Fordítástudomány X. (2008) 1. szám 18–41.

Újonnan bevezetett szóalakok és lemmák Dan Brown: *The Da Vinci Code* című művében és fordításaiban

Csernoch Mária

1. Bevezetés

Valamennyi természetes nyelven írott szöveg szellemi termék, és mint ilyen, jelen számítástudományi ismereteink nem elegendőek ahhoz, hogy teljes egészében modellezhető legyen egyetlen számítógépes programmal. A megoldást nem az egyetlen, minden szövegre alkalmazható modell keresése jelenti, hanem a részfeladatok megoldása lehet a közvetlen cél. Olyan kisebb problémák megoldására kell törekednünk, amelyek modellezése nem ütközik technikai akadályokba, és később ezen részmegoldások összeilleszthetők, hogy egy egységes nagyobb rendszert tudjunk létrehozni.

Egyik ilyen egyszerűsítése a problémának, hogy nem törekszünk a szövegek pontos modellezésére, hanem első-rendű statisztikai modelleket (Herdan 1960) használunk, amelyek figyelmen kívül hagyják a szövegnek mind a szintaktikai, mind a szemantikai kötöttségeit, és csakis az egyes szavak előfordulási gyakoriságát veszik figyelembe. Az eddigi tapasztalatok azt mutatják, hogy a szöveg szókészletével, a szókészlet gazdagságával kapcsolatos paraméterek igen jól reprodukálhatóak a különböző első-rendű statisztikai modellekkel (összefoglalót lásd: Oakes 1998 és Baayen 2001). Ahhoz azonban, hogy a szókészlet folyamatos változását is tudjuk követni nem elegendőek a korábban alkalmazott statikus első-rendű modellek (Baayen 1996, 2001), sokkal inkább a dinamikus modellek azok, amelyek a folyamatot is képesek követni (Csernoch 2006b, 2007b).

Ahogy az korábban ismertetésre került (Csernoch 2006b, 2007b), egy dinamikus modell segítségével képesek vagyunk tetszőleges számú mesterséges szöveg előállítására. Ezen mesterséges szövegeknek két alapvető tulajdonságát emelném ki, amelyek egyenes következményei annak, hogy egy első-rendű statisztikai modell alapján készülnek. A szavak véletlenszerűen vannak válogatva, így a keletkezett szöveg sem szintaktikailag, sem szemantikailag nem koherens, ugyanakkor megvan az a tulajdonsága, hogy a szavak száma és azok gyakorisága megegyezik az eredeti szövegben számolt adatokkal. Ez utóbbi tulajdonságot figyelembe véve képesek vagyunk az eredeti szöveggel összehasonlítható mesterséges szövegek előállítására. Mivel a mesterséges szöveg csak egy közelítése az eredeti szövegnek, ezért természetesen adódik, hogy eltérés van a két szöveg szókészletének alakulásában. Vizsgálatainkban éppen ezekre az eltérésekre keresünk rá, keressük a szövegnek azokat a szeleteit, ahol az eredeti szövegben mérhető változás tapasztalható a szókészlet gazdagságában.

Szövegek korábbi elemzése során, függetlenül attól, hogy statikus vagy dinamikus modellt alkalmaztunk, azt tapasztaltuk, hogy az eredeti szöveg és a modell (Baayen 1996, 2001), valamint az eredeti és a mesterséges szöveg (Csernoch 2006b, 2007a) közötti eltérés nem tulajdonítható a mondaton belüli szintaktikai, illetve szemantikai kötöttségeknek. Ezek az eltérések sokkal inkább szövegszerkezeti változásokból adódnak.

Az eltérések okának pontos meghatározására azonban a szókészletre vonatkozó globális vizsgálatok már nem bizonyultak elegendőnek, ezért a szókészlet méretére vonatkozó elemzéseket le kellett váltsa az újonnan bevezetett szavak számának statisztikai elemzése (Csernoch 2006b, 2007a).

Baayen (1996, 2001) korábbi kutatásai arra engednek következtetni, hogy egyrészt a szövegszintű változások azok, amelyek maguk után vonhatnak szókészletbeli változásokat, illetve a téma megváltozása is hasonló eredménnyel jár, történjenek ezek fejezethatáron vagy szövegek összefűzésekor. Korábbi, szövegek nem számítógéppel segített feldolgozása is ezen két megállapítás valamelyikét eredményezte.

Genette (1995) elképzelése szerint akkor történik mérhető változás a szókészlet gazdagságában, ha megjelenik a szövegben egy 'hosszabb' leírás, vagy stílusváltás történik. Ezzel szemben Petőfi (1990) úgy gondolja, hogy a fejezethatároknak olyan meghatározó szerepe van, hogy ezek is eredményezhetnek emelkedést az újonnan bevezetett szavak számában. Az alkalmazott dinamikus modell (Csernoch 2005, 2006b, 2007a) segítségével meg lehetett mutatni, hogy sem a témaváltásnak, sem a fejezethatárok megjelenésének nincs olyan hatása a szókészlet változására, mint a stílusváltásnak, illetve a leírások megjelenésének. A számítógéppel segített elemzések tehát Genette állításaival vannak összhangban.

A dinamikus modellek alkalmazásával kapott eredményeket felhasználva arra kerestem a választ, hogy az eredeti szöveget összehasonlítva annak fordításaival milyen változások tapasztalhatók az újonnan bevezetett szavak számának alakulásában, és hogy továbbra is fellelhetők lesznek-e az eredeti szöveg új szavakban gazdag szegmensei a fordításokban. Jelen projektben fordításon a fordításelmélet azon meghatározását értem, amely a szöveg bármilyen típusú módosítását fordításként fogadja el. Ennek megfelelően, az eredeti szöveg három különböző típusú adaptációjának, fordításának az elemzését végeztem el. A három adaptáció magába foglalja az eredeti szöveg

– idegen nyelvű fordításait,
– a lemmatizálás után keletkezett szövegeket, valamint
– a rövidített (condensed) szövegeket.

A statisztikai elemzéshez Dan Brown: *The Da Vinci Code* című művét és a mű néhány fordítását választottam (1. táblázat).

1. táblázat
Dan Brown *The Da Vinci Code* című művének az analízis során feldolgozott különböző fordításai

	teljes hosszúságú szöveg		rövidített szöveg	
	nem lemmatizált	lemmatizált	nem lemmatizált	lemmatizált
angol	+	+	+	+
magyar	+	+	+	+
francia	+			
német	+			

Mind az idegen nyelvű fordítás, mind a szöveg rövidítése olyan folyamat, melynek során a célközönség igényeit is figyelembe véve egy szöveget kell létrehozni, amelyben a lehetőségekhez mérten hűek maradunk az eredeti szöveghez. Ezen cél elérése érdekében a fordítás során számos olyan döntést kell meghozni, amely befolyásolhatja a fordítási folyamat eredményességét (Hatim és Mason 1990). Ezen két fordítás mindegyike olyan, hogy a célközönség befolyásolhatja a fordítás eredményét. Ezzel szemben a szöveg lemmatizálása pusztán elméleti kérdés, mivel ebben az esetben nem az olvasói célközönség az, akinek az igényeihez szeretnénk a művet alakítani, hanem egy érdekes nyelvészeti probléma megoldásához adhat magyarázatot. A három különböző fordítás elemzése így három különböző kérdésre keresi a választ.

Hasonlóan a korábbi vizsgálatokhoz, a *The Da Vinci Code* fordításainak statisztikai elemzésekor arra kerestem a választ, hogy a fordítások mennyiben követik az eredeti szöveg szókészletének változásait. Itt szeretném azonban megjegyezni, hogy a feldolgozásra kerülő műveket nagyban befolyásolta elérhetőségük. Ahogy az az 1. táblázatban is látható, nem sikerült az eredeti elképzelésem szerint valamennyi teljes hosszúságú szöveg rövidített változatát megszerezni, illetve a francia és a német szöveg lemmatizálását sem tudtam technikai okok miatt elvégezni.

Elsőként arra kerestem a választ, hogy az idegen nyelvű fordítások milyen hasonlóságot, esetleg eltérést mutatnak az eredeti műhöz, illetve az egymáshoz történő összehasonlításban. A teljes hosszúságú eredeti mű három idegen nyelvű fordításának az elemzését végeztem el (1. táblázat). Azért esett a választás a francia nyelvű szövegre, mert a regény 'másodlagos' nyelve a francia, és a szereplők számos esetben franciául beszélnek, valamint azért, mert az események egy jelentős hányada Franciaországban játszódik. A magyar fordítás esetében a nyelv sajátosságai játszottak döntő szerepet. Arra voltam kíváncsi, hogy egy agglutináló nyelvre történő fordítás mennyiben befolyásolhatja a szógazdag szegmensek megjelenését. Végezetül a német fordítás azért került a feldolgozott szövegek közé, mert mindkét fent említett szempontból semleges.

A rövidített szövegek létrehozása egy soha véget nem érő vitát indított el, amely folyamatosan megkérdőjelezi ezen szövegek létjogosultságát (lásd ké-

sőbb). Jelen tanulmány szerzője nem kíván csatlakozni a vitában résztvevőkhöz, egyetlen célja megmutatni, hogy a rövidítés hogyan befolyásolja a kiválasztott mű esetén a szógazdag szövegrészek jelenlétét. A rövidített szövegek a Reader's Digest sorozatban jelentek meg. Ellentétben a nyelvtanulásban használt rövidített szövegekkel, a szókészlet nagyságára nincs semmiféle megkötés, tetszőleges arányú rövidítés alkalmazható ezekben a szövegekben.

