Irodalmi művek szókészletének statisztikai elemzése és matematikai modellezése

Doktori (PhD) értekezés

Csernoch László Józsefné

Debreceni Egyetem
Természettudományi Kar
Debrecen, 2005.

Irodalmi művek szókészletének statisztikai elemzése és matematikai modellezése

Doktori (PhD) értekezés

Csernoch László Józsefné

Témavezető: Dr. Arató Mátyás

Debreceni Egyetem
Természettudományi Kar
Debrecen, 2005.

Ezen értekezést a Debreceni Egyetem TTK Matematika és Számítástudományok Doktori Iskola Didaktika programja keretében készítettem a Debreceni Egyetem TTK doktori (PhD) fokozatának elnyerése céljából.

Debrecen, 2005. június 25.

a jelölt aláírása

Tanúsítom, hogy Csernoch László Józsefné doktorjelölt 2003-2005. között a fent megnevezett Doktori Iskola Didaktika programjának keretében irányításommal végezte munkáját. Az értekezésben foglalt eredményekhez a jelölt önálló alkotó tevékenységével meghatározóan hozzájárult. Az értekezés elfogadását javasolom.

Debrecen, 2005. június 25.

a témavezető aláírása

Tartalomjegyzék

Közlemények jegyzéke

I. Bevezetés

> „A szövegnyelvészet nem lehet valamely egységes elmélet vagy módszer neve. Inkább minden olyan munkára vonatkozik a nyelvvel foglalkozó tudományágak körén belül, amely érdeklődésének középpontjába a szöveget állítja."
>
> Teun van Dijk (1979)

A számítógépes nyelvészet mozgatója a kezdetektől, a múlt század ötvenes éveitől, a gépi fordítás megvalósítása (machine translation) iránti igény volt, mivel már a számítógépek megjelenése előtt is keresték azokat a módszereket, amelyekre az egyhangú munkát végző fordítók régóta várták a megoldást. Szemben a korábbi elképzelésekkel, már az ötvenes évek végére megfogalmazódott, hogy egy egyszerű szavankénti átírás nem adhat megfelelő kimenetet egy fordítási problémára (I.B.M., 1959). A hatvanas évek közepére az is nyilvánvalóvá vált, hogy a számítógép még sokáig nem lesz képes emberi felügyelet nélkül jó minőségű fordítást készíteni egy szövegről (Prószéky, 1989; Church és Mercer, 1994; Prószéky és Kis, 1999).

Az ezredfordulóhoz közeledve, amikor a számítógépes nyelvészet már nem kizárólag az angol nyelvterületre korlátozódott, ismét felerősödött a fordítás iránti igény. A gépi fordítást ugyan nem, de a gépi fordítás során felmerülő számos részfeladatot sikerült megoldani. A részfeladatok a későbbiekben a számítógépes nyelvészet egy-egy résztudományává nőttek ki magukat. Ezek közül néhány: korpusznyelvészet, lexikográfia, lexikológia, digitális-szöveg

kódolás, morfológiai elemzések, stylometry, szerző azonosítás, stilisztika, beszédfeldolgozás, információ visszanyerés, szemantikai elemzések, fordítás.

- Korpusznyelvészet: a szűkebb értelemben vett korpusznyelvészet a számítógéppel feldolgozható szövegkorpuszok előállításának tudománya.
- Lexikográfia (lexicography): a számítógépes szótárkészítés tudománya.
- Lexikológia (lexicology): a szövegkorpuszok alapján végzett főképp szavak és kifejezések használatára vonatkozó szintaktikai és szemantikai információ szerzés.
- Digitális-szöveg kódolás: a digitalizált szövegek egységes kódolására vonatkozó törekvések.
- Morfológiai elemzések: ide sorolhatók a ténylegesen morfológiai elemzést, lemmatizálást, annotálást végző programok, valamint a szóhatárokon túl nyúló elemzők fejlesztése. A mindennapi életben talán az egyik leggyakrabban használt alkalmazás, hiszen a szövegszerkesztő programok (áldott és átkos) helyesírás ellenőrzői mind ezen tudományterület eredményeit használják fel.
- Információ visszanyerés (information retreaval) tudománya napjainkra szervesen összefonódott a szemantikai jellegű elemzésekkel. A talán leglátványosabb, sokak által ismert, a gyakorlatban is bárki által felhasználható eredménye ennek a tudományterületnek az internetes tematikus keresők. Nem kisebb jelentőségű, de talán nem annyira népszerű másik nagy „fogyasztója” az eddigi eredményeknek az EU, ahol a keletkező dokumentumok tartalmi kivonatolása mára már nemcsak emberi munka eredménye.
- Szemantikai elemzésekkel foglalkozók két nagy csoportba oszthatók.
  Az egyik csoport a szemantikai hálók megépítésével, speciális információkat tartalmazó machine-readable szótárak létrehozásával,

próbálja a szavak, kifejezések közötti szemantikai kapcsolatokat tárolni, majd ezekből a szótárakból visszakeresni. Az egyik legnagyobb horderejű, ezen az elven működő program a WordNet (Fellbaum, 1998).

A másik csoport hívei azt tartják, hogy valamennyi kapcsolatot tárolni lehetetlenség. Ezért ezek un. knowledge-poor megközelítések a digitális adatbázisokban tárolt hatalmas mennyiségű szövegekből próbálnak szemantikai jellegű információt visszanyerni. Azon a hipotézisen alapul ez a megközelítés, hogy a hasonló módon használt szavak hasonló szemantikai információt hordoznak, tehát statisztikai alapon meghatározhatók. Napjaink egyik legjelentősebb képviselője ennek az elképzelésnek Grefenstette (1994).

- Fordítás: mára már csak egy tudományterület a sok között és nem az egyetlen mozgató rugója a számítógépes nyelvészetnek. Az azonban továbbra is igaz, hogy a fordítás a fentebb említett részterületek eredményeit intenzíven használja. Továbbra is cél olyan fordító programok létrehozása, amelyek automatikusan állítják elő a célnyelvi szöveget (teljesen automatizált gépi fordítás, fully automatic machine translation, FAMT), de napjainkban még messze vagyunk ettől az állapottól. Napjainkban reálisan olyan programokról beszélhetünk, amelyek vagy emberi segítséggel működnek helyesen (ember támogatta gépi fordítás, human aided machine translation, HAMT), vagy amelyek segíthetik az ember fordítói munkáját (gép támogatta emberi fordítás, machine aided human translation, MAHT) (Prószéky és Kis, 1999).
- Stylometry, stilisztika, szerzőazonosítás: irodalmi művek számítógépes elemzésével foglalkozó területek. Szemben a korábban említett tudományterületekkel nem elsősorban a nagy méretű, általános célú korpuszok kerülnek feldolgozásra, hanem olyan művek, amelyek szerzője meghatározható (még ha nem is ismerjük).

  - A stylometry elsősorban az irodalmi művek statisztikai elemzésével foglalkozik,
  - a stilisztika (stylistic) fő célja a különböző stílusok számítógéppel segített meghatározása, beazonosítása, míg a
  - szerzőazonosítás (authorship attribution) ismeretlen szerzőjű művek szerzőjének, illetve kérdéses esetekben a valódi szerző meghatározása.
- Beszédfeldolgozás: beszélt nyelvi szövegek rögzítése, feldolgozása, mesterséges beszéd generálása.

Ezen tudomány területek számára is szükségessé vált egy fogalomrendszer létrehozása, amely részben a számítástechnika és a nyelvészet egyes fogalmainak átvételét, részben újak megalkotását, esetleg a régiek átfogalmazását, pontosítását jelentette.

A számítógépes nyelvészettel kapcsolatban mindenképpen említést kell tennünk a szövegfeldolgozások során elkerülhetetlen fogalomnak, a szónak a definiálásáról. A későbbiekben látni fogjuk, hogy a szóhatárok meghatározása alapvető fontosságú, különösen írott szövegek esetén, hiszen a különböző szintű szövegfeldolgozások kiinduló pontja az lesz, hogy mit is értünk a szó alatt (Bevezetés – Függelék).

I.1. Nyelvi modellek

A nyelvészeti modell, mint egy természetes nyelvi jelenség absztrakt reprezentációja, definícióját Edmundson (1963) fogalmazta meg. A modellek létrehozásához kvantitatív adatokra van szükség, így szükségképpen korpusz alapúak. Egy nyelvi modell minden esetben egy közelítése a természetes nyelvi szövegnek. A napjainkra számos nyelvi modell látott napvilágot, melyek közül néhány, különböző elven működő modellt szeretnék megemlíteni. Markov

(1916) sztohasztikus modellje, Shannon (1949) kommunikációs elmélete és Zipf (1935) nagyság-szerinti gyakorisági törvénye.

A nyelvi modelleket két nagy csoportba szokás sorolni:

- jövőbeni eseményeket leíró (predictive) modellek, illetve
- azok, amelyek megmagyarázzák a tapasztalt jelenségeket (explicative).

I.1.1. Nyelvi modellek - kezdetek

I.1.1.1. Zipf törvénye

Zipf (1935) törvénye azt fogalmazza meg, hogy egy szó helye egy olyan szógyakorisági listában, amely a gyakoriságok szerinti csökkenő sorrendben van rendezve, inverz kapcsolatban van a szó gyakoriságával. Egy szó gyakoriságát meg tudjuk határozni abból, hogy mely pozíciót foglalja el a nagyság-szerinti gyakorisági listán.

I.1.1.2. Shannon kommunikációs elmélete

Shannon kommunikáció elmélete az információ elméletén alapul és azt próbálja megadni, hogy mennyi információt továbbít az a forrás, amely előállítja az üzenetet, mint például egy emberi lény, a telefon, az újság. Az információ fogalma nem az üzenet jelentésére, illetve szemantikai tartalmára vonatkozik, hanem az üzenetnek statisztikailag ritkán bekövetkező részére. Az elmélet alapján minél ritkábban következik be egy jelenség, annál több információt közvetít.

I.1.1.3. Markov modell

Markov modell is rendelkezik explicative tulajdonságokkal, mivel felhasználja az információ elméletet, de ugyanakkor van predikative tulajdonsága is. Ennek a tulajdonságának köszönhetően napjainkra a Markov

modell leginkább statisztikai alapon működő szófaj meghatározások (Part of Speech, POS) algoritmusaként használatos, szemben a szabály-alapú determinisztikus (például: INTEX[1] (Silberztein, 1993; Roche és Schabes, 1997; Váradi, 2004)) POS algoritmusokkal.

A statisztikai alapú modellek az eseményeket azok statisztikai viselkedése alapján írják le, leginkább a megjelenés valószínűségét, mint az idő függvényét figyelembe véve. Statisztikai alapú, a Markov modellt felhasználó POS tagger (jelölő) például a TnT[2] (Trigrams'n Tags). A program tanítható bármilyen nyelven és gyakorlatilag bármilyen tag-gyűjteményt felhasználva. A paraméterek előállítása egy tanító korpusz segítségével történik, amelyet előzőleg el kell látni a helyes tagekkel (jelekkel) (Brants, 1997; 2000).

I.1.2. Természetes nyelvi szövegek statisztikai közelítései

A természetes nyelvi szövegeket leíró statisztikai modellek aszerint is csoportosíthatók, hogy a nyelv elemeit mennyire tekintjük egymástól függőnek, illetve függetlennek (Oakes, 1998). Ezen szempontot figyelembe véve a következő kategorizálást lehet elvégezni:

I.1.2.1. Nullad-rendű közelítés

Ez a legegyszerűbb modell, amelyben a szimbólumok egymástól függetlenek és egyenlő valószínűségűek (például: az egymást követő kocka dobások)

I.1.2.2. Első-rendű közelítés

[1] http://www.nyu.edu/pages/linguistics/intex;
http://intex.univ-fcomte.fr
[2] http://www.coli.uni-sb.de/~thorsten/tnt

A szimbólumok függetlenek, de az előfordulási valószínűségük a szövegbeli gyakoriságokból számolható.

I.1.2.3. Másod-rendű közelítések

Az egymást követő szimbólumok nem függetlenül kerülnek kiválasztásra, hanem a valószínűségük függ az előtte álló szimbólumtól. Ezek a közelítések adják egy nyelv bigram struktúráját, amelyben a szimbólum párok gyakorisága az alapján van meghatározva, hogy természetes nyelvi szövegben milyennek mérték. A Markov modellben minden egyes következő állapot csak és kizárólag a jelen állapottól függ.

I.1.2.4. Harmad-rendű közelítések

A természetes nyelv trigram struktúrájának az előállítása lehetséges ezzel a közelítéssel (lsd. TnT, I.5.1.3.), amelyben három egymást követő elem gyakoriságát veszik figyelembe.

I.1.2.5. n-edrendű közelítések

Az előbbiek általánosításaként beszélhetünk n-edrendű közelítésekről is, de természetes nyelvű szövegek modellezéséhez harmadrendű modelleknél magasabb számút nem szokás használni. Ekkor ugyanis már számolnunk kell az értékes kimenetek számának csökkenésével és a nagyon nagy, szerteágazó, elemző mátrixokkal, gráfokkal. Egy összetett Markov modellben a függőség messzebbre nyúlik vissza, egy lánc, amely megelőzi a jelen állapotot. A Markov választás, amely a kiválasztott állapot előtti *n* állapotot veszi figyelembe (n+1)-edrendű közelítése a vizsgált nyelvnek, amelyből az átmeneti valószínűségeket számoltuk. Ezt a modellt n-edrendű Markov modellnek hívják.

> „Bármennyire is biztatónak tűnik a Markov modellek használata elég nagy egyetértés van abban a tekintetben, hogy a statisztikai

> valószínűségi modellje a természetes nyelvi kommunikációra nem alkalmazható maradéktalanul. Képtelenség megszámolni az összes elemsorozatát egy nyelvnek. De még ha ez nem lenni is lehetetlen, a legtöbb elem előfordulása más tényezőktől függ, nem az előtte előforduló elemtől. Például a nyelvtani függőségek gyakran olyan kifejezések között állnak fenn, amelyek nem szomszédosak egymással. Nem beszélve arról, hogy a statisztikai módszer figyelmen kívül hagyja a szövegek értelmének és a diskurzusban betöltött szerepének, céljának a legtöbb tényezőjét."
>
> (Beaugrande és Dressler, 2000)

Mindennek tudatában azonban azt mondhatjuk, hogy

> „…a prabobilisztikus (valószínűségen alapuló) modellek megfelelőbbek és valósághűbbek, mint a determinisztikus (meghatározottságon alapuló) modellek. A szerkezetépítő műveletek dinamikus jellemzése produktívabb, mint maguknak a szerkezeteknek a leírása. Arra kell törekednünk, hogy a szabályszerűségeket, stratégiákat, indítékokat, preferenciákat és alapeseteket fedezzük fel, nem pedig arra, hogy szabályokat vagy törvényeket."
>
> (Beaugrande és Dressler, 2000)

I.1.3. Modellek napjainkban

Napjaink szövegnyelvészetében párhuzamosan vannak jelen a prabobilisztikus és a determinisztikus modellek. Ebben az átmeneti időszakban, amely hossza jelen pillanatban nem megjósolható, a két tábor egymástól csaknem teljesen szeparáltan dolgozik. A determinisztikus modelleket építők azzal érvelnek, hogy jelen eszközeinkkel az ő munkájuk legalább olyan

eredményes, mint a másik táboré, amit el is kell fogadjunk eredményeik alapján. Személyes tapasztalataim is ezt támasztják alá. Két, a két különböző elven alapuló, magyar nyelvű, jelenleg fejlesztés alatt álló, egyértelműsítő programot összehasonlítva azt találtam, hogy gyakori mondatok esetén mindkettő helyesen működött, de a *Pista sír*. és *A sír a temetőben van*. mondatokban szereplő *sír* szóalakra az egyik program mindkét esetben főnév tag-et illesztett, míg a másik ige tag-et.

I.1.4. Irodalmi művek statisztikai elemzése

Irodalmi művek statisztikai módszerekkel történő elemzésével, szemben a korábban kizárólagosan alkalmazott szubjektív módszerekkel, a művek objektív értékelése végezhető el. Az olvasói intuícióra alapozott eredmények már korábban sem találtak maradéktalanul elfogadásra az irodalmárok között, így folyamatosan keresték azokat a megoldásokat, amelyek kevésbé szubjektívek. Az egyik lehetséges megoldás, hogy az irodalmi műveket számokkal próbáljuk meg leírni. A számítógép jelentette/jelenti, ahogy sok más probléma esetén is a megoldást napjainkban. A szóalakok, mint egy lehetséges minimális egység számának a pontos ismeretében további olyan formulák határozhatók meg, amelyek képesek a szövegek egy-egy tulajdonságának a leírására. Lehet arról vitatkozni, hogy a nyers adatok/szóalakok mennyire alkalmasak egy irodalmi mű stilisztikai leírására, de úgy tűnik, hogy ezek statisztikai vizsgálatánál mostanáig nem sikerült megbízhatóbb módszert találni az irodalmi művek leírására (Holmes, 1994b).

Az irodalmi művek statisztikai elemzéséhez felhasznált korpuszból legkönnyebben nyerhető információ a szavak előfordulási gyakorisága. Így tehát nem meglepő, hogy számos, a szavak gyakoriságán alapuló formula és modell látott mára napvilágot.

I.1.5. Szavak csoportosítása gyakoriságuk alapján

Szavakat gyakoriságuk szerint kétféleképpen szokás csoportosítani:

- nagyság szerinti gyakoriság eloszlás (rank-frequency distribution),
- csoportosított gyakorisági eloszlás (group-frequency distribution).

I.1.5.1. Nagyság szerinti gyakoriság eloszlás

Szavak nagyságszerinti gyakorisági eloszlását kapjuk, amikor az $f(j,N)$ gyakoriság a j-edik legnagyobb gyakoriságú szót jelöli, ahol N a szöveg hosszát, j az adott szó gyakoriság szerinti sorszámát jelöli és fennáll

$$f(j,N) \geq f(j+1,N) \tag{1.1.}$$

egyenlőtlenség bármely j-re, ha $j = 1, \ldots, V(N)$.

$$V(N) = \sum_{k='a'}^{'z'} m_k \,, \tag{1.2.}$$

ahol m_k a karakterkészlet k-adik karakterével kezdődő szavainak száma (3.9. ábra).

I.1.5.2. Csoportosított gyakorisági eloszlás

A szavak csoportosított gyakorisági eloszlását kapjuk, amikor összeszámláljuk azokat a szavakat, amelyek előfordulási gyakorisága megegyezik. Az r gyakorisági csoportot azok a szavak adják, amelyek előfordulási gyakorisága egy N szövegszó hosszúságú szövegben pontosan r. Ezen szavak számát $V(r,N)$-nek szokás jelölni.

$$V(r,N) = \sum_{i=1}^{V(N)} I_{[f(i,N)=r]} \tag{1.3.}$$

$$I_{[\alpha]} = \begin{cases} 1, & ha\ \alpha \quad igaz \\ 0, & ha\ \alpha \quad hamis \end{cases} \tag{1.4.}$$

I.2. Large Number of Rare Events

Szóalakok gyakorisági eloszlásának egyik legkarakterisztikusabb jellemzője, hogy nagyon magas a ritkán előforduló szavak száma, ezért ezek az eloszlások a nagyszámú, de ugyanakkor rendkívül alacsony gyakoriságú eseményeket leíró (Large Number of Rare Events (LNRE)) osztályba tartoznak (Khmaladze, 1987). A gyakoriságon alapuló szókészletet leíró görbe általános alakja lényegesen eltér a nem-LNRE típusú eloszlások hasonló görbéjének alakjától. A nem-LNRE típusú eloszlásoknál a spektrum elemek viszonylag gyorsan elérik a maximumot, míg az LNRE típusú eloszlások nem. Az LNRE zónát úgy lehet jellemezni, hogy azon mintaméret tartományok, amelyekben az eloszlás függvény alakjából látszik, hogy még csak most kezdtük a teljes populáció elemeit feltérképezni.

Khmaladze két definíciót vezetett be az LNRE koncepció meghatározásához. Legyen a

$$\upsilon(N) = (f(1,N), f(2,N), \ldots, f(S,N)) \tag{1.5.}$$

a szóalakok gyakoriságának a vektora, ahol N a szövegszók száma. Az N növelésével az ilyen vektoroknak egy sorozatát kapjuk

$$\{\upsilon(N)\},\ N = 1,2,3,\ldots \tag{1.6.}$$

Kezdetben (kis N értékek esetén) a gyakoriságok többsége 0 lesz, de ahogy növeljük a mintaszöveg méretét egyre több és több szó jelenik meg nem-nulla gyakorisággal. Az első definíció azt állítja, hogy

Definíció: $\{\upsilon(N)\}$ sorozat akkor nevezhető nagyszámú alacsony gyakoriságú eseménynek, ha

$$\lim_{N\to\infty} \frac{\mathrm{E}[V(1,N)]}{N} > 0\,, \tag{1.7.}$$

ahol $V(1,N)$ jelöli az N szövegszó hosszúságú szövegben az egyszer előforduló szavak (hapax legomena) számát, $\mathrm{E}[V(1,N)]$ pedig ezen szavak számának várható értékét.

Ezen definíciónak megfelelően akkor beszélhetünk egy LNRE eloszlásról, ha a szókészlet növekedése tetszőleges N esetén is nagyobb mint nulla. Könnyű belátni, hogy rögzített véges S és rögzített előfordulási valószínűségek $\pi_i, i = 1,2,3,\ldots,S$ esetén

$$\lim_{N\to\infty} f(i,N) = \infty \tag{1.8.}$$

ebből

$$\lim_{N\to\infty} \mathrm{E}[V(1,N)] = 0 \tag{1.9.}$$

és

$$\lim_{N\to\infty} \mathrm{E}[V(N)] = S\,. \tag{1.10.}$$

Ezen feltételek mellett a szókészlet növekedés nulla lesz, azonban véges szóalak számmal (*S*) rendelkező eloszlásoknál sincs garantálva a nem-nulla növekedés $N \to \infty$ esetén, mivel *N* túl gyorsan növekedhet az egyszer előforduló szavak száma miatt. Khmaladze második definíciója kevésbé szigorú:

Definíció: $\{\upsilon(N)\}$ sorozat akkor nevezhető nagyszámú alacsony gyakoriságú eseménynek, ha

$$\lim_{N\to\infty} \frac{\mathrm{E}[V(1,N)]}{N} > 0 \text{ és } \lim_{N\to\infty} \mathrm{E}[V(N)] = \infty \qquad (1.11.)$$

Ez a definíció megköveteli, hogy S legyen végtelen és, hogy az egyszer előforduló szavak nem-elhanyagolható hányadát képezzék a teljes szókészletnek.

