$w_i^x \in \mathrm{trans}\left(w_j^y\right) \Rightarrow w_i^x \equiv w_j^y$$

Hasonló módon

$$w_j^y \notin \mathrm{trans}\left(w_i^x\right) \Rightarrow w_i^x \not\equiv w_j^y$$

$$w_i^x \notin \mathrm{trans}\left(w_j^y\right) \Rightarrow w_i^x \not\equiv w_j^y$$

Számos előny származik az így definiált hasonlósági metrikából. Az automatikus fordítókra épülő algoritmusok esetében a fordítás során jelentés-egyértelműsítést kell alkalmazni, annak érdekében, hogy minden a szónak a megfelelő fordítása kerüljön bele a fordított szövegbe (Apple: alma vagy Apple?). Ez nehéz feladat: akár több száz megoldás létezhet rá, a legjobb rendszerek 90% feletti pontosságot is el tudnak érni, de ezt csak nagyon szűk tartományokban teljesítik, néha csak egy tucat betanított szóra. (Craggs 2011) Az egynyelvű plágiumkeresők és az arra épülő többnyelvű algoritmusok azért, hogy egy egyezést akkor is megtaláljanak, ha egy szónak a szinonimája található a másik szövegben, gyakran egy közös szóval vagy jellel helyettesítenek egy-egy szinonimacsoportba (sinset) tartozó szavakat. Ha belegondolunk, ezt is egy jelentés-egyértelműsítés kell, hogy megelőzze, hiszen a többértelmű szavak esetén előbb el kell dönteni, hogy melyik jelentésről van szó, és csak utána lehet helyettesíteni. Ezekkel a problémákkal szemben a fenti algoritmus ezt a két nehéz lépést egyben megoldja azzal, hogy képes különböző nyelvű szövegeket közvetlenül összehasonlítani, és az összes lehetséges fordítás, szinonima egyszerre szerepel az összehasonlításban, nem kell előre kitalálni, hogy melyik lehet a helyes, melyik az amit a fordító használt. Az algoritmus nem érzékeny továbbá a szavak sorrendjére, ami a magyar nyelv esetében kiemelten fontos, hiszen míg például angolul a „the dog chases the cat” nem azonos a „the cat chases the dog” kifejezéssel. Magyarul ez közel azonos jelentéssel bír: „a kutya kergeti a macskát” illetve „a macskát kergeti a kutya” vagy akár a „kergeti a macskát a kutya” és a „kergeti a kutya a macskát” is használható. (Kiss 1983)

A fenti előnyökön felül hátrányai is vannak az algoritmusnak, méghozzá a keresési tér nagysága. Mint később látni fogjuk, egy nagy szótár alkalmazása mellett egy-egy szónak átlagban akár 20 jelentése is lehet, ami nagyban megnehezíti, lassítja a keresést. A másik hátránya, hogy a keresési idő lineáris, azaz elméletileg minden gyanús, keresett mondatot össze kell hasonlítanunk az adatbázisunkban lévő összes mondattal, ami már egy közepes méretű adatbázis esetén is elfogadhatatlanul lassú futási időkhöz vezetne. Később látni fogjuk, hogy ezt a problémát miként lehet kiküszöbölni, megkerülni indexált keresés használatával.

4.2.2. Implementációs döntések

A fentiekben az új algoritmus alapgondolatának a leírása és kialakulásának a fontosabb lépései, döntései találhatók, amely ezek alapján már implementálható és tesztelhető. A

gyakorlatban pár további megkötést, egyszerűsítést, illetve implementációs lépést is be kellett iktatni az algoritmus használhatóságának, sebességének növelése érdekében. Most röviden ezeket foglaljuk össze.

Az egyik első lépés, amely a legtöbb számítógépes nyelvészeti elemzés első lépése, a stopszavak, azaz a leggyakoribb szavak kiszűrése. A legtöbb szöveg esetében annak jelentős részét az adott nyelv leggyakoribb szavai teszik ki. Ilyenek többek közt a névelők, igekötők, létigék. Crystal (2003) szerint a leggyakoribb 15 szó teszi ki a szöveg 25%-át, és a leggyakoribb 100 szó a 60%-át. Utóbbi már szövegről, szövegre változni fog, az előbbi majdnem teljesen állandó, az adott nyelvre jellemző.

Nem teljesen egyértelmű, hogy mikor érdemes és mikor nem érdemes a nagybetűs szavakat figyelembe venni. Ha a mondatkezdő nagybetűre gondolunk, akkor az fordítás szempontjából nem hordoz információt, ugyanakkor tulajdonnevek esetében lényeges információt hordoz. Az alábbi mondatban: „*Az osztályt dr. Kovács László vezeti.*" a *Kovács* szó egy tulajdonnév és mindkét nyelvben ugyanúgy kell szerepeljen (nem írhatjuk, hogy *dr. Smith*), de kisbetűvel írva: „*A kovács megpatkolta a lovat.*" már le kell fordítani és itt a *smith* a helyes fordítás. Ez a különbség nyelvfüggő, hiszen egyes nyelvek esetén, ahol minden főnevet nagybetűvel írnak, komolyabb nyelvtani elemzés szükséges ahhoz, hogy megállapítsuk, mik a tulajdonnevek és mik nem azok. (Faruqui 2010) Mivel a fordító valószínűleg könnyedén el tudja dönteni, melyik esetről van szó, ezért az algoritmusnak nem kell figyelembe vennie a kis- és nagybetűket, csupán egy új szabály bevezetésére van szükség, miszerint „bármely szó fordítása önmagának bármely másik nyelven". Ez nemcsak tulajdonnevek, de szakkifejezések esetén is gyakran alkalmazható, főleg angol-magyar nyelvpár esetén, de német-angol esetében is igen gyakoriak az angolból átvett szavak (Neudeutsch). Ezt az alábbi képlettel íratjuk le:

$$w_i^x \in \mathrm{trans}\left(w_i^x\right) \Rightarrow w_i^x \equiv w_i^x$$

minden esetben.

Bizonyos szófajok jobban jellemeznek egy mondatot mint mások, illetve nagyobb valószínűséggel kerülnek fordításra, mint mások. Például főnév vagy ige majdnem biztos, hogy mindkét mondatban megtalálható lesz, míg egy igekötő, vagy elöljárószó már nem biztos, vagy bizonyos nyelvpárok esetén biztos nem.

Aarfy [tul] was always (alway, always) [hsz] trying (trying, try) [i] to [TO] help [fn] Nately [tul] because [ksz] Nately s [fn] father [fn] was rich [nn] and [ksz] prominent [nn] and [ksz] in [elölj] an [ksz] excellent [nn] position [fn] to [TO] help [fn] Aarfy after (aft, after) [nn] the [hsz] war [fn]

Aarfy mindig (mindig, mind) igyekezett (igyekezik, igyekezett) Natelyn segíteni (segít) mert (mert, mer) Nately apja (apa) gazdag és befolyásos ember volt (volt, van) aki kitűnő állása (állás) révén (révén, rév) segíthetett (segít) volna (van) Aarfyn a háború után

4.5. ábra: Megtalált szavak és jelölt szófajok

Ezért érdemes kisebb súllyal kezelni azon szavak hiányát, amelyek általában csak az egyik nyelvre jellemzőek, míg nagyobb súllyal kell figyelembe venni azokat, amelyek általában le szoktak fordulni. Ugyanakkor fontos megjegyezni, hogy ez a lépés már nyelvpárfüggő, hiszen a két nyelv nyelvtanának hasonlósága nagyban befolyásolja, hogy mely szófajú szavak maradnak meg, és melyek azok, amelyek átalakulnak vagy eltűnnek a fordítás során. Egy rövid teszt során megnéztük, hogy a SzegedParalell Korpusz (Tóth et al. 2008) illetve a Hunglish Korpusz (Varga et al. 2007) esetében mely szavak azok, amelyeknek a leggyakrabban nincs párja a másik nyelvben. Ehhez érdemes hozzátenni, hogy egyes esetekben ez szótárunk hiányossága is lehet, de legtöbbször nem erről van szó, illetve az algoritmus szempontjából ez nem is lényeges: az adott szótárat használva ezek a szavak tulajdonképpen úgy viselkednek, mint a stopszavak.

```
a:62616, hogy:15665, az:14511, is:11627, s:10595, nem:9572,
és:7707, meg:4516, el:4410, egy:4395, már:4255, sem:3419,
még:2812, e:2492, csak:2480, mint:2469, ha:2379, úgy:2341,
ez:2282, fel:2140, A:2097, de:2076, ki:1973, pedig:1857,
olyan:1305, ezt:1294, ott:1291, be:1286, majd:1278, azt:1239,
arra:1236, minden:1223, hát:1186, maga:1169, volt:1160,
vele:1138, Nem:1133, őket:1067, nagy:1050, akkor:1030,
De:960, volna:958, én:903, hozzá:890, így:883, le:878,
aki:856,mert:856, benne:855, itt:836
```

4.6. ábra: Magyar szavak, melyeknek nincs megfelelője az angol fordításban és az előfordulási gyakoriságuk a SzegedParalell Korpuszban (99 300 mondatpár)

```
a:512614, az:140951, hogy:113419, nem:76761, is:69974,
és:64415, s:45574, meg:43880, A:34303, egy:33716, el:32589,
már:27852, sem:21244, Nem:20740, ki:19500, e:19346,
még:18677, ha:18414, csak:18251, úgy:16522, mint:15389,
ez:15180, van:14233, olyan:13232, fel:12995, be:12938,
de:12736, azt:11592, kell:11460, Az:10428, minden:10045,
```

```
akkor:9908, őket:9346, pedig:9291, ezt:9119, volt:8816,
ott:8760, De:8726, majd:8704, arra:8273, én:7946, le:7845,
vele:7770, vagyok:7583, itt:7430, amit:7331, amely:7270,
te:7267, Ez:7003, hozzá:6897
```

4.7. ábra: Magyar szavak, melyeknek nincs megfelelője az angol fordításban és az előfordulási gyakoriságuk a Hunglish Korpuszban (1 301 700 mondatpár)

```
the:111917, of:64082, and:60256, to:49716, a:39573, in:33119,
I:32823, that:22202, it:18408, he:16688, with:16314,
his:16130, for:14279, had:14190, as:14180, not:12508,
you:12502, is:12437, her:11577, my:11406, be:11322,
The:11240, s:10586, at:10444, on:9794, have:9505, which:9419,
by:9273, but:8902, me:8602, from:8521, him:7742, this:7713,
she:7291, all:6710, so:6693, they:6575, an:5730, their:5709,
or:5600, one:5514, could:5228, said:5113, are:5090,
been:5048, we:4951, them:4944, no:4768, He:4727, there:4458
```

4.8. ábra: Angol szavak, melyeknek nincs megfelelője a magyar fordításban és az előfordulási gyakoriságuk a SzegedParalell Korpuszban (99 300 mondatpár)

```
the:818321, of:432800, to:377146, and:361797, a:283826,
in:236198, I:217910, that:155791, you:146520, it:132097,
for:113081, he:111064, s:110159, with:103306, is:100833,
be:99180, his:96176, on:90164, not:86259, The:81924,
had:80337, as:79995, at:70875, by:62949, this:62771,
have:59397, said:58528, t:58187, her:56259, him:55057,
or:54581, me:53811, from:52924, all:51366, my:49994,
which:48537, they:47193, are:46732, He:44626, but:44621,
shall:44616, You:40440, It:40039, one:38852, she:38414,
out:37944, an:37512, we:35832, up:34772, them:34295
```

4.9. ábra: Angol szavak, melyeknek nincs megfelelője a magyar fordításban és az előfordulási gyakoriságuk a Hunglish Korpuszban (1 301 700 mondatpár)

Ezek alapján frissítettük a stopszavak listáját, és az alábbi szavak is bekerültek a listánkba:

- magyar: e, úgy, fel, pedig, olyan, ezt, ott, be, majd, arra, hát, maga, vele, őket, akkor, volna, én, hozzá, így, le, mert, benne, itt, , sem, vagyok, amit, te, hogyan, ban, ben, nak, nek, ig, tól, os, es, án, én, ra, re
- angol: had, her, me, him, she, so, could, said, them, no, there, shall, would, then, d, ve, things, didn, wasn, couldn, doesn, isn, wouldn

Ezenfelül megnéztük azt is, hogy mely szavaknak nincs fordítása, és nem szerepel az általunk használt szótárban sem (SZSzótár) Érdekes módon a Hunglish esetében 1 676 alkalommal a „hogy” szó megjelent az angol oldalon, valamint 2 112 alkalommal az „az” szó. Ez a korpusz hibája, erre oda kell figyelni a jövőben, de ez nem okozott számunkra különösebb nehézséget.

Ahhoz, hogy a szavakat meg tudjuk feleltetni egymásnak, illetve ki tudjuk keresni egy szótárból, szótövezni kell a azokat. Ennek a részleteiről a következő fejezetben lesz szó, de már most lényeges kiemelni, hogy számos szónak lehet ugyanaz a szótöve (állnak, álltok, álla), és egy szónak több lehetséges szótöve is lehet (alma: alom, alma). Mivel nem használunk jelentésegyértelműsítést, bizonyos esetekben ez téves hasonlóságot jelezhet szavak, illetve mondatok között, de mint látni fogjuk, ez nem befolyásolja negatívan az algoritmus működését. Ugyanakkor előnye is van az ilyen kétértelműségnek, az angol nyelvben például számos főnév lehet ige is, sőt, van hogy fordításnál bizonyos esetekben változik a szófaj, ezekre az algoritmus nem érzékeny, hiszen nem veszi figyelembe a szófajokat.

Mivel szózsákot alkalmazunk, ezért a szavak megfeleltetése egymásnak nem egyértelmű, ugyanakkor fontos, hogy két mondatot akkor tekinthetünk egymás lehetséges fordításának, ha minél több szó feleltethető meg a másik zsákban lévő szavakkal. Ahhoz, hogy ez érthető legyen, képzeljük el az alábbi két mondatot:

- A kobold kószált az erdőben
- He trolled around the pile of wood.

Ha szavanként végigmegyünk a magyar mondaton (kobold, kószál, erdő), és összehasonlítjuk az angol szózsákunkkal (troll, around, pile, wood), akkor látni fogjuk, hogy mindhárom szó egy-egy fordítása megtalálható az angol szózsákban.

- kobold → troll
- kószál → troll
- erdő → wood

Ugyanakkor könnyű belátni, hogy ez nem helyes, hiszen a *troll* szót kétszer vettük figyelembe, így hibásan feltétezzük, hogy minden szónak megvan a fordítása a másik oldalon. Ezért az algoritmus kiszedi a zsákból azokat a szavakat, amelyeket már

felhasznált egy azonossághoz, így küszöbölve ki a hibás találatokat. Így már megfelelő eredményt kapunk:

- kobold → troll
- kószál →
- erdő → wood

Mivel a szózsákok összehasonlítása, a szótár használata erőforrás-igényes művelet, ezért létrehoztunk gyorsszűrő metrikákat is, amelyek kizárják, hogy a két mondat egymás fordítása legyen, és ezek lefutnak, a tényleges algoritmus előtt. Ha ezek azt adják eredményül, hogy a két mondat egymás lehetséges fordítása, akkor kiszámoljuk a tényleges hasonlóságot, ha nem, akkor kihagyjuk a további, erőforrásigényes számításokat.

Az első ilyen gyorsszűrő metrika a mondatok hossza alapján tesztel. A mondatok hosszát a stopszavak nélküli szavak számával (*n*') definiáljuk. A teszt igen egyszerű, és egy nagyon biztos felső mértéket ad. Semelyik mondat nem lehet több mint kétszer olyan hosszú, mint a fordítása, kivéve az egész rövid (5 vagy kevesebb szó) mondatokat.

$$n' = | W^x \setminus W^x_{stop} |$$

ahol W^x a mondat összes szava és W^x_{stop} az adott nyelv stopszavai. Azaz akkor dobunk el egy mondatpárt, ha $m' > 5$ és $m' > 2 \cdot n'$

Ezeknek a mértékét egy rövid teszt keretében a SzegedParalell és Hunglish korpuszokra ki is számoltuk (lásd 4.10. és 4.11. táblázat). A két táblázatban aláhúztuk azokat a cellákat, amelyek még megfelelnek a kritériumainknak, és ezáltal jól látszik, hogy csak szélsőséges esetben dobunk el jó találatot, ráadásul az ennyire különböző hosszúságú mondatokat a hasonlósági metrika se találná hasonlónak, a sok hiányzó szó miatt.

A másik gyorsteszt már a fordítás után, de még a részletes algoritmus kiszámítása előtt fut le és azt nézi, hogy van-e elég azonos szó a két zsákban ($S^x \cap S^y$).

$$T_{min} = n' / 3 - 1 \text{ ha } n' \geq 6$$

$$T_{min} = 1 \text{ ha } n' < 6$$

$$T_{min} \leq | S^x \cap S^y |$$

hun	eng																			
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
1	1864	823	151	28	5	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0
2	558	2668	1193	413	123	25	10	3	2	2	2	1	0	0	0	0	0	1	0	0
3	168	1119	2541	1531	583	191	60	20	7	4	1	2	0	2	0	1	0	1	0	0
4	54	432	1298	2141	1548	677	275	109	31	15	6	4	2	0	1	1	0	0	0	0
5	18	113	578	1326	1925	1330	710	299	105	44	13	9	2	2	0	1	1	1	0	0
6	5	52	195	647	1216	1585	1199	718	333	143	55	32	10	6	1	3	0	1	1	0
7	2	16	71	247	629	1118	1358	1089	669	356	153	76	28	15	9	7	3	3	2	1
8	2	6	22	107	298	597	1047	1143	997	626	364	175	79	31	18	13	6	1	4	2
9	0	4	11	31	125	315	579	898	980	874	554	334	145	104	44	23	18	6	5	1
10	0	1	1	26	61	131	325	579	785	876	714	485	321	177	107	51	21	12	9	4
11	0	2	0	10	20	91	176	347	507	789	765	679	470	323	173	108	61	35	21	8
12	0	0	1	2	6	30	80	164	305	496	658	669	578	432	278	152	80	41	30	12
13	0	0	0	5	2	17	30	103	168	277	513	572	608	511	381	243	150	91	58	24
14	0	0	2	2	1	7	15	55	95	160	315	405	509	513	465	328	242	139	91	51
15	0	0	0	1	5	3	6	15	47	89	195	253	338	447	444	401	305	207	123	57
16	0	0	0	0	2	1	4	6	35	47	110	182	252	312	382	394	340	241	185	113
17	0	0	0	0	2	2	3	9	6	22	65	99	161	248	320	340	311	298	199	173
18	0	0	0	1	3	2	3	3	10	14	24	50	93	133	218	278	286	243	225	173
19	0	0	0	0	0	1	4	2	3	1	14	37	57	98	152	185	237	279	218	211
20	0	0	0	0	1	1	0	2	0	2	12	14	32	50	84	112	171	209	190	201
21	0	0	0	0	0	0	3	3	0	7	5	8	9	24	52	74	105	136	167	158
22	0	0	0	0	0	1	2	1	1	2	2	6	7	17	29	51	87	104	129	134
23	0	0	0	0	1	0	1	0	1	2	4	1	5	8	18	38	48	73	81	92
24	0	0	0	0	1	1	1	1	0	2	1	2	4	8	12	18	32	33	48	60

4.10. táblázat: SzegedParalell Korpusz angol és magyar mondatainak a szóhossza, stopszavak nélkül, egymáshoz viszonyítva, előfordulási gyakoriság (99 000 mondat)

hun	eng															
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**
1	110839	32008	5325	763	89	13	2	2	0	0	0	0	0	0	0	2
2	23819	122233	39884	10651	2170	385	63	9	12	14	2	0	0	0	0	0
3	4470	33088	72362	34051	11367	3076	707	155	47	25	22	9	4	2	0	0
4	852	7829	27871	45984	25844	10212	3417	1062	283	101	42	32	12	2	1	0
5	159	1699	8970	22770	29854	19506	8743	3641	1225	422	115	61	27	21	11	2
6	27	416	2554	8913	19090	20992	15412	7772	3312	1379	532	221	77	28	15	8
7	11	106	661	3050	8173	13718	15487	11381	6353	3143	1463	581	261	86	38	25
8	5	20	165	940	3150	7208	11024	12687	8910	5550	2832	1367	654	259	110	56
9	7	13	190	310	1183	3091	6308	8792	9713	8371	4594	2522	1308	668	280	113
10	6	21	160	129	421	1336	3051	5397	7568	9067	6290	3940	2221	1269	579	340
11	1	13	48	60	163	554	1335	2936	4625	6016	5967	4895	3430	2001	1108	597
12	0	1	19	66	90	216	597	1421	2711	4038	4804	5030	4422	2987	1802	1040
13	0	0	2	29	52	106	295	718	1398	2361	3371	4115	4095	3346	2374	1552
14	0	0	0	9	13	41	131	299	702	1257	2230	2961	3462	3290	2725	1968
15	0	0	0	3	8	26	51	140	367	663	1218	1969	2523	2909	2736	2223
16	0	0	0	0	2	6	23	68	188	351	725	1106	1686	2153	2344	2258
17	0	0	0	0	0	2	7	30	83	173	375	681	1009	1487	1815	2012
18	0	0	0	0	0	5	0	7	37	69	182	377	582	923	1250	1603
19	0	0	0	1	0	0	3	5	26	35	93	191	364	578	857	1125
20	0	0	0	0	0	0	1	6	6	22	57	103	194	347	578	732
21	0	0	0	0	0	1	0	1	2	7	24	63	93	215	285	448
22	0	0	0	0	0	0	0	1	2	2	14	28	60	127	196	311
23	0	0	0	0	0	0	1	2	0	1	7	13	38	55	102	190
24	0	0	0	0	0	0	0	0	0	3	1	2	16	34	58	98

4.11. táblázat: Hunglish Korpusz angol és magyar mondatainak a szóhossza, stopszavak nélkül, egymáshoz viszonyítva, előfordulási gyakoriság (1 300 000 mondat)

Azaz a szavak majdnem harmadához kell, hogy legyen fordítás. Rövid mondatok esetében ez a képlet negatív számot ad vissza, ilyenkor 1-gyel számolunk. Ez a képlet is empirikus kutatás eredménye, és csak annak a feltételnek kell megfelelnie, hogy jó fordítást nem zár ki. Ennek tökéletesen megfelel: nem csak logikailag látható be, hogy ha a szavak harmadánál kevesebb rendelkezik fordítással, akkor az valószínűleg nem jó fordítás, de a tesztek is azt mutatták, hogy ez nem okoz kimutatható találatvesztést.

Az utolsó gyorsszűrő sokkal egyszerűbb. A *k* darab legmagasabb *Sim* értékű találatot nyilvántartjuk, és az adott mondat csak akkor lesz egy potenciális jelölt, amennyiben van esélye bekerülni ebbe a listába. A listába a mondatok a *Sim* értékük szerint, rendezve kerülnek be, és egy magasabb értékű találat kiszorít egy alacsonyabbat. A *Sim* értékre $\alpha \geq 1$ esetében egy jó felső becslés a korábban említett és már kiszámolt azonos szavak száma a két szózsákban ($S^x \cap S^y$), hiszen ennél csak kisebb lehet a tényleges érték, mert

$$\mathrm{Sim}\left(S_n^x, S_m^y\right) = \min\left(\alpha \cdot \left|S_n^x \cap S_m^y\right| - \beta \cdot \left|S_n^x \setminus S_m^y\right|, \alpha \cdot \left|S_m^y \cap S_n^x\right| - \beta \cdot \left|S_m^y \setminus S_n^x\right|\right)$$

ezért

$$\mathrm{Sim}\left(S_n^x, S_m^y\right) \leq \alpha \cdot \left|S_n^x \cap S_m^y\right| - \beta \cdot \left|S_n^x \setminus S_m^y\right|$$

ezért

$$\mathrm{Sim}\left(S_n^x, S_m^y\right) \leq \alpha \cdot \left|S_n^x \cap S_m^y\right|$$

ha $\alpha \geq 1$ akkor

$$\mathrm{Sim}\left(S_n^x, S_m^y\right) \leq \left|S_n^x \cap S_m^y\right|$$

Ezzel a három teszttel jelentősen tudjuk csökkenteni az algoritmus futási idejét. Ugyanakkor a futási idő még mindig lineáris, azaz a bemeneti mondatot minden esetben össze kell hasonlítani az adatbázisban lévő összes többi mondattal. Ez semmiképp se megoldhatatlan feladat, ugyanakkor jó lenne szublineáris időben, indexáltan keresni, hiszen ezt az algoritmust a gyakorlatban igen nagy adabázisokra is alkalmazni kell, például teljes egyetemi digitális könyvtárakra, Wikipédia adatbázisa vagy akár egy webes kereső által visszaadott oldalakra. A keresések száma nem lesz olyan nagy, mint az adatbázis maga, hiszen terveink szerint nem az adatbázisunkon belül keresünk majd hasonlóságokat, hanem külső dokumentumokat, például diplomákat hasonlítunk össze

az adatbázissal. Viszont így is fel kell készülnünk nagy mennyiségű dokumentum egy időben való beérkezésére (diplomavédési időszak), így fontos, hogy a keresési időt minél jobban leszorítsuk. Ezzel a problémával részletesen a „Keresési idő csökkentése indexált kereséssel" című 4.3.2 fejezetben foglalkozunk. Jelenleg csak a megoldás irányát vetítjük előre, hogy a következő fejezetek tartalma érthető legyen. Az imént ismertetett algoritmus elé tettünk egy előszűrőt, egy indexált keresést, amely a mondat szavait használja fel ahhoz, hogy a nagy méretű adatbázisból kiszűrje azokat, amelyek a legnagyobb valószínűséggel hasonlók lesznek a bemeneti mondathoz.

4.2.3. A hasonlósági eredmények értelmezése

Az utolsó lépésben azt kell eldönteni, hogy két mondat, illetve két dokumentum hasonlít-e egymásra vagy sem, azaz azt, hogy mikor jelezzük ki a hasonlóságot a felhasználónak. Ehhez az alábbi egyszerű metrikát vezettük be. Megszámoljuk egy adott forrásdokumentumon belül (jelen esetben: Wikipédia-szócikk) a találatok számát és azok egymáshoz viszonyított elhelyezkedését. Ez a megoldás különbözik Kasprzak (2010) rendszerében implementáltaktól, hiszen itt nem egy bináris találat/nem-találat döntést kapunk, hanem egy folytonos skálán egy értéket, hogy két mondat mennyire hasonlít egymáshoz.

Jelölje Sim_a és Sim_b a hasonlósági metrika által adott hasonlóság mértékét az *a*-adik és *b*-edik mondatok esetében. Definiáljunk két állandót, két hasonlósági mértéket SIM_1 és SIM_2, ahol $SIM_1 < SIM_2$ valamint egy távolság-értéket, *d*-t. Ezek alapján leírhatjuk, hogy akkor tekintjük az adott dokumentumot találatnak, ha az alábbi igaz:

- $\mathrm{Sim_a} \geq \mathrm{SIM_2}$ vagy
- $\mathrm{Sim_a} \geq \mathrm{SIM_1}$ és $\mathrm{Sim_b} \geq \mathrm{SIM_1}$ és $|a - b| < d$

Azaz: az adott töredék értéke nagyobb SIM_2-nél, vagy kisebb ugyan, de nagyobb SIM_1-nél és *d* távolságon belül van egy másik SIM_1-nél nagyobb hasonlósági értéket kapott mondat is. A jelenlegi rendszer SIM_1, SIM_2 és *d* értékeire sorban 0, 8 illetve 10 értéket használ, de ez a felhasználási területtől függően állítható.

4.3. Az új fordításiplágium-kereső algoritmus vizsgálata

Mielőtt részletesen belemennénk az algoritmus eredményeinek az elemzésébe, vegyük sorra, hogy milyen tesztadatokat, eszközöket és adatbázisokat használunk. Ezek mind befolyásolhatják az eredményeket, ezért érdemes részletesen foglalkozni velük.

4.3.1. Tesztkörnyezet kialakítása

Ahhoz, hogy az algoritmust tesztelni tudjuk, szükségünk van olyan szövegekre, amelyeknek ismerjük a fordítását, valamint egy olyan hatalmas szöveges adatbázisra, amely lehetővé teszi, hogy hamis pozitív találatokat is teszteljünk, azaz egy olyan, amely már biztos tartalmaz hasonló mondatokat, hiszen 10 mondatból kiválasztani egy adott mondat fordítását egy igen rosszul teljesítő algoritmusnak se lenne gond. Az első szempontnak – azaz, hogy ismert legyen a fordítása – eleget tesz a Hunglish-korpusz (Varga et al. 2007), amely egy párhuzamos korpusz. Nagy mennyiségű adatbázisnak pedig a Wikipédiát választottuk, abból is az angolt. (Wikipedia) Amennyiben egy algoritmus képes egy Wikipédia méretű adatbázisból kiválasztani, a megfelelő mondato(ka)t, akkor elmondhatjuk, hogy jól működik. Utóbbira azért is esett a választás, mert sokan idéznek, illetve sokan plagizálnak is sajnos a Wikipédiából, így gyakorlati haszna is van egy olyan keresőnek, amely kiemeli a Wikipédiából átvett részeket egy dolgozatban. Ezen felül szükségünk volt még egy szótövező algoritmusra, egy mondatra bontó algoritmusra, egy szószedetre valamint egy automatikus fordítóra, utóbbival nagymennyiségű „plagizált” tartalmat állítottunk elő a teszteléshez. A továbbiakban részletesen ismertetjük a tesztkörnyezetben alkalmazott programokat és adatbázisokat.

4.3.1.1. A Hunglish korpusz

A Hunglish egy kétnyelvű, magyar-angol párhuzamos korpusz, ahol a magyar és az ebből fordított angol mondatok meg vannak feleltetve egymásnak. Az oldalukon ezt olvashatjuk az adatbázisról:

> „The sentence pairs were created from document pairs by automatic methods. Note that sometimes parts of the documents are not in perfect correspondence. (Due to liberal translation, or even skipping of some segments by the translator.) These may lead to erroneous sentence pairs. As our database consists of more than two million sentence pairs, manual elimination of these errors is unfortunately infeasible.”

A Hunglish adatbázist átvéve, néhány fájlról kiderült, hogy nem magyar nyelvű fordításokat tartalmazott, ezeket töröltük. Nem dolgoztuk fel továbbá azokat a sorokat se, amelyek csupa nagybetűvel íródtak, mert ezek jelentős része cím és egyéb

viszonylag értelmetlen szöveg volt, ilyenek főleg az *EU_Law* fájlokban fordulnak elő. Nem ellenőriztük kézzel a fájlokat, így továbbra is biztos maradtak hibák, csak a legalapvetőbb gépi teszteket végeztük el, hogy feldolgozható legyen az állomány.

A tisztítás után 2055 fájl maradt, ezekről egy lista található a 9.7 mellékletben. Ezekben a fájlokban összesen 1 296 912 sor található, a fájl formátumából (bi) adódóan ez annyit jelent, hogy ennyi mondatpár található bennük. Ez magába foglalja az általunk fel nem dolgozott hibás sorokat is, mint például:

- Üres, vagy majdnem üres sorok / mondatok
- Csak számokat tartalmazó mondatok (fejezetcímek)
- Csupa nagybetűvel írt mondatok (fejezetcímek)
- Mondatok, amelyek magyarul és angolul ugyanúgy vannak (le nem fordított, egyszavas mondatok)

Feldolgozás, tisztítás után 2 281 730 mondatot kaptunk, melynek fele angol, fele magyar, azaz 1 140 865 mondatpárt. Azaz a tisztítás során 156 047 mondatpárt töröltünk.

Ennek a gyűjteménynek előnye, hogy vegyesen tartalmaz szépirodalmi és hétköznapi fordításokat is; profi fordítók munkáját, és amatőr fordításokat is. A filmfeliratok (subtitles.bi) például általában amatőr fordítás eredménye, és sok félrefordítást, szlenget tartalmaz.

4.3.1.2. A SzegedParalell korpusz

A Szegedi Tudományegyetem Informatikai Tanszékcsoport Nyelvtechnológiai Csoportja által létrehozott párhuzamos korpusz, amely összesen 63 810 mondatpárt tartalmaz. (Tóth 2008) A honlapjukon az alábbi adatokat lehet megtudni róla:

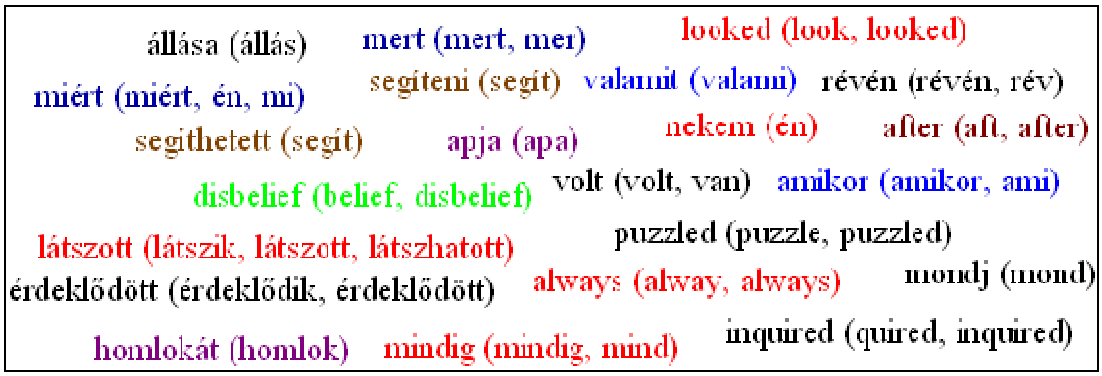
- Nyelvkönyvi mondatok: 2 937
- EU-ról szóló anyagok: 1 518
- Horizon Magazine: 3 980
- Ingatlan Info: 1 340
- Irodalom: 53 985
- Egyéb: 50

Igen kis méretű, de jó minőségű tesztállomány, amelyet érdemes megvizsgálni az algoritmussal.

4.3.1.3. Hunspell

Szótövező szoftverek közül kettőt próbáltunk ki, amelyek elérhetőek voltak és tudtak magyarul. A Magyarláncot (Zsibrita et al. 2009), és a Hunspell-t (Németh et al. 2004). Előbbiről gyakorlatilag semmilyen információ nem érhető el az interneten, a szegediek fejlesztették, de sajnos már nem felel meg a kor elvárásainak, legalábbis az a változat, amit sikerült letölteni. Mivel a Hunspell használhatónak bizonyult, így a Magyarláncot a továbbiakban nem használtuk.

A Hunspellt a MOKK fejlesztette, 2003-2005 között. Elérhető hozzá az általunk használt magyar, angol és német nyelvekhez is megfelelő minőségű szótövező adatbázisok. Nem hibátlan a program, de megfelel annak amire nekünk szükségünk van. Számos szolgáltatása közül nekünk csak a szótövezésre van szükségünk. A 4.12. ábrán látható egy kis szófelhő, a szavakkal és azok mögé írt szótöveivel. Látható pár hibás megoldás is (always → alway), de a legtöbb szótő jó. Ez a szófelhő nem reprezentatív, szándékosan szerepelnek benne hibás és több szótővel rendelkező szavak.



4.12. ábra: Hunspell szótövező

Az szoftver használatának az elején olyan problémába ütköztünk, hogy egy szó tövének lekérdezése – az általunk használt gépen – 100 ms-ot vett igénybe, ami nem tűnik soknak, de ha a teljes Wikipédiát szeretnénk szótövezni, akkor belátható, hogy még egy multiprocesszoros rendszeren is túl sokat futott volna az alkalmazás. A mi esetükben ez 2 év lett volna. A tesztek során kiderült, a program betöltése veszi el a legtöbb időt, így elkezdtük kötegelt módban meghívni a programot, gyakorlatilag előbb kiszedve az összes szót a Wikipédia egy nagyobb szeletéből, majd arra meghívva a szótövezőt. Így

már egy nap alatt elkészült a teljes szótövezési feladat, ami elfogadható, hiszen egy 30GB-os adathalmaz szótövezéséről van szó. Az eredmény közel 50GB lett, de ha kihagyjuk az ismétlődő szavakat, akkor csak 500MB. Ez a fordítási cache összesen 46 millió sort tartalmazott, rendezve viszont, és azokaz a szavakat kiszűrve, amelyeknek van szótöve 230 717 különböző szót és hozzájuk tartozó 311 361 (nem feltétlen különböző) szótövet kaptunk.

4.3.1.4. Wikipédia

A Wikipédia elvitathatatlan előnye, hogy hihetetlen mennyiségű tudásanyag halmozódott fel szócikkeiben az évek során. Ezt csak úgy tudták elérni, hogy bárki hozzáadhatot tartalmat, illetve egymást lektorálták a felhasználók. Amikor használjuk a Wikipédiát, és itt most elsősorban az angol nyelvű Wikipédiára gondolunk, legtöbbször megfelelő válaszokat kapunk a kérdéseinkre, elég jó minőségben. A Wikimedia Alapítvány, amelyik a Wikipédia fenntartója, elérhetővé tette letölthető formában a teljes adatbázist. (WikiDump) Ez az adatbázis XML formátumban van, és pár kisebb hiányosságtól eltekintve az összes oldalt tartalmazza wiki formátumban. Ezzel el is érkeztünk az adatbázis árnyoldalához. Ennek a szabadságnak, a kötetlen szerkesztésnek és a sok képzetlen szerkesztőnek az eredménye, hogy az adatbázis rengeteg hibát tartalmaz.

4.3.1.5. Wiki2Txt

A kereséshez és az adatbázis feltöltéséhez nekünk szöveges formátumban van szükségünk erre a nagy mennyiségű adatra. Jelentős mennyiségű időt fordítottunk arra, hogy olyan szoftvereket kerestünk, amelyek képesek a wiki formátumú szöveget – pontosabban MediaWiki formátumú XML fájlokat – átalakítani a mi igényeinknek megfelelően sima szöveges formátumúvá, anélkül, hogy egy teljes MediaWikit fel kéne telepítenünk. Utóbbit azért zártuk ki, mert olyan lassú, hogy lehetetlen lett volna a teljes adatbázis legenerálása, átalakítása HTML formátumú szöveggé. Végül egy saját konvertert írtunk, amely képes az XML formátumú MediaWiki dumpból a tartalommal bíró szócikkeket kiszedni, és azokat sima szöveggé konvertálni. A konverter PHP-ban íródott, a könnyű fejlesztés, és az eddigi komponensekkel való kompatibilitás érdekében. Egyszerűen csak a Wiki2Txt nevet kapta a fejlesztés során.

A konverter által visszaadott eredményt kézzel teszteltük, hiszen nincs egy kész átalakított adatbázisunk, amihez viszonyítani tudnánk. A konverternek számos apró dologra oda kellett figyelnie, amire az általunk kipróbált többi program sajnos nem figyelt, vagy nem mindenre, így mind hibás eredményt adtak. Többek között ezekkel a problémákkal szembesültünk:

- Lezáratlan elemek (tagek)
- Hibás elemek
- Ugyanarra a célra párhuzamos jelölések alkalmazása (pl.: Wiki és HTML)
- Egymásba ágyazott, azonos viselkedésű elemek pl.: <pre>, <code>, <nowiki>, <source>
- HTML megjegyzések
- Template-ek
- Szövegbe ágyazott bináris fájlok
- Táblázatok (Wiki és HTML, bármilyen sorrendben és mélységben egymásba ágyazva)

A Wiki2Txt előnye, hogy Windows és Linux alatt is megy és mindenképp visszaad valami szöveges eredményt, akkor is, ha nagyobb hibával találkozik. Sokkal gyorsabb, mint a MediaWiki: egy nap alatt át tudja alakítani a teljes Wiki adatbázist. Ezen felül megtartja a wiki oldalak címét, így ez később is rendelkezésünkre áll, amikor tovább dolgozunk a szövegeken.

A Wikipédia általunk, 2011 elején letöltött változata XML formátumban 28GB. Ebből a Wiki2Txt 13GB-nyi szöveget állít elő. Ebből a szövegből egy következő lépésben eltávolítottunk minden nem alfanumerikus karaktert, és így 12GB-nyi tiszta szöveget kaptunk. Ez a szöveg a bemenete az előző fejezetben ismertetett szótövezőnek, amely 28GB-nyi szótőállományt generált ebből. Ezek képezik a bemenetét az algoritmusunknak, amelyik a sima szövegből és a szótőlistából előállítja az adatbázis-feltöltéshez szükséges fájlokat.

4.3.1.6. Mondatra bontás

A fejlesztés során az alábbi okok, de főleg a könnyű testreszabhatóság miatt saját mondatszegmentáló algoritmust készítettünk:

a) A Wikipédia szövege – még szöveges formátumra alakítás után is – tartalmazott hibákat, például mondatok rendszeresen egybe vannak írva a következővel (hiányzik a szóköz a mondatot lezáró írásjel után).

b) Olyan algoritmusra volt szükségünk, ami gyors, és segítségével elkerülhetjük az újabb köztes fájlok létrehozását.

c) Mivel ekkor már látszott, hogy a teljes folyamat igen erőforrás-igényes, ezért szerettünk volna minél kevesebb külső programot használni, hogy a program minél több gépen képes legyen futni.

A Wikipédia hibás szövege és szakszócikkei komoly kihívás elé állítják a szegmentálót, de a kézi átnézés során, és a tesztek alapján – ahol a hibás találatok esetében kézzel ellenőriztük az eredményt – úgy találtuk, hogy a célnak megfelel. A hibák jelentős része abból adódik, hogy egy mondatot olyan helyen is felbont két részre, ahol nincs mondathatár, például nem ismer egy rövidítést.

Az angol Wikipédia általunk használt mentése több mint 3,5 millió szócikket tartalmaz. Ez a 3,5 millió szócikk összesen 183 759 808 mondatra bomlik, azaz átlagosan 50 mondat szócikkenként. Mivel a mi átalakításaink során minden cím, táblázatsor stb. egy mondatra bomlik, és a mondatra bontás is valamennyire felfelé torzít így ez az 50 mondat egy reális szám.

4.3.1.7. Google fordító

Az algoritmus teszteléséhez, mesterséges plágiumok előállításához a Google fordítót használtuk. A Hunglish korpusz jó kiindulási alap a tesztekhez, de mint korábban említettük, számos hibát tartalmaz, ezen felül elérhető még a SzegedParalell korpusz, de az csak néhány tízezer mondatpárt tartalmaz, így annak a mérete se elegendő ahhoz, hogy nagyobb mennyiségű automatikus teszteket futtassunk.

Ahhoz, hogy tesztelni tudjuk az új algoritmust, annak teljesítményét össze kell hasonlítani a mások által használt algoritmusokkal. A korábban már említett CLEF konferencián szereplő fordításiplágium-kereső algoritmusok automatikus fordítót

használnak a működésük során. Ezt az algoritmust is implementáltuk, és a tesztek eredményét a 4.4 fejezetben ismertetjük.

A könnyű elérhetőség és az API felület miatt esett a választás a Google fordítójára. (GoogleT) A Google fordító egy statisztikai, és nem szabály alapú fordító, azaz nem egy belső nyelvtani értelmezést, és szótárat használ a fordításra, hanem párhuzamos korpuszok segítségével – mint amilyen a Hunglish és a SzegedParalell – tanul meg fordítani. Ez azért lényeges, mert így a tesztek, amelyeket a Google fordító segítségével végzünk el, biztos, hogy független az általunk használt algoritmustól, szószedettől, szótártól, és nem egyszerűen csak „visszafejtjük" a Google algoritmusát.

A Gooogle egy könnyen elérhető API felületet biztosít a fordítóhoz, így egyszerűen fel lehet készíteni bármilyen programot gépi fordításra. Jelenleg 100 000 karakterben korlátozza a naponta lefordítható szövegek mennyiségét egy hozzáféréssel, de bárki, akinek van Google fiókja, lefordíthat ennyit egy API-kulcs segítségével. Ezért ez a korlátozás nem okozott gondot, csak megfelelő mennyiségű kulcsra volt szükségünk. A tesztek elvégzésének a végére a Google már megszüntette ezt az ingyenes elérést, de addigra a gépi fordítással készült tesztadatbázis már rendelkezésünkre állt.

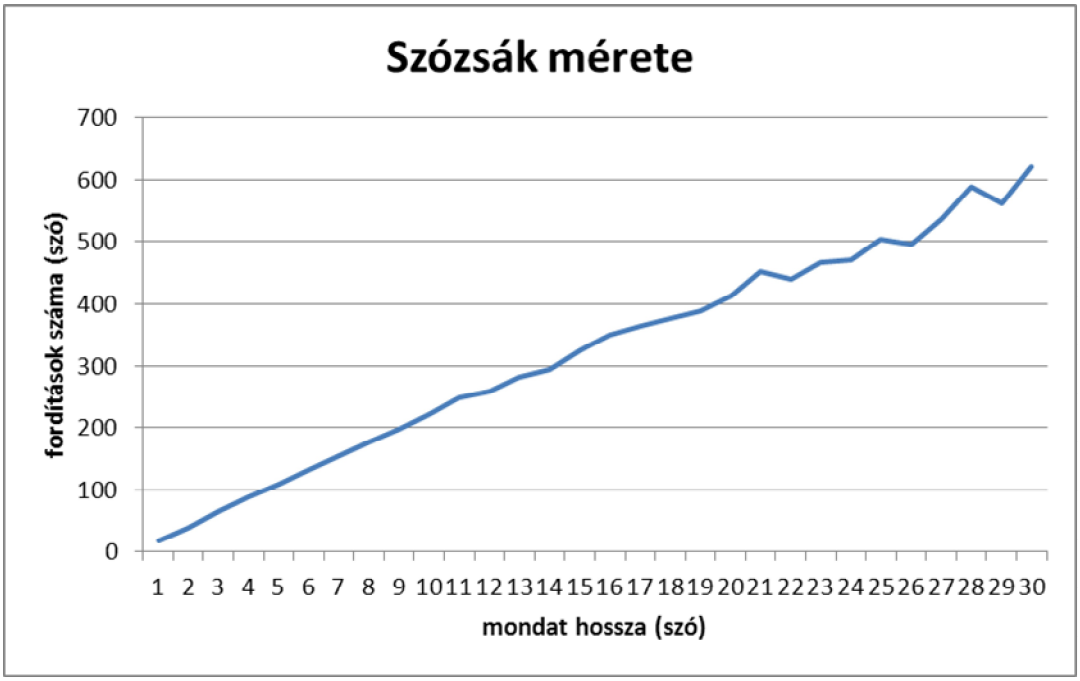
4.3.1.8. Angol-magyar párhuzamos szószedet

Ahhoz, hogy egy angol és egy magyar szó azonosságát (ahogy azt a 4.2.1 fejezetben definiáltuk) meg tudjuk állapítani, szükségünk van egy szószedetre, egy egyszerű szótárra. Ehhez a SZTAKI online szótárát (SZSzótár) használtuk. Mivel azt is tesztelni szeretnénk, hogy a szótár mérete, illetve a hiányzó fordítások mennyire befolyásolják az algoritmust, ezért egyéb, online elérhető szótárakkal illetve szószedetekkel is kísérleteztünk. A kutatás jelentős részét az összes szótár uniójával végeztük; ahol nem, ott erre külön utalunk.

A teljes szótárban közel egymillió angol-magyar szó- illetve kifejezéspár található. Ez magában foglalja többek között a Hunglish szótárát is, amely egy automatikus eszközökkel a Hunglish párhuzamos korpuszból készült szótár, és igen sok szlenget és hibát, félrefordítást tartalmaz. A tesztek során az derült ki, hogy ezek a hibák nem befolyásolták hátrányosan az algoritmusunkat, nem okozott nagy mennyiségű hibás találatot. A tapasztalatok alapján inkább a hiányzó fordítások érintették negatívan az eredményt, mint a hibásak.

4.3.2. Keresési idő csökkentése indexált kereséssel

Ebben a fejezetben a keresési tér csökkentésére használt információ-visszakeresési (information retrieval) algoritmust mutatjuk be, és annak hatását a keresésre. Mivel a használt hasonlósági metrika lineáris, így nagyobb adatbázis esetén már nem végezne elfogadható időn belül az algoritmus, ha előtte nem keresnénk meg egy sokkal gyorsabb (de pontatlanabb) algoritmussal azokat a lehetséges mondatokat, amelyekhez a legnagyobb eséllyel hasonlít a bemeneti mondatunk. Ez a keresés tulajdonképpen egy olyan ritka mátrixban való keresés, ahol a mátrix sorai az adatbázisban tárolt mondatok, az oszlopai az adott nyelv szavai, és „1” jelöli azt, ha egy adott mondatban megtalálható egy adott szó, és „0” amennyiben nem. Amennyiben teljes egyezést keresnénk – azaz pontosan tudnánk, hogy mely szavak vannak benne a mondatunkban – akkor könnyebb dolgunk lenne megtalálni a mátrix megfelelő sorait, ugyanakkor nem pontos egyezést keresünk, hiszen a fordítás során nemcsak hogy nem tudjuk mely fordítását választotta az adott szónak a fordító, de bizonyos szavak „el is vesznek”, illetve „keletkeznek” a fordítás során, a nyelvek közötti különbségek miatt. Így tulajdonképpen azokat a sorokat keressük, amelyekben az általunk keresett szavakból (szózsákból) a legtöbb szó megtalálható. A szózsák mérete egyenes arányban függ az eredeti mondat szavainak számától, mint azt a 4.13. ábra szemlélteti.