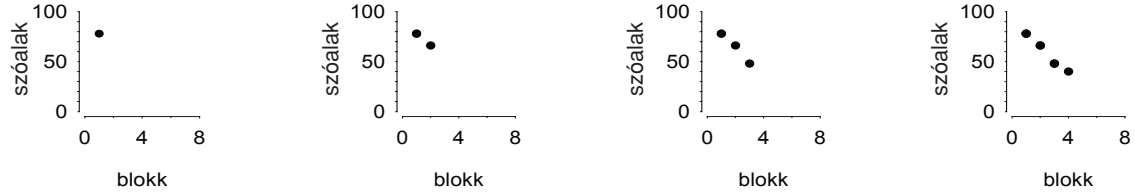
Végül a harmadik típusú fordítást, a lemmatizálást végeztük el a szöveggel. Szemben az előző két fordítással, a lemmatizálást nem hús-vér fordítók végezték, hanem erre alkalmas programok. Az angol szövegek lemmatizálásához szükséges programot (Rayson 2003, 2005) és tárterületet Paul Rayson, Lancaster University, biztosította számomra, míg a magyar nyelvű szövegek lemmatizálását Oravecz Csaba, a Nyelvtudományi Intézet munkatársa végezte el (Oravecz és Dienes 2002a, 2002b). Az eredeti szöveg a szóalakokat tartalmazza, amelyek magukkal hordozzák a saját affixumaikat. Jelen projekt kezdetén komoly vitát váltott ki, hogy a szókészlet méretében végbemenő változások elemzéséhez elegendő-e a szóalakok vizsgálata, vagy szükséges a szöveg lemmatizálása, és a lemmák számának a változását kell nyomon követni. Korábban elemzett szövegekkel sikerült megmutatni, hogy sem az angol, sem a magyar nyelvű szövegek ilyen jellegű vizsgálata nem indokolja a szövegek lemmatizálását (Csernoch 2006a). Ahhoz azonban, hogy további bizonyságot nyerjünk, úgy gondoltam, hogy a *The Da Vinci Code* angol és magyar nyelvű szövegén is elvégezem a lemmatizálást. Korábban arra is fény derült, hogy a nem lemmatizált és lemmatizált szövegek összehasonlításából nyert eredmények rámutathatnak a szövegnek azon szegmenseire, ahol a szerző az affixumokat használta stílusváltásra. A nem lemmatizált és lemmatizált szöveg további összehasonlítása azonban túlmutat jelen tanulmány keretein.

2. Módszerek

A szövegek digitalizálása után, függetlenül a nyelvtől, a szöveg hosszától, függetlenül attól, hogy lemmatizált a szöveg vagy sem, ugyanazt az eljárást alkalmaztuk. Minden esetben a DyMoCASAT (Dynamic Model for Computer Aided Statistical Analysis of Texts, Csernoch 2005, 2007b) nevű program végezte el az elemzést a megfelelő paraméterek beállításával.

Ezen eljárás első lépése a szöveg elemzése. Meg kellett határozni a szöveg hosszát (N), majd a különböző szavak (szóalak vagy lemma) számát ($V(N)$) és gyakoriságát. A második lépésben a szöveget feldaraboltuk egyenlő szélességű szeletekre, szövegintervallumokra, amelyeket blokkoknak neveztünk el. A blokkok hosszát h-val jelöltük, és rendszerint 100 szövegszó hosszúságúra állítottuk. Ezek után minden egyes blokkban meghatároztuk az újonnan megjelenő szavak számát. Újonnan bevezetett szón azt a szót értjük, amelyik az éppen aktuális blokkot megelőzően egyetlen blokkban sem jelent meg, tehát ez a blokk az, amelyben először fordul elő a szó az adott szövegben. Így minden egyes blokkhoz hozzárendelünk egy értéket ($ff(i)$, ahol $i = 1, \ldots, n$ és n a blokkok száma a szövegben), és ezen pontok együttesen alkotnak egy diszkrét

pontokból álló függvényt. Általánosságban az újonnan bevezetett szavak számának görbéje egy szigorúan monoton csökkenő függvény (1. ábra), de minden egyes szövegben találni olyan szövegszegmenseket, ahol ez a monotonitás megszakad, és egy hirtelen emelkedés, majd egy gyors csökkenés tapasztalható a függvényen (2A ábra és 3–8 ábrák).



1. ábra. Az újonnan bevezetett szóalakok száma az első négy blokkban (*The Da Vinci Code*, angol nyelvű, rövidített verzió)

A szövegek kezdetén az újonnan bevezetett szavak száma monoton csökkenő tendenciát követ, ahogy az 1. ábrán az első négy blokkhoz tartozó függvényértékek mutatják.

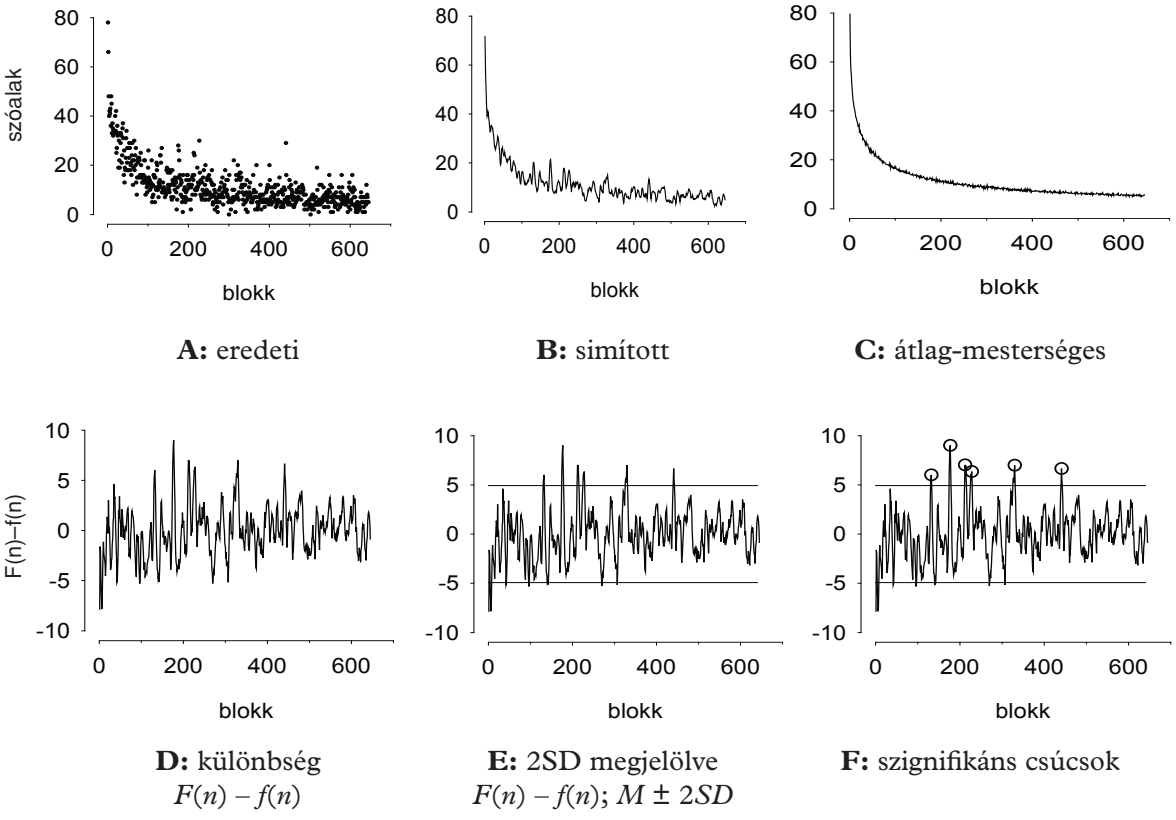
A következő lépésben a görbén jelenlévő zaj csökkentésére az *ff*(*i*) függvény simítását kellett elvégezni (2B ábra). Az eredmény az *F*(*i*) függvény, amely követi az *ff*(*i*) függvény menetét, de a véletlenszerű kiugrásokat már nem tartalmazza. A simítás után a korábbi diszkrét pontokból álló *ff*(*i*) függvényt az *F*(*i*) függvény helyettesíti, amelyen már tisztán kivehető az újonnan bevezetett szavak számának változása.

2.1. A szignifikáns változások az újonnan bevezetett szavak számában

A szöveg elemzése után a modell építését végeztük el, amelyet, az elemzéshez hasonlóan, a DyMoCASAT program végez el. A modell azon a feltételezésen alapszik, hogy egy szövegben a szavak megjelenése hipergeometrikus eloszlást követ (Csernoch 2006a, 2006b). A modell megépítéséhez a szavak visszatevés nélküli véletlenszerű válogatását kell elvégezni, mely válogatásnak a végeredménye egy mesterséges szöveg. Ez a mesterséges szöveg pontosan annyi szövegszót és szóalakot tartalmaz, mint az eredeti szöveg, és az egyes szóalakok gyakorisága is megegyezik az eredeti szöveg szóalakjainak gyakoriságával.