I.2.1. Szövegek tulajdonságait leíró konstansok

I.2.1.1. Zipf törvénye

Definíció: A Zipf-rang, *z*, megadja egy szó pozícióját a szavak csökkenő sorrendű gyakorisági listájában elfoglalt helye alapján.
Ennek megfelelően a leggyakoribb szó kerül a legelső pozícióra a Zipf rangsorban ($z = 1$), a következő leggyakoribb szó a második ebben a listában, stb.

Definíció: $f_z(z,N)$ megadja annak a szónak a gyakoriságát, amely a Zipf rangsorban a *z*. pozíciót foglalja el.
Zipf törvénye (1935) kimondja, hogy

$$f_z(z,N) = \frac{C}{z^a} \tag{1.12.}$$

ahol *a* egy szabad paraméter, ami megadja a regressziós egyenes meredekségét, míg *C* egy normalizáló konstans.
Zipf törvénye a következő formában vált ismertté:

$$\lg f_z(z,N) = \lg C - a \lg z \tag{1.13.}$$

A szövegből nyerhető értékek, mint az átlagos gyakoriság, a szókészlet mérete, a Zipf-rang, stb. mind függ a mintaszöveg méretétől, így különböző hosszúságú szövegek összehasonlítása ezen paraméterek segítségével nem megoldható. Így nem meglepő, hogy számos kutatás arra irányult, hogy találjanak olyan konstansokat, amelyek alkalmasak a szöveg jellemzésére, de függetlenek a mintaszöveg hosszától.

I.2.1.2. Yule féle K

A legkorábbi, mintaszövegtől független paramétert Yule (1944) találta, amely a szavak ismétlési számát adja meg. A megoldás érdekessége, hogy Yule számításait számítógép nélkül, kézi munkával végezte (1944!), hasonlóan a korai szótár készítőkhöz, akik kis papír szeletkéket használtak az egyes információk lejegyzéséhez.

$$K = 100000 \frac{\sum_m m^2 V(m,N) - N}{N^2}. \tag{1.14.}$$

Nem sokkal később, egy a Yule féle *K*-hoz szorosan kapcsolódó *D* is megjelent (Simpson, 1949)

$$D = \sum_{m} V(m,V)\frac{m}{N}\frac{m-1}{N-1} \qquad (1.15.)$$

Természetesen számos újabb és újabb konstans jelent meg, mivel ennek a kettőnek nem sikerült maradéktalanul bizonyítani a stabilitását (Guiraud, 1954; Herdan, 1964; Sichel, 1975; Brunet, 1978), de számos olyan próbálkozás is napvilágot látott, amely ezek kombinációját használja. Burrows (2003) szintén a *D*-val (Delta) jelölte a kutatásai során talált formulát, amellyel stílusbeli különbségek írhatók le és úgy hivatkozunk rá, hogy Burrow's Deltája. Döntését azzal magyarázta, hogy egyrészt a különbségre (distance) akart utalni, másrészt pedig „… egy gesztus, Yule és a többi úttörő felé, akik egyszerű kifejezéseket próbáltak találni stílusbeli eltérések leírására. Yule *K*-ja az egyik legjelentősebb eredmény ezen a téren." (Burrows, 2003).

I.2.2. Urna-modell

Szövegek modellezésére is számos megoldás látott napvilágot, amelyek közül az urna-modellt szeretném megemlíteni, mivel ennek a modellnek egy tovább fejlesztett változata képezi a dolgozatban ismertetett eredmények alapját.

Az urna-modellben a szóalakokat úgy lehet elképzelni, mint különböző színű golyók, amelyekből pontosan annyi darab van, ahányszor az adott szóalak előfordult a kiválasztott szövegben. Tegyük fel, hogy az urna *S* darab különböző szóalakot tartalmaz, ω_i, $i = 1,2,3,\ldots,S$. Minden egyes szóalakhoz hozzárendeljük az előfordulási valószínűségét, $\pi_i, i = 1,2,3,\ldots,S$. Ha ebből az urnából visszatevéses válogatással húzzuk a golyókat, szóalakokat, akkor egy binomiális eloszlással leírható folyamatot kapunk. Ezek alapján meg tudjuk határozni egy adott (ω_i) szóalak várható gyakoriságát:

$$E\left[f\left(i,N\right)\right]=N\pi_i, \tag{1.16.}$$

majd azoknak a szóalakoknak a számát, amelyek gyakorisága *m*

$$E\left[V\left(m,N\right)\right]=\sum_{i=1}^{S}\binom{N}{m}\pi_i^m\left(1-\pi_i\right)^{N-m} \tag{1.17.}$$

és végül a különböző szóalakok számát

$$E\left[V\left(N\right)\right]=S-\sum_{i=1}^{S}\left(1-\pi_i\right)^N \tag{1.18.}$$

(Baayen, 1996a; 1996b; 2001).

Nem szabad azonban megfeledkeznünk arról, hogy ezek a lexikai statisztikai modellek valamennyien azzal az alapfeltételezéssel dolgoznak, hogy a szövegen belül a szavak egymástól függetlenül követik egymást.

I.2.3. A szavak egymástól független előfordulásának feltételezése (the randomness assumption)

Szövegek statisztikai elemzésénél mindenképpen említést kell tennünk a szövegben előforduló szavak egymástól nem-független használatáról. Egy természetes nyelvi szövegben a szavak nem függetlenül követik egymást, tehát a szöveg visszaállítására nem lehetnek alkalmasak a szavakat random válogató modellek, de nem is ez volt a céljuk. Ennek természetes következménye, hogy az említett vizsgálatoknál különbség van az eredeti szöveg és a modell szóalakjainak száma között. Az eltérés mértéke függhet egyrészt a választott modell típusától, ugyanakkor a korábbi szubjektív vélemények alapján főleg

szintaktikai, esetleg szemantikai, illetve szövegszerkezeti megkötések is okozhatják az eltérést.

I.2.4. Szöveg szintek

Számos, különböző szinten megtörténhet a szavak nem-független előfordulása. Ezeket, eredetük szerint, három csoportba szokás sorolni (Balázs, 1985; Tolcsvai Nagy, 1993; Beaugrande és Dressler, 2000):

- szintaktikai,
- szemantikai,
- szöveg szint.

I.2.4.1. Szintaktikai szint

A mondaton belüli kohézió, amely az adott nyelv szintaktikai kötöttségeiből adódik.

> „Maguk a hallott vagy látott szavak között milyen kölcsönös összefüggések vannak egy adott szósorozaton belül. A felszíni összetevők nyelvtani alakzatok és konvenciók alapján függnek egymástól, vagyis a kohézió a grammatikai függőségeken alapul.” (Beaugrande és Dressler, 2000)

I.2.4.2. Szemantikai szint

A mondatok egymás közötti koherenciájából adódó kötöttségek. Ezen a szinten a szöveget témájának azonossága teszi összefüggővé. Mindaddig, amíg a közlés elemei ugyanazon témához vagy egymáshoz szervesen illeszkedő résztémákhoz kapcsolódnak, a szöveg szemantikailag egységet alkot (Kiefer, 1983; Beaugrande és Dressler, 2000; Levinson, 2000).

> „Ez a szövegvilág összetevői, vagyis a szövegfelszín alatt meghúzódó fogalmak és viszonyok kölcsönösen elérhető és releváns voltára utalnak. A fogalom esetünkben úgy határozható meg, mint valamely tudás-alakzat, amely többé-kevésbé egységesen és következetesen elérhető vagy aktivizálható a tudatban. A viszonyok azon fogalmak közötti kapcsolatot jelentik, amelyek együtt jelennek meg a szövegvilágban: minden ilyen kapcsolat magán viseli annak a fogalomnak a jelzését, amelyhez kötődik. … A viszonyok néha nincsenek explicitté téve a szövegben, vagyis nincsenek közvetlenül felszíni kifejezések révén. Az ember annyi viszonyt tesz hozzá az előtte álló szöveghez, amennyi csak szükséges ahhoz, hogy a szöveg értelmes legyen.” (Beaugrande és Dressler, 2000)

I.2.4.3. Szöveg szint

> „A kohézió és a koherencia szöveg-központú fogalmak, amelyek a szövegek anyagára irányuló műveleteket jelölnek.” (Beaugrande és Dressler, 2000)

Ezzel szemben szöveg szinten a szöveg, mint teljes egésznek a szerkezeti kötöttségei jelennek meg, magában foglalva a szöveg szituációs jelentését.

> „A szövegszerűség harmadik ismérvét szándékoltságnak nevezhetjük. Ez a szöveg létrehozójának arra az igyekezetére vonatkozik, hogy a létrehozott közlés kohézióval rendelkező és koherens szöveget alkosson, amely teljesíteni képes a létrehozójának szándékait, vagyis például ismereteket tudjon közvetíteni, vagy pedig egy tervben meghatározott célt tudjon elérni.” (Beaugrande és Dressler, 2000)

I.2.5. Nem a mondaton belüli kohézió felelős az eredeti és a modell közötti eltérésért

Baayen (1993; 2001) foglalkozott a modellek összehasonlításával, azok elfogadhatóságával, valamint vizsgálta, hogy az általa használt modell a szókészlet meghatározására mennyiben tér el a mért értékektől és ez az eltérés mivel magyarázható (Baayen, 1996a; 1996b; 2001). Mivel a szintaktikai megkötések a legnyilvánvalóbbnak ezért annak megmutatására, hogy mondat szinten jelenlévő megkötésekkel magyarázható-e a modell és az eredeti szöveg közötti eltérés kísérletében azt a módszert használta, hogy előállított egy mesterséges szöveget, amelyben meghagyta a mondatokon belüli eredeti szórendet, de random válogatta a mondatokat. Azt tapasztalta, hogy a modell és a mesterséges szöveg közötti különbségek csökkentek, tehát nem a mondatszinten meglévő megkötések felelősek az eltérésért. További kérdés volt tehát, hogy mivel magyarázható az eltérés, amire Baayen vizsgálatai sem adnak maradéktalanul választ.

Célkitűzések

Kutatásaink elsődleges célja az volt, hogy irodalmi művekben megjelenő különböző szóalakok lexikai statisztikai elemzése alapján a mű sajátosságaira tudjunk következtetni, a mű szerkezetéről, felépítéséről információt tudjunk szerezni, és azt további feldolgozásra elő tudjuk készíteni. Főként angol és magyar nyelvű irodalmi művek egy speciális tulajdonságának meghatározását tűztük ki célul: arra keressük a választ, hogy az írók mikor, a szöveg mely pontján találják indokoltnak olyan szavak bevezetését, amelyek korábban nem szerepeltek az adott műben. Jellemző-e az íróra, a zsánerre, a mű hosszára, hogy mikor, a szöveg mely pontján jelenik meg egy olyan szóalak, amely addig még nem fordult elő, illetve mennyi ezeknek az újonnan megjelenő szóalakoknak a száma abszolút értékekben, valamint a kiválasztott szövegpont egy adott nagyságú környezetében.

Korábbi kutatások eredményei alapján ismert, hogy a szóalakok vizsgálata önmagában nem alkalmas szerzőazonosításra, kérdés volt tehát, hogy a szóalakok bevezetésére irányuló elemzések segítségével milyen újabb információkhoz juthatunk.

Anyanyelvi irodalmi művek olvasása során is, de főleg idegen nyelvű szövegek esetén megtapasztalhatjuk, hogy a regényt olvasva folyamatosan csökken az újonnan bevezetésre kerülő különböző szóalakok száma, így előre haladva a könyvben egyre könnyebb annak olvasása. Az ilyen és hasonló jellegű olvasói intuíciók azonban nem minden esetben nyertek bizonyítást, mivel egy könyv olvashatósága nemcsak a felhasznált szóalakok függvénye, hanem számos más tényező is befolyásolhatja.

Az újonnan bevezetésre kerülő szavak, nagy általánosságban, valóban monoton csökkenő tendenciát mutatnak. A művek többségénél azonban találni olyan intervallumokat, amelyekben hirtelen megemelkedik a különböző szóalakok száma. Vizsgálatainkban arra kerestük a választ, hogy mivel magyarázható a monoton csökkenő tendenciától való eltérés, tehát mikor és miért következik be, hogy az újonnan bevezetésre kerülő szavak száma lényegesen magasabb, mint az azt megelőző periódusokban.

Kísérleteink elvégzéséhez szükség volt egy olyan dinamikus vizsgálati módszer kidolgozására, amely mind az angol, mind a magyar szövegekben képes az újonnan megjelenő szóalakok számának viselkedését a lehető legjobb közelítéssel visszaadni. A szókészlet nagyságára és gazdagságára vonatkozó statikus modelleknél is alkalmazott elméleti meggondolások közül kettő tűnt alkalmazhatónak. Ezek egyike a szavak egymástól függetlenül történő megjelenésének a feltételezése (randomness assumption), továbbá, hogy a szavak egy adott szövegen belül polinomiális, annak speciális eseteként binomiális eloszlást követnek. Ezeket felhasználva, illetve továbbiakkal kiegészítve olyan dinamikus modell megépítését tűztük ki célul, amely az eredeti szövegben meglévő trendek és szezonalitások leírására is alkalmas lehet.

Munkánk során angol, magyar és német nyelvű irodalmi művek, azon belül is regények és novellák feldolgozását tűztük ki célul.

Angol nyelvű szövegekre azért esett a választás, hogy eredményeinket össze tudjuk hasonlítani korábbi, a szókészlet méretére vonatkozó, statikus modellek alapján kapott eredményekkel. A magyar agglutináló és a német flektáló tulajdonsága miatt az előforduló szóalakok száma magasabb, mint az angol nyelvű szövegekben. A nyelveknek ezt a tulajdonságát figyelembe véve, a magyar és a német nyelvű irodalmi művek elemzése során arra voltunk kíváncsiak, hogy a szóalakok bevezetése mutat-e hasonlóságot az angol szövegeknél megfigyeltekhez, illetve mennyiben tér el azoktól, a magasabb

szóalakszám befolyásolja-e az angol nyelvű szövegeknél megfigyelt másodlagos események leírását jellemző szóalakszám emelkedést. Magyar szövegek ilyen jellegű számítógépes feldolgozására, tudomásunk szerint, ez idáig nem történtek kísérletek. Érdemesnek tűnt tehát megvizsgálni, hogy egy agglutináló nyelv (Prószéky, 1989; O'Grady et al., 1993; Kiefer, 1998; Laczkó, 2000) esetén hogyan alkalmazhatóak a szavak függetlenségét feltételező modellek.

A korábban megjelent szubjektív vélemények arra engedtek következtetni, hogy megoszlik a témával foglalkozók véleménye abban, hogy mikor jelennek meg új szavak egy irodalmi műben. Egyes vélemények szerint a fejezet határok azok a helyek, ahol látványosan emelkedik az újonnan bevezetett szavak száma, míg mások szerint a szövegekben megjelenő hosszabb leírások okoznak ilyen jellegű változásokat. Baayan eredményeinek és sejtéseinek ismeretében ez utóbbi vélemények tűntek elfogadhatónak, így az általunk kidolgozott módszert annak a hipotézisnek az igazolására kívántuk felhasználni, hogy az újonnan bevezetett szóalakok száma akkor emelkedik meg, ha a szöveg menetében, a szöveg teljes hosszához viszonyítva, egy viszonylag rövid változás következik be. Ezt az állításunkat úgy is megfogalmazhatjuk, hogy a szavak egymástól független megjelenését feltételező modell és az eredeti szöveg közötti eltérések szöveg szinten bekövetkező változások eredményei.

Vizsgálatainkban a dinamikus statisztikai modell megépítésén túl egy eddig nem, vagy igen ritkán alkalmazott módszert, az eredeti mű és a fordításainak az összehasonlítását alkalmaztuk. Hasonló, szavak gyakoriságán alapuló módszereknél, korábban azért nem tűnt alkalmazhatónak a különböző nyelveken írt szövegek összehasonlítása, mert a nyelvek szintaktikai, szemantikai szabályai, kötöttségei, a fordításból származó eltérések más és más szószámot eredményeztek a szöveg különböző verzióiban. Mivel kutatásainknak nem az volt az elsődleges célja, hogy a szókészlet nagyságára, gazdagságára, a felhasznált szavak pontos meghatározására találjunk formulát, magyarázatot,

hanem azt próbáltuk meghatározni, hogy mikor jelenik meg a szövegben egy új szóalak, ezért az eredeti mű és fordításainak összehasonlítása egy szokatlan, de jól alkalmazható eljárásnak bizonyult.

Annak további igazolásához, hogy az újonnan bevezetett szóalakok számának változása szöveg szinten következik be, az egyszer előforduló szavak (hapax legomena) megjelenésének vizsgálatát is elvégeztük.

III. Módszerek

III.1. Szövegek elektronikus formában

III.1.1. Források

Angol és magyar nyelvű irodalmi művek, azon belül is regények és novellák elemzését végeztük el. A szövegek feldolgozásához szükség volt azok digitális verziójára, amelyek elsődleges forrása az Internet volt.

- Az angol nyelvű könyvek
  - a Project Gutenberg[3] és
  - a University of Virginia's E-book Library[4],
- míg a magyar nyelvűek
  - a Magyar Elektronikus Könyvtár[5] (MEK), illetve
  - a Digitális Irodalmi Akadémia[6]

elektronikus könyvtárakból kerültek letöltésre.

A feldolgozásra került művek kiválasztását azonban nagyban befolyásolta, hogy melyek voltak ingyenesen elérhetőek. A szerzői jogok védelme nem teszi lehetővé, hogy ingyenesen hozzáférjünk XX. századi angol nyelvű szépirodalmi művekhez. A magyar elektronikus könyvtáraknak köszönhetően azonban napjaink irodalmi művei is elérhetőek, így ezek közül is feldolgozásra került néhány.

---

[3] http://promo.net/pg/
[4] http://etext.lib.virginia.edu/ebooks/ebooklist.html
[5] http://www.mek.iif.hu, http://www.mek.oszk.hu
[6] http://www.irodalmiakademia.hu/scripts/DIATxcgi?infile=diat_vm_main_menu.html

III.1.2. Digitalizálás

Az elektronikus formában nem elérhető irodalmi művek egy részét, valamint a nyelvkönyvek elektronikus formára alakítását kézi szkeneléssel sikerült pótolni. A szövegbevitelt és felismerést a Recognita[7] cég *RECOGNITA* szoftverével végeztük (Recognita Plus 5.0), ami mind magyar, mind angol nyelvű szövegek kezelésére alkalmas. Az irodalmi művek feldolgozása minimális számú hibával történt, így nem volt szükség a könyvek újra olvasására. A helyesírás ellenőrző által jelzett hibák kijavítása elegendőnek bizonyult. Ezzel szemben a nyelvkönyvek szöveges anyagának bevitelekor a könyvek teljes újraolvasására, ellenőrzésére szükség volt. Mindez azzal magyarázható, hogy a napjaikban piacra kerülő nyelvkönyvek tudásszinttől és életkortól függetlenül rendkívül látványosak, színekben, képanyagban igen gazdagok. A látványgazdagság abban is megnyilvánul, hogy a kiadók szakítottak a megszokott fehér alapon fekete karakterekkel, valamint a tipográfia hagyományos elemeivel (Virágvölgyi, 1996). Ily módon ezek a könyvek tartalmaznak sötét, illetve mintás alapon sötét karaktereket, valamint világos alapon világos karaktereket, amit a szkenner nem tudott olvasni, tehát gépeléssel kellett pótolni. Az esetek többségében a kiadók igyekeztek a szöveg témájának megfelelő betűtípust választani, ennek megfelelően a betűtípusok rendkívül széles skálájával találkozhatunk a könyvekben. Nehézséget okoztak még a szövegfelismerésnél azok a szövegdarabkák, amelyek nem vízszintesen kerültek a könyvbe, hanem attól különböző szögben elforgatva, esetleg 180º-kal is, ami ismételten azt jelentette, hogy a hagyományos gépelés volt a szövegbevitel egyetlen járható útja.

[7] http://www.recognita.hu

III.1.3. Felhasznált karakterkészlet

A szövegek feldolgozása, kiértékelése, modellezése a saját fejlesztésű, Windows operációs rendszerek alatt futtatható *DYMOCASAT*-tel (Dynamic Model for Computer Aided Statistical Analysis of Texts) történt. Mivel a végső cél a szövegekben előforduló különböző szóalakok vizsgálata volt, ezért a feldolgozás alapját a szó definiálása, a szöveg szavakra bontása képezte. A feldolgozás első lépéseként definiálni kellett azt a karakterkészletet, amellyel a program dolgozni fog, amely alapján el fogja dönteni, hogy a szöveg mely karaktersorozata tekinthető szónak.

A program alapértelmezésben angol és magyar nyelvű szövegek feldolgozására alkalmas, de a felhasználó definiálhatja saját karakterkészletét is, ily módon más nyelven írt szövegek is feldolgozhatók a program segítségével (3.1. ábra).

Az érvényes karakterek készletét mindkét nyelv esetén az adott nyelv ábécéje alapján hoztuk létre és alakítottuk úgy, hogy az alkalmas legyen a szövegek számítógépes feldolgozására.

Minden más karakter, mint a szóköz, az arab számok, a mondatvégi írásjelek, a vessző, az idézőjel, kötő- és gondolatjel, kettőspont, pontosvessző, a zárójelek és minden egyéb speciális karakter elválasztójelként használatos.

A program két elválasztó jel közötti karakter sorozatot tekint szónak, nem téve különbséget a kis- és nagybetűk között. Ennek következménye, hogy a tulajdonnevekben megjelenő szavak köznévként kerültek feldolgozásra.

III.1.3.1. Angol karakterkészlet

Az alapértelmezett angol karakterkészlet tartalmazza mind a huszonhat kis- és nagybetűt, valamint az aposztrófot (3.1/A ábra).

Különösnek tűnhet az aposztróf használata, mint érvényes karakter, ugyanolyan tulajdonságokkal, mint az angol ábécé betűi. Ha azonban nem soroltuk volna az aposztrófot az érvényes karakterek közé, akkor elválasztó karakterként működne. Ennek viszont egyenes következménye lenne, hogy az angolban megengedett aposztróf segítségével képzett rövidítések, valamint a birtokviszonyt kifejező aposztrófos kifejezések két szóba törnének.

Elfogadhatónak tartom, hogy kifejezések, mint *it's*, *don't*, *children's*, továbbá azok a szavak, amelyek az eltérő nyelvhasználatból adódnak (3.1. táblázat) nem két-két szóba törve jelennek meg, hanem egy-egy szóként. Ugyanis, ha megtörnénk őket, akkor indokolatlanul magas számban tartalmazna a szöveg pl. *s* és *t* karaktereket, valamint megjelennének karaktersorozatok, amelyek nem is érvényes angol szavak, mint pl. *don*.