4.13. ábra: A szózsák mérete az eredeti mondat hosszának függvényében

Egy 700 000 szópárt tartalmazó szótárral, egy angol mondatból képzett magyar szózsák mérete átlagosan 20-szor annyi szót tartalmaz, mint az eredeti mondat (lásd 4.13. ábra).

Ilyen keresésre képes szoftverek készen letölthetőek az internetről, így a kutatás elvégzéséhez is egy ilyen, ingyenesen elérhető, széles körben használt kész szoftvermegoldást, az Apache Solr-t, használtunk (Solr), melynek honlapján ezt lehet olvasni:

```
Solr is the popular, blazing fast open source enterprise
search platform from the Apache Lucene project. Its major
features include powerful full-text search, hit highlighting,
faceted search, dynamic clustering, database integration,
rich document (e.g., Word, PDF) handling, and geospatial
search. Solr is highly scalable, providing distributed search
and index replication, and it powers the search and
navigation features of many of the world's largest internet
sites.

Solr is written in Java and runs as a standalone full-text
search server within a servlet container such as Tomcat. Solr
uses the Lucene Java search library at its core for full-text
indexing and search, and has REST-like HTTP/XML and JSON APIs
that make it easy to use from virtually any programming
language. Solr's powerful external configuration allows it to
be tailored to almost any type of application without Java
coding, and it has an extensive plugin architecture when more
advanced customization is required.
```

A kutatás során két dologra kellett választ kapni. Az első, hogy a Solr alkalmas-e erre a feladatra, a második, hogy a találati arányt illetve a pontosságot mennyiben rontja az információ-visszakeresésre használt Solr és mennyiben az utána alkalmazott új hasonlósági metrika. A következő fejezetekben ezt vizsgáljuk meg.

4.3.2.1. Indexált és nem indexált keresések összehasonlítása

A tesztekhez a teljes angol Wikipédiát, és egy abból véletlenszerűen kiválasztott kisebb, 65 000 mondatot tartalmazó párhuzamos korpuszt használtunk. A párhuzamos korpuszban az eredeti angol mellett a mondatoknak a Google fordítóval készült magyar és német fordítása található meg még. A szózsák legenerálásához, valamint a hasonlósági metrikához a 4.3.1.8 fejezetben ismertetett 700 000 szópárt tartalmazó angol-magyar, és a 150 000 szópárt tartalmazó angol-német szótárat használtuk.

A rendszert kézzel teszteltük egy igen rövid, körülbelül 100 mondatos, kézzel fordított Wikipédia korpuszon, amely megtalálható a 9.4 mellékletben. Az eredmények igen hasonlóak lettek, mint a gépi fordítású nagyobb korpusz esetében, de mivel a kézi korpusz mérete kicsi, így a továbbiakban a sokkal nagyobb gépi fordítású korpuszon elért eredményeket ismertetjük.

Az angol Wikipédiát – amely jelenleg 3,8 millió szócikkből áll – mondatonként, szótövezve és a stopszavakat kiszűrve feltöltöttük a Solr adatbázisba. Innét az első 50 találatot kérdeztük le a kísérletekhez, keresőkérdésnek a magyar, illetve német mondatok angolra fordított szózsákját használva.

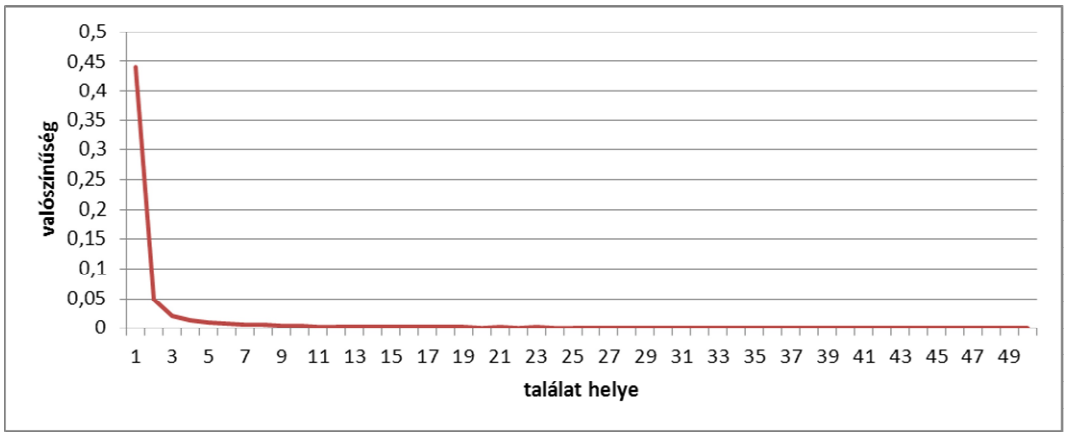
A továbbiakban a fedés (recall) és pontosság (precision) mértékeket az alábbi definíció alapján fogjuk alkalmazni:

$$fedés = \frac{|releváns\_dokumentumok \cap visszakapott\_dokumentumok|}{|releváns\_dokumentumok|}$$

$$pontosság = \frac{|releváns\_dokumentumok \cap visszakapott\_dokumentumok|}{|visszakapott\_dokumentumok|}$$

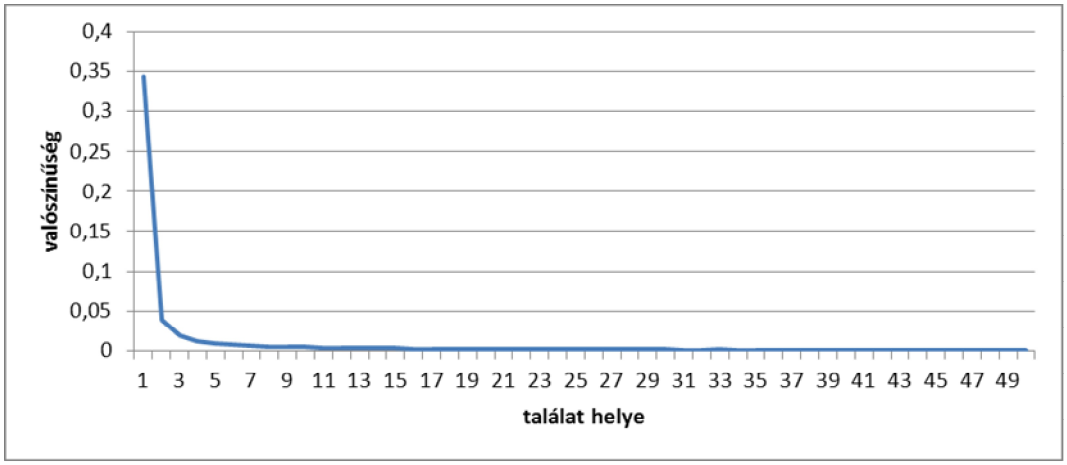
Azaz a fedés azt határozza meg, hogy a jó találatoknak, amiket vissza kellet volna kapnunk, mekkora hányadát találta meg ténylegesen a rendszer. A pontosság pedig azt, hogy a visszakapott eredményben mennyi a jó, releváns találat. Előbbi mérése triviális a tesztadatbázison, utóbbi nehezebb, mert nem biztos, hogy csak az általunk ismert találat az egyetlen érvényes és jó találat, hiszen például a Wikipédiában több oldalon is találunk ismétlődő, más oldalakról származó szövegrészeket.

A kísérlet eredménye a magyar mondatok esetében azt mutatja (lásd. 4.14. ábra), hogy az első 10 találatot is elég lett volna lekérdezni, mert a jó találat 44% eséllyel az első, 13%-kal a 2. és a 10. között van, és csak 6% annak az esélye, hogy a 11. és 50. között található, valamint 37%, hogy egyáltalán nincs az első 50 között. A német korpusz esetében (lásd. 4.15. ábra) – valószínűleg a kisebb szótár miatt – valamivel eltérő eredményeket kapunk: 34%; 11%; 6% és 49% ugyanezekre az intervallumokra.



4.14. ábra:Az indexált keresés által visszaadott jó találatok helyezése (angol-magyar)

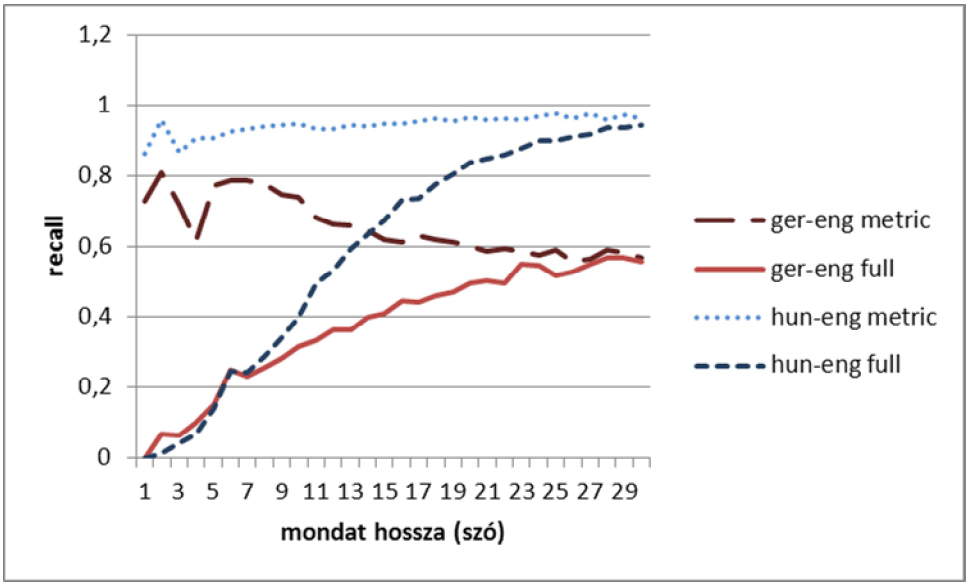
Ez a 49% nem talált mondat jól látszik a végeredményen is (4.16. ábra). A fedés értéke növekszik a keresett mondat hosszának növekedésével: a folyamatos piros vonal a hasonlósági metrika által visszaadott fedés értéket mutatja, a mondat szavakban mért hosszának a függvényében; a szaggatott barna a teljes rendszer (információ-visszakeresés+hasonlósági metrika) fedés értékét ugyanazokra a mondatokra (német-angol nyelvpárra).



4.15. ábra: Az indexált keresés által visszaadott jó találatok helyezése (német-magyar)

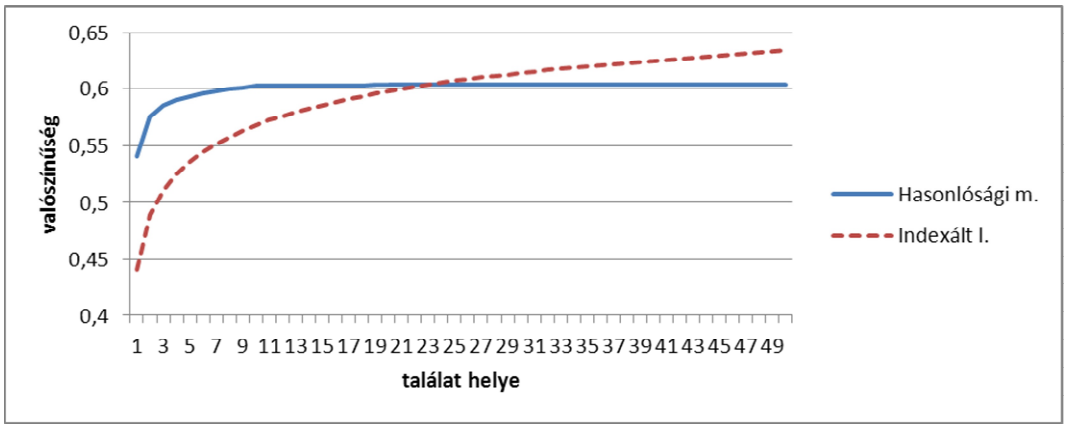
A 4.16-os ábrán látható grafikonon jól látszik, hogy az információ-visszakeresés során jelentős mennyiségű találat elvész, főleg a rövidebb mondatok esetében. Míg a hasonlósági metrika nagyjából érzéketlen a mondat hosszára, addig a Solr gyakorlatilag használhatatlan a néhány szavas mondatok esetében. Láthatólag a két-két összetartozó görbe egy-egy közös pontba tart, aminek az oka, hogy a Solr a hosszú mondatokra nagyon nagy eséllyel visszaadja a jó megoldást így a hosszú mondatoknál a teljes

rendszer fedés értéke megközelíti a hasonlósági metrika fedés értékét. Természetesen nem tud fölé menni, hiszen a teljes rendszer tartalmazza magát a hasonlósági metrikát is.



4.16. ábra: A hasonlósági metrika és a teljes rendszer fedés értéke a mondat hosszának függvényében (angol-német és angol-magyar nyelvpárokra)

A hasonlósági metrikával a visszakapott 50 találat közül hatékonyan ki lehetett szűrni a tényleges fordításokat, mint azt a 4.17.ábrán is látni lehet. A grafikon azt mutatja, hogy mekkora az esélye annak, hogy egy jó találatot a hasonlósági metrika az első *n* helyre besorol. Az első helyre (*n=1*) 10%-kal nagyobb eséllyel sorolja be, mint a Solr, és a meredeken emelkedő görbe azt bizonyítja, hogy jól meg tudja különböztetni a helyes találatot a hamis pozitívoktól. Ugyanakkor az is látszik a görbék végénél, hogy az esetek 3%-ában nem ismeri fel a jó találatot és elveti.



4.17. ábra: Annak a valószínűsége, hogy egy mondat elér minimum egy adott helyezést a két algoritmussal (angol-magyar)

A 4.17. ábrán található görbe azt mutatja, hogy mekkora valószínűséggel hányadik helyre sorolja a jó találatokat a metrika, ugyanakkor a végleges rendszer nem a sorrendet használja, hanem két adott értéket (lásd **Hiba! A hivatkozási forrás nem található.** fejezet), ami felett találatnak minősíti az adott mondatot. Ha ezek alapján nézzük meg, akkor is nagyon hasonló értéket kapunk (4.18. ábra).

	1. hely	1-10. hely	≥ Sim_1	≥ Sim_2
WP hun-eng	0,54	0,60	0,62	0,52
WP ger-eng	0,31	0,34	0,40	0,23

4.18. táblázat: Helyezés és fedés értékek a magyar és német Wikipédia fordításokra

4.3.3. A szótár hatása a fedésre

A 4.16 ábrán látható, hogy míg a hasonlósági metrika fedés értéke a magyar-angol nyelvpár esetében nagyjából konstans, addig a német-angol nyelvpárnál csökkenést mutat.

A hasonlósági metrika a két mondat összehasonlításánál az alábbi képleten alapul:

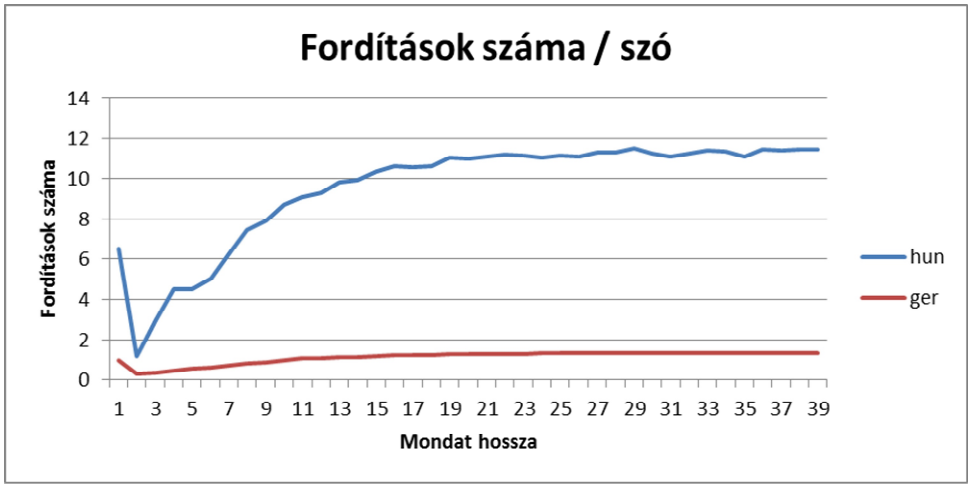
- α × közös szavak - β × hiányzó szavak

Konkrét, a tesztek során használt, értékekkel számolva:

- 2 × közös szavak - 1 × hiányzó szavak

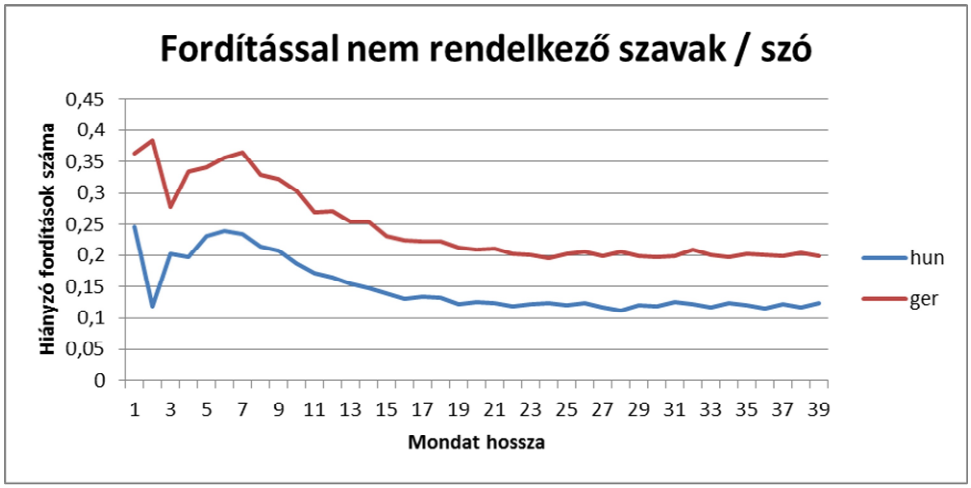
Elvégeztünk egy újabb kísérletet, hogy meglássuk mi a különbség a magyar-angol és a német-angol szótárak között (4.16b ábra). Ha megnézzük, a két szótár igen nagy eltérést mutat az egy szóra eső fordítások számában, ami logikus is, hiszen a magyar-angol

szótár körülbelül egy nagyságrenddel több szókapcsolatot tartalmaz, mint a német-angol.



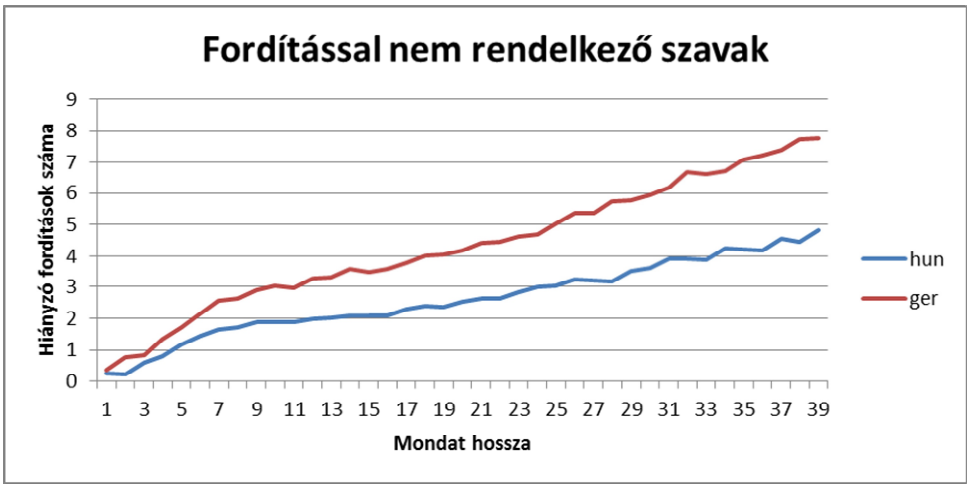
4.16b ábra: Egy szóra jutó átlagos fordítások száma a mondat hosszának függvényében

Ezek után azt nézzük meg, hogy a hosszabb mondatokban a szavakra több ritka, fordítással nem rendelkező szó esik-e (4.16c ábra), de ez konstans, a mondat hosszától független, sőt a rövidebb mondatoknál inkább magasabb.



4.16c ábra: Egy szóra jutó, fordítással nem rendelkező szavak a magyar-angol és a német-angol szótárban a mondat hosszának függvényében

Amennyiben ezt abszolút értékben, nem szavakra ábrázoljuk, akkor viszont jól látszik, hogy a német-angol nyelvpárnál egy meredekebb görbét kapunk, mint a magyar-angol esetében (4.16d ábra).



4.16d ábra: A mondatban előforduló, fordítással nem rendelkező szavak száma a mondat hosszának függvényében

Ehhez hozzávehetjük, hogy az egy szóra jutó átlagos fordítások száma németben 2 alatt van, így a hosszabb mondatokban sokkal több lesz a másik nyelvben fordítással nem rendelkező szavak száma. Ezért előfordulhatott, hogy a szavak számának a növekedésével a mondatban a hiányzó szavak több mint kétszer annyival nőttek, mint a megtalált szavak, így a hosszabb mondatok nagyobb eséllyel kaptak rosszabb értéket. Ezek alapján el lehet gondolkozni azon, hogy α és β értékét a szótárhoz igazítsuk.

4.3.4. A szótár méretének hatása a plágiumkeresésre

A szótár mérete kétféleképpen is befolyásolja a plágiumkeresést, először is a hiányzó fordítások – ha éppen azt használta ott a fordító – csökkentik a találati esélyt, azaz rontják a pontosságot. Ugyanakkor egy kisebb méretű szótár növeli a sebességet, hiszen minél kevesebb fordítása van egy szónak, annál kisebb lesz a keresőkérdés, és annál gyorsabban fut le a mondat-összehasonlító algoritmus is. Ebben a fejezetben ezt a két szempontot vizsgáljuk meg részletesebben, hogy lássuk, hogyan érdemes kialakítani a szótárat, és van-e olyan pont, amikortól már az újabb szavak, illetve jelentések nem növelik a pontosságot jelentősen, csak a sebességre vannak negatív hatással.

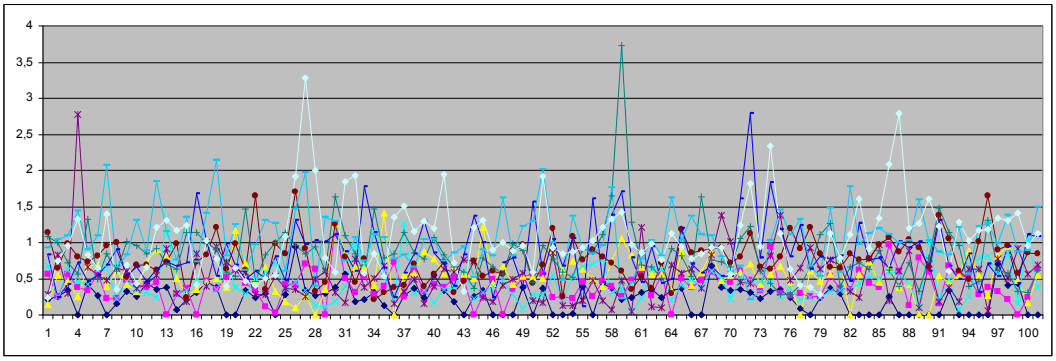
Először nézzük az indexált keresésnél elérhető sebességet. Ez a rész szorosan összefügg a 4.3.2 fejezettel, összefoglalásul csak annyi álljon itt, hogy ugyan az új algoritmus lineáris sebességű – azaz elvileg az összes az adatbázisban lévő mondattal össze kell hasonlítani a kereséshez használt mondatot – ugyanakkor könnyű belátni, hogy le lehet szűrni azokat a mondatokat egy indexált keresés segítségével, amelyek sokkal nagyobb

valószínűséggel lehetnek jó fordítások. Ugyanakkor ez az indexált keresés sem konstans idő alatt fut le.

A tesztekhez egy Dell PowerEdge 2950 gépet használtunk az alábbi paraméterekkel:

- 2* Intel Xeon E5345 @ 2.33GHz (4 mag) processzor
- 8GB memória
- 2TB háttértár

Sajnos ez a gép nem állt teljes egészében a rendelkezésünkre, más szolgáltatások is futottak rajta a tesztek során. Valószínűleg ezt tükrözi egy-két kiugró keresési érték, illetve a viszonylag nagy szórás a keresési időben, ugyanakkor ez fontos adat, hiszen azt jelenti, hogy erre egy leterhelt szerver esetén számítani kell. Ez viszont semmiben sem befolyásolta az eredmények értékelhetőségét: mint látni fogjuk, a trendek jól látszanak így is.



4.19. ábra: Lekérdezés sebessége (mp) 1-10 szóig 100 próbálkozás

A 4.19. ábrán jól látszik, hogy a sebesség igen ingadozó, és ez nem írható mind a szerver leterheltségének a rovására. Az átlagsebességek és a szórás az 4.20. táblázat szerint alakulnak 1-20 szavas lekérdezésekre.

Szavak száma	Átlag	Szórás
1	0,2231291	0,190574922
2	0,407189667	0,213336076
3	0,552207151	0,268039935
4	0,452547352	0,251823843
5	0,536414057	0,363305438
6	0,780704475	0,298026441
7	0,843644614	0,428144216
8	0,836708149	0,40908436
9	1,048863718	0,390252266
10	1,068794513	0,518749782
11	1,320166534	0,561434351

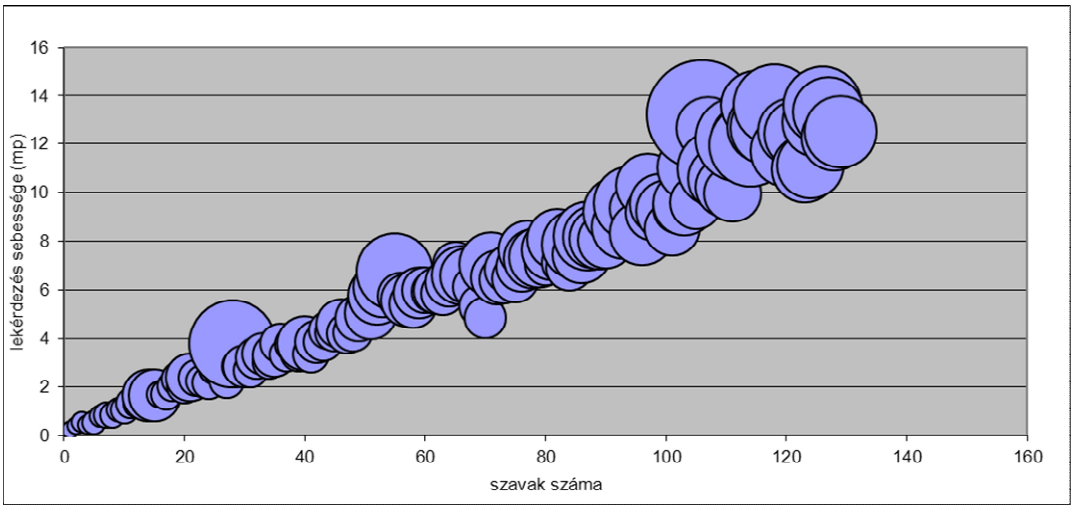
Szavak száma	Átlag	Szórás
12	1,308967463	0,468942174
13	1,564688546	0,61686659
14	1,647964079	1,830034407
15	1,627764607	1,831342686
16	1,674288263	0,469458253
17	1,691041177	0,644338773
18	2,032163346	0,60317575
19	2,322987077	0,788666072
20	2,024361214	0,788597686

4.20. táblázat: Lekérdezés átlagsebessége (mp) és a szórás a szavak számának függvényében

A lekérdező-szavakat ehhez a teszthez a Wikipédia egyik darabjából vettük folyamatosan, szótövezve, az összes szótövet felhasználva. 100 lekérdezés nem elég ahhoz, hogy ez elég véletlenszerű legyen, így a gyakori szavaknál sokkal több adattal kellett a szervernek megbirkóznia, mint a ritkáknál; ezért is fordulhatnak elő kiugróan magas és alacsony értékek a keresés során. A 4.21. ábrán a keresési sebesség és a lekérdezésre használt szavak számának az összefüggése látszik. A pontok mérete a szórásnak felel meg. Jól látszik, hogy a függvény lineáris, ennek az oka az, hogy az indexált kereséshez használt szoftver (Solr) minden bemenő szóval végez egy keresést, és ezek összesített eredményét adja vissza.

Ezen az adott hardveren az összefüggés körülbelül így alakul, egy *n* szóból álló lekérdezésre: $t_n \approx n \cdot 0{,}1$ mp

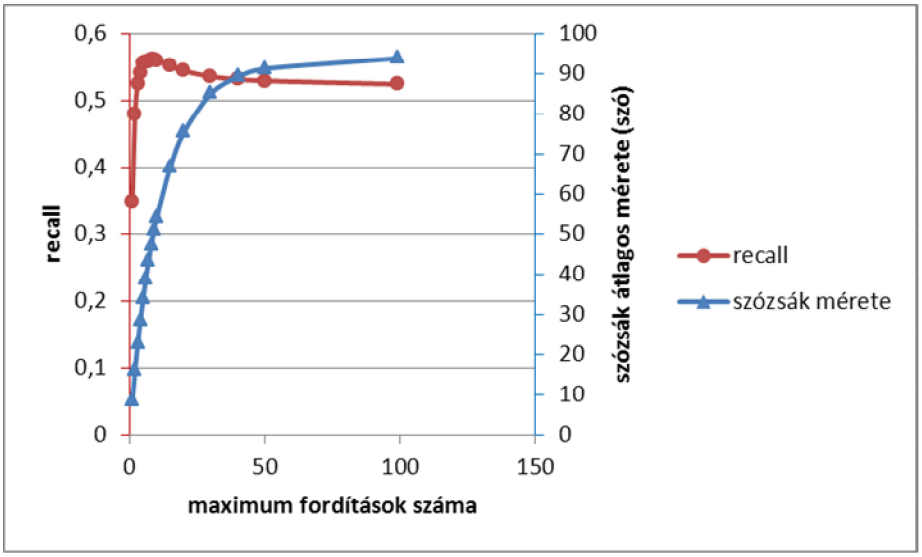
Igazából számunkra most nem is érdekes, hogy ez pontosan mennyi, hiszen feladatunk nem egy produkciós szerver felállítása, inkább csak azt érdemes kiemelni, hogy a keresési idő lineárisan függ a keresőkérdés szavainak számától, máshogyan fogalmazva, a keresési idő lineárisan függ a szótárunk méretétől. Ez egy nagyon fontos megállapítás, és ebből következik, hogy valószínűleg nem érdemes nagyon ritka szavat, illetve ritka jelentéseket a szótárunkban tartani, mert az komolyan rontja a keresési időt, de a pontossághoz nem járul hozzá nagyban. Egy szöveg leggyakoribb szavai közül 4000 adja ki annak 97,5%-át (Crystal 2003), azaz valószínűleg igen pontos eredményt lehet elérni már egy kisebb szótárral is. A következőkben ezt elemezzük részletesen.



4.21. ábra: A keresés sebességének (mp) alakulása a keresőkérdés hosszának (szó) függvényében

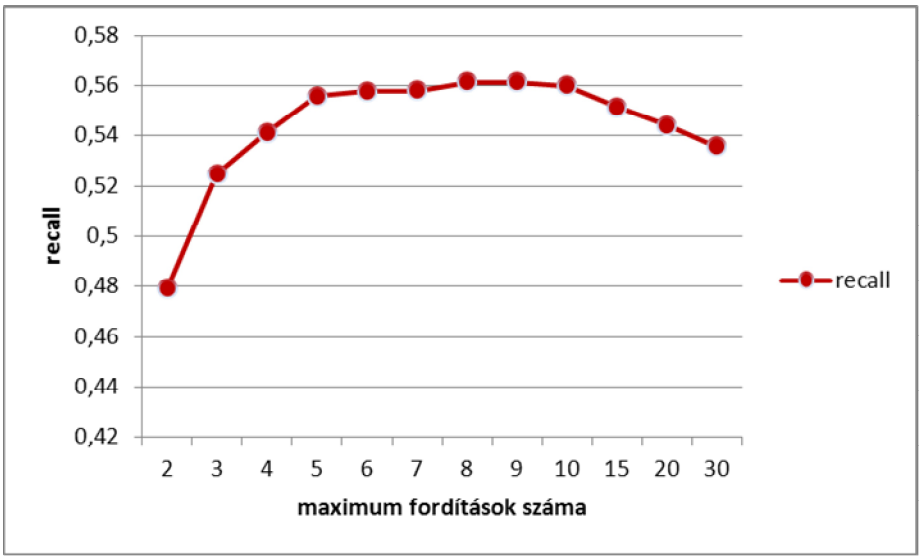
Most nézzük meg, hogy miként tudjuk csökkenteni a bemenő lekérdezés szavainak számát, vagyis a szótár méretét. Mivel nem állt rendelkezésre egy igazi nagyszótár, ahol a szavak jelentései fordítási gyakoriság szerint vannak sorba téve, így egy közelítő megoldást alkalmaztunk: a szavak előfordulási gyakoriságát vettük alapul, azt feltételezve, hogy egy gyakori szó jobb fordítás, mint egy ritkább. Ez természetesen nem feltétlen igaz, de mint látni fogjuk erre a felhasználási célra ez is megfelelő eredményt adott.

A kísérlet az alábbiak szerint került kialakításra: az előző fejezetben is használt 65 000 mondatos párhuzamos korpusz felhasználásával, 16 különböző szótármérettel lefuttattuk ugyanazt a 65 000 keresést a teljes angol Wikipédiát tartalmazó adatbázison, és mértük a fedés értékét és a szózsák méretét. A szótár méretét úgy befolyásoltuk, hogy egy szónak maximum hány fordítása található meg benne. Az 1-es paraméter azt jelenti, hogy csak a leggyakoribb fordítása (illetve a korábban ismertetett okból a fordításai közül a leggyakoribb szó), az 5-ös paraméter pedig azt jelenti, hogy a leggyakoribb 5, amennyiben van 5 fordítása. Érthető módon egy idő után a szótár mérete nem növekszik, hiszen a szavaknak véges számú fordításuk van.



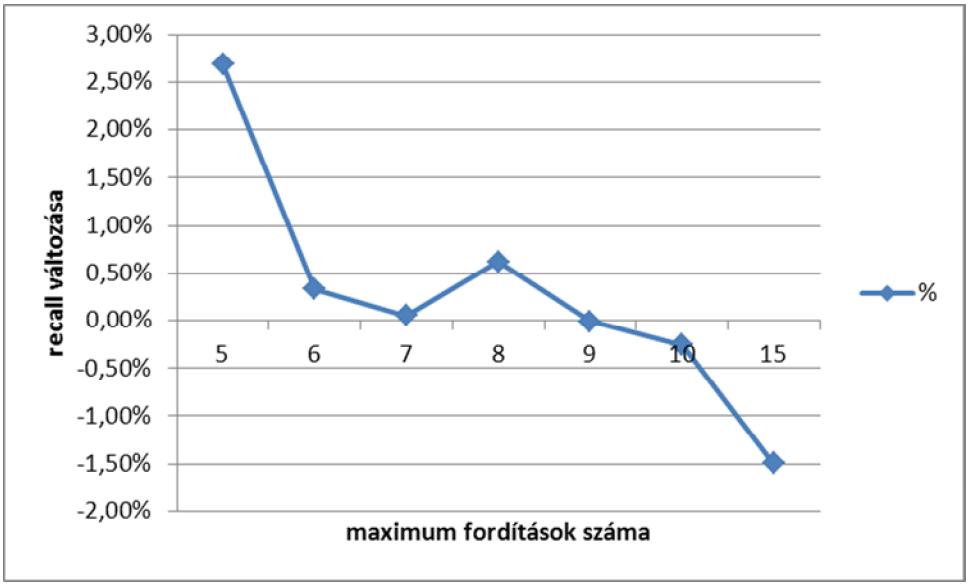
4.22. ábra: A fedés értéke és a szózsák mérete a szótár méretének a függvényében

Mint az a 4.22. ábrán látszik, a fedés értékének van egy maximuma, amikortól már hiába növeljük a szótárunk méretét, az algoritmus teljesítménye nemhogy nem javul, de még romlik is. Ez érdekes eredmény, és magyarázata valószínűleg a hasonlósági metrikában rejlik, amely a nagyon ritka fordítások miatt rosszul párosítja össze a szavakat, illetve olyan szavakat is azonosnak talál, amelyek nem azok, csak azért, mert az egyik szó ritka jelentése egyezik egy másik szó gyakoribb jelentésével.



4.23. ábra: A fedés értéke a szótár méretének a függvényében

Kinagyítva a 4.23. illetve 4.24. ábrán azt láthatjuk, hogy a legmagasabb értéke a fedésnek 8-nál van ugyan, de 5 és 10 között gyakorlatilag alig változik, míg a szózsák mérete egészen 30-40-ig meredeken emelkedik.



4.24. ábra: A fedés értékének a változása a szótár méretének a függvényében

Ezek alapján egy produkciós rendszerben az 5 fordítást tartalmazó szótár használata javasolt, amely 1%-kal alacsonyabb fedés mellett 39%-kal alacsonyabb futási időt eredményez, mint a legjobb, maximum 8 fordítást tartalmazó megoldás.

	1	2	3	4	5	6	7	8	9	10	15	20
szózsák mérete	8,6	16,0	22,7	28,6	33,9	38,9	43,3	47,3	50,8	54,1	66,6	75,3
fedés	0,349	0,479	0,525	0,541	0,556	0,558	0,558	0,561	0,561	0,560	0,551	0,544
%		37,17%	9,52%	3,14%	2,68%	0,34%	0,05%	0,62%	-0,01%	-0,26%	-1,50%	-1,32%

4.25. táblázat: A fedés értéke a szótár méretének a függvényében

4.3.5. Az algoritmus eredményének értékelése

Az előző fejezetben ismertetett fedési értékek egy-egy mondatra vonatkoztak. Mivel a plágiumkeresés során nem az a célunk, hogy egy lefordított mondatot megtaláljunk, hanem, hogy nagy mennyiségű fordított szöveget hatékonyan tudjunk detektálni, így az eredményeket ennek tükrében kell megnézni. Ehhez azt feltételezzük, hogy egy mondat megtalálásának a valószínűsége független az előző mondattól. A 4.26. táblázat megmutatja, hogy mekkora eséllyel találunk meg minimum x darabot egy y mondat hosszú szövegből, a rosszabb találati arányt adó angol-német korpuszra kapott, Sim_1-hez tartozó *0,4* fedési értéket alapul véve.

y	*x=1*	*x=2*	*x=3*	*x=4*	*x=5*
1	0,4				
2	0,64	0,16			
3	0,784	0,352	0,064		
4	0,8704	0,5248	0,1792	0,0256	
5	0,92224	0,66304	0,31744	0,08704	0,01024
6	0,953344	0,76672	0,45568	0,1792	0,04096
7	0,972006	0,84137	0,580096	0,289792	0,096256
8	0,983204	0,893624	0,684605	0,405914	0,17367
9	0,989922	0,929456	0,768213	0,51739	0,266568
10	0,993953	0,953643	0,83271	0,617719	0,366897
11	0,996372	0,969767	0,881083	0,703716	0,467226
12	0,997823	0,980409	0,916557	0,774663	0,561822
13	0,998694	0,987375	0,942098	0,83142	0,646958
14	0,999216	0,991902	0,960208	0,875691	0,720743

4.26. táblázat: Annak a valószínűsége, hogy y mondatot megtalálunk egy x mondat hosszú szövegben

Mint ahogy azt látni lehet, az eredmények igen bíztatóak. Egy 10 mondat hosszú fordítást (kb. 1 oldal) már 99%-os eséllyel észreveszünk, sőt több mint 80% annak az esélye, hogy a harmadát megtaláljuk.

4.4. A hasonlósági metrikán és az automatikus fordítón alapuló algoritmusok összehasonlítása

Mindenképpen ki kell emelni, hogy az ebben a fejezetben tárgyalt témának (a jelenlegi algoritmusok összehasonlítása és értékelése) a teljes kifejtése egy külön dolgozat témája lehetne, hiszen annyi paramétere van minden algoritmusnak, hogy itt csak a legalapvetőbb beállításokkal és paraméterekkel tudunk kísérletezni. A szakirodalomban található leírások és saját legjobb tudásom szerint készítettem el az n-gramon és gépi fordításon alapuló algoritmust, hogy össze tudjuk hasonlítani az előző fejezetben ismertetett új algorimtussal. Ennek a disszertációnak nem témája a legjobb algoritmus megtalálása, csupán az általam létrehozott algoritmus bemutatása és elhelyezése a plágiumkereső algoritmusok világában. Ez a fejezet ez utóbbit hivatott szolgálni.

Mint azt a bevezetőben említettük, a szakirodalomban található (Potthast 2010), széles körben használt, fordításiplágium-kereső algoritmusok általában automatikus fordítót és n-gram darabolást használnak a különböző nyelvű szövegek összehasonlítására. Most röviden nézzük meg, hogy egy ilyen algoritmus miként működhet, és hogyan tudjuk összehasonlítani a teljesítményét a korábban felvázolt új algoritmussal.

Először is meg kell határozni, hogy miként működjön a n-gram algoritmus, milyen transzformációkat alkalmazzon a szövegek összehasonlítása előtt, hogy minél jobb eredményt érjen el. Az alábbi, elég elterjedt algoritmus tűnik a legcélravezetőbbnek az adatbázis építésére:

1. a szövegből töröljük a stopszavakat
2. szótövezést alkalmazunk minden szóra
3. veszünk minden szó szinonimái közül egy meghatározottat (legyen pl. a szinonimái közül az ábécében legelső, melyben maga is szerepel)
4. feldaraboljuk a szöveget átlapolódó n-gramokra, n-szavakra, ahol n értéke 3-5 között lehet
5. az n-gramokon belül ábécé sorrendbe tesszük a szavakat
6. ezekből az n-gramokból képezünk ujjlenyomatokat (MD5)
7. az ujjlenyomatokat feltöltjük az adatbázisba

A lekérdezésnél nagyon hasonló lépéseket teszünk, csak beiktatunk egy 0. lépést:

0. **lefordítjuk a szöveget egy automatikus fordítóval az adatbázis nyelvére**
1. a szövegből töröljük a stopszavakat
2. szótövezést alkalmazunk minden szóra
3. veszünk minden szó szinonimái közül egy meghatározottat (legyen pl. a szinonimái közül az ábécében legelső, melyben maga is szerepel)
4. feldaraboljuk a szöveget átlapolódó n-gramokra, n-szavakra, ahol n értéke 3-5 között lehet (*a szövegek eltolódása miatt kell átlapolódó darabolás*)
5. az n-gramokon belül ábécé sorrendbe tesszük a szavakat
6. ezekből az n-gramokból képezünk ujjlenyomatokat (MD5)

7. az ujjlenyomatokat **lekérdezzük** az adatbázisból
8. **eldobjuk azokat a találatokat, amelyek több mint 50 szócikkben találhatóak meg**

Ennek az algoritmusnak három gyenge pontja van: az automatikus fordító (amely nagyon eltérő minőséget produkál különböző nyelvekre), a szórend (amit az ábécébe tétellel próbálunk meg ellensúlyozni), valamint a szinonimaszótár használata. Utóbbihoz jó lenne pontosan tudni, hogy mely szó melyik értelmében szerepel ott, de ez erőforrásigényes, drága művelet, nem is elég megbízható, így nem került implementálásra. Az utolsó, 8. pont a gyakorlati tapasztalat miatt, az első próbálkozás után került bele, a szakirodalomban nem találtam erre vonatkozó utalást, de a 3-gramok körülbelül 10%-a több mint 50 szócikkben is megvan, azaz olyan általános, hogy semmi plusz információval nem rendelkezik és rengeteg hamis találatot eredményez.

Azért, hogy az automatikus fordítók minősége közötti különbséget és ennek hatását megnézhessük a kísérletet angol-magyar és angol-német nyelvpárokra fogjuk elvégezni. Előbbire sokkal gyengébb fordítási eredményt kapunk, utóbbi pedig – bizonyos esetekben –annyira jó tud lenni, hogy egy nem anyanyelvi beszélő nem is tudja megkülönbözetni egy kézi fordítástól.

Mivel ez az algoritmus használ gépi fordítást, így itt a tesztadatbázist csak kézi fordítással állíthatjuk elő, hogy az független legyen a plágiumkereső algoritmustól.

4.4.1. Az n-gram paraméterek kiválasztása

Az algoritmusok teszteléséhez az angol Wikipédiából kézzel fordított szövegre van szükségünk. Ennek az alapját a 12 véletlenszerűen kiválasztott, körülbelül fél-fél oldalas szövegek adták, melyek megtalálhatóak a magyar fordítással együtt a 9.4 mellékletben. Két fontos paramétert kell néznünk az összehasonlítás során: a fedést és a pontosságot. A fedést sokkal könnyebb vizsgálni, hiszen csak attól függ, hogy ennek a 12 Wikipédia-szócikknek a neve közte lesz-e az eredmények között. A pontosság sokkal nehezebb, két okból is: először nem minden olyan szócikk számít hibás találatnak, amit visszaad a rendszer de nem volt közte a 12 eredeti cikk között. A Wikipédiában vannak ismétlődések, és hasonlóságok a szócikkek között, és ezért nem lehet automatikusan hamis pozitívnak tekinteni minden „rossz” találatot. A másik nehézség, hogy hamis pozitívokat nagy mennyiségű keresés során van értelme csak

vizsgálni, hiszen annak nem nagy a jelentősége, hogy szűk 6 oldalnyi szöveg esetében nem túl sok a hamis találat. Mindezekért olyan dokumentumokkal is tesztelni kell a két algoritmust, amelyek függetlenek a Wikipédiától, és elég hosszúak ahhoz, hogy nagy valószínűséggel kijelenthessük, hogy az eredmények reprezentatívak. Ehhez a feladathoz megfelelőek szakmai cikkek, amelyekben nem szoktak Wikipédia-idézetek lenni, és így biztosak lehetünk benne, hogy a találatok (egy esetleges közös forrásból származó rövid idézeten kívül), mind hamis pozitív találatok lesznek.

Az első körben az adatbázis felépítéséhez az n-gram paraméterét, a lehető legkisebbre, *n=3*-ra vettük. Könnyen belátható, hogy ez adja a legtöbb találatot, mind valósat, mind hamis pozitívat. Mielőtt kipróbálnánk a 4-es paramétert, érdemes megnézni, hogy milyen eredményre jutunk a 3-assal, hiszen jól meg lehet állapítani, hogy ott milyen eredményt kapnánk. Négy egymás után szereplő szóból, amely megtalálható azonos módon a másik dokumentumban is, két 3-gram egyezés lesz, illetve egy 4-gram (átlapolódó szavas darabolást használunk). Két közvetlen egymás után szereplő 3-gram találat két dolgot jelenthet:

1. 4 egymás utáni szó azonosan szerepel a másik dokumentumban
2. ez a két 3-gram nem egymás után, hanem elszórtan szerepel, véletlen osztoznak 2 szón

Ez azt jelenti, hogy az 1 távolságra lévő 3-gramok egy jó felső becslést adnak a 4-gramos darabolás eredményére. Azaz ahány ilyen közeli 3-gram található egy dokumentumban, maximum annyi lehet ugyanebben 4-gram darabolást használva. Ezért most nézzük meg előbb a később ismertetett 3-gramos darabolás eredménye alapján a 4-gramosra kapott felső korlátot. Az aláhúzott címek a helyes találatok, a számok a zárójelben a 4-gramok maximális lehetséges számát jelölik.

1. University of Oxford (1)
2. Foreign relations of Pakistan (4)
3. Politics of Pakistan (2)
4. Pete Seeger (9)
5. Charles Seeger (4)
6. Mike Seeger (7)
7. Seeger (1)
8. Meitetsu 5000 series (2008) (1)
9. Upper 10 (9)
10. Fruktsoda (1)
11. Dr Pepper Snapple Group (1)
12. Dr Pepper/Seven Up (1)
13. Munich Philharmonic (6)

14. Castle Rock (Pineville, West Virginia) (3)
15. Tumulus (1)
16. Dragon-and-Tiger Pagoda (1)
17. Golden Twenties (3)
18. GpsGate (1)
19. Wialon (10)
20. GPS tracking server (8)
21. Web accessibility (5)

A 12 szócikkből 2-t egyáltalán nem tudna így megtalálni, 2 pedig csak 1-1 n-gramban egyezik, amit a hamis pozitív találatok kiszűrése, csökkentése miatt biztos, hogy el kéne dobnunk, véletlen egyezésnek minősítve. Ez azt jelenti, hogy maximum a szócikkek kétharmadát találnánk meg 4-gramokat használva. Természetesen ez nagyon kicsi adatbázis ahhoz, hogy messzemenő következtetést tudjunk leszűrni ebből, de be lehet látni, hogy annak az esélye, hogy egy oda-vissza fordítás után 4 szó is megtartja a helyét, és valahogy egymás mellé kerülnek, elég kicsi.