A következő lépésben a mesterséges szöveg elemzését végzi el a program. Hasonlóan az eredeti szöveghez, minden egyes blokkban meg kell számolni az újonnan bevezetett szóalakok számát és ezt a számot hozzá kell rendelni a megfelelő blokkhoz. A véletlen válogatás azonban éppen a véletlenszerűsége miatt okozhat váratlan kiugrásokat a függvényen. Ezen váratlan kiugrások elkerülése érdekében nem egyetlen mesterséges szöveget készíttettünk el a programmal, hanem összesen 100 ilyen szöveget, majd vettük minden egyes blokkban a 100 mesterséges szöveg újonnan bevezetett szavainak az átlagát, és az így létrehozott átlagfüggvény (*f*(*n*) függvény a 2C ábrán) volt az, amelyet a továbbiakban használtunk vizsgálatainkban.

A mesterséges szövegek átlagának képzése után az eredeti ($F(n)$) és a mesterséges ($f(n)$) szöveg különbségét vettük ($F(n) - f(n)$ függvény a 2D ábrán). A különbségfüggvény megrajzolásával már szemléltethető, hogy melyek azok a szövegszegmensek, amelyekben magasabb az újonnan bevezetett szavak száma, mint a közvetlen környezetükben. Ezek a szövegszegmensek úgy azonosíthatók, hogy a különbségfüggvényen megkeressük a függvény helyi maximumait.



2. ábra. Az újonnan bevezetett szavak számának elemzésénél alkalmazott eljárás főbb lépései

A helyi maximumok azonban nem minden esetben szignifikáns változás eredményei. Annak eldöntésére, hogy egy helyi maximum szignifikáns változást képvisel-e vagy sem, szükség volt a szignifikancia-szint meghatározására. Jelen projektben a sziginifikancia-szintet a különbségfüggvény átlagából (M) és szórásából (SD) számoltuk ki, és azokat az értékeket tekintettük szignifikánsnak, amelyek meghaladják az $M + 2SD$ értéket (2F ábra). Korábbi elemzések során azt tapasztaltuk, hogy akkor emelkedett meg az újonnan bevezetett szavak száma, ha a szövegben megjelent egy 'hosszabb' leírás, vagy stílusbeli váltás történt. Ezek a változások, az őket szemléltető kiugrásokkal a görbén, egyediek minden szövegben, és alkalmasak a szöveg azonosítására (Csernoch 2005, 2007b).

Az elemzés eredménye minden esetben egy függvény (2D–2F ábrák). Ennek a függvénynek a menete megjósolhatatlan, mivel nem tudni, hogy a mű szerzője mikor él azzal a lehetőséggel, hogy megnöveli az újonnan bevezetett

szavak számát. Cook (1994) szerint egy szövegfolyam, amelyik a 'schema refreshing' kategóriába sorolható, természetes módon eltér minden korábban ismert és használt kategória szerinti elvárástól. Ennek megfelelően egy irodalmi mű nem feltétlenül sorolható be egyetlen korábban ismert kategóriába sem, így természetesen a 'schema refreshing', mint jelző használható a nem létező kategóriába való besorolásra (éppen ez az, ami olyan értékessé tehet egy irodalmi művet). Így egy irodalmi műhöz létrehozott, az újonnan bevezetett szavak megjelenését leíró függvény mindig egyedi, az egyediségét éppen a függvény kiugrásai mutatják. A függvény kiugrásai legalább két paraméterrel jellemezhetők, ezek egyike a kiugrás intenzitása, a másik pedig a kiugrás szélessége. A kiugrás intenzitása a kiugrás magassága, ami nem más, mint az $F(n) - f(n)$ függvénynek a helyi maximuma. A kiugrás szélességén azoknak a blokkoknak a számát értjük, amelyekhez tartozó függvényérték a szignifikancia-szint felett van. Ha nem a blokkok, hanem a kiugrásban szereplő szövegszók számával kívánjuk megadni a kiugrás szélességét, akkor ez a blokkok száma szorozva a blokkok hosszával.

2.2. Szóalakok vagy lemmák

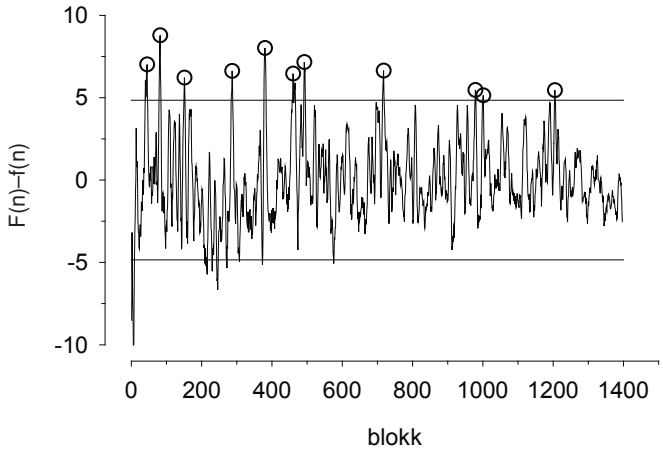
Ahogy a Bevezetésben már említettem, a projekt kezdeti szakaszában a szóalakok számának a változását vizsgáltuk. Felmerült azonban a kérdés, hogy mennyiben befolyásolhatja az analízis eredményét, ha a szóalakokat lemmákkal helyettesítjük. Számos vélemény hangzott el pro és kontra.

Az angol és a magyar szövegek teljes szófaji elemzését és lemmatizálását végül a Bevezetésben említett programokkal végeztük el. Mindkét esetben vertikális formában érkezett meg a lemmatizálás eredménye, amely oszlopok tartalmaztak az elemzés szempontjából irreleváns információkat: azonosító karakterek, mondat elejét-végét jelző információ, az eredeti szövegszó, valamint az eredeti szövegszó szófaji besorolása. Az elemzéshez meg kellett tartani a szövegszóból generált lemmát, valamint elő kellett állítani a lemma szófaji besorolását, tehát az eredeti szövegszó szófaji besorolásáról el kellett távolítani azokat az információs tageket, amelyek a szövegszót megkülönböztetik a hozzá tartozó lemmától. Az ily módon előállított lemmákból és szófaji tagekből összefűzéssel egy olyan mesterséges 'szót' tudtunk generálni, amelynek első fele a lemmát, a második pedig a lemma szófaji tagjét tartalmazza. Hasonlóan az eddig leírtakhoz, az így előállított szövegnek az elemzését, modellezését, az eredeti és a mesterséges szövegek összehasonlítását kellett elvégezni, ami ismételten a DyMoCASAT programmal történt.

3. Eredmények

Az elemzés az eredeti, teljes hosszúságú, nem lemmatizált angol nyelvű *The Da Vinci Code* című művel kezdődött, majd a további eredményeket ehhez hasonlítottuk. A szöveg hossza 139 700 és 139 800 szövegszó között van, amelyet a program lekerekített 139 700 szövegszóra, így 1397 darab, egyenként 100

szövegszót tartalmazó blokkot kaptunk. Ahogy a 3. ábra és a 2. és 3. táblázat mutatja, a függvény kiugrásai között 11 szignifikáns kiugrást sikerült azonosítani. Ezek a kiugrások igen eltérőek mind hosszban, mind intenzitásban. A 11 kiugrás közül 8 olyan, amelyik 500 vagy annál kevesebb szövegszó hosszúságú, és csak 3 olyan van, amelyik ennél szélesebb (3. táblázat). A széles kiugrások mind a szöveg első felében jelennek meg. A kiugrások intenzitását vizsgálva láthatjuk, hogy a szöveg első felében található kiugrásoknak nagyobb az intenzitása, mint azoknak, amelyek a szöveg második felében helyezkednek el. Ezek az adatok azt mutatják, hogy a szöveg végéhez közeledve nem találni olyan szövegszegmenseket, amelyek új szavak bevezetését indokolják, vagyis a korábban bevezetett szavak újrahasználatát tapasztalhatjuk.



3. ábra. A teljes hosszúságú, angol, nem lemmatizált *The Da Vinci Code* és a neki megfelelő mesterséges szöveg újonnan bevezetett szavai közötti eltérés

A 3. ábrán a vízszintes vonalak a szignifikancia-szintet mutatják. Jelen tanulmányban a szignifikancia-szintet az $M \pm 2SD$ értékre állítottuk. Azokat a kiugrásokat tekintjük szignifikánsnak, amelyek meghaladják az $M + 2SD$ értéket. Ennek megfelelően, az eredeti *The Da Vinci Code*-ban 11 szignifikáns kiugrás található, amelyek inkább a szöveg első felében helyezkednek el. Azok a szövegszegmensek generáltak szignifikáns kiugrásokat, amelyek a regény fővonalától eltérően valamilyen történelmi jellegű vagy akár jelenkori esemény, egy szereplő, egy helyszín részletes bemutatását, leírását tartalmazzák.

A 2. táblázatban azok az események vannak felsorolva, amelyek szignifikáns kiugrásokat eredményeztek. A táblázat 1. oszlopa a szövegszegmens könyvbeli fejezetszámát tartalmazza, a 2. oszlop magát az eseményt, a 3. oszlopban a kiugrás szélessége adott blokkokban meghatározva szóalakok esetén, míg a 4. oszlopban szintén a kiugrások szélessége adott a lemmatizált szöveg esetén.

A táblázat tartalmazza még a mért $2SD$ értéket mind a nem lemmatizált, mind a lemmatizált szöveg esetén. Ez az érték magas, ha a szöveg tartalmaz olyan kiugrást/kiugrásokat, amelyek intenzitása kiemelkedően magas a többi kiugráshoz viszonyítva. Egy magasabb $2SD$ érték gazdagabb, változatosabb

szókészlet jelenlétére utal, mint egy alacsonyabb. A *The Da Vinci Code*-ban az újonnan bevezetett szavak száma nem igazán magas, ami kiválóan szemléltethető a viszonylag alacsony 2*SD* értékkel.