A szókezdő és -záró aposztrófok szerepe sem elhanyagolható, ha figyelembe vesszük, hogy mindkettő szolgálhat jelentés hordozóként. Ismételten, ha az aposztróf elválasztó karakterként szolgálna, akkor a következő kifejezéseket megcsonkítanánk, ha elhagynánk belőlük az aposztrófot: *'cause, 'ting*, *thinkin'*, *likin'*. Kapnánk olyan szavakat, amelyek nem léteznek az angol nyelvben, pl. *ting, thinkin*, *likin*, de el is elveszthetnénk néhányat azzal, hogy az aposztróf elhagyása olyan szót eredményez, amely más jelentéssel, de része az angol nyelvnek, ilyen pl. a *cause*.

III.1.3.2. Magyar karakterkészlet

A magyar karakterkészlet a magyar ábécé (MTA, 1996) harmincöt egyjegyű kis- és nagybetűjét tartalmazza, de szemben az angol karakterkészlettel az aposztrófot nem. Ez azzal magyarázható, hogy a magyar szövegekben az aposztrófnak csak nagyon ritkán van jelentés hordozó szerepe és használata is kevésbé gyakori, mint angol szövegekben.

III.1.3.3. Felhasználó által definiált karakterkészlet

A felhasználó további nyelvek karakterkészletét is definiálhatja azzal, hogy megadja az adott nyelvben használni kívánt karaktertereket vagy azok ASCII kódját (3.1/B ábra). Az így létrehozott karakterkészletek fájlba menthetők (3.1/D) és további használatra bármikor betölthetők (3.1/B-C). A használni kívánt nyelv a program menüjén keresztül érhető el. Az angol és a magyar karakterkészleteken túl az eredeti szöveg által meghatározott, egy további, a felhasználó által definiált karakterkészlet betöltése lehetséges (3.1/C ábra).

III.2. Szövegek feldolgozása

III.2.1. Szövegek előkészítése

III.2.1.1. Szövegfájlok

A szövegek tényleges feldolgozását azonban megelőzte egy előkészítő folyamat, amelynek fő célja a szövegek egységesítése volt. Az egységesítésére azért volt szükség, hogy összehasonlítható eredményeket kapjunk különböző forrásból származó szövegek esetén is, valamint azért, hogy a program egységesen tudja kezelni őket. A végső formátum, amelyre a szövegeket alakítottuk szöveges típusú volt. Ez az a fájlformátum, amit a program bemenő adatként el tud fogadni.

A számítógépes feldolgozást meg kellett előzze a művek teljes szövegének előállítása. Az esetek többségében ez az egyes fejezetek konkatenációját jelentette, míg novellák esetén az egyes művek szétválasztását és ezek külön fájlba mentését.

Ezt követően törölni kellett azokat a bekezdéseket, amelyek nem a történet részét képezik, ilyenek a láb- és végjegyzetek, valamint azokat, amelyek

a nyomtatott szövegnek nem, de az elektronikus szövegnek részei, ilyenek például a licence agreement-ek, valamint a hyperlinkek, amelyek képhez, térképhez stb. vihetik azokat, akik on-line olvassák ezeket a műveket.

III.2.1.2. Tipográfiai szabályok

A feldolgozásra kerülő szövegek végső formátuma text fájl, de az előkészítés során szükség volt egy olyan szövegszerkesztő programra is, amely makrók kezelésére alkalmas. A makrók elsősorban az angol nyelvű szövegek aposztróf karakterét kezelték. A különböző forrásból származó könyvek, a tipográfiai szabályok (Virágvölgyi, 1996) időben történő változása, valamint az országonként eltérő tipográfiai szabályok nem tették lehetővé az angol nyelvű könyvek egységes kezelését. Minden egyes szöveget egyénileg kellett feldolgozni, ami több lépésből állt.

Az aposztrófok egységes kezelését azonban meg kellett előzze a sorvégi elválasztójelek és Enter karakterek törlése. A két karakter kezelése mind elméletileg, mind gyakorlatilag eltérő feladatot jelentett az aposztrófok kezeléséhez képest. Míg az aposztrófok kezelése, használata konvenció kérdése, addig a szavakon belüli elválasztójelek és Enter karakterek törlése nem, ezeket el kellett távolítani a szövegből. Nem maradhat a szövegben az *asz-tal* karakter sorozat, mivel ezt a program az „-” (elválasztójel) elválasztó karakter miatt két szónak fogja számolni: *asz* és *tal*, ami nyilvánvalóan nem megengedhető. Abban az esetben lehet megengedett az elválasztó karakter használata, ha az eredeti nyomtatott szövegben is szándékosan – nem sorvégi – elválasztó jeleket helyezett el az író (*E-gész nap ró-bó-tó-lunk / és majd' hogy é-hen nem veszünk … A pró-le-tá-rok i-fjú-gár-dája mi vagyunk…*, Kertész Imre: SORSTALANSÁG, 1975).

III.2.1.2.1. Aposztrófok kezelése

Az aposztrófok kezelésére számos megoldás létezik. A szöveg különböző szintű feldolgozása különböző megoldást igényel. Jelen munka végső célja a különböző szóalakok számlálása és ezek viselkedésének leírása, így az aposztrófokat, az aposztrófokat tartalmazó szavakat is eredeti formájukban kellett megőrizni. Ahhoz, hogy ezt meg tudjuk tenni szükség volt az aposztrófok és az idézőjelek megkülönböztetésére. Egyszerűsítette a problémát, hogy a későbbiekben az idézőjelekre nem lesz szükség, tehát azután, hogy sikerült megtalálni őket ki lehetett törölni, mivel úgy is csak elválasztó karakterként funkcionálnak.

A „valódi” aposztrófok kiválasztásához elsőként szükség volt az eredeti szövegből azon szavak listájára, amelyek aposztróffal kezdődnek, illetve végződnek. Ezt a listát a *DYMOCASAT* készítette el. A lista ismeretében már eldönthető, hogy egy szó csak azért kezdődik, végződik aposztróffal, mert az aposztróf az idézőjelet helyettesíti, vagy valóban indokolt a használata és feldolgozás folyamán meg kell őrizni (3.1. táblázat).

Ki kellett törölni a hamis szókezdő és szóvégi aposztrófokat, úgy, hogy közben megtartjuk azokban a szavakban, amelyekbe a szerző szándékosan helyezte el ezeket. A feldolgozásnál meg kellet oldani, hogy szókezdő aposztrófok olykor bekezdés kezdők is, tehát a szóköz + ' kapcsolatot ilyen esetben ¶ + ' kapcsolat helyettesíti. Hasonló módon, amikor az ' + szóköz kapcsolatot a bekezdések végén ' + ¶ kapcsolat helyettesíti.

III.2.2. A szövegben előforduló szóalakok számlálása

A szövegek előkészítése és text fájllá történő alakítása után következett a tényleges feldolgozás. Hasonlóan az előkészítéshez, a szövegek feldolgozása is

egy többlépcsős folyamat. A tényleges feldolgozást továbbra is a *DYMOCASAT* végzi. A program fő profilja a szövegben előforduló különböző szóalakok számlása és ezek ismeretében a szöveg további feldolgozása.

III.2.3. Előfeldolgozás (Preprocessing)

Szövegek feldolgozásánál mindig kérdés, hogy mennyire legyen az adott szöveg előkészítve, melyek azok az inflexiók, amelyeket el kell távolítani és melyek azok, amelyek a további feldolgozás során is információtartalommal bírnak. Érvek és ellenérvek sorakoztathatók fel az egyik vagy a másik módszer használata mellett, de egyetlen üdvözítő út nem létezik, mindig az elérendő célok adják meg, hogy a szöveget milyen formában érdemes feldolgozni.

Jelen vizsgálataink a szövegben előforduló szóalakokra irányultak, arra kerestük a választ, hogy az eredeti szöveg bármiféle manipulálása, ún. előkészítése nélkül milyen, a szöveg szerkezetére vonatkozó következtetésekre juthatunk.

III.2.3.1. Az eredeti szóalakok feldolgozásának előnyei

Már az eredeti szövegekben megjelenő különböző szóalakok – a szavaknak az a formája, amelyet a szövegből nyerünk és nem végzünk rajta semmilyen csonkítást, átalakítást – is tartalmaznak számos olyan információt, amelyek feldolgozása során betekintést kaphatunk a mű belső szerkezetébe, az egyes szereplők jellemvilágába, az események helyszínéről (Rybicki, 2003), stb. Mindezek olyan információk, amelyek a lemmatizálás vagy bármilyen átmenetinek tekinthető átalakítás során elveszhetnek.

Különösen fontos lehet az eredeti szóalakok megőrzése az egyes szereplők azonosítása során, mert nagyon sok esetben, amikor az író beszélteti őket azért teszi, hogy megkülönböztesse a többi szereplőtől.

- Sokat mondó lehet angol nyelvű szövegekben, hogy hogyan használja az író az összevonásokat (*it's, don't vagy it is, do not*),
- az aposztrófoknak milyen szerepet szán, mennyire jelentéshordozó az aposztróf (*'tink, likin', per'aps vagy think, liking, perhaps* (3.1. táblázat)),
- szándékosan mikor és hol használja hibásan a szavakat (3.2. táblázat),
- a különböző nyelvhasználatok (tájnyelv, sleng, nem-anyanyelvi beszélő, stb.) mekkora szerepet kapnak a műben (pl. *aint gonna* használata szemben az *I'm not going to* kifejezéssel).

Hasonló módon, a szó- és mondathossz, a használt nyelvtani szerkezetek is csonkulnak, ha lemmatizálást végzünk a szövegen, amely információk szintén jelentéshordozóként viselkednek.

III.2.4. Szóalakok számának alakulása angol és magyar nyelvű szövegekben

Kutatásaink során angol és magyar nyelvű irodalmi művekben vizsgáltuk a különböző szóalakok megjelenését. Mivel a magyar agglutináló nyelv ezért kettő, de inkább több morfémát (a szótő és a hozzácsatolt egy vagy több képző és/vagy rag) tartalmazó szavak a gyakoriak (O'Grady et al, 1993; Farber, 1991; Kiefer, 1998; Laczkó, 2000). Magyar nyelvű szövegekben előforduló morfémákban gazdag szavak számának alakulását hasonlítjuk össze az angol megfelelőjükkel a 3.3. táblázatban. Az angol nyelvben a morfémák jelentős hányada önálló egységként, szóként jelenik meg. Ennek következménye, hogy angol és magyar nyelven írott szövegek szövegszóinak és szóalakjainak számát összehasonlítva azt tapasztaljuk, hogy a magyar nyelvű szövegek össz-szószáma, tehát a szövegszók száma kisebb, mint angol nyelvű megfelelőjének,

de a felhasznált különböző szóalakok száma, éppen fordítva, magyar nyelvű szövegekben nagyobb (3.3. táblázat).

Egy másik jellegzetessége a magyar nyelvnek, hogy a szórend nem kötött, tehát a szavak szinte tetszőleges sorrendje értelmes magyar mondatot alkothat (Kugler, 2000; É. Kiss, 1998; 3.4. táblázat).

A magyar nyelv ezen sajátosságának részletes vizsgálata azonban túl mutat jelen tanulmány céljain.

III.2.4.1. A szöveg szavainak számlálása, az adatok tárolása

A szövegek feldolgozását meg kellett előzze a különböző szóalakok számának és megjelenési helyének pontos meghatározása. Mindezt a *DYMOCASAT* végezte.

Az N szövegszó hosszúságú szöveget feldaraboltuk egyenlő hosszúságú, azonos számú szövegszót (h) tartalmazó intervallumokra, blokkokra (b_i).

$$b_i,\ i = 1, \ldots, n, \text{ ahol } n = \left\lfloor \frac{N}{h} \right\rfloor . \qquad (3.1.)$$

III.2.4.2. A szövegek csonkítása

A szövegek ily módon történő feldolgozásánál mindig számolni kell valamennyi veszteséggel, mivel a szöveg végének csonkításakor (az N/h hányados egészrészének a képzése miatt) a szöveg n. blokkot követő részének szavai (v) nem kerülnek feldolgozásra.

$$v = N - h \cdot \left\lfloor \frac{N}{h} \right\rfloor . \qquad (3.2.)$$

Az így bevezetett v egy egyenletes eloszlású valószínűségű változó (Hajtman, 1971; Solt, 1971; Meszéna és Ziermann 1981). Ennek megfelelően a feldolgozásra nem kerülő szavak száma, a szóveszteség várható értéke a lehetséges értékek számtani közepe:

$$\bar{v} = \sum_{i=1}^{r} v_i \ , \tag{3.3.}$$

ahol a r a feldolgozott könyvek száma.

Regények esetén, ahol N, a szövegszók száma általában meghaladja 40000 és nem több, mint 400000 (a feldolgozott művek közül egyedül Tolsztoj HÁBORÚ ÉS BÉKE című műve tartalmazott több, mint 400000 szövegszót) az átlagos relatív veszteség (v_r)

$$\frac{\bar{v}}{400000} < v_r < \frac{\bar{v}}{40000} \ , \tag{3.4.}$$

azaz közelítőleg 10^{-4} és 10^{-3} közé esik.

III.2.4.3. Szavak száma az egyes blokkokban

A blokkok hosszúsága az esetek többségében száz szövegszó hosszúságúra volt állítva, tehát $h = 100$. A végső cél az volt, hogy minden egyes száz szövegszó hosszúságú blokkhoz egy egész számot rendeljünk, az adott blokkban újonnan bevezetésre került különböző szóalakok számát y_i (y_i, $i = 1, \ldots, n$). Az y_i definíciójából következik, hogy bármely i-re

$$0 \leq y_i \leq h \tag{3.5.}$$

Tárolásra azonban nemcsak ezek az értékek kerültek, hanem minden egyes szó szövegen belüli pozíciója, a blokk sorszámával és a szó ezen blokkon belüli előfordulási gyakorisága is. Valamennyi érték tárolása szöveg fájlokban (.txt) történt. A program legfeljebb annyi szöveg fájlt hozott létre az aktuális könyvtárban, ahány karakterből áll a karakter készlet (k = 'a', …, 'z'). (Az aktuális könyvtár beállítása is a programon belül történik, alapértelmezés szerint a WINDOWS\TEMP könyvtár.) A fájlok a szavak kezdőbetűinek az ASCII kódja alapján vannak azonosítva.

Minden egyes szöveg fájl annyi bekezdést (s_k) tartalmaz ahány azzal a karakterrel kezdődő szót (m_k) talált a program a szövegben.

$$s_k = 1, \ldots, m_k, \text{ ahol } m_k = \max('k\ldots'),\ k = \text{'a'},\ldots, \text{'z'}. \tag{3.6.}$$

Az egyes bekezdések pedig legfeljebb n számú karakterből állhatnak. A bekezdések az egyes pozíciókon vagy a szóalak előfordulásának számát vagy annak hiányát jelölik az adott sorszámú blokkban.

A különböző szóalakok tárolására egy hármas indexű elem ($X = \{x_{ksi}\}$ k, s, i) alkalmas (3.2. és 3.3. ábra), ahol az egyes elemek a különböző szóalakokat jelölik, azok pontos megjelenési helyével. k jelöli az ábécé betűit, s a szó ábécébeli sorrendjének a számát az adott betűn belül, míg i a blokkok sorszáma.

$$N = \sum_{k='a'}^{'z'} \sum_{s=1}^{m} \sum_{i=1}^{n} x_{ksi} \ . \tag{3.7.}$$

Az újonnan megjelenő különböző szóalakok meghatározásához, azonban nincs szükségünk sem a szavak előfordulási gyakoriságára, sem az összes előfordulásra. Egy adott szóalak esetén csak az első előfordulását kell

megjegyezni, valamint össze kell számlálni a különböző szóalakok első előfordulását egy adott blokkon belül.

Az egyes blokkokban újonnan bevezetésre kerülő különböző szóalakok számát, az átalakított *XT* tömbben (3.4. ábra) a blokkonkénti (a táblázat oszlopai) T-k száma adja, ha összegezzük ezeket valamennyi karakterre.

$$y_i = \sum_{k=a}^{z} \sum_{s=1}^{m} xt_{ksi} \qquad (3.8.)$$

A számok ábrázolása azonban nem tízes számrendszerben történt, mert előfordulhat, hogy egy szó egy blokkon belül tíznél több alkalommal fordul elő. A számokat (x_{ksi}) ASCII kódok helyettesítik $x + 63$ formátumban. Ennek megfelelően: 1→A; 2→B; stb. (3.5. és 3.6. ábra).

III.2.5. A szóalakok megjelenésének ábrázolása *DYMOCASAT*-tel

Kutatásaink elsődleges célja az volt, hogy szépirodalmi művekben vizsgáljuk a különböző szóalakok megjelenésének szabályszerűségeit, ezért a program egyik feladata, hogy olyan ábrát készítsen, amellyel szemléltethető, hogy az egyes blokkokban hány új szó jelenik meg az előző blokkokhoz képest. A viszonyítási pont mindig az addig vizsgált blokkok összessége. Két ábrázolási módot is használtunk:

- az új szóalakok száma az adott blokkban (y_i) (3.7/A, 3.8/C és D ábra),
- az addigi összes szóalak száma, a teljes szókészlet nagysága (V_i) (3.7/B, 3.8/A és B ábra).

III.2.6. A szavak további feldolgozása

A fent ismertetett módszer, a szavak szövegfájlokban történő tárolása,

további feldolgozásra is alkalmassá teszi a kapott értékeket.

- Az analízist végző program elsőként megszámolja a szóalakok számát, így megkapjuk, hogy az adott szövegben hány különböző szóalak található.
- Lekérdezhető és külön fájlban tárolható, ezen túl, a szavak gyakorisága és relatív gyakorisága is számuk szerint csökkenő, illetve ábécés rendben (3.9. ábra),
- az egyes blokkok szövege,
- az egyes blokkokban újonnan megjelenő szóalakok, illetve
- az egyszer előforduló szavak (hapax legomena) listája blokkonként, stb.

III.2.7. Blokkok hosszának megválasztása

A blokkok hosszának (h) megválasztásakor az elsődleges szempont az volt, hogy a szóalakok megjelenésének szabályszerűségeiről információt szerezhessünk és a különböző hosszúságú szövegeket össze tudjuk hasonlítani. Az összehasonlításhoz mindenképpen egyenlő hosszúságú blokkokra volt szükség. A hosszt úgy kellett meghatározni, hogy

- mind rövid, mind hosszú művek esetén feldolgozható számú blokkot kapjunk, tehát a blokkok száma ne legyen túl nagy,
- a művek végén a csonkítás ne okozzon nagy veszteséget, tehát az egyes blokkok ne legyenek túlságosan hosszúak,
- kellően rövidek legyenek a blokkok ahhoz, hogy a műben bekövetkező változásokra érzékeny maradjon a vizsgálat és végül, hogy
- a blokk elég hosszú legyen ahhoz, hogy a mondaton belüli kötöttségek ne érvényesüljenek a blokkokon belül, tehát az egyes szavak egymástól függetlennek legyenek tekinthetők.

III.3. Szövegek modellezése

A különböző szóalakok számlálása az eredeti művekben, a kapott értékek blokkonkénti tárolása további feldolgozásra ad módot. A használt grafikus ábrázolások már alkalmasak arra, hogy áttekinthető képet kapjunk a műben bevezetésre került szóalakok megjelenésének általános jellemzőiről, de mindez még nem elegendő ahhoz, hogy pontos leírást tudjunk adni ezen szóalakok viselkedéséről, szabályszerűségéről.

A probléma megoldásához egy, a szóalakok előfordulási gyakoriságán alapuló dinamikus modellt próbáltunk építeni, a modell alapján egy mesterséges szöveget előállítani, amely a lehető legjobb közelítéssel vissza tudja adni az eredeti szöveget.

III.3.1. Statikus modellek

A korábban ismertetett lexikai statisztikai modellek valamennyien statikus modellek (összefoglaló értékelés Baayen, 2001-ben) voltak, amelyek azon túl, hogy feltételezték a szavak egymástól független megjelenését egy-egy matematikai formulával próbálták a problémát leírni. Egy egzakt formula megtalálása azonban maga után vonta, hogy ezek a modellek nem képesek visszaadni sem az eredeti szövegben jelenlévő trendeket, sem a szezonalitásokat. Alkalmasnak bizonyultak viszont arra, hogy vizsgálják a szavak nem-független megjelenésének forrásait. Segítségükkel arra a következtetésre jutottak (Baayen, 1996a; 1996b; 2001), hogy ugyan a mondaton belüli kötöttségek a legnyilvánvalóbbak, mégsem ezek a legfőbb forrásai a teljes szöveg szavai nem-véletlenszerű megjelenésének. Sokkal inkább meghatározóak a bekezdés vagy szöveg szinten bekövetkező változásokkal.

III.3.2. Dinamikus modellek

A szavak előfordulási gyakoriságán alapuló dinamikus modellek, hasonlóan a statikus modellekhez, élnek azzal a nyilvánvaló egyszerűsítéssel, hogy a szavak egymástól függetlenül jelennek meg egy szövegben. Szemben azonban a statikus modellekkel képesek visszaadni a szövegben meglévő trendeket.

Tehát egyik típusú modellnél sem az a cél, hogy bebizonyítsuk, hogy a szavak egymástól függetlenül jelennek meg a szövegben, hanem sokkal inkább annak a vizsgálata, hogy mennyiben tér el egy szöveg a modelltől és mivel magyarázhatóak ezek az eltérések.

III.3.3. Modell építése

Egy, a szavak előfordulási gyakoriságán alapuló mesterséges szöveg létrehozásánál elsőként a szókészlet nagyságát célszerű meghatározni. Ez egy természetes elvárás, mivel az írók is rendelkeznek egy elvileg lehatárolható, amikor létrehozzák műveiket. Ennek megfelelően az tűnik ésszerűnek, hogy vesszük az író szókészletét és ezt a szókészletet tekintve kiindulási halmaznak, válogatunk belőle, ahogy azt az író is tette. Az író teljes szókészletének meghatározása azonban szinte lehetetlen feladat. Még nagyon termékeny írók valamennyi művét feldolgozva sem állíthatjuk bizton, hogy hozzájutottunk a teljes szókészlethez. Ez két okkal magyarázható. Az egyik, hogy a szókészletünk folyamatosan változik, így nem rendelkezünk azzal az információval, hogy a kiválasztott mű írásakor mi volt az író aktuális szókészlet (Nation és Waring, 1997; Singleton, 1999).