Most nézzük meg a 3-gramra kapott eredményeket: ha semmit se módosítunk az algoritmuson, akkor 5 477 Wikipédia-szócikkel talál egyezést. Ha kizárjuk az 1 db n-gramot tartalmazó találatokat, akkor már csak 386-tal, de még ez is elfogadhatatlanul sok. A 3 db közös töredékkel rendelkező találatokat is kizárva 78 szócikk marad, ami nem tűnik túl soknak, de ha pl. az egyik, mindössze 15 oldalas angol nyelvű cikkemet (Pataki 2012) megadjuk ennek az algoritmusnak, akkor így is 240 Wikipédia szócikkel talál egyezést. Mivel ebben a cikkben nincs Wikipédia-idézet, ez azt jelenti, hogy ennyi hamis pozitív találatunk van. Az a legnagyobb probléma, hogy a 15 oldalas cikk 2712 töredékéből 1360-hoz talált hasonlót. Ez gyakorlatilag egy átlapolódó szavas darabolás esetén 100%-os egyezés a teljes szövegre (de ha ez mind az első felében lenne, akkor is 50%-os). Az a gond, hogy 3 jelentéssel bíró szó egymás után még nem elég egyedi ahhoz, hogy bármi egyebet meg lehetne belőle állapítani, mint hogy ugyanarról szól a két szöveg. Három szó sokkal inkább jellemzi a témát, a nyelvet, a szóösszetételeket, mint az író stílusát. Mivel láttuk, hogy a 4-es paraméter már valószínűleg alkalmatlan a találatok kimutatására, így ezeket az eredményeket kell megfelelően megszűrni, hogy kezelhető eredményt kapjunk. Ahogy a hasonlósági metrikán alapuló algoritmusnál, itt is szűrni kell a kapott találatokat, és csak a közelieket értékelni. Mivel nincs a hasonlóságnak mértéke, így csak az egyező darabok egymáshoz viszonyított helyzetére tudunk szűrni: $|i - j| < d$ Ezen kívül meg tudjuk határozni, hogy minimum hány ilyen közeli darab legyen: $\# d_{ij} \geq d_{num}$ Még egy további paramétert változtattunk: azt, hogy

hány töredéknek kell *d* távolságon belül lennie. E mögött az a meggondolás áll, hogy esetenként 2 töredék véletlen is egymás mellé kerülhet, de 4 vagy 5 talán már biztosabban jelzi a találatot.

Az algoritmust lefuttattuk a 12 Wikipédia-cikk magyar fordításának az angol visszafordítására (Google fordítóval), az említett 15 oldalas cikkre és, hogy legyen egy hosszabb dokumentum is, a Harry Potter első kötetére (annak is az eredeti angol verziójára) így zárva ki a gépi fordítás torzításait, és hozva ki a maximumot az algoritmusból. Induljunk ki abból a (majdnem helyes) feltevésből, hogy a 15 oldalas dokumentumban, illetve a Harry Potter első kötetében lévő találatok mind hamis pozitívak (*F*+), és a 12 Wikipédia-cikkből megtaláltak a valós pozitív értékek (*T*+): ebből már könnyen ki tudjuk számolni a pontosság és a fedés értékét. Mivel számunkra a tényleges egyezések megtalálása fontosabb, így érdemes F_2 értékét kiszámolnunk, azaz a fedés értékének kétszer akkora súlyt adunk, mint a pontosságnak.

A 15 oldalas cikk esetében a legjobb paraméterek a 4.27. táblázatban találhatóak, *d*+ jelöli azt, hogy d távolságon belül hány töredékkel nézünk előre, azaz azt számoljuk, hogy d_i után a sorban hányadik elem d_j. A $d+ = 5$ azt jelenti, hogy minimum 6 töredéknek kell lennie *d* távolságon belül. A d_{num}/D pedig azt jelenti, hogy ilyen töredékhalmazból hánynak kell lennie egy dokumentumon belül, hogy az adott forrást kijelezzük.

F_2	d+	d	d_{num}/D	T+	F+
0.78947368421053	7	120	1	9	0
0.78947368421053	7	140	1	9	0
0.78947368421053	7	160	1	9	0
0.78947368421053	7	180	1	9	0
0.78947368421053	7	200	1	9	0
0.78947368421053	7	220	1	9	0
0.78947368421053	7	240	1	9	0
0.78947368421053	7	260	1	9	0
0.78947368421053	7	280	1	9	0
0.78947368421053	7	300	1	9	0
0.77586206896552	6	80	1	9	1
0.77586206896552	5	80	1	9	1
0.77586206896552	5	100	1	9	1
0.77586206896552	6	100	1	9	1
0.77586206896552	5	120	1	9	1
0.77586206896552	6	120	1	9	1
0.77586206896552	5	140	1	9	1
0.77586206896552	6	140	1	9	1
0.77586206896552	6	160	1	9	1
0.77586206896552	5	160	1	9	1
0.77586206896552	6	180	1	9	1

0.77586206896552	5	180	1	9	1
0.77586206896552	5	200	1	9	1
0.77586206896552	6	200	1	9	1
0.77586206896552	6	220	1	9	1
0.77586206896552	5	220	1	9	1
0.77586206896552	5	240	1	9	1
0.77586206896552	6	240	1	9	1
0.77586206896552	5	260	1	9	1
0.77586206896552	6	260	1	9	1
0.77586206896552	6	280	1	9	1
0.77586206896552	5	280	1	9	1
0.77586206896552	5	300	1	9	1
0.77586206896552	6	300	1	9	1

4.27. táblázat: Lehetséges paraméterek, az F_2 maximalizációjára törekedve (15 oldalas cikk)

Az érdekesség kedvéért kiszámoltuk az F_4 értékét is a különböző paraméterekre. Ezt a 9.8 mellékletben lehet megtekinteni, a két eredmény gyakorlatilag megegyezik, ugyanazokat a paramétereket preferálja.

F_2	**d+**	**d**	**d_{num}/D**	**T+**	**F+**
0.65217391304348	7	120	1	9	12
0.65217391304348	7	140	1	9	12
0.65217391304348	7	160	1	9	12
0.64285714285714	6	80	1	9	13
0.64285714285714	7	180	1	9	13
0.64285714285714	7	200	1	9	13
0.64285714285714	7	220	1	9	13
0.64285714285714	7	240	1	9	13
0.63380281690141	7	260	1	9	14
0.63380281690141	7	280	1	9	14
0.63380281690141	7	300	1	9	14
0.625	10	100	1	8	8
0.61643835616438	6	100	1	9	16
0.61643835616438	6	120	1	9	16
0.61643835616438	6	140	1	9	16
0.61643835616438	6	160	1	9	16
0.61643835616438	6	180	1	9	16
0.61538461538462	8	80	1	8	9
0.61538461538462	9	100	1	8	9
0.61538461538462	10	120	1	8	9
0.61538461538462	10	140	1	8	9
0.61538461538462	10	160	1	8	9
0.61538461538462	10	180	1	8	9

4.28. táblázat: Lehetséges paraméterek, az F_2 maximalizációjára törekedve (Harry Potter)

A legjobb paraméternek mindkét esetben a $d+ = 7$, $d = 120$ illetve $d_{num}/D = 1$ bizonyult (4.27. és 4.28. táblázat). Ez annyit jelent, hogy azt az adatbázisban lévő szócikket tekintjük találatnak, amelyben: minimum 1-szer 8 töredék 120 szó távolságon belül van. Látható módon a 140 illetve a 160 is megfelelő paraméterek lettek volna, de minél

kisebb *d* értéke annál kevesebb a hamis pozitív találatok száma, ezért döntöttünk úgy, hogy a továbbiakban a 120-ast választjuk.

Ezzel a paraméterrel az alábbi találatokat kaptuk a 12 Wikipédia-cikkre:

1. Mike Seeger (19)
2. Pete Seeger (18)
3. Upper 10 (18)
4. Wialon (15)
5. Web accessibility (14)
6. Castle Rock (Pineville, West Virginia) (13)
7. Munich Philharmonic (13)
8. Foreign relations of Pakistan (13)
9. Golden Twenties (12)
10. GPS tracking server (11)
11. Politics of Pakistan (10)
12. Avidius Cassius (8)

A Harry Potterre pedig az alábbi találatokat:

1. Places in Harry Potter (145)
2. Portal:Harry Potter/Quotes/Archive (103)
3. List of Harry Potter characters (101)
4. Hogwarts (83)
5. Harry Potter (character) (78)
6. Harry Potter universe (57)
7. Portal:Harry Potter/Quotes (57)
8. List of fictional books (56)
9. Treacle (13)

Kimondottan jónak tűnik, hiszen megtaláltuk azokat a cikkeket, amelyek erről a könyvről szólnak, ugyanakkor érdemes megnézni a 9.9 mellékletben a részletes találati listát. Számos kisebb, lényegtelen egyezés van, mint például: „when Harry saw this" vagy „what had happened when he". Egyébként ezek a találatok már egész jók, hiszen ezért választottuk az adott paramétert, hogy a hamis pozitívakat minél inkább csökkentsük.

Az érdekesség kedvéért megnéztük, hogy amennyiben nagyobb fedést szeretnénk elérni, akkor milyen egyéb paraméterek jöhetnek szóba. A teljes táblázat megtalálható a 9.8 mellékletben, itt most csak a legjobb értékek szerepelnek adott fedés (*T+*) értékhez.

F_6	d+	d	d_{num}/D	T+	F+
0.82723577235772	2	60	1	11	49
0.77405857740586	3	40	1	10	36
0.75510204081633	7	120	1	9	0

4.29. táblázat: Lehetséges paraméterek, az F_6 maximalizációjára törekedve (15 oldalas cikk)

F_6	d+	d	d_{num}/D	T+	F+
0.75356415478615	3	40	1	10	49
0.73509933774834	7	120	1	9	12

4.30. táblázat: Lehetséges paraméterek, az F_6 maximalizációjára törekedve (Harry Potter)

A 4.29. és 4.30. táblázatban jól lehet látni, hogy ha több valós pozitív találatot szeretnénk visszakapni, akkor sokkal több hamis pozitív találatunk lesz. 11 valós találatot kapunk vissza, ha $d+ = 2$, $d = 60$ illetve $d_{num}/D = 1$, viszont ez 49+185 hamis pozitív találatot ad a két másik műben, ami elfogadhatatlan. 10 valós találat esetén, $d+ = 3$, $d = 40$ illetve $d_{num}/D = 1$ paraméterekkel 36+49 hamis pozitív találatot kapunk. Az érdekesség kedvéért egy listába összegyűjtöttük ez utóbbiakat, melyek megtalálhatóak a 9.14 mellékletben. Ez a lista már nagyon hosszú, számos, még a témában se egyező találatot ad, ugyanakkor, olyan helyen, ahol van türelme a felhasználónak átnéznie egy nagyon hosszú listát, el lehet gondolkodni a használatán.

4.4.2. Angol-magyar irányú keresések összehasonlítása

Miután beláttuk, hogy a $n=3$ paraméter jó választás –ugyan elvileg a 4-es 5-ös paraméter is alkalmas egynyelvű keresésnél jó eredményeket elérni, a fordításnál láthatólag nem működik jól (rövidebb egyezésekre) – valamint a $d+ = 7$, $d = 120$ illetve $d_{num}/D = 1$ adják a legjobb eredményeket, most nézzük meg, hogy ténylegesen hogyan teljesít a két algoritmus.

4.4.2.1. Automatikus fordítón és n-gramon alapuló algoritmus 12 szócikkre

A 4.4.1 fejezetben már láttuk, hogy mi az eredménye a céljainknak legjobban megfelelő paraméterekkel, de most az összehasonlíthatóság kedvéért ismételjük meg. A zárójelben mögé írt számok a közös töredékek számát jelölik.

1. Mike Seeger (19)
2. Pete Seeger (18)
3. Upper 10 (18)
4. Wialon (15)
5. Web accessibility (14)
6. Castle Rock (Pineville, West Virginia) (13)
7. Munich Philharmonic (13)
8. Foreign relations of Pakistan (13)
9. Golden Twenties (12)
10. GPS tracking server (11)
11. Politics of Pakistan (10)
12. Avidius Cassius (8)

A 12-ből 9 szócikket talált meg: a legelső egy olyan, amelyik tényleg hasonlít a második találatra, de azért jó lett volna, ha nem egy ilyen félig hamis találatot minősít a legjobbnak az algoritmus.

Sajnos, mint azt a korábbiakban is láttuk, ez az algoritmus, hasonló témában íródott szövegekre nagyon érzékeny, és azokat is megtalálja. Ezt támasztja alá az alábbi megállapítás is, amelyet a Turnitin szoftvert oktatási célra használó egyetemi tanár mondott el a 5th International Plagiarism Coference keretében. (Guerin et al. 2012)

> „In order to develop novice research writers' understanding of acceptable use of sources and mastery of disciplinary language, we have developed a process that uses concordancing software alongside Turnitin. Here we present textual analyses of two cases using this process: in one, the student's percentage of matches decreased as he developed his authorial voice; in the second, **the percentage of matches actually increased as the student's language choices came to reflect more closely the expectations of the discipline.**"

Csak attól, hogy a megfelelő kifejezéseket használja valaki, hasonlítani fog a műve a többiekére. Ez nagy hátrány és számos felhasználó jelezte a konferencián, hogy igazából olyan sok az egyezés, hogy nehéz kiválogatni a ténylegesen jókat. Természetesen a szó szerinti másolat akkora egyezést okoz, hogy az egyezés kiugró lesz, de egy olyan mű, amely maga gyengén lett megírva (nem használja a megfelelő

nyelvezetet), ugyanakkor plagizál egy másik műből egy kisebb részt, könnyen elcsúszik a sok jó nyelvezettel megírt dolgozat között, ha a részletes eredményt nem nézi meg az oktató. Guerin et al. (2012) arra használta a konkordancia-szoftvereket, hogy kikeresse bennük, hogy az adott kiemelt találat, három-négy szavas kifejezés, másolat-e, vagy így szokott szerepelni az ilyen témában íródott cikkekben, ez jó megoldás, de nagyon időigényes. Ennek az algoritmusnak a használata ki kell, hogy egészüljön egy n-gram gyakorisági szótárral, amely jelezni tudja a felhasználónak, hogy az adott egyezés mekkora eséllyel a véletlen műve, és mekkora eséllyel lehet másolat.

4.4.2.2. Hasonlósági metrikán alapuló algoritmus 12 szócikkre

A hasonlósági metrikán alapuló algoritmusunk az alábbi 25 találatot adta vissza, a zárójelben mögé írt számok a hasonló mondatok számát jelölik.

1. Pete Seeger (7)
2. Golden Twenties (4)
3. Castle Rock (Pineville, West Virginia) (4)
4. Wialon (4)
5. Munich Philharmonic (3)
6. Upper 10 (3)
7. Meitetsu 5000 series (2008) (3)
8. Mike Seeger (3)
9. Avidius Cassius (2)
10. Foreign relations of Pakistan (2)
11. Politics of Pakistan (2)
12. GPS tracking server (2)
13. Web accessibility (2)
14. Portal:University of Oxford/Intro (2)
15. Castle Rock (2)
16. MOPAC (1)
17. Kosmo (1)
18. VTune (1)
19. List of office suites (1)
20. Pakistan (1)
21. LMMS (1)
22. FEMtools (1)
23. Nautilus (secure telephone) (1)
24. Monolith (1)
25. Charles Seeger (1)

Itt a 12-ből 10 szócikket megtalált, és mindezeket az első 13 találat közé sorolta. A 14. találat az egyik hiányzó szócikkről szóló másik oldal, amit nem vesszük találatnak, de az adott részhez így is talált hasonlót. Az utolsó 10 találatot, csakúgy mint az előzőekben, az 1 darab hasonlóság miatt el is dobhatnánk, de itt ezek esetében sokkal erősebb okunk van jó találatot feltételezni, hiszen az algoritmusunk egy mondat egyezése esetén magasabb hasonlósági mértéket követel meg (Sim_2), mint több találat

esetén (Sim_1). A részletes találati listát a mondatokkal együtt a 9.9 melléklet tartalmazza.

4.4.2.3. Hasonlósági metrikán alapuló algoritmus a 15 oldalas cikkre

A 15 oldalas cikkre összesen 1 találatot ad vissza az új algoritmus, mégpedig az alábbit:

1. Word-sense disambiguation (1)
   In Proceedings of the 20th National Conference on Artificial Intelligence (AAAI, Pittsburgh, PA).
   - Proceedings of the 13th International Conference on Artificial Intelligence. pp. 83–92. (8)

Ez jó találat, ami az egy mondatos találatok kiszűrésével eltűnne, ezt most az algoritmus működésének megértése miatt nem tettük. Egy éles rendszerben ezt valószínűleg meg kell tenni, hasonló módon, mint az n-gram algoritmusnál.

4.4.2.4. Hasonlósági metrikán alapuló algoritmus a Harry Potter könyvre

Ahogy már az n-gram algoritmusnál is kiderült, vannak tényleges, érvényes találatok a Harry Potter első kötetéből, ugyanakkor egy ilyen hosszú könyvben már a hasonlósági metrikán alapuló algoritmus is talál hamis pozitív találatokat. És ugyanaz okozza a gondot, mint az n-gram esetében: a rövid mondatok, a kifejezések és szóösszetételek. A részletes eredmény megtekinthető a 9.12-es mellékletben. Összegezve az alábbi felsorolás tartalmazza, aláhúzva a valós pozitív találatok, a többi pedig annak ellenére, hogy mind témában megfelelő, inkább hamis pozitívnak tekinthető.

1. Portal:Harry Potter/Quotes/Archive (13)
2. Places in Harry Potter (10)
3. Wikipedia:WikiProject Harry Potter/PS Differences (10)
4. Harry Potter and the Philosopher's Stone (8)
5. List of Harry Potter characters (8)
6. Magical objects in Harry Potter (8)
7. Portal:Harry Potter/Quotes (6)
8. Quidditch (6)
9. Wizard People, Dear Reader (6)
10. List of supporting Harry Potter characters (5)
11. Harry Potter and the Order of the Phoenix (video game) (5)

12. Potter Puppet Pals (5)
13. Ron Weasley (4)
14. Hogwarts staff (4)
15. Death Eater (4)
16. Hogwarts (4)
17. Harry Potter and the Philosopher's Stone (film) (4)
18. A Very Potter Musical (3)
19. Harry Potter and the Prisoner of Azkaban (3)
20. Harry Potter and the Chamber of Secrets (film) (3)
21. Muggle (3)
22. A Very Potter Sequel (3)
23. Wikipedia:In the news/Candidates/July 2011 (3)
24. Harry Potter and the Order of the Phoenix (film) (3)
25. Wikipedia:Administrators' noticeboard/Moulton (2)
26. List of Deadliest Warrior episodes (2)
27. Aunt Petunia (2)
28. Portal:Harry Potter/Featured Character/Archive (1)
29. Albus Dumbledore (1)
30. Artemisia (plant) (1)
31. Muggle Quidditch (1)
32. Magic in Harry Potter (1)
33. Sonic's Rendezvous Band (album) (1)
34. St. Joseph's School, Bhagalpur (1)
35. Viper (Six Flags Great America) (1)
36. Shell (theater) (1)
37. Order of the Phoenix (organisation) (1)

Az első 10 találat között 9 valós és 1 hamis van. A teljes listára számolva 54%-a jó a találatoknak (20 a 37-ből), azaz a pontosság 54%. Ez növelhető lenne más paraméterezéssel, de az algoritmus fejlesztésénél az volt a cél, hogy a jó találatok előre kerüljenek, és a felhasználó fentről lefelé nézhesse a találatokat. Ha az az általa megkövetelt szint alá megy, akkor fontos, hogy biztos lehessen benne, hogy már semmi nagyobb találat nem lesz lentebb. Ez a KOPI rendszer üzemeltetése során szerzett tapasztalat alapján a legjobb megoldás, mert minden felhasználó mást ítél meg „nagy egyezésnek”. Ezt a célt teljesíti a rendszer.

4.4.3. Angol-német irányú keresések összehasonlítása

A 12 Wikipédia-cikket tartalmazó korpuszt angolról német nyelvre lefordította egy fordító, és ezzel is kipróbáltuk az n-gramon alapuló és az új algoritmust.

4.4.3.1. Automatikus fordítón és n-gramon alapuló algoritmus 12 szócikkre

A Microsoft Bing fordítóval készült gépi visszafordítás után az alábbi hasonlóságokat találta meg az algoritmus, a $d+ = 7$, $d = 120$ illetve $d_{num}/D = 1$ paraméterekkel:

1. Munich Philharmonic (45)
2. Web accessibility (39)
3. Golden Twenties (30)
4. Wialon (28)
5. Upper 10 (26)
6. Castle Rock (Pineville, West Virginia) (26)
7. Meitetsu 5000 series (2008) (19)
8. Pete Seeger (19)
9. GPS tracking server (18)
10. Foreign relations of Pakistan (18)
11. Mike Seeger (18)
12. Politics of Pakistan (13)
13. Avidius Cassius (12)
14. University of Oxford (11)
15. Charles Seeger (8)

Illetve $d+ = 3$, $d = 40$ illetve $d_{num}/D = 1$ paraméterekkel:

1. Munich Philharmonic (45)
2. Web accessibility (39)
3. Golden Twenties (30)
4. Wialon (28)
5. Upper 10 (26)
6. Castle Rock (Pineville, West Virginia) (26)
7. Meitetsu 5000 series (2008) (19)
8. Pete Seeger (19)
9. Foreign relations of Pakistan (18)
10. GPS tracking server (18)
11. Mike Seeger (18)
12. Politics of Pakistan (13)
13. Avidius Cassius (12)
14. University of Oxford (11)
15. Charles Seeger (8)
16. Portal:University of Oxford/Intro (7)
17. Names for soft drinks (4)

Kimondottan jó eredményt ért el, 11-et megtalált mindkét esetben a 12-ből és ezeket előre is sorolta. Láthatóan az automatikus fordítók jobban működnek német-angol nyelvpárra, és ezért az automatikus visszafordíts sokkal kevesebb hibát visz bele a szövegbe, mint magyar-angol nyelvpár esetén.

4.4.3.2. Hasonlósági metrikán alapuló algoritmus 12 szócikkre

A hasonlósági metrikával az alábbi eredményt kaptuk:

1. Pete Seeger (7)
2. Munich Philharmonic (6)
3. Upper 10 (5)
4. Foreign relations of Pakistan (5)
5. Golden Twenties (4)
6. Wialon (4)
7. Politics of Pakistan (4)
8. Web accessibility (4)
9. Castle Rock (Pineville, West Virginia) (4)
10. Mike Seeger (3)
11. University of Oxford (2)
12. GPS tracking server (2)
13. Canberra 400 (1)

Ennek az eredménye egyezik a magyar-angol nyelvpárral, belátható módon ez az algoritmus nem függ annyira a nyelvpároktól, mint az n-gramon alapuló. Ugyanakkor érdemes megnézni a 9.15 mellékletben található listát, hogy milyen sok szó nem volt meg a szótárában (összesen 258), azaz egy jobb szótárral mennyit lehetett volna még javítani az eredményen. Ebben természetesen számos tulajdonnév van, amely mindkét nyelven azonos lenne, de a nagy része azért nem ilyen.

Az előző kísérletek alapján látható, hogy míg az n-gram alapú algoritmus eredménye nagyon függ az automatikus fordítás minőségétől, addig az új, hasonlósági metrikán alapuló algoritmus viszonylag egyenletes eredményt nyújt minden nyelven. Számos javítási lehetőség van még, mint láttuk korábban: 5 lehetséges fordításig még meredeken növekszik a fedés értéke, azaz a hiányzó fordításokat érdemes felvinni a rendszerbe, ezt akár be is lehet kérni a felhasználótól, hiszen a rendszer pontosan tudja mely szavak ezek, és még a keresés elvégzése előtt rá tud kérdezni a felhasználónál.

4.5. Fordításiplágium-kereső algoritmus – új eredmények összefoglalása

Létrehoztam egy új fordítási plágiumok megtalálására képes algoritmust, amely az n-szavas darabolás helyett mondatokra bontja a szöveget, és egy a fordítás menetét utánzó hasonlósági metrika segítségével hasonlítja össze a mondatokat egymással, hogy megállapítsa, mekkora eséllyel fordításai azok egymásnak. Az új algoritmus lényeges tulajdonsága, hogy nem kell hozzá gépi fordító, elég csupán egy szótár is, amit sokkal könnyebb beszerezni és folyamatosan fejleszteni. Megmutattam, hogy az adatbázis méretéhez viszonyított lineáris futási ideje konstansra csökkenthető információ-visszakereső algoritmus használatával. Megmutattam, hogy a gyakorlatban is használható, más algoritmusokkal összemérhető eredményt ér el, valamint, hogy a fedés értéke nem függ az adott nyelvpártól. A magyar-angol nyelvpár esetén a fedése 83%, míg a pontossága 40% volt, a német-angol nyelvpárnál a fedése 83% míg a pontossága 77% a 12 Wikipédia-cikket tartalmazó tesztkorpuszon.

5. Mondat alapú hasonlóság- és plágiumkeresés egy nyelven belül

5.1. Bevezetés

Az új, a 4. fejezetben ismertetett, fordítási plágiumok megkeresésére használt algoritmust akár egynyelvű szövegek összehasonlítására is alkalmazhatjuk. Ebben az esetben a szótári azonosság, illetve a fordítási függvény helyett szinonima-, esetleg antonima-, hiponima- és hipernima-azonosságokat vezethetünk be, és ezek alapján értékelhetjük két szöveg azonosságát. A korábban bemutatott képlethez hasonlóan ezt így definiálhatjuk:

$$w_j^y \in \mathrm{syno}\left(w_i^x\right) \Rightarrow w_j^y \equiv w_i^x$$

$$w_j^y \in \mathrm{anto}\left(w_i^x\right) \Rightarrow w_j^y \equiv w_i^x$$

$$w_j^y \in \mathrm{hypo}\left(w_i^x\right) \Rightarrow w_j^y \equiv w_i^x$$

$$w_j^y \in \mathrm{hyper}\left(w_i^x\right) \Rightarrow w_j^y \equiv w_i^x$$

Az, hogy ezekből a műveletekből melyeket használjuk, a felhasználástól függ. A szinonima használata egyértelműen elkerülhetetlen, antonimákat használni nagy mennyiségben egy szöveg átdolgozásánál hatalmas munka és valószínűleg sok tagadószó kerülne a szövegbe. A hipernimák használata sokkal egyszerűbb és kevésbé specifikussá is teszi a szöveget, például ahelyett, hogy:

Egy nagy tölgy alatt telepedtek le.

könnyen írhatjuk azt, hogy:

Egy nagy fa alatt telepedtek le.

Fordítva valószínűleg ritkábban fordul elő plágiumok esetében, azaz a hiponima használata ritkább, ugyanakkor ahhoz, hogy tudjuk, hogy melyik irányt kell alkalmaznunk, ahhoz tudnunk kéne, hogy melyik az eredeti, és melyik a másolt mű, ez viszont nem mindig van így. Egy egyetemi környezetben például nem lehetünk biztosak, hogy az eredeti mű került be hamarabb az adatbázisunkba, lehet, hogy a plagizált művet adták be hamarabb, vagy a korábbi dolgozatokat lassabban dolgozták

fel a könyvtárban. Hasonlóan, a Wikipédia az első számú forrása a plagizálásnak (Turnitin 2011), ugyanakkor volt már arra is példa, hogy a Wikipédiába kerültek be máshonnét átvett, lopott tartalmak. (Wikihu 2011)

Most nézzük meg négy angol szónak a lehetséges szino-, anto-, hiper-, és hiponimáit, hogy lássuk, mennyi hasonló szóról beszélünk. Ehhez a WordNet (Miller 1995) adatázist használtuk:

```
apple
Synonyms (2): malus pumila, orchard apple tree
Antonyms (0):
Hypernyms (4): edible fruit, false fruit, pome, apple tree
Hyponyms (5): crab apple, crabapple, dessert apple, eating
apple, cooking apple

cold
Synonyms (12): coldness, frigidity, frigidness, low
temperature, common cold, cold-blooded, inhuman, insensate,
frigid, dusty, moth-eaten, stale
Antonyms (1): hot
Hypernyms (8): communicable disease, respiratory disease,
respiratory disorder, respiratory illness, pressor,
vasoconstrictive, vasoconstrictor, temperature
Hyponyms (9): head cold, chill, gelidity, iciness,
chilliness, coolness, nip, frostiness, cool

go
Synonyms (71): go game, crack, fling, offer, pass, whirl,
adam, cristal, disco biscuit, ecstasy, hug drug, x, xtc,
spell, tour, turn, become, get, proceed, buy the farm, cash
in one's chips, choke, conk, croak, decease, die, drop dead,
exit, expire, give-up the ghost, kick the bucket, pass away,
perish, pop off, snuff it, break, break down, conk out, fail,
give out, give way, go bad, move, run, plump, run low, run
short, rifle, function, operate, work, locomote, travel,
depart, go away, get going, start, sound, endure, hold out,
hold up, last, live, live on, survive, fit, extend, lead,
blend, blend in, belong
Antonyms (1): no-go
Hypernyms (11): duty period, shift, work shift, mdma,
methylenedioxymethamphetamine, attempt, effort, endeavor,
endeavour, try, board game
Hyponyms (0):

ten
Synonyms (5): ten-spot | 10 | decade | tenner | x
Antonyms (0):
Hypernyms (2): large integer | spot
Hyponyms (0):
```

Több érdekes dolgot is felfedezhetünk: először is rengeteg kifejezés van, főleg a hiper-, és hiponimák között. Ez azért érdekes, mert ha az algoritmusunk nem támogatja a

kifejezéseket, vagy nem kezeli azokat megfelelően, akkor ez újabb hibákat vezethet be a rendszerbe. A második, amit észre lehet venni, hogy mivel nem használtunk jelentés-egyértelműsítést, így a különböző jelentésekhez tartozó szino-, anto-, hiper-, és hiponimák keverednek, amint a *go*, a *cold* és *ten* szavaknál megfigyelhetjük ezt.

5.2. Hasonlósági metrikán alapuló algoritmus tesztelése azonos nyelvű szövegek összehasonlítására

Miután részletesen teszteltük az algoritmust az angol WordNet adatbázisból a szinonimák, antonimák valamint a hipernimák összességét választottuk ki a fordításbeli *trans* függvény helyettesítésére. Az alárendelt hiponimákat azért hagytuk ki, mert a WordNetben legtöbb esetben a hétköznapi, gyakran használt szavak alatt is vannak még specifikusabb szintek, amelyek már nem igazán értelmesek erre a felhasználásra. Az *apple* (alma) szóra például az alábbi hiponimákat adja: *crab apple, crabapple, eating apple, dessert apple, cooking apple*. Ezen belül pedig az almafajtákat. A hiponimák kihagyása helyett sokkal elegánsabb lett volna megnézni, hogy mely gyakori szavak találhatóak meg a WordNetben, ezeket összekötni (akár pár köztes viszonyt kihagyva), és egy ilyen szűk szótárból dolgozni, de erre nem volt lehetőségünk. Ez egy következő kutatás témája lehet.

Most nézzük meg, hogy az előző fejezetben használt tesztdokumentumokra miként teljesít az egynyelvű kereső.

A **15 oldalas cikk** esetében az alábbi találatokat kapjuk:

1. Books LLC (1)
   A blog about plagiarism from a German professor, written in English.
   - A blog about plagiarism from a German professor, written in English. (18)
2. Word-sense disambiguation (1)
   In Proceedings of the 20th National Conference on Artificial Intelligence (AAAI, Pittsburgh, PA).
   - Proceedings of the 13th International Conference on Artificial Intelligence. pp. 83–92. (8)

3. Wolfgang Wendland (1)
   Google machine translation
   - Machine translation was done with Google Translate API. (8)
4. Tepper Aviation (1)
   Machine translation by Google.
   - Machine translation was done with Google Translate API. (8)

A rendszer láthatólag jól működött, és megtalálta azt az egy mondatot, ami az irodalomjegyzékben található Weber-Wulff *Copy, Shake, and Paste* blogjának (CSPblog) jellemzésére. Egy időben ez volt a blog alcíme, ami mára már kicsit megváltozott: „A blog about plagiarism and scientific misconduct". A Google 2,570 találatot ad erre az egy mondatra, hiszen a Wikipédián és a fenti cikken kívül is számos szerző így hivatkozik a blogra, ez az egy mondat tömören összefoglalta a blog lényegét. A többi egyezés láthatólag hasonlóság, de viszonylag lényegtelen mondatok esetében. A mögé írt számból látszik, hogy ezek épp a minimumot, 8 hasonlósági pontot, kaptak – pont elérték Sim_2 szintjét – míg az első, a tényleges átvétel, 18 pontot kapott.

12 Wikipédia cikk magyar fordításának Google Translate fordítóval történt angol visszafordítására is lefuttattuk az algoritmust, a teljes lista a találatok szövegével együtt megtalálható a 9.11.1 mellékletben.

1. Pete Seeger (7)
2. Foreign relations of Pakistan (5)
3. Politics of Pakistan (5)
4. Web accessibility (5)
5. Munich Philharmonic (4)
6. Wialon (4)
7. Avidius Cassius (4)
8. Mike Seeger (4)
9. Golden Twenties (3)
10. Upper 10 (3)
11. GPS tracking server (2)
12. Castle Rock (Pineville, West Virginia) (2)
13. Meitetsu 5000 series (2008) (2)
14. Charles Seeger (2)
15. University of Oxford (1)
16. Portal:University of Oxford/Intro (1)
17. Portal:South East England/Selected article/9 (1)
18. Portal:Oxfordshire/Selected article/8 (1)
19. Portal:University/Previous articles (1)
20. Kosmo (1)

A 12 szócikkből 11-et megtalált és az első 15 helyen szerepel is, ez egy kimondottan jó eredmény ahhoz képest, hogy ez egy kétszer fordított szöveg.

A 20. utáni találatok nagy része arra a mondatra illeszkedik, hogy a "The Wialon Linux and Windows operating system as available.". Ezek tulajdonképpen nem jó találatok, és egy utólagos szűréssel el lehet távolítani, hiszen az a mondat, amelyiknek ennyi szócikkben van megfelelője, nem jó találat.

Az érdekesség kedvéért egyébként a Microsoft Bing fordítójával is visszafordítottuk a magyar szöveget angolra, és így is megnéztük. Az eredmény nagyon hasonló (lásd 9.11.2 melléklet), valamivel rosszabb: 10 helyes találatot ad az első 17 találat között.

5.3. Azonos nyelvű szövegek összehasonlítása – új eredmények összefoglalása

Ebben a fejezetben bebizonyítottam, hogy az új, hasonlósági metrikán alapuló algoritmusom nem csak fordítások, hanem azonos nyelven írt szövegek összehasonlítására is alkalmas. Egy kétszeres fordításon átesett szöveg esetén 92% illetve 83% volt a mondat szintű fedés.

A mondaton belüli szórendre ez az algoritmus teljesen érzéketlen, ellentétben az n-gramon alapuló algoritmussal, ahol a szavak sorrendjének a változása a fedés csökkenését eredményezi. A találatok sorrendezése a mondatok száma és a találatok értéke szerint a 11 jó találatot az első 15 hely közé sorolta be, a hamis pozitív találatok egy kivétellel mind a lista végén szerepelnek. Ez lehetővé teszi, hogy az algoritmust egy valós rendszerbe is beépítsük, mert a felhasználó könnyen át tudja tekinteni a legfontosabb találatokat.

6. Az algoritmus implementálása és használata a gyakorlatban

6.1. Bevezetés

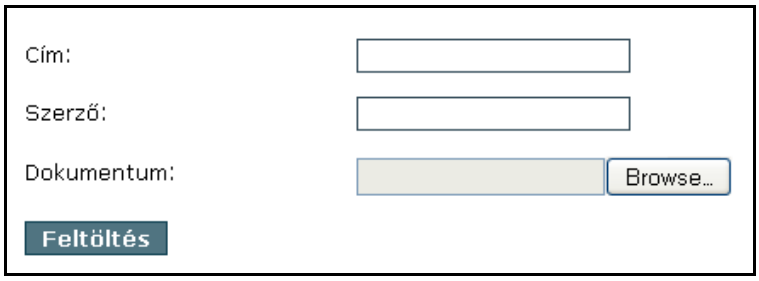
A fentebb ismertetett fordítási plágiumok keresésére irányuló kutatásnak az volt a célja, hogy kiderítsük, lehetséges-e, és ha igen, milyen hatásfokkal, angol és magyar nyelvek között fordítási plágiumokat felismerni. Mivel az eredmények nagyon bíztatóak, és az algoritmus a gyakorlatban is használhatónak bizonyult, lehetővé vált, hogy beépítésre kerüljön a SZTAKI KOPI Plágiumkeresőbe is. 2011. év végén a világon elsőként nyújtott fordítási plágiumkereső szolgáltatást a KOPI Portál.

6.2. A felhasználói felület

A KOPI Plágiumkeresőt elsősorban a hazai felsőoktatásban is elharapódzott plagizálás visszaszorítása fejlesztettük ki, de a házi feladatok és diplomadolgozatok, cikkek összehasonlításán felül még sok egyéb célra is alkalmas a rendszer. Oktatók számára lehetőséget biztosít a Rendszer a házidolgozatok és diplomák összehasonlítására, akár a korábbi évek hasonló műveivel, akár a KOPI adatbázisával és a Wikipédia teljes szövegével is. Diákok ellenőrizhetik a művüket, megnézhetik, hogy az összes idézet mennyisége nem haladta-e meg az oktatási intézményben megengedettet. Diplomájukat ők maguk feltölthetik, hogy másolás esetén látható legyen az eredeti szerző. Szerzők feltölthetik az eredeti művüket a KOPI portálba, és utána szabadon publikálhatják, közzétehetik, árulhatják. A KOPI védi a szerzői jogokat, ha valaki idéz a műből, akkor pillanatok alatt megtalálható az eredeti forrás. Bírálók használhatják a Plágiumkeresőt arra, hogy a szerző korábbi műveivel és a forrásként megjelölt cikkekkel összehasonlítsák a beadott cikket, így kiszűrhetőek a nem jelölt idézetek, és az önplagizálás – egy cikk, gondolat többszöri eladása – is. Konferenciaszervezőknek segít a cikkek minél egyedibbé, értékesebbé tételében, és a hasonló témájú cikkek, szerzők megtalálásában. A SZTAKI KOPI plágiumkereső szolgáltatása a http://kopi.sztaki.hu címen érhető el.

6.2.1. Dokumentum feltöltése

A plágiumkeresés öt fő lépésből áll, melyet a felhasználói felület is tükröz. A felhasználónak először fel kell töltenie azt a dokumentumot, amelyet össze szeretne hasonlítani más forrásokkal. A rendszer jelenleg html, doc, docx, rtf, txt és pdf formátumú dokumentumokat kezel. Érdemes kitölteni a dokumentum címét és szerzőjét, hogy pontosan lehessen látni a keresés eredményénél, hogy ugyanaz a dokumentum szerepel kétszer a rendszerben, vagy tényleges egyezésről van szó.



6.1. ábra: Dokumentum feltöltése

6.2.2. Dokumentum(ok) kiválasztása

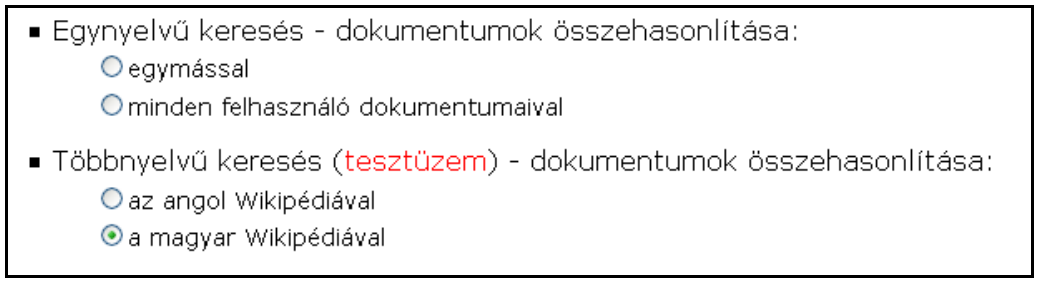
A feltöltés után ki kell választani egy dokumentumot (6.2. ábra), amelyet az adatbázissal, vagy több dokumentumot, amelyeket egymással szeretnénk összehasonlítani.

6.2. ábra: Feltöltött dokumentum

Amennyiben a dokumentum melletti jelölőnégyzet helyén egy kis ikon van, akkor nem választható ki, és nem indítható vele keresés. A kis háromszög () jelentése, hogy a dokumentumot a rendszer nem tudta értelmezni, nem tudta konvertálni, és ezért nem használható plágiumkeresésre. Ilyenkor érdemes a dokumentumot más formátumban feltölteni. A kis óra () jelentése, hogy a dokumentum feldolgozás alatt van, ez általában pár perc alatt megtörténik, de ha nagyon le van terhelve a rendszer, akkor elképzelhető, hogy egy órát is várni kell rá.

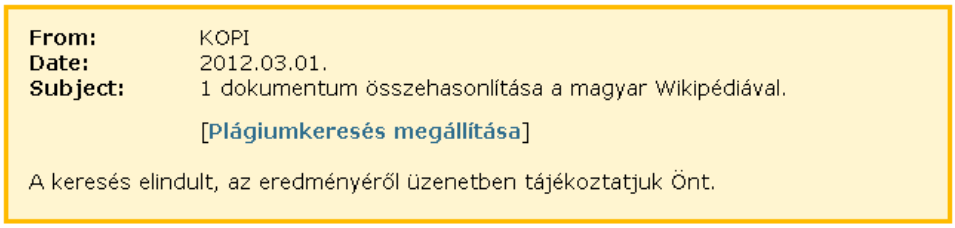
6.2.3. Keresési lehetőségek kiválasztása

Attól függően jelennek meg a választható keresési lehetőségek (6.3. ábra), hogy hány dokumentumot választottunk ki. Egy dokumentum esetén azt összehasonlíthatjuk a KOPI adatbázisával (minden felhasználó dokumentumával), ez jelenleg körülbelül 35 000 dokumentumot jelent. Ugyancsak lehetőségünk van a dokumentumot összehasonlítani az angol vagy magyar Wikipédiával, melyek a 4. és 5. fejezetekben ismertettet algoritmusokkal való keresést jelentik. Amennyiben több dokumentumot választottunk ki, akkor azokat egymással is összehasonlíthatjuk – ez a funkció alkalmas egy dolgozatban található szakirodalmak mennyiségének megállapítására vagy hasonló témában íródott dokumentumok összehasonlítására is.



6.3. ábra: Plágiumkeresési lehetőségek

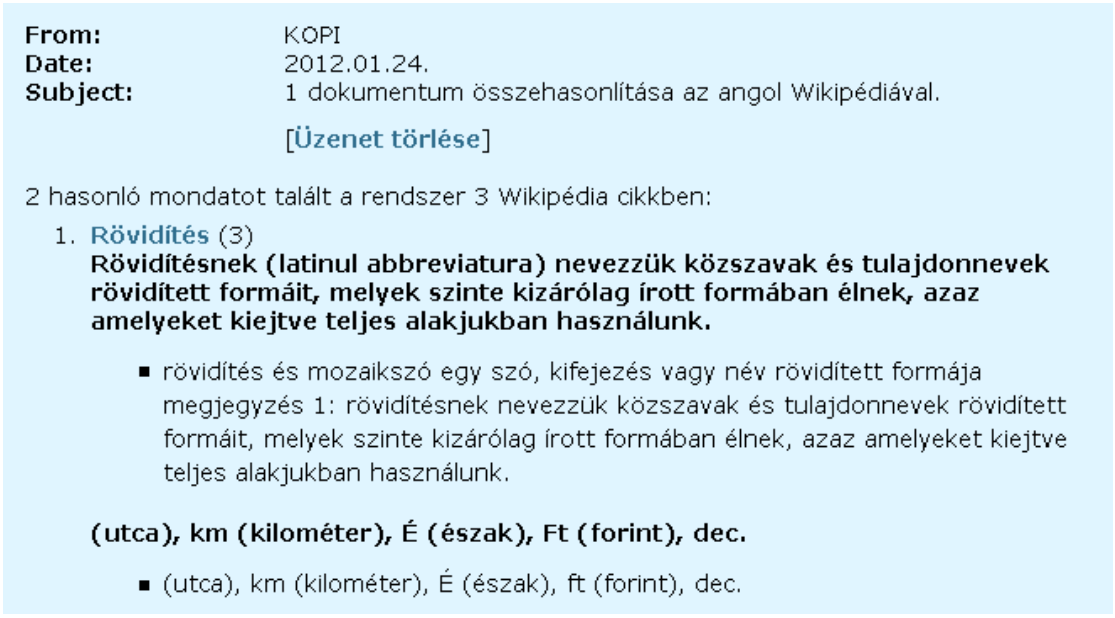
Miután kiválasztottuk a megfelelő keresést, például a magyar Wikipédiát, elindíthatjuk a plágiumkeresést. Erről a rendszer egy kis üzenetben tájékoztat minket.



6.4. ábra: Plágiumkeresés fut

A keresés eredménye

A kereséseket a kereső beérkezési sorrendben dolgozza fel. A rendszer leterheltségétől függően az eredmény pár perc vagy pár óra múlva jelenik meg az üzenetek között, és ha a felhasználó nem tiltotta le, akkor az eredményről egy email üzenetet is kap.



6.5. ábra: A plágiumkeresés eredménye, magyar-magyar keresés

A Wikipédiával történő összehasonlításkor az üzenet tartalmazza a Wikipédia szócikk nevét, a szócikkben talált mondatokat, valamint azokat a mondatokat, amelyekhez a dokumentumon belül hasonlított (lásd 6.5. és 6.6. ábra). Ez történhet egy nyelven is, mint a 6.5. ábrán látható, de lehet a cikk magyar nyelvű és a dokumentum angol nyelvű, vagy fordítva.

6.6. ábra: A plágiumkeresés eredménye, magyar-angol keresés

Fontos kiemelni, hogy ezeket a rendszereket plágiumkeresőnek hívjuk, de tulajdonképpen hasonlóságot keres, azaz nem különbözteti meg az idézetet a plágiumtól, ennek eldöntését mindig a felhasználóra bízza.

6.3. Az algoritmus implementálásának tapasztalatai

Az új fordítási plágiumkereső 2011 decemberi integrálása után két hónappal derül ki, hogy Schmitt Pál plagizálta doktori disszertációját. Ez nagyban megnövelte a felhasználóink számát, és elősegítette a közvélemény felhívását erre az igen fontos problémára. A jelenlegi statisztikák alapján körülbelül havi 1 500 plágiumkeresést végez a rendszer: 57%-ban a rendszerbe feltöltött dokumentumokhoz viszonyítanak egy másikat a felhasználók, 17%-ban dokumentumokat egymáshoz és a maradék 26 százalékban használják a fordítási plágiumkeresőt. Ez utóbbi jelenleg az angol és a magyar Wikipédiát takarja, amelyhez magyar, angol és német nyelvű szövegeket hasonlíthatnak.

Az algoritmus kialakítása egy évet vett igénybe. A visszajelzések alapján jól működik, jobban használható, és kevesebb panasz érkezik rá, mint az egynyelvű keresésre. Ennek az is az oka, hogy képes megjeleníteni az egyező részek szövegét.

7. Összefoglalás, továbbfejlesztési lehetőségek

A dolgozatban bemutatásra került a **félig átlapolódó szavas darabolás**, amely az egynyelvű plágiumkeresés esetén teszi lehetővé, hogy az adatbázisunk jelentősen kisebb legyen, miközben a pontossága nem romlik jelentősen. Ez az új darabolási eljárás képezi a KOPI Plágiumkereső Rendszer egynyelvű algoritmusának az alapját, és az elmúlt 8 évben a gyakorlatban is bizonyította, hogy jól teljesít.

A **többnyelvű dokumentumok nyelvének megállapításá**ra is mutattunk egy újszerű megoldást, amely egy ismert algoritmus továbbfejlesztése. A tesztek során bebizonyosodott, hogy ez jól alkalmazható olyan szövegek azonosítására, melyek minimum 30%-ban tartalmaznak más nyelven íródott részeket is, még akkor is, ha ezek nem tömbösítve, hanem elszórtan szerepelnek, mint egy szótárban például.

A **fordítási plágiumkeresés**re kifejlesztett **hasonlósági metrikán alapuló algoritmus** egy alternatívája az n-gram alapú – az egynyelvű keresésre használt adatbázisra, illetve gépi fordító algoritmusra épülő – megoldásoknak. Már most lehet látni, több továbbfejlesztési irányt. Az információ-visszakeresési részben van a legtöbb tartalék, annak a lépésnek javításával a fedés értéke valamint az algoritmus sebessége is sokat javulhat. A WordNet megfelelő megszűrésével valószínűleg még több átírást képes lesz kezelni. A kifejezéseket, szóösszetételeket nem kezeli egyáltalán jelenleg az algoritmus, ez mindenképp javítaná a hasonlósági metrika pontosságát. Fordítás során előfordul, hogy egy mondatból kettő vagy több lesz, illetve ennek a fordítottja is. Ennek a kezelésére egy adaptív megoldást lehetne alkalmazni, melyben a nem talált rövid mondatokat összevonja a rendszer, illetve a hosszúaknak egy-egy részét is megpróbálja megkeresni az adatbázisában. Az is elképzelhető, hogy egy olyan adatbázist építünk, amelyben a mondatok kettesével összevonva is szerepelnek már, de ez valószínűleg túlságosan megnövelné a rendszer erőforrásigényét, így annak ellenére, hogy gyors és pontos megoldás lehet, a mondatok darabolása valószínűleg a gyakorlatban jobban megvalósítható megoldás.