2. táblázat

A szignifikáns kiugrások helye az angol, teljes hosszúságú, nem lemmatizált és lemmatizált *The Da Vinci Code*-ban

		angol, teljes hosszúságú	
		nem lemmatizált	lemmatizált
		n = 1397 2SD = 4,85	n = 1415 2SD = 4,4
fejezet	szövegszegmens	pozíció	pozíció
3	Párizs	40-46	37-47
5	Opus Dei	80-84	81-84
7	Sandrine telefonhívása		122-124
9	Sophie	150-152	152-153
20	Phi természete	285-288	287+291
28	boszorkányüldözés	378-383	383-387
34	Pápa, Castel Gandolfo	458-467	463-470
36	üldözés		486-488
37	Priory of Sion	491-495	497-499
54	Remy + Teabing háza		705
55	Jézus, kereszténység	715-719	723-726
76	Baphomet, Hebrew ábécé	979-980	989-993
79	Grand Masters	1001	1011-1013
87	padlás		1132-1135
97	Westminster	1204-1206	

Megjegyzés: A kiugrások helye azon blokkoknak a sorszámával van megadva, amelyeknél a kiugrás elérte a szignifikancia-szintet, valamint amelyeknél elhagyta azt.

A 3. táblázat összesítve mutatja azoknak a szövegszegmenseknek a darabszámát és szélességét, amelyek szignifikáns kiugrásokat eredményeztek. Az 1. oszlopban találni a kiugrások mért szélességét, a 2. oszlop azoknak a szövegszegmenseknek a számát mutatja, amelyek az adott szélességgel rendelkeznek, míg a 3. oszlop az 1. és 2. oszlopban szereplő számok szorzata, vagyis az adott szélességgel rendelkező szövegszegmensek össz-szélessége.

3. táblázat

Az eredeti, teljes hosszúságú *The Da Vinci Code* kiugrásainak szélessége és össz-szélessége

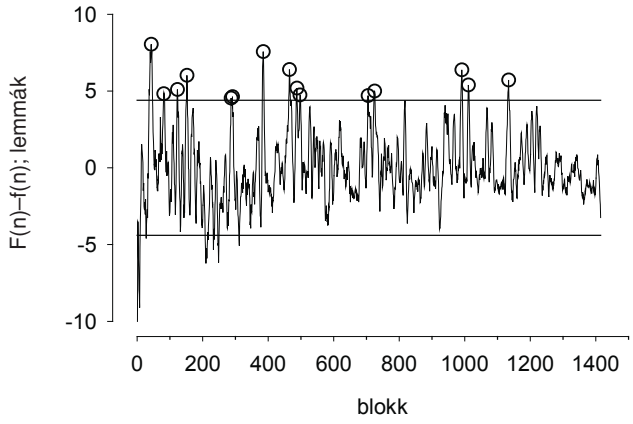
kiugrások szélessége	adott szélességű szövegszegmensek száma	adott szélességű szövegszegmensek össz-szélessége
1	1	1
2	1	2
3	2	6
4	1	4
5	3	15
6	1	6
7	1	7
10	1	10
	11	51

3.1. Teljes hosszúságú lemmatizált szöveg

A lemmatizált szöveg elemzése, ahogy már korábbi szövegek esetében tapasztalhattuk (Csernoch 2006a), nem mutatott számottevő eltérést az eredeti, nem lemmatizált szöveg elemzése során kapott eredményektől, ha továbbra is az újonnan bevezetett szavak számát vizsgáljuk. Kisebb eltérések tapasztalhatók, de összességében azt látni, hogy a kiugrások ugyanazoknál a szövegszegmenseknék jelennek meg, és a különbséget inkább a kiugrások intenzitásában lehet tapasztalni. A lemmatizált *The Da Vinci Code*-ban 3 új szignifikáns kiugrást találtunk, míg egyet elvesztettünk. Az eredeti szöveg további 10 szignifikáns kiugrása szignifikáns kiugrásként jelent meg a lemmatizált szövegben is. Új szignifikáns kiugrások megjelenése a lemmatizált szövegben azt jelenti, hogy a szóalakok affixumai felelősek a kiugrás elvesztéséért. Más oldalról viszont a lemmatizálás során is veszíthetünk el kiugrást, ám ezekben az esetekben éppen az affixumok hordozták az információt.

A 3. és a 4. ábrát, valamint a 2. táblázat 3. és 4. oszlopának adatait összehasonlítva láthatjuk, hogy a nem lemmatizált és a lemmatizált szöveg kiugrásai között csak kisebb eltérések fedezhetők fel.

Annak eldöntésére, hogy szükséges-e a szöveg lemmatizálása, nem lehet egyértelmű igennel vagy nemmel válaszolni. A kapott eredmények alapján azt viszont bizton állíthatjuk, hogy annak eldöntésére, hogy mi okozza a kiugrásokat, nem szükséges a szöveg lemmatizálása. Ahhoz azonban, hogy megtaláljuk valamennyi kiugrást, mind a nem lemmatizált, mind a lemmatizált szöveg elemzését el kell végezni.



4. ábra. A teljes hosszúságú, lemmatizált angol *The Da Vinci Code* szignifikáns kiugrásai

Korábbi eredményekhez hasonlóan (Csernoch 2006a) a magyar szövegek lemmatizálása sem eredményezett lényeges eltéréseket. A magyar nyelv agglutináló jellege miatt azonban két érdekességre felhívnám a figyelmet. Egyik oldalról igaz, hogy a szóalakok vizsgálatakor annak a valószínűsége, hogy a szöveg elején veszítünk el kiugrásokat nagyobb, mint annak, hogy a szöveg hátralevő részében. A nem lemmatizált szövegben a szöveg elején azok a kiugrások maradnak szignifikánsak, amelyek rendkívül 'erősek'. A magyar *The Da Vinci Code*-ban az első kiugrást ('Párizs') teljesen elveszítettük, a harmadik ('Sophie') és a negyedik ('Silas') kiugrás létezik, de nem éri el a szignifikancia-szintet, míg a második kiugrás ('Opus Dei') szignifikáns kiugrást eredményezett mind a lemmatizált, mind a nem lemmatizált szövegben (4. táblázat). Másrészről viszont a lemmatizálással információt veszíthetünk. A magyar szövegben magasabb azoknak a csúcsoknak a száma, amelyek szignifikánsak a nem lemmatizált szövegben és hiányoznak a lemmatizált szövegből, mint ahogy az angol szövegek esetén tapasztaltuk. Magyar szövegeknél tehát fokozottan igaz az, hogy az összes kiugrás megtalálásához szükséges lehet mind a nem lemmatizált, mind a lemmatizált szövegek elemzése.

4. táblázat

A szignifikáns kiugrások helye a nem lemmatizált és a lemmatizált magyar *The Da Vinci Code*-ban

		magyar, teljes hosszúságú	
		nem lemmatizált	lemmatizált
		n = 1199 2SD = 7,36	n = 1189 2SD = 5,99
fejezet	szövegszegmens	pozíció	pozíció
3	Párizs		33-36
5	Opus Dei	71	69-73
7	Sandrine telefonhívása	—	—

9	Sophie	133*	132
12	Silas gyerekkora	148*	146-150
20	Phi természete	—	—
28	boszorkányüldözés	325-328	321-326
34	Pápa, Castel Gandolfo	390-394	387-391
36	üldözés	409-411	
37	Priory of Sion	420	
51	Teabing	576	
54	Remy + Teabing háza		590
55	Jézus, kereszténység	610-614	608
66	keresés Teabing házában	744-745	
72	tükörírásos vers, jambikus pentameter	799-802	791-796
76	Baphomet, Hebrew ábécé	839-840	830-834
79	Grand Masters	—	—
84	Remy elengedi Silast	920-921	
87	padlás	963-966	
95	keresés a hálózaton	1023*	1013-1014
97	Westminster	—	—

Jelmagyarázat: — azok a szövegszegmensek, amelyek az angol szövegben kiugrásokat eredményeztek, de a magyar szövegben nem, * azok a szövegszegmensek, amelyekhez tartozó kiugrás csúcsa nem éri el a szignifikancia-szintet

3.2. Idegen nyelvű fordítások

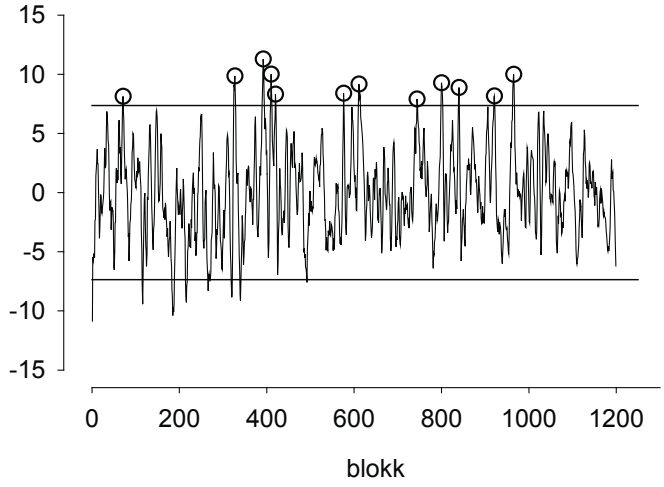
Irodalmi művek fordítása esetén mindig számolnunk kell azzal, hogy a szöveg tartalmazhat 'lefordíthatatlan' elemeket. Egy meghatározott elem 'visszaadhatatlansága' viszont nem mond ellent a fordíthatóság elvének, mivel a fordíthatóság a mű egészére kell hogy vonatkozzon (Simigné 2006). Az újonnan bevezetett szavak (akár szóalakok és lemmák) számának vizsgálata alkalmas lehet arra, hogy a fordítás egészéről, tehát szövegszinten kapjunk információt a fordítás milyenségéről. A veszteségek kompenzálása annyit jelent, hogy a fordítás során az eredeti szöveg valamely le nem fordítható elemét felcseréljük egy más elemmel úgy, hogy a fordítás összhangban maradjon az eredeti szöveg általános eszmei-művészi tartalmával (Bart és Klaudy 1986). Az újonnan bevezetett szavak számában bekövetkezett változások vizsgálatával meg lehet találni a szövegnek azokat a szegmenseit, ahol a fordítás eltér az eredeti szövegtől, meg tudjuk mutatni, hogy az eredeti szöveg szógazdag szegmenseit sikerült-e átörökíteni a fordított szövegbe.