A másik magyarázat, hogy az aktív és a passzív szókészlet különböző méretű, míg az ismert művek feldolgozása is csak az aktív szókészletről ad

információkat. Valamennyiünk számára nyilvánvaló azonban, hogy a válogatás nemcsak kizárólag az aktív szókészletből történhetett, hanem a passzív szókészletet is hozzávéve, az így keletkezett, de már jóval nagyobb, a két halmaz uniójából összeállt halmaz elemeiből.

Vizsgálataink elvégzéséhez két modellt építettünk. Mindkét modell dinamikus, hiszen a szavak ténylegesen végrehajtott statisztikailag független válogatásán alapszik. Az első egy korábban ismertetett, az urna modellt alapul vevő statikus modell (Baayen, 1993; 1996a; 2001) mintájára készült. Az említett szerző a szavak válogatását visszatevéses válogatással modellezte, így az N méretű mintában a p_i valószínűségű ω_i szóalakok előfordulása (N,p_i) polinomiális (speciális esetben binomiálisra redukált) eloszlást mutatott. A másik modellünk az egyes szóalakok (ω_i) számára visszatevés nélküli válogatáson alapszik, így egy hipergeometrikus eloszlást eredményező dinamikus modell.

III.3.3.1. Visszatevéses válogatás (P1)

A modell megépítéséhez az eredeti mű szóalakjainak gyakoriságát használtuk fel. Ennek megfelelően először az egyes szavak gyakoriságát (*f(j,N)*), majd a relatív gyakoriságát (*frel(j,N)*) határoztuk meg.

$$frel(j,N) = \frac{f(j,N)}{N} \quad . \tag{3.9.}$$

A szóalakok relatív gyakoriságának ismeretében meg tudtunk határozni az adott eloszláshoz tartozó empirikus eloszlásfüggvényt (*Femp*), ahol minden egyes szóalaknál a relatív gyakoriságok összege szerepel:

$$Femp(j) = \sum_{i=1}^{j} frel(i, N) \ . \qquad (3.10.)$$

Ezen relatív gyakoriságok és a hozzájuk tartozó empirikus eloszlás függvény alapján állítottunk elő egy mesterséges szöveget, amelyben a szóalakok előfordulási gyakorisága megegyezett az eredeti szöveg szóalakjainak relatív gyakoriságával.

Feltételezve, hogy a könyv szóalakjai egymástól függetlenül adott valószínűséggel követik egymást, valamint azt, hogy egy szó felhasználása nem jelenti a szó törlését a szókészletből az eloszlás függvény értékkészletéből véletlenszerűen válogattunk elemeket. A válogatáshoz a számítógép beépített RANDOMIZE és RANDOM függvényét használtuk. A RANDOMIZE függvény inicializálását nagy prímekkel végeztük. Azért választottuk ezt a módszert a számok előállítására, mert így láttuk biztosítottnak, hogy a számok előállítására használt algoritmus független a szövegben előforduló szavak rendszerétől (Ashby, 1972).

Ezt az eljárást annyiszor ismételtük meg, ahány szövegszót tartalmazott az eredeti szöveg. Ennek az eljárásnak azonban az a hátránya, hogy nem pontosan annyi különböző szóalakot állít elő, mint amennyit az eredeti szöveg tartalmazott. A 3.12-13. ábrákon az eredeti szöveg szókészletének nagyságát ($V(N)$) a folyamatos vonal, míg a polinomiális eloszlást feltételező modellel előállított szöveg szókészletének nagyságát ($EP1V(N)$) a szaggatott vonal jelöli.

III.3.3.2. Visszatevéses válogatás, módosított modell (P2)

A szóalakok számának az eredetitől való eltérése az egyszer előforduló szavak (hapax legomena, $V(1,N)$) esetében volt a legnagyobb. Ahhoz, hogy az eredeti és a mesterséges szöveg szóalakjainak száma közötti eltérést csökkenteni

tudjuk a modellt módosítani kellett. Ez a legegyszerűbben úgy történhet meg, hogy megnöveljük azoknak a szóalakoknak a számát, amelyekből a válogatás történt. Ezt azonban úgy kellett elvégezni, hogy az eredeti könyvből nyert relatív gyakoriságok ne változzanak meg. A modell módosított verziójában megnöveltük az egyszer előforduló szavak számát csökkentve ezzel azok relatív gyakoriságát, úgy, hogy az összes egyszer előforduló szó relatív gyakorisága ne változzék (3.13. ábra).

Míg az eredeti műben és modell első verziójában az összes egyszer előforduló szó relatív gyakorisága

$$\mathrm{rel}\big(V(1,N)\big) = \frac{V(1,N)}{N}\ , \tag{3.11.}$$

addig a módosított modellben a rel(V(1,N*)) a

$$\frac{1}{N\cdot\left(1+\dfrac{V2}{V(1,N)}\right)} = \frac{V(1,N)}{N\cdot\big(V(1,N)+V2\big)}, \tag{3.12.}$$

kifejezéssel adható meg, ahol *V2* a hozzáadott szóalakok száma, tehát

$$N^* = N + V2. \tag{3.13.}$$

A módosított modell alapján előállított szöveg szókészletének nagyságát (*EP2V*(*N*)) a 3.12 és 13. ábrán a pontozott görbe jelöli. Az eltérés az eredeti és a mesterséges szöveg között azonban nem lényegesen kisebb, mint a korábban használt statikus modellek esetén (Baayen, 1993; 1996a; 2001; 3.13. ábra). Az eredeti és a mesterséges szöveg közötti különbség csökkentésére ezért egy újabb modellt építettünk.

III.3.3.3. Visszatevés nélküli válogatás (H)

Ebben a modellben a szövegszókat egy vektorban tároltuk majd az így tárolt elemeket véletlenszerűen válogattuk, de ebben az esetben visszatevés nélkül. A már felhasznált szövegszó nem került vissza a vektorba azután, hogy lejegyeztük, hogy melyik volt kihúzva.

Ezt a módszert használva megoldódott az a korábbi probléma, hogy az eredeti és a mesterséges szöveg különböző szóalakjainak a száma nem egyezett meg, ugyanis pontosan annyi szóalak volt tárolva, ahányat az eredeti szöveg tartalmazott, pontosan annyiszor, ahányszor az eredeti szövegben előfordultak.

A visszatevés nélküli válogatás még a módosított (*P2*) polinomiális eloszláson alapuló modellnél is jobb közelítését adta az eredeti szövegeknek.

A visszatevés nélküli válogatással készült modell nemcsak az angol, de a magyar nyelvű szövegek szókészletének közelítő leírására is alkalmasnak bizonyult, függetlenül a két nyelv közötti eltérésektől. Annak ellenére, hogy magyar szövegekben magasabb a különböző szóalakok száma, az eredeti szöveg és a modell között nem nagyobb az eltérés, mint angol nyelvű szövegek esetén.

III.4. A szöveg szezonalitásainak meghatározása

III.4.1. Szóalakok számát megjelenítő grafikonok

Vizsgálatainkhoz tehát az eredeti művekben újonnan megjelenő szóalakok számát használtuk kiindulásként. Megszámoltuk, hogy száz szövegszó hosszúságú blokkokban hány új szóalak (y_i, i = 1, …, h) jelenik meg az előzőekhez képest és az így kapott értékeket ábrázoltuk (3.7/A, 3.8/C-D, 3.10. ábra). Ezek a függvények azonban még nem alkalmasak arra, hogy megbízható következtetéseket vonjunk le a szavak megjelenésének szabályszerűségeire vonatkozóan, mert az újonnan bevezetésre kerülő szavak

számát leíró függvény monoton csökkenő tendenciáját megtörő kiugrások közül nehezen választhatóak ki azok, amelyek szignifikáns eltérés következményei.

A függvény menetének megváltozása, a monoton csökkenő tendencia átmeneti visszafordulása, két okkal is magyarázható. Az elsődleges kiugrások a függvényen jelenlévő trendek, a másodlagos kiugrások pedig az ettől jól elkülöníthető, valamilyen rendkívüli eseménynek a következménye a szövegben, tehát a szezonalitások jelenlétére utalnak. A grafikonról az esetek többségében jól leolvasható, hogy melyek azok a pontok, ahol ezek a rendkívüli események bekövetkeznek, de a grafikon alapján nehéz megmondani, hogy mely változások tekinthetők szignifikánsnak (3.7/A, 3.8/C-D, 3.10. ábra). További feldolgozásra volt szükség tehát annak eldöntésére, hogy az újonnan megjelenő szavakat leíró görbe mely csúcsai jelennek meg a szezonális hatások következtében, melyek azok, amelyek a szövegben végbemenő előre nem jelezhető változás következményei és ezek közül melyek azok, amelyek szignifikáns változás következményei.

Ennek eldöntésére elsőként a mért adatok alapján az újonnan bevezetésre kerülő szóalakok számát ábrázoló görbe simítását kellett elvégezni, az így kapott értékek (yp_i) az *fp* simított görbe függvényértékei. A száz szövegszó hosszúságú blokkok ugyanis kellően rövidek ahhoz, hogy visszaadják a szöveg finomabb változásait is, de éppen e miatt a jelentéktelen változásokra is érzékenyek. Amennyiben a szövegben bekövetkezett változás jelentéktelen, csak abban az egy blokkokban érezteti hatását, úgy az a simítás során eltűnik, ugyanakkor a jelentős változások a simítás után is megfigyelhetők a görbén (3.15/A ábra).

III.4.2. Az eredeti és a mesterséges szövegek összehasonlítása

Ezt a simított görbét hasonlítottuk a modell által előállított mesterséges szöveg szóalakjait ($fm(b_i) = ym_i$) leíró görbék sorozatához (fm_k, $k = 1, \ldots, 100$),

ahol fm_{ki} jelöli a *k*. mesterséges függvény *i*. blokkjában megjelenő szóalakok számát. A modell alapján előállítottunk száz mesterséges szöveget, megszámoltuk ezen szövegekben az újonnan megjelenő szavak számát a száz szövegszó hosszúságú blokkokban és vettük az így kapott függvények átlagát ($F(b_i) = Y_i$) (3.15/B ábra) (Ashby, 1972).

A következő lépésben vettük a simított függvény (*fp*) és az átlag függvény (*F*) különbségét, majd a különbségek átlagát (*M*) és szórását (σ) (Yule, 1950; Hajtman, 1971; Solt, 1971; Nemetz és Kusolitsch, 1999)

$$fp - F\,, \tag{3.14.}$$

$$\Delta y_i = yp_i - Y_i,\ i = 1, \ldots, n, \tag{3.15.}$$

$$M = \frac{1}{n}\sum_{i=1}^{n} \Delta y_i \tag{3.16.}$$

$$\sigma = \sqrt{\frac{(y_i - M)^2}{n}} \tag{3.17.}$$

Azokat az eltéréseket tekintettük szignifikánsnak, amelyek az átlagnál 2σ-val nagyobbak. A 3.15/B és a 4.3., 4.5., 4.10-16., 4.18-19. ábrák mindegyikén tisztán kivehetőek a görbéknek azok a pontjai, amelyek az $M \pm 2\sigma$ tartományon kívül esnek.

III.4.3. A szóalakok számát megjelenítő görbék simítása

Mint azt korábban jeleztük, az újonnan bevezetett szóalakok számát leíró görbéken megfigyelhető zaj jelentősen megnehezíti azon kiugrások helyének meghatározását, melyek szignifikáns eltérés következményei. A zaj csökkentésére egy, a legkisebb négyzetek módszerével, több pontra illesztett

polinomiális simítást alkalmaztunk. Választásunk azért esett erre a módszerre, mert az ilyen típusú simítás

- a függvény görbe alatti területét (esetünkben tehát a könyvben előforduló szóalakok számát, $V(N)$) és
- a csúcsok relatív helyzetét

nem változtatja meg (Worthing és Geffner, 1959; Guest, 1961). Fontos kiemelni, hogy a módszer különösen alkalmas diszkrét, azon belül is egyenletesen változó abszcisszájú (esetünkben a blokk sorszáma), függvények simítására.

A görbeillesztés szempontjából legyen a koordinátarendszerünk origója az i-edik blokk és s a futóindex, amellyel a szomszédos blokkokat jelöljük. Ekkor egy n-ed fokú polinom alakja:

$$fp'_{i+s} = \sum_{k=0}^{n} a_{nk} s^k \tag{3.18.}$$

Ha az illesztést $2m+1$ pontra végezzük (azaz $s = -m, \ldots, 0, \ldots, m$) és az i-edik blokkban a szóalakok száma, a korábbi jelöléseknek megfelelően y_i, akkor a legkisebb négyzetek elve szerint:

$$\frac{\partial}{\partial a_{nk}} \sum_{s=-m}^{m} \left(fp'_{i+s} - y_{i+s} \right)^2 = 0 \tag{3.19.}$$

Tetszőleges a_{nl} ($l = 0, 1, \ldots, n$) együtthatóra elvégezve a deriválást kapjuk, hogy:

$$\sum_{s=-m}^{m} \sum_{k=0}^{m} a_{nk} s^{k+l} = \sum_{s=-m}^{m} y_{i+s} s^l \tag{3.20.}$$

A keresett *fp*' polinom együtthatóit a 3.18. lineáris egyenletrendszer együtthatói adják. Vegyük figyelembe, hogy *fp*' polinomnak csak az *i*-edik blokk helyén felvett értékére van szükségünk, ez adja ugyanis az *i*-edik blokkhoz tartozó simított értéket, az fp_i-t. Ennek megfelelően:

$$fp_i = fp'_i = a_{n0} \tag{3.21.}$$

A 3.18. egyenletrendszert csak a_{n0}-ra kell tehát megoldani. A simított érték így:

$$fp_i = \frac{1}{\sum_{s=-m}^{m} C_s^m} \sum_{s=-m}^{m} C_s^m y_{i+s} \tag{3.22.}$$

ahol a C_s^m együtthatók a 3.18. egyenletrendszer 3.19. feltétel melletti megoldásából adódnak. Figyeljük meg, hogy a simított érték függ a polinom fokszámától (n) és a figyelembe vett eredeti értékektől (y_{i+s}), tulajdonképpen az utóbbiak súlyozott átlaga.

A 3.18. egyenletrendszerben szereplő *n* és *m* (simítási paraméterek) értékeit a simítandó függvény tulajdonságait figyelembe véve kell megválasztani. Adott *m* esetén annál jobban illeszkedik a polinom a kis görbületi sugarú helyekre is, minél nagyobb az *n*. A fokszám növelése azonban oda vezet, hogy a polinom „képes lesz” a zajt is követni, tehát nem szűri meg azt. Ha azonban *n* túl kicsi hasznos részleteket tüntethetünk el. Számos függvény megvizsgálása után a 7 pontra illesztett, másodfokú polinom használata mellett döntöttünk (3.15. ábra).

III.4.4. A szóalakok számát megjelenítő görbék Fourier spektrumának meghatározása

Annak bemutatására, hogy a simítás milyen hatással van a görbéken megfigyelhető zajra, meghatároztuk az újonnan bevezetett szóalakokat leíró függvények Fourier-spektrumát. Ehhez a MicroCal[8] cég kereskedelmi forgalomban kapható *ORIGIN* programjának beépített gyors Fourier transzformációt (Fast Fourier Transform) végző rutinját használtuk.

Mint azt a 3.16. ábra mutatja, a simítás érintetlenül hagyta az alacsony frekvenciákhoz tartozó komponenseket, míg a magasabb frekvenciájú komponenseket szinte teljesen kiszűrte.

[8] http://www.microcal.org

IV. Eredmények

IV.1. Szóalakok megjelenése irodalmi művekben

IV.1.1. Angol nyelvű irodalmi művek újonnan bevezetésre kerülő szóalakjainak elemzése

A III.1-2. fejezetekben megadott módszer segítségével száz szövegszó hosszúságú blokkokat használva ábrázoltuk az újonnan megjelenő szóalakok számát. A vizsgálatoknak ebben a szakaszában arra kerestük a választ, hogy mivel magyarázható, hogy az író bővíti az eddig használt szókészletét, tehát bevezet egy új szóalakot. Találni-e valamiféle kapcsolatot az író stílusa és a bevezetésre kerülő új szavak száma és a bevezetés helye között. Esetleg a mű zsánere vagy talán hossza szabályozhatja, hogy mennyi új szó kerül bevezetésre a szöveg egy meghatározott pontján (3.7.; 3.8.; 3.10. ábra).

A függvények ábrázolásánál a szokásos száz szövegszó hosszúságú blokkokat használtuk, amit két szempontból is fontosnak tartottunk.

- Ahogy azt már korábban említettem (3.10. ábra), hosszabb blokkhosszúságokat használva az új szóalakok száma kiátlagolódhat és így a másodlagos kiugrások eltűnhetnek a függvényről, holott ezek a kiugrások valamiféle változásra utalnak a szövegben.
- A művek egyenlő hosszúságú blokkokra történő felbontásának annyiban van jelentősége, hogy ezeket használva különböző hosszúságú szövegeket össze tudunk hasonlítani. Nem fordulhat elő, hogy egy rövidebb és egy hosszabb szöveget összevetve a rövidebb szöveg ábrázolása sokkal pontosabb, mint a hosszabb szövegé.

Ahogyan az várható volt, az olvasói intuíciók és a teljes szókészlet vizsgálatából kapott eredmények alapján, az újonnan bevezetésre kerülő szóalakok monoton csökkenő tendenciát mutatnak, ahogy haladunk előre a műben (3.7., 3.8.; 3.10., 4.1/A, 4.1/C és 4.2. ábra). Az is igaz azonban, hogy az újonnan bevezetésre kerülő szóalakok nem írhatók le tisztán egy monoton csökkenő függvénnyel, mert bizonyos kiugrások is megfigyelhető ezeken a függvényeken, ami ebben az esetben annyit jelent, hogy a szöveg egy későbbi pontján több új szóalak került bevezetésre, mint az azt megelőzőekben. Felmerült az a kérdés, hogy a szavak egymástól független megjelenését feltételezve a görbék menetének tendenciáját megszakító visszafordulások visszaadhatók-e. A korábbi statikus modelleken (Mandelbrot, 1962; Carroll, 1967; Sichel, 1986; Baayen, 1996a) nem jelentek meg ezek a kiugrások, míg az általunk használt dinamikus modell a kiugrások többségét képes visszaadni. (4.1/A-B ábra). Ezek a kiugrások a műben jelenlévő trendek jelenlétére utalnak és a későbbiekben elsődleges kiugrásként kerülnek említésre.

Vannak azonban művek (4.1/C-D ábra), amelyben olyan kiugrások jelentek meg, amelyek nem írhatók le maradéktalanul a dinamikus modellel sem, amennyiben feltételezzük a szavak függetlenségét. Ezek a kiugrások lesznek a másodlagos kiugrások, amelyek szezonalitások jelenlétére utalnak.

Továbbra is igaz azonban, hogy szövegnek ezen változásai, mind az elsődleges, mind a másodlagos kiugrások többsége, csak akkor jelennek meg a grafikonon, ha a blokkok hossza megfelelően kicsire van választva.

IV.1.2. Szóalakok megjelenése magyar nyelvű regényekben

Ahogy az a magyar nyelv agglutináló tulajdonságát ismerve várható volt hasonló hosszúságú angol és magyar szövegeket összehasonlítva a különböző

szóalakok száma magyar nyelvű szövegekben lényegesen magasabb, mint angol nyelvű szövegekben (3.3. táblázat, 3.8. ábra, 4.2. ábra).

A magas szószámmal együtt a zaj is nagyobb a magyar szövegekben, aminek következménye, hogy kevésbé tisztán látni azokat a pontokat, ahol valóban hirtelen megemelkedik a szóalakok száma.

IV.1.3. A szezonalitások és a szóalakok számának változása

A szezonalitásokra utaló másodrendű kiugrások akkor jelentek meg az újonnan bevezetésre kerülő szóalakokat leíró görbén, mikor valamilyen „rendkívüli" esemény következett be.

Ilyen jellegű rendkívüli esemény lehet

- egy hosszabb lélegzetű leírás megjelenése a műben, bevezetve ezzel egy új helyszínt, szereplőt stb.,
- amikor olyan új szereplőt, szereplőket vezet be az író, akinek, akiknek a beszédstílusa lényegesen eltér a korábbi szereplőkétől,
- idegen kifejezéseket, mondatokat használtat az író a szereplőkkel, vagy
- a történethez csak marginálisan kapcsolódó szövegrész jelenik meg a műben, amely szövegrészek rendszerint a műre egyébként jellemző szókészlettől eltérő szókészletet használnak.

Mindezeknek megfelelően tehát nemcsak azok a változások okoztak törést a monoton csökkenő függvényben, amelyek a történet természetes folyásából következtek, hanem azok is, amelyek valamilyen „rendkívüli" eseménynek tudhatók be. A szubjektív véleményeken alapuló várakozásokkal ellentétben a felsorolt „rendkívüli" események nagyobb kiugrásokat eredményeztek, mint egy újabb fejezet kezdete.

A felhasznált modell képes volt követni a monoton csökkenő menetét az újonnan bevezetésre kerülő szóalakokat leíró függvénynek. A trendek a modell által generált szövegben hasonlóak voltak az eredeti szövegéhez. Ezek a változások a történet logikus folyásából származtak és relatíve kicsi kiugrásokat eredményeztek a grafikonon. Ezzel szemben azonban, a váratlan, a történethez csak marginálisan kapcsolódó szövegrészek okozta kiugrások a mesterséges szövegben nem jelentek meg, ami egybe esett előzetes várakozásainkkal. Kérdés volt, hogy összehasonlítva az eredeti és a mesterséges szöveget meg tudjuk-e adni azokat a pontokat, esetleg rövid intervallumokat, ahol ezek a történethez szervesen nem tartozó szövegdarabok megjelennek.

IV.2. Az eredeti és a mesterséges szöveg összevetése

A kiválasztott példák mutatják, hogy sem a szöveg hossza, szerzője, zsánere, nyelve nem befolyásolja az eredeti és a mesterséges szövegek összehasonlításából kapott eredményeket. Ehhez az összehasonlításhoz kiválasztottunk egy angol (Mark Twain: THE ADVENTURES OF TOM SAWYER, mostantól TOM SAWYER) és egy magyar (Kertész Imre: SORSTALANSÁG) regényt, egy angol novellás kötetet, amelyben minden mű ugyanattól szerzőtől származik (Rudyard Kipling: THE JUNGLE BOOK) és egy olyan gyűjteményt, amelyben hasonló zsánerű művek szerepelnek, de különböző szerzőktől (AMERICAN MYSTERY STORIES): F. Marion Crawford: BY THE WATERS OF PARADISE; Mary E. Wilkins Freeman: THE SHADOWS ON THE WALL; Melville D. Post: THE CORPUS DELICTI; Ambrose Bierce: AN HEIRESS FROM REDHORSE, THE MAN AND THE SNAKE; Edgar Allan Poe: THE OBLONG BOX, THE GOLD-BUG; Washington Irving: WOLFERT WEBBER, OR GOLDEN DREAMS, ADVENTURE OF THE BLACK FISHERMAN.