Az 5. fejezetben megnéztük, hogy az **egynyelvű keresés**t miként lehetne kiváltani a **hasonlósági metrikán alapuló algoritmus**sal, és azt láthattuk, hogy kétszer fordított szövegben is megtalálta a hasonlóságot az eredeti művekkel. Ennek a dolgozatnak nem

volt célja, hogy részletesen összevesse az egynyelvű keresést és az n-gram algoritmust, de egy további cikkben ezt mindenképp érdemes lehet megtenni.

Az utolsó részben ismertettük, hogy ezek az **algoritmusok implementálásra kerültek**, ez ugyan nem kutatás volt, de jól bizonyítja, hogy az eredmények a gyakorlatban is hasznosultak, és mindenki számára kipróbálhatóvá, tesztelhetővé teszi az ismertetett algoritmusokat, és nem utolsósorban egy hasznos szolgáltatássá, amit nap-mint nap használnak tanárok, diákok és kutatók egyaránt.

8. Köszönetnyilvánítás

Isaac Newton szavaival élve „If I have seen further, it is by standing on the shoulders of giants.”, aki amikor ezt mondta, éppen Bernard of Chartres vállán állt.

Köszönöm szépen szüleimnek a támogatást, és hogy olyan magasra tették a lécet. Feleségemnek és Édesapámnak a folyamatos és fáradhatatlan ösztönzést.

Köszönöm Monostori Krisztiánnak, hogy több mint 10 éve bevezetett a plágiumkeresés témakörébe; Hodász Gábornak a közös munkát, az első lépéseket ezen a területen; Arkady Zaslavskynak az ausztrál ösztöndíjat, ahol még jobban beleáshattam magam ebbe a témába.

Köszönöm szépen Kovács Lászlónak, hogy meghívott, hogy dolgozzak a SZTAKI-ban; Micsik Andrásnak a szakmai támogatást; Tóth Zoltánnak a KOPI rendszer elkészítésében, Vajna Miklósnak a nyelvfelismerő algoritmus tesztelésében és implementálásában, Pataki Balázsnak pedig a fordításiplágium-kereső algoritmus implementációja során nyújtott segítséget; Pallinger Péternek a rendszergazdai támogatását a kísérletekhez; Zsivnovszki Magdolnának és Virág Évának az angol cikkek angolosítását és a magyar cikkek magyarosítását; Inzelt Péternek a belső pályázati lehetőséget, aminek a segítségével be tudtam fejezni a kutatásomat.

Köszönöm szépen Prószéky Gábornak támogatását a Pázmányon töltött két év alatt és a részletes, mindenre kiterjedő lektorálásokat.

9. Mellékletek

9.1. A Szeretet himnusza három fordításban

Károli Gáspár fordítása	**Református fordítás**	**Katolikus fordítás**
Ha embereknek vagy angyaloknak nyelvén szólok is, szeretet pedig nincsen én bennem, olyanná lettem, mint a zengő ércz vagy pengő czimbalom.	Ha emberek vagy angyalok nyelvén szólok is, szeretet pedig nincs bennem, olyanná lettem, mint a zengő érc vagy pengő cimbalom.	Szólhatok az emberek vagy az angyalok nyelvén, ha szeretet nincs bennem, csak zengő érc vagyok vagy pengő cimbalom.
És ha jövendőt tudok is mondani, és minden titkot és minden tudományt ismerek is; és ha egész hitem van is, úgyannyira, hogy hegyeket mozdíthatok ki helyökről, szeretet pedig nincsen én bennem, semmi vagyok.	És ha prófétálni is tudok, ha minden titkot ismerek is, és minden bölcsességnek birtokában vagyok, és ha teljes hitem van is, úgyhogy hegyeket mozdíthatok el, szeretet pedig nincs bennem: semmi vagyok.	Lehet prófétáló tehetségem, ismerhetem az összes titkokat és mind a tudományokat, hitemmel elmozdíthatom a hegyeket, ha szeretet nincs bennem, mit sem érek.
És ha vagyonomat mind felétetem is, és ha testemet tűzre adom is, szeretet pedig nincsen én bennem, semmi hasznom abból.	És ha szétosztom az egész vagyonomat, és testem tűzhalálra szánom, szeretet pedig nincs bennem: semmi hasznom abból.	Szétoszthatom mindenemet a nélkülözők közt, odaadhatom a testemet is égőáldozatul, ha szeretet nincs bennem, mit sem használ nekem.
A szeretet hosszútűrő, kegyes; a szeretet nem irígykedik, a szeretet nem kérkedik, nem fuvalkodik fel.	A szeretet türelmes, jóságos; a szeretet nem irigykedik, a szeretet nem kérkedik, nem fuvalkodik fel.	A szeretet türelmes, a szeretet jóságos, a szeretet nem féltékeny, nem kérkedik, nem is kevély.
Nem cselekszik éktelenül, nem keresi a maga hasznát, nem gerjed haragra, nem rójja fel a gonoszt,	Nem viselkedik bántóan, nem keresi a maga hasznát, nem gerjed haragra, nem rója fel a rosszat.	Nem tapintatlan, nem keresi a maga javát, nem gerjed haragra, a rosszat nem rója fel.
Nem örül a hamisságnak, de együtt örül az igazsággal;	Nem örül a hamisságnak, de együtt örül az igazsággal.	Nem örül a gonoszságnak, örömét az igazság győzelmében leli.
Mindent elfedez, mindent hiszen, mindent remél, mindent eltűr.	Mindent elfedez, mindent hisz, mindent remél, mindent eltűr.	Mindent eltűr, mindent elhisz, mindent remél, mindent elvisel.
A szeretet soha el nem fogy: de legyenek bár jövendőmondások, eltöröltetnek; vagy akár nyelvek, megszünnek; vagy akár ismeret, eltöröltetik.	A szeretet soha el nem múlik. De legyen bár prófétálás: el fog töröltetni; legyen nyelveken való szólás: meg fog szűnni; legyen ismeret: el fog töröltetni.	S a szeretet nem szűnik meg soha. A prófétálás végetér, a nyelvek elhallgatnak, a tudomány elenyészik.
Mert rész szerint van bennünk az ismeret, rész szerint a prófétálás:	Mert töredékes az ismeretünk és töredékes a prófétálásunk.	Most megismerésünk csak töredékes, és töredékes a prófétálásunk is.
De mikor eljő a teljesség, a rész szerint való eltöröltetik.	Amikor pedig eljön a tökéletes, eltöröltetik a töredékes.	Ha azonban elérkezik a tökéletes, ami töredékes, az véget ér.
Mikor gyermek valék, úgy szóltam, mint gyermek, úgy gondolkodtam, mint gyermek, úgy értettem, mint gyermek: minekutána pedig férfiúvá lettem, elhagytam a gyermekhez illő dolgokat.	Amikor gyermek voltam, úgy szóltam, mint gyermek, úgy éreztem, mint gyermek, úgy gondolkoztam, mint gyermek; amikor pedig férfivá lettem, elhagytam a gyermeki dolgokat.	Gyermekkoromban úgy beszéltem, mint a gyerek, úgy gondolkoztam, mint a gyerek, úgy ítéltem, mint a gyerek. De amikor elértem a férfikort, elhagytam a gyerek szokásait.
Mert most tükör által homályosan látunk, akkor pedig színről-színre; most rész szerint van bennem az ismeret, akkor pedig úgy ismerek majd, a mint én is megismertettem.	Mert most tükör által homályosan látunk, akkor pedig színről színre; most töredékes az ismeretem, akkor pedig úgy fogok ismerni, ahogyan engem is megismert az Isten.	Ma még csak tükörben, homályosan látunk, akkor majd színről színre. Most még csak töredékes a tudásom, akkor majd úgy ismerek mindent, ahogy most engem ismernek.

Most azért megmarad a hit, remény, szeretet, e három; ezek között pedig legnagyobb a szeretet.	Most azért megmarad a hit, a remény, a szeretet, e három; ezek közül pedig a legnagyobb a szeretet.	Addig megmarad a hit, a remény és a szeretet, ez a három, de közülük a legnagyobb a szeretet.

9.2. A bibliai tesztdokumentumok hasonlóságai

9.2.1. Átlapolódó szavas darabolás

file1	file2	o1	o2	o3	o4	o5	o6	o7	o8	o9	o10	o11	o12	o13	o14	o15	o16	o17	o18	o19	o20
1kor_kat.txt	1kor13_karoli.txt	2	0	0	0																
1kor_kat.txt	1kor13_kat.txt	3	3	3	3	3	3	3	2	2	2	2	2	2	2	2	2	2	2	2	2
1kor_kat.txt	1kor13_ref.txt	2	0	0	0	0															
1kor_kat.txt	1moz_kat.txt	32	4	0																	
1kor_kat.txt	2kor_kat.txt	35	6	0	0	0															
1kor_kat.txt	2moz20_karoli.txt	3	0	0																	
1kor_kat.txt	2moz20_kat.txt	3	0	0																	
1kor_kat.txt	2moz20_ref.txt	3	0	0																	
1kor13_karoli.txt	1kor_kat.txt	73	18	5	0																
1kor13_karoli.txt	1kor13_kat.txt	43	14	5	0																
1kor13_karoli.txt	1kor13_ref.txt	67	43	30	22	16	12	9	6	4	2	1	0								
1kor13_karoli.txt	1moz_kat.txt	52	3																		
1kor13_karoli.txt	2kor_kat.txt	54	4																		
1kor13_karoli.txt	2moz20_karoli.txt	25	0																		
1kor13_karoli.txt	2moz20_kat.txt	21																			
1kor13_karoli.txt	2moz20_ref.txt	22																			
1kor13_kat.txt	1kor_kat.txt	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
1kor13_kat.txt	1kor13_karoli.txt	48	15	5	0																
1kor13_kat.txt	1kor13_ref.txt	58	26	11	2	0															
1kor13_kat.txt	1moz_kat.txt	50	6	0																	
1kor13_kat.txt	2kor_kat.txt	53	6																		
1kor13_kat.txt	2moz20_karoli.txt	23	0																		
1kor13_kat.txt	2moz20_kat.txt	27	0																		
1kor13_kat.txt	2moz20_ref.txt	26	0																		
1kor13_ref.txt	1kor_kat.txt	78	28	10	2	0															
1kor13_ref.txt	1kor13_karoli.txt	68	44	31	22	17	12	9	6	4	2	1	0								
1kor13_ref.txt	1kor13_kat.txt	54	24	10	2	0															
1kor13_ref.txt	1moz_kat.txt	53	4																		
1kor13_ref.txt	2kor_kat.txt	56	3																		
1kor13_ref.txt	2moz20_karoli.txt	25	0																		
1kor13_ref.txt	2moz20_kat.txt	23	0																		
1kor13_ref.txt	2moz20_ref.txt	24	0																		
1moz_kat.txt	1kor_kat.txt	33	4	0																	
1moz_kat.txt	1kor13_karoli.txt	1	0																		
1moz_kat.txt	1kor13_kat.txt	1	0	0																	
1moz_kat.txt	1kor13_ref.txt	1	0																		
1moz_kat.txt	2kor_kat.txt	23	3	0																	
1moz_kat.txt	2moz20_karoli.txt	3	0	0	0	0	0														
1moz_kat.txt	2moz20_kat.txt	3	0	0	0	0	0														
1moz_kat.txt	2moz20_ref.txt	3	0	0																	
2kor_kat.txt	1kor_kat.txt	56	11	0	0	0															
2kor_kat.txt	1kor13_karoli.txt	2	0																		
2kor_kat.txt	1kor13_kat.txt	2	0																		
2kor_kat.txt	1kor13_ref.txt	2	0																		
2kor_kat.txt	1moz_kat.txt	35	4	0																	
2kor_kat.txt	2moz20_karoli.txt	5	0	0																	
2kor_kat.txt	2moz20_kat.txt	4	0	0																	
2kor_kat.txt	2moz20_ref.txt	4	0	0																	
2moz20_karoli.txt	1kor_kat.txt	52	11	0																	
2moz20_karoli.txt	1kor13_karoli.txt	12	0																		
2moz20_karoli.txt	1kor13_kat.txt	10	0																		
2moz20_karoli.txt	1kor13_ref.txt	12	0																		
2moz20_karoli.txt	1moz_kat.txt	54	10	1	0	0	0														
2moz20_karoli.txt	2kor_kat.txt	47	7	0																	
2moz20_karoli.txt	2moz20_kat.txt	46	19	10	5	3	1	0	0	0											
2moz20_karoli.txt	2moz20_ref.txt	59	32	19	11	7	5	3	2	2	2	1	1	1	1	0	0	0	0		
2moz20_kat.txt	1kor_kat.txt	55	8	0																	
2moz20_kat.txt	1kor13_karoli.txt	12																			
2moz20_kat.txt	1kor13_kat.txt	14	0																		
2moz20_kat.txt	1kor13_ref.txt	13	0																		
2moz20_kat.txt	1moz_kat.txt	56	13	3	1	0	0														
2moz20_kat.txt	2kor_kat.txt	51	6	0																	
2moz20_kat.txt	2moz20_karoli.txt	53	22	12	6	3	1	1	0	0											
2moz20_kat.txt	2moz20_ref.txt	58	27	12	5	2	1	0													
2moz20_ref.txt	1kor_kat.txt	54	9	0																	
2moz20_ref.txt	1kor13_karoli.txt	12																			
2moz20_ref.txt	1kor13_kat.txt	13	0																		
2moz20_ref.txt	1kor13_ref.txt	13	0																		
2moz20_ref.txt	1moz_kat.txt	56	13	1																	
2moz20_ref.txt	2kor_kat.txt	48	7	0																	
2moz20_ref.txt	2moz20_karoli.txt	66	35	21	12	8	5	4	2	2	2	2	1	1	1	1	0	0	0		
2moz20_ref.txt	2moz20_kat.txt	57	26	12	5	2	1	0													

9.2.2. Mondatonkénti darabolás

file1	file2	Sentence
1kor_kat.txt	1kor13_karoli.txt	0
1kor_kat.txt	1kor13_kat.txt	3
1kor_kat.txt	1kor13_ref.txt	0
1kor_kat.txt	1moz_kat.txt	0
1kor_kat.txt	2kor_kat.txt	1
1kor_kat.txt	2moz20_karoli.txt	
1kor_kat.txt	2moz20_kat.txt	
1kor_kat.txt	2moz20_ref.txt	0
1kor13_karoli.txt	1kor_kat.txt	5
1kor13_karoli.txt	1kor13_kat.txt	5
1kor13_karoli.txt	1kor13_ref.txt	28
1kor13_karoli.txt	1moz_kat.txt	
1kor13_karoli.txt	2kor_kat.txt	
1kor13_karoli.txt	2moz20_karoli.txt	
1kor13_karoli.txt	2moz20_kat.txt	
1kor13_karoli.txt	2moz20_ref.txt	
1kor13_kat.txt	1kor_kat.txt	100
1kor13_kat.txt	1kor13_karoli.txt	5
1kor13_kat.txt	1kor13_ref.txt	9
1kor13_kat.txt	1moz_kat.txt	
1kor13_kat.txt	2kor_kat.txt	
1kor13_kat.txt	2moz20_karoli.txt	
1kor13_kat.txt	2moz20_kat.txt	
1kor13_kat.txt	2moz20_ref.txt	
1kor13_ref.txt	1kor_kat.txt	10
1kor13_ref.txt	1kor13_karoli.txt	31
1kor13_ref.txt	1kor13_kat.txt	10
1kor13_ref.txt	1moz_kat.txt	
1kor13_ref.txt	2kor_kat.txt	
1kor13_ref.txt	2moz20_karoli.txt	
1kor13_ref.txt	2moz20_kat.txt	
1kor13_ref.txt	2moz20_ref.txt	
1moz_kat.txt	1kor_kat.txt	0
1moz_kat.txt	1kor13_karoli.txt	
1moz_kat.txt	1kor13_kat.txt	
1moz_kat.txt	1kor13_ref.txt	
1moz_kat.txt	2kor_kat.txt	0
1moz_kat.txt	2moz20_karoli.txt	
1moz_kat.txt	2moz20_kat.txt	
1moz_kat.txt	2moz20_ref.txt	
2kor_kat.txt	1kor_kat.txt	1
2kor_kat.txt	1kor13_karoli.txt	
2kor_kat.txt	1kor13_kat.txt	
2kor_kat.txt	1kor13_ref.txt	
2kor_kat.txt	1moz_kat.txt	0
2kor_kat.txt	2moz20_karoli.txt	
2kor_kat.txt	2moz20_kat.txt	
2kor_kat.txt	2moz20_ref.txt	
2moz20_karoli.txt	1kor_kat.txt	
2moz20_karoli.txt	1kor13_karoli.txt	
2moz20_karoli.txt	1kor13_kat.txt	
2moz20_karoli.txt	1kor13_ref.txt	
2moz20_karoli.txt	1moz_kat.txt	
2moz20_karoli.txt	2kor_kat.txt	
2moz20_karoli.txt	2moz20_kat.txt	4
2moz20_karoli.txt	2moz20_ref.txt	15
2moz20_kat.txt	1kor_kat.txt	
2moz20_kat.txt	1kor13_karoli.txt	
2moz20_kat.txt	1kor13_kat.txt	
2moz20_kat.txt	1kor13_ref.txt	
2moz20_kat.txt	1moz_kat.txt	
2moz20_kat.txt	2kor_kat.txt	
2moz20_kat.txt	2moz20_karoli.txt	4
2moz20_kat.txt	2moz20_ref.txt	9
2moz20_ref.txt	1kor_kat.txt	1
2moz20_ref.txt	1kor13_karoli.txt	
2moz20_ref.txt	1kor13_kat.txt	
2moz20_ref.txt	1kor13_ref.txt	
2moz20_ref.txt	1moz_kat.txt	
2moz20_ref.txt	2kor_kat.txt	
2moz20_ref.txt	2moz20_karoli.txt	15
2moz20_ref.txt	2moz20_kat.txt	8

9.2.3. Hash-kódon alapuló darabolás

Fájl1	Fájl2	h1	h2	h3	h4	h5	h6	h7	h8	h9	h10	h11	h12	h13	h14	h15	h16	h17	h18	h19	h20
1kor_kat.txt	1kor13_karoli.txt	2	1	0	0	0		0	0		0	0					0				0
1kor_kat.txt	1kor13_kat.txt	3	3	2	3	3	2	3	3	2	4	4	2	2	3	3	2	1	3	2	4
1kor_kat.txt	1kor13_ref.txt	2	0	0	0	0		0	0		0	0		0			0				0
1kor_kat.txt	1moz_kat.txt	32	6	8	3	4	0	0	1	0	1	3		0			0				0
1kor_kat.txt	2kor_kat.txt	35	12	8	8	5	1	0	3	0	3	3	0	0			2	0		1	2
1kor_kat.txt	2moz20_karoli.txt	3	0	0	0	0	0		0	0	0	0									0
1kor_kat.txt	2moz20_kat.txt	3	0	0	0	0			0		0	0		0						0	
1kor_kat.txt	2moz20_ref.txt	3	0	0	0	0	0		0	0	0	0									0
1kor13_karoli.txt	1kor_kat.txt	73	29	20	18	14		3	11		6	11					4				8
1kor13_karoli.txt	1kor13_kat.txt	43	16	4	8	12	3	3	8		6	2			5	7	4				4
1kor13_karoli.txt	1kor13_ref.txt	67	36	22	25	26	6	21	22		21	22		8	5	7	23	15			16
1kor13_karoli.txt	1moz_kat.txt	52	17	15	13	8		6	2		3	8									
1kor13_karoli.txt	2kor_kat.txt	54	19	7	15	8			5		3	13									
1kor13_karoli.txt	2moz20_karoli.txt	25	6	5	5	2			5			5									
1kor13_karoli.txt	2moz20_kat.txt	21	5	2	3				2			2									
1kor13_karoli.txt	2moz20_ref.txt	22	7	4	6	2			5			2									
1kor13_kat.txt	1kor_kat.txt	100	100	98	96	95	95	91	94	84	94	93	83	87	81	80	87	75	77	77	92
1kor13_kat.txt	1kor13_karoli.txt	48	17	5	9	12	4	4	8		5	3			9	10	6				4
1kor13_kat.txt	1kor13_ref.txt	58	17	9	7	18	4	4	8		11	3		6	9	10	6				8
1kor13_kat.txt	1moz_kat.txt	50	17	7	11	12			2		5	3					6				
1kor13_kat.txt	2kor_kat.txt	53	18	7	13	12			5		5	3					6				4
1kor13_kat.txt	2moz20_karoli.txt	23	1	7	1	2			2		2	3									
1kor13_kat.txt	2moz20_kat.txt	27	1	5		4						3									
1kor13_kat.txt	2moz20_ref.txt	26	5	5	5	2			2		2	3									
1kor13_ref.txt	1kor_kat.txt	78	25	17	12	22		2	10		11	6		9			7				7
1kor13_ref.txt	1kor13_karoli.txt	68	42	28	30	27	10	19	28		19	18		9	5	10	35	15			14
1kor13_ref.txt	1kor13_kat.txt	54	18	8	8	18	5	2	10		11	2		9	5	10	7				7
1kor13_ref.txt	1moz_kat.txt	53	11	10	6	14		5	3		5	6									
1kor13_ref.txt	2kor_kat.txt	56	15	7	8	12			3		5	9									
1kor13_ref.txt	2moz20_karoli.txt	25	2	5	2	4			3		5	4									
1kor13_ref.txt	2moz20_kat.txt	23	1	3								2									
1kor13_ref.txt	2moz20_ref.txt	24	4	5	4	4			3		5	2									
1moz_kat.txt	1kor_kat.txt	33	7	8	4	5	0	0	2	0	2	3		0			1				0
1moz_kat.txt	1kor13_karoli.txt	1	0	0	0	0		0	0		0	0									
1moz_kat.txt	1kor13_kat.txt	1	0	0	0	0			0		0	0					0				
1moz_kat.txt	1kor13_ref.txt	1	0	0	0	0		0	0		0	0									
1moz_kat.txt	2kor_kat.txt	23	5	5	2	3	0	0	1	0	1	2		0			0				0
1moz_kat.txt	2moz20_karoli.txt	3	0	0	0	0	0	0	0	0	0	0	0								0
1moz_kat.txt	2moz20_kat.txt	3	0	0	0	0		0	0			0									
1moz_kat.txt	2moz20_ref.txt	3	1	0	0	0	0	0	1		0	0									0
2kor_kat.txt	1kor_kat.txt	56	18	13	11	9	1	1	5	1	5	5	0	1			3	1		3	3
2kor_kat.txt	1kor13_karoli.txt	2	1	0	0	0			0		0	1									
2kor_kat.txt	1kor13_kat.txt	2	0	0	0	0			0		0	0					0				0
2kor_kat.txt	1kor13_ref.txt	2	0	0	0	0			0		0	0									
2kor_kat.txt	1moz_kat.txt	35	7	7	3	4	0	1	1	0	1	4		0			0				0
2kor_kat.txt	2moz20_karoli.txt	5	0	0	0	0		0	0		0	0									0
2kor_kat.txt	2moz20_kat.txt	4	0	1	0	0			0			0		0							
2kor_kat.txt	2moz20_ref.txt	4	1	0	0	0			0		0	0								0	0
2moz20_karoli.txt	1kor_kat.txt	52	8	13	5	8	1		5	4	5	3									3
2moz20_karoli.txt	1kor13_karoli.txt	12	3	3	3	1			3			3									
2moz20_karoli.txt	1kor13_kat.txt	10	0	3	1	1			1		2	1									
2moz20_karoli.txt	1kor13_ref.txt	12	1	2	1	2			1		5	3									
2moz20_karoli.txt	1moz_kat.txt	54	10	13	8	7	1	3	7	2	5	4	2								3
2moz20_karoli.txt	2kor_kat.txt	47	4	9	3	8		1	5		5	3									3
2moz20_karoli.txt	2moz20_kat.txt	46	13	11	7	2	1		3		2	9		5							
2moz20_karoli.txt	2moz20_ref.txt	59	25	24	20	14	12	1	23	16	10	14	14						17	9	7
2moz20_kat.txt	1kor_kat.txt	55	10	13	8	6			2		3	3		9						4	
2moz20_kat.txt	1kor13_karoli.txt	12	3	1	2				2			1									
2moz20_kat.txt	1kor13_kat.txt	14	0	2		2						1									
2moz20_kat.txt	1kor13_ref.txt	13	0	1								1									
2moz20_kat.txt	1moz_kat.txt	56	13	13	6	2		3	2			7									
2moz20_kat.txt	2kor_kat.txt	51	9	14	8	2			2			3		3							
2moz20_kat.txt	2moz20_karoli.txt	53	17	12	9	2	2		5		3	11		3							
2moz20_kat.txt	2moz20_ref.txt	58	16	12	9	2	4	3	5			9									
2moz20_ref.txt	1kor_kat.txt	54	14	11	8	6	1		10	3	8	4									7
2moz20_ref.txt	1kor13_karoli.txt	12	4	2	3	1			3			1									
2moz20_ref.txt	1kor13_kat.txt	13	3	2	2	1			1		2	1									
2moz20_ref.txt	1kor13_ref.txt	13	2	2	1	3			1		5	1									
2moz20_ref.txt	1moz_kat.txt	56	15	14	11	11	3	2	11		11	4									11
2moz20_ref.txt	2kor_kat.txt	48	12	9	8	6			8		8	4								7	7
2moz20_ref.txt	2moz20_karoli.txt	66	28	25	19	16	14	2	22	21	11	14	15						25	7	7
2moz20_ref.txt	2moz20_kat.txt	57	14	12	6	3	3	2	3			7									

9.2.4. Átlapolódó hash-kódon alapuló darabolás

file1	file2	m7,8,9	m7,9	m5,9
1kor_kat.txt	1kor13_karoli.txt	0	0	0
1kor_kat.txt	1kor13_kat.txt	3	3	3
1kor_kat.txt	1kor13_ref.txt	0	0	0
1kor_kat.txt	1moz_kat.txt	1	0	3
1kor_kat.txt	2kor_kat.txt	2	0	4
1kor_kat.txt	2moz20_karoli.txt	0	0	0
1kor_kat.txt	2moz20_kat.txt	0		0
1kor_kat.txt	2moz20_ref.txt	0	0	0
1kor13_karoli.txt	1kor_kat.txt	5	1	9
1kor13_karoli.txt	1kor13_kat.txt	4	1	8
1kor13_karoli.txt	1kor13_ref.txt	16	12	18
1kor13_karoli.txt	1moz_kat.txt	4	5	5
1kor13_karoli.txt	2kor_kat.txt	2		5
1kor13_karoli.txt	2moz20_karoli.txt	2		1
1kor13_karoli.txt	2moz20_kat.txt	1		
1kor13_karoli.txt	2moz20_ref.txt	2		1
1kor13_kat.txt	1kor_kat.txt	91	89	93
1kor13_kat.txt	1kor13_karoli.txt	5	2	9
1kor13_kat.txt	1kor13_ref.txt	5	2	14
1kor13_kat.txt	1moz_kat.txt	1		9
1kor13_kat.txt	2kor_kat.txt	2		9
1kor13_kat.txt	2moz20_karoli.txt	1		1
1kor13_kat.txt	2moz20_kat.txt			3
1kor13_kat.txt	2moz20_ref.txt	1		1
1kor13_ref.txt	1kor_kat.txt	5	1	17
1kor13_ref.txt	1kor13_karoli.txt	18	13	20
1kor13_ref.txt	1kor13_kat.txt	5	1	14
1kor13_ref.txt	1moz_kat.txt	3	3	11
1kor13_ref.txt	2kor_kat.txt	1		9
1kor13_ref.txt	2moz20_karoli.txt	1		3
1kor13_ref.txt	2moz20_kat.txt			3
1kor13_ref.txt	2moz20_ref.txt	1		
1moz_kat.txt	1kor_kat.txt	1	0	3
1moz_kat.txt	1kor13_karoli.txt	0	0	0
1moz_kat.txt	1kor13_kat.txt	0		0
1moz_kat.txt	1kor13_ref.txt	0	0	0
1moz_kat.txt	2kor_kat.txt	0	0	2
1moz_kat.txt	2moz20_karoli.txt	0	0	0
1moz_kat.txt	2moz20_kat.txt	0	0	0
1moz_kat.txt	2moz20_ref.txt	0	0	0
2kor_kat.txt	1kor_kat.txt	3	1	6
2kor_kat.txt	1kor13_karoli.txt	0		0
2kor_kat.txt	1kor13_kat.txt	0		0
2kor_kat.txt	1kor13_ref.txt	0		0
2kor_kat.txt	1moz_kat.txt	1	0	3
2kor_kat.txt	2moz20_karoli.txt	0	0	0
2kor_kat.txt	2moz20_kat.txt	0	0	0
2kor_kat.txt	2moz20_ref.txt	0		0
2moz20_karoli.txt	1kor_kat.txt	3	3	7
2moz20_karoli.txt	1kor13_karoli.txt	1		0
2moz20_karoli.txt	1kor13_kat.txt	0		0
2moz20_karoli.txt	1kor13_ref.txt	0		1
2moz20_karoli.txt	1moz_kat.txt	4	3	7
2moz20_karoli.txt	2kor_kat.txt	2	1	5
2moz20_karoli.txt	2moz20_kat.txt	1		1
2moz20_karoli.txt	2moz20_ref.txt	13	8	16
2moz20_kat.txt	1kor_kat.txt	1		5
2moz20_kat.txt	1kor13_karoli.txt	1		
2moz20_kat.txt	1kor13_kat.txt			2
2moz20_kat.txt	1kor13_ref.txt			
2moz20_kat.txt	1moz_kat.txt	3	3	3
2moz20_kat.txt	2kor_kat.txt	2	1	2
2moz20_kat.txt	2moz20_karoli.txt	2		2
2moz20_kat.txt	2moz20_ref.txt	3	1	2
2moz20_ref.txt	1kor_kat.txt	6	2	5
2moz20_ref.txt	1kor13_karoli.txt	1		1
2moz20_ref.txt	1kor13_kat.txt	0		1
2moz20_ref.txt	1kor13_ref.txt	0		2
2moz20_ref.txt	1moz_kat.txt	6	1	8
2moz20_ref.txt	2kor_kat.txt	3		4
2moz20_ref.txt	2moz20_karoli.txt	16	11	19
2moz20_ref.txt	2moz20_kat.txt	2	1	2

9.3. Szövegtár: gépi vs. kézi fordítás

9.3.1. Eredeti angol nyelvű Wikipédia szócikk: Johann Haller

Johann Haller or Jan Haller (1463–1525) is considered one of the first commercial printers in Poland.

Born in Rothenburg, Haller is perhaps best known for publishing in 1509 a volume of poems by Theophylact Simocatta which had been translated from Byzantine Greek by Nicolaus Copernicus. At the time there was no printing press in Copernicus' area—Lidzbark (Heilsberg), Frombork (Frauenburg), Toruń (Thorn)[3]—therefore Copernicus' translation could have been printed only in Breslau (Wrocław), Kraków or farther afield. Copernicus, who had studied in Kraków, opted for Johann Haller, who together with Kasper Hochfeld had already published the first illustrated work in Poland, Jan Łaski's Statutes (1506),[4] and one of 25 works by Laurentius Corvinus (1508). Corvinus had lectured at the Kraków Academy while Copernicus studied there, and they were well acquainted. Corvinus took a job at Thorn, but in June 1509 traveled to the printer Haller in Kraków, bringing with him the manuscript entrusted to him by Copernicus. Corvinus (Rabe) added a poem, and Copernicus wrote a dedication to his uncle, Prince-Bishop of Warmia Lucas Watzenrode. Haller published the book before the end of 1509. Its cover featured the arms of Poland, Lithuania and Kraków.

After his studies at the Kraków Academy, Haller had become a merchant in wine, copper and tin, thus enabling himself to engage, at a later time, in the production of printing elements and finally establishing a printing press in Kraków. His first printing products were almanacs, followed by a breviary for the clergy. Haller acquired a partial monopoly on them, thereby protecting himself from competition. He soon expanded his business to include scientific and scholarly books inn astronomy, mathematics, philosophy and law, as well as royal and church statutes.[3]

Altogether Haller produced 3,530 prints. His masterpieces are illustrated books containing 354 sheets of woodcuts.

9.3.2. Kézi fordítás magyarra: Johann Haller

Johann Haller, vagy Jan Haller, (1463-1525) az, akit az első hivatásos nyomdásznak tartanak Lengyelországban.

Haller Rothenburgban született, és talán arról a legismertebb, hogy 1509-ben kiadta Theophylact Simocatta verseskötetét, melyet bizánci görögről Nicolaus Copernicus fordított le. Abban az időben nem volt nyomda ott, ahol Copernicus élt – Lidzbarkban (Heilsberg), Fromborkban (Frauenburg), és Torunban (Thorn) – így Copernicus fordítása csak Breslauban (Wroclaw), Krakkóban vagy még messzebb kerülhetett nyomtatásra. Copernicus, aki korábban Krakkóban tanult, Johann Hallert választotta, aki Kasper Hochfeld segédletével már kiadta Lengyelországban az első illusztrált művet, Jan Laski "Törvények" c. művét (1506), valamint Laurentius Corvinus 25 művének egyikét (1508). Corvinus előadott a Krakkói Akadémián mikor Copernicus ott tanult, ezáltal jól ismerték egymást. Corvinus Thornban vállalt munkát, de 1509 júniusában Krakkóba utazott Haller nyomdájához, magával hozva a kéziratot, melyet

Copernicus rábízott. Corvinus (Rabe) hozzáírt egy verset, és Copernicus ajánlást írt nagybátyjához, Lucas Watzenrode-hoz, Warmia hercegi érsekéhez. Haller még 1509-ben kiadta a könyvet. Borítóján Lengyelország, Litvánia és Krakkó címere szerepelt.

Miután befejezte tanulmányait a Krakkói Akadémián, Haller bor-, réz- és önkereskedő lett, ezáltal biztosítva magának azt, hogy később a nyomtatással foglalkozhasson és végül Krakkóban nyomdát alapíthasson. Első nyomtatott kiadványai évkönyvek (almanak) voltak, melyeket a papság számára készített breviáriumok követtek. Haller részleges monopóliumra tett szert ennek kapcsán, így védve magát a konkurrenciától. Hamarosan bővítette tevékenységét és tudományos könyveket is nyomtatott asztronómia, matematika, filozófia és jog témákban, akárcsak királyi és egyházi rendeleteket. Haller összesen 3,530 nyomtatást készített. Mesterművei illusztrálással díszített könyvek melyek 354 oldalnyi fametszetet tartalmaznak.

Virág Éva fordítása

9.3.3. Gépi fordítás magyarra: Johann Haller

Johann Haller és Jan Haller (1463-1525) tartják az egyik első kereskedelmi nyomtatók Lengyelországban.

Született Rothenburg, Haller talán legismertebb megjelentetésére 1509-ben egy kötet verset Theophylact Simocatta volt, amely fordított a bizánci görög Nicolaus Copernicus. Abban az időben nem volt nyomda Kopernikusz terület-Lidzbark (Heilsberg), Frombork (Frauenburg), Toruń (Thorn) [3]-tehát Kopernikusz fordítás lehetett volna kinyomtatni csak Breslau (Wrocław), Kraków vagy távolabbi mezőn . Kopernikusz, aki tanult Krakkóban, választott Johann Haller, aki együtt Kasper Hochfeld már kiadta az első illusztrált munkát Lengyelországban, Jan Laski alapszabálya (1506), [4] és egy 25 művei Laurentius Corvinus (1508). Corvinus volt, előadásokat tartott a Krakkói Akadémia, míg Kopernikusz tanult ott, és ők jól ismerik. Corvinus állást a Thorn, de június 1509 utaztak a nyomtató Haller Krakkóban, és magával hozta a kéziratot bízott rá Kopernikusz. Corvinus (Rabe) vett fel egy verset, és Kopernikusz írt egy elkötelezettség, hogy nagybátyja, Prince-püspöke Warmia Lucas Watzenrode. Haller megjelent a könyv vége előtt 1509. A fedél szerepelt a karok Lengyelország, Litvánia és Krakkó.

Miután tanulmányait a krakkói Akadémián, Haller vált kereskedő bor, réz, ón, ezáltal lehetővé téve magát, hogy vegyenek részt, egy későbbi időpontban, a termelési nyomtatás elemek és végül létrehozásáról nyomda Kraków. Az első nyomdai termékek évkönyvek, majd egy papi zsolozsmakönyv a papság. Haller szerzett részleges monopólium őket, megvédve ezáltal magát a versenyt. Hamarosan bővítette üzleti tartalmazza a tudományos és tudományos könyvek fogadóban csillagászat, a matematika, a filozófia és a jog, valamint a királyi és egyházi törvények [3].

Összesen Haller előállított nyomatok 3530. Ő remekművek is illusztrált könyvek tartalmazó 354 darab fametszet.

9.3.4. Magyar kézi fordítás visszafordítása géppel: Johann Haller

Johann Haller, Haller and Jan, (1463-1525), the one of the first professional typography held in Poland.

Haller was born in Rothenburg, and, perhaps best known for sure that in 1509 he published volumes of poetry Theophylact Simocatta by Nicolaus Copernicus translated from Greek Byzantium. At that time there was no printing, where Copernicus lived - Lidzbarkban (Heilsberg) Fromborkban (Frauenburg) and Torun (Thorn) - Translation of Copernicus so only Breslau (Wroclaw) and Krakow, or even further could be printed. Copernicus, who had studied in Krakow, Johann Haller chose Hochfeld Kasper, who has been granted with the help of Poland, the first illustrated work, Jan Laski, "Laws" in work (1506), and 25 works of one of Laurentius Corvinus (1508). Corvinus in the Cracow Academy when Copernicus studied there, so were familiar with each other. Corvinus Thorne worked, but in June 1509 he traveled to Krakow Haller printing house, bringing the manuscript, which was entrusted to Copernicus. Corvinus (Rabe), add in a poem written recommendation to Copernicus and his uncle, Lucas Watzenrode to, Prince Archbishop of Warmia. Haller has published a book in 1509. Cover of Poland, Lithuania and the coat of arms appeared in Cracow.

After completing studies at the Cracow Academy, Haller wine, copper and önkereskedő was, thus ensuring himself that later, deal with printing presses based in Krakow and then be available. First printed yearbooks (apple) were prepared for the priesthood, which brevériumok followed. Haller has gained a partial monopoly in this connection that protects itself konkurrenciától. Soon expanded its activities and scientific books printed in astronomy, mathematics, philosophy and law issues, as well as ecclesiastical and royal decrees. Haller made a total of 3.530 printing. Masterpieces illusztrálással decorated with books that contain 354 pages of woodcut.

9.3.5. Eredeti angol nyelvű Wikipédia szócikk: London Underground

The London Underground (also known as the Tube or The Underground) is a rapid transit system serving a large part of Greater London and some parts of Buckinghamshire, Essex and Hertfordshire in England. It is the oldest underground railway in the world, the first section of which opened in 1863 on what are now the Circle, Hammersmith & City and Metropolitan lines.[3] In 1890 it became the first to operate electric trains.[4] The whole network is commonly referred to by Londoners and in official publicity as the Tube,[5] although that term originally applied only to the deep-level bored lines, along which run trains of a smaller and more circular cross-section, to distinguish them from the sub-surface "cut and cover" lines that were built first.

The earlier lines of the present London Underground network were built by various private companies. They became part of an integrated transport system in 1933 when the London Passenger Transport Board (LPTB) or London Transport was created. The underground network became a single entity in 1985, when the UK Government created London Underground Limited (LUL).[6] Since 2003 LUL has been a wholly owned subsidiary of Transport for London (TfL), the statutory corporation responsible for most aspects of the transport system in Greater London, which is run by a board and a commissioner appointed by the Mayor of London.[7]

The Underground serves 270 stations and has 402 kilometres (250 mi) of track,[1] making it the second largest metro system in the world in terms of route miles after the Shanghai Metro.[8] It also has one of the largest numbers of stations. In 2007, more than one billion passenger journeys were recorded,[2] making it the third busiest metro system in Europe after Moscow and Paris. The tube is an international icon for London,

with the tube map, considered a design classic, having influenced many other transport maps worldwide. Although also shown on the Tube map, the Docklands Light Railway (DLR) and London Overground are not part of the LUL network.

Currently, about half the London Underground's costs are met from passenger fares collected and half from a grant by the Department for Transport.[9]

History

Railway construction in the United Kingdom began in the early 19th century, and six railway terminals had been built just outside the centre of London by 1854: London Bridge, Euston, Paddington, London King's Cross, Bishopsgate and Waterloo.[10] At this point, only Fenchurch Street station was located in the actual City of London. Traffic congestion in the city and the surrounding areas had increased significantly in this period, partly due to the need for rail travellers to complete their journeys into the city centre by road. The idea of building an underground railway to link the City of London with the mainline terminals had first been proposed in the 1830s, but it was not until the 1850s that the idea was taken seriously as a solution to traffic congestion

The first underground railways

In 1855 an Act of Parliament was passed approving the construction of an underground railway between Paddington Station and Farringdon Street via King's Cross which was to be called the Metropolitan Railway. The Great Western Railway (GWR) gave financial backing to the project when it was agreed that a junction would be built linking the underground railway with its mainline terminus at Paddington. GWR also agreed to design special trains for the new subterranean railway.

9.3.6. Kézi fordítás magyarra: London Underground

A Londoni Metró (amit “Csőnek” (the Tube) vagy “Földalattinak (the Underground) is hívnak) egy gyors közlekedési hálózat mely szinte egész Nagy-London területén, valamint Buckinghamshire, Essex és Hertfordshire egyes részei között biztosít összeköttetést. Ez a világ legöregebb földalatti vasútja, első szakaszát 1863-ban nyitották meg ott, ahol ma a Circle, Hammersmith&City és a Metropolitan vonalak futnak. 1890-ben ez a hálózat üzemeltetett először villanyvonatokat. A londoniak valamint a hivatalos kiadványok is az egész hálózatot Csőnek (the Tube) hívják, bár a kifejezés eredetileg csak azokra a mélyen a felszín alatt futó vonalakra vonatkozott ahol kisebb és kerekebb átmérőjű vonatok jártak, hogy megkülönböztessék őket az épp csak a föld alatt futó, először épült vonalaktól.

A jelenlegi Londoni Metró korábbi vonalait különböző magáncégek építették. 1933-ban ezek egységes közlekedési hálózattá váltak, amikor létrehozták a Londoni Utasszállító Bizottságot (London Passenger Transport Board – LPTB) vagy más néven Londoni Közlekedési Vállalatot (London Transport). A földalatti hálózat 1985-ben vált egy egységgé, mikor a brit kormány létrehozta a Londoni Közlekedés Korlátolt Felelősségű Társaságot (London Underground Limited – LUL). 2003 óta e társaság a Londoni Közlekedési Vállalat (Transport for London – TfL) tulajdonában lévő leányvállalat, az anyacég a Nagy-London területén lévő közlekedési hálózat legtöbb területéért felel, a céget pedig egy bizottság és a London polgármestere által kinevezett biztos vezet.

A Földalatti 270 állomást köt össze, pályájának hossza 402 kilométer (250 mérföld), vagyis pályahossz alapján a világ második legnagyobb metróhálózata a Sanghai Metró

után. Emellett egyike a legtöbb állomással rendelkező földalattiknak is. 2007-ben több mint egybillió utazást regisztráltak, ezzel Európában a harmadik legforgalmasabb metróhálózat Moszkva és Párizs után. A metró London nemzetközi jelképévé vált együtt a metrótérképpel, amelyet design-klasszikusnak tartanak, mivel világszerte nagyon sok közlekedési térkép készült mintájára. Bár a londoni metrótérkép mindkettőt jelöli, a Dockland Felszíni Vasút (Dockland Light Railway – DLR) és a Londoni Felszíni Vasút sem része a Londoni Közlekedési Vállalat hálózatának. Jelenleg a londoni metró költségeinek mintegy felét a viteldíj bevételek fedezik, másik felét a Közlekedési Minisztérium által adott támogatás biztosítja.

Története

A vasútépítés a 19. század elején indult Angliában, és 1854-re már 6 vasúti pályaudvar épült London középpontja köré: London Bridge, Euston, Paddington, London King's Cross, Bishopsgate és Waterloo. Ebben az időben csak a Fenchurch Street-i pályaudvar helyezkedett el az akkori London területén. A forgalmi torlódások a városban és közvetlen környékén egyre gyakoribbá váltak, részben annak köszönhetően, hogy a vonattal érkezőknek közúton kellett megtenniük az út hátralévő részét a városközpontig. Már az 1830-as években felmerült annak ötlete, hogy földalatti vasúttal kellene összekötni Londont a fő pályaudvarokkal, de csak az 1850-es években kezdték ezt az ötletet komolyan megfontolni, mint a forgalmi torlódások lehetséges megoldását.

Az első földalatti vasút

1855-ben a Parlament törvényben hagyta jóvá egy földalatti vasút megépítésének tervét, a Paddington Station-től a King's Crosson keresztül a Farringdon Streetig futó vonal a Metropolitan Vonal nevet kapta. A Nagy Nyugati Vasúttársaság (Great Western Railway – GWR) pénzügyileg támogatta a projektet miután megegyezés született, hogy egy csomópontot is építenek Paddingtonnál, amely összekapcsolja a földalatti vasutat a felszínen futó vasúti pályaudvarral. A GWR azt is vállalta, hogy speciális vonatokat tervez az új földalatti vasúthoz.

Virág Éva fordítása

9.3.7. Gépi fordítás magyarra: London Underground

A londoni metró (más néven a cső vagy a földalatti) gyors tranzit szolgáló rendszer nagy részét a Nagy-London és egyes részeit Buckinghamshire, Essex és a Hertfordshire-ben Angliában. Ez a legrégebbi földalatti vasút a világon, az első szakasza, amely 1863-ban megnyitott, amit most a Circle, Hammersmith & City és a Metropolitan vonalak [3]. 1890-ben lett az első működtetéséhez villamos motorvonatok. [4] A az egész hálózat nevükön a londoniak, és a hivatalos nyilvánosságot, mint a metrótól, [5], bár ezt a kifejezést eredetileg csak olyan mély szinten unatkozik vonalak, amelyek mentén futnak a vonatok a kisebb és kör keresztmetszetű, hogy megkülönböztessük őket A felszín alatti "cut és fedél" vonalakat építettek először.

A korábbi vonalak a jelenlegi londoni földalatti hálózat épült különböző magánvállalatok. Ezek részévé vált egy integrált közlekedési rendszer 1933-ban, amikor a londoni Passenger Transport Board (LPTB) vagy London Transport jött létre. A földalatti hálózat lett egyetlen szervezet 1985-ben, amikor az Egyesült Királyság kormánya létrehozott London Underground Limited (LUL) [6]. 2003-tól LUL már teljes tulajdonú leányvállalata, a Transport for London (TfL), a törvényes vállalat felelős a

legtöbb szempontjai A közlekedési rendszer a Greater London, amely működteti a fedélzeten és a biztos által kijelölt London polgármestere [7].

Az Underground szolgálja 270 állomás és 402 km (250 mi) a pálya, [1] ezzel a második legnagyobb metró rendszer a világon az útvonal mérföld után a Shanghai Metro. [8] is az egyik legnagyobb számban állomások. 2007-ben több mint egymilliárd utazások jegyeztek, [2] és ezzel a harmadik legforgalmasabb metróhálózat után Európa Moszkvában és Párizsban. A cső egy nemzetközi ikon a londoni, a cső térkép, úgy tervezési klasszikus, miután befolyásolja sok egyéb közlekedési térképek világszerte. Bár még látható a cső térképen, a Docklands Light Railway (DLR) és London Overground nem része a LUL hálózat.

Jelenleg körülbelül a fele a londoni metró költségei teljesülnek személygépkocsik viteldíjak gyűjtött, másik fele pedig a támogatást a Közlekedési Minisztérium [9].

Történelem

Vasútépítés az Egyesült Királyságban indult a 19. század elején, és hat vasúti terminál épült csak kívülről központja London 1854: London Bridge, Euston, Paddington, London Kings Cross, Bishopsgate és Waterloo. [10] Ezen a ponton csak Fenchurch Street állomás található, a tényleges City of London. Forgalmi torlódások a városban és a környező területeken is jelentősen nőtt ebben az időszakban, részben a szükséges vasúti utazók teljes útjuk a városközponttól közúton. Az az elképzelés, épület földalatti vasút összekapcsolása a City of London a fővonal terminál már előbb javasolt a 1830-as években, de ez nem volt egészen a 1850-es években, hogy az ötlet az volt, komolyan veszik, mint a megoldás a forgalmi dugók

Az első földalatti vasút

1855-ben törvénnyel fogadtak el jóváhagyásáról építése egy földalatti vasút között Paddington pályaudvar és Farringdon Street keresztül Kings Cross volt, hogy az úgynevezett Metropolitan Railway. A Great Western Railway (GWR) nyújtott pénzügyi támogatást nyújtsanak a projekt során megállapodás született arról, hogy egy csomópont épülne összekötő földalatti vasút és a fővonal végállomása a Paddington. GWR is megegyeztek, hogy tervezési speciális vonatok az új földalatti vasút.