3.2.1. A *The Da Vinci Code* magyar nyelvű fordítása

A 2. és a 4. táblázat, valamint a 3. és az 5. ábra adatai alapján össze tudjuk hasonlítani az eredeti angol szöveg és a magyar fordítás kiugrásait. Összességében

elmondható, hogy a magyar fordítás igen jó közelítéssel követi az angol szöveg szókészletében bekövetkezett változásokat, találni azonban eltéréseket is. Ahogy az előző fejezetben említettem, a magyar nem lemmatizált szöveg elején hiányoznak kiugrások, amelyek elvesztése a magyar nyelv agglutináló jellegével magyarázható. 'Párizs bemutatása' és 'Sophie megjelenése' nem eredményezett szignifikáns kiugrásokat a nem lemmatizált szövegben, míg a lemmatizálás eredményeként ezek a kiugrások megjelentek. Ezzel szemben az 'Opus Dei bevezetése' mindkét magyar szövegben megjelent szignifikáns kiugrásként, és az előzőekkel összhangban a lemmatizált szövegben szélesebb ez a kiugrás, mint a nem lemmatizált szövegben. Az angol szöveg kiugrásai kettő kivételtől – 'Grand Masters' és a 'Westminster' – megjelentek a magyar fordításban is. A 'Westminster leírásához' tartozó kiugrás csúcsa csaknem eléri a szignifikancia-szintet, tehát jelenléte erőteljes, míg a 'Grand Masters' alig kivehető kiugrást eredményezett a magyar szövegben.

Másrészről viszont a magyar szöveg tartalmaz olyan kiugrásokat, amelyek az eredeti angol szövegben nem jelentek meg szignifikáns kiugrásként. Ezen kiugrásokra továbbra is igaz, hogy egy-egy hosszabb leírás felelős a jelenlétükért, tehát ugyanúgy viselkednek, mint az eredeti szöveg kiugrásai. A magyar szövegben megjelent kiugrások a szövegbeli sorrendjükben a következők: 'Silas gyerekkora', 'üldözés Párizson keresztül', 'Teabing és Remy bemutatása', 'Teabing háza', 'tükörírásos vers', 'Remy elengedi Silast' és végül a 'könyvtári keresés'. Ezek a szignifikáns kiugrások arra engednek következtetni, hogy a kiugrásokhoz tartozó szövegszegmensek gazdagabb szókészlettel rendelkeznek, mint az eredeti szöveg ugyanezen szegmensei. A 'tükörírásos versrészlet' azonban eltérést mutat a többi, leírást tartalmazó szövegrésztől. A magyar szöveg tartalmazza a verset mind magyarul, mind angolul, és a verset közvetlenül követi a 'jambikus pentameter' bemutatása. A 'jambikus pentameter' önállóan nem volt képes szignifikáns kiugrást generálni, de az angol verssel együtt ez az egyébként is szógazdag szövegszegmens elegendő volt ahhoz, hogy a magyar fordításban megjelenjen ez a kiugrás.



5. ábra. A *The Da Vinci Code* magyar fordításának szignifikáns kiugrásai.

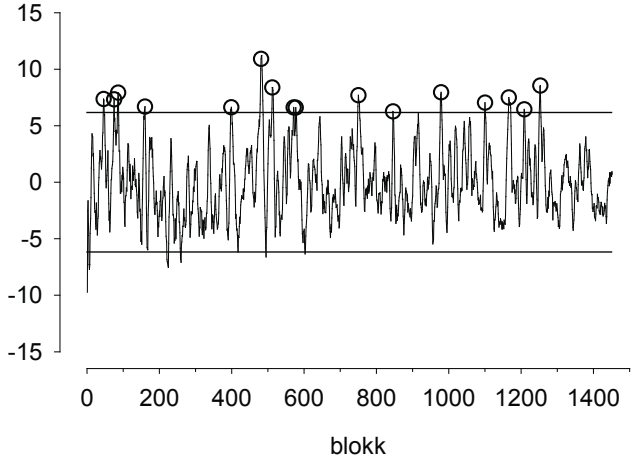
3.2.2. A *The Da Vinci Code* német fordítása

A német fordítás, szemben az összes többivel, volt az egyetlen, amely nem tartotta meg az eredeti *The Da Vinci Code* címet, hanem a *Sakrileg* szót választották a fordítók. Az elemzés során azonban egyértelművé vált, hogy nincs közvetlen kapcsolat a megváltoztatott cím és a szöveg szókészlete között.

A német fordítás egyetlen kivételtől eltekintve visszaadja az eredeti szöveg valamennyi szignifikáns kiugrását. Ezek a kiugrások, ahogy már korábban is láthattuk, mind egy-egy hosszabb szógazdag leírás eredményei. Az egyetlen hiányzó kiugrás a 76. fejezet 'Baphomet, Hebrew ábécé' szövegszegmense. Ennek a kiugrásnak a hiánya meglepő, mivel minden más teljes hosszúságú szövegben (2., 4. és 5. táblázat), és még a rövidített verziókban is megtalálható.

Hasonlóan a magyar fordításhoz, a német fordítás is tartalmaz olyan szignifikáns kiugrásokat, amelyek az eredeti angol szövegben nem érték el a szignifikancia-szintet. Ezen kiugrások megjelenését, mint a szöveg vizsgálata során minden esetben, szógazdag szövegszegmensek eredményezték. A szöveg elején, 'Párizs bemutatása' után a 'múzeum bemutatása" következik. A kiugrás érdekessége, hogy ez a szövegszegmens csak a német és a francia fordításban generált szignifikáns kiugrást, míg sem az angol, sem a magyar teljes hosszúságú szövegben nem. Az angol rövidített szövegben nyomaiban fellelhető a kiugrás, de a csúcsa nem éri el a szignifikancia-szintet. Találni még három kiugrást a német fordításban, amely egyetlen másik szövegben sem jelenik meg. Ezek a következők: 'Depository Bank of Zurich', 'Disney hatása a világra' és végezetül a 'Temple Church'.

Összességében elmondható, hogy a német fordítás igen jó megközelítését adta az eredeti szöveg szógazdag szövegszegmenseinek.



6. ábra. A *The Da Vinci Code* német fordításának szignifikáns kiugrásai

A 3. és a 6. ábrát, valamint a 2. és az 5. táblázatot összehasonlítva láthatjuk, hogy a német fordítás az eredeti angol szövegnek csak egyetlen kiugrását nem adta vissza, ezzel szemben találni négy olyan szövegszegmenst, amely gazdagabb szókészlettel rendelkezik, mint az eredeti angol szöveg.

5. táblázat

A teljes hosszúságú *The Da Vinci Code* német és francia fordításainak kiugrásai

		német	**francia**
		n = 1452 2SD = 6,17	n = 1361 2SD = 5,04
fejezet	szövegszegmens	pozíció	pozíció
3	Párizs	45-47	
4	múzeum	75	60
5	Opus Dei	85-87	71
9	Sophie	159-160	141
20	Phi természete	—	—
28	boszorkányüldözés	397-400	
34	Pápa, Castel Gandolfo	479-486	430
37	Priory of Sion	511-515	461
42	belépés a bankba, bank	571; 577	
55	Jézus, kereszténység	749-753	
61	Disney	846	
72	tükörírásos vers	977-980	
74	Hieros Gamos		922
76	Baphomet, Hebrew ábécé		944
79	Grand Masters	—	—
84	Temple Church	1100	
87	padlás	1165-1171	
92	keresés a hálózaton	1209	
95	keresés a hálózaton		1152
97	Westminster	1251	1162

3.2.3. A *The Da Vinci Code* francia fordítása

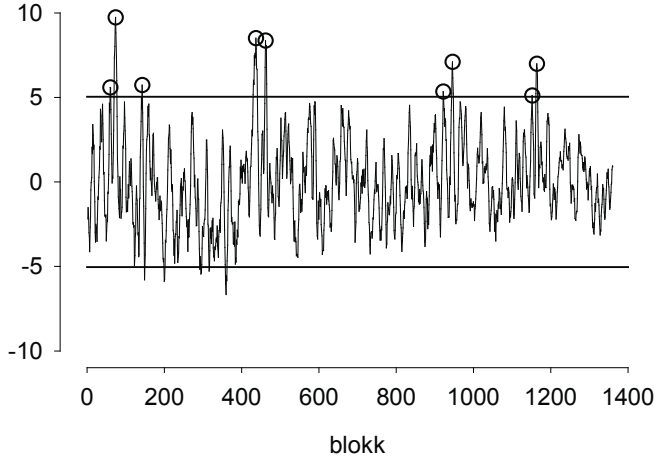
Számos érvet lehet felsorakoztatni a francia fordításnak az analízisbe történő integrálása mellett. Ezen érvek között kiemelt helyet foglal el az a tény, hogy az események egy jelentős része Franciaországban játszódik. A következő igen figyelemreméltó tény, hogy a regény 'másodlagos' nyelve a francia. A fordítások összehasonlításánál mindenképpen érdemes figyelembe venni az 'elsődleges' és egy 'másodlagos' nyelv együttes jelenlétét. Egy korábbi regényelemzés során (Csernoch 2006a) azt tapasztaltuk, hogy a műnek a 'másodlagos' nyelvre fordítása nagyon jól követte az eredeti mű szókészletbeli változásait, tehát közel azonos helyen és számban jelentek meg a kiugrások a két szövegben. Egyetlen szöveg vizsgálata azonban nem elegendő ahhoz, hogy bizonyítsuk, hogy a 'má-

sodlagos' nyelvre történő fordításnak egyenes következménye a szókészletbeli hasonlóságok jelenléte.