IV.2.1. THE ADVENTURES OF TOM SAWYER kiugrásainak elemzése

A műben (4.3/A ábra) három olyan kiugrást találunk, amely eléri vagy nagyobb, mint az $M + 2\sigma$ (4.1. táblázat), tehát egy olyan szövegrész jelenik, amelyik nem tartozik szervesen a történethez és stílusában is eltér a könyv egészének stílusától.

Az egyetlen igazán szignifikáns eltérés az átlagtól akkor következik be a TOM SAWYER-BEN, amikor a gyerekek a tanév végére elkészítendő vizsgamunkáikat, verseket, novellákat felolvassák. Ez tehát egy olyan szövegrész, amely stílusában, így a felhasznált szókészletben is eltér a regény egészétől.

IV.2.2. THE JUNGLE BOOK kiugrásainak elemzése

A THE JUNGLE BOOK hét mesét és hét verset tartalmaz. Ezzel szemben négy olyan csúcsot találtunk, amely egyértelműen szignifikáns eltérésre utal és ebből a négyből is csak három esik egybe egy új mese kezdésével vagy közel van ahhoz (4.3/C ábra, 4.2. táblázat). Ez a három mese a WHITE SEAL, RIKKI-TIKKI-TAVI és TOOMAI OF THE ELEPHANTS. Mindhárom érdekessége, hogy új helyszínt vezet be az író és ezzel magyarázható az újonnan megjelenő szóalakok magas száma. (Az ötödik csúcs éppen csak megközelíti a szignifikancia szintet.)

A négy közül, a sorban a legelső kiugrás még a dzsungelben történik, de hasonlóan a másik háromhoz, itt is új helyszínt vezet be az író, a királyi palotát írja le. Itt is, hasonlóan az előző műhöz, a királyi palota leírása stílusában eltér a dzsungelben játszódó történetek stílusától. Sem a többi mese kezdetén, sem a verseknél nem találtunk kiugrást, tehát nem jellemző, hogy hasonló zsánerű műveknél, műrészleteknél megemelkedik az újonnan bevezetett szóalakok száma.

IV.2.3. AMERICAN MYSTERY STORIES novellás kötet elemzése

Ez előző megállapításainkat támasztja alá a különböző szerzőktől származó AMERICAN MYSTERY STORIES gyűjtemény is (4.3/D ábra). Három jól megkülönböztethető csúcsot eredményezett a mesterséges szövegek átlagának és az eredeti szövegnek az összehasonlítása (4.3. táblázat). Mindhárom csúcs a novellák elején lévő egy-egy részletes leírás eredményeként jelent meg a görbén. A három csúcs közül a második igazán figyelemre méltó, ugyanis ez a csúcs Edgar Allen Poe THE GOLD-BUG című történetének kezdeténél jelent meg. Az érdekességét viszont az adja, hogy az ezt megelőző történetnek is Poe a szerzője, tehát itt is látszik, hogy a zsáner a szerzőt is felülmúlja az újonnan bevezetett szóalakok tekintetében.

Ezek a megfigyelések egyértelműen mutatják, hogy az újonnan bevezetésre kerülő szóalakok száma abban az esetben emelkedik meg hirtelen, ha megváltozik a mű korábbi stílusa, valami új kerül ismertetésre, bevezetésre. Sem a szerző, sem az új fejezetek nem eredményeznek olyan látványos emelkedést, mint a zsáner, vagy a regiszter váltása.

IV.2.4. A szöveg hosszának a szerepe

Felmerülhet a kérdés, hogy miért nem okoz a novellák összefűzése kiugrást a szövegben annak ellenére, hogy az önállóan vizsgált szövegek mindegyik magas szószámmal indul.

Azonos nyelvű szövegeket vizsgálva azt tapasztaltuk, hogy az újonnan bevezetett szóalakok száma a szöveg bármely tetszőlegesen kiválasztott pontján nem függ a szöveg teljes hosszától. Ez a megfigyelés egybevág korábban kapott eredményekkel (Muller, 1964; Holmes; 1994; for review see Baayen, 2001), amikor ugyan nem az újonnan bevezetésre kerülő szóalakokat vizsgálták, hanem a szókészlet nagyságát.

Az újonnan bevezetésre kerülő szóalakok száma és a szöveg hossza közötti összefüggés szemléltetésére kiválasztottunk kilenc különböző hosszúságú, különböző szerzőtől származó művet és mindegyiket lecsonkítottuk 150 blokk-hosszúságúra. A 4.4. ábrán a műveket hosszuk szerint sorokba és szerzőik szerint oszlopokba rendeztük.

A 4.4. ábrán az első sorban ábrázolt szövegeket, rövid történeteket nem csonkítottuk, azok eredeti hosszukban láthatóak. Ezért van az, hogy ezeknél a műveknél a blokkok száma nem feltétlenül éri el a százötvenet. A függvényeket összehasonlítva láthatjuk, hogy sem a művek teljes hossza, sem a szerző nem befolyásolja szignifikánsan a bevezetésre kerülő szóalakok számát, ezzel szemben fontos szerepet kap, hogy hanyadik szó olvasásánál tartunk.

Ennek következtében, ha rövid, néhány ezer, esetleg tízezer szavas szövegeket, amelyeknek még a végén is magas az újonnan bevezetésre kerülő szóalakok száma, fűzünk össze, az új novella szóalakjainak a száma nem eredményez szignifikáns változást. Ezzel magyarázható, hogy összefűzött rövid történetek úgy viselkednek, mint egy megfelelő hosszúságú regény, amelynek hossza a rövid történetek hosszának az összege. Az egyetlen különbség, ami felfedezhető egy regény és egy összefűzött rövid történetekből előállított szöveg között, hogy az összefűzött szövegben az átlagos szószám magasabb, mint egy regényben.

A 4.4. ábráról az is leolvasható, hogy a szerzőnek ebből a szempontból sincs olyan jelentősége, amely alapján az így kapott információkat szerzőazonosításra tudnánk használni. Természetesen vannak szerzők, akik az átlagosnál magasabb szószámmal, tehát kisebb ismétlési rátával dolgoznak, de ezek az értékek sem térnek el lényegesen egymástól. Az olvasói intuíció alapján azt várnánk, hogy különböző szerzők eltérő szókészletet használnak, így műveiket összefűzve alacsonyabb ismétlődési rátát kapunk, mint egy szerző műveit vizsgálva. Ezzel szemben azt tapasztaltuk, hogy különböző szerzők

összekapcsolt művei alacsonyabb ismétlési rátát is adhatnak, mint egy sok szót használó szerzőé (4.4. táblázat).

$$rep = \frac{N}{V(N)} \tag{4.4.}$$

IV.3. A hapax legomena szerepe

Az angol nyelvű szövegek vizsgálatakor kapott másodlagos kiugrások tehát akkor következnek be, amikor a soron következő szövegrész sem az előzményekhez nem kötődik, sem a későbbiekhez való szerves kapcsolódást nem készíti elő. Olyan szövegrészek, amelyekhez nem találni a mű más pontjain vele rokon témát/témákat, olyanokat, amelyek az itt találthoz hasonló szókészletet használnak. A 3.14, a 4.3., 4.10-18. ábrákon jól látható kiugrásokon túl ugyanezt támasztja alá az egyszer előforduló szavak vizsgálata is (4.5. ábra). Ugyanazokon a helyeken növekedett meg az egyszer előforduló szavak száma, ahol az eredeti műben szintén magas volt az újonnan bevezetett szavak száma. Ez a megfigyelés is arra enged következtetni, hogy a görbéken található kiugrások a szöveghez szervesen nem kapcsolódó részeknél jelennek meg.

IV.3.1. A hapax legomena előfordulása

A hapax legomena szövegen belüli előfordulásának vizsgálatára, hasonlóan a szóalakok számához, a száz szövegszó hosszúságú blokkokat használtuk. Megszámoltuk, hogy hány darab hapax legomena szerepel egy blokkokban (yh_i, $i = 1,\ldots,n$) és ezt ábrázoltuk (4.5. ábra). A hapax legomena $V(1,N)$ számának és a szöveg hosszának (N) ismeretében meghatároztuk a hapax legomena relatív gyakoriságát. Mint azt a III.3.3.3. fejezetben ismertettük, modellünk egy visszatevés nélküli válogatásnak tekinthető. Ha ebben a

modellben az *A* esemény bekövetkezése az, hogy *h* húzásból *k* darab egyszer előforduló szót választunk ki, akkor *A* hipergeometrikus eloszlást követ. Ezt kihasználva meghatároztuk az így definiált valószínűségű változó várható értékét (*Mh*) és szórását (*SDh*). Az $Mh \pm 2 \cdot SDh$ értéket ábrázoltuk 4.5. ábrán vízszintes vonalakkal.

Összehasonlítva az eredeti műben megjelenő szóalakok számát (4.10. és 4.11. ábra) és a hapax legomena előfordulását a 4.5. ábra nyilai mutatják az egyezést. A fekete nyilak a szövegnek azokat a pontjait mutatják, ahol mindkét esetben kaptunk kiugrást, míg a szürke nyilak azt jelzik, hogy az eredeti szövegben volt kiugrás, de a hapax legomena vizsgálatánál nem.

IV.3.2. Hapax legomena eloszlása

Annak igazolására, hogy az egyszer előforduló szavak az eredeti művekben követik-e a modell által jósolt hipergeometrikus eloszlást a 4.6-8. ábrákon a hapax legomena eloszlását (•), valamint az adott műhöz tartozó *N*, *V*(1,*N*) és *h* segítségével számolt hipergeometrikus eloszlást (■) ábrázoltuk.

Tekintettel arra, hogy korábbi irodalmi adatok és egyik korábbi modellünkben (III.3.3.1-2.) mi is a visszatevéses válogatást használtuk, az ábrákon az ebből következő binomiális eloszlást (▲) is bemutatjuk.

Ha yh_i-vel ($i = 1,\ldots, n$) jelöljük a blokkonkénti hapax legomena számát, akkor meg tudjuk adni, hogy hány olyan blokk van, amelyekben *r* darab hapax legomena fordul elő, ha a blokkok hossza *h* ($r = 0,\ldots, mh$, ahol $mh = \max(yh_i)$). Legyen azoknak a blokkoknak a száma

hap(0,*h*), amelyekben 0 darab hapax legomena,

hap(1,*h*), amelyekben 1 darab hapax legomena, … és

hap(*mh*,*h*), amelyikben/amelyekben a legtöbb hapax legomena van.

$$hap(r,h) = \sum_{i=1}^{n} I_{[yh_i=r]} \tag{4.5.}$$

$$V(1,N) = \sum_{j=1}^{mh} hap(j,h) \tag{4.6.}$$

Hasonló módon, a hipergeometrikus eloszlás feltételezése esetén kapott blokkok számát jelöljük *haphip*(0,*h*),…, *haphip*(mhhip,*h*), illetve a binomiális eloszlás esetén *hapbi*(0,*h*),…, *hapbi*(*mhbi*,*h*),

A viszonylag egyenletes eloszlású The ADVENTURES OF ROBINSON CRUSOE-ban (mostantól ROBINSON) $hap_{Robinson}(mh,100) = 10$, $i = 8$ és 815, míg a TOM SAWYER-ben, amelyben találtunk nagy kiugrást, $hap_{Tom}(mh,100) = 25$, $i = 440$.

Függetlenül a *hap*(*mh*,*h*) értékétől azt találtuk, hogy azon blokkok gyakoriságát, amelyek nagy számú hapax legomenát tartalmaznak alul becsüli a modell. A különbség az egyenletesebb eloszlású ROBINSON és a nagy kiugrást tartalmazó TOM SAWYER között annak az intervallumnak a hosszában van, ahol *haphip*(*k*,*h*)<*hap*(*k*,*h*), ahol $k = l,\ldots, mh$. l az a szám, hapax legomena száma, ahol a mért és a modell értékei között a relatív eltérés kisebb, mint 50% ($l_{Robinson} = 6$, $l_{Tom} = 11$).

Magyar szöveg, Gárdonyi Géza EGRI CSILLAGOK, egyszer előforduló szavainak számát vizsgálva is hasonló eredményt kaptunk: $hap_{Egri}(mh,100) = 17$ és $l_{Egri} = 11$.

Az így kapott l értékek segítségével meg tudjuk határozni azoknak a blokkoknak a számát, ahol a hapax legomena száma magasabb, mint az a modell alapján várható lenne, és ebből meghatározhatóak azok a blokkok, ahol megemelkedett az újonnan bevezetett szóalakok száma.

IV.4. Magyar nyelvű irodalmi művek

Mivel a magyar nyelv alapjaiban agglutináló nyelv legalább két olyan tulajdonsággal rendelkezik, amelyeket vizsgálatainknál érdemes figyelembe venni.

- A különböző szóalakok száma magasabb magyar nyelvű szövegek esetén, mint egy hasonló hosszúságú angol szövegben, mivel a ragok és a képzők a szótőhöz kapcsolódnak (O'Grady et al, 1993; Laczkó, 2000; Kiefer, 1998; Farber, 1991; 3.3. táblázat).
- Magyar nyelvben a mondaton belüli szórend nem kötött, tehát a szavaknak szinte bármilyen sorrendje elfogadható, attól függően, hogy melyik szót tesszük fókuszba (É. Kiss, 1998; Kugler, 2000; 3.4. táblázat).

Figyelembe véve a magyar nyelv sajátosságait a kérdés az volt, hogy

- a különböző szóalakok bevezetése és a zaj a függvényen eltér-e az angol szövegektől vagy hasonlít azokhoz,
- a modell használható lesz-e magyar szövegek leírására is,
- a másodlagos, csak marginális szerepet játszó eseményeket leíró szószám-növekedés hasonló módon vizsgálható-e, mint angol szövegek esetén,
- a kötetlen szórend miatt jobb közelítést ad-e a modell, mint kötött szórendű nyelvek esetén?

Azt találtuk, hogy a szóalakok száma valóban magasabb magyar nyelvű, mint ugyanolyan hosszúságú angol szövegekben (3.8. ábra, 4.5. táblázat). A szóalakok bevezetését leíró függvény monoton csökkenő tendenciája megmaradt és a függvényen meglévő zaj is hasonló mintájú (pattern), de nagyobb, mint az angol szövegeken (3.8, 3.10; 4.2. ábra).

A szavak függetlenségét feltételező modell az angol nyelvű szövegekhez hasonlóan magyar nyelvű szövegek leírására is alkalmasnak bizonyult.

Összehasonlítva az eredeti és a mesterséges szöveg szóalakjainak a számát hasonló mértékű eltérést tapasztalunk az angol és a magyar nyelvű szövegek esetén (3.13-14. és 4.9. ábra).

IV.4.1. SORSTALANSÁG másodlagos kiugrásainak elemzése

Kertész Imre SORSTALANSÁG című művében hat olyan pontot találtunk, amelyeknél hirtelen megemelkedik az újonnan bevezetésre kerülő szóalakok száma (3.15. és 4.3/B ábra, 4.7. táblázat). Ezt a hat szignifikánsnak tekinthető eltérést vizsgáltuk meg részletesen. A kiugrások valamennyien olyan helyen fordultak elő, ahol a szöveghez szervesen nem kapcsolódó, a korábbi eseményektől függetleníthető leírás jelent meg a szövegben. A hat esemény, a megjelenés sorrendjében, a következő volt: megérkezés a koncentrációs táborba, megérkezés a második táborba, reggeli események és az üzem leírása, kórház leírása, Pjetyka főz, haza indulás. Ebben az esetben is arra a következtetésre jutottunk tehát, hogy azok a helyek azonosíthatók be az ismertetett módszerrel, amelyek váratlanul szakítanak a mű korábbi folyásával.

IV.5. Szöveg és fordításainak összehasonlítása

A kérdés az volt, hogy a mondatok belső kohéziója, tehát a szintaktikai szabályok befolyásolják-e, s ha igen mennyiben az új szóalakok megjelenését, illetve származhatnak-e más forrásokból az eredeti és a mesterséges szöveg közötti eltérések.

Annak további bizonyítására, hogy az eredeti szöveg másodlagos kiugrásait a szöveg szinten bekövetkező változások idézik elő a szöveget annak fordításaival hasonlítottuk össze (Fodor, 1999; Fodor, 2003). A választás egy

- magyar regényre (Kertész Imre: Sorstalanság) és annak

- német (ROMAN EINES SCHICKSALLOSEN[9]) és
- angol nyelvű fordításaira (FATELESS[10]),

– egy angol novellás kötetre (Rudyard Kipling. THE JUNGLE BOOKS) és annak

- magyar fordítására (A DZSUNGEL KÖNYVE[11]), valamint

– két angol meseregényre (Lewis Caroll: ALICE'S ADVENTURES IN WONDERLAND és THROUGH THE LOOKING GLASS, mostantól ALICE)

- és ezek magyar fordításaira (ALICE CSODAORSZÁGBAN[12] és ALICE TÜKÖRORSZÁGBAN[13]) esett.

A művek kiválasztását nagyban befolyásolta, hogy melyek azok, amelyeknek létezik fordítása legalább nyomtatott formában, valamint az, hogy a kiválasztott nyelvek szerkezetükben eltérőek legyenek. A szerint csoportosítva ugyanis, hogy a morfémákból hogyan képzi a nyelv a szóalakokat a három nyelv három különböző kategóriába sorolható. A német a flektáló, a magyar az agglutináló nyelvek csoportjába tartozik, míg az angol leginkább az izoláló nyelvek kategóriájába tartozik, de amiatt, hogy több különböző kategória eszközeit is felhasználja, így igazán egyikbe sem illik bele (Prószéky, 1989; O'Grady, 1993; Quirk et al., 1995; Uzonyi, 1996; É. Kiss, 1998; Kiefer, 1998; Kugler, 2000; Laczkó, 2000;).

---

[9] Aus dem Ungarischen von Christina Viragh. (1996) Rowohlt Taschenbuch Verlag, Hamburg

[10] Translated by Christopher C. Wilson and Katharina M. Wilson (1992) Hydra Books, Northwestern University Press, Evanston, Illinois

[11] A fordítás a MACMILLAN, LONDON KIADÓ 1930. évi kiadásából készült. Benedek Marcell fordítása, a verseket Weöres Sándor fordította (1976) Móra Könyvkiadó, Budapest

[12] Fordította Kosztolányi Dezső. A fordítást az eredetivel egybevetve átdolgozta Szobotka Tibor (1974) Móra Ferenc Könyvkiadó, Budapest

[13] Fordította Révbíró Tamás. A versbetéteket Tótfalusi István fordította (1980) Móra Ferenc Könyvkiadó, Budapest

A 4.6. táblázat értékei mutatják, hogy az egyes nyelvek sajátosságaiból, valamint a fordításból adódóan a szövegszók, a különböző szóalakok és az egyszer előforduló szavak száma között lényeges eltérések mutatkoznak az egymásnak megfelelő szövegek esetén. A szintaktikai és szemantikai szinten bekövetkező változások elemzésére ez a módszer tehát nem lehet alkalmas, abban az esetben viszont, ha a szignifikáns eltérések a szöveg szinten következnek be, akkor állításunk bizonyítást nyer.

IV.5.1. A ROMAN EINES SCHICKSALLOSEN másodlagos kiugrásainak elemzése

A német nyelvű szövegben hét kiugrás található (4.10. ábra, 4.7. táblázat, Csernoch, 2004), amelyek közül az első nem a táborba érkezést, hanem egy korábbi eseményt, a vonatra szállást írja le. Várhatóan azért nem kaptunk a német szövegben újabb kiugrást a táborba érkezéskor, mert a vonatra szállás, a vonat leírására használt szavak nagyban fedik a tábor jellemzésére használt szavakat.

A második és a harmadik kiugrás ugyanannál a szövegrésznél következett be, mint a magyar szövegben. A német szövegben akkor jelenik meg a negyedik kiugrás, amikor a főszereplő pillanatnyi lelkiállapotáról következik egy leírás. Ez a leírás a magyar szövegben szignifikáns eltérést nem, csak ahhoz közelit eredményezett (402. blokk).

Végül az utolsó három kiugrás újra teljes egészében megegyezik a magyar szöveg kiugrásaival. (A német szöveg utolsó kiugrása még éppen az elfogadhatósági intervallumon belül esik, de ez várhatóan annak tudható be, hogy a digitalizálás során egy ének a magyar szövegben szótagolva került be, míg a német szövegben a szavak egybe vannak írva.)

IV.5.2. A FATELESS másodlagos kiugrásainak elemzése

Az angol szöveg elemzésekor is hasonló eredményeket kaptunk (4.11. ábra, 4.7. táblázat). Olyan helyeken jelentkeztek a görbén kiugrások, ahol a műbe egy hosszabb lélegzetű leírás került. Ezek nagy része most is megegyezett a magyar (német) szöveg kiugrásaival. Annyiban történt változás, hogy az angol szövegben összesen nyolc csúcs tekinthető lényeges eltérésnek a szöveg megszokott menetéhez képest. A magyar és a német nyelvű szöveghez képest megjelent a szöveg elején két kiugrás, amely további részletes leírást ad. A középső négy kiugrás megegyezik a másik két szöveg kiugrásaival, míg a két utolsó olyan leírás, amely csak az angol szövegben okozott szignifikáns eltérést, de jellegét tekintve ezek is hasonlóak az előzőekhez: valamiféle, a szöveg egészét tekintve váratlan leírás jelent meg a műben. Az utolsó és az utolsó előtti szignifikáns kiugrások között az ábrán jól látható további két kiugrás, amelyek közel szignifikánsak. Ezen kiugrások helyén talált események egybe esnek a magyar és a német szövegben találtakkal.

Az angol és a német fordítás közötti eltérés már a hapax legomena vizsgálatánál is szembetűnő volt. Az eddigi eredményeket felhasználva lehetőségünk nyílik arra, hogy összehasonlítsunk irodalmi műveket azok fordításaival, és a felhasznált szókészletek alapján következtetéseket vonjunk le a fordítás minőségére vonatkozóan.