9.3.8. Magyar kézi fordítás visszafordítása géppel: London Underground

The London Underground (which is "bone" (the Tube) or "below the earth (the Underground) is called) is a rapid transit network, which is almost the whole area of Greater London and Buckinghamshire, Hertfordshire and Essex providing connections between different parts. This is the world's oldest underground railway, the first stage opened in 1863, where it is now the Circle, Hammersmith & City and Metropolitan lines run. In 1890, this network operated by the first electric trains. The Londoners and the official publications for the entire network of pipes (the Tube) is called, although the term was originally only for those beneath the surface running lines covered with smaller and rounder with a diameter of trains had to distinguish them from just under the ground running first line was built from.

The current London Underground was built in earlier lines of various private companies. In 1933, they became a single transport network, which created the London Passenger Committee (London Passenger Transport Board - LPTB), also known as

London Transport Company (London Transport). The subway system in 1985, became a unit, when the British government set up the London Transport Company Limited (London Underground Limited - LUL). Since 2003 this company on the London Transport Company (Transport for London - TfL) owned subsidiary, the parent company of the Greater London area in the transport network for the area most responsible for the company by a committee appointed by the Mayor of London and certain leads.

The Underground station 270 connects the path length of 402 kilometers (250 miles) length of track that is the world's second largest underground network after the Shanghai Metro. In addition, one of the most földalattiknak station as well. In 2007, more than a billion trips were registered, with Europe's busiest subway system in the third after Moscow and Paris. The London Underground has become an international symbol along with the subway map, which is held in a classic design, because a lot of traffic around the world map made of the model. Although both London Underground map represents the surface Docklands Railway (Docklands Light Railway - DLR) and London Rail Surface Transport for London is not part of the network. Currently, about half the cost of the London Underground fare revenue covers the other half of the support provided by the Ministry of Transport.

History

The railway construction in the 19th century began in England, and from 1854 has six railway stations built around the center of London: London Bridge, Euston, London King's Cross, Waterloo and Bishopsgate. At this time, only the Fenchurch Street railway station was located at that time in London. The traffic congestion in the city and its immediate surroundings become increasingly common, partly due to the road for those arriving by train had to go in the way the rest of the city center. Already in the 1830s raised the idea of an underground railway to be linked to the main London railway stations, but only in the 1850s began to seriously consider this idea as a possible solution to traffic congestion.

The first underground railway

In 1855, the Parliament Act was approved plan of constructing an underground railway to Paddington Station from King's Cross to Farringdon Street, a line running through the Metropolitan Line was named. The Great Western Railway Company (Great Western Railway - GWR) is financially supported the project after it was agreed that a node can build Paddingtonnál that connects the underground railway station, on the surface running rail. The GWR also undertook to train a special design of the new underground railway.

9.3.9. Eredeti angol nyelvű Wikipédia szócikk: Mozartkugel

The Mozartkugel (English: Mozart ball), originally known as the “Mozartbonbon”, was created by the Salzburg confectioner, Paul Fürst, in 1890 and named after Wolfgang Amadeus Mozart.

The confectionery Fürst still produces the original Salzburg Mozartkugeln by hand according to the original recipe and only sells them in its shops or over its website. As the Fürst confectionery does not own a trademark for Mozartkugeln, there are numerous imitation products, most of which are produced using industrial techniques.

The original

The master confectioner, Paul Fürst, came to Salzburg in 1884 and opened his own shop at number 13, Brodgasse. He presented the Mozartbonbon for the first time in 1890, later producing and selling it in greater quantities as Mozartkugeln. Fürst's achievement was the production of a perfectly rounded chocolate, with no flat areas. The production process used by the confectionery Fürst has not changed to this day.

Paul Fürst presented the Mozartkugel at a fair in Paris in 1905 and was awarded a gold medal for it.

Today, the confectionery Fürst sells the original Salzburg Mozartkugeln exclusively in its four shops in Salzburg (at the Old Market, with branches in the Ritzerbogen, the Getreidegasse and near the Castle Mirabell), and via a direct service, but not in other shops. Mozartkugeln can be bought from the confectionery Fürst individually and in packages of several pieces.

Original recipe

The "Original Salzburg Mozartkugeln" are still produced manually by the confectionery Fürst according to the original recipe and using the original technique: First, a ball of green pistachio marzipan covered in a layer of nougat is produced. This ball is then placed on a small wooden stick and dunked in a dark chocolate coating. Next, the stick is placed vertically, with the ball at the top, on a platform to allow the chocolate to cool off and harden. Finally, the stick is removed; the hole that it leaves behind is filled with chocolate coating, and the ball is wrapped in blue-silver tin foil by hand. According to the Fürst company, their employees produce approximately 1.4 million Mozartkugeln by hand using this technique every year. In the firm's air-conditioned salerooms, the balls remain fresh for about eight weeks.

Prizes

The specialist magazine, Der Feinschmecker (English: The Gastronome), gave the original Salzburg Mozartkugel first place in a comparison test of different Mozartkugeln in its January 2006 edition. It was remarked that the original Salzburg Mozartkugel is handmade and that it has a nougat taste with a note of slightly bitter pistachio marzipan. The original Salzburg Mozartkugel was awarded a gold medal at the second international truffle competition during the confectionery fair ÖKONDA in Wels in September 2005.

Rights to the name

The existence of numerous imitation Mozartkugeln finally led Paul Fürst's descendants to initiate a court process. At stake were the rights to the name, not the Mozartkugeln recipe itself. At first, the dispute concerned only confectionery producers in Salzburg, but later spread to include the competition from Germany. The result was an agreement which obliged Fürst's competitors to use other names. The Mirabell firm, based in Grödig near Salzburg, chose the name, "Real Salzburg Mozartkugeln". The Bavarian producer, Reber, opted for "Real Reber Mozartkugeln". In 1996, a dispute between Fürst and a subsidiary of the Swiss food producer, Nestlé, which wanted to market "Original Austria Mozartkugeln", was decided in the third instance. Only Fürst products may be called original Salzburg Mozartkugeln.

9.3.10. Kézi fordítás magyarra: Mozartkugel

A Mozart-golyót, eredeti nevén “Mozart-bonbont”, 1890-ben Paul Fürst salzburgi cukrász készítette és Wolfgang Amadeus Mozartról nevezte el.

A Fürst cukrászda a mai napig kézzel, az eredeti recept alapján készíti a salzburgi Mozartgolyót, és csak saját üzleteiben vagy weboldalán keresztül árusítja azt. Mivel azonban a Fürst cukrászda nem birtokolja a Mozartgolyó márkanevet, számos utánzat készül, legtöbbjük ipari technológiával és üzemi gyártás keretében.

Az eredeti

Paul Fürst mestercukrász 1884-ben költözött Salzburgba és a Brodgasse 13. alatt megnyitotta saját üzletét. Először 1890-ben kínálta a Mozartbonbont, később nagyobb mennyiségben készítette és árulta ezt az édességet Mozartgolyó néven. Fürst újítása az volt, hogy egy teljesen gömbölyű csokoládét tudott készíteni, melynek nem voltak egyenes felületei. A Fürst cukrászat által alkalmazott gyártási folyamat a mai napig nem változott. 1905-ben Paul Fürst Párizsban is bemutatta a Mozartgolyót, mellyel aranyérmet nyert.

Ma a cukrászat az eredeti salzburgi Mozartgolyót kizárólag saját négy salzburgi boltjában (a Régi Piactéren, Ritzerbogenben, a Getreidegasse-n és a Mirabell Kastély mellett), valamint egy direkt weboldalon keresztüli szolgáltatáson keresztül árusítja, de egyetlen más üzletben sem. A Mozartgolyó a Fürst cukrászdában egyenként és néhány darabos kiszerelésekben is megvásárolható.

Az eredeti recept

Az “Eredeti Salzburgi Mozartgolyót” a Fürst cukrászat a mai napig kézzel, az eredeti recept és elkészítési eljárás szerint készíti: először egy zöld pisztáciamarcipán golyót készítenek amit egy réteg nugátkrém fed. Ezt a golyót aztán egy kis fapálcikára tűzik és étcsokoládéba mártják. Ezután a pálcikát, a marcipángolyóval a tetején függőlegesen egy állványra helyezik, hogy a csokoládé kihűljön és megszilárduljon. Legvégül a pálcikát kihúzzák, a lyuk amely utána marad a csokoládébevonattal töltödik fel, ezután az édességet kék-ezüst fóliába csomagolják, szintén kézzel. A Fürst cég szerint alkalmazottaik évente körülbelül 1,4 millió Mozartgolyót készítenek a fenti technikával. A cég légkondicionált árusítóhelyén a golyók mintegy 8 hétig frissek maradnak.

Díjak

A Der Feinschmecker (kb. Ízmester) című szaklap az eredeti salzburgi Mozartgolyónak ítélte az első helyet a 2006 januári számában közölt Mozartgolyók összehasonlító tesztjében. Kiemelték, hogy az eredeti salzburgi Mozartgolyó kézzel készül és a nugátkrém ízében felfedezhető az enyhén kesernyés pisztáciamarcipán is. Emellett 2005 szeptemberében a Wels-ben (város Felső-Ausztriában) zajló ÖKONDA cukrászati kiállításon aranyérmet nyert az ott megrendezett második nemzetközi trüffelversenyen is.

Jogok a névre

A Mozartgolyó egyre több utánzatának feltűnése végül arra késztette Paul Fürst utódait hogy jogi eljárást indítsanak. A névhez való jog forgott kockán, nem maga a Mozartgolyó receptje. Kezdetben a vita csak a Salzburgban élő cukrászokat érintette, de később már a Németországból érkező konkurrensekre is kiterjedt. A jogi eljárás eredményeként megszületett a határozat, mely a Fürst család vetélytársait más név

használatára kötelezte. A Salzburg melletti Grödig-ben bejegyzett Mirabell cég az "Igazi Mozartgolyó" nevet, a cseh Reber cég pedig az "Igazi Rebel Mozartgolyó" nevet választotta. 1996-ban egy per a Fürst cég és a svájci élelmiszergyártó Nestlé egyik leányvállalata között, mely "Eredeti Mozartgolyó" néven kívánta forgalmazni a terméket, azzal zárult hogy csak a Fürst termékek használhatják az "Eredeti Salzburgi Mozartgolyó" nevet.

Virág Éva fordítása

9.3.11. Gépi fordítás magyarra: Mozartkugel

A Mozartkugel (angolul: Mozart golyó), eredeti nevén a "Mozartbonbon", hozta létre a salzburgi cukrász, Paul Fürst, 1890-ben és névadója Wolfgang Amadeus Mozart.

A cukrász Fürst mai napig az eredeti salzburgi Mozartkugeln kézzel szerint az eredeti recept, és csak adja el őket saját üzletekben, vagy több mint a honlapján. Mivel a Fürst édesipari nem saját védjegyet a Mozartkugeln, számos utánzat termékeket, amelyek többsége felhasználásával az ipari technikák.

Az eredeti

A mester cukrász, Paul Fürst, eljött Salzburg 1884-ben, és megnyitotta saját bolt száma 13, Brodgasse. Ő mutatta be Mozartbonbon először 1890-ben, később gyártó és forgalmazó is nagyobb mennyiségben Mozartkugeln. Fürst az eredmény volt a termelés az egy tökéletesen kerek csokoládé, nem sík területek. A termelési folyamat által használt édesipari Fürst nem változott ezen a napon.

Paul Fürst bemutatta a Mozartkugel egy kiállításon Párizsban 1905-ben aranyérmet szereztek rá.

Ma a cukrászda Fürst eladja az eredeti salzburgi Mozartkugeln kizárólag a négy üzletek Salzburg (az Old Market, fiókteleppel a Ritzerbogen, a Getreidegasse és közel a Vár Mirabell), illetve egy közvetlen járat, de más üzletekben. Mozartkugeln lehet vásárolni a cukrász Fürst egyénileg és csomagok több darabból.

Eredeti recept

Az "Eredeti Salzburg Mozartkugeln" a mai napig kézzel a cukrász Fürst szerint az eredeti recept, és az eredeti technika: Először is, a labdát a zöld pisztácia marcipán szabályozni egy réteg nugát keletkezik. Ez a labdát, majd helyezni egy kis fából készült botot, és dunked egy sötét csokoládé bevonattal. Ezután a botot is függőlegesen, a labda a felső, a platform lehetővé teszi a csokoládét, hogy lehűtse magát, és megkeményedik. Végül, a botot is el kell távolítani a lyuk, hogy hagy maga után tele van csokoládé bevonattal, és a labda csomagolva kék-ezüst alufólia kézzel. Szerint a Fürst cég, az alkalmazottak termelnek mintegy 1.400.000 Mozartkugeln kézzel, ezzel a technikával minden évben. A cég légkondicionált salerooms, a golyók maradnak friss mintegy nyolc héten át.

Díjak

A szakmai magazin, a Der Feinschmecker (angol: A gasztronómus), adta az eredeti Salzburg Mozartkugel első helyen egy összehasonlító tesztet a különböző Mozartkugeln a 2006 januárjában kiadás. Azt megjegyezte, hogy az eredeti Salzburg Mozartkugel a kézzel készített, és hogy egy nugát íz és egy jegyzet a kesernyés pisztácia marcipán. Az

eredeti salzburgi Mozartkugel nyert aranyérmet a második nemzetközi szarvasgomba verseny során a cukrászda tisztességes ÖKONDA Wels 2005 szeptemberében.

Jogokat a név

Megléte számos utánzat Mozartkugeln végül vezetett Paul Fürst leszármazottai kezdeményezni bírósági eljárást. A tét volt a jogokat a név, nem a Mozartkugeln recept is. Az első, a vita csak az érintett édesipari gyártó salzburgi, de később elterjedt, hogy tartalmazza a versenyre Németországból. Az eredmény olyan megállapodás, amely köteles Fürst versenytársainak, hogy más neveket. A Mirabell cég székhelye Grödig közelében Salzburg, úgy döntött, a név, "Real Salzburg Mozartkugeln". A bajor gyártó, Reber, választotta "Real Reber Mozartkugeln". 1996-ban közötti vita Fürst és leányvállalata a svájci élelmiszer-termelő, a Nestlé, amely szerette volna piacra "Eredeti Ausztria Mozartkugeln", úgy döntöttek, a harmadik esetben. Csak Fürst termékek nevezhető eredeti Salzburg Mozartkugeln.

9.3.12. Magyar kézi fordítás visszafordítása géppel: Mozartkugel

The Mozart balls, originally called "Mozart chocolates," in 1890, Salzburg confectioner Paul Fürst made and named by Wolfgang Amadeus Mozart.

The cake is still Fürst hand, the original recipe given by the Salzburg Mozartgolyót, and only through their own stores or websites selling it. However, as the confectionery Fürst Mozartgolyó not own the brand, made a number of imitations, most of them within the framework of factory production and industrial technology.

The original

Master confectioner Paul Fürst in 1884, he moved to Salzburg and the 13th Brodgasse During opened his own business. First offered in 1890, the Mozartbonbont, and later in larger quantities produced and sold under the name of this candy Mozartgolyó. Fürst-renewal was that chocolate could produce a completely round, which did not have a flat surface. The manufacturing process used by the confectionery Fürst has still not changed. In 1905, Paul Fürst in Paris presented the Mozartgolyót, which won the gold medal.

Today, the original pastry Mozartgolyót Salzburg, Salzburg, only four own store (in the Old Market Square, Ritzerbogenben the Getreidegasse Mirabell Palace and the n-on), as well as through the website via a direct selling, but not with any other business. The Mozartgolyó the confectionary Fürst and a few pieces per pack in also available.

The original recipe

The "Original Salzburger Mozartgolyót" the confectionery Fürst is still by hand, the original recipe and prepare according to the procedure of preparation: first a green ball is made by a layer pisztáciamarcipán nugátkrém Fed. This ball was placed and then a little fapálcikára étcsokoládéba icing. Then stick to the top of the marcipángolyóval placed vertically on a rack to cool and solidify the chocolate. Finally, the stick is pulled out of the hole that remains after the coating of chocolate is filled, then the candy blue and silver foil wrapped by hand also. Fürst The company said about 1.4 million employees annually Mozartgolyót made in the above technique. The company is selling air-conditioned place the balls remain fresh up to 8 weeks.

Awards

Der Feinschmecker (about Ízmester) was the original magazine Salzburg Mozartgolyónak took first place in the 2006 January issue published Mozartgolyók in comparative tests. They pointed out that the original hand-made in Salzburg Mozartgolyó nugátkrém and discover the taste is slightly bitter pisztáciamarcipán. In addition, in September 2005 in Wels (Upper Austria town) ÖKONDA confectionery exhibition taking place in gold medal at the Second International trüffelversenyen also held there.

Rights to the name

Mozartgolyó the emergence of an increasing number of imitations that eventually led to his descendants, Paul Fürst launch legal action. The name of the right at stake is not the recipe for Mozartgolyó. Initially, the only discussion of the Salzburg confectioner living in affected, but later also from Germany konkurrensekre extensive. The legal action resulted in the decision, which rivals other Fürst family is required to use. The Grödig near Salzburg Mirabell company registered in the "real Mozartgolyó 'name, the Czech company Reber was the" real Mozartgolyó Rebel "name is selected. In 1996, a lawsuit the company Fürst and Swiss food producer Nestle's subsidiary in which "Original Mozartgolyó" intended to be marketed under the name of the product, that only ended with the Fürst products you can use the "Original Salzburger Mozartgolyó" name.

9.4. Kézzel angolról magyarra fordított tesztkorpusz

Az alábbi angol Wikipédia idézeteket Virág Éva fordította magyarra.

University of Oxford

The University of Oxford (informally Oxford University or Oxford) is a university located in Oxford, United Kingdom. It is the second-oldest surviving university in the world and the oldest in the English-speaking world. Although its exact date of foundation is unclear, there is evidence of teaching as far back as 1096. The University grew rapidly from 1167 when Henry II banned English students from attending the University of Paris.

After disputes between students and Oxford townsfolk in 1209, some academics fled north-east to Cambridge, where they established what became the University of Cambridge.

http://en.wikipedia.org/wiki/University_of_Oxford

Az Oxford-i Egyetem (röviden csak "Oxford") egy állami egyetem Oxfordban, az Egyesült Királyságban. Ez a második legöregebb, ma is működő egyetem, és a legrégibb egyetem az angol-anyanyelvű világban. Bár alapításának pontos időpontja nem ismert, bizonyított, hogy már a 11. században folyt itt tanítás. Az Egyetem látványos növekedésnek indult 1167 után, amikor II. Henrik megtiltotta, hogy angol diákok a Párizs-i Egyetemre járjanak.

1209-ben, a diákok és az oxfordi városlakók közti viszály következtében néhány egyetemi tanár északkeletre, Cambridge-be menekült, ahol megalapították a ma Cambridge-ként ismert egyetemet.

Foreign relations of Pakistan

Pakistan is an active member of the United Nations. It was a member of the CENTO and SEATO military alliances. Its alliance with the United States was especially close after the Soviets invaded the neighboring country of Afghanistan. In 1964, Pakistan signed the Regional Cooperation for Development (RCD) Pact with Turkey and Iran, when all three countries were closely allied with the U.S., and as neighbors of the Soviet Union, wary of perceived Soviet expansionism. To this day, Pakistan has a close relationship with Turkey. RCD became defunct after the Iranian Revolution, and a Pakistani-Turkish initiative led to the founding of the Economic Cooperation Organisation (ECO) in 1985.

http://en.wikipedia.org/wiki/Foreign_relations_of_Pakistan

Pakisztán az Egyesült Nemzetek aktív tagja. Korábban tagja volt a CENTO (Központi Szerződés Szervezete) és a SEATO (Délkelet-Ázsiai Szerződés Szervezete) katonai szövetségeknek. Az Egyesült Államokkal, mint szövetségessel kapcsolata különösen szorossá vált miután a szovjetek lerohanták a szomszédos Afganisztánt. 1964-ben Pakisztán Törökországgal és Iránnal együtt aláírta a Térségi Együttműködés a Fejlődésért Paktumot (Regional Cooperation for Development Pact), melynek értelmében e három ország szoros köteléket vállal az Egyesült Államokkal, és mint a Szovjetunió szomszédállamai, igyekeznek megállítani a látható szovjet terjeszkedést. Pakisztán a mai napig szoros kapcsolatokat ápol Törökországgal. A fentebb hivatkozott Paktum az iráni felkelést követően hatályát vesztette, azonban egy pakisztáni-török kezdeményezés elvezetett az ECO (Gazdasági Együttműködési Szervezet) 1985-ös megalapításához.

Pete Seeger

Seeger was born in French Hospital, Midtown Manhattan. His parents were living with his grandparents in Patterson, New York, from 1918 to 1920. His father, Charles Louis Seeger Jr., was a composer and pioneering ethnomusicologist investigating both American folk and non-Western music. His mother, Constance de Clyver Edson, was a classical violinist and teacher. His parents divorced when Seeger was seven. His stepmother, Ruth Crawford Seeger, was one of the most significant female composers of the twentieth century. His eldest brother, Charles Seeger III, was a radio astronomer, and his next older brother, John Seeger, taught in the 1950s at the Dalton School in Manhattan and was the principal from 1960 to 1976 at Fieldston Lower School in the Bronx. His uncle, Alan Seeger, a noted poet, was killed during the First World War. His half-sister, Peggy Seeger, also a well-known folk performer, was married for many years to British folk singer Ewan MacColl. Half-brother Mike Seeger went on to form the New Lost City Ramblers, one of whose members, John Cohen, was married to Pete's other half-sister, singer Penny Seeger, also a highly talented singer.

http://en.wikipedia.org/wiki/Pete_Seeger

Pete Seeger Manhattan közepén, a Midtown-nak is hívott városrész francia kórházában született. Szülei 1918 és 1920 között a nagyszülőkkel együtt a New York állambeli Pattersonban éltek. Apja, ifj. Charles Louis Seeger, zeneszerző és zenetudós volt, aki az

elsők között vizsgálta mind az amerikai népzenét, mind a nem-európai gyökerekből fakadó zenét. Édesanyja, Constance de Clyver Edson klasszikus hegedűművész és tanár volt. A szülők Seeger hétéves korában elváltak. Nevelőanyja, Ruth Crawford Seeger egyike volt a huszadik század legkiemelkedőbb női zeneszerzőinek. Legidősebb bátyja, Charles Seeger III, rádiós csillagász volt, másik bátyja, John Seeger pedig az 1950-es években a manhattan-i Dalton School-ban tanított, majd 1960-tól 1976-ig a bronx-i Fieldston Lower School igazgatója lett. Nagybátyját, Alan Seegert, aki neves költő volt, megölték az első világháborúban. Féltestvére, Peggy Seeger, aki szintén ismert népzenei előadó volt, hosszú évekig Ewan MacColl brit folkénekessel élt házasságban. Másik féltestvére, Mike Seeger megalapította a New Lost City Ramblers-t (egy hagyományos zenei stílusban játszó vonósegyüttest), melynek egyik tagja, John Cohen Pete féltestvérét, az énekes Penny Seegert vette feleségül, és maga is igen tehetséges énekes volt.

Avidius Cassius

In 175 he was proclaimed Roman Emperor after the erroneous news of the death of Marcus Aurelius; the sources also indicate he was encouraged by Marcus's wife Faustina, who was concerned about her husband's ill health, believing him to be on the verge of death, and felt the need for Cassius to act as a protector in this event, since her son Commodus was still young (13). The evidence, including Marcus's own Meditations, supports the idea that Marcus was indeed quite ill, but by the time Marcus recovered, Cassius was already fully acclaimed by the Egyptian legions of II Traiana Fortis and XXII Deitoriana.

At first, according to Cassius Dio, Marcus, who was on campaign against tribes in the north, tried to keep the rebellion a secret from his soldiers, but after the news had spread among them, he addressed them. In this speech that Dio attributes to Marcus, he laments the disloyalty of "a dearest friend", while at the same time expressing his hope that Cassius would not be killed or commit suicide, so that he could show mercy. The Senate declared Cassius a public enemy.

It is known that Cassius was recognized as emperor by May 3, since a document of that date is recorded as being in the first year of Cassius's reign. The beginning of his rebellion have been in April 175.

http://en.wikipedia.org/wiki/Avidius_Cassius

175-ben Római Császárrá kiáltották ki, (az akkori császár) Marcus Aurelius elhamarkodott halálhírének következtében: a források szerint Marcus felesége, Faustina is segíthette ebben, aki nagyon aggódott férje rossz egészségi állapota miatt, azt gondolva, hogy már a halál szélén vergődik. Így szükségesnek érezte, hogy Cassius mint védelmezője lépjen fel ebben a helyzetben, mivel fia, Commodus még kiskorú, mindössze tizenhárom éves volt. A bizonyítékok, ideértve Marcus "Elmélkedések" című feljegyzéseit is, azt az elgondolást támasztják alá, mely szerint Marcus valóban nagyon beteg volt, de a Cassiushoz eljutó hírek ellenére mégsem halt meg. Bár Cassius értesült arról, hogy Marcus életben van, a lázadás folytatása mellett döntött.

Kezdetben, Cassius Dio írásaira támaszkodva, Marcus, aki éppen az északi törzsek ellen folytatott hadjáratot, próbálta a lázadás hírét eltitkolni katonái elől, de miután a hír

terjedni kezdett, ő maga mondta ezt el nekik. A beszédben, melyet Dio Marcusnak tulajdonít, a császár "egy legdrágább barát" hűtlenségét fájlalja, miközben azon reményét is kifejezi, hogy Cassiust nem ölik meg, és önkezével sem vet véget életének, hiszen így ő majd könyörületet mutathat. A Szenátus Cassiust kikiáltotta közellenségnek.

Bizonyított, hogy Cassiust május 3-án kiáltották császárrá, mivel egy ilyen keltezésű dokumentumot feljegyeztek, mint Cassius uralkodásának első évéből származó irat. Lázadását 175 áprilisában indította.

Meitetsu 5000 series (2008)

The interior has a light gray color scheme. As with the earlier 3150 series, individual seats are 470 mm wide. They are upholstered with a blue moquette. The seats use a cantilever structure, i.e. attached to the walls, and in each car there is one folding seat. The straps are 1,630 mm from the floor.

The number of priority seats, for the aged or infirm, was increased to 10 seats in each car. The priority seats use red moquette and orange straps and poles to hold on to to clarify the distinction between those seats and regular seats. Wheelchair spaces are located behind the driver's cab in both cars on each end of the train.

http://en.wikipedia.org/wiki/Meitetsu_5000_series_(2008)

A belső tér halványszürke színösszeállítású. Mint a korábbi 3150-es sorozatnál, az egyes ülések 470 mm szélesek, kék mokett (plüss-szerű szövet) kárpitozással. Az ülések konzol-struktúrával készültek, vagyis a falra erősítettek, és minden egyes kocsiban egy összecsukható ülés is található. Az ülésszíjak / övek a padlótól 1,630 mm magasságban találhatók.

Az időseknek, mozgássérülteknek fenntartott ún. megkülönböztetett ülések számát kocsinként 10-re emelték. Ezen ülések piros mokett borításúak, narancsszínű szíjakkal és kapaszkodókkal, hogy egyértelműen meg lehessen őket különböztetni a hagyományos ülésektől. Kerekesszékkel utazóknak fenntartott hely a vonat mindkét végén az utolsó kocsiban, a vezetőfülke mögött található.

Canberra 400

The Canberra 400, also known as both the GMC 400, Stegbar Canberra 400 and in its infancy, National Capital 100 was a V8 Supercar race run on the streets of Australia's capital, Canberra.

Unfortunately, the Canberra 400 only lasted 3 of its 5 year contract. Gary Humphries's Liberal Government was replaced in 2002 by Jon Stanhope's Labor Government. The new Chief Minister allowed the race to run in 2002, but decided to pull the plug for the 2003 race. The main reason given for the cancellation of the contract was the amount of money being spent on the race. Kate Carnell's initial estimation on cost blew out as the years went on, and some Canberrans believed that this money was better spent elsewhere. The race wasn't making as much money as had been expected. The motels and hotels around Canberra were full and having the best winter period ever, but the

crowd at the track dropped from 101,000 in 2000 to 89,000 in 2002. This was put down the time of year and the weather. In Canberra, in June the temperature can become lower than 5°C during the day. To most Canberrans this is just a normal winter, but to interstate visitors who weren't used to such cold conditions, it was too cold.

http://en.wikipedia.org/wiki/Canberra_400

A Canberra 400 egy V8 Supercar verseny volt, mely az ausztrál főváros, Canberra utcáin zajlott.

A verseny igen rövid életű volt, csupán 2000 és 2002 között került megrendezésre a Királynő születésnapjának hétvégéjén, júniusban.

Így sajnos a Canberra 400 ötéves szerződéséből mindössze 3 év valósult meg. Gary Humphries liberális kormányát 2001-ban Jon Stanhope munkáspártja váltotta. Az új miniszterelnök engedélyezte a versenyt 2002-ben, de úgy döntött, hogy a 2003-as versenyhez már nem ad támogatást. A szerződés érvénytelenítésének fő oka ugyanis a verseny költsége volt. Kate Carnell kezdeti költségbecslései az évek múlásával egyre inkább emelkedtek, és sok canberra-i gondolta úgy, hogy ezt a pénzt sokkal jobban is el lehetne költeni. A verseny emellett nem hozott annyi hasznot, mint ahogy várták. Míg a Canberra környéki motelek és szállodák megteltek és a legjobb téli szezonjaikat zárhatták, addig a versenyre érkező látogatók száma 2002-ben már csak 89,000 volt a 2000-ben ide érkező 101,000 látogatóval szemben. Ezt a csökkenést az évszaknak és az időjárásnak tulajdonították. Canberrában a júniusi hőmérséklet napközben akár az 5°C-t sem haladja meg. A legtöbb Canberrában élő embernek ez ugyan a normális tél, de az ország belsejéből érkező látogatók ilyen körülményekhez nem voltak hozzászokva, nekik egész egyszerűen túl hideg volt.

Upper 10

Upper 10 is a caffeinated lemon-lime soft drink, similar to Sprite, 7 Up, Sierra Mist, and Bubble Up. It was bottled by RC Cola.

The Upper 10 brand debuted in 1933 as a product of the Nehi Corporation (later Royal Crown Corporation). Upper 10 was one of RC Cola's flagship brands throughout the company's history. However, with the acquisition of RC Cola by Cadbury Schweppes plc in 2000 and subsequent folding of company operations into Dr Pepper/Seven Up, Inc., bottlers have gradually discontinued bottling Upper 10 in favor of the similar, more popular and non-caffeinated 7 Up (which is also owned by Dr Pepper Snapple Group).

Upper 10 is still sold outside of North America by Cott Beverages, the same company that sells RC Cola internationally

http://en.wikipedia.org/wiki/Upper_10

Az Upper 10 egy koffeintartalmú, a Sprite-hoz és 7UP-hoz hasonló, citrom-zöldcitrom-ízű üdítőital, melyet az RC Cola palackozott.

Az Upper 10 márka 1933-ban került piacra, mint a Nehi Corporation (később Royal Crown Corporation) terméke. Az Upper 10 az RC Cola cég történetének egyik

zászlóshajója lett. Azonban az RC Cola 2000-ben történt, a Cadbury Schweppes plc általi felvásárlását követően a cég inkább a Dr Pepper/Seven Up Inc. cégbe olvasztotta tevékenységét, mely együttjárt azzal, hogy az Upper 10 helyett az ahhoz hasonló, de népszerűbb és koffeinmentes 7 UP palackozása és forgalmazása került előtérbe. (a 7 Up-ot szintén a Dr Pepper cégcsoport tulajdonolja)

Az Upper 10-et azonban Észak-Amerikán kívül a Cott Beverages, az RC Cola termékeinek nemzetközi forgalmazója továbbra is árusítja.

Munich Philharmonic

The orchestra was founded in Munich in 1893 by Franz Kaim, son of a piano manufacturer, as the Kaim Orchestra. In 1895, it took up residence in the city's Tonhalle (concert hall). It soon attracted distinguished conductors: Gustav Mahler first directed the group in 1897 and premiered his Symphony No. 4 and Symphony No. 8 with the orchestra, while Bruno Walter directed the orchestra for the posthumous premiere of Mahler's Das Lied von der Erde. Felix Weingartner was music director from 1898 to 1905, and the young Wilhelm Furtwängler made his auspicious conducting debut there in 1906. Meanwhile Anton Bruckner pupil Ferdinand Löwe established an enduring tradition of Bruckner performance which continues to this day.

Throughout this time the orchestra, which by 1910 was known as the Munich Konzertverein Orchestra, was privately funded, but during World War I finances became tight and players were called for military service, forcing the orchestra to cease operation. After the war, the orchestra was taken over by the city of Munich and restarted under the leadership of composer Hans Pfitzner, soon replaced by Bruckner pioneer Siegmund von Hausegger. In 1928, the orchestra acquired its current name.

After the rise of the Nazi party in 1933, the orchestra stamped its scores with swastikas and the words "The Orchestra of the Fascist Movement". (The swastikas weren’t removed until the early 1990s.)

http://en.wikipedia.org/wiki/Munich_Philharmonic

A zenekart 1893-ban Franz Kaim, egy zongorakészítő fia alapította Kaim Zenekar néven. 1895-ben a zenekar a városi koncertterembe költözött, és hamarosan olyan kiváló karmestereket sikerült megnyernie mint Gustav Mahler, aki 1897-ben vezényelt itt először és a zenekarral mutatta be Negyedik és Nyolcadik Szimfoniáját, majd Bruno Walter, aki Mahler "A Föld dala" című művének premierjét vezényelte a zeneszerző halálát követően. 1898 és 1905 között Felix Weingartner volt a zenei igazgató, 1906-ban pedig a fiatal Wilhelm Furtwangler debütált sikeresen mint karmester. Mindeközben Anton Bruckner tanítványa, Ferdinand Löwe elindította azt a ma is élő hagyományt, hogy a zenekar Bruckner műveket mutat be.

Ezen idő alatt a zenekar, mely 1910-ben már Müncheni Koncertzenekar néven volt ismert, magántőkéből tartotta fent magát. Az első világháború pénzügyi nehézségei miatt, és mivel a zenészeket katonai szolgálatra hívták be, a zenekar működésének felfüggesztésére kényszerült. A háború után München városa vette át a zenekart és újraindította azt a zeneszerző Hans Pfitzner vezetésével, akit hamarosan a Bruckner nyomdokain járó Siegmund von Hausegger követett. 1928-ban a zenekar megkapta mai nevét, mint a Müncheni Filharmonikusok.

A náci párt 1933-as hatalomra kerülésével, a zenekar kottáit horogkeresztes pecséttel látta el és e szavakkal: "a fasiszta mozgalom zenekara". (A horogkeresztet az 1990-es évek elejéig nem távolították el).

Castle Rock (Pineville, West Virginia)

Castle Rock is a geological feature located in Pineville, West Virginia next to the Pineville Public Library. Named for its resemblance to a castle, it rises about 200 feet above Rock Castle Creek, a branch of the Guyandotte River. Its base is estimated to be 100 feet in diameter. Midway up there is a stone terrace, with a narrower shale formation rising out of it. The shale formation is approximately 20.23 feet in diameter at its base, and between 25 to 30 feet at the top.

The formation of Castle Rock began about two hundred million years ago. Over time water eroded a way the surrounding rock creating its unique shape. Castle Rock was know to early explores of the area simply as the "castle". At one time ladders provided access to the top of the rock, but they were removed in 1911, after Virgil Senter fell to his death. Steps and hand rails leading to the terrace were added later. In 2001 a sign explaining how the rock was formed was placed in front of Castle Rock.

http://en.wikipedia.org/wiki/Castle_Rock_%28Pineville,_West_Virginia%29

A Castle Rock egy geológiai képződmény Pineville-ben, Nyugat-Virginiában. Nevét kastélyhoz hasonló alakjáról kapta, és mintegy 65 méterrel magasodik a Rock Castle Creek, a Guyandotte folyó egyik ága fölé. Alapjának átmérőjét 30 méterre becsülik. Felfelé haladva középtájon egy kőterasz található, itt egy vékonyabb palaképződmény emelkedik ki a sziklából. E palaképződmény az alján mintegy 7 méter, a tetején pedig 8-10 méter átmérőjű.

A Castle Rock kialakulása kb. 200 millió évvel ezelőtt kezdődött. Az idők során a víz errodálta a szomszédos sziklákat, létrehozva e sajátos alakú kőtömböt. A Castle Rock a vidék első felfedezői számára egyszerűen mint a"kastély" volt ismert. Egy időben létrák biztosították a feljutást a szikla tetejére, azonban 1911-ben eltávolították őket, miután Virgil Senter leesett és halálra zúzta magát. Később a kőteraszra vezető lépcsőket és korlátokat alakítottak ki. 2001-ben egy, a szikla kialakulását magyarázó táblát állítottak a Castle Rock elé.

Golden Twenties

Golden Twenties or Happy Twenties is a term, mostly used in Europe, to describe the 1920s, in which most of the continent had an economic boom following the First World War and the severe economic downturns that took place between 1919–1923, and before the Wall Street Crash in 1929.

It is often applied to Germany, which during the early 1920s, experienced, like most of Europe, record-breaking levels of inflation of one trillion percent between January 1919 and November 1923. The inflation was so severe that printed currency was often used for heating and other uses, and everyday requirements like food, soap, electricity cost a wheelbarrow full of banknotes. Such events, among many other factors, triggered the

rise of fascism in Italy, as well as the ill-fated Beer Hall Putsch, masterminded by a young Adolf Hitler.

Before long, the Weimar Republic under Chancellor Gustav Stresemann managed to tame the extreme levels of inflation by the introduction of a new currency, the Rentenmark, with tighter fiscal controls and reduction of bureaucracy, leading to a relative degree of political and economic stability.

http://en.wikipedia.org/wiki/Golden_Twenties

Az Arany Húszas Évek vagy Boldog Húszas Évek elnevezést elsősorban Európában használják az 1920-as évekre, amikor a kontinens nagyrészén gazdasági fellendülés követte az első világháborút és az 1919-1923 közötti gazdasági visszaesést, és e fellendülés eltartott egészen az 1929-es gazdasági világválságig.

Gyakran egyenesen Németországgal kapcsolatban használják ezt a kifejezést, ahol Európa legtöbb országához hasonlóan, az infláció mértéke elérte az egybillió százalékot is 1919 januárja és 1923 novembere között. Az infláció olyan súlyos volt, hogy a papírpénzzel sok esetben begyújtottak, fűtöttek vagy más célra használták azt, illetve a mindennapi szükségletek - mint az étel, a szappan vagy a villanyáram - kifizetéséhez egy talicskányi bankjegyre volt szükség. Az ilyen helyzetek, sok más tényező mellett, elősegítették Olaszországban a fasizmus térnyerését, akárcsak a hírhedt müncheni sörpuccsot amelyet a fiatal Adolf Hitler irányított.

Nemsokkal később azonban a Weimar-i Köztársaság Gustav Stresemann kancellár vezetésével megfékezte az inflációt egy új fizetőeszköz, a Rentenmark bevezetésével, szigorúbb pénzügyi szabályozással és a bürokrácia csökkentésével, ezáltal viszonylagos politikai és gazdasági stabilitást teremtve.

Wialon

Wialon is a web-based GPS tracking software platform with some fleet management features, developed by a Belarusian company Gurtam. Wialon is available for both Linux and Windows operating systems. Wialon is notable because of its compatibility with different GLONASS and GPS tracking units, counting 131 in the beginning of July 2010.

Wialon was introduced internationally in 2010 at CeBIT computer expo. Before that Wialon GPS tracking software was mainly used in Post-Soviet states. According to Gurtam website, there are about 200 GPS tracking services worldwide who have already implemented Wialon GPS tracking platform.

http://en.wikipedia.org/wiki/Wialon

A Wialon egy internet-alapú GPS nyomkövető szoftverplatform, néhány flottakezelési tulajdonsággal, amit a fehérorosz Gurtam cég fejlesztett ki. A Wialon mind Linux mint Windows operációs rendszerre elérhető. A Wialon azért említésre méltó, mert különböző GLONASS és GPS követőegységekkel is kompatibilis, ezek száma 2010 júliusában elérte a 131-et.

A Wialont a 2010-es CeBIT-en mutatták be a nemzetközi közönségnek. Ezt megelőzően a Wialon GPS nyomkövető szoftvert elsősorban a szovjet utódállamokban használták. A Gurtam honlapja szerint a világon mintegy 200 GPS nyomkövető szolgáltatás alkalmazza már a Wialon GPS nyomkövető platformot.

Web accessibility

Web accessibility refers to the inclusive practice of making websites usable by people of all abilities and disabilities. When sites are correctly designed, developed and edited, all users can have equal access to information and functionality. For example, when a site is coded with semantically meaningful HTML, with textual equivalents provided for images and with links named meaningfully, this helps blind users using text-to-speech software and/or text-to-Braille hardware. When text and images are large and/or enlargable, it is easier for users with poor sight to read and understand the content. When links are underlined (or otherwise differentiated) as well as coloured, this ensures that color blind users will be able to notice them. When clickable links and areas are large, this helps users who cannot control a mouse with precision. When pages are coded so that users can navigate by means of the keyboard alone, or a single switch access device alone, this helps users who cannot use a mouse or even a standard keyboard. When videos are closed captioned or a sign language version is available, deaf and hard of hearing users can understand the video. When flashing effects are avoided or made optional, users prone to seizures caused by these effects are not put at risk. And when content is written in plain language and illustrated with instructional diagrams and animations, users with dyslexia and learning difficulties are better able to understand the content. When sites are correctly built and maintained, all of these users can be accommodated while not impacting on the usability of the site for non-disabled users.

http://en.wikipedia.org/wiki/Web_accessibility

A webes akadálymentesítés arra az átfogó gyakorlatra utal amikor a weblapokat minden felhasználó, ép vagy bármilyen fogyatékos ember számára is egyaránt használhatóvá, elérhetővé teszik. Amikor egy honlapot megfelelően terveznek, fejlesztenek és szerkesztenek, minden felhasználó számára egyformán elérhető az ott található összes információ és funkcionalitás. Például, ha az oldal szemantikai jelentéssel bíró HTML-ben van kódolva, a képekhez, fotókhoz megfelelő szöveges megjelenést biztosít, és a linkeknek értelmes elnevezést adtak, ez segíti a látássérült felhasználókat, hogy az oldalt felolvasó-szoftverrel és/vagy szöveget Braille nyelvvé alakító hardverrel használják. Ha a szöveg és a képek nagyok vagy felnagyíthatók, ez a gyengénlátó felhasználók számára könnyíti meg az oldal olvasását és az ott lévő tartalom megértését. Amennyiben a linkeket aláhúzzák (vagy más módon jelölik őket) és más színű betűvel jelenítik meg őket, ez a színvak felhasználókat segíti abban, hogy észrevegyék azokat. Ha a klikkelhető linkek és területek nagyok, ez azon felhasználóknak könnyíti meg az oldal használatát, akik az egeret nem tudják kellő finomsággal, pontossággal mozgatni. Amikor az oldalakat oly módon kódolják, hogy azon csupán a billentyűzet vagy egy egyszerű kapcsolóeszköz segítségével lehet navigálni, ezzel az egeret vagy a normál billentyűzetet használni egyáltalán nem képes felhasználókat segítik. A videók feliratozása vagy hozzájuk jelnyelvi verzió biztosítása a siket és hallássérült

felhasználóknak biztosítja azt, hogy megértsék a tartalmat. Ha a villanó effekteket kihagyják vagy használatukat opcionálissá teszik, akkor azon felhasználókat, akiknél ezek a fényhatások epilepsziás rohamot válthatnak ki, nem teszik ki ilyen veszélynek. Ha a tartalmat egyszerű mondatokkal írják le amit diagramokkal, animációkkal is illusztrálnak, akkor a dyslexiás vagy tanulási nehézséggel küzdő felhasználókat segítik a tartalom minél jobb megértésében. Mikor egy weboldalt megfelelően hoznak létre és tartják azt karban, minden felhasználó használhatja azt anélkül, hogy a teljesen ép felhasználókra bármilyen hatással lennének ezek a változások.

9.5. Kézzel angolról németre fordított tesztkorpusz

University of Oxford

Die Universität in Oxford (informell Universität Oxford oder Oxford) ist eine sich in Oxford befindlichen Universität, in dem Vereinigten Königreich. Die ist die zweitälteste erhaltene Universität der Welt und die älteste in der Englisch sprechenden Welt. Obwohl ihres genaue Gründungsdatum nicht bekannt ist, hier wurde schon seit 1096 nachgewiesen gelehrt. Die Universität wuchs schnell vom 1167, als Heinrich II. englischen Studenten das Lernen an der Pariser Universität verboten hat.

Nach einem Streit zwischen Studenten und den Stadteinwohnern flüchteten einige Akademiker im Jahre 1209 in Richtung Nordost nach Cambridge, wo sie die spätere Universität in Cambridge gegründet haben.

http://en.wikipedia.org/wiki/University_of_Oxford

Foreign relations of Pakistan

Pakistan ist ein aktives Mitglied der Vereinigten Nationen.

Es war Mitglied der Militärbündnisse CENTO und SEATO. Sein Bündnis mit der Vereinigten Staaten war besonders dann eng, als die Sowjets in das Nachbarland Afghanistan eingedrungen waren. Im Jahre 1964, als die drei Länder sehr eng mit der USA verbunden waren und als Nachbars der Sowjetunion eine sowjetische Expansion befürchteten, unterschrieb Pakistan den Vertrag zur Regionalen Kooperation für Entwicklung (RCD) mit der Türkei und Iran. Bis zu diesen Tagen hatte Pakistan eine sehr enge Verbindung mit der Türkei. RCD wurde nach der Iranischen Revolution gelöscht und die pakistanisch-türkische Initiativen führten im Jahre 1985 zu der Gründung der Organisation für wirtschaftliche Zusammenarbeit (CEO).

http://en.wikipedia.org/wiki/Foreign_relations_of_Pakistan

Pete Seeger

Seeger wurde in dem französischen Krankenhaus in Midtown, einem Stadtteil von Manhatten geboren. Seine Eltern lebten mit seinen Großeltern in Patterson, in New York von 1918 bis 1920. Sein Vater Charles Louis Seeger Jr. war Komponist und der Vorreiter der ethnomusikologischen Forschung der amerikanischen und der nicht westlichen Musikwissenschaft. Seine Mutter, Constance de Clyver Edson war eine klassische Geigerin und Lehrerin. Seine Eltern haben sich geschieden als Seeger sieben Jahre alt war. Seine Stiefmutter, Ruth Crawford Seeger war die bedeutendste

Frauenkomponistin des zwanzigsten Jahrhunderts. Sein ältester Bruder, Charles Seeger III, war Radioastronom und sein nächst älterer Bruder, John Seeger unterrichtete in den 50-er jahren in der Dalton School, in Manhatten und war der Rektor der Fieldston Lower School in Bronx von 1960 bis 1976. Sein Onkel, Alan Seeger war ein berühmter Dichter, der während des ersten Weltkrieges umgebracht wurde. Seine Halbschwester, die ebenfalls eine bekannte Volksdarstellerin war, war seit vielen Jahren mit dem englischen Volkssänger Ewan MacColl verheiratet. Halbbruder Mike Seeger hat sich die New Lost City Ramblers gegeründet, deren Mitglied, John Cohen Pete's andere Halbschwester, die Sängerin Penny Seeger, eine ebenfalls sehr talentierte Sängerin geheiratet hat.

http://en.wikipedia.org/wiki/Pete_Seeger

Avidius Cassius

Er wurde nach der falschen Meldung über den Tod von Mark Aurels im Jahre 175 zum Römischem Kaiser proklamiert; die Quellen zeigen auch, dass er durch Faustina, die Frau von Marcus, ermuntert wurde, die über die schwache Gesundheit ihres Mannes besorgt war, und glaubte, er wäre am Rande des Todes und die hat es für notwendig gehalten Cassius als Stützer in diesem Fall zu haben, da ihr Sohn Commodus noch sehr jung (13) war. Die Beweise, darunter Marks eigene Meditation, untermauern die Vermutung, dass Mark tatsächlich krank war, aber zu der Zeit als er sich erholte, war Cassius schon von den ägyptischen Legionen II Traiana fortis und XXII Deitoriana völlig anerkannt.

Gemäß Cassius Diro hat Mark, der gegen den nördlichen Stämmen einen Kampf führte, zuerst den Aufstand vor seinen Soldaten geheim zu halten versucht, aber nach dem die Nachricht unter denen verbreitet hat, hat er sie angesprochen. In dieser Rede, die Dio Mark zuschreibt, beklagt er die Illoyalität "eines lieben Freundes", zugleich hofft er, dass Cassius nicht getötet oder Selbsmord begehen würde, so, dass er ihn begnadigen könnte. Der Senat hat Cassius zum Staatsfeind deklariert.

Seit einem Dokument mit dem Vermerk dieses Datums, als Anfang des ersten Jahres der Herrschaft von Cassius, ist es bekannt, dass Cassius am 03. Mai als Kaiser anerkannt wurde. Der Anfang seines Aufstandes war im April im Jahre 175.

http://en.wikipedia.org/wiki/Avidius_Cassius

Meitetsu 5000 series (2008)

Der Innenraum hat einen hellgrauen Überblick. Mit der früheren Serie 3150 entsprechend sind die Einzelsitze 470 mm weit. Sie sind mit einem blauen Mokett bezogen. Die Sitze haben eine freitragende Struktur, d.h. sie sind an die Wände befestigt und in jedem Wagen gibt es ein klappbarer Sitz. Die Haltegriffe sind 1,630 mm von dem Fußboden entfernt.