Az eredeti mű elején található, 'Párizst bemutató' szövegszegmens hiányzik a francia fordításból, míg a 'múzeum bemutatása' kiugrást eredményezett a görbén (7. ábra, 5. táblázat). A fordítónak azt a döntését, hogy nem találta indokoltnak a francia fordításban Párizs részletes bemutatását, el lehet fogadni. Ugyanakkor nehezen magyarázható, hogy a hasonlóan ismert múzeum részletes leírását, szemben az eredeti angol szöveggel, miért illesztette be a francia fordításba.

Az elemzés során nyolc további szignifikáns kiugrást sikerült találni (7. ábra, 5. táblázat). Az eredeti szöveg három szignifikáns kiugrása nem jelent meg a francia fordításban, míg létezik két olyan, amely az eredeti szövegben nem jelent meg. Az 'új' kiugrások egyike a 'könyvtári keresés', amely megjelent mind a német, mind a magyar fordításban. A másik 'új' kiugrás, a 'Hieros Gamos' a 74. fejezetben, csak ebben az egyetlen teljes hosszúságú szövegben jelent meg. A kiugrás jelenléte, hasonlóan a múzeumhoz tartozó kiugrás esetén, azzal magyarázható, hogy szemben a többi teljes hosszúságú szöveggel, nincs a kiugrást megelőzően olyan szövegszegmens, amellyel közös a szókészlete.

Az itt elmondottak alapján látható, hogy a francia fordítás kilenc kiugrása igen eltérő a többi teljes hosszúságú szöveg kiugrásaitól. A 7. ábra és az 5. táblázat adatait összehasonlítva az eredeti szöveg és a többi fordítás adataival az is látható, hogy a francia szöveg valamennyi szignifikáns kiugrása csak egyetlen blokk hosszúságú és ezek intenzitása is lényegesen kisebb a korábbi szövegeknél mért értékeknél. Ezek a tények együttesen azt jelzik, hogy a francia fordítás kiugrásai igen jelentéktelenek, alig találni a francia fordításban szógazdag szövegszegmenseket.



7. ábra. A *The Da Vinci Code* francia fordításának szignifikáns kiugrásai.

Az ábráról leolvasható, hogy mindössze kilenc szignifikáns detektálható a szövegben. Ezek a kiugrások is igen jelentéktelenek, mivel csak egyetlen blokk szélesek és intenzitásuk is gyenge.

3.2.4. Az idegen nyelvű fordítások összehasonlítása

A három idegen nyelvű fordítás közül a német fordítás követte leginkább az eredeti szöveg szókészletében bekövetkező változásokat. A magyar fordítás elemzése hasonló eredményt mutat, mivel csak eggyel több kiugrás hiányzott a német fordításhoz hasonlítva. Ezen túl mind a német, mind a magyar fordításban találni néhány olyan kiugrást, amelyek az eredeti szövegben nem jelentek meg szignifikáns kiugrásként. Az ezekhez a kiugrásokhoz tartozó szövegszegmensekben a fordítások gazdagabb szókészlettel rendelkeznek, mint az eredeti szöveg. Két olyan kiugrást is találtunk, amelyek jelen vannak mind a német, mind a magyar fordításban. Ezek egyike a 'tükörírásos vers', míg a másik a 'keresés a hálózaton'.

A fordításokban angolul megjelenő 'tükörírásos vers' a 'pentameter' leírásával együtt volt képes a szignifikáns csúcs generálására.

> *An ancient word of wisdom frees this scroll … and helps us keep her scatter'd family whole … a headstone praised by templars is the key … and atbash will reveal the truth to thee.*

A vers összesen 32 szövegszót és 30 különböző szóalakot tartalmaz. Ez a hosszúságú szövegszegmens egymagában nem lenne képes egy szignifikáns kiugrás generálására, mivel a görbe simítása éppen az ilyen jellegű kiugrások eliminálására szolgál. A pentameter leírásával együtt azonban egy négy blokk hosszúságú szövegszegmens keletkezett, amely már elég erős ahhoz, hogy szignifikáns kiugrásként jelenjen meg a görbén.

Amennyiben azt szeretnénk demonstrálni, hogy egyetlen szógazdag, ám rövid szövegszegmens nem képes szignifikáns kiugrást eredményezni, nem kell mást tennünk, mint a magyar és a német fordítást összehasonlítani az angol és a francia szöveggel. A 'pentameter' mind a négy szövegben szerepel, de sem az angol, sem a francia szövegben nem generált szignifikáns kiugrást. Az angol szövegben a vers megjelenése nem jelentette új szavak bevezetését, mivel olyan szavak szerepelnek a versben, amelyek többsége már korábban előfordult a műben, míg a francia fordítás nem adja meg a verset angolul, csak annak francia fordítása jelenik meg.

A francia szöveg bizonyult szógazdagságban a legszegényesebbnek, amit meglepőnek találtam. A fordító nem tudott előnyt kovácsolni abból, hogy az eredeti szerző a francia helyszíneket, a francia nyelvet intenzíven használja, így végeredményként egy az eredetinél szűkebb szókészlettel rendelkező szöveget hozott létre. Számos bizonyítékát találni azonban, hogy ez nem azért történt, mert a fordító nem volt tudatában a közös nyelvnek, kultúrának, a francia olvasóközönségnek. Ennek első bizonyítéka a fordítás elején hiányzó Párizs bemutatása. Nagyon jól reprezentálja a fordítói tudatosságot a 35. fejezet egyik jelenete, ahol a főszereplők vonatjegyet próbálnak vásárolni a Saint-Lazare pályaudvaron a következő vonatra. Az eredeti angol szövegben a következő vonat 3:06-kor induló Lille – Rapide. A német szövegben ugyanezt találjuk, de a magyar és francia fordításban nem. A magyar fordításban az indulási idő ugyan

ez, de a célállomás nem Lille, hanem Lyon. Első olvasáskor azt gondoltam, hogy a fordító egész egyszerűen összekeverte Lille-t Lyonnal, mivel az 54. fejezetben az eredeti 'Remy is Lyonnais' kifejezés helyett 'Rémy lille-i' kifejezés található. (Továbbra is maradt ez az érzésem, mivel valamennyi szöveg megtartotta a 'Lyonnais' szót, ami egy komornyik esetében, aki főz is, többet jelenthet, mint valaki, aki lyoni. A kifejezés utalhat a komornyik kulináris képességeire is.) A francia fordításban mind az indulási idő, mind a célállomás megváltozott. Az indulás időt 3:06-ról 5:38, a célállomást pedig Lille-ről Caenre változtatta a fordító. Ekkor már nem tudtam, hogy melyik vonat hova és mikor indul, és úgy döntöttem, hogy ellenőrzöm a vonatindulásokat az interneten.[1] A menetrend szerint a Saint-Lazare pályaudvarról nem indul vonat sem Lillebe, sem Lyonba, és biztosan nem ebben az időpontban, mivel éjszaka nem járnak személyszállító vonatok. Összességében tehát elmondható, hogy francia fordító tudatában van a műben jelenlévő francia kultúrának, nyelvnek, szokásoknak, de ennek ellenére nem tudta ezt kihasználni annak érdekében, hogy legalább olyan szógazdag művet hozzon létre, mint az eredeti.

Végül még egyetlen kiugrásról szeretnék említést tenni, amelyről az idegen nyelvű fordítások elemzésénél nem szóltam. Az ok, hogy egyetlen fordításban sem jelent meg ez a kiugrás. Ez a kiugrás az eredeti angol szövegben akkor jelent meg, amikor szerző a 'phi természetéről' tart előadást. A kiugrás mindössze négy blokk hosszúságú, ami nem különösebben hosszú, de nagyon erős az intenzitása (2. táblázat), és emiatt úgy gondolom, érdemes róla említést tenni. Rendkívül érdekes módon egyetlen fordító sem gondolta úgy, hogy ez a szövegszegmens olyan jelentőségű, hogy egy gazdagabb szókészlet megjelenése indokolt lenne.