IV.5.3. THE JUNGLE BOOKS és A DZSUNGEL KÖNYVE összehasonlítása

Vizsgálatainknak ebben a szakaszában nemcsak a THE JUNGLE BOOKS első kötetét dolgoztuk fel, hanem mindkettőt. Már az első kötet elemzésénél is láttuk, hogy a modell és az eredeti szöveg közötti szignifikáns eltéréseket azok a

leírások eredményezték, amelyek a történethez csak marginálisan kapcsolódnak, és inkább formai, hangulati mintsem tartalmi, a szöveg megértéséhez nélkülözhetetlen szerepet töltenek be.

Az eredeti novellás kötetet összehasonlítva a magyar fordítással az első figyelemre méltó eltérés, hogy nem egyezik az elbeszélések sorrendje. A magyar fordításban nem tartották meg az eredeti sorrendet, hanem előre kerültek a dzsungelben játszódó és a könyv végére a más helyszínű történetek. A történetek relatív sorrendje egy-egy eltéréstől eltekintve megegyezik az eredetivel (4.8. táblázat).

A 4.12. és a 4.15. ábrák az eredeti, angol, míg a 4.13. és a 4.14. ábrák a magyar sorrendű szövegek alapján készültek. Az angol sorrendű ábrákon a négyzetek (□) jelzik azokat a helyeket, ahol vagy az angol vagy a magyar nyelvű szövegben szignifikáns eltérést találunk az újonnan bevezetett szóalakok számában. A magyar sorrendű ábrákon a körök (○) jelzik ugyanezeket a pontokat. Az ábrákon megjelölt helyek sorszámát a megfelelő táblázatban az első oszlop számai adják.

Mivel a magyar fordítás nem a megjelenés sorrendjét, hanem valamiféle logikai sorrendet követ, így nem meglepő, hogy a King's Ankus sokkal nagyobb kiugrást eredményezett, mint eredetileg. Az eredeti sorrendet tartva ugyanis a King's Ankus történetét már megelőzte egy emberekkel kapcsolatos történet, The Miracle of Purun Bhagat. A magyar sorrendnél a szórás ($\sigma = 3{,}1866$) is nagyobb, mint az angolnál ($\sigma = 3{,}0275$). A magyar sorrendnél sokkal kiegyensúlyozottabb a novellák elrendeződése, így egy, a sorba nem illő novella sokkal nagyobb kiugrást eredményezhet.

Az egyező sorrendű műveket összehasonlítva azt tapasztaljuk, hogy a szignifikáns, vagy közel szignifikáns kiugrások a szövegnek ugyanazon a

pontján jelennek meg, akkor, amikor a szöveg stílusától eltérő szövegrész jelenik meg a műben.

IV.5.4. THE ADVENTURES OF ROBINSON CRUSOE elemzése

A 4.16. és a 4.17. ábrákat összehasonlítva azt tapasztaljuk, hogy az eredeti, a simítás előtti görbéből kivonva a mesterséges szövegek átlagát egy olyan függvényt kapunk, amelyről szinte teljesen eltűnnek a széles, több blokk hosszúságú kiugrások. Ezek helyén rövid, az esetek többségében egy blokk hosszúságúak a kiugrások. Hozzávetőleg így is látni, hogy hol történt szignifikáns változás a szóalakok számában, de nehéz elkülöníteni a valóban szignifikáns változásokat és jelen vizsgálat szempontjából jelentékteleneket.

IV.5.5. ALICE'S ADVENTURES IN WONDERLAND és THROUGH THE LOOKING GLASS művek elemzése

Azért esett a választás az ALICE történetekre, mert Petőfi S. János (1990) ezek egy részletes elemzését adja, összevetve az eredeti angol szöveget és annak fordítását. Arra voltunk kíváncsiak, hogy a műben megfogalmazott, a szóhasználatra vonatkozó szubjektív vélemények mennyiben támaszthatók alá a számítógépes feldolgozás eredményeit figyelembe véve. (Itt csak azokat az észrevételeket említeném, amelyek vizsgálataink szempontjából érdekesek.) Petőfi úgy ítéli meg, hogy a magyar fordítások nem követik hűen az eredeti szöveget sem szöveg-, sem képanyagban. Ezen túl azt találta, hogy vannak a műben olyan fejezetek, amelyek között az átmenetek sokkal gördülékenyebbek, mint ahogy az egy fejezet határon várható. Ezen, a fejezetek közötti szokatlan átmeneteket rendhagyó formai elemekkel jelzi a szerző.

Hasonlóan más művekhez azt találtuk, hogy akkor emelkedett meg az újonnan bevezetett szóalakok száma, amikor a szövegben olyan szövegrész

jelenik meg, ami eltér a mű stílusától. Ezeknél a műveknél nem mondhatjuk, hogy a versek, amelyeknél megnövekedett a szóalakok száma csak kiegészítő szerepet töltenek be műben, itt a stílusváltásnak van meghatározó szerepe. A fejezet határok viszont abban az esetben sem hoztak látványos szóalakszám emelkedést, amikor a hagyományos módon kapcsolódtak egymáshoz. Így tehát a formabontó fejezet összecsatolások nem eredményezték a szóalakok számának változását, azok rendhagyó viselkedését.

A nagyobb kiugrások mind a két nyelven megjelentek ugyanazokon a helyeken. A magyar szövegben találtunk a szöveg végén egy kiugrást, ami hiányzott az angolból, ez várhatóan a fordító szóhasználatának következménye. Az angol szövegben találtunk négy olyan kiugrást (4.13. táblázat 2-5. sorok), amely a magyar szövegben nem jelent meg. Ezeket a kiugrásokat elemezve azt kaptuk, hogy nagyon rövid intervallumon jelentek meg, maximum három blokkot érintve, és a méretük is jelentéktelen, alig érik el az $M + 2\sigma$ értéket. Úgy gondolom, hogy kétféle magyarázat is adható erre az eltérésre. Az egyik, hogy szövegnek a fordítása ezen a darabon tényleg nem adja vissza az eredeti művet, míg a másik lehetséges magyarázat, hogy a magyar szöveg zajosabb, mint az angol a magyar agglutináló tulajdonsága miatt. Ennek következtében a rövid intervallumra kiterjedő, alig jelentős változások a magyar szövegben eltűnnek.

A szignifikáns eltéréseket okozó kiugrások, az újonnan bevezetett szavak számának hirtelen emelkedése, szerkezetükben is eltérőek lehetnek. Az egyes kiugrások kiterjedése, a függvényen megjelenő lokális csúcsok szélessége, nagyban befolyásolja, hogy egyes kiugrások okoznak-e szignifikáns eltérést.

A TOM SAWYER-ben egy igazán látványos és kettő kisebb kiugrást találtunk. A nagy kiugrás tizenhat (433-448 blokkok) blokkot érint, míg a két kisebb négy-négy blokkot.

Ha megnézzük a görbét (4.3/A ábra), akkor azt látni, hogy további pontok is közel vannak a szignifikancia küszöbhöz, és hasonlóan a ROBINSON-hoz kiugrásokat eredményeztek volna egy sokkal egyenletesebb eloszlás esetén.

Az ALICE történetek kiugrásai két csoportba sorolhatók aszerint, hogy milyen hosszúságú intervallumon vannak jelen (4.14. táblázat). Az egy blokk hosszúságú kiugrások csak kisebb, a stílustól nem vagy csak kevésbé eltérő szövegrészekre utalnak. Ezek többsége a magyar fordításban meg sem jelent, ami lehet a magyar szövegeken lévő zaj következménye, de a fordításból is eredhet.

V. Megbeszélés

Jelen tanulmányban bemutatott módszerek és eredmények a számítógépes nyelvészet egy igen speciális részterületéhez tartoznak, amely az irodalmi művek számítógéppel segített lexikai statisztikai elemzése elnevezést kapta.

Ez a tudományterület, mint a bemutatott eredményekből is látható, az emberi gondolkodás részeként megjelenő szövegalkotás menetét próbálja matematikai, statisztikai módszerekkel modellezni. Valamennyiünk számára ismert azonban, hogy a szövegalkotás folyamatát befolyásolják a szintaktikai és szemantikai kötöttségek. Ezen kötöttségek meg-/betartása nem törvényszerű, de feltételezhetjük, hogy az írók többsége igazodik ezekhez a szabályokhoz.

Három különböző szintjét szokás a szövegnek megadni, de a mondat és a bekezdés szint egymástól nehezen szétválasztható fogalom. Mindkettőnek megvan a maga határoló jele, ami megkülönböztetné őket, de ettől többről van szó, hiszen a szintaktikai formák szerveződésének elsősorban szemantikai okai vannak (Dobi, 2002), míg a másik oldalról az igaz, hogy a szintaktikai kötöttségek túl nyúlnak a mondat határain, de viszonylag ritka azoknak az eseteknek a száma, amikor a bekezdés határt is átlépik előre és visszamutató hivatkozások (Kiefer, 1983).

Első pillanatra azt lehetne gondolni, hogy egy, a szavak függetlenségét feltételező (randomness assumption) modell azért nem képes visszaadni a szóalakok eredeti eloszlását, mert az író követte a szintaktikai és szemantikai szabályokat, míg a random válogatás ezt nem képes megtenni.

Ha azonban nemcsak felületes szemlélőként nézzük a problémát, akkor kiderül, hogy az eredeti szöveg és a modell közötti eltérést nem ezek a

megkötések okozzák, hanem a szöveg szinten bekövetkező változások. Baayen megmutatta, hogy valóban nem a mondat szinten jelenlévő megkötések a felelősek az eltérésért, de ugyanakkor nem vette figyelembe a szintaktikai és szemantikai eszközök kölcsönös egymásra hatását, amely már eleve kizárja, hogy a bekezdés szinten zajló események meghatározzák az eredeti szöveg és a modell közötti eltérést.

Az irodalmi művek, az egyes szerzők szóhasználatára vonatkozó olvasói intuíciók alapján kialakult szubjektív vélemények nem minden esetben hitelesek. A művek szóhasználatának elemzésére irányuló kísérleteket a számítógéppel segített feldolgozások nagyban megkönnyítik és a kapott eredmények lényegesen objektívebbek, mint azok, amelyek a szöveg olvasása közben fogalmazódnak meg (Holmes, 1994). Példaként említeném azokat a korábban végezett vizsgálatokat, amelyekben azt kérdezték az olvasóktól, hogy szerintük mennyire gazdag a kiválasztott művek szókészlete (Hoover, 2003). A vizsgálat részletei, körülményei sajnos nem ismertek, de azt lehet tudni, hogy a szubjektív vélemények és a számítógépes feldolgozáson alapuló pontos értékek egészen más sorrendjét adták a műveknek.

Az újonnan bevezetett szóalakok számának vizsgálatára ilyen jellegű összehasonlító felmérések (tudomásom szerint) nem készültek. Azt tudjuk, hogy anyanyelvünkön írt könyvek olvasásakor nem igazán érzékeljük az újonnan bevezetett szóalakok számának változását. Ezzel szemben, ha idegen nyelvű szöveget olvasunk, akkor a könyv vége felé közeledve egyre könnyebbé válik az olvasás. Ezt tesztelni egy nem túl egyszerű módszerrel, de lehetséges: ha az olvasó minden, számára ismeretlen szót kikeres a szótárból, akkor előre haladva a szövegben egyre ritkábban kell a szótárhoz forduljon, azt sejtetve tehát, hogy a szövegben előre haladva csökken az újonnan bevezetett szóalakok száma. Az újonnan bevezetett szóalakok számítógépes feldolgozása azonban, függetlenül az olvasó anyanyelvétől, objektív képet ad.

V.1. Az újonnan bevezetett szóalakok számának változása

Az újonnan bevezetett szóalakok száma teoretikusan egy monoton csökkenő, diszkrét pontokból álló függvény. A gyakorlatban ezzel szemben azt tapasztaltuk, hogy bár a függvény valójában monoton csökkenő tendenciát mutat, vannak intervallumok, amelyekben a folyamat visszafordul: a szöveg egy adott pontján magasabb az újonnan megjelenő szóalakok száma, mint az azt közvetlenül megelőző pontokban.

Olvasói intuíciók és korábban megjelent szubjektív vélemények alapján azt várnánk, hogy a szövegben bekövetkező szignifikáns változások fejezet határokra esnek. Ez azt jelentené, hogy az újonnan bevezetett szóalakok száma akkor változik meg látványosan, amikor az író egy új fejezetet kezd. Ezekkel az előzetes várakozásokkal (Balázs, 1985; Petőfi, 1990) szemben azonban az újonnan megjelenő szavakat leíró görbe kiugrásai nem fejezet határokat jelölnek, tehát eredményeink nem támasztják alá ezeket a véleményeket.

Ezt a THE JUNGLE BOOKS egymást követő meséinél jól láthattuk. A SORSTALANSÁG és fordításaiban az újonnan megjelenő szavakat leíró görbék vizsgálatán túl a különböző nyelvű szövegekben a fejezet határokat is egyeztetve azt tapasztaltuk, hogy a FATELESS fejezet határai a történetnek nem ugyanazon a pontjaira esnek, mint a SORSTALANSÁGban. Az ALICE történetek különböző tipográfiai eszközökkel történő fejezet váltásai sem okoztak eltérést az újonnan megjelenő szóalakok számának a viselkedésében. Ezek a példák mind mutatják, hogy a fejezet határoknak nem az az elsődleges funkciója, hogy magas szóalak-számmal új információkat közöljenek.

Ezt a szerepet a fejezet határoktól függetlenül megjelenő, terjedelmesebb leírások töltik be, amelyek akkor jelennek meg a szövegben, amikor váratlanul

egy hosszabb, a történethez nem szervesen kapcsolódó leírást szúr be az író. Ezeknek a leírásoknak inkább formai szerepe van, amelyek kihagyásával, esetleg rövidítésével a mű még mindig teljesen érthető lenne. Genette (1980) szerint ezeknek a szövegrészeknek gyakorlati haszna nincs, és csak esetlegesen jelennek meg a szövegben („functionally useless and contingent details"). Erre a következtetésre Homérosz ILIÁSZ című művének és Platon az ILIÁSZ alapján készített rövidített verziójának összehasonlítása alapján jutott. Nagyon fontos azonban hangsúlyozni, hogy ez a kijelentés nem azt állítja, hogy az irodalmi művekben ezekre az eszközökre nincs szükség. Igenis, nagyon fontos szerepet töltenek be, hiszen ettől lesznek az egyes művek egyediek, ezek teszik a szöveget egy teljes egésszé, irodalmi művé. Genette annyit mond csak, hogy a szöveg megértéséhez nem elengedhetetlenül szükségesek.

Azt, hogy egy szövegrész fölösleges, vagy csak gyakorlati haszna nincs, leginkább azzal lehet megadni, hogy mennyire releváns a közölt információ (Sperber és Wilson, 1988), ez azonban már egy más tudományterülethez tartozik, amely túl mutat jelen tanulmány keretein.

V.2. A fordítások és az egyszer előforduló szavak

Korábbi vizsgálatainkat azzal egészítettük ki, hogy különböző nyelveken írt irodalmi művek összehasonlítását végeztük el. Ahhoz, hogy összehasonlítható eredményeket kapjunk olyan műveket kerestünk, amelyek több különböző nyelven is elérhetőek. A különböző nyelvű szövegek feldolgozásával kapott értékek mutatják, hogy az egyes nyelvek sajátosságaiból, valamint a fordításból adódóan a szövegszók, a különböző szóalakok és az egyszer előforduló szavak száma között lényeges eltérések mutatkoznak az egymásnak megfelelő szövegek esetén. Mivel az eredeti állításunk az volt, hogy a modell és az eredeti szöveg

közötti eltérések a szöveg szinten bekövetkező változásokkal magyarázhatóak, ezért kérdés volt az is, hogy a művek különböző nyelvi reprezentációi ugyanazoknál a témáknál eredményeznek-e kiugrásokat, tehát hirtelen növekedést a szavak számában.

Eredményeink ismételten azokkal az előzetes várakozásokkal, leginkább szubjektív véleményekkel egyeztek, amelyek a hosszabb lélegzetű, a műhöz szervesen nem kapcsolódó szövegrészeknél érzékelték a szóalakok számának emelkedését, szemben azokkal, akik fejezet határokra várták ezeket. Különös tekintettel arra, hogy a fordításokban nem feltétlenül ugyanott vannak a fejezet határok, mint az eredeti szövegben vagy egy másik fordításban.

Az említett kiugrások tehát a nyelvi reprezentációtól függetlenül akkor következnek be, amikor a soron következő mondatok sem az előzményekhez nem kötődnek, sem a későbbiekhez való szerves kapcsolódást nem készítik elő. Olyan szövegrészek, amelyekhez nem találni olyan témát a mű más részein, amelyhez a bennük foglaltak kapcsolódnának.

Ugyanezt támasztja alá az egyszer előforduló szavak vizsgálata is. Ezen vizsgálatok elvégzéséhez azt a feltételezést vettük alapul, hogy az egyszer előforduló szavak hipergeometrikus eloszlást követnek. Ugyanazokon a helyeken növekedett meg az egyszer előforduló szavak száma, ahol az eredeti műben szintén magas volt az újonnan bevezetett szavak száma. Ez a megfigyelés is arra enged következtetni, hogy a görbéken található kiugrások a szöveghez szervesen nem kapcsolódó részeknél jelennek meg.

Művek és fordításainak összehasonlítása, az egyszer előforduló szavak vizsgálata, a dolgozatban ismertetett módszer segítségével, nemcsak arra adott választ, hogy a szöveg szinten bekövetkező változásokkal magyarázható az eredeti és a mesterség szöveg közötti eltérés, hanem alkalmas lehet a fordítások összevetésére is.

A vizsgálatokhoz elvégzéséhez a SORSTALANSÁG német és angol fordítását használtuk fel. A megjelenő szóalakok és a blokkonként megjelenő hapax legomena összehasonlítása során azt tapasztaltuk, hogy míg a német nyelvű szövegben pontos volt az egyezés, addig az angol fordításban nem teljesen. Ez azt jelenti, hogy az angol szövegben az adott intervallumokon nem jelent meg annyi új szóalak, tehát már a korábban bevezetett szóalakok újra felhasználása történik meg, így relatíve szegényesebb a szókészlete. Csak az angol szöveget ismerve mondhatnánk, hogy azért nem jelent meg az egyszer előforduló szavaknál minden várt kiugrás, mert olyan információt közölt a szerző, amely kapcsolódik a szöveg folytatásához. Így viszont, hogy elvégeztük a német szöveg elemzését is és annak során pontos egyezést találtunk, bizton állíthatjuk, hogy az angol szöveg tényleg egy szűkebb szókészletet használt a vizsgált intervallumokban.

V.3. További lehetséges felhasználások

A görbéken megjelent kiugrások további vizsgálatának része volt, hogy a Módszerek fejezetben ismertetett eljárást, amely szerint a simított görbét hasonlítottuk a mesterséges szövegek átlagához, módosítottuk. Ekkor a simított görbe helyett az eredetileg mért függvény értékeit hasonlítottuk az átlag függvényhez. bár az így kapott különbség-függvény sokkal zajosabb, mint az általunk elemzett függvények, elemzése mégsem teljesen haszontalan. Jelen munka során nem volt célunk ennek a függvénynek a vizsgálata, de azt azért látni, részletes elemzések nélkül is, hogy a különbség-függvény ezen formája sokkal több, de ugyanakkor rövidebb intervallumra kiterjedő kiugrást tartalmaz. A kétféle ábrázolást összevetve eredményeink alkalmasak lehetnek a szövegen

belüli szóalak függő téma-réma (Petőfi és Szikszainé, 2002) kapcsolatok felfedésére.

Korábban már említettem a sokak által támogatott, de legalább annyi szakember által elutasított előfeldolgozás kérdését. Kétségtelen tény, hogy az előfeldolgozás nélkül elemzett szövegek olyan szóalakokat is újnak tekintenek, amelyek csak számban, időben, módban térnek el egymástól, de ugyanaz a szótári alakjuk, ugyanakkor éppen ezek az információk lehetnek alkalmasak arra, hogy a szóalakok számának változásával a szövegbeli változásokra utaljanak (pl. narratív szövegről párbeszédre váltás). Ezek a grammatikai jellemzők is alkalmasak lehetnek a téma-réma megkülönböztetésére.

Az előfeldolgozás nélküli szövegek elemzése során nem teszünk különbséget a grammatikai és a lexikai eszközökkel történő szóalakszám-emelkedés között, ezért ezen módszekerek alkalmazása egyszerű, mert így egységesen kezeli a kétféle megközelítést, ugyanakkor információ gazdag is.

Láttuk azonban, hogy különösen magyar nyelvű szövegek esetén, a nyelv agglutináló tulajdonsága miatt magasabb a szóalakok száma és zajosabb az újonnan megjelenő szóalakokat leíró függvény, mint angol nyelvű szövegek esetén. Ez az eltérés nem okozott nehézségeket a feldolgozás során, de ha összehasonlítjuk a magyar és az angol szövegek ábráit, akkor láthatjuk, hogy a szöveg elején a magyar szövegek nem adtak szignifikánsnak tekinthető eltéréseket. Ennek a morfológiailag gazdag nyelvek viselkedése lehet magyarázata.

Vizsgálatainknak nem volt célja valamennyi kiugrás hiánytalan megtalálása, hiszen annak bizonyítására, hogy szöveg szinten jelennek meg a kiugrások a rendelkezésre álló adatok is elegendőnek bizonyultak. Abban az esetben viszont, ha valamennyi ilyen kiugrás meghatározására szükség lenne, akkor valószínűleg érdemes lenne a szavak lemmatizálásának és szófaji tagekkel történő felcímkézésének az elvégzése is. Nem szabad azonban elfelejteni, hogy

így információt veszíthetünk el, ezért a két módszer, az előfeldolgozás nélküli és az előfeldolgozott együttes alkalmazása lehet egy lehetséges megoldás.

V.4. Szignifikancia szint meghatározása

Nem szabad elmennünk amellett a kérdés mellett, hogy mikor tekintünk egy kiugrást szignifikánsnak. Vizsgálataink során az átlagtól legalább 2σ-val való eltéréseket tekintettük szignifikánsnak. Ezt a küszöböt használva a pontok 3,9-7,5%-a került az elfogadhatósági tartományon kívülre (TOM SAWYER: 3,9%-THE JUNGLE BOOKS: 7,5%), a művek többsége esetén ez az érték 5% körüli. Első ránézésre tehát, a 2σ érték egy szerencsés választásnak tűnik. A kapott eredmények azonban azt mutatják, hogy a szignifikancia küszöb megválasztásának problémáját is érdemes lehet tovább gondolni. Láttuk a simított görbe és az átlag függvény különbségéből kapott görbéknél, valamint a fordítások összehasonlításakor, hogy a szignifikancia küszöb változtatásával hogyan nyerhetünk, illetve veszíthetünk el kiugrásokat. Vizsgálataink során megfogalmazott állítások bizonyításához megfelelt a 2σ szignifikancia küszöb, további vizsgálatokhoz, például fordítások összehasonlításához, stílus jegyek szerzőazonosításhoz, meghatározásához, stb. indokolt lehet a megváltoztatása.