Die Anzahl der bevorzugten Sitzplätze für Ältere und Kranke wurden in jedem Wagen auf 10 Sitze erhöht. Die bevorzugten Sitzplätze haben einen roten Mokett und orange Haltegriffe und -stangen zum Festhalten und für die Klarstellung der Differenz zwischen diesen und den normalen Sitzplätzen. Die Rollstuhlplätze befinden sich in beiden Wagen hinter der Fahrerkabine an jedem Ende des Zuges.

http://en.wikipedia.org/wiki/Meitetsu_5000_series_(2008)

Canberra 400

Die Canberra 400, auch als GMC 400 und Stegbar Canberra 400 oder in den Anfangsjahren als National Capital 100 bekannt, war eine V8-Supercar-Rennen auf den Straßen der australischen Hauptstadt Canberra.

Unglücklicherweise hat die Canberra 400 nur 3 Jahre aus dem 5-jährigen Vertrag gedauert. Im Jahre 2002 wurde die liberale Regierung von Gary Humphries durch die Regierung der Arbeiter-Partei in der Leitung von Jon Stanhope abgelöst. Der neue Ministerpräsident hat das Rennen in 2002 zugelassen, hat aber in 2003 dagegen entschieden. Der Hauptgrund für die Vertragslösung war der für das Rennen zu bezahlende Geldbetrag. Die anfängliche Kostenschätzung von Kate Carnell hat sich mit den Jahren geplatzt und einige aus Canberra haben gedacht, dass das Geld für etwas anders besser ausgegeben werden könnte. Das Rennen hat nicht das erwartete Geld gebracht. Die Motels und Hotels in der Umgebung von Camberra waren voll und haben das beste Wintersaison überhaupt aber die Anzahl der Masse um die Rennbahn hat sich von 101.000 im Jahre 2000 auf 89.000 im Jahre 2002 gesunken. Das wurde auf die Jahreszeit und das Wetter zurückgeführt. In Canberra kann die Temperatur Tags über weniger als um 5°C im Juni liegen. Das ist für die meisten Canberras ein ganz normaler Winter, aber für die interstaatliche Besucher, die an solche Kälte nicht wohnt waren, war das zu kalt.

http://en.wikipedia.org/wiki/Canberra_400

Upper 10

Upper 10 ist ein koffeinhaltiges Zitrone-Lime Erfrischungsgetränk ähnlich zu Sprite, 7 Up, Sierra Mistund Bubble Up. Abgefüllt war durch RC Cola.

Die Upper 10 Marke hat im Jahre 1933 als ein Produkt der Nehi Corporation (später Royal Crown Corporation) debütiert. Das Upper 10 war einer der Flaggenschiffe der RC Cola durchweg in der Geschichte der Firma. Jedenfalls, als RC Cola durch Cadbury Schweppes plc im Jahre 2000 gekauft und anschließend in die Firmenoperationen der Firma Dr Pepper/Seven Up, Inc. eingegliedert wurde, die Abfüllfirmen haben die Abfüllung von Upper 10 zu Gunsten der ähnlichen, populärer und nicht koffeinhaltiges 7 Up (was ebenfalls im Eigentum der Dr Pepper Snapple Group stand) allmählich eingestellt .

Upper 10 wird nach wie vor durch die Firma Cott Beverages, dieselbe Firma, die RC Cola international verteilt, außerhalb von Nord-Amerika verkauft.

http://en.wikipedia.org/wiki/Upper_10

Munich Philharmonic

Das Orchester wurde in München im Jahre 1893 durch Franz Kaim, den Sohn eines Klavierbauers als das Kaim-Orchester gegründet. Im Jahre 1895 hat seine Residenz in der städtischen Tonhalle (Konzerthalle) gesetzt. Es hat bald solche hervorragenden Dirigenten angezogen wie: Gustav Mahler der erste Dirigent der Gruppe im Jahre 1897

und der seine Symphony No. 4 und Symphonie No. 8 mit dem Orchester vorgeführt hat, Bruno Walter hat das Orchestra bei der posthumen Premiere von Mahlers Das Lied von der Erde dirigiert. Felix Weingartner war Musikdirektor von 1898 bis 1905 und der junge Wilhelm Furtwängler hat hier sein vielversprechendes Dirigentendebüt im Jahre 1906. Inzwischen hat Anton Bruckner, ein Schüler von Ferdinand Löwe eine bleibende Tradition der Bruckner-Vorstellungen geschafft, die auch bis heute fortgesetzt wird.

Durch diese Zeit war das Orchesters, das im Jahre 1910 als Orchester des Münchner Konzertvereins hieß, privat finanziert aber während des Ersten Weltkrieges gab es finanzielle Schwierigkeiten und die Musikanten wurden in militärischen Dienst gezogen, das Orchester war gezwungen seine Tätigkeit zu beenden. Nach dem Krieg wurde das Orchester von der Münchner Stadt übernommen und wurde neugestrartet unter der Führung von dem Komponist Hans Pfitzner, der bald durch den Bruckner-Vorkämpfer Siegmund von Hausegger aufgelöst wurde. Im Jahre 1928 hat das Orchester seinen heutigen Name erworben.

Nach der Entstehung der Nazi Partei im Jahre 1933, das Orchester stempelte ihre Partituren mit Hakenkreuz und den Wörtern "Das Orchester der faschistischen Bewegung". (die Hakenkreuze waren bis zu den frühen 90-er Jahren nicht entfernt.)

http://en.wikipedia.org/wiki/Munich_Philharmonic

Castle Rock (Pineville, West Virginia)

Castle Rock ist eine geologische Formation, die sich in Pineville, West Virginia in der Nähe der Öffentlichen Bibliothek von Pineville befindet. Hat ihren Name wegen seiner Ähnlichkeit zu einer Burg bekommen, die erhebt sich über 200 Fuß über Rock Castle Creek, eine Abzweigung des Flusses Guyandotte. Ihr Grunddurchmesser wird etwa für 100 Fuß geschätzt. Auf der Mitte gibt es eine Steinterrasse mit einer eingeengten Schieferformation, die sich aus ihr erhebt. Die Schieferformation ist im Grunddurchmesser etwa 20 bis 23 Fuß breit und etwa 25 bis 30 Fuß an der Spitze.

Die Formation von Castle Rock hat mehr els vor zweihundert Millionen Jahren begonnen. Mit der Zeit hat das Wasser die umgebenden Berge erodiert, und formte ihre eigenartige Gestalt. Castle Rock war unter den jüngsten Erforschern einfach als "Burg" genannt. In einer Zeit haben Leiter das Erreichen der Felsenspitze unterstützt, diese wurden aber im Jahre 1911, nachdem Virgil Senter tödlich verunglückte entfernt. Zu der Terrasse führende Treppen und Ranken wurden später erbaut. Man hat im Jahre 2001 eine Zeichenerklärung über Entstehung der Felsen in Front von Castle Rock gesetzt.

http://en.wikipedia.org/wiki/Castle_Rock_%28Pineville,_West_Virginia%29

Golden Twenties

Die Goldenen Zwanziger oder Glücklichen Zwanziger Jahren ist ein Ausdruck, der meist in Europa dafür gebraucht wird, die 1920-er Jahren zu beschreiben, in denen in den meisten Gebieten des Kontinents ein wirtschaftlicher Aufschwung dem Ersten Weltkrieg folgend gab und heftige Konjunkturschwächen zwischen 1919 und 1923, bevor des Krachs der Wall Street im Jahre 1929, ereigneten.

Man bezieht es oft an Deutschland, wo während der frühen 20-er Jahre, wie meist in Europa, ein rekordbrechendes Niveau der Inflation von ein Billion Prozent zwischen Januar 1919 und November 1923 erfahren wurde. Die Inflation war so stark, dass die gedruckte Währung oft für Heizen und andere Zwecke benutzt war und die alltäglichen Produkte wie Lebensmittel, Seife, Strom kosteten Schubkarren voller Geldscheine. Unter anderen Faktor haben solche Ereignisse das Emporsteigen des Faschismus in Italien ausgelöst, wie auch den Hitler-Putsch, der durch den jungen Adolf Hitler geleitet wurde.

Lange davor hat die Weimarer Republik unter Kanzler Gustav Stresemann das extremes Niveau der Inflation durch die Einführung einer neuen Währung, der Rentenmark, mit harten Finanzkontrolle und Reduzierung der Bürokratie zu zähmen versucht, zu einem relativen Grad der politischen und wirtschaftlichen Stabilität führend.

http://en.wikipedia.org/wiki/Golden_Twenties

Wialon

Wialon ist eine auf webbasierte GPS Verfolgung-Software-Plattform mit einigen Flottenmanagement-Funktionalitäten, die von der belarussischen Firma Gurtam entwickelt wurde. Wialon ist sowohl für Linux als auch für Windows Oparationssysteme erhältlich. Wialon ist beachtenswert wegen ihrer Kompatibilität mit verschiedenen GLONASS und GPS Verfolgung-Einheiten, genau 131 am Anfang Juli 2010.

Wialon wurde im Jahre 2010 an der CeBIT IT-Messe international vorgestellt. Zuvor hat man Wialon GPS Verfolgung-Software hauptsächlich in den sowjetischen Nachfolgestaaten benutzt. Gemäß der Webseite von Gurtam, es sind mehr als 200 GPS Verfolgung-Services weltweit, die alle schon Wialon GPS Verfolgung-Plattform implementiert haben.

http://en.wikipedia.org/wiki/Wialon

Web accessibility

Die Web-Zugänglichkeit bezieht sich auf die Praxis der Nutzbarmachung von Webseiten für Leute mit allen Fähigkeiten und Behinderten. Wenn die Webseiten richtig entworfen, entwickelt und editiert sind, können alle Benutzer gleichermaßen zu den Informationen und den Funktionalitäten zugreifen. Zum Beispiel, wenn eine Webseite mit semantisch aussagefähigen HTML, mit den Bildern entsprechenden Texten unterstützt und mit sinnvoll benannten Links codiert sind, das kann blinden Benutzern helfen eine Text-in-Sprache Umwandlungs-Software und/oder eine Text-to-Braille Hardware zu benutzen. Wenn die Texte und Bilder umfangreich/groß und/oder vergrößerbar sind, ist für die Benutzer mit schwachen Sehvermögen leichter den Inhalt zu lesen und zu verstehen. Wenn die Links unterstrichen (oder in anderer Weise unterscheidet) und gefärbt sind, das sichert, dass die farbblinden Benutzer fähig sind es zu merken. Wenn die anklickbaren Links und Bereichen sind groß, das hilft Benutzern, die die Maus nicht präzis kontrollieren können. Wenn Seiten so codiert sind, dass die Benutzer allein mit der Nutzung der Tastatur oder allein mit einer einzelnen Schalt-Zugangsvorrichtung navigieren können, was solchen Benutzern hilft, die weder die Maus noch eine Standard-Tastatur benutzen können. Wenn Videos geschlossen

untertitelt sind oder eine Zeichensprachen-Version erreichbar ist, so können taube und schwerhörige Benutzer die Videos verstehen. Wenn leuchtende Effekte vermieden oder optional gemacht werden, werden Benutzer mit durch diese Effekten ausgelösten Krampfanfällen nicht in Gefahr gebracht. Und wenn der Inhalt in einfacher Sprache oder mit Lehrdiagramm und Animationen illustriert geschrieben sind, können Benutzer mit Dyslexia und Lernschwierigkeiten den Inhalt besser verstehen. Wenn die Seiten richtig aufgebaut und gewartet sind, können diese all diesen Benutzern angepasst werden und die Benutzbarkeit wird nicht auf nicht behinderte Benutzer eingeengt.

http://en.wikipedia.org/wiki/Web_accessibility

9.6. Az algoritmus által használt stopszavak

Magyar

a, az, és, van, hogy, nem, is, egy, ez, meg, ha, kell, de, csak, már, volt, amely, azt, még, el, aki, minden, mint, tud, ki, ami, nagy, illetve

Angol

the, of, and, to, a, in, for, is, on, that, by, s, with, i, or, not, you, be, are, this, at, it, its, as, from, your, have, was, an, will, all, can, more, has, we, one, but, about, which, do, their, our, they, up, my, out, if, new, any, his, he, been, were, t

Francia

alors, au, aucuns, aussi, autre, avant, avec, avoir, bon, car, ce, cela, ces, ceux, chaque, ci, comme, comment, dans, des, du, dedans, dehors, depuis, deux, devrait, doit, donc, dos, droite, début, elle, elles, en, encore, essai, est, et, eu, fait, faites, fois, font, force, haut, hors, ici, il, ils, je, juste, la, le, les, leur, la, ma, maintenant, mais, mes, mine, moins, mon, mot, meme, ni, nommés, notre, nous, nouveaux, ou, ou, par, parce, parole, pas, personnes, peut, peu, piece, plupart, pour, pourquoi, quand, que, quel, quelle, quelles, quels, qui, sa, sans, ses, seulement, si, sien, son, sont, sous, soyez, sujet, sur, ta, tandis, tellement, tels, tes, ton, tous, tout, trop, tres, tu, valeur, voie, voient, vont, votre, vous, vu, ça, étaient, état, étions, été, etre"));

Spanyol

un, una, unas, unos, uno, sobre, todo, también, tras, otro, algún, alguno, alguna, algunos, algunas, ser, es, soy, eres, somos, sois, estoy, esta, estamos, estais, estan, como, en, para, atras, porque, por qué, estado, estaba, ante, antes, siendo, ambos, pero, por, poder, puede, puedo, podemos, podeis, pueden, fui, fue, fuimos, fueron, hacer, hago, hace, hacemos, haceis, hacen, cada, fin, incluso, primero, desde, conseguir, consigo, consigue, consigues, conseguimos, consiguen, ir, voy, va, vamos, vais, van, vaya, gueno, ha, tener, tengo, tiene, tenemos, teneis, tienen, el, la, lo, las, los, su, aqui, mio, tuyo, ellos, ellas, nos, nosotros, vosotros, vosotras, si, dentro, solo, solamente, saber, sabes, sabe, sabemos, sabeis, saben, ultimo, largo, bastante, haces, muchos, aquellos, aquellas, sus, entonces, tiempo, verdad, verdadero, verdadera, cierto, ciertos, cierta, ciertas, intentar, intento, intenta, intentas, intentamos, intentais, intentan, dos, bajo, arriba, encima, usar, uso, usas, usa, usamos, usais, usan, emplear, empleo, empleas, emplean, ampleamos, empleais, valor, muy, era, eras, eramos, eran, modo, bien, cual, cuando, donde,

mientras, quien, con, entre, sin, trabajo, trabajar, trabajas, trabaja, trabajamos, trabajais, trabajan, podria, podrias, podriamos, podrian, podriais, yo, aquel

Német

aber, als, am, an, auch, auf, aus, bei, bin, bis, bist, da, dadurch, daher, darum, das, daß, dass, dein, deine, dem, den, der, des, dessen, deshalb, die, dies, dieser, dieses, doch, dort, du, durch, ein, eine, einem, einen, einer, eines, er, es, euer, eure, für, hatte, hatten, hattest, hattet, hier, hinter, ich, ihr, ihre, im, in, ist, ja, jede, jedem, jeden, jeder, jedes, jener, jenes, jetzt, kann, kannst, können, könnt, machen, mein, meine, mit, muß, mußt, musst, müssen, müßt, nach, nachdem, nein, nicht, nun, oder, seid, sein, seine, sich, sie, sind, soll, sollen, sollst, sollt, sonst, soweit, sowie, und, unser, unsere, unter, vom, von, vor, wann, warum, was, weiter, weitere, wenn, wer, werde, werden, werdet, weshalb, wie, wieder, wieso, wir, wird, wirst, wo, woher, wohin, zu, zum, zur, über

Forrás: http://www.ranks.nl/stopwords/

9.7. Hunglish korpusz fájljai

bi_Andersen_1.bi, Austen_1.bi, Austen_2.bi, Austen_3.bi, Balzac_1.bi, Balzac_2.bi, Barrie_1.bi, Barrie_2.bi, Baum_1.bi, Bible_1.bi, Bible_2.bi, Bible_3.bi, Blake_1.bi, Boccaccio_1.bi, Bronte_1.bi, Carroll_1.bi, Chekhov_1.bi, Chekhov_2.bi, Collodi_1.bi, Cooper_1.bi, Dickens_1.bi, Dickens_2.bi, Dickens_3.bi, Dostoyevsky_1.bi, DTM_1.bi, DTM_2.bi, EU_Law_0.bi, ..., EU_Law_2641.bi, France_1.bi, Goethe_1.bi, Gogol_1.bi, Gogol_2.bi, Grimm_1.bi, Hawthorne_1.bi, Hawthorne_2.bi, James_1.bi, Joyce_1.bi, Kafka_1.bi, Kipling_1.bi, Kipling_2.bi, Lamb_1.bi, Lermontov_1.bi, Lofting_1.bi, Machiavelli_1.bi, opensource_1.bi, opensource_2.bi, opensource_3.bi, opensource_4.bi, opensource_5.bi, opensource_6.bi, opensource_7.bi, opensource_8.bi, opensource_9.bi, Quincey_1.bi, Racine_1.bi, Rostand_1.bi, Schiller_1.bi, Scott_1.bi, Scott_2.bi, Shakespeare_10.bi, Shakespeare_11.bi, Shakespeare_12.bi, Shakespeare_13.bi, Shakespeare_14.bi, Shakespeare_1.bi, Shakespeare_2.bi, Shakespeare_3.bi, Shakespeare_4.bi, Shakespeare_6.bi, Shakespeare_7.bi, Shaw_1.bi, Shuffle.bi, Stevenson_1.bi, Stevenson_2.bi, subtitles.bi, Swift_1.bi, test.bi, Tolstoy_1.bi, Twain_1.bi, Twain_2.bi, Twain_3.bi, Twain_4.bi, Twain_5.bi, Twain_6.bi, Verne_1.bi, Verne_2.bi, Verne_3.bi, Verne_4.bi, Verne_5.bi, Verne_6.bi, Verne_7.bi, VOA_1.bi, Webster_1.bi, Wells_1.bi, Whitman_1.bi

Összesen 1962 db EU_Law fájlt használtunk fel, ezek felsorolásáról most eltekintünk.

9.8. Fordításon és n-gramon alapuló algoritmus paramétereinek optimalizálása

F_4	d_{num}/d	d	d_{num}/D	T+	F+
0.76119402985075	7	120	1	9	0
0.76119402985075	7	140	1	9	0
0.76119402985075	7	160	1	9	0
0.76119402985075	7	180	1	9	0
0.76119402985075	7	200	1	9	0
0.76119402985075	7	220	1	9	0
0.76119402985075	7	240	1	9	0
0.76119402985075	7	260	1	9	0
0.76119402985075	7	280	1	9	0
0.76119402985075	7	300	1	9	0
0.75742574257426	5	80	1	9	1
0.75742574257426	6	80	1	9	1
0.75742574257426	5	100	1	9	1
0.75742574257426	6	100	1	9	1
0.75742574257426	5	120	1	9	1
0.75742574257426	6	120	1	9	1
0.75742574257426	6	140	1	9	1
0.75742574257426	5	140	1	9	1
0.75742574257426	6	160	1	9	1
0.75742574257426	5	160	1	9	1
0.75742574257426	6	180	1	9	1
0.75742574257426	5	180	1	9	1
0.75742574257426	6	200	1	9	1
0.75742574257426	5	200	1	9	1
0.75742574257426	5	220	1	9	1
0.75742574257426	6	220	1	9	1
0.75742574257426	6	240	1	9	1
0.75742574257426	5	240	1	9	1
0.75742574257426	6	260	1	9	1
0.75742574257426	5	260	1	9	1
0.75742574257426	5	280	1	9	1
0.75742574257426	6	280	1	9	1
0.75742574257426	5	300	1	9	1
0.75742574257426	6	300	1	9	1
0.75369458128079	4	40	1	9	2
0.75369458128079	4	60	1	9	2
0.75369458128079	4	80	1	9	2
0.75369458128079	4	100	1	9	2
0.75369458128079	4	120	1	9	2
0.74634146341463	3	60	2	9	4
0.74634146341463	3	80	2	9	4
0.74634146341463	3	100	2	9	4
0.74634146341463	3	120	2	9	4
0.74634146341463	4	140	1	9	4
0.74634146341463	3	140	2	9	4
0.74634146341463	3	160	2	9	4
0.74634146341463	3	180	2	9	4
0.74634146341463	3	200	2	9	4
0.74634146341463	3	220	2	9	4
0.74634146341463	3	240	2	9	4
0.74634146341463	3	260	2	9	4

9.1. táblázat: Lehetséges paraméterek, az F_4 maximalizációjára törekedve (15 oldalas cikk)

F_6	d+	d	d_{num}/D	T+	F+
0.82723577235772	2	60	1	11	49
0.82555780933063	2	80	1	11	50
0.81075697211155	2	100	1	11	59
0.80914512922465	2	120	1	11	60
0.80914512922465	2	140	1	11	60
0.80753968253968	2	160	1	11	61
0.80753968253968	2	180	1	11	61
0.80434782608696	2	200	1	11	63
0.80434782608696	2	220	1	11	63
0.8011811023622	2	240	1	11	65
0.79960707269155	2	260	1	11	66
0.79960707269155	2	280	1	11	66
0.79647749510763	2	300	1	11	68
0.77405857740586	3	40	1	10	36
0.77405857740586	3	60	1	10	36
0.77405857740586	3	80	1	10	36
0.77244258872651	3	100	1	10	37
0.77244258872651	3	120	1	10	37
0.77244258872651	3	140	1	10	37
0.77244258872651	3	160	1	10	37
0.77244258872651	3	180	1	10	37
0.77244258872651	3	200	1	10	37
0.77244258872651	3	220	1	10	37
0.77244258872651	3	240	1	10	37
0.77244258872651	3	260	1	10	37
0.75975359342916	2	20	1	10	45
0.75819672131148	3	280	1	10	46
0.75819672131148	3	300	1	10	46
0.75510204081633	7	120	1	9	0
0.75510204081633	7	140	1	9	0
0.75510204081633	7	160	1	9	0
0.75510204081633	7	180	1	9	0
0.75510204081633	7	200	1	9	0
0.75510204081633	7	220	1	9	0
0.75510204081633	7	240	1	9	0
0.75510204081633	7	260	1	9	0
0.75510204081633	7	280	1	9	0
0.75510204081633	7	300	1	9	0

9.2. táblázat: Lehetséges paraméterek, az F_6 maximalizációjára törekedve (15 oldalas cikk)

F_6	d+	d	d_{num}/D	T+	F+
0.75356415478615	3	40	1	10	49
0.74148296593186	3	60	1	10	57
0.74	3	80	1	10	58
0.73852295409182	3	120	1	10	59
0.73852295409182	3	100	1	10	59
0.73509933774834	7	120	1	9	12
0.73509933774834	7	160	1	9	12
0.73509933774834	7	140	1	9	12
0.73348017621145	6	80	1	9	13
0.73348017621145	7	180	1	9	13
0.73348017621145	7	220	1	9	13
0.73348017621145	7	200	1	9	13

0.73348017621145	7	240	1	9	13
0.73186813186813	7	300	1	9	14
0.73186813186813	7	280	1	9	14
0.73186813186813	7	260	1	9	14
0.72866520787746	6	140	1	9	16
0.72866520787746	6	180	1	9	16
0.72866520787746	6	120	1	9	16
0.72866520787746	6	100	1	9	16
0.72866520787746	6	160	1	9	16

9.3. táblázat: Lehetséges paraméterek, az F_6 maximalizációjára törekedve (Harry Potter)

9.9. Fordításon és n-gramon alapuló algoritmus részletes találati listája a Harry Potterre

- Places in Harry Potter (145)
  - ...Eeylops Owl Emporium ..... Malkin's Robes for All Occasions ...
  - ...3 Platform Nine and Three Quarters ...
  - ...School of Witchcraft and Wizardry[..... 4, Privet Drive, Little ..... Petunia, uncle Vernon, and ...
  - ...two streets away from 4 Privet Drive in the ..... keep an eye on Harry. In Harry Potter and the ...
  - ...Harry, Ron, and Hermione, they are brought to ...
  - ...Dark Arts. While ..... Defence Against the Dark Arts, ...
  - ...through The Leaky Cauldron (a ..... through The Leaky Cauldron to a ...
  - ...Eeylops Owl Emporium Eeylops Owl Emporium ..... rustling and the flickering of "jewel-...
  - ...safest place in the world for ..... Higher security vaults ...
  - ...once every 10 ... high security vaults ...
  - ...When Harry, Ron, and Hermione ...
  - ...Makers of Fine Wands ... gold letters over the ..... cushion in the dusty window. ... boxes piled neatly ...
  - ...copper, brass, pewter, silver, self-stirring, ...
  - ...bad smell (a mixture of bad ..... barrels of slimy stuff on the ... herbs, dried roots and ..... strings of fangs and snarled ..... silver unicorn horns (for twenty-one Galleons each) and ... beetle eyes (five ...
  - ...Harry, Ron, and Hermione. ..... Fred and George Weasley in the ...
  - ...Defence Against the Dark Arts ...
  - ...Bott's Every Flavour Beans, ...
  - ...Platform Nine and Three Quarters ..... platforms 9 and 10, ...
  - ...between platforms 9 and 10. ..... platforms 9 and 10 at ...
  - ...Professor Flitwick's classroom)[... King's Cross Station, ...
  - ...Harry Potter and the Sorcerer's ..... Quidditch Through the Ages. ...
- Portal:Harry Potter/Quotes/Archive (103)
  - ...Longbottom, if brains were gold, you'd be poorer than Weasley, and that's ...
  - ...believe your friends Misters Fred and George Weasley were responsible for trying to send you a toilet seat. No doubt they ..... ter turn

him into a pig, but I suppose he was so much like a pig anyway there wasn't much left ter do." ..... teach you how to bottle fame, brew glory, ... aren't as big a bunch of dunderheads as I usually have to teach." ...
    - ...find out that some wizarding ... than others, Potter. You don't want to go making friends with the wrong ..... sort for myself, thanks." ... Ah, go boil ..... Malfoy?" Draco Malfoy: " Longbottom, if brains were gold you'd be poorer than Weasley, and that's ..... mind, I'm going to ...
    - ...deal of courage to stand up to your ..... evil: only power and those too weak to seek it." ... increases fear of the thing ...
    - ...handy. I have one myself above my left knee that is a perfect map of the London Underground." ...
- List of Harry Potter characters (101)
    - ...Bott's Every Flavour Beans ...
    - ...Defence Against the Dark Arts for one ...
    - ...Nicholas de Mimsy- Porpington/ Nearly Headless Nick – ...
    - ...Magical Drafts and Potions Angelina Johnson – Gryffindor ...
    - ...Defense Against the Dark Arts ...
    - ...Defence Against the Dark Arts in ...
    - ...Defence Against the Dark Arts ...
    - ...Defence Against the Dark Arts ...
    - ...Harry Potter Harry ..... James Potter and Lily ...
    - ...Defence Against the Dark Arts ...
    - ...Defence Against the Dark Arts ... Head of Slytherin House, ...
    - ...Defence Against the Dark Arts ...
    - ...Befuddle Your Enemies with the Latest Revenges: Hair Loss, Jelly-Legs, Tongue- Tying and Much, Much More) ...
    - ...Harry Potter in Harry ..... Quidditch Through the Ages ...
    - ...Gryffindor common room and the ...
- Hogwarts (83)
    - ...Defence Against the Dark Arts ...
    - ...first year students are ..... glass or crystal phials, a ...
    - ...King's Cross station in London. The train ..... first year students are ...
    - ...important ceremony because, while you are here, your House will be something like your family within Hogwarts. You will have classes with the rest of your House, sleep in your House dormitory, and spend free time in your House ...
    - ...chivalry set Gryffindors ..... Nicholas de Mimsy- Porpington, more ... Nearly Headless Nick. ..... Gryffindor common room is ...
    - ...Professor McGonagall, the head of the ..... Sorting Hat said in Harry ...
    - ...Defence Against the Dark Arts, ...
    - ...into a pig and back in ... Defence Against the Dark Arts Defence Against the Dark Arts, ... against the Dark Arts, and to be ..... Defence Against the Dark Arts ...
    - ...o'clock. First year students ...
    - ...Head Boy or Girl or ..... Boy or Head Girl, they are not ..... Boy and Head Girl, are ...
    - ...corridor on the third floor, and ..... three-headed dog ...
    - ...Defence Against the Dark Arts ...
    - ...Nearly Headless Nick) ..... when Draco Malfoy ...

    - ...Defence Against the Dark Arts ...
    - ...School of Witchcraft and Wizardry on ...
- Harry Potter (character) (78)
    - ..... Harry Potter Harry ..... Harry Potter in Harry ...
    - ...School of Witchcraft and Wizardry to ... headmaster Albus Dumbledore and ...
    - ...Harry Potter and the Sorcerer's ..... powerful Dark Wizard, ...
    - ...Defence Against the Dark Arts ...
    - ...defence against the dark arts as ...
    - ...Defence Against the Dark Arts ...
    - ...shaped scar on his forehead. He is ..... Harry liked about his own ...
    - ...Defence Against the Dark Arts, and ...
    - ...Defence Against the Dark Arts and would have ...
    - ...Harry at times. I know that ..... Nimbus Two Thousand, was ..... first-year student. This ...
    - ...Nearly Headless Nick, ... Nearly Headless Nick's ...
    - ...Harry Potter on Harry ...
- Harry Potter universe (57)
    - ...young Sirius Black's ..... Wizards and witches who are ...
    - ...Beasts and Where to Find Them, it is said that the £ ..... Quidditch Through the Ages). This ...
    - ...when Harry saw his ..... what had happened. When he ...
    - ...Dumbledore, Nicolas Flamel, ... Morgana, Hengist of Woodcroft, Alberic Grunnion, Circe, ...
    - ...Bott's Every-Flavour Beans, ... Bott's Every-Flavour Beans ...
    - ...Quidditch Through the Ages ^ ...
- Portal:Harry Potter/Quotes (57)
    - ...Longbottom, if brains were gold, you'd be poorer than Weasley, and that's ...
    - ...money and life as you could want! The two things most human beings would choose above all. The trouble is, humans do have a knack of choosing precisely those things that are ...
    - ...first time in your life, you' ..... Voldemort: "Harry.....Potter..... ... Harry Potter: "Yes.". ...
- List of fictional books (56)
    - ...Quidditch Through the Ages by ... Beasts and Where to Find Them by ...
    - ...Defence Against the Dark Arts ...
    - ...Forces: A Guide to Self- Protection by Quentin Trimble ...
    - ...Thousand Magical Herbs and Fungi by Phyllida Spore ..... Magic by Bathilda Bagshot ..... Magical Drafts and Potions by ...
    - ...Wizards of the Twentieth Century ..... Magical Names of Our Time The ... Fall of the Dark Arts The ..... Great Wizarding Events of the ..... Modern Magical Discoveries A Study of Recent Developments in ..... Breeding for Pleasure and Profit Dragon Species of Great ..... Inferno: a Dragon-Keeper's ...
    - ...Professor Vindictus Viridian The ..... Against the Dark Arts ...
    - ...1 See also ...
- Treacle (13)

- ...moment later the desserts appeared. Blocks of ice cream in every flavor you could think of, apple pies, ...

9.10. Hasonlósági metrikán alapuló algoritmus találati listája a 12 Wikipédia cikkre

1. Pete Seeger (7)
   Seeger was born in French Hospital, Midtown Manhattan, the youngest of three sons.
   - Pete Seeger Manhattan közepén, a Midtown-nak is hívott városrész francia kórházában született. (8)

   His father, Charles Louis Seeger Jr. was a prominent musicologist, composer, and music professor.

   - Apja, ifj. Charles Louis Seeger, zeneszerző és zenetudós volt, aki az elsők között vizsgálta mind az amerikai népzenét, mind a nem-európai gyökerekből fakadó zenét. (1)

   His stepmother, Ruth Crawford Seeger, was one of the most significant female composers of the 20th century.

   - Nevelőanyja, Ruth Crawford Seeger egyike volt a huszadik század legkiemelkedőbb női zeneszerzőinek. (8)

   His eldest brother, Charles Seeger III, was a radio astronomer, and his next older brother, John Seeger, taught in the 1950s at the Dalton School in Manhattan and was the principal from 1960 to 1976 at Fieldston Lower School in the Bronx.

   -), rádiós csillagász volt, másik bátyja, John Seeger pedig az 1950-es években a manhattan-i Dalton School-ban tanított, majd 1960-tól 1976-ig a bronx-i Fieldston Lower School igazgatója lett. (24)

   His uncle, Alan Seeger, a noted poet, was killed during the First World War.

   - Nagybátyját, Alan Seegert, aki neves költő volt, megölték az első világháborúban. (0)

   His half-sister, Peggy Seeger, also a well-known folk performer, was married for many years to British folk singer Ewan MacColl.

   - Féltestvére, Peggy Seeger, aki szintén ismert népzenei előadó volt, hosszú évekig Ewan MacColl brit folkénekessel élt házasságban. (16)

   Half-brother Mike Seeger went on to form the New Lost City Ramblers, one of whose members, John Cohen, was married to Pete's other half-sister, singer Penny Seeger, also a highly talented singer.

- Másik féltestvére, Mike Seeger megalapította a New Lost City Ramblers-t (egy hagyományos zenei stílusban játszó vonósegyüttest), melynek egyik tagja, John Cohen Pete féltestvérét, az énekes Penny Seegert vette feleségül, és maga is igen tehetséges énekes volt. (25)

2. Golden Twenties (4)
   It is often applied to Germany, which during the early 1920s, experienced, like most of Europe, record-breaking levels of inflation of one trillion percent between January 1919 and November 1923.
   - Gyakran egyenesen Németországgal kapcsolatban használják ezt a kifejezést, ahol Európa legtöbb országához hasonlóan, az infláció mértéke elérte az egybillió százalékot is 1919 januárja és 1923 novembere között. (17)

   The inflation was so severe that printed currency was often used for heating and other uses, and everyday requirements like food, soap, electricity cost a wheelbarrow full of banknotes.

   - Az infláció olyan súlyos volt, hogy a papírpénzzel sok esetben begyújtottak, fűtöttek vagy más célra használták azt, illetve a mindennapi szükségletek - mint az étel, a szappan vagy a villanyáram - kifizetéséhez egy talicskányi bankjegyre volt szükség. (15)

   Such events, among many other factors, triggered the rise of fascism in Italy, as well as the ill-fated Beer Hall Putsch, masterminded by a young Adolf Hitler.

   - Az ilyen helyzetek, sok más tényező mellett, elősegítették Olaszországban a fasizmus térnyerését, akárcsak a hírhedt müncheni sörpuccsot amelyet a fiatal Adolf Hitler irányított. (10)

   Before long, the Weimar Republic under Chancellor Gustav Stresemann managed to tame the extreme levels of inflation by the introduction of a new currency, the Rentenmark, with tighter fiscal controls and reduction of bureaucracy, leading to a relative degree of political and economic stability.

   - Nemsokkal később azonban a Weimar-i Köztársaság Gustav Stresemann kancellár vezetésével megfékezte az inflációt egy új fizetőeszköz, a Rentenmark bevezetésével, szigorúbb pénzügyi szabályozással és a bürokrácia csökkentésével, ezáltal viszonylagos politikai és gazdasági stabilitást teremtve. (21)

3. Castle Rock (Pineville, West Virginia) (4)
   Castle Rock is a geological feature located in Pineville, West Virginia next to the Pineville Public Library.
   - A Castle Rock egy geológiai képződmény Pineville-ben, Nyugat-Virginiában. (1)

   Named for its resemblance to a castle, it rises about 200 feet above Rock Castle Creek, a branch of the Guyandotte River.

- Nevét kastélyhoz hasonló alakjáról kapta, és mintegy 65 méterrel magasodik a Rock Castle Creek, a Guyandotte folyó egyik ága fölé. (11)

The formation of Castle Rock began about two hundred million years ago.

- A Castle Rock kialakulása kb. 200 millió évvel ezelőtt kezdődött. (12)

At one time ladders provided access to the top of the rock, but they were removed in 1911, after Virgil Senter fell to his death.

- Egy időben létrák biztosították a feljutást a szikla tetejére, azonban 1911-ben eltávolították őket, miután Virgil Senter leesett és halálra zúzta magát. (19)

4. Wialon (4)
Wialon is a web-based GPS tracking software platform with some fleet management features, developed by a Belarusian company Gurtam.
   - A Wialon egy internet-alapú GPS nyomkövető szoftverplatform, néhány flottakezelési tulajdonsággal, amit a fehérorosz Gurtam cég fejlesztett ki. (5)

Wialon is available for both Linux and Windows operating systems.

- A Wialon mind Linux mint Windows operációs rendszerre elérhető. (16)

Wialon is notable because of its compatibility with different GLONASS and GPS tracking units, counting 131 in the beginning of July 2010.

- A Wialon azért említésre méltó, mert különböző GLONASS és GPS követőegységekkel is kompatibilis, ezek száma 2010 júliusában elérte a 131-et. (7)

According to Gurtam website, there are about 200 GPS tracking services worldwide who have already implemented Wialon GPS tracking platform.

- A Gurtam honlapja szerint a világon mintegy 200 GPS nyomkövető szolgáltatás alkalmazza már a Wialon GPS nyomkövető platformot. (8)

5. Munich Philharmonic (3)
Meanwhile Anton Bruckner pupil Ferdinand Löwe established an enduring tradition of Bruckner performance which continues to this day.
   - Mindeközben Anton Bruckner tanítványa, Ferdinand Löwe elindította azt a ma is élő hagyományt, hogy a zenekar Bruckner műveket mutat be. (4)

After the war, the orchestra was taken over by the city of Munich and restarted under the leadership of composer Hans Pfitzner, soon replaced by Bruckner pioneer Siegmund von Hausegger.

- A háború után München városa vette át a zenekart és újraindította azt a zeneszerző Hans Pfitzner vezetésével, akit hamarosan a Bruckner

nyomdokain járó Siegmund von Hausegger követett. 1928-ban a zenekar megkapta mai nevét, mint a Müncheni Filharmonikusok. (22)

After the rise of the Nazi party in 1933, the orchestra stamped its scores with swastikas and the words "The Orchestra of the Fascist Movement".

- A náci párt 1933-as hatalomra kerülésével, a zenekar kottáit horogkeresztes pecséttel látta el és e szavakkal: "a fasiszta mozgalom zenekara". (18)

6. Upper 10 (3)
The Upper 10 brand debuted in 1933 as a product of the Nehi Corporation (later Royal Crown Corporation).
   - Az Upper 10 márka 1933-ban került piacra, mint a Nehi Corporation (később Royal Crown Corporation) terméke. (25)

   Upper 10 - RC Cola International home site

   - Az Upper 10 az RC Cola cég történetének egyik zászlóshajója lett. (0)

   Upper 10 is still sold outside of North America by Cott Beverages, the same company that sells RC Cola internationally.

   - (a 7 Up-ot szintén a Dr Pepper cégcsoport tulajdonolja) Az Upper 10-et azonban Észak-Amerikán kívül a Cott Beverages, az RC Cola termékeinek nemzetközi forgalmazója továbbra is árusítja. (10)
7. Meitetsu 5000 series (2008) (3)
As with the earlier 3150 series, individual seats are 470 mm wide.
   - Mint a korábbi 3150-es sorozatnál, az egyes ülések 470 mm szélesek, kék mokett (plüss-szerű szövet) kárpitozással. (7)

   The seats use a cantilever structure, i.e. attached to the walls, and in each car there is one folding seat.

   - Az ülések konzol-struktúrával készültek, vagyis a falra erősítettek, és minden egyes kocsiban egy összecsukható ülés is található. (11)

   Wheelchair spaces are located behind the driver's cab in both cars on each end of the train.

   - Kerekesszékkel utazóknak fenntartott hely a vonat mindkét végén az utolsó kocsiban, a vezetőfülke mögött található. (16)
8. Mike Seeger (3)
Seeger was born in New York and grew up in Maryland and Washington D.C. His father, Charles Louis Seeger Jr., was a composer and pioneering ethnomusicologist, investigating both American folk and non-Western music.
   - Apja, ifj. Charles Louis Seeger, zeneszerző és zenetudós volt, aki az elsők között vizsgálta mind az amerikai népzenét, mind a nem-európai gyökerekből fakadó zenét. (4)

His next older half brother is Pete Seeger.

- Legidősebb bátyja, Charles Seeger III (? (3)

His sister Peggy Seeger, also a well-known folk performer, was married for many years to British folk singer Ewan MacColl.

- Féltestvére, Peggy Seeger, aki szintén ismert népzenei előadó volt, hosszú évekig Ewan MacColl brit folkénekessel élt házasságban. (18)

9. Avidius Cassius (2)
At first, according to Cassius Dio, Marcus, who was on campaign against tribes in the north, tried to keep the rebellion a secret from his soldiers, but after the news had spread among them, he addressed them.
   - Kezdetben, Cassius Dio írásaira támaszkodva, Marcus, aki éppen az északi törzsek ellen folytatott hadjáratot, próbálta a lázadás hírét eltitkolni katonái elől, de miután a hír terjedni kezdett, ő maga mondta ezt el nekik. (28)

   In this speech that Dio attributes to Marcus, he laments the disloyalty of "a dearest friend", while at the same time expressing his hope that Cassius would not be killed or commit suicide, so that he could show mercy.

   - A beszédben, melyet Dio Marcusnak tulajdonít, a császár "egy legdrágább barát" hűtlenségét fájlalja, miközben azon reményét is kifejezi, hogy Cassiust nem ölik meg, és önkezével sem vet véget életének, hiszen így ő majd könyörületet mutathat. (13)
10. Foreign relations of Pakistan (2)
Pakistan is an active member of the United Nations.
   - Pakisztán az Egyesült Nemzetek aktív tagja. (15)

   To this day, Pakistan has a close relationship with Turkey.

   - Pakisztán a mai napig szoros kapcsolatokat ápol Törökországgal. (11)
11. Politics of Pakistan (2)
It is also an active member of the United Nations.
   - Pakisztán az Egyesült Nemzetek aktív tagja. (10)

   To this day, Pakistan has a close relationship with Turkey.

   - Pakisztán a mai napig szoros kapcsolatokat ápol Törökországgal. (11)
12. GPS tracking server (2)
* Wialon — GPS tracking software platform with some fleet management features, developed by a Belarusian company Gurtam.
   - A Wialon egy internet-alapú GPS nyomkövető szoftverplatform, néhány flottakezelési tulajdonsággal, amit a fehérorosz Gurtam cég fejlesztett ki. (4)

   Wialon is available for both Linux and Windows operating systems.

- A Wialon mind Linux mint Windows operációs rendszerre elérhető. (16)

13. Web accessibility (2)
    For example, when a site is coded with semantically meaningful HTML, with textual equivalents provided for images and with links named meaningfully, this helps blind users using text-to-speech software and/or text-to-Braille hardware.
    - Például, ha az oldal szemantikai jelentéssel bíró HTML-ben van kódolva, a képekhez, fotókhoz megfelelő szöveges megjelenést biztosít, és a linkeknek értelmes elnevezést adtak, ez segíti a látássérült felhasználókat, hogy az oldalt felolvasó-szoftverrel és/vagy szöveget Braille nyelvvé alakító hardverrel használják. (15)

    When pages are coded so that users can navigate by means of the keyboard alone, or a single switch access device alone, this helps users who cannot use a mouse or even a standard keyboard.

    - Amikor az oldalakat oly módon kódolják, hogy azon csupán a billentyűzet vagy egy egyszerű kapcsolóeszköz segítségével lehet navigálni, ezzel az egeret vagy a normál billentyűzetet használni egyáltalán nem képes felhasználókat segítik. (4)

14. Portal:University of Oxford/Intro (2)
    Although the exact date of foundation remains unclear, there is evidence of teaching there as far back as the 11th century.
    - Bár alapításának pontos időpontja nem ismert, bizonyított, hogy már a 11. században folyt itt tanítás. (15)

    After disputes between students and Oxford townsfolk in 1209, some academics fled north-east to Cambridge, where they established what became the University of Cambridge.

    - 1209-ben, a diákok és az oxfordi városlakók közti viszály következtében néhány egyetemi tanár északkeletre, Cambridge-be menekült, ahol megalapították a ma Cambridge-ként ismert egyetemet. (2)

15. Castle Rock (2)
    *Castle Rock, West Virginia: Wyoming County, inside the small town of Pineville
    - A Castle Rock egy geológiai képződmény Pineville-ben, Nyugat-Virginiában. (9)

    *Castle Rock, Vermont: Also known as Castle Rock Peak.

    - A Castle Rock a vidék első felfedezői számára egyszerűen mint a"kastély" volt ismert. (4)

16. MOPAC (1)
    MOPAC2007 is available for both Windows and Linux operating systems.
    - A Wialon mind Linux mint Windows operációs rendszerre elérhető. (11)

17. Kosmo (1)
    It is available for Windows and Linux operating systems.
    - A Wialon mind Linux mint Windows operációs rendszerre elérhető. (11)

18. VTune (1)
    It is available for both Linux and Microsoft Windows operating systems.
    - A Wialon mind Linux mint Windows operációs rendszerre elérhető. (11)
19. List of office suites (1)
    Available for Windows / Linux operating systems
    - A Wialon mind Linux mint Windows operációs rendszerre elérhető. (11)
20. Pakistan (1)
    The country is an active member of the United Nations.
    - Pakisztán az Egyesült Nemzetek aktív tagja. (10)
21. LMMS (1)
    LMMS is available for the Linux, OpenBSD, and Microsoft Windows operating systems.
    - A Wialon mind Linux mint Windows operációs rendszerre elérhető. (9)
22. FEMtools (1)
    The program is available on on Microsoft Windows, Linux and Mac operating systems.
    - A Wialon mind Linux mint Windows operációs rendszerre elérhető. (9)
23. Nautilus (secure telephone) (1)
    It runs from a command line and is available for the Linux and Windows operating systems.
    - A Wialon mind Linux mint Windows operációs rendszerre elérhető. (9)
24. Monolith (1)
    * Castle Rock, Pineville, West Virginia
    - A Castle Rock egy geológiai képződmény Pineville-ben, Nyugat-Virginiában. (9)
25. Charles Seeger (1)
    His first wife was Constance de Clyver Edson, a classical violinist and teacher; they divorced in 1927.
    - Édesanyja, Constance de Clyver Edson klasszikus hegedűművész és tanár volt. (8)

9.11. Hasonlósági metrikán alapuló algoritmus találati listája a 12 Wikipédia cikk angol visszafordítására

9.11.1. Google Translate fordítóval

1. Pete Seeger (7)
   Seeger was born in French Hospital, Midtown Manhattan, the youngest of three sons.
   - Pete Seeger middle of Manhattan, has also called the Midtown district born in hospitals in France. (3)

   His mother, Constance de Clyver Edson, was a violinist and teacher, raised in Tunisia and trained at the Paris Conservatory of Music and the Juilliard School.

   - His mother, Constance de Clyver Edson classical violinist and teacher. (3)

His stepmother, Ruth Crawford Seeger, was one of the most significant female composers of the 20th century.

- Her stepmother, Ruth Crawford Seeger was one of the most prominent twentieth-century women composers. (18)

His eldest brother, Charles Seeger III, was a radio astronomer, and his next older brother, John Seeger, taught in the 1950s at the Dalton School in Manhattan and was the principal from 1960 to 1976 at Fieldston Lower School in the Bronx.

-), Radio astronomer, another brother, John Seeger in the 1950s, the Dalton School in Manhattan, i was taught, and from 1960 to 1976, the Bronx Fieldston Lower School Director was. (29)

His uncle, Alan Seeger, a noted poet, was killed during the First World War.

- His uncle, Alan Seeger, who was a famous poet, was killed in World War II. (20)

His half-sister, Peggy Seeger, also a well-known folk performer, was married for many years to British folk singer Ewan MacColl.

- Half-sister, Peggy Seeger, who is also known folk artist was for many years British folk singer Ewan MacColl lived marriage. (31)

Half-brother Mike Seeger went on to form the New Lost City Ramblers, one of whose members, John Cohen, was married to Pete's other half-sister, singer Penny Seeger, also a highly talented singer.

- Another half-brother Mike Seeger founded the New Lost City Ramblers, t (a string band playing traditional music style), with one member, John Cohen, Pete's half-brother, the singer married Penny Seeger, and himself a very talented singer. (35)

2. Foreign relations of Pakistan (5)
Pakistan is an active member of the United Nations.
   - Pakistan is an active member of the United Nations. (15)

Its alliance with the United States was especially close after the Soviets invaded the neighboring country of Afghanistan.

- The United States, especially the close relationship as allies became after the Soviets invaded the neighboring Afghanistan. (28)

In 1964, Pakistan signed the Regional Cooperation for Development (RCD) Pact with Turkey and Iran, when all three countries were closely allied with the U.S., and as neighbors of the Soviet Union, wary of perceived Soviet expansionism.

- In 1964, Pakistan, Turkey and Iran together signed the Regional Cooperation for Development Pact (Regional Cooperation for

Development Pact), under which these three countries have close ties with companies in the United States and the Soviet Union's neighboring states, attempt to stop the apparent Soviet expansionism. (36)

To this day, Pakistan has a close relationship with Turkey.