3.3. A *The Da Vinci Code* rövidített verziói

Az elemzés ezen szakaszában a Reader's Digest gondozásában megjelent angol és magyar rövidített verziók digitalizálására és elemzésére került sor. Az irodalmi művek rövidített verzióinak kiadása egy folyamatos, véget nem érő vita témája, amelyben megkérdőjelezik az így keletkezett művek irodalmi értékét és ezen túl a kiadók által megcélzott olvasóközönséget. Az eredeti nemes cél az volt, hogy olyanok számára is elérhetővé tegyük az irodalmi műveket, akik nem tudják, vagy nem akarják valamilyen ok miatt az eredeti művet elolvasni, és ahelyett, hogy rájuk erőszakolnánk az eredeti szöveget, alternatíva lehet a rövidített szöveg létrehozása és olvasásuk. Azzal azonban számolnunk kell, hogy az irodalmi művek rövidítése veszteséggel jár, és mindenképpen számolnunk kell azzal a felmerülő jogos kérdéssel is, hogy az így született szöveg továbbra is tekinthető-e irodalmi műnek vagy sem. A következő megválaszolandó kérdés, hogy ezek a művek valóban a megcélzott olvasóközönséghez jutnak-e el, vagy azok is olvassák, akiknek lehetőségük lenne az eredeti, teljes hosszúságú mű elolvasására.

---

[1] http://www.sncf.com

3.3.1. A *The Da Vinci Code* rövidített angol verziója

A rövidített angol szöveg az eredeti teljes hosszúságú szöveg 46%-a. Nem kétséges, hogy egy ilyen volumenű rövidítés az eredeti szöveg durva csonkítása, mivel a szövegszók több mint felét elveszítettük (6. táblázat).

6. táblázat
A szövegszók, a különböző szóalakok és lemmák száma a teljes hosszúságú és a rövidített angol és magyar *The Da Vinci Code*-ban

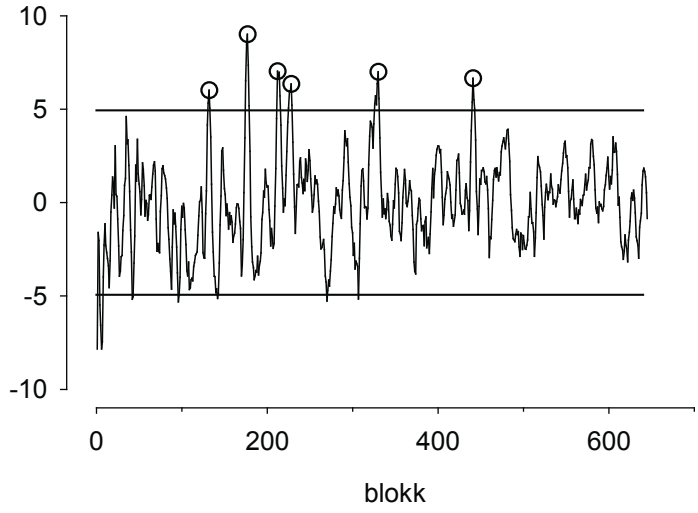
nyelv	hossz	szövegszó		szóalak		lemma	
angol	teljes	1397	46%	11442	64%	10300	64%
	rövidített	645		7333		6584	
magyar	teljes	1199	45%	25394	55%	13888	61%
	rövidített	543		14041		8414	

Az elemzés végső célja az volt, hogy a rövidített szövegek szókészletét össze tudjuk hasonlítani a teljes hosszúságú szövegekkel és egymással is. Ennek megfelelően az elemzés minden egyes lépése ismét a DyMoCASAT-tal került kivitelezésre.

A rövidített nem lemmatizált angol szövegben 6 szignifikáns kiugrást találni, míg a lemmatizált szövegben, a szöveg elején megjelent még egy kiugrás (egy blokk hosszúságú). A rövidített angol szövegben a kiugrások száma a felére csökkent a teljes hosszúságú szöveghez képest. Nemcsak a kiugrások száma, de azok hossza is csökkent, míg a kiugrások intenzitása nem feltétlenül (8. táblázat, 3. és 8. ábra). A megmaradt kiugrások közül négy olyan szövegszegmenst reprezentál, amely egy-egy történelmi eseményt ír le, míg három jelenkori eseményt, vagy történelmi és jelenkori eseményt együttesen ír le.

Az azonos nyelvű szövegek esetén egy másik összehasonlítható paraméter a kiugrások össz-szélessége (3. és 8. táblázat). Az eredeti szövegben a kiugrások össz-szélessége a teljes szöveg 3,65%-a volt, míg a rövidített verzió esetén ez az érték 4,19%, ami azt jelenti, hogy ezt a paramétert tekintve nem nagyobb a veszteség, mint a szavak számában.

A teljes hosszúságú és a rövidített szövegek különbségfüggvényeinek kétszeres szórását összehasonlítva azt tapasztalhatjuk, hogy a rövidítés nem eredményezte a szórás csökkenését, inkább egyfajta gyenge növekedés mérhető. Ez a két paraméter azt mutatja, hogy annak ellenére, hogy a szövegszók számában igen komoly csökkenés tapasztalható, a változatosságra utaló szórás akár emelkedhet is. Összességében ez azt jelenti, hogy a 'nem-kiugrásokhoz' tartozó szövegszegmensek azok, amelyek komoly csonkítást szenvedtek el, ami azt eredményezi, hogy a kiugrásoktól eltekintve, a maradék szövegszegmensek közelebb állnak a véletlen válogatáshoz, mint az eredeti szöveg (az összehasonlításhoz nézzük meg a 3. és a 8. ábrát).



8. ábra. A rövidített angol *The Da Vinci Code* szignifikáns kiugrásai

3.3.2. A magyar és az angol rövidített szöveg összehasonlítása

Mind az angol, mind a magyar rövidített szöveg 14 fejezetre van osztva, szemben az eredeti 105-tel. Mivel mind a teljes hosszúságú, mind a rövidített szövegekben a kiugrások ugyanazokhoz a szövegszegmensekhez, eseményekhez tartoznak, így ismételten sikerült megmutatni, hogy a fejezethatárok puszta jelenlétükkel nem képesek a szóalakok számának az emelésére.

A magyar rövidített verzióban sem 'Sophie', sem a 'Baphomet' leírása nem eredményezett szignifikáns kiugrást, másrészről viszont megjelent az a jelenet, amikor Aringarosa átveszi a kötvényeket. Ennek a kiugrásnak az az érdekessége, hogy egyetlen más fordításban sem generált szignifikáns kiugrást, csak a magyar rövidített, nem lemmatizált és lemmatizált szövegekben.

Arra voltam még kíváncsi, hogy az újonnan bevezetett szavak számának vizsgálata alkalmas lehet-e arra, hogy a másodrendű fordításoknál meghatározzuk közvetlen forrásukat. Jelen esetben az volt a kérdés, hogy a rövidített magyar szöveg közvetlen forrása a teljes hosszúságú magyar szöveg, vagy a rövidített angol szöveg volt-e. Mivel mind a teljes hosszúságú magyar szöveg, mind a rövidített angol szöveg egy-egy adaptációja az eredeti szövegnek, a hasonlóságok és esetleges eltérések az új szavak számának alakulásában segítségünkre lehetnek a másodrendű fordítások közvetlen forrásának meghatározásában.

7. táblázat

A rövidített angol és magyar *The Da Vinci Code* szignifikáns kiugrásai

	angol		**magyar**	
	nem lemmatizált	lemmatizált	nem lemmatizált	lemmatizált
	n = 645 2SD = 4,93	n = 654 2SD = 4,62	n = 543 2SD = 7,28	n = 538 2SD = 5,79
szövegszegmens	pozíció	pozíció	pozíció	pozíció
Párizs		21		
Sophie	131-133	132-134		
boszorkány-üldözés	175-179	177-181	147-148	145-147
Pápa, Castel Gandolfo	212-215	214-217	214-217	1750177
Priory of Sion	225-229	2270229	186-191	186-188
Aringarosa			207-208	206-206
Jézus, kereszténység	326-331	329-334	275-278	272-275
Baphomet, Hebrew ábécé	440-443			

8. táblázat

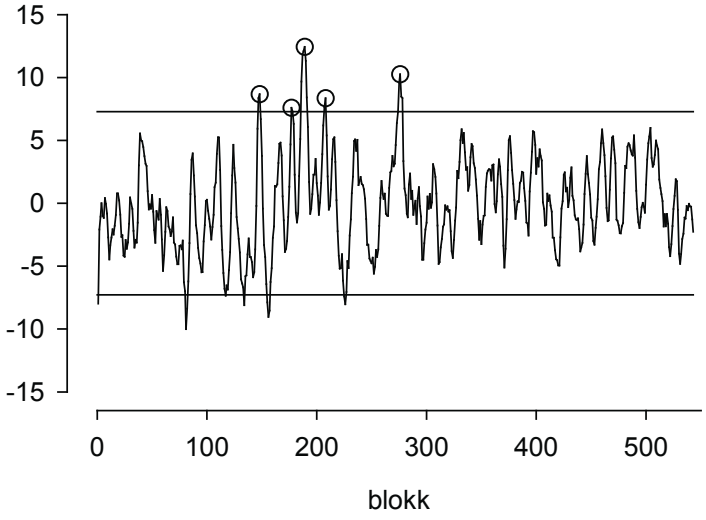
A rövidített angol szöveg kiugrásainak szélessége és össz-szélessége

kiugrások hossza	az azonos szélességű kiugrások száma	össz-szélesség
3	1	3
4	2	8
5	2	10
6	1	6
	6	27

Három olyan eseményt találni, amely mindkét teljes hosszúságú szövegben megtalálható, de nem szerepelnek a rövidített verziókban (az 'Opus Dei', a 'padlás' és a 'Westminster', 2. és 4. táblázat). Másrészről viszont találni egy olyan kiugrást, ami alig észrevehető a teljes hosszúságú magyar szövegben, de tisztán jelen van a rövidített verzióban. Ez a szövegszegmens ad leírást a 'Priory of Sion'-ról. Továbbá az is igaz, hogy a teljes hosszúságú magyar szövegnek azon kiugrásai közül, amelyek az eredeti szövegben nem szerepeltek, egyetlen-egy sem jelent meg a rövidített verzióban.