Továbbgondolkodásra ad lehetőséget a szignifikancia szint meghatározása olyan szövegek esetén, mint a TOM SAWYER, ahol egyetlen nagy kiugrás miatt számos más kiugrást elveszítettünk. Ezeknek a kiugrásoknak a visszanyeréséhez megoldás lehet a szignifikancia szint csökkentése, de Platon módszeréhez hasonlóan, az is megfontolandó, hogy az ilyen jellegű szövegrészeket leíró kiugrásokat kihagyjuk a görbéből, illetve, hogy szakaszonként eltérő, lokálisan meghatározott σ-t használjunk.

Köszönetnyilvánítás

Ezúton szeretnék köszönetet mondani mindazoknak, akiknek a segítsége, megértő türelme és gondoskodása nélkül ezekre a kutatásokra nem került volna sor, és így ez a munka sem születhetett volna meg.

Hálával és köszönettel tartozom Dr. Kormos Jánosnak, az Információ Technológia Tanszék vezetőjének, Dr. Pap Gyulának, az Alkalmazott Matematikai és Valószínűségszámítási Tanszék vezetőjének, Dr. Hollósy Bélának, az Angol Nyelvészeti Tanszék vezetőjének, valamint Dr. Lajkó Károlynak, a Matematika- és számítástudományok doktori iskola Matematika – didaktika programvezetőjének.

Köszönetemet szeretném kifejezni Dr. Arató Mátyás professzor emeritusnak, témavezetőmnek munkám támogatásáért, kritikai megjegyzéseiért, tanácsaiért.

Külön köszönet illeti szerzőtársaimat, Dr. Hunyadi Lászlót és Dr. Korponayné dr. Nagy Ildikót, akik osztoztak velem az elvégzett munka minden örömében.

Végezetül hálával és köszönettel tartozom férjemnek, akire ezzel a munkával óhatatlanul együtt járó extra áldozatvállalás hárult, valamint családom többi tagjának, akik a sikertelen időszakokban is értelmet adtak minden további erőfeszítésnek.

Hivatkozások jegyzéke

Allen, V. F. (1983) Techniques in Teaching Vocabulary. Oxford University Press, Oxford, UK

Arató, M. és Knuth, E. (1970) Sztochasztikus folyamatok elemei. Tankönyvkiadó, Budapest

Ashby, W. R. (1972) Bevezetés a kibernetikába. Akadémiai Kiadó, Budapest

Aston, G. és Burnard, L. (1998) The BNC Handbook: Exploring the British National Corpus with SARA. Edinburgh, UK: Edinburgh University Press

Baayen R. H. (1993) Statistical Models for Word Frequency Distributions: A Linguistic Evaluation. *Computers and the Humanities* 26. 347-363.

Baayen, R. H. (1996a) The Randomness Assumption in Word Frequency Statistics. In Perissinotto, G. (ed), Research in Humanities Computing 5. Oxford: Oxford University Press: 17-31.

Baayen R. H. (1996b) The Effect of Lexical Specialization on the Growth Curve of the Vocabulary. *Computational Linguistics* 22: 455-480.

Baayen, R. H. (2001) Word Frequency Distributions. Kluwer Academic Publishers, Dordrecht, Netherlands

Baayen, R. H.; Halteren, H. és Tweedie F. (1996) Outside the Cave of Shadows: Using Syntactic Annotation to Enhance Authorship Attribution, *Literary and Linguistic Computing* 11: 121-131.

Balázs, J. (1985) A szöveg. Gondolat, Budapest

Barnbrook, G. (1996) Language and Computers. A Practical Introduction to the Computer Analysis of Language. Edinburgh University Press, Edinburgh

Beaugrande, R. és Dressler, W. (1981) Bevezetés a szövegnyelvészetbe. Siptár, P. (ford.) (2000) Corvina, Budapest

Biber, D.; Conrad, S. és Reppen, R. (1998) Corpus Linguistics. Investigating language structure and use. Cambridge University Press, Cambridge

Brants, T. (1997) Internal and external tagsets in part-of-speech tagging. In Proc. of Eurospeech. Rhodes, Greece

Brants, T. (2000) TnT – a statistical part-of-speech tagger. In Proceedings of the sixth conference on applied natural language processing ANLP-2000. Seattle, WA

Brunet, E. (1978) Le Vocabulaire de Jean Giraudoux, Vol. 1 of TLQ, Slatkine, Geneve

Burrows, J. (2003) Question of Authorship: Attribution and Beyond. *Computers and the Humanities* 37: 5-32.

Carroll, J. B. (1967) On Sampling from a Lognormal Model of Word Frequency Distribution. In Kucera, H. and Francis, W. N. (eds), Computational Analysis of Present-Day American English. Providence: University Press of New England.

Carter, R. és McCarthy, M. (1991) Vocabulary and Language Teaching. London: Longman Group UK

Church, K. W. és Mercer, R. L. (1994) Introduction to the Special Issue on Computational Linguistics Using Large Corpora. In Armstrong (ed.) Using Large Corpora. A Bredford Book The MIT Press Cambridge, Massachusetts London, England

Csernoch, M. (2004) Another Method to Analyze the Introduction of Word-Types in Literary Works and Textbooks. Conference Abstract, The 16th Joint International Conference of the Association for Literary and Linguistic Computing and the Association for Computers and the Humanities Göteborg University, Sweden

Csernoch, M. (2004) A szavak véletlenszerű megjelenésén alapuló modellek és az irodalmi művek közötti eltérések magyarázata. Magyar Számítógépes Nyelvészeti Konferencia, Szeged

Csernoch, M. és Hunyadi L. (2003) Szótípusok bevezetésének szabályszerűsége magyar és angol nyelvű nyomtatott szövegekben. Magyar Számítógépes Nyelvészeti Konferencia, Szeged

Csernoch, M. és Korponayné, N. I. (2005) A New Headway sorozat szókészletének számítógépes feldolgozása. MANYE, Nyíregyháza

Cunningsworth, A. (1995) Choosing your Coursebook. Oxford, UK Heinemann Publishers

de Haan, P. (1984) Problem-Oriented tagging of English Corpus Data. In Aarts and Meijs

Demetrovics, J., Denev, J. és Pavlov, R. (1985) A számítástudomány matematikai alapjai. Nemzeti Tankönyvkiadó, Budapest

Dijk, Teun van (1979) The structure and functions of discourse. Lectures at the University of Puerto Rico, Rio Piedras

Dobi, E. (2002) A pragmatika szerepe a nyelvi rendszer egységeinek leírásában. In Andor, J.; Benkes, Zs. és Bókay, A. (eds.) Szöveg az egész világ. Petőfi S. János 70. születésnapjára. Tinta, Budapest

É. Kiss, K. (1998) Mondattan. In É. Kiss, K., Kiefer, F. Siptár, P. (eds), Új magyar nyelvtan. Osiris Kiadó, Budapest

Edmundson, H. P. (1963) A Statistician's View of Linguistic Models and Language Data Processing, in Garvin, P. L. (ed.) Natural Language and the Computer, McGraw Hill, New York

Farber, B. (1991) How to Learn Any Language. Barnes & Noble Books, New York

Fellbaum, C. (ed.) (1998) WordNet. An Electronic Lexical Database. The MIT Press, Cambridge, Massachusetts. London, England

Fodor, I. (1999) A világ nyelvei. Akadémiai Kiadó, Budapest

Fodor, I. (2003) A világ nyelvei. Tinta Kiadó, Budapest

Fry, E. B.; Kress, J. E. és Fountoukidis, D. L. (2000) The Reading Teacher's Book of Lists. New Jersey, USA: Jossey-Bass Hoboken

Genette, G. (1980) Narrative Discourse. An Essay in Method. Lewin, J. E. (trans.) "Discours du recit" a portion of Figures III (1972). Cornell University Press Ithaca, New York

Gósy, M. (ed.) (1993-2003) Beszédkutatás-sorozat. MTA Nyelvtudományi Intézete

Grant, N. J. H. (1994) Making the most of your Textbook. London, UK Longman

Grefenstette, G. (1994) Explorations in Automatic Theasaurus Discovery. Kluwer Academic Publishers, Boston/Dordrecht/London

Guest, P. G. (1961) Numerical Methods of Curve Fitting. Cambridge University Press, Cambridge, UK

Guiraud, H. (1954) Les Caracteres Statistiques du Bocabulaire, Presses Universitaires de France, Paris

Hajtman, B. (1971) Bevezetés a matematikai statisztikába. Akadémiai Kiadó Budapest

Herdan, G. (1964) Quantitative Linguistics. Buttersworths, London

Holmes, D. I. (1994a) Authorship Attribution. *Computers and the Humanities* 28: 87-106.

Holmes, D. I. (1994b) Vocabulary Richness and the Book of Mormon: A Stylometric Analysis of Mormon Scripture. In Research in Humanities Computing. Hockey, S.; Ide, N.; Ross; D.; Brink, D. (eds.) Clarendon Press, Oxford

Hoover D. L. (2003) Another Perspective on Vocabulary Richness. *Computers and the Humanities* 37: 151-178.

I.B.M. (1959) Final report on computer set AN/GSQ-16 (XW-1). I.B.M. Research. Cited in Sparck Jones. 1986

Kennedy, G. (1998) An Introduction to Corpus Linguistics. Longman, London and New York

Khmaladze, E. V. (1987) The statistical analysis of large number of rare events, technical Report MS-R8804, Dept. of Mathematical Statistics, CWI. Amsterdam: Center for Mathematics and Computer Science.

Kiefer, F. (1983) Az előfeltevések elmélete. Akadémiai Kiadó, Budapest

Kiefer, F. (1998) Alaktan. In É. Kiss, K., Kiefer, F. és Siptár, P. (eds.) Új magyar nyelvtan Osiris Kiadó, Budapest

Kugler, N. (2000) Alaktan. In Balogh, J., Haader, L., Keszler, B., Kugler, N., Laczkó, K. és Lengyel, K. (eds.) Magyar grammatika. Nemzeti Tankönyvkiadó, Budapest

Laczkó, K. (2000) Alaktan. In Balogh, J., Haader, L., Keszler, B., Kugler, N., Laczkó, K. és Lengyel, K. (eds.) Magyar grammatika. Nemzeti Tankönyvkiadó, Budapest

Leech, J.; Myers, G. és Thomas, J. (eds.) (1995) Spoken English on Computer. Transcription, mark-up and application. Longman Publishing, New York

Levinson, S. C. (2000) Presumptive Meanings. The Theory of Generalized Conversational Implicature. A Bradford Book, The MIT Press, Cambridge, Massachusetts London, England

Mandelbrot, B. (1962) On the Theory of Word Frequencies and on Related Markovian Models of Discourse. In Jakobson, R. (ed), Structure of Language and its Mathematical Aspects. Providence: University Press of New England.

Markov, A. A. (1916) An Application of Statistical Method. Izvestiya Imperialisticheskoj akademii nauk, 6(4): 281-97.

McDonough, J. és Shaw, C. (1994) Materials and Methods in ELT. Blackwell, Oxford UK & Cambridge USA

Meszéna, Gy. és Ziermann, M. (1981) Valószínűség elmélet és matematikai statisztika. Közgazdasági és Jogi Könyvkiadó, Budapest

Meyer, C. (2002) English Corpus Linguistics. An Introduction. Cambridge University Press, Cambridge

MTA (1996) A magyar helyesírás szabályai. Akadémiai Kiadó, Budapest

Muller C. (1964) Calcul des Probabilités at Calcul d'un Vocabulaire. Travaux de Linguistique et de Littérature, 235-244

Nation P. és Waring R. (1997) Vocabulary size, text coverage and word list. In Schmitt N. és McCarthy M. (eds) Vocabulary: Description, acquisition, and pedagogy, Cambridge University Press, Cambridge, UK

Nemetz, T. és Kusolitsch, N. (1999) Guide to the empire of random. TypoTEX, Budapest

Oakes, M. P. (1998) Statistics for Corpus Linguistics. Edinburgh University Press

O'Grady, W.; Dobrovolsky, M. és Aronoff, M. (1993) Contemporary Linguistics, An Introduction New York: St. Martin's Press.

Olaszy, G.; Kiss, G.; Németh, G. és Olaszi, P. (2000) Profivox: a legkorszerűbb hazai beszédszintetizátor. In Gósy, M. (ed.) Beszédkutatás 2000. MTA Nyelvtudományi Intézete

Oxford, R. L. és Scarcella, R. C. (1994) Second Language Vocabulary Learning Among Adults: State of the Art in Vocabulary Instruction. System Vol.22. No. 2.

Papp, F. (1974) Szövegszó, szóalak, lexéma. Magyar Nyelvőr 98. kötet: 76-82

Petőfi, S. J. (1990) Szöveg, szövegtan, műelemzés. Országos Pedagógiai Intézet, Budapest

Petőfi, S. J. és Szikszainé, N. I. (2002) A kontrasztív szövegnyelvészet aspektusai. Linearizáció: téma-réma szerkezet. Officina Textologica 7. Debreceni Egyetem Kossuth Egyetemi Kiadója

Prószéky, G. (1989) Számítógépes nyelvészet. Számítástechnika-Alkalmazási Vállalat, Budapest

Prószéky, G. és Kis, B. (1999) Számítógéppel – emberi nyelven. Intelligens szövegkezelés számítógéppel. SZAK Kiadó, Budapest

Quirk, R.; Greenbaum, S.; Leech, G. és Svartvik, J. (1995) A Comprehensive Grammar of the English Language Longman Group UK Limited, London and New York

Ribycki, J. (2003) Burrowing into Translation: Character Idiolects in Henryk Sienkiewicz's Trilogy and its Two English Translations. Conference Abstract, The 16th Joint International Conference of the Association for Literary and Linguistic Computing and the Association for Computers and the Humanities Göteborg University, Sweden

Rocha, M. (1997) A Probabilistic Approach to Anaphora Resolution in Dialogues in English. In Ljung

Roche, E. és Schabes, Y. (eds.) (1997) Finite-State Language Processing. The MIT Press: Cambridge, Mass./London

Scarborough, J. B. (1966) Numerical Mathematical Analysis. The Johns Hopkins Press, Baltimore

Schmitt, N. (2000) Vocabulary in Language Teaching. Cambridge University Press, Cambridge, UK

Scurfield, S. (2003) Give Vocabulary Learning a Chance. TESOL-Spain Newsletter. http://www.tesol-spain.org/newsletter/scurfield.html

Shannon, C. E. (1949) The Mathematical Theory of Communication. In Shannon, C. and Weaver, W. (eds.), The Mathematical Theory of Communication, Urbana, IL, The University of Illinois Press

Sichel, H. S. (1975) On a distribution law for word frequencies. *Journal of the American Statistical Association* 70, 542-547.

Sichel, H. S. (1986). Word Frequency Distributions and Type-Token Characteristics. *Mathematical Scientist* 11: 45-72.

Silberztein (1993) Dictionnaires electroniques et analyse automatique de textes: le systeme INTEX. 240 p., Masson Ed.: Paris.

Simpson, E. H. (1949) Measurement of diversity. *Nature* 163: 168

Sinclair, J. (1995) Corpus, Concordance, Collocation. Oxford University Press, Oxford

Singleton, D. (1999) Exploring the Second Language Mental Lexicon. Cambridge: Cambridge University Press

Solt, Gy. (1971) Valószínűségszámítás. Műszaki Könyvkiadó, Budapest, Hungary

Sparck Jones, K. (1986) Synonymy and Semantic Classification. Edinburgh: Edinburgh University Press. PhD thesis delivered by University of Cambridge in 1964

Sperber, D. és Wilson, D. (1988) Relevance. Harvard Univesity Press Cambridge, Massachusetts

Sperberg-McQueen, C. M. és Burnard, L. (eds.) (1994) TEI P3: Guidelines for Electronic Text Encoding and Interchange. Chicago, Oxford

Sperberg-McQueen, C. M. és Burnard, L. (eds.) (2002). TEI P4: Guidelines for Electronic Text Encoding and Interchange. Text Encoding Initiative Consortium. XML Version: Oxford, Providence, Charlottesville, Bergen

Sproat, R. (1992) Morphology and Computation. A Bradford Book The MIT Press Cambridge, Massachusetts London, England

Tolcsvai Nagy, G. (1993) A szövegek világa. Nemzeti Tankönyvkiadó, Budapest

Tweedie, F. J. és Baayen, R. H. (1998) How Variable May a Constant be? Measures of Lexical Richness in Perspective. *Computers and the Humanities* 32: 323-352.

Uzonyi, P. (1991) Számítógépes szótárak és egyéb nyelvi adatbázisok. KLTE, Debrecen

Uzonyi, P. (1996) Rendszeres német nyelvtan. AULA Kiadó Budapest, Hungary

Váradi, T. (2004) A magyar INTEX fejlesztéséről. In Alexin, Z. & Csendes, D. (eds.) A Magyar Számítógépes Nyelvészeti Konferencia 2004 rendezvényen elhangzott előadások kötete, Szegedi Tudományegyetem Nyomdája, Szeged. p. 3-10.

Virágvölgyi, P. (1996) A tipográfia mestersége – számítógéppel. Tölgyfa Kiadó, Budapest

Wilson, A. & Thomas, J. (1997) Semantic Annotation. In Garside, Leech and McEnery ZZZ

Worthing, A. G. és Geffner, J. (1959) Treatment of Experimental Data. John Wiley and Sons, Inc., New York

Yongqi, P. G. (2003) Vocabulary Learning in a Second Language: Person, Task, Context and Strategies. *Teaching English as a Second or Foreign Language* Vol. 7. No.2

Yule, G. U. (1944). The Statistical Study of Literary Vocabulary. Cambridge: Cambridge University Press

Yule, G. U. (1950) An Introduction to the Theory of Statistics. Charles Griffin & Company Limited, London, UK

Zipf, G. K. (1935) The Psycho-Biology of Language, Boston, MA, Houghton Mifflin

Közlemények jegyzéke

Közlemények

Csernoch, M. és Hunyadi, L. (2003) Szótípusok bevezetésének szabályszerűsége magyar és angol nyelvű nyomtatott szövegekben. Magyar Számítógépes Nyelvészeti Konferencia, Szeged p 24-30.

Csernoch, M. (2004) Another Method to Analyze the Introduction of Word-Types in Literary Works and Textbooks, The 16th Joint International Conference of the Association for Literary and Linguistic Computing and the Association for Computers and the Humanities Göteborg University, Göteborg p 44-45.

Csernoch, M. és Korponayné, N. I. (2004) A New Headway sorozat szókészletének számítógépes feldolgozása. MANYE, Nyíregyháza

Csernoch, M. Dinamikusan kezelhető statisztikai modellek irodalmi művek szóalakjainak vizsgálatára. Alkalmazott Matematikai Lapok (közlésre elfogadva)

Csernoch, M. Frequency-based Dynamic Model for the Analysis of English and Hungarian Literary Works and Coursebooks. Teaching Mathematics and Computer Science (közlésre elfogadva)

Csernoch, L-né (1996) Hogyan készítik fel az egyetemek, főiskolák a tanárszakos hallgatókat az informatika, számítástechnika tantárgy tanítására. Informatika a Felsőoktatásban ’96 – Networkshop ’96, Debrecen, p 499-503.

Csernoch, M. (1997) Methodological Questions of Teachig Word Processing. 3rd International Conference on Applied Informatics, Eger-Noszvaj, p 375-382.

Csernoch, L-né (2001) Multimédia alkalmazása a gyermekkori nyelvoktatásban. A Let's Play English oktató program bemutatása. Computer Panoráma, 2001/8, lemezmelléklet

Csernoch, L-né (2003) Híd a tantárgyak között. Az informatika és az idegen-nyelvoktatás hatékony összekapcsolásának egy lehetséges módja. Mit? Kinek? Hogyan? Vezetőtanárok I. Országos Módszertani Konferenciája, Bába és Társai Kft. Szeged, p 254-264.

Csernoch, L-né (2003) Szoftver, melynek segítségével nyelvtanárok digitális oktatási segédanyagot készíthetnek az általuk használt tananyaghoz. Mit? Kinek? Hogyan? Vezetőtanárok I. Országos Módszertani Konferenciája, Bába és Társai Kft. Szeged, p 265-272.

Csernoch, M. (2004) The Accuracy of Target Group Definitions in Language Teaching Software. Novelty 11. p 65-72.

Csernoch, M. (2004) Language Games for Young Learners of English. First Central European International Multimedia and Virtual Reality Conference, Veszprém, p. 233-238.

Csernoch, M. The evaluation system of language teaching programs: a comparative analysis. Novelty (közlésre elfogadva)

Csernoch, M. Vocabulary richness, the variety of tasks and their technical support in language teaching software. Novelty (közlésre elfogadva)

Tankönyvek

Csernoch László, Csernoch Lászlóné (1998) Word 6.0 gyakorlatok I-II., Nemzeti Tankönyvkiadó, Budapest

I.3. Mi a szó? – A szó meghatározása

Annak ellenére, hogy a szó központi szerepet játszik a nyelv használatában, elemzésében egy rendkívül nehezen definiálható fogalom. Számtalan nyelvész próbálkozott már a pontos definíció megadásával, de egyetlen, minden feltételnek megfelelő definíciót mind a mai napig nem sikerült találni (összefoglaló mű: Singleton, 1999; Schmitt, 2002; Papp, 1974). Sikeresnek bizonyultak ezzel szemben azok a kísérletek, amelyek több különböző szintű szódefiníció megadásával próbálkoztak.

Ezek közül azokat szeretném csak megemlíteni, amelyek jelen tanulmány szempontjából fontosak.

I.3.1. Karakterek

A szó definiálását meg kell azonban előzze a karakter, illetve karakterkészlet definiálása, hiszen az egyetlen nyelvben sem egyezik meg, feltétlenül, az adott nyelv betűkészletével. Fontos kiemelni továbbá, hogy a számítógépes szövegfeldolgozás alapegysége mindig a karakter, és nem a betű.

A számítógépes szövegfeldolgozás során a karaktereket, ha a szó meghatározásához akarjuk őket felhasználni, két nagy csoportba szokás sorolni:

- érvényes vagy értékes, illetve
- elválasztó karakterek.