- Pakistan is still a close relationship with Turkey. (10)

RCD became defunct after the Iranian Revolution, and a Pakistani-Turkish initiative led to the founding of the Economic Cooperation Organisation (ECO) in 1985.

- The above security pact after the Iranian uprising was repealed, however, a Pakistani-Turkish initiative led to the ECO (Economic Cooperation Organization) in 1985, the foundation. (24)

3. Politics of Pakistan (5)
   It is also an active member of the United Nations.
   - Pakistan is an active member of the United Nations. (10)

   Its alliance with the United States was especially close after the Soviets invaded the neighbouring country of Afghanistan.

   - The United States, especially the close relationship as allies became after the Soviets invaded the neighboring Afghanistan. (23)

   In 1964, Pakistan signed the Regional Cooperation for Development (RCD) Pact with Turkey and Iran, when all three countries were closely allied with the U.S., and as neighbours of the Soviet Union, wary of perceived Soviet expansionism.

   - In 1964, Pakistan, Turkey and Iran together signed the Regional Cooperation for Development Pact (Regional Cooperation for Development Pact), under which these three countries have close ties with companies in the United States and the Soviet Union's neighboring states, attempt to stop the apparent Soviet expansionism. (31)

   To this day, Pakistan has a close relationship with Turkey.

   - Pakistan is still a close relationship with Turkey. (10)

   RCD became defunct after the Iranian Revolution, and a Pakistani-Turkish initiative led to the founding of the Economic Cooperation Organisation (ECO) in 1985.

   - The above security pact after the Iranian uprising was repealed, however, a Pakistani-Turkish initiative led to the ECO (Economic Cooperation Organization) in 1985, the foundation. (24)

4. Web accessibility (5)
   When sites are correctly designed, developed and edited, all users can have equal access to information and functionality.

- When a website properly planned, developed and edited, all users are equally available to all of the information contained therein and functionality. (14)

For example, when a site is coded with semantically meaningful HTML, with textual equivalents provided for images and with links named meaningfully, this helps blind users using text-to-speech software and/or text-to-Braille hardware.

- For example, if the page is a semantic meaning in HTML encoded images, photographs corresponding text appearance, and for links to a meaningful name given, this will help visually impaired users to the page reading software and / or text to Braille language shaping hardware is used. (19)

When pages are coded so that users can navigate by means of the keyboard alone, or a single switch access device alone, this helps users who cannot use a mouse or even a standard keyboard.

- When the pages are encoded in a way that's just a keyboard or a simple switching tool to navigate, with the mouse or the keyboard to normal use is not at all able to help users. (4)

When videos are closed captioned or a sign language version is available, deaf and hard-of-hearing users can understand the video.

- The video subtitling or sign language versions of them to ensure the deaf and hard of hearing users to ensure that they understand the content. (22)

And when content is written in plain language and illustrated with instructional diagrams and animations, users with dyslexia and learning difficulties are better able to understand the content.

- If the content is simple sentences to describe what diagrams, animations also illustrate, the dyslexiás or learning difficulties to help users better understand the content. (16)

5. Munich Philharmonic (4)
Felix Weingartner was music director from 1898 to 1905, and the young Wilhelm Furtwängler made his auspicious conducting debut there in 1906.
   - Between 1898 and 1905, Felix Weingartner was music director in 1906 and the young Wilhelm Furtwängler's successful debut as a conductor. (17)

Meanwhile Anton Bruckner pupil Ferdinand Löwe established an enduring tradition of Bruckner performance which continues to this day.

- Meanwhile, Anton Bruckner pupil Ferdinand Löwe started the still living tradition of the orchestra presents works of Bruckner. (14)

After the war, the orchestra was taken over by the city of Munich and restarted under the leadership of composer Hans Pfitzner, soon replaced by Bruckner pioneer Siegmund von Hausegger.

- After the war, the city of Munich took over the band and I restarted the composer Hans Pfitzner leadership, who will soon be following in the footsteps of the Bruckner going Siegmund von Hausegger followed. (35)

After the rise of the Nazi party in 1933, the orchestra stamped its scores with swastikas and the words "The Orchestra of the Fascist Movement".

- The Nazi party came to power in 1933, the band saw it sealed scores swastika and the words, "bands of the fascist movement. (17)

6. Wialon (4)
Wialon is a web-based GPS tracking software platform with some fleet management features, developed by a Belarusian company Gurtam.
   - The Wialon an Internet-based GPS tracking software platform, some fleet management features, which the company developed in Belarus belts. (30)

   Wialon is available for both Linux and Windows operating systems.

   - The Wialon Linux and Windows operating system as available. (16)

   Wialon is notable because of its compatibility with different GLONASS and GPS tracking units, counting 131 in the beginning of July 2010.

   - The Wialon is noteworthy because different GLONASS and GPS is also compatible with following units, their number reached 131 in July 2010, et. (19)

   According to Gurtam website, there are about 200 GPS tracking services worldwide who have already implemented Wialon GPS tracking platform.

   - The website, the belts around the world 200 GPS tracking service is used to Wialon GPS tracking platform. (13)

7. Avidius Cassius (4)
At first, according to Cassius Dio, Marcus, who was on campaign against tribes in the north, tried to keep the rebellion a secret from his soldiers, but after the news had spread among them, he addressed them.
   - At first, relying on provisions of Dio Cassius, Marcus, who was the campaign against the northern tribes, trying to conceal news of the revolt from his soldiers, but after the news began to spread, he said this to them. (19)

   In this speech that Dio attributes to Marcus, he laments the disloyalty of "a dearest friend", while at the same time expressing his hope that Cassius would not be killed or commit suicide, so that he could show mercy.

- The speech, which was attributed to Marcus Dio, the Emperor "is a most precious friend" regrets infidelity, while also expressing the hope that Cassius is not killed, and önkezével not put an end to his life in a way that he will show compassion. (24)

In the meantime, the Senate declared Cassius a public enemy.

- The Senate declared public enemy Cassius. (13)

It was in Egypt that Cassius made his base of operations, and it is known that Cassius was recognized as emperor there by May 3, since a document of that date is recorded as being in the first year of Cassius's reign.

- There is evidence that Cassius was proclaimed emperor on May 3, dated as such a document recorded as Cassius records from the early years of his reign. 175 rebellion started in April. (5)

8. Mike Seeger (4)
Seeger was born in New York and grew up in Maryland and Washington D.C. His father, Charles Louis Seeger Jr., was a composer and pioneering ethnomusicologist, investigating both American folk and non-Western music.
   - Charles Louis Seeger, composer and musicologist who was among the first to examine both the American folk music and the music arising from non-European roots. (4)

   His eldest half-brother, Charles Seeger III, was a radio astronomer, and his next older half-brother, John Seeger, taught for years at the Dalton School in Manhattan.

   -), Radio astronomer, another brother, John Seeger in the 1950s, the Dalton School in Manhattan, i was taught, and from 1960 to 1976, the Bronx Fieldston Lower School Director was. (9)

   His uncle, Alan Seeger, a poet, was killed during the First World War.

   - His uncle, Alan Seeger, who was a famous poet, was killed in World War II. (15)

   His sister Peggy Seeger, also a well-known folk performer, was married for many years to British folk singer Ewan MacColl.

   - Half-sister, Peggy Seeger, who is also known folk artist was for many years British folk singer Ewan MacColl lived marriage. (28)
9. Golden Twenties (3)
It is often applied to Germany, which during the early 1920s, experienced, like most of Europe, record-breaking levels of inflation of one trillion percent between January 1919 and November 1923.
   - Germany is often directly related to the use of this phrase, which, similarly to most European countries, the inflation rate reached one trillion percent between November 1919 and January 1923. (22)

Such events, among many other factors, triggered the rise of fascism in Italy, as well as the ill-fated Beer Hall Putsch, masterminded by a young Adolf Hitler.

- In such situations, while many other factors, contributed to the spread of fascism in Italy, as may the infamous Munich beer coup led by a young Adolf Hitler. (15)

Before long, the Weimar Republic under Chancellor Gustav Stresemann managed to tame the extreme levels of inflation by the introduction of a new currency, the Rentenmark, with tighter fiscal controls and reduction of bureaucracy, leading to a relative degree of political and economic stability.

- Not much later, however, the Weimar Republic of the leadership of Gustav Stresemann, Chancellor of inflation reined in a new currency, the Rentenmark introduction of stricter financial regulation and cutting red tape, so the relative political and economic stability created. (21)

10. Upper 10 (3)
The Upper 10 brand debuted in 1933 as a product of the Nehi Corporation (later Royal Crown Corporation).
    - Upper 10 brand was launched in 1933 as the Nehi Corporation (later Royal Crown Corporation) products. (30)

    U.S. Patent and Trademark Office Upper 10 was one of RC Cola's flagship brands throughout the company's history.

    - Upper 10 on the RC Cola company's history became one of the flagships. (7)

    Upper 10 is still sold outside of North America by Cott Beverages, the same company that sells RC Cola internationally.

    - (7-Up was also owned by Dr Pepper Group) Upper 10 in North America but also in Cott Beverages, the RC Cola products in the international distributor continues to sell. (13)
11. GPS tracking server (2)
* Wialon — GPS tracking software platform with some fleet management features, developed by a Belarusian company Gurtam.
    - The Wialon an Internet-based GPS tracking software platform, some fleet management features, which the company developed in Belarus belts. (29)

    Wialon is available for both Linux and Windows operating systems.

    - The Wialon Linux and Windows operating system as available. (16)
12. Castle Rock (Pineville, West Virginia) (2)
Castle Rock is a geological feature located in Pineville, West Virginia next to the Pineville Public Library.
    - The Castle Rock is a geological formation in Pineville, West Virginia. (6)

At one time ladders provided access to the top of the rock, but they were removed in 1911, after Virgil Senter fell to his death.

- Ladders at the same time assured the ascent of the cliff top, but removed them in 1911, after Virgil Senter dropped and smashed itself to death. (18)

13. Meitetsu 5000 series (2008) (2)
    The seats use a cantilever structure, i.e. attached to the walls, and in each car there is one folding seat.
    - The seats with console-made structure, that is attached to the wall, and each car is also a folding seat. (11)

    Wheelchair spaces are located behind the driver's cab in both cars on each end of the train.

    - Wheelchair travelers reserved space at each end of the last car, behind the cab. (8)
14. Charles Seeger (2)
    His first wife was Constance de Clyver Edson, a classical violinist and teacher; they divorced in 1927.
    - His mother, Constance de Clyver Edson classical violinist and teacher. (11)

    * Charles Seeger entry - at NNDB.com

    - Eldest brother, Charles Seeger III (? (0)
15. University of Oxford (1)
    After disputes between students and Oxford townsfolk in 1209, some academics fled north-east to Cambridge, where they established what became the University of Cambridge.
    - In 1209, students and townspeople in Oxford due to some conflict between the north-east university professor, fled to Cambridge, where they founded the now known as Cambridge University. (32)
16. Portal:University of Oxford/Intro (1)
    After disputes between students and Oxford townsfolk in 1209, some academics fled north-east to Cambridge, where they established what became the University of Cambridge.
    - In 1209, students and townspeople in Oxford due to some conflict between the north-east university professor, fled to Cambridge, where they founded the now known as Cambridge University. (32)
17. Portal:South East England/Selected article/9 (1)
    According to legend, after riots between students and townsfolk broke out in 1209, some of the academics at Oxford fled north-east to the town of Cambridge, where the University of Cambridge was founded.
    - In 1209, students and townspeople in Oxford due to some conflict between the north-east university professor, fled to Cambridge, where they founded the now known as Cambridge University. (23)
18. Portal:Oxfordshire/Selected article/8 (1)
    According to legend, after riots between students and townsfolk broke out in

1209, some of the academics at Oxford fled north-east to the town of Cambridge, where the University of Cambridge was founded.
    - In 1209, students and townspeople in Oxford due to some conflict between the north-east university professor, fled to Cambridge, where they founded the now known as Cambridge University. (23)
19. Portal:University/Previous articles (1)
    The University of Cambridge is the second-oldest university in the English-speaking world.
    - This is the second oldest, still functioning university, and is the oldest university in the English-speaking world. (14)
20. Kosmo (1)
    It is available for Windows and Linux operating systems.
    - The Wialon Linux and Windows operating system as available. (13)
21. List of office suites (1)
    Available for Windows / Linux operating systems
    - The Wialon Linux and Windows operating system as available. (13)
22. VTune (1)
    It is available for both Linux and Microsoft Windows operating systems.
    - The Wialon Linux and Windows operating system as available. (11)
23. Monolith (1)
    * Castle Rock, Pineville, West Virginia
    - The Castle Rock is a geological formation in Pineville, West Virginia. (11)
24. Shared web hosting service (1)
    For example, the Plesk control panel is available for two operating systems, Linux and Windows.
    - The Wialon Linux and Windows operating system as available. (10)
25. Hazel Dickens (1)
    She met Mike Seeger, younger half-brother of Pete Seeger and founding member of the New Lost City Ramblers and became active in the Baltimore-Washington area bluegrass and folk music scene during the 1960s.
    - Another half-brother Mike Seeger founded the New Lost City Ramblers, t (a string band playing traditional music style), with one member, John Cohen, Pete's half-brother, the singer married Penny Seeger, and himself a very talented singer. (10)
26. Pakistan (1)
    The country is an active member of the United Nations.
    - Pakistan is an active member of the United Nations. (10)
27. Foreign relations of the United Kingdom (1)
    Both UK and Pakistan are active members of the Commonwealth of Nations.
    - Pakistan is an active member of the United Nations. (10)
28. Castle Rock (1)
    *Castle Rock, West Virginia: Wyoming County, inside the small town of Pineville
    - The Castle Rock is a geological formation in Pineville, West Virginia. (10)
29. Osama Ali Maher (1)
    He has worked for the United Nations in Darfur and in Pakistan.
    - Pakistan is an active member of the United Nations. (10)

30. Nautilus (secure telephone) (1)
    It runs from a command line and is available for the Linux and Windows operating systems.
    - The Wialon Linux and Windows operating system as available. (9)
31. MINTO (1)
    It can run on both Linux and Windows operating system.
    - The Wialon Linux and Windows operating system as available. (8)
32. VariCAD (1)
    It runs on the Windows and Linux operating systems.
    - The Wialon Linux and Windows operating system as available. (8)
33. Movial (1)
    It is available on Windows, Mac OS, Linux, Symbian S60 and Windows Mobile operating systems.
    - The Wialon Linux and Windows operating system as available. (8)

9.11.2. Microsoft Bing fordítóval

1. Pete Seeger (6)
   Seeger was born in French Hospital, Midtown Manhattan, the youngest of three sons.
   - Pete Seeger in the middle of Midtown Manhattan, also called in the French suburbs, was born in 2004. (3)

   His mother, Constance de Clyver Edson, was a violinist and teacher, raised in Tunisia and trained at the Paris Conservatory of Music and the Juilliard School.

   - His mother, Constance de Clyver was a classical violinist and teacher Edson. (3)

   His stepmother, Ruth Crawford Seeger, was one of the most significant female composers of the 20th century.

   - Ruth Crawford Seeger, his was one of the most prominent women in the twentieth century, zeneszerzőinek. (8)

   His eldest brother, Charles Seeger III, was a radio astronomer, and his next older brother, John Seeger, taught in the 1950s at the Dalton School in Manhattan and was the principal from 1960 to 1976 at Fieldston Lower School in the Bronx.

   -), radio astronomer, was another brother, John Seeger in the 1950s, the Dalton School in manhattan-i taught, and then in 1976 1960-től bronx Fieldston Lower School became the Director. (29)

   His half-sister, Peggy Seeger, also a well-known folk performer, was married for many years to British folk singer Ewan MacColl.

   - His half-sister, Peggy Seeger, who is also known as the folk singer Ewan MacColl was long years she lived in wedlock, British folkénekessel. (26)

Half-brother Mike Seeger went on to form the New Lost City Ramblers, one of whose members, John Cohen, was married to Pete's other half-sister, singer Penny Seeger, also a highly talented singer.

- Another half-brother, Mike Seeger founded the New Lost City Ramblers (a traditional musical style which, as vonósegyüttest), which is a member of John Cohen, the singer Penny féltestvérét Pete Seegert, and itself has been married to a very talented singer. (33)

2. Wialon (4)
Wialon is a web-based GPS tracking software platform with some fleet management features, developed by a Belarusian company Gurtam.
   - In an internet-based GPS Wialon tracking software platform, some of the flottakezelési property, which is developed by the Belarusian company Gurtam. (27)

   Wialon is available for both Linux and Windows operating systems.

   - The both Linux as Windows Wialon operating system available. (21)

   Wialon is notable because of its compatibility with different GLONASS and GPS tracking units, counting 131 in the beginning of July 2010.

   - This is notable because Wialon different GLONASS and GPS, the number of these követőegységekkel also compatible with 2010 reached 131-et in July. (14)

   According to Gurtam website, there are about 200 GPS tracking services worldwide who have already implemented Wialon GPS tracking platform.

   - According to the website of Gurtam around 200 GPS tracking service around the world use Wialon GPS tracking is already in the platform. (23)
3. Golden Twenties (4)
It is often applied to Germany, which during the early 1920s, experienced, like most of Europe, record-breaking levels of inflation of one trillion percent between January 1919 and November 1923.
   - Often used in relation to Germany directly to this phrase, like most of Europe, where new, inflation reached egybillió per cent between January and November, 1919, 1923. (17)

   The inflation was so severe that printed currency was often used for heating and other uses, and everyday requirements like food, soap, electricity cost a wheelbarrow full of banknotes.

   - Inflation is a serious, was that in many cases, the papírpénzzel begyújtottak was used to heat their homes or for any other purpose, or the everyday needs such as food, SOAP, or a bank to pay the electric talicskányi was needed. (5)

Such events, among many other factors, triggered the rise of fascism in Italy, as well as the ill-fated Beer Hall Putsch, masterminded by a young Adolf Hitler.

- In such situations, in addition to many other factors, have promoted fascism in Italy in General, as well as the infamous Munich sörpuccsot directed by a young Adolf Hitler. (15)

Before long, the Weimar Republic under Chancellor Gustav Stresemann managed to tame the extreme levels of inflation by the introduction of a new currency, the Rentenmark, with tighter fiscal controls and reduction of bureaucracy, leading to a relative degree of political and economic stability.

- Shortly afterwards, however, the Weimar Republic under the leadership of Chancellor Gustav Stresemann, a new currency and inflation in megfékezte, with the introduction of the Rentenmark, regulations and tighter financial red tape reduction, thereby making the relative political and economic stability. (41)

4. Web accessibility (4)
Web accessibility refers to the inclusive practice of making websites usable by people of all abilities and disabilities.
   - Web Accessibility refers to the practice of comprehensive when the Web pages for all users, any handicapped people intact or can use, make available. (10)

When sites are correctly designed, developed and edited, all users can have equal access to information and functionality.

- When you develop a website developed by and in accordance with the planned, for all users are equally available to all the information and functionality. (13)

For example, when a site is coded with semantically meaningful HTML, with textual equivalents provided for images and with links named meaningfully, this helps blind users using text-to-speech software and/or text-to-Braille hardware.

- For example, if the page is semantic meaning is encoded in HTML, text, pictures, gallerys, and corresponds to the appearance of the links provide a meaningful name should have been added, this will help partially sighted users to the page reading software and/or hardware used for shaping the language text for a Braille. (25)

When pages are coded so that users can navigate by means of the keyboard alone, or a single switch access device alone, this helps users who cannot use a mouse or even a standard keyboard.

- When the pages are coded in such a way that only the keyboard or you can use a simple switching device to navigate, you can use this mouse or use the normal keyboard users may not be able to help at all. (19)

5. Foreign relations of Pakistan (4)
   Pakistan is an active member of the United Nations.
   - Pakistan is an active member of the United Nations. (15)

   In 1964, Pakistan signed the Regional Cooperation for Development (RCD) Pact with Turkey and Iran, when all three countries were closely allied with the U.S., and as neighbors of the Soviet Union, wary of perceived Soviet expansionism.

   - An ally of the United States, and in particular relations with the EU as it became after the Soviet press upon the neighboring Afghanistan. in 1964, together with Iran and Turkey, Pakistan signed the Pact on Regional cooperation for development (Regional Cooperation for Development Pact), according to which of these three countries undertake a close ties with the United States and the Soviet Union, as they seek to stop the szomszédállamai, shows Soviet expansion. (7)

   To this day, Pakistan has a close relationship with Turkey.

   - Pakistan maintains close links with the present day Turkey. (6)

   RCD became defunct after the Iranian Revolution, and a Pakistani-Turkish initiative led to the founding of the Economic Cooperation Organisation (ECO) in 1985.

   - The Pact was referred to above was repealed after Iran, but a Pakistani-Turkish initiative led to the ECO (Economic Cooperation Organization) in 1985 for the launching. (23)
6. Upper 10 (3)
   The Upper 10 brand debuted in 1933 as a product of the Nehi Corporation (later Royal Crown Corporation).
   - In 1933, the Upper 10 brand has been on the market, such as the Nehi Corporation (later Royal Crown Corporation). (25)

   Upper 10 - RC Cola International home site

   - The Upper 10 is one of the RC Cola company's flagship. (4)

   Upper 10 is still sold outside of North America by Cott Beverages, the same company that sells RC Cola internationally.

   - (the 7 Up also in the Dr Pepper company Smith) However, the Upper 10 outside of North America, the Cott Beverages, the RC Cola international distributor continues to sell its products. (21)
7. Castle Rock (Pineville, West Virginia) (3)
   Castle Rock is a geological feature located in Pineville, West Virginia next to the Pineville Public Library.
   - The Castle Rock is a geological formation in Pineville, West Virginia. (6)

Named for its resemblance to a castle, it rises about 200 feet above Rock Castle Creek, a branch of the Guyandotte River.

- Name of the castle-like look, and approximately 65 metres above the sampler of Rock Castle Creek, one of the arms of the Guyandotte River. (14)

At one time ladders provided access to the top of the rock, but they were removed in 1911, after Virgil Senter fell to his death.

- At the same time ladders to the top of the rock, were granted a promotion, but they were removed in 1911, after Virgil Senter fell and were crushed to death. (25)

8. Politics of Pakistan (3)
It is also an active member of the United Nations.
   - Pakistan is an active member of the United Nations. (10)

To this day, Pakistan has a close relationship with Turkey.

- Pakistan maintains close links with the present day Turkey. (6)

RCD became defunct after the Iranian Revolution, and a Pakistani-Turkish initiative led to the founding of the Economic Cooperation Organisation (ECO) in 1985.

- The Pact was referred to above was repealed after Iran, but a Pakistani-Turkish initiative led to the ECO (Economic Cooperation Organization) in 1985 for the launching. (23)

9. Mike Seeger (3)
His eldest half-brother, Charles Seeger III, was a radio astronomer, and his next older half-brother, John Seeger, taught for years at the Dalton School in Manhattan.
   -), radio astronomer, was another brother, John Seeger in the 1950s, the Dalton School in manhattan-i taught, and then in 1976 1960-től bronx Fieldston Lower School became the Director. (3)

His sister Peggy Seeger, also a well-known folk performer, was married for many years to British folk singer Ewan MacColl.

- His half-sister, Peggy Seeger, who is also known as the folk singer Ewan MacColl was long years she lived in wedlock, British folkénekessel. (21)

His sister, singer Penny Seeger, married John Cohen, a member of Mike's musical group, New Lost City Ramblers.

- Another half-brother, Mike Seeger founded the New Lost City Ramblers (a traditional musical style which, as vonósegyüttest), which is a member of John Cohen, the singer Penny féltestvérét Pete Seegert, and itself has been married to a very talented singer. (13)

10. GPS tracking server (2)
    * Wialon — GPS tracking software platform with some fleet management features, developed by a Belarusian company Gurtam.
        - In an internet-based GPS Wialon tracking software platform, some of the flottakezelési property, which is developed by the Belarusian company Gurtam. (22)
        - According to the website of Gurtam around 200 GPS tracking service around the world use Wialon GPS tracking is already in the platform. (4)

    Wialon is available for both Linux and Windows operating systems.

        - The both Linux as Windows Wialon operating system available. (21)
11. Munich Philharmonic (2)
    After the war, the orchestra was taken over by the city of Munich and restarted under the leadership of composer Hans Pfitzner, soon replaced by Bruckner pioneer Siegmund von Hausegger.
        - After the war, the city of Munich was taken over by the band and restarted under the leadership of composer Hans Pfitzner I, who will soon be the Bruckner of Siegmund von Hausegger's example. in 1928, the Orchestra has received today's name, such as the Munich Philharmonic Orchestra. (32)

    After the rise of the Nazi party in 1933, the orchestra stamped its scores with swastikas and the words "The Orchestra of the Fascist Movement".

        - The Nazi party came to power in 1933, the Orchestra's passage kottáit horogkeresztes affixing its stamp and (e), the words: "the fascist movement Orchestra". (13)
12. Meitetsu 5000 series (2008) (2)
    The seats use a cantilever structure, i.e. attached to the walls, and in each car there is one folding seat.
        - The meetings were structured, i.e. for console on the wall, and each confirmed case of a folding seat in the car. (8)

    Wheelchair spaces are located behind the driver's cab in both cars on each end of the train.

        - Wheelchair space reserved for travellers on both ends of the train, the last car is located behind the cab. (18)
13. Charles Seeger (2)
    His first wife was Constance de Clyver Edson, a classical violinist and teacher; they divorced in 1927.
        - His mother, Constance de Clyver was a classical violinist and teacher Edson. (11)

    * Charles Seeger entry - at NNDB.com

        - The eldest brother, Charles Seeger III (? (0)

14. Seeger (2)
    :(ii) Ruth Crawford Seeger (1901 - 1953), a modernist composer and an American folk music specialist; second wife of Charles; 4 children
        - Charles Louis Seeger, composer and zenetudós was among the first who examined both the American folk music and the music of non-European origins. (1)

    ::* Charles Seeger III, (1912-2002) astronomer

        - The eldest brother, Charles Seeger III (? (3)
15. VTune (1)
    It is available for both Linux and Microsoft Windows operating systems.
        - The both Linux as Windows Wialon operating system available. (16)
16. MOPAC (1)
    MOPAC2007 is available for both Windows and Linux operating systems.
        - The both Linux as Windows Wialon operating system available. (14)
17. Avidius Cassius (1)
    At first, according to Cassius Dio, Marcus, who was on campaign against tribes in the north, tried to keep the rebellion a secret from his soldiers, but after the news had spread among them, he addressed them.
        - Initially, Cassius Dio, Marcus, the Aquarium who is currently based in the northern tribes against the campaign, trying to hide from the soldiers a reputation as the rebellion, but after the news began to spread, he himself said this to them. (14)
18. Portal:South East England/Selected article/9 (1)
    According to legend, after riots between students and townsfolk broke out in 1209, some of the academics at Oxford fled north-east to the town of Cambridge, where the University of Cambridge was founded.
        - 1209-ben, students and as a result of hostilities between the towns of Oxford, some university lecturers to the North, Cambridge, where he was granted refugee status in the University, known as the Cambridge, ma. (13)
19. Portal:Oxfordshire/Selected article/8 (1)
    According to legend, after riots between students and townsfolk broke out in 1209, some of the academics at Oxford fled north-east to the town of Cambridge, where the University of Cambridge was founded.
        - 1209-ben, students and as a result of hostilities between the towns of Oxford, some university lecturers to the North, Cambridge, where he was granted refugee status in the University, known as the Cambridge, ma. (13)
20. Kosmo (1)
    It is available for Windows and Linux operating systems.
        - The both Linux as Windows Wialon operating system available. (11)
21. MINTO (1)
    It can run on both Linux and Windows operating system.
        - The both Linux as Windows Wialon operating system available. (11)
22. Monolith (1)
    * Castle Rock, Pineville, West Virginia

- The Castle Rock is a geological formation in Pineville, West Virginia. (11)

23. List of office suites (1)
    Available for Windows / Linux operating systems
    - The both Linux as Windows Wialon operating system available. (11)
24. Pakistan (1)
    The country is an active member of the United Nations.
    - Pakistan is an active member of the United Nations. (10)
25. Osama Ali Maher (1)
    He has worked for the United Nations in Darfur and in Pakistan.
    - Pakistan is an active member of the United Nations. (10)
26. Castle Rock (1)
    *Castle Rock, West Virginia: Wyoming County, inside the small town of Pineville
    - The Castle Rock is a geological formation in Pineville, West Virginia. (10)
27. Foreign relations of the United Kingdom (1)
    Both UK and Pakistan are active members of the Commonwealth of Nations.
    - Pakistan is an active member of the United Nations. (10)
28. Shared web hosting service (1)
    For example, the Plesk control panel is available for two operating systems, Linux and Windows.
    - The both Linux as Windows Wialon operating system available. (10)
29. Nautilus (secure telephone) (1)
    It runs from a command line and is available for the Linux and Windows operating systems.
    - The both Linux as Windows Wialon operating system available. (9)
30. Movial (1)
    It is available on Windows, Mac OS, Linux, Symbian S60 and Windows Mobile operating systems.
    - The both Linux as Windows Wialon operating system available. (8)

9.12. Hasonlósági metrikán alapuló algoritmus találati listája a Harry Potter könyvre

1. Portal:Harry Potter/Quotes/Archive (13)
   " Professor Minerva McGonagall: "Is that a student?
   - " Professor McGonagall sniffed angrily. (2)
   - Professor McGonagall gasped. (2)
   - " faltered Professor McGonagall. (2)
   - " "Yes," said Professor McGonagall. (2)
   - " whispered Professor McGonagall. (2)
   - " growled Professor McGonagall. (2)

   I have one myself above my left knee that is a perfect map of the London Underground.
   - **I have one myself above my left knee that is a perfect map of the London Underground. (24)**

   ***Professor Rubeus Hagrid: "Ah, go boil yer heads, both of yeh.**

- **"Ah, go boil yet heads, both of yeh," said Hagrid. (11)**

***Rubeus Hagrid: "Meant ter turn him into a pig, but I suppose he was so much like a pig anyway there wasn't much left ter do.**

- **Meant ter turn him into a pig, but I suppose he was so much like a pig anyway there wasn't much left ter do. (12)**

*Harry Potter: "Yes.

- " "Yes," said Harry. (4)
- "Yes," said Harry. (4)
- " "Yes," said Harry. (4)
- " "Yes," said Harry. (4)

*Harry Potter: "I don't go looking for trouble.

- "Don' you worry, Harry. (3)

***Draco Malfoy: "You'll soon find out that some wizarding families are better than others, Potter.**

- **"You'll soon find out some wizarding families are much better than others, Potter. (24)**

" Harry Potter: "I think I can tell the wrong sort for myself, thanks.

- **"I think I can tell who the wrong sort are for myself, thanks," he said coolly. (14)**

***Severus Snape: "I can teach you how to bottle fame, brew glory, even put a stopper to death — if you aren't as big a bunch of dunderheads as I usually have to teach.**

- **I can teach you how to bottle fame, brew glory, even stopper death -- if you aren't as big a bunch of dunderheads as I usually have to teach. (34)**

" George Weasley: "What are Fred and I?

- What about you, Weasley? (2)

***Draco Malfoy: "Longbottom, if brains were gold, you'd be poorer than Weasley, and that's saying something.**

- **" "Longbottom, if brains were gold you'd be poorer than Weasley, and that's saying something. (18)**

***Lord Voldemort: "There is no good and evil: only power and those too weak to seek it.**

- **There is no good and evil, there is only power, and those too weak to seek it.... (11)**

***Albus Dumbledore: "I believe your friends Misters Fred and George Weasley were responsible for trying to send you a toilet seat.**

- **I believe your friends Misters Fred and George Weasley were responsible for trying to send you a toilet seat. (29)**

2. Places in Harry Potter (10)

In addition, according to Hagrid, apart from Hogwarts, Gringotts is considered "the safest place in the world for anything you want to keep safe".

- **Gringotts is the safest place in the world fer anything yeh want ter keep safe -- 'cept maybe Hogwarts. (12)**

==Diagon Alley==

- "Welcome," said Hagrid, "to Diagon Alley. (2)

Potage's Cauldron Shop sells different varieties and sizes of cauldrons, including copper, brass, pewter, silver, self-stirring, collapsible, and solid gold, according to a sign outside the shop in Philosopher's Stone.

- Cauldrons -- All Sizes - Copper, Brass, Pewter, Silver -- Self-Stirring -- Collapsible, said a sign hanging over them. (2)

Some of the ingredients available are silver unicorn horns (for twenty-one Galleons each) and glittery-black beetle eyes (five Knuts a scoop).

- **While Hagrid asked the man behind the counter for a supply of some basic potion ingredients for Harry, Harry himself examined silver unicorn horns at twenty-one Galleons each and minuscule, glittery-black beetle eyes (five Knuts a scoop). (17)**

Inside, it is dark and full of a low, soft hooting, rustling and the flickering of "jewel-bright eyes.

- " Twenty minutes later, they left Eeylops Owl Emporium, which had been dark and full of rustling and flickering, jewel-bright eyes. (7)

Ollivanders is a fine wands shop described as "narrow and shabby, with a sign that reads Ollivanders: Makers of Fine Wands since 382 BC in peeling gold letters over the door.

- Peeling gold letters over the door read Ollivanders: Makers of Fine Wands since 382 B.C. A single wand lay on a faded purple cushion in the dusty window. (18)

When Snatchers capture Harry, Ron, and Hermione, they are brought to Malfoy Manor.

- " When Malfoy had gone, Ron and Harry looked at each other. (2)

Dumbledore sent Hagrid to retrieve it while he escorted Harry.

- "Dunno what Harry thinks he's doing," Hagrid mumbled. (1)

Nearing the end of the summer holidays, Harry meets Ron and Hermione here.

- As the match drew nearer, however, Harry became more and more nervous, whatever he told Ron and Hermione. (3)

Hagrid talked Harry out of buying a solid gold cauldron.

- " Harry couldn't speak, but Hagrid understood. (1)

3. Wikipedia:WikiProject Harry Potter/PS Differences (10)

Harry is further down the hall looking into the fire.

- **" Harry looked into the fire. (6)**

* Hagrid and Harry are on the London Underground.

- Hagrid grinned at Harry. (2)
- " Harry couldn't speak, but Hagrid understood. (2)

Hagrid takes Harry shopping for his supplies at Diagon Alley.

- "Welcome," said Hagrid, "to Diagon Alley. (1)
- " Harry explained about their meeting in Diagon Alley. (1)

Ron is going through his Chocolate Frog cards.

- " Harry unwrapped his Chocolate Frog and picked up the card. (3)

" Hermione looks pretty pleased when Ron says this.

- It was hard to tell whether Ron or Hermione was angrier about this. (1)

However, in the book Snape takes two points from Gryffindor; in the film he takes five.

- " Harry left, before Snape could take any more points from Gryffindor. (0)

They are both trying on wizard’s robes, and Draco is friendly to Harry.

- " "That's really nice of her," said Harry, trying the fudge, which was very tasty. (1)

Hermione warns Ron about studying for his exams when Neville limps in with his legs in a leg locker curse.

- Little did Harry know that Ron and Hermione had been secretly practicing the Leg-Locker Curse. (1)

* Harry, Ron and Hermione are walking down a corridor.

- " Harry knew Ron and Hermione were thinking the same as he was. (3)

Dumbledore's is using the Mirror of Erised.

- It was the Mirror of Erised. (2)

4. Harry Potter and the Philosopher's Stone (8)

Professor Quirrell is also featured in the novel.

- "Professor Quirrell! (0)
- Quirrell was there, too. (0)

The school's caretaker, Filch, knows the school's secret passages better than anyone else except possibly the Weasley twins.

- **Filch knew the secret passageways of the school better than anyone (except perhaps the Weasley twins) and could pop up as suddenly as any of the ghosts. (25)**

While on the train Harry makes friends with Ron Weasley, who tells him that someone tried to rob a vault at Gringotts.

- **Harry remembered Ron telling him on the train that someone had tried to rob Gringotts, but Ron hadn't mentioned the date. (10)**

Harry and Ron rescue her, but are caught by Professor McGonagall.

- Professor McGonagall was looking at Ron and Harry. (6)
- Professor McGonagall turned to Harry and Ron. (6)
- "It's tonight," said Harry, once he was sure Professor McGonagall was out of earshot. (1)

However, they are caught, and Harry loses the Invisibility Cloak.

- "We've got the invisibility cloak," said Harry. (3)

She dashes over to the Professors' stand, knocking over Professor Quirrell in her haste, and sets fire to Snape's robe.

- Your friend Miss Granger accidentally knocked me over as she rushed to set fire to Snape at that Quidditch match. (0)

Snape had been trying to protect Harry and suspected Quirrell.

- " "Quirrell said Snape --" "Professor Snape, Harry. (1)

Dumbledore had foreseen that the Mirror would show Voldemort/Quirrell only themselves making the elixir of life, as they wanted to use the Philosopher's Stone; Harry was able to see the Stone in the Mirror because he wanted to find it but not to use it.

- You see, only one who wanted to find the Stone -- find it, but not use it -- would be able to get it, otherwise they'd just see themselves making gold or drinking Elixir of Life. (8)

5. List of Harry Potter characters (8)

***Vindictus Viridian – Author of Curses and Counter-curses (Bewitch Your Friends and Befuddle Your Enemies with the Latest Revenges: Hair Loss, Jelly-Legs, Tongue-Tying and Much, Much More)**

- **Hagrid almost had to drag Harry away from Curses and Countercurses (Bewitch Your Friends and Befuddle Your Enemies**

with the Latest Revenges: Hair Loss, Jelly-Legs, Tongue- Tying and Much, Much More) by Professor Vindictus Viridian. (32)

*Ptolemy – Famous wizard, featured on a Chocolate Frog card

- " "Oh, of course, you wouldn't know -- Chocolate Frogs have cards, inside them, you know, to collect -- famous witches and wizards. (1)
- " Harry unwrapped his Chocolate Frog and picked up the card. (1)

*Hannah Longbottom – See Hannah Abbott

- "Abbott, Hannah! (0)

*Lisa Turpin – Ravenclaw student in Harry's year

- "Turpin, Lisa," became a Ravenclaw and then it was Ron's turn. (3)

*Mrs Wood – Mother of Oliver Wood

- "Potter, this is Oliver Wood. (0)

*Marcus Flint – Slytherin Quidditch Chaser and Captain, five years above Harry.

- Harry noticed that she seemed to be speaking particularly to the Slytherin Captain, Marcus Flint, a sixth year. (8)

*Terence Higgs – Slytherin Quidditch Seeker during Harry's first year

- Slytherin Seeker Terence Higgs had seen it, too. (2)

*Bertie Bott – Creator of Bertie Bott's Every Flavour Beans

- Bettie Bott's Every Flavor Beans! (4)

6. Magical objects in Harry Potter (8)

Such cores have been mentioned to include phoenix tail feathers, unicorn tail hairs, Veela's hair, and dragon heartstrings.

- We use unicorn hairs, phoenix tail feathers, and the heartstrings of dragons. (7)

They also meet Hermione Granger when she comes to ask if they have seen Neville Longbottom's toad, who is called Trevor.

- When Neville Longbottom, the boy who kept losing his toad, was called, he fell over on his way to the stool. (0)

On it is inscribed "erised stra ehru oyt ube cafru oyt on wohsi.

- **There was an inscription carved around the top: Erised stra ehru oyt ube cafru oyt on wohsi. (16)**

The stone is legendary in that it changes all metals to gold, and can be used to brew a potion called the Elixir of Life that can make the drinker immortal.

- It also produces the Elixir of Life, which will make the drinker immortal. (2)

In the epilogue of the movie, the scar has faded to a normal looking scar on Harry's forehead.

- He looked carefully at Harry, his eyes lingering on the scar that stood out, livid, on Harry's forehead. (0)

Within the Harry Potter universe, an invisibility cloak is used to make the wearer invisible.

- "I'll use the invisibility cloak," said Harry. (0)

===The Mirror of Erised===

- It was the Mirror of Erised. (6)

Dumbledore hid the Mirror and hid the Stone inside it, knowing that only a person who wanted to find but not use the stone would be able to obtain it.

    - **You see, only one who wanted to find the Stone -- find it, but not use it -- would be able to get it, otherwise they'd just see themselves making gold or drinking Elixir of Life. (6)**
7. Portal:Harry Potter/Quotes (6)
   *Moaning Myrtle: "Oh, Harry.
    - " Harry groaned. (2)

   *Rubeus Hagrid: "Yer a wizard Harry"
    - Hagrid grinned at Harry. (0)
    - " Harry couldn't speak, but Hagrid understood. (0)

   *Harry Potter: "I don't go looking for trouble.
    - "Don' you worry, Harry. (3)

   ***Draco Malfoy: "Longbottom, if brains were gold, you'd be poorer than Weasley, and that's saying something.**
    - **" "Longbottom, if brains were gold you'd be poorer than Weasley, and that's saying something. (18)**

   " Hermione gasped, pointing into the trunk.
    - " "But what can we --" Hermione gasped. (0)

   Very wise, Harry.
    - "Very," said Harry. (4)
8. Quidditch (6)
   On a quidditch team there are 7 players
    - "What's your Quidditch team? (0)

   Chasers score by sending the red, football-sized Quaffle through any of the three goal hoops.
    - **"The Chasers throw the Quaffle to each other and try and get it through one of the hoops to score a goal. (3)**
    - " "The Chasers throw the Quaffle and put it through the hoops to score," Harry recited. (3)

   In this respect, the game is similar, as Harry suggests in the first book, to "basketball on broomsticks with six hoops".
    - **"So -- that's sort of like basketball on broomsticks with six hoops, isn't it? (3)**

   The Beaters are armed with wooden clubs that are similar to, but shorter than, baseball bats.
    - " He handed Harry a small club, a bit like a short baseball bat. (7)

   They are tasked with protecting their team-mates and the seeker (mainly) from the Bludgers by knocking these balls off course or towards opponents.
    - That's why you have two Beaters on each team -- the Weasley twins are ours -- it's their job to protect their side from the Bludgers and try and knock them toward the other team. (3)

   Centre Chaser Angelina Johnson (Captain)
    - "And women," said Chaser Angelina Johnson. (5)
9. Wizard People, Dear Reader (6)
   Professor Quirrell Professor Queerman
    - "Professor Quirrell! (2)

   Diagon Alley Calgon Alley
    - "Welcome," said Hagrid, "to Diagon Alley. (2)

   Mr Ollivander Ed Vanders
    - " said Mr. Ollivander sharply. (0)

Oliver Wood Major Wood
- "Potter, this is Oliver Wood. (2)

Gryffindor common room Gryffindor parlor
- The Gryffindor common room was very noisy that evening. (3)

The Mirror of Erised The Gate of Heaven
- It was the Mirror of Erised. (2)

10. List of supporting Harry Potter characters (5)

Vernon is described as a big, beefy man, with hardly any neck, and a large moustache.
- **He was a big, beefy man with hardly any neck, although he did have a very large mustache. (15)**

She and Draco bump into Harry, Ron and Hermione.
- I think he's been knocked out," Ron said to Harry. (1)

===Oliver Wood===
- "Potter, this is Oliver Wood. (4)

Goyle and Malfoy are left mourning Crabbe's death.
- " Malfoy looked at Crabbe and Goyle, sizing them up. (3)
- " Malfoy grinned broadly at Crabbe and Goyle. (3)

===Crabbe and Goyle===
- Crabbe and Goyle chuckled. (4)

11. Harry Potter and the Order of the Phoenix (video game) (5)

*Harry Melling – Dudley Dursley
- " Dudley asked Harry in amazement. (2)
- " Harry was strongly reminded of Dudley. (2)

*Charlotte Skeoch – Hannah Abbott
- "Abbott, Hannah! (2)

On the back of the box it says one can be 'Crabbe, Goyle, Draco Malfoy, Bellatrix.
- " Draco Malfoy and his friends Crabbe and Goyle sniggered behind their hands. (4)

*Shefali Chowdhury – Parvati Patil
- "Shut up, Malfoy," snapped Parvati Patil. (0)

*Petrificus Totalus: Freezes opponent temporarily.
- "Petrificus Totalus! (0)

12. Potter Puppet Pals (5)

\" Dumbledore advises him to go to Hagrid.
- "Hagrid," said Dumbledore, sounding relieved. (0)
- Hagrid would never betray Dumbledore. (0)

Ron asks if Harry wants to go to his house for Christmas.
- "You want to be careful with those," Ron warned Harry. (1)

\" Snape overhears and attempts to take 500,000 points from Gryffindor, but they run away.
- **" Harry left, before Snape could take any more points from Gryffindor. (0)**

He is then joined in song by Dumbledore, Ron, Harry and Hermione.
- " Ron asked as Harry joined them. (1)

Snape reappears and Harry, Ron, Hermione and Dumbledore hug him.
- " Harry, Ron, and Hermione looked at one another, wondering what to tell him. (4)

- "We want to see Professor Dumbledore," said Hermione, rather bravely, Harry and Ron thought. (0)

13. Ron Weasley (4)

Rowling introduces Ron as "tall, thin and gangling, with freckles, big hands and feet, and a long nose.

- **He was tall, thin, and gangling, with freckles, big hands and feet, and a long nose. (21)**

This has a much more profound effect on Ron than it seems to have on Hermione or Harry.

- " Harry, Ron, and Hermione looked at one another, wondering what to tell him. (4)

During his funeral, Ron comforts a weeping Hermione.

- "It'll all be over at midnight on Saturday," said Hermione, but this didn't soothe Ron at all. (1)

At the Leaving Feast, the last dinner of the school year, Albus Dumbledore, Hogwarts' Headmaster, awards Ron fifty House points to Gryffindor for "the best-played game of chess Hogwarts has seen in many years.

- **"...for the best-played game of chess Hogwarts has seen in many years, I award Gryffindor house fifty points. (17)**

14. Hogwarts staff (4)

===Nearly Headless Nick===

- "My brothers told me about you -- you're Nearly Headless Nick! (3)
- "I've never asked," said Nearly Headless Nick delicately. (3)

The Bloody Baron is the Slytherin House ghost.

- **The Bloody Baron's becoming almost unbearable -- he's the Slytherin ghost. (6)**

Head of Slytherin House in Deathly Hallows.

- "Snape's Head of Slytherin House. (5)

Filch has a cat named Mrs.

- Filch, Snape, and Mrs. (2)

15. Death Eater (4)

Regulus Arcturus Black is the younger brother of Sirius Black.

- "Young Sirius Black lent it to me. (1)

===Draco Malfoy===

- "And my name's Malfoy, Draco Malfoy. (2)
- Draco Malfoy looked at him. (4)

Leader of the Snatcher Gang that captured Harry, Ron, and Hermione.

- " Hermione asked him, leading him over to sit with Harry and Ron. (6)

===Severus Snape===

- " "But I thought -- Snape --" "Severus? (4)

16. Hogwarts (4)

The students sleep in their House dormitories, which branch off from the common rooms.

- You will have classes with the rest of your house, sleep in your house dormitory, and spend free time in your house common room. (1)

The ghost of Slytherin house is The Bloody Baron.

- **The Bloody Baron's becoming almost unbearable -- he's the Slytherin ghost. (6)**

The class is taught by Professor Flitwick.

- Professor Flitwick put the class into pairs to practice. (1)

Divination is described by Professor McGonagall as "one of the most imprecise branches of magic".

- It sounds like fortune-telling to me, and Professor McGonagall says that's a very imprecise branch of magic. (3)

17. Harry Potter and the Philosopher's Stone (film) (4)

* Fiona Shaw as Petunia Dursley, Harry's Muggle aunt.

- " said Aunt Petunia, looking furiously at Harry as though he'd planned this. (1)

He winds up in Gryffindor, along with Ron and Hermione.

- "Gryffindor," said Ron. (0)

' And I thought, You're gonna say Ron.

- "Say you're ill," said Ron. (3)

However, Ron is nearly killed in the match and Hermione stays with Ron as Harry goes on ahead, alone.

- As the match drew nearer, however, Harry became more and more nervous, whatever he told Ron and Hermione. (6)

18. A Very Potter Musical (3)

Harry Potter, Ron Weasley, Hermione Granger, and the other students of Hogwarts School of Witchcraft and Wizardry rejoice that they are going back for their second year ("Goin' Back to Hogwarts").

- " Harry unfolded a second piece of paper he hadn't noticed the night before, and read: HOGWARTS SCHOOL of WITCHCRAFT and WIZARDRY UNIFORM First-year students will require: 1. (5)

* "Not Alone" – Ginny Weasley, Harry Potter, Ron Weasley, Hermione Granger

- This is Harry Potter an' Hermione Granger, by the way. (2)

Harry, Ron, and Hermione arrive at Dumbledore's office for the meeting.

- "We want to see Professor Dumbledore," said Hermione, rather bravely, Harry and Ron thought. (5)

19. Harry Potter and the Prisoner of Azkaban (3)

Ron is furious at Hermione.

- " said Ron furiously. (4)
- It was hard to tell whether Ron or Hermione was angrier about this. (3)

Ron, Hermione, and Harry are reconciled in their efforts to help Hagrid.

- " Harry knew Ron and Hermione were thinking the same as he was. (1)

This map leads Harry through a secret passageway into Hogsmeade, where he rejoins Ron and Hermione.

- With one last desperate look back at Ron, Harry and Hermione charged through the door and up the next passageway. (1)

20. Harry Potter and the Chamber of Secrets (film) (3)

* Fiona Shaw as Petunia Dursley, Harry's Muggle aunt.