Következésként elmondhatjuk, hogy a szövegek szignifikáns kiugrásainak összehasonlítása segítség lehet annak eldöntésében, hogy egy másodlagos fordí-

tásnak mi a közvetlen forrása. Az állítás további igazolására a hagyományosnak számító módszer, a szövegek részleteikben történő összehasonlítása alkalmazható. Ahogy már korábban is említettem, a vonatjegy-vásárlási epizód összehasonlítását érdemes elvégezni. A rövidített magyar verzióban, a teljes hosszúságú és a rövidített angol verzióhoz hasonlóan, Lille-be kérik a jegyet, szemben a teljes hosszúságú magyar szöveggel.



9. ábra. A magyar rövidített *The Da Vinci Code* kiugrásai

4. Összegzés

Korábbi elemzések során sikerült megmutatni, hogy az első-rendű statisztikai modellek alkalmasak arra, hogy a szövegben bekövetkezett szókészletet érintő változásokra rámutassunk. Elsődleges célom az volt, hogy egy szövegben az újonnan bevezetett szavak számának a változásait vizsgáljam egy olyan dinamikus modellel, amely a szavaknak a szövegen belüli hipergeometrikus eloszlását feltételezi. Egy olyan modell, amely bemenő paraméterként az eredeti szöveg szóalakjainak gyakoriságát veszi figyelembe, természetesen nem képes a szöveget hűen visszaadni. Következésképpen az eredeti szöveg és a modell által generált szöveg között mindig lesznek eltérések. A két szöveg összehasonlításakor azonban hozzájuthatunk az eredeti szövegnek azokhoz a szógazdag szegmenseihez, ahol ezen eltérések számokban kifejezhetőek. Genette (1995) korábbi észrevételeivel összhangban azt tapasztaltuk, hogy ezek a szógazdag szegmensek akkor jelennek meg egy irodalmi műben, amikor az esemény illusztrációként szerepel, vagy olyan esemény, amely nem kapcsolódik szervesen a történethez, nincs a későbbiekben folytatása.

Az elemzéshez Dan Brown *The Da Vinci Code* című művét választottam. Az eredeti mű elemzésén túl még elvégeztem az elemzést három idegen nyelvű fordításán, az angol és magyar lemmatizált szövegen, valamint az angol és magyar rövidített verziókon is.

Az eredeti angol szöveg elemzését a lemmatizált verzió elemzése követte. Korábbi tapasztalataink azt mutatták, hogy a nem lemmatizált és a lemmatizált szövegek nem mutatnak lényeges eltérést. Így volt ez a *The Da Vinci Code* esetében is. A magyar nyelvű szövegek esetében is hasonló eredményekhez jutottunk. Ennek elvégzésére azért volt szükség még a vizsgálat elején, hogy meg tudjuk mutatni, hogy a továbbiakban elegendő a szóalakok vizsgálata, nem feltétlenül szükséges a szövegek lemmatizálása. A nem lemmatizált szövegek elemzése mellett a lemmatizált szövegek elemzésére abban az esetben lehet szükség, ha a szöveg összes új, szóban gazdag szegmensét meg szeretnénk találni.

Annak bizonyítására, hogy az újonnan bevezetett szavak számában bekövetkezett változások a szöveg mondatszintű kötöttségein túlmutatnak, a szövegek idegen nyelvű fordításait hasonlítottuk össze. Az eredeti elképzelés az volt, hogy ha a különböző idegen nyelvű fordításokban a szógazdag szövegszegmensek a szövegnek ugyanazon pontjain jelennek meg, akkor az újonnan bevezetett szavak számát a nyelv szintaktikai megszorításai nem befolyásolják (legalábbis számottevően nem). Ahogyan már korábbi vizsgálataink során is tapasztaltuk, a *The Da Vinci Code* szógazdag szövegszegmensei a különböző fordításokban közel azonos helyen jelentek meg. Módszerünk segítségével azonban sikerült találni olyan szövegszegmenseket is, ahol az egyes fordításokban szegényebb, esetleg gazdagabb szókészlettel dolgozott a szerző, mint az eredeti műben. A következő lépésként igyekeztünk magyarázatot találni az eltérések okaira. Az eltéréseket a természetes nyelvű szövegek és a mesterséges szöveg összehasonlításával sikerült megtalálni. Ezek az eltérések arra adnak számszerű összefüggést, hogy az egyes fordítások mennyire követik az eredeti szöveg szókészletbeli változásait.

Végezetül az angol és a magyar szövegek rövidített verzióinak az elemzésére került sor. Azt találtuk, hogy a kiugrások száma, a kiugrások össz-szélessége követte a szövegszók számában bekövetkezett változásokat. Ezzel szemben a kiugrások intenzitása nem mutatott eltérést a rövidített szövegekben az eredeti szövegekben mért értékekhez képest. Ez azt jelenti, hogy a kiugrások közötti szövegszegmensek rendkívül szegényes szókészletet használnak, amely a puszta tények ismertetésére elegendő.

Irodalom

Baayen, R. H. 1996. The Effect of Lexical Specialization on the Growth Curve of the Vocabulary. *Computational Linguistics* Vol. 22. 455-480.

Baayen, R. H. 2001. *Word Frequency Distributions*. Dordrecht, Netherlands: Kluwer Academic Publishers.

Bart I., Klaudy K. (szerk.) 1986. *A fordítás tudománya*. Budapest: Tankönyvkiadó.

Cook, G. 1994. *Discourse and Literature: The Interplay of Form and Mind*. Oxford: Oxford University Press.

Csernoch M. 2005. Természetes nyelvi szövegek összehasonlítása első-rendű statisztikai modellekkel. *Publicationes Universitatis Miskolcinensis, Sectio Philosophica, Tomus X. – Fasciculus 3*. Miskolc: Miskolci Egyetem. 3–26.

Csernoch M. 2006a. The introduction of word types and lemmas in novels, short stories and their translations. http://www.allc-ach2006.colloques.paris-sorbone.fr/DHs.pdf.

Digital Humanities 2006. The First International Conference of the Alliance of Digital Humanities Organisations. (5–9 July 2006, Paris)
Csernoch M. 2006b. Frequency-based Dynamic Models for the Analysis of English and Hungarian Literary Works and Coursebooks for English as a Second Language. *Teaching Mathematics and Computer Science*. Debrecen: University of Debrecen. 53-70.
Csernoch M. 2007a. Seasonalities in the Introduction of Word-types in Literary Works. *Publicationes Universitatis Miskolcinensis, Sectio Philosophica, Tomus XI. – Fasciculus 3.* Miskolc: Miskolci Egyetem.11–34.
Csernoch M. 2007b. Dinamikusan kezelhető statisztikai modellek irodalmi művek szóalakjainak vizsgálatára. *Alkalmazott Matematikai Lapok* 24. 57–77.
Genette, G. 1995. *Narrative Discourse*. Ithaca, NY: Cornell University Press.
Hatim, B., Mason, I. 1990. *Discourse and the Translator.* Harlow: Longman.
Herdan, G. 1960. *Type-token Mathematics*. The Hague: Mouton.
Oakes, M. P. 1998. *Statistics or Corpus Linguistics*. Edinburgh University Press.
Oravecz Cs., Dienes P. 2002a. Large scale morphosyntactic annotation of the Hungarian National Corpus. In: Hollósi, B., Kiss-Gulyás, J. (eds.) *Studies in Linguistics*, Volume VI., pages 277–298, Debrecen, 2002.
Oravecz Cs., Dienes P. 2002b. Efficient Stochastic Part-of-Speech tagging for Hungarian. In: *Proceedings of the Third International Conference on Language Resources and Evaluation*, Las Palmas, 2002. 710–717
Petőfi S. J. 1990. *Szöveg, szövegtan, műelemzés*. Budapest: Országos Pedagógiai Intézet.
Rayson, P. 2003. *Matrix: A statistical method and software tool for linguistic analysis through corpus comparison*. Ph.D. thesis, Lancaster University.
Rayson, P. 2005. *Wmatrix: a web-based corpus processing environment*. Computing Department, Lancaster University. http://www.comp.lancs.ac.uk/ucrel/wmatrix/
Simigné F. S. 2006. *A fordítás mint közvetítés*. Miskolc: Studió.

Források

Brown, D. 2003. *The Da Vinci Code*. New York: Doubleday.
Brown, D. 2003. *The Da Vinci Code*. Translated by Bori E. 2004. *A Da Vinci-kód*. Budapest: Gabo.
Brown, D. 2003. *The Da Vinci Code*. Translated by Roche, D. 2004. *Da Vinci Code*. Paris: JC Lattes.
Brown, D. 2003. *The Da Vinci Code*. Translated by van Poll, P. 2004. *Sakrileg*. Germany: Gustav Lübbe Verlag.
Brown, D. 2003. *The Da Vinci Code.* 2004. United States of America: Reader's Digest Association Inc.
Brown, D. 2003. *The Da Vinci Code*. *A Da Vinci-kód*. 2005. Budapest: Reader's Digest Kiadó Kft.