I.3.1.1. Érvényes karakterek

Az érvényes karakterek készletét az adott nyelv ábécéje alapján szokás meghatározni, rendszerint megtartva az egyjegyű betűket, míg a két- vagy többjegyű betűket felbontva megfelelő számú karakterre. Az így meghatározott

alap karakterkészlet tovább bővíthető/bővítendő ha a feldolgozott szövegek nem egynyelvűek, például, ha angol nyelvű szöveg francia vagy német nyelvű idézeteket tartalmaz. Angol nyelvű szövegek esetén megfontolandó az aposztróf érvényes karakterként való kezelése is.

I.3.1.2. Elválasztó karakterek

Első számú elválasztó karakter a szóköz, amely alapértelmezésben elválasztja egymástól a szöveg egymást követő szóalakjait. Minden más, a szövegekben előforduló karakter érvényes vagy elválasztó karakterként való kezelése a feladat függvénye és a felhasználó felelősége ezek eldöntése. Elválasztó karakter lehet valamennyi számjegy, az írásjelek, a nyelvben idegen betűk, műveleti jelek, pénzegységek jele, stb.

I.3.2. Szó definíciók (zárójelben az angol szakirodalomban elterjedt elnevezések)

Szövegszó (token, running word): két elválasztó karakter közötti karaktersorozat.
Szóalak (word-type): egyedi szövegszó.
Szótári alak, lexéma (lexeme): a szónak az a formája, amely szótári cím vagy alcímként jelenik meg.
Szótő (lemma): a szónak az inflexiók eltávolítása után kapott formája. (Igen magas azoknak a szavaknak a száma, ahol a szónak a szótári alakja és a szótöve megegyezik.)

I.3.2.1. Példák különböző szóhasználatokra

Mind magyar, mind angol szövegekben használhatjuk ugyanazokat a szódefiníciókat.

I.3.2.1.1. Szövegszó

A *Ment, ment mendegélt* vagy a *Tomorrow, and tomorrow, and tomorrow* kifejezések három (*ment, ment, mendegélt*), illetve öt (*tomorrow, and, tomorrow, and, tomorrow*) szót tartalmaznak, ha a szövegszók számát vesszük.

I.3.2.1.2. Szóalak

Ha azonban a különböző szóalakok számát nézzük, akkor kettőre (*ment, mendegélt*, illetve *tomorrow, and*) jön ki a szavak száma mindkét esetben. Hasonló módon, a *Going, going, gone* kifejezés három szót jelent a szövegszók szintjén, de csak kettőt a szóalakok szintjén.

I.3.2.1.3. Szótő

Egy további szintet jelent a szavak lemmatizálásával kapható alak, a szótő. Ha újra a *Going, going, gone* kifejezést tekintjük, akkor egyetlen, a *go* szóval, mint szótári alakkal helyettesíthető mind a *going*, mind a *gone*.

I.3.2.1.4. Szótári alak

Az azonban, hogy a szó milyen alakban kerül a szótárba nagyban függ a szótár felépítésétől, szerkezetétől, céljaitól, ezért a szótári alak definíciója, ha lehetséges, egy még inkább összetett kérdés. A probléma összetettségét egy példán keresztül szeretném bemutatni, az *-ing* végződésű angol szavakat vizsgálva. Nemcsak igei, hanem névszói, melléknévi és határozói szerepben is megjelenhet egy *-ing* végződésű szó. A betöltött szereptől függően indokolt lehet az *-ing* végződésű szó szótőre redukálása, de nem minden esetben helyes. Például a *talking donkey* kifejezésben a *talking* melléknév leredukálása a *talk* igére szóveszteséget jelenthet, amely azzal a következménnyel járna, hogy egy szövegben megnőne az igék száma, míg a melléknevek száma csökkenne.

I.4. Számítógépes szótárak – szöveges adatbázisok

Számítógépes szótárnak nevezzük azt a képződményt, amely valamilyen lexikális elemeknek (szavaknak, szókapcsolatoknak stb.) gépileg tárolt halmaza (Uzonyi, 1991).

A szótár elnevezés esetleg zavaró lehet, ha a hagyományos értelemben vett szótárakra gondolunk, itt azonban a szótár adathalmaz jelentésben jelenik meg. A számítógépes szótárak tehát adatbázisok, amelyekben a felhasználó igénye szerint (természetesen a szótár korlátain belül) információkat kereshet vissza. Attól függően, hogy a ki/mi a felhasználó a szótárakat további kisebb csoportokba sorolhatjuk. Az első szempont, amely szerint meg szokás különböztetni a számítógépes szótárakat az az, hogy a felhasználó az ember vagy egy számítógép(es program).

I.4.1. Nem-gépi szövegfeldolgozásra létrehozott szótárak

A hagyományos értelemben vett szótár fogalom alatt mára a fordítást, megértést segítő szótárakat értjük, és nem-gépi szövegfeldolgozásra létrehozott szótáraknak is szokás azokat nevezni. A szótár felhasználója tehát nem egy számítógépen futó program, hanem az ember. Abban az esetben is ebbe a kategóriába soroljuk a szótárat, ha annak elektronikus formáját használjuk, tehát nem kizárólagosan a nyomtatott formában megjelenő szótárakat értjük ide. A keresés módjában van csak eltérés, de a keresést minden esetben az ember indítványozza.

I.4.2. Gépi szövegfeldolgozásra létrehozott szótárak

A gépi szövegfeldolgozásra létrehozott szótárak számítógépes programok számára olvasható inputot, illetve ezen programok által előállított outputot jelentenek, amely formátumok az ember számára nem vagy csak nehézkesen olvashatók. Továbbra is igaz azonban, hogy az ilyen típusú adatbázisok felhasználásánál is a végtermék, a szótárból nyert információ emberi kezdeményezésre jön létre és az ember számára ad információt.

Ahhoz, hogy egy számítógépes program írni és olvasni tudja ezeket a szótárakat un. machine/computer-readable formátumba kell hozni az adatokat. Számos kisebb-nagyobb adatbázis látott mára napvilágot. Ezek között megtalálhatóak az egészen speciális célokat kielégítő kisebb adatbázisok és a nagyobb, több-százmillió-szó nagyságrendű általános célú adatbázisok is. A speciális igényeket kielégítő gyűjtemények rendszerint egyedi formátummal rendelkeznek, az adott rendszerben a feladatnak leginkább megfelelő formátumot igyekeznek megtalálni. Nagy méretű adatbázisoknál azonban célszerű lenne egy egységes formátumot találni.

I.4.3. Digitális szövegek kódolása

Napjainkra már megindult egy folyamat, amelynek célja, hogy a különböző típusú rendszerek által létrehozott adatbázisok mások, más programok számára is elérhetőek legyenek. Az egységesítési törekvés folyamatában az egyik legjelentősebb vállalkozás a Text Encoding Initiative[14] (TEI) projekt. A program 1987-ben indult és fő célja az volt, hogy létrehozzon egy olyan nemzetközi, több tudományterületet is érintő sztenderdet irodalmi és nyelvészeti szövegek online kutatásaihoz, tanításához segédeszközként, amelyet fel tudnak használni könyvtárak, múzeumok, kiadók, kutatók, stb. A rendszer létrehozásakor egy könnyen érthető és minimálisan elavuló kódrendszer

[14] http://www.tei-c.org

használatát javasolták (Sperberg-McQueen és Burnard, 1994; 2002), amely így alkalmassá teszi a szövegek feldolgozását nemcsak számítástechnikai szakemberek számára, és nem kell tartani a viszonylag gyors elévüléstől és az ebből adódó nehézségekből. Ezen szempontokat szem előtt tartva a Standard Generalized Markup Language-et (SGML) választották a kódolásra.

I.5. Szöveg korpuszok

A szöveges adatbázisok, szótárak létrehozásához szükség van az írott szövegek elektronikus verziójára. A korpusznyelvészet az, amely ezen korpuszok létrehozásával és feldolgozásával foglalkozik. A korpusznyelvészet mára külön tudományággá fejlődött számos résztudományágat magában foglalva.

A számítógépen tárolható korpuszok méretének és az őket feldolgozó programok sebességének köszönhetően számos, főleg irodalmi és nyelvészeti kutatásban új irányzat és eredmény született, amelyek korábban nem tűntek elérhetőnek, lehetségesnek.

Az első és legnyilvánvalóbb változás, hogy átalakult a szótárkészítések módja – a nem-gépi feldolgozásra létrehozott szótárak esetén is –, melynek eredményeként a napjainkban megjelenő akár elektronikus, akár nyomtatott szótárak korpusz alapúak. A nem-gépi feldolgozásra szánt szótárak esetén mind az egynyelvű, mind a kétnyelvű szótárakban megjelentek a magyarázatok mellett a korpuszból nyert példamondatok, amelyek nagyban segíthetik a szótárhasználót. Az elektronikus formában megjelenő nem-gépi feldolgozásra szánt szótáraknak további nagy előnye, hogy szakítani tudtak a nyomtatott szótárak lineáris felépítésével és hypertext szervezésűek.

A mindenkori nyelvhasználattal, a nyelv változásával, alakulásával foglalkozó nyelvészeti kutatásoknak is újabb horizontot nyitott az elektronikus korpuszok megjelenése. Ide tartozik, egyebek között, a speciális nyelvtani szerkezetek vizsgálata, a referencia nyelvtanok létrehozása, a nyelvi különbözőségek (language variation) tanulmányozása, a történelmi nyelvészet, a fordítás (elmélet), a nyelvtanulás, és a nyelvpedagógia (Meyer, 2002). Ez a felsorolás is mutatja, hogy az elektronikus korpuszok rendkívül széles körben felhasználhatók, mind elméleti, mind gyakorlati kutatásokban, mivel a korpusznyelvészet egy módszer, amelyet bármely szöveggel foglalkozó tudományterület alkalmazhat.

A korpuszok azonban a szövegek nyers formájában nem minden felhasználó számára adnak használható információt, ezért a korpusznyelvészetnek nemcsak a korpuszok létrehozásának nem kis feladata, hanem a korpuszok feldolgozása is része.

I.5.1. Korpuszok létrehozása

A korpuszok létrehozása egy rendkívül összetett feladat, amelynek főbb összetevői

- a tervezés,
- a digitalizálás,
- a szöveg, szövegrészek annotálása.

(Sinclair, 1995; Barnbrook, 1996; Biber et al., 1998; Kennedy, 1998; Meyer, 2002):

I.5.1.1. Korpuszok tervezése

- A várható felhasználók körének meghatározása.

- A korpusz méretének meghatározása (rendelkezésre álló idő, az anyagi lehetőségek, személyi feltételek, számítógépes erőforrások).
- Azoknak a szövegtípusoknak meghatározása, amelyekből válogatunk (írott és beszélt nyelvi szövegek és ezen két nagy kategórián belül további kisebb kategóriák meghatározása).
- A szövegdarabok hosszának meghatározása.
- A szövegek számának és a beszélők/írók kiválasztása (mennyire tudják reprezentálni a kiválasztott zsánert).
- Időkeretek meghatározása, amelyből kiválasztjuk a beszélőt/írót.
- Anyanyelvi, nem-anyanyelvi beszélők arányának meghatározása.
- Szociolingvisztikai szempontok figyelembe vétele (nem, életkor, iskolázottság, tájnyelvi eltérések, szociális szövegkörnyezet, szociális kapcsolatok).

(Biber, 1993; Meyer 2002)

I.5.1.2. Elektronikus verzió létrehozása – digitalizálás

Szövegek számítógépes, számítógéppel segített feldolgozásához természetesen szükség van a szövegek digitalizált verziójára. Az elektronikus szövegeket eredetük szerint két csoportba szokás sorolni:

- számítógépen előállított dokumentumok,
- nyomtatott formában keletkezett dokumentumok, amelyek digitalizálása egy későbbi folyamat eredménye.

Mindkét eredetű szöveg esetén nagyban befolyásolhatja a kutatás során feldolgozásra került szövegeket, hogy melyek azok, amelyek elérhetőek és nem védik szerzői jogok a kiválasztott műveket, illetve vannak-e anyagi korlátok, amelyek megakadályozhatják a szövegekhez való hozzáférést.

I.5.1.3. Elektronikus formában megtalálható szövegek

- elektronikus könyvtárak,
- CD-ROM-on megvásárolható szövegek,
- szövegszerkesztőkkel generált, napjainkban publikált szövegek.

I.5.1.4. Nyomtatott szöveg digitalizálása

Az eredeti nyomtatott szöveg típusa, minősége, a rendelkezésre álló hardver és szoftver eszközök határozzák meg, hogy a hagyományosnak számító gépelést vagy inkább a szkennelést választjuk a szövegek digitalizálásához.

I.5.1.4.1. Gépelés

A gépelés az a módszer, amely bármikor használható. Az egyik nagyon komoly probléma a gépeléssel, hogy rendkívül időigényes. Nagyon nagy sebességgel és pontossággal kell gépelni ahhoz, hogy belátható időn belül megfelelő mennyiségű és minőségű szöveghez jussunk. Így, szinte kizárólagosan profi gépírókkal szokás ezt a munkát végeztetni, akik nagyon pontosan, szinte hiba nélkül végzik a szövegbevitelt. Ez a megoldás ugyanakkor rendkívül költséges, mert nagyon magas a leütések száma, ami után a gépírók kiszámítják költségeiket.

I.5.1.4.2. Szkennelés

A szkennelés kiválthatja a gépelést, ha a szöveg minősége ezt lehetővé teszi. Ahhoz viszont, hogy a szkennelési munkát meg tudjuk kezdeni szükség van egy komolyabb anyagi befektetésre. Meg kell vásárolni a szkennert és egy karakter felismerő programot is (OCR, Optical Character Recognition). A szkennelés folyamata is időigényes, mivel egy oldal bevitele nem kevesebb, mint egy-másfél perc, ezt követi a felismerésre szánt zónák kijelölése, amelyre

fordítandó idő a szöveg minőségétől nagyban függ (ez a folyamat újabb perceket jelenthet, ha sok a kép, a táblázat, esetleg váltakozó a hasábok száma az adott oldalon), majd ezt követi a tényleges karakter felismerés. Azt, hogy a karakterek felismerése mennyire pontos még további apróságnak tűnő, ugyanakkor meghatározó fontosságú tényezők befolyásolják. Ilyenek lehetnek például a háttérszínek, a font típusa, mérete, a kiemelésekre használt módszerek. A szkennelés tehát kevésbé időigényes, de nagyobb hibaszázalékkal dolgozik, mint a gépelők. Ennek következtében a szkenneléssel történt szövegbevitelt minden esetben egy nagyon alapos ellenőrzés (proof-reading) kell, hogy kövesse, ami szintén egy időigényes folyamat (Barnbrook, 1996; Prószéky és Kis, 1999; Meyer, 2002).

A nyomtatott szövegek digitalizálásának röviden ismertetett folyamata még nem meríti ki a papíron megjelent dokumentumok teljes körét, ugyanis nem szabad megfeledkeznünk a kézzel írott dokumentumok felhasználásáról sem, amely további, az előzőeknél még nehezebben megoldható problémákat vet föl.

I.5.2. Hang korpuszok

Mivel a dolgozatnak nem témája és a későbbiekben is csak marginálisan kerül említésre a beszélt nyelvi szövegek használata, így csak röviden térnék ki a problémára.

Az, hogy az írott szövegek mellett beszélt nyelvi szövegek feldolgozására is szükség van ez nem kérdéses. Megnehezíti azonban a problémát, hogy míg az írott nyelvi szöveg egy maradandó képződmény, addig a beszélt nyelvi nem (Leech et al. 1995). A beszéd maradandóvá tételéhez szükséges eszközök között még a legegyszerűbbek is lényegesen bonyolultabbak, mint a papír és az írógép. További probléma, a spontán beszéd rögzítése, hiszen azzal, hogy a beszélővel tudatjuk, hogy rögzítjük az elhangzottakat megszűnik a beszéde spontán lenni és

újra csak ál-szöveget sikerül felvenni (Sinclair, 1995; Meyer, 2002), a titokban történő hangfelvétel pedig napjainkban nem megengedett, hiszen a szereplők személyiségi jogait sértheti.

Beszélt nyelvi szövegek megőrzésében is a számítógépek és azok fejlődése hozott lényeges változást. Mára már nemcsak az az elsődleges cél, hogy felvételeket tudjunk készíteni beszélt nyelvi szövegekből, hanem írott szövegekhez hasonlóan, ezek is további feldolgozásra kerülnek.

Egyik nagyon gyakori felhasználási terület, hogy a beszélt nyelvi, esetleg ál-beszélt nyelvi (film, dráma, stb.) szövegekkel egészítik ki a szövegkorpuszokat. Ez a folyamat maga után vonta/vonja a beszélt nyelvi szövegfelismerés, -kódolás tudományának kialakulását, fejlődését, amelynek napjainkra számos további felhasználása is napvilágot látott.

A beszédfelismerés napjainkra csak egyik részterülete a beszélt nyelvi szövegek feldolgozásával foglalkozó tudományoknak, ugyanis sikerült jól működő beszédszintetizátorokat is előállítani, amelyek képesek valamilyen bemenet alapján hangot előállítani (Olaszy et al., 2000).

Mindkét irányú folyamat számos „ipari” alkalmazása is folyamatban van. (Ipari alkalmazás alatt értek minden olyan felhasználást, amely nemcsak tudományos célok érdekében jön (jött) létre, például oktatás, valamilyen fogyatékkal élők számára használható alkalmazások, mobil telefonos kommunikáció stb.)

I.5.3. Korpusz annotálása

Ahhoz, hogy egy korpusz a lehetséges felhasználók számára használható legyen annak nyers formája nem minden esetben elegendő, a korpusz annotálására van szükség. Az annotálás folyamata azt jelenti, hogy a korpusz

meghatározott egységeit tagekkel (jelekkel) látjuk el. Több különböző típusú annotálásról szokás beszélni:

- structural

  A szöveg, mint teljes egészre vonatkozó információk meghatározása (pl. bibliográfia, etnográfiai jellemzők), bekezdéshatárok megjelölése, egymást átfedő szegmensek jelölése),

- part-of-speech

  Ezt a típusú annotálást tagger (jelölő) programokkal szokás végezni, melyek feladata szavak szófajának meghatározás (Prószéky, 1989; Sproat, 1992; Church és Mercer, 1994; Biber et al., 1998; Kennedy, 1998; Prószéky és Kis, 1999; Meyer, 2002;). Napjainkra a tagger programok többsége, gyakorlatilag tetszőleges input fájl esetén, legalább 95%-os pontossággal dolgozik elfogadható idő és hely ráfordítással.

- grammatical

  Ezt a típusú annotálást parser (elemző) programokkal szokás végezni, melyek feladata szónál nagyobb egységek (kifejezések, mondatok) címkézése (Prószéky, 1989; Sproat, 1992; Biber et al., 1998; Kennedy, 1998; Prószéky és Kis, 1999; Meyer, 2002;).

- semantic tagging (Wilson és Thomas, 1997)
- discourse tagging (Rocha,1997)
- problem oriented tagging (De Haan, 1984)

I.5.4. Lemmatizálás (szótövezés)

A lemmatizálás az a folyamat, amely során a szóalakokról eltávolítjuk az inflexiókat – ragokat, képzőket – és végeredményként a szótőt kapjuk.

A lemmatizálásnak számos különböző fázisa és számos különböző kimenete is elképzelhető. Ennek megfelelően ismertek olyan eredmények,

amelyek csak a szótőt adják vissza, míg mások megadják a feltételezett szótőt és hozzá a lehetséges szófajokat, míg megint mások felvállalják a szófaj meghatározását is, tehát az annotálást. Azt azonban nem szabad elfelejtenünk, hogy a programok még nem 100%-os pontossággal dolgoznak, így az elsődleges gépi feldolgozás után még manuális kiegészítésre van szükség, ahhoz, hogy megbízható eredményeket kapjunk.

Mindig az adott célkitűzések szabják meg a felhasznált módszereket. Ennek megfelelően sem az az állítás, hogy minden szöveget lemmatizálni kell, sem az, hogy lemmatizálás nélkül is megmagyarázható minden természetes nyelven íródott szöveggel kapcsolatos probléma nem fogadható el. Mindkét kijelentés szélsőséges álláspontnak tekinthető.

A lemmatizálás eredményeként kaphatunk a szövegről olyan információkat, amelyeket az eredeti szóalakok, a képzők és ragok hatványozott használata miatt, nem adhatnak vissza. Különösen igaz ez, ha különböző nyelveken írt szövegeket hasonlítunk össze. Ebben az esetben ugyanis a különböző nyelveken írt szövegben a szóalakok száma lényegesen eltérő is lehet. Ezzel szemben a lemmatizálás mélységétől függően elveszíthetünk olyan információkat, melyeket a ragok és a képzők hordoznak magukban.

A magyar fejlesztésű *SZÓSZABLYA – HUNSTEM*[15] szótövező program mintaszövegén mutatom be a lemmatizálási folyamat egy lehetséges eredményét (1.1. táblázat).

1.1. táblázat. *SZÓSZABLYA – HUNSTEM* szótövező outputja három rövid példamondat alapján (honlapról letöltve). A minta is mutatja, hogy ilyen output esetén további kézi feldolgozásra van szükség, mivel nincs egyértelműen megadva a végeredmény, és a szófaji meghatározások teljesen hiányoznak.

Alaposan felöntött a garatra.
Budapesti műtrágya-gyártósorok

[15] http://szoszablya.sourceforge.net/hunstem.html

összeszerelésével foglalkozott Bruckner.
A Hunstem futtatása és kimenete az állományon:
$ hunstem példa.txt

```
példa.txt               alaposan
példa.txt               alapos

példa.txt               felöntött
példa.txt               felönt

példa.txt               a

példa.txt               garatra
példa.txt               garat

példa.txt               budapesti
példa.txt               Budapest

példa.txt               műtrágya
példa.txt               gyártósorok
példa.txt               sor
példa.txt               gyártósor
példa.txt               gyártó
példa.txt               gyár
példa.txt               tó

példa.txt               összeszerelésével
példa.txt               összeszerelés
példa.txt               összeszerel

példa.txt               foglalkozott
példa.txt               foglalkozik

példa.txt               # Bruckner
```

I.6. Gépi fordítás

A MorphoLogic[16] cég által készített programokról szeretnék említést tenni, mivel véleményem szerint ma Magyarországon ez a cég végzi a legnagyobb fejlesztéseket a számítógépes, számítógéppel segített fordítás területén több más szervezettel együtt működve, ahol mindenképpen megemlíteném a Magyar Tudományos Akadémia Nyelvtudományi Intézetét.

[16] http://www.morphologic.hu

Az itt következő információk a MorphoLogic honlapjáról kerültek letöltésre.

I.6.1. *METAMORPHO* programcsalád