- " said Aunt Petunia, looking furiously at Harry as though he'd planned this. (1)

Harry and the Weasleys travel to Diagon Alley by Floo Powder.

- " Harry explained about their meeting in Diagon Alley. (1)
- Harry was remembering his trip to Diagon Alley -how could he have been so stupid? (6)

Harry suspects the Heir is Malfoy.

- "You're worth twelve of Malfoy," Harry said. (0)

21. Muggle (3)
    *Petunia Dursley, Harry's Aunt
        - " "Thirty-nine, sweetums," said Aunt Petunia. (0)
        - " said Aunt Petunia, looking furiously at Harry as though he'd planned this. (1)
        - " "On vacation in Majorca," snapped Aunt Petunia. (0)
        - " he asked Aunt Petunia. (2)
        - " "DotA be stupid," snapped Aunt Petunia. (0)
        - Aunt Petunia gave a gasp of horror. (0)
        - " shrieked Aunt Petunia suddenly. (2)
    *Vernon Dursley, Harry's Uncle
        - But Uncle Vernon didn't believe him. (2)
    *Tobias Snape, the father of Severus Snape
        - " "But I thought -- Snape --" "Severus? (0)
22. A Very Potter Sequel (3)
    The map leads Harry, Ron, and Hermione to a room containing a large mirror.
        - " Hermione asked him, leading him over to sit with Harry and Ron. (2)
    Richard Campbell Neville LongbottomPast Hermione
        - " Hermione urged Neville. (0)
        - " Hermione and Neville were suffering, too. (0)
    * "No Way" – Harry Potter, Draco Malfoy, Ron Weasley, Hermione Granger
        - This is Harry Potter an' Hermione Granger, by the way. (5)
23. Wikipedia:In the news/Candidates/July 2011 (3)
    Yes, Harry Potter's culturally significant.
        - " "Yes," said Harry. (0)
        - "Yes," said Harry. (0)
        - " "Yes," said Harry. (0)
        - " "Yes," said Harry. (0)
    At the moment it looks like you're on some sort of crusade to get this posted.
        - He hadn't expected something like this the moment they arrived. (2)
    BTW: were you typing with your feet, or what?
        - "What's that at its feet? (2)
24. Harry Potter and the Order of the Phoenix (film) (3)
    With Dumbledore gone, Umbridge becomes the new Headmistress.
        - I suppose he really has gone, Dumbledore? (0)
    *Fiona Shaw as Petunia Dursley, Harry's Muggle aunt.
        - " said Aunt Petunia, looking furiously at Harry as though he'd planned this. (1)
    Harry and Sirius duel Lucius Malfoy.
        - "You're worth twelve of Malfoy," Harry said. (0)
25. Wikipedia:Administrators' noticeboard/Moulton (2)
    :Draco Malfoy: You'll soon find out some wizarding families are much better than others, Potter.
        - **"You'll soon find out some wizarding families are much better than others, Potter. (27)**
    :Harry Potter: I think I can tell who the wrong sort are for myself, thanks.
        - **"I think I can tell who the wrong sort are for myself, thanks," he said coolly. (17)**

26. List of Deadliest Warrior episodes (2)
    He slowly walks through the trees and makes his way to a darker part of the forest.
        - " They walked on through the dense, dark trees. (7)
    The Mongol, feeling as if he's being watched, looks to the top of the ridge.
        - He had the nasty feeling they were being watched. (6)
27. Aunt Petunia (2)
    Aunt Petunia can refer to:
        - " "Thirty-nine, sweetums," said Aunt Petunia. (0)
        - " "On vacation in Majorca," snapped Aunt Petunia. (0)
        - " he asked Aunt Petunia. (4)
        - " "DotA be stupid," snapped Aunt Petunia. (0)
        - Aunt Petunia gave a gasp of horror. (0)
        - " shrieked Aunt Petunia suddenly. (2)
    * Harry Potter's aunt, Petunia Dursley
        - " said Aunt Petunia, looking furiously at Harry as though he'd planned this. (1)
28. Portal:Harry Potter/Featured Character/Archive (1)
    Dumbledore is described as tall and thin, with long silver hair that looked long enough to tuck into his belt and a long beard.
        - **He was tall, thin, and very old, judging by the silver of his hair and beard, which were both long enough to tuck into his belt. (20)**
29. Albus Dumbledore (1)
    Albus Dumbledore was tall and thin, with silver hair and beard (auburn in his youth) so long that they could be tucked into his belt.
        - **He was tall, thin, and very old, judging by the silver of his hair and beard, which were both long enough to tuck into his belt. (15)**
30. Artemisia (plant) (1)
    *** In Harry Potter, the Draught of Living Death, an extremely powerful sleeping potion, is made from powdered root of asphodel added to an infusion of wormwood.**
        - **"For your information, Potter, asphodel and wormwood make a sleeping potion so powerful it is known as the Draught of Living Death. (13)**
31. Muggle Quidditch (1)
    ***Chasers are responsible for passing the Quaffle and scoring points by throwing the Quaffle through one of the opponent's goals.**
        - **"The Chasers throw the Quaffle to each other and try and get it through one of the hoops to score a goal. (8)**
        - " "The Chasers throw the Quaffle and put it through the hoops to score," Harry recited. (3)
32. Magic in Harry Potter (1)
    In addition, the drinking of Unicorn blood will keep a person alive even if death is imminent, but at the terrible price of being cursed forever.
        - **The blood of a unicorn will keep you alive, even if you are an inch from death, but at a terrible price. (10)**
33. Sonic's Rendezvous Band (album) (1)
    #"Slow Down (Take a Look)" (Morgan)
        - The train did seem to be slowing down. (10)

34. St. Joseph's School, Bhagalpur (1)
Meritorious houses are awarded house points and the house with the most points at the end of the year wins the House cup.
- **At the end of the year, the house with the most points is awarded the house cup, a great honor. (9)**

35. Viper (Six Flags Great America) (1)
This turn around leads to a sharp second drop, before rising over another hill, and passing through a left hand second turn around.
- The Potters smiled and waved at Harry and he stared hungrily back at them, his hands pressed flat against the glass as though he was hoping to fall right through it and reach them. (9)

36. Shell (theater) (1)
Often shells are designed to be removable, either rolling away on wheels or flying into a flyspace.
- Hagrid was sitting in an armchair outside his house; his trousers and sleeves were rolled up, and he was shelling peas into a large bowl. (8)

37. Order of the Phoenix (organisation) (1)
Rubeus Hagrid Care of Magical Creatures Teacher at Hogwarts as well as Keeper of Keys and Grounds.
- Rubeus Hagrid, Keeper of Keys and Grounds at Hogwarts. (8)

9.13. Hasonlósági metrikán alapuló algoritmus találati listája a 12 Wikipédia cikk német fordítására

1. Pete Seeger (7)
Seeger was born in French Hospital, Midtown Manhattan, the youngest of three sons.
- Seeger wurde in dem französischen Krankenhaus in Midtown, einem Stadtteil von Manhatten geboren. (7)

His father, Charles Louis Seeger Jr. was a prominent musicologist, composer, and music professor.

- Sein Vater Charles Louis Seeger Jr. war Komponist und der Vorreiter der ethnomusikologischen Forschung der amerikanischen und der nicht westlichen Musikwissenschaft. (2)

His mother, Constance de Clyver Edson, was a violinist and teacher, raised in Tunisia and trained at the Paris Conservatory of Music and the Juilliard School.

- Seine Mutter, Constance de Clyver Edson war eine klassische Geigerin und Lehrerin. (3)

His stepmother, Ruth Crawford Seeger, was one of the most significant female composers of the 20th century.

- Seine Stiefmutter, Ruth Crawford Seeger war die bedeutendste Frauenkomponistin des zwanzigsten Jahrhunderts. (2)

His eldest brother, Charles Seeger III, was a radio astronomer, and his next older brother, John Seeger, taught in the 1950s at the Dalton School in Manhattan and was the principal from 1960 to 1976 at Fieldston Lower School in the Bronx.

- Sein ältester Bruder, Charles Seeger III, war Radioastronom und sein nächst älterer Bruder, John Seeger unterrichtete in den 50-er jahren in der Dalton School, in Manhatten und war der Rektor der Fieldston Lower School in Bronx von 1960 bis 1976. (26)

His uncle, Alan Seeger, a noted poet, was killed during the First World War.

- Sein Onkel, Alan Seeger war ein berühmter Dichter, der während des ersten Weltkrieges umgebracht wurde. (5)

Half-brother Mike Seeger went on to form the New Lost City Ramblers, one of whose members, John Cohen, was married to Pete's other half-sister, singer Penny Seeger, also a highly talented singer.

- Halbbruder Mike Seeger hat sich die New Lost City Ramblers gegeründet, deren Mitglied, John Cohen Pete's andere Halbschwester, die Sängerin Penny Seeger, eine ebenfalls sehr talentierte Sängerin geheiratet hat. (10)

2. Munich Philharmonic (6)
The orchestra was founded in Munich in 1893 by Franz Kaim, son of a piano manufacturer, as the Kaim Orchestra.
   - Das Orchester wurde in München im Jahre 1893 durch Franz Kaim, den Sohn eines Klavierbauers als das Kaim-Orchester gegründet. (13)

Felix Weingartner was music director from 1898 to 1905, and the young Wilhelm Furtwängler made his auspicious conducting debut there in 1906.

- Felix Weingartner war Musikdirektor von 1898 bis 1905 und der junge Wilhelm Furtwängler hat hier sein vielversprechendes Dirigentendebüt im Jahre 1906. (7)

Meanwhile Anton Bruckner pupil Ferdinand Löwe established an enduring tradition of Bruckner performance which continues to this day.

- Inzwischen hat Anton Bruckner, ein Schüler von Ferdinand Löwe eine bleibende Tradition der Bruckner-Vorstellungen geschafft, die auch bis heute fortgesetzt wird. (2)

After the war, the orchestra was taken over by the city of Munich and restarted under the leadership of composer Hans Pfitzner, soon replaced by Bruckner pioneer Siegmund von Hausegger.

- Nach dem Krieg wurde das Orchester von der Münchner Stadt übernommen und wurde neugestrartet unter der Führung von dem Komponist Hans Pfitzner, der bald durch den Bruckner-Vorkämpfer Siegmund von Hausegger aufgelöst wurde. (10)

In 1928, the orchestra acquired its current name.

- Im Jahre 1928 hat das Orchester seinen heutigen Name erworben. (6)

After the rise of the Nazi party in 1933, the orchestra stamped its scores with swastikas and the words "The Orchestra of the Fascist Movement".

- Nach der Entstehung der Nazi Partei im Jahre 1933, das Orchester stempelte ihre Partituren mit Hakenkreuz und den Wörtern "Das Orchester der faschistischen Bewegung". (4)

3. Upper 10 (5)
Upper 10 is a caffeinated lemon-lime soft drink, similar to Sprite, 7 Up, Sierra Mist, and Bubble Up.
   - Upper 10 ist ein koffeinhaltiges Zitrone-Lime Erfrischungsgetränk ähnlich zu Sprite, 7 Up, Sierra Mistund Bubble Up. (9)

   It was bottled by RC Cola.

   - Abgefüllt war durch RC Cola. (0)

   The Upper 10 brand debuted in 1933 as a product of the Nehi Corporation (later Royal Crown Corporation).

   - Die Upper 10 Marke hat im Jahre 1933 als ein Produkt der Nehi Corporation (später Royal Crown Corporation) debütiert. (15)

   U.S. Patent and Trademark Office Upper 10 was one of RC Cola's flagship brands throughout the company's history.

   - Das Upper 10 war einer der Flaggenschiffe der RC Cola durchweg in der Geschichte der Firma. (10)

   Upper 10 is still sold outside of North America by Cott Beverages, the same company that sells RC Cola internationally.

   - Upper 10 wird nach wie vor durch die Firma Cott Beverages, dieselbe Firma, die RC Cola international verteilt, außerhalb von Nord-Amerika verkauft. (13)
4. Foreign relations of Pakistan (5)
Pakistan is an active member of the United Nations.
   - Pakistan ist ein aktives Mitglied der Vereinigten Nationen. (10)

   It was a member of the CENTO and SEATO military alliances.

- Es war Mitglied der Militärbündnisse CENTO und SEATO. (0)

Its alliance with the United States was especially close after the Soviets invaded the neighboring country of Afghanistan.

- Sein Bündnis mit der Vereinigten Staaten war besonders dann eng, als die Sowjets in das Nachbarland Afghanistan eingedrungen waren. (11)

In 1964, Pakistan signed the Regional Cooperation for Development (RCD) Pact with Turkey and Iran, when all three countries were closely allied with the U.S., and as neighbors of the Soviet Union, wary of perceived Soviet expansionism.

- Im Jahre 1964, als die drei Länder sehr eng mit der USA verbunden waren und als Nachbars der Sowjetunion eine sowjetische Expansion befürchteten, unterschrieb Pakistan den Vertrag zur Regionalen Kooperation für Entwicklung (RCD) mit der Türkei und Iran. (4)

RCD became defunct after the Iranian Revolution, and a Pakistani-Turkish initiative led to the founding of the Economic Cooperation Organisation (ECO) in 1985.

- RCD wurde nach der Iranischen Revolution gelöscht und die pakistanisch-türkische Initiativen führten im Jahre 1985 zu der Gründung der Organisation für wirtschaftliche Zusammenarbeit (CEO). (18)

5. Golden Twenties (4)
It is often applied to Germany, which during the early 1920s, experienced, like most of Europe, record-breaking levels of inflation of one trillion percent between January 1919 and November 1923.
   - Man bezieht es oft an Deutschland, wo während der frühen 20-er Jahre, wie meist in Europa, ein rekordbrechendes Niveau der Inflation von ein Billion Prozent zwischen Januar 1919 und November 1923 erfahren wurde. (15)

   The inflation was so severe that printed currency was often used for heating and other uses, and everyday requirements like food, soap, electricity cost a wheelbarrow full of banknotes.

   - Die Inflation war so stark, dass die gedruckte Währung oft für Heizen und andere Zwecke benutzt war und die alltäglichen Produkte wie Lebensmittel, Seife, Strom kosteten Schubkarren voller Geldscheine. (16)

   Such events, among many other factors, triggered the rise of fascism in Italy, as well as the ill-fated Beer Hall Putsch, masterminded by a young Adolf Hitler.

   - Unter anderen Faktor haben solche Ereignisse das Emporsteigen des Faschismus in Italien ausgelöst, wie auch den Hitler-Putsch, der durch den jungen Adolf Hitler geleitet wurde. (5)

Before long, the Weimar Republic under Chancellor Gustav Stresemann managed to tame the extreme levels of inflation by the introduction of a new currency, the Rentenmark, with tighter fiscal controls and reduction of bureaucracy, leading to a relative degree of political and economic stability.

- Lange davor hat die Weimarer Republik unter Kanzler Gustav Stresemann das extremes Niveau der Inflation durch die Einführung einer neuen Währung, der Rentenmark, mit harten Finanzkontrolle und Reduzierung der Bürokratie zu zähmen versucht, zu einem relativen Grad der politischen und wirtschaftlichen Stabilität führend. (26)

6. Wialon (4)
Wialon is a web-based GPS tracking software platform with some fleet management features, developed by a Belarusian company Gurtam.
   - Wialon ist eine auf webbasierte GPS Verfolgung-Software-Plattform mit einigen Flottenmanagement-Funktionalitäten, die von der belarussischen Firma Gurtam entwickelt wurde. (5)

   Wialon is available for both Linux and Windows operating systems.

   - Wialon ist sowohl für Linux als auch für Windows Oparationssysteme erhältlich. (1)

   Wialon is notable because of its compatibility with different GLONASS and GPS tracking units, counting 131 in the beginning of July 2010.

   - Wialon ist beachtenswert wegen ihrer Kompatibilität mit verschiedenen GLONASS und GPS Verfolgung-Einheiten, genau 131 am Anfang Juli 2010. (21)

   According to Gurtam website, there are about 200 GPS tracking services worldwide who have already implemented Wialon GPS tracking platform.

   - Gemäß der Webseite von Gurtam, es sind mehr als 200 GPS Verfolgung-Services weltweit, die alle schon Wialon GPS Verfolgung-Plattform implementiert haben. (8)

7. Politics of Pakistan (4)
It was a member of the CENTO and SEATO military alliances.
   - Es war Mitglied der Militärbündnisse CENTO und SEATO. (0)

   Its alliance with the United States was especially close after the Soviets invaded the neighbouring country of Afghanistan.

   - Sein Bündnis mit der Vereinigten Staaten war besonders dann eng, als die Sowjets in das Nachbarland Afghanistan eingedrungen waren. (11)

   In 1964, Pakistan signed the Regional Cooperation for Development (RCD) Pact with Turkey and Iran, when all three countries were closely allied with the U.S., and as neighbours of the Soviet Union, wary of perceived Soviet expansionism.

- Im Jahre 1964, als die drei Länder sehr eng mit der USA verbunden waren und als Nachbars der Sowjetunion eine sowjetische Expansion befürchteten, unterschrieb Pakistan den Vertrag zur Regionalen Kooperation für Entwicklung (RCD) mit der Türkei und Iran. (4)

RCD became defunct after the Iranian Revolution, and a Pakistani-Turkish initiative led to the founding of the Economic Cooperation Organisation (ECO) in 1985.

- RCD wurde nach der Iranischen Revolution gelöscht und die pakistanisch-türkische Initiativen führten im Jahre 1985 zu der Gründung der Organisation für wirtschaftliche Zusammenarbeit (CEO). (18)

8. Web accessibility (4)
Web accessibility refers to the inclusive practice of making websites usable by people of all abilities and disabilities.
   - Die Web-Zugänglichkeit bezieht sich auf die Praxis der Nutzbarmachung von Webseiten für Leute mit allen Fähigkeiten und Behinderten. (0)

When sites are correctly designed, developed and edited, all users can have equal access to information and functionality.

- Wenn die Webseiten richtig entworfen, entwickelt und editiert sind, können alle Benutzer gleichermaßen zu den Informationen und den Funktionalitäten zugreifen. (3)

For example, when a site is coded with semantically meaningful HTML, with textual equivalents provided for images and with links named meaningfully, this helps blind users using text-to-speech software and/or text-to-Braille hardware.

- Zum Beispiel, wenn eine Webseite mit semantisch aussagefähigen HTML, mit den Bildern entsprechenden Texten unterstützt und mit sinnvoll benannten Links codiert sind, das kann blinden Benutzern helfen eine Text-in-Sprache Umwandlungs-Software und/oder eine Text-to-Braille Hardware zu benutzen. (20)

When clickable links and areas are large, this helps users who cannot control a mouse with precision.

- Wenn die anklickbaren Links und Bereichen sind groß, das hilft Benutzern, die die Maus nicht präzis kontrollieren können. (2)

9. Castle Rock (Pineville, West Virginia) (4)
Castle Rock is a geological feature located in Pineville, West Virginia next to the Pineville Public Library.
   -) Castle Rock ist eine geologische Formation, die sich in Pineville, West Virginia in der Nähe der Öffentlichen Bibliothek von Pineville befindet. (11)

Named for its resemblance to a castle, it rises about 200 feet above Rock Castle Creek, a branch of the Guyandotte River.

- Hat ihren Name wegen seiner Ähnlichkeit zu einer Burg bekommen, die erhebt sich über 200 Fuß über Rock Castle Creek, eine Abzweigung des Flusses Guyandotte. (5)

The formation of Castle Rock began about two hundred million years ago.

- Die Formation von Castle Rock hat mehr els vor zweihundert Millionen Jahren begonnen. (5)

In 2001 a sign explaining how the rock was formed was placed in front of Castle Rock.

- Man hat im Jahre 2001 eine Zeichenerklärung über Entstehung der Felsen in Front von Castle Rock gesetzt. (3)

10. Mike Seeger (3)
His eldest half-brother, Charles Seeger III, was a radio astronomer, and his next older half-brother, John Seeger, taught for years at the Dalton School in Manhattan.
    - Sein ältester Bruder, Charles Seeger III, war Radioastronom und sein nächst älterer Bruder, John Seeger unterrichtete in den 50-er jahren in der Dalton School, in Manhatten und war der Rektor der Fieldston Lower School in Bronx von 1960 bis 1976. (12)

His uncle, Alan Seeger, a poet, was killed during the First World War.

- Sein Onkel, Alan Seeger war ein berühmter Dichter, der während des ersten Weltkrieges umgebracht wurde. (7)

His sister, singer Penny Seeger, married John Cohen, a member of Mike's musical group, New Lost City Ramblers.

- Halbbruder Mike Seeger hat sich die New Lost City Ramblers gegeründet, deren Mitglied, John Cohen Pete's andere Halbschwester, die Sängerin Penny Seeger, eine ebenfalls sehr talentierte Sängerin geheiratet hat. (0)

11. University of Oxford (2)
It is the second-oldest surviving university in the world and the oldest in the English-speaking world.
    - Die ist die zweitälteste erhaltene Universität der Welt und die älteste in der Englisch sprechenden Welt. (2)

The University grew rapidly from 1167 when Henry II banned English students from attending the University of Paris.

- Die Universität wuchs schnell vom 1167, als Heinrich II. englischen Studenten das Lernen an der Pariser Universität verboten hat. (7)

12. GPS tracking server (2)
* Wialon — GPS tracking software platform with some fleet management features, developed by a Belarusian company Gurtam.

- Wialon ist eine auf webbasierte GPS Verfolgung-Software-Plattform mit einigen Flottenmanagement-Funktionalitäten, die von der belarussischen Firma Gurtam entwickelt wurde. (7)

Wialon is available for both Linux and Windows operating systems.

- Wialon ist sowohl für Linux als auch für Windows Oparationssysteme erhältlich. (1)

13. Canberra 400 (1)
    The Canberra 400, also known as both the GMC 400, Stegbar Canberra 400 and in its infancy, National Capital 100 was a V8 Supercar race run on the streets of Australia's capital, Canberra.
    - Die Canberra 400, auch als GMC 400 und Stegbar Canberra 400 oder in den Anfangsjahren als National Capital 100 bekannt, war eine V8-Supercar-Rennen auf den Straßen der australischen Hauptstadt Canberra. (12)

9.14. N-gram algoritmus által visszaadott találatok F_6 maximalizációjára törekedve

12 Wikipédia cikk: Mike Seeger (19), Upper 10 (18), Pete Seeger (18), Wialon (15), Web accessibility (14), Foreign relations of Pakistan (13), Munich Philharmonic (13), Castle Rock (Pineville, West Virginia) (13), Golden Twenties (12), GPS tracking server (11), Politics of Pakistan (10), Avidius Cassius (8), Charles Seeger (7), Meitetsu 5000 series (2008) (4), Dr Pepper/Seven Up (4)

15 oldalas cikk: Wikipedia:Articles for deletion/Characters in the Sims 2 (10), Books LLC (10), SIM2 (10), Basic helix-loop-helix (10), SIM1 (10), Stretched grid method (6), Reduced dimensions form (5), Heritability (5), Wavelength (5), Sexual dimorphism measures (5), Genetic algorithm scheduling (5), Arabic Letter Keyboard Intellark (5), Non-radiative dielectric waveguide (5), Backpressure Routing (5), Mixing patterns (4), Thermometric titration (4), EXPRESS (data modeling language) (4), Size effect on structural strength (4), Metadata modeling (4), Reflector sight (4), Proton-transfer-reaction mass spectrometry (4), Harmonic pitch class profiles (4), Multiplication table (4), Gábor N. Sárközy (4), SHIWA project (4), Robert Lovas (4), Péter Kacsuk (4), Optical neural network (4), Wikipedia:Suspected copyright violations/2012-03-21 (4), The Computer and Automation Research Institute, Hungarian Academy of Sciences (4), Crossbar latch (4), Basal body temperature (4), Lagrange multiplier (4), András Gyárfás (4), Wikipedia:TLAs from UA0 to XZ9 (4), Photoacoustic imaging in biomedicine (4)

Harry Potter: Places in Harry Potter (145), Harry Potter and the Philosopher's Stone (135), Magical objects in Harry Potter (118), Portal:Harry Potter/Quotes/Archive (103), List of Harry Potter characters (101), Hogwarts (83), Harry Potter (character) (78), Hogwarts staff (78), List of supporting Harry Potter characters (63), Harry Potter universe (57), Portal:Harry Potter/Quotes (57), List of fictional books (56), Albus Dumbledore (49), Order of the Phoenix (organisation) (48), List of spells in Harry Potter (45), Ron Weasley (44), Quidditch (26), Potter Puppet Pals (25), Wikipedia:Administrators' noticeboard/Moulton (24), Politics of Harry Potter (24), Harry Potter influences and analogues (23), Harry and the Potters (23),

Wikipedia:Reference desk/Archives/Miscellaneous/2010 August 31 (22), Pottermore (19), The Chronicles of Narnia (17), Harry and the Potters (album) (16), Wikipedia:Articles for deletion/Ealing Broadway Platform 9 (15), Wikipedia:Articles for deletion/Platform 10 (14), Wikipedia:Upload log/Archive 2 (14), Wikipedia:WikiProject Harry Potter/Images/Watchlist (14), Wikipedia:Reference desk/Archives/Language/2010 August 5 (13), Treacle (13), Wikipedia:Administrators' noticeboard/IncidentArchive729 (13), Aathichoodi (10), Wikipedia:Administrators' noticeboard/IncidentArchive178 (9), English cuisine (8), Surrey (7), Wikipedia:Articles for deletion/Trevor (Harry Potter) (7), Portal:Literature/Did you know/Week 5 (6), Spoonerism (5), Wikipedia:Reference desk/Archives/Language/2007 June 4 (5), Down South Flava (4), Super Survivor (4), Who U Wit? (4), We Have the Right to Remain Violent (4), Cupid (Cupid album) (4), We Got This (4), Tha Boss (4), WPEG (4)

9.15. Angol-német hasonlósági metrikán alapuló keresés során nem talált szavak

Seeger, Cassius, Canberra, Wialon, GPS, Castle, Upper, Fuß, Oxford, Cola, fuß, Up, RC, währen, Sitzplätze, sei, Bruckner, spät, Schieferformation, Weltkrieges, School, Gurtam, gemäß, Halbschwester, Hitler, John, Kaim, Gustav, koffeinhaltiges, Corporation, Pepper, No, Münchner, mm, Formation, Mokett, Haltegriffe, Dr, Pineville, Grunddurchmesser, Cambridge, RCD, II, Webseite, Charles, New, Manhatten, Webseiten, sowjetisch, groß, Felix, Mahlers, Bruno, symphonie, Symphonie, Weingartner, Walter, Orchestra, Wilhelm, Ferdinand, geschafft, Konzertvereins, hieß, vergrößerbar, Anton, Symphony, Furtwängler, Dirigentendebüt, Musikdirektor, farbblinden, Abfüllfirmen, Dyslexia, Lehrdiagramm, Snapple, Group, eingegliedert, Inc, Firmenoperationen, Benutzbarkeit, Lernschwierigkeiten, Seven, Cott, Krampfanfällen, Klavierbauers, Tonhalle, Konzerthalle, Braille, anklicken, Franz, Beverages, außerhalb, navigieren, Zugangsvorrichtung, Mahler, Hans, Adolf, Service, Weimarer, Stresemann, Rentenmark, rekordbrechend, Street, Virgil, Senter, erbauen, Konjunkturschwächen, Finanzkontrolle, service, GLONASS, Linux, Oparationssysteme, windows, belarussischen, Flottenmanagement, webbasiert, Nachfolgestaaten, IT, CeBIT, Felsenspitze, zweihundert, heutig, partitur, Partitur, HTML, Hausegger, Siegmund, to, Windows, Pfitzner, umwandlungs, semantisch, west, anschließen, Steinterrasse, gleichermaßen, els, Guyandotte, Creek, West, Virginia, Information, information, neugestrartet, Sprite, Ewan, Volkssänger, MacColl, Mike, City, losen, Volksdarstellerin, umgebracht, Fieldston, Dalton, Lower, Bronx, Alan, Ramblers, gegeründet, wären, ermuntern, Stützer, Commodus, ägyptisch, Marcus, Faustina, Pete, Cohen, s, talentiert, Aurel, Radioastronom, III, USA, Nachbarland, Expansion, iranisch, Midtown, CEO, SEATO, CENTO, Gründungsdatum, informell, Heinrich, Stadteinwohnern, Militärbündnisse, Großeltern, Patterson, Edson, Clyver, Ruth, Crawford, Frauenkomponistin, de, Constance, Louis, York, Jr, ethnomusikologischen, Musikwissenschaft, Traiana, fortis, Kostenschätzung, Vertragslösung, Kate, Carnell, Wintersaison, Camberra, Hauptgrund, Stanhope, australisch, Straße, Gary, Humphries, Jon, c, interstaatliche, debütieren, Crown, Flaggenschiffe, Cadbury, Schweppes, royal, Nehi, Erfrischungsgetränk, Lime, Sierra, Mistund, Bubble, straße, straßen, begnadigen, ihn, Innenraum, hellgrau, Einzelsitze, begehen, Selbsmord, Deitoriana, XXII, Diro, Dio, Illoyalität, d, h, Anfangsjahren, Stegbar, Capital, v, Supercar, GMC, Fahrerkabine, Ältere, Fußboden, klarstellen, Klarstellen, sitzplätze, plc

10. Irodalomjegyzék

Alattar 2004	Adnan M. Alattar and Osama M. Alattar: Watermarking electronic text documents containing justified paragraphs and irregular line spacing, Proc. SPIE 5306, Security, Steganography, and Watermarking of Multimedia Contents VI, 685 (June 22, 2004); doi:10.1117/12.527147; http://dx.doi.org/10.1117/12.527147
Alzahrani 2010	Salha Alzahrani and Naomie Salim: Fuzzy Semantic-Based String Similarity for Extrinsic Plagiarism Detection - Lab Report for PAN at CLEF 2010
Amazon 2012	Kindle ebook sales have overtaken Amazon print sales, The Guardian, 2012 http://www.guardian.co.uk/books/2012/aug/06/amazon-kindle-ebook-sales-overtake-print
Baeza-Yates 1999	Ricardo Baeza-Yates, Berthier Ribeiro-Neto: Modern Information Retrieval, Addison Wesley, 1999.
Bailey 2011a	Jonathan Bailey: The Problem with Detecting Translated Plagiarism, 2011 http://www.plagiarismtoday.com/2011/02/24/the-problem-with-detecting-translated-plagiarism/
Bailey 2011b	Jonathan Bailey: Using Citations to Detect Plagiarism, Plagiarism Today http://www.plagiarismtoday.com/2011/08/08/using-citations-to-detect-plagiarism/
Bailey 2012	Jonathan Bailey: 5 Ways Web Scraping Has Already Affected You, Plagiarism Today, 2012 http://www.plagiarismtoday.com/2012/04/25/ways-web-scraping-has-already-affected/
Barrón-Cedeño 2008	Alberto Barrón-Cedeño, Paolo Rosso, David Pinto and Alfons Juan: On Crosslingual Plagiarism Analysis Using a Statistical Model. In: ECAI'08 PAN Workshop Uncovering Pla-giarism, Authorship and Social Software Misuse, pp. 9–13, Patras, Greece (2008)
Barrón-Cedeño 2009	Alberto Barrón-Cedeño and Paolo Rosso: Monolingual and Crosslingual Plagiarism Detection. Towards the Competition @ SEPLN09. In: Proc. III Jornadas PLN-TIMM, Madrid, Spain, February 5-6, pp.29-32
Békés 1951	Békés Gábor, Dalos Péter ford., Újszövetségi szentírás. A külföldi katolikus magyar akció kiadása, Róma, 1951 http://www.cadvision.com/mayl/bd-ind.html

Benedetto 2002	D. Benedetto; E. Caglioti, V. Loreto: Language trees and zipping. Physical Review Letters (2002) 88:4
Brin 1995	S. Brin, J. Davis, and H. Garcia-Molina . Copy detection mechanisms for digital do cuments. In Proceedings of the ACM SIGMOD Annual Conference, San Francisco, CA, May 1995.
Cavnar 1994	W. B. Cavnar, J. M. Trenkle: N-Gram-Based Text Categorization. Proceedings of Third Annual Symposium on Document Analysis and Information Retrieval, Las Vegas, NV, UNLV Publications/Reprographics (1994) 161-175
Ceska2008	Zdenek Ceska, Michal Toman and Karel Jezek, „Multilingual Plagiarism Detection", Artificial Intelligence: Methodology, Systems, and Applications Lecture Notes in Computer Science, 2008, Volume 5253/2008, 83-92
Copyscape	Copyscape by Indigo Stream Technologies http://www.copyscape.com/
Costa-jussà 2010	Marta R. Costa-jussà, Rafael E. Banchs, Jens Grivolla and Joan Codina: Plagiarism Detection Using Information Retrieval and Similarity Measures Based on Image Processing Techniques - Lab Report for PAN at CLEF 2010
Craggs 2011	David Justin Craggs: An Analysis and Comparison of Predominant Word Sense Disambiguation Algorithms, Edith Cowan University, 2011 http://ro.ecu.edu.au/cgi/viewcontent.cgi?article=1003&context=theses_hons
Crystal 2003	David Crystal, A nyelv enciklopédiája, Osiris, Bp. 2003.
Csernoch 2003	Csernoch Mária: A szavak véletlenszerű megjelenésén alapuló modellek és az irodalmi művek közötti eltérések magyarázata, II. Magyar Számítógépes Nyelvészeti Konferencia, 2004
CSPblog	Copy, Shake, and Paste blog, 2012. A blog about plagiarism and scientific misconduct from a German professor, written in English [online] Available at: http://copy-shake-paste.blogspot.com/ [Accessed 30 March 2012].
DRM	Digital rights management, Wikipedia http://en.wikipedia.org/wiki/Digital_rights_management
DSD	MTA SZTAKI Elosztott Rendszerek Osztály http://dsd.sztaki.hu
Dunning 1994	Dunning, T.: Statistical Identification of Language. Technical Report MCCS 94-273, New Mexico State University (1994)
EVE	EVE Plagiarism Detection System http://www.canexus.com

Faruqui 2010	Manaal Faruqui and Sebastian Pad: Training and Evaluating a German Named Entity Recognizer with Semantic Generalization, Proceedings of Konvens 2010, Saarbrücken, Germany.
Fischer 2008	Fischer Márta, Fordítás és közvetítés a nyelvoktatásban – mit nyújthat a nyelvoktatásnak a fordítástudomány? , 2008. http://ecml.opkm.hu/files/FischerM.doc
Fry 1989	Stephen Fry: A Bit Of Fry And Laurie, BBC2, Series 1 episode 3, 1989
GImage	Google image search, Feltöltött kép alapján lehet hasonló képeket keresni http://images.google.com/
Gipp 2010	Bela Gipp and Jöran Beel: Citation Based Plagiarism Detection – A New Approach to Identify Plagiarized Work Language Independently, Proceedings of the 21st acm conference on hyptertext and hypermedia (ht'10), New York, 2010, pp. 273-274.
GoogleT	Google Translate http://translate.google.com/
Gottron 2010	Thomas Gottron: External Plagiarism Detection Based on Standard IR Technology and Fast Recognition of Common Subsequences - Lab Report for PAN at CLEF 2010
GPSP	Glatt Plagiarism Screening Program, http://www.plagiarism.com/
Grozea 2010	Cristian Grozea and Marius Popescu: Encoplot-Performance in the Second International Plagiarism Detection Challenge - Lab Report for PAN at CLEF 2010
GTSearch	Google keresés idegen nyelvű szövegekben http://www.google.com/language_tools
Guerin 2012	Guerin, C.& Picard, M.: To match or not to match? Voice, concordancing and textmatching in doctoral writing, 5th International Plagiarism Coference, 2012
Gupta 2010	Parth Gupta, Sameer Rao, and Prasenjit Majumder: External Plagiarism Detection: N-Gram Approach Using Named Entity Recognizer - Lab Report for PAN at CLEF 2010
Guttenberg 2011	Guttenberg stürzt über Plagiat-Affäre, N24, 2011 http://www.n24.de/news/newsitem_6696857.html
Harris 1954	Zellig Harris: Distributional Structure, Word, Vol. 10, Nr. 23, 1954
Heintze 1996	Nevin Heintze: Scalable Document Fingerprinting, Proceedings Usenix Workshop On Electronic Commerce, 1996

iTrace — iTrace, Search by image and check visual arts submissions against a database of existing visual works
http://www.itrace.ac.uk/

iTunes 2012 — Brandon Griggs: Can Bruce Willis leave his iTunes music to his kids?, CNN, 2012
http://edition.cnn.com/2012/09/03/tech/web/bruce-willis-itunes/index.html

Juola 2006 — Patrick Juola, John Sofko and Patrick Brennan: A Prototype for Authorship Attribution Studies, Literary and Linguistic Computing 21:169-178
http://www.mathcs.duq.edu/~juola/papers.d/jsb-aut.pdf

Juola 2012 — Patrick Juola: An Overview of the Traditional Authorship Attribution
Subtask, Notebook for PAN at CLEF, 2012
http://www.uni-weimar.de/medien/webis/research/events/pan-12/pan12-papers-final/pan12-author-identification/juola12-overview-of-the-traditional-authorship-attribution-subtask.pdf
http://www.uni-weimar.de/medien/webis/research/events/pan-12/pan12-web/authorship.html

Károli 1591 — Károli Gáspár ford., Szent Biblia, 1591
http://www.cab.u-szeged.hu/WWW/books/Biblia/

Kasprzak 2010 — Jan Kasprzak and Michal Brandejs: Improving the Reliability of the Plagiarism Detection System - Lab Report for PAN at CLEF 2010

Kim 2003 — Young-Won Kim, Kyung-Ae Moon and Il-Seok Oh: A text watermarking algorithm based on word classification and inter-word space statistics, Proceedings Seventh International Conference on Document Analysis and Recognition, 2003

Kiss 1983 — É. Kiss Katalin: A magyar mondatszerkezet generatív leírása. Nyelvtudományi Értekezések 116. Akadémiai Kiadó, Budapest, 1983

KOPI — Kopi Online Plágiumkereső és Információs Portál
http://kopi.sztaki.hu

Mahmoud 2006 — Abdulmoneim Mahmoud, Translation and Foreign Language Reading Comprehension: A Neglected Didactic Procedure, English Teaching Forum, Volume 44, Number 4, 2006
http://eca.state.gov/forum/vols/vol44/no4/p28.htm

Martins 2005 — Bruno Martins and Mário J. Silva: Language identification in web pages, In Proceedings of the 2005 ACM symposium on Applied computing (SAC '05), Lorie M. Liebrock (Ed.), ACM, New York,

NY, USA, 764-768., 2005

MDR — Match Detect Reveal, Document Overlapping Detection

http://www.csse.monash.edu.au/projects/MDR/

Miháltz 2010 — Miháltz Márton: OpinHu - online szövegek többnyelv véleményelemzése, VII. Magyar Számítógépes Nyelvészeti Konferencia, Szeged, 2010.

http://opinhu.com/wp-content/uploads/2011/06/mihaltz_mszny2010.pdf

Miller 1995 — George A. Miller: WordNet - A Lexical Database for English, Communications of the ACM Vol. 38, No. 11: 39-41. (1995)

Monash — Monash University Melbourne, Faculty of Information Technology

http://www.infotech.monash.edu.au/about/schools/clayton/

Muhr 2010 — Markus Muhr, Roman Kern, Mario Zechner, and Michael Granitzer: External and Intrinsic Plagiarism Detection Using a Cross-Lingual Retrieval and Segmentation System - Lab Report for PAN at CLEF 2010

Németh 2004 — László Németh, Viktor Trón, Péter Halácsy, András Kornai, András Rung, and István Szakadát: Leveraging the open source ispell codebase for minority language analysis. In Proceedings of the SALTMIL Workshop at LREC 2004, pages 56–59.

Nida1964 — Eugene A. Nida, Toward a Science of Translating, 1964., Leiden: E. J. Brill.

Oxford — How many words are there in the English language?, Oxford University Press

http://oxforddictionaries.com/page/93

PAN 2010 — PAN 2010 Lab, Uncovering Plagiarism, Authorship, and Social Software Misuse

http://www.uni-weimar.de/medien/webis/research/events/pan-10/

Pataki 2002 — Pataki Máté: Szöveges dokumentumok darabolása és tömörítése hash-kódolással - darabolási technikák és másolatkeresés, Budapesti Műszaki és Gazdaságtudományi Egyetem, diplomadolgozat, 2002

http://dsd.sztaki.hu/people/mate_pataki/200201_DiplomaM25.pdf

Pataki 2003 — Máté Pataki: Plagiarism detection and document chunking methods, Proceedings of the twelfth international conference on World Wide Web, Budapest, 2003

http://www2003.org/cdrom/papers/poster/p186/p186-Pataki.html

Pataki 2011 — Pataki Máté, Vajna Miklós: Többnyelvű dokumentum nyelvének megállapítása, VIII. Magyar Számítógépes Nyelvészeti

Konferencia. Szeged, 2011.

Pataki 2012 — Máté Pataki: A new approach for searching translated plagiarism, 5th International Plagiarism Conference, Newcastle upon Tyne, UK, 2012

PCheck — Plagiarism Check using Google's Search API

http://hip2b2.yutivo.org/2006/03/25/plagiarism-check-using-googles-search-api

Pereira 2010 — Rafael Corezola Pereira, Viviane P. Moreira, and Renata Galante: UFRGS@PAN2010: Detecting External Plagiarism - Lab Report for PAN at CLEF 2010.

PFind — Plagiarism Finder

http://www.m4-software.de/en-index.htm

Picture-shark — Picture-shark

http://www.picture-shark.com

Ponta 2012 — Conflicting verdicts on Romanian prime minister's plagiarism, Nature, 2012

http://www.nature.com/news/conflicting-verdicts-on-romanian-prime-minister-s-plagiarism-1.11047

Potthast 2008 — Martin Potthast, Benno Stein and Maik Anderka, „A Wikipedia-Based Multilingual Retrieval Model", Advances in Information Retrieval, Lecture Notes in Computer Science, 2008, Volume 4956/2008, 522-530

Potthast 2010 — Martin Potthast, Alberto Barrón-Cedeño, Andreas Eiselt, Benno Stein, and Paolo Rosso, Overview of the 2nd International Competition on Plagiarism Detection

http://www.clef2010.org/resources/proceedings/clef2010labs_submission_125.pdf

Potthast 2011 — Martin Potthast, Andreas Eiselt, Alberto Barrón-Cedeño, Benno Stein, and Paolo Rosso: Overview of the 3rd International Competition on Plagiarism Detection, CLEF, 2011

http://www.uni-weimar.de/medien/webis/publications/papers/stein_2011t.pdf

PPKE 2011 — A szellemi alkotások védelméről és a plágium tilalmáról, PPKE, 2011

http://www.btk.ppke.hu/frontend_dev.php/hallgatoinknak/plagium tajekoztato

Prager 1999 — John M. Prager: Linguini: Language Identification for Multilingual Documents, Proceedings of the 32nd Annual Hawaii International Conference on Systems Sciences, 1999

PSearch — Plagiarism Search V 1.0.0,

http://baltic.cse.msu.edu/heynige1/Search/

Református 1993 Magyar Bobliatársulat ford., Szent Biblia, Budapest, 1993

Řehůřek 2009 Radim Řehůřek and Milan Kolkus: Language Identification on the Web: Extending the Dictionary Method, 10th International Conference on Intelligent Text Processing and Computational Linguistics (2009)

RFC Request For Comments, szabad terjesztésű ajánlások gyűjteménye, amelyek de facto szabványnak tekinthetők. A kutatás kezdetekor (2001. április) 2590 fájlt tartalmazott a gyűjtemény.

http://www.rfc-editor.org

Schmitt 2012 Súlyos plágiumgyanú Schmitt Pál doktori értekezése körül, HVG, 2012

http://hvg.hu/itthon/20120111_Schmitt_doktori_disszertacio_plagium

Solr Apache Solr, http://lucene.apache.org/solr/

Stamatatos 2008 Efstathios Stamatatos, A Survey of Modern Authorship Attribution Methods

http://www.icsd.aegean.gr/lecturers/stamatatos/papers/survey.pdf

Sváby 2012 Ellopták a népszerű magyar weboldalt, HVG, 2012

http://hvg.hu/Tudomany/20121007_videoklinika_koppintas

SzegedDe Wikipedia: Szeged szócikk német nyelven, http://de.wikipedia.org/wiki/Szeged (2011)

SzegedEn Wikipedia: Szeged szócikk angol nyelven, http://en.wikipedia.org/wiki/Szeged (2011)

SzegedFr Wikipedia: Szeged szócikk francia nyelven, http://fr.wikipedia.org/wiki/Szeged (2011)

SzegedHu Wikipedia: Szeged szócikk magyar nyelven, http://hu.wikipedia.org/wiki/Szeged (2011)

SzegedIt Wikipedia: Szeged szócikk olasz nyelven, http://it.wikipedia.org/wiki/Seghedino (2011)

SzegedP SzegedParalell Párhuzamos Korpusz

http://www.inf.u-szeged.hu/rgai/nlp?lang=en&page=corpus_paralell

Szókincs Szókincsméretek összehasonlító listája, Wikipédia

http://hu.wikipedia.org/wiki/Szókincsméretek_összehasonlító_listája

Szótár 2005 Szótárháború plágium gyanúja miatt, Magyar Nemzet, 2005. november

http://www.mno.hu/portal/321395

SZSzótár SZTAKI Szótár, MTA SZTAKI DSD

http://szotar.sztaki.hu/

Shivakumar 1995 Narayanan Shivakumar, Hector Garcia-Molina: SCAM: A Copy Detection Mechanism for Digital Documents, Department of Computer Science, Stanford University, CA 94305-2140, Proceedings of 2nd International Conference in Theory and Practice of Digital Libraries (DL'95), Austin, Texas, 1995.

Shivakumar 1995b Narayanan Shivakumar, Hector Garcia-Molina: The SCAM Approach To Copy Detection in Digital Libraries. Technical Report. Stanford InfoLab. (Publication Note: D-lib Magazine, November 1995.)

Shivakumar 1996 Narayanan Shivakumar, Hector Garcia-Molina: Building a Scalable and Accurate Copy Detection Mechanism, Department of Computer Science, Stanford, CA 94305, 1996.

Torrejón 2010 Diego Antonio Rodríguez Torrejón and José Manuel Martín Ramos: CoReMo System (Contextual Reference Monotony) - Lab Report for PAN at CLEF 2010

Tóth 2005 Tóth Péter, Fordításelmélet, 2005

http://dettk.ucoz.com/load/0-0-0-93-20

Tóth 2008 Krisztina Tóth, Richárd Farkas, András Kocsor: Hybrid algorithm for sentence alignment of Hungarian-English parallel corpora, Acta Cybernetica 18(3):463-478. (2008)

Turnitin Turnitin, Plagiarism Prevention and Online Grading

http://turnitin.com/

Turnitin 2011 A Comparison of Internet Sources for Secondary and Higher Education Students, White Paper , Plagiarism and the Web

UDHR The Universal Declaration of Human Rights, UN, 1948

http://www.un.org/en/documents/udhr/index.shtml

Unideb 2010 Plágium miatt visszavontak egy doktori címet a Debreceni Egyetemen, Lánchíd Rádió, 2010

http://www.lanchidradio.hu/kronika/plagium_miatt_visszavontak_egy_doktori_cimet_a_debreceni_egyetemen_20100309

Varga 2007 Dániel Varga, Péter Halácsy, András Kornai, Viktor Nagy, László Németh, Viktor Trón: Parallel corpora for medium density languages, Recent Advances in Natural Language Processing IV. Selected papers from RANLP-05 John Benjamins, 2007, 247-258

WaterMarks WaterMarks.ws: a tool that allows you to watermark your photos online

http://www.watermark.ws/

Weber 2010 Dr. Debora Weber-Wulff, „Results of the Plagiarism Detection System Test 2010”, http://plagiat.htw-berlin.de/software-en/2010-

2/, http://www.plagiarismtoday.com/2011/01/13/plagaware-takes-top-honors-in-plagiarism-checker-showdown/

Webfordítás — Online weblapfordító, szövegfordító, fordítási kereső és szótár
http://www.webforditas.hu/google-kereses.php

WikiDump — A Wikipedia teljes adatbázisa, XML formátumban
http://en.wikipedia.org/wiki/Wikipedia:Database_download

Wikihu 2011 — Több ezer szócikket töröltek a magyar Wikipédiából, Index, 2011
http://index.hu/tech/2011/01/06/tobb_ezer_szocikket_toroltek_a_magyar_wikipediabol/
http://hu.wikipedia.org/wiki/Wikip%C3%A9dia:A_VIL_copy_%C3%BCgy

Wikipedia — Wikipedia the free encyclopedia
http://en.wikipedia.org/

Xafopoulos 2004 — A. Xafopoulos, C. Kotropoulos, G. Almpanidis, I. Pitas: Language identification in web documents using discrete HMMs, Pattern Recognition, Volume 37, Issue 3, pp. 583-594, March 2004

Zou 2010 — Du Zou, Wei-jiang Long, and Zhang Ling: A Cluster-Based Plagiarism Detection Method - Lab Report for PAN at CLEF 2010

Zsibrita 2009 — Zsibrita János, Nagy István, Farkas Richárd, Magyar nyelvi elemző modulok az UIMA keret-rendszerhez, Magyar Számítógépes Nyelvészeti Konferencia 2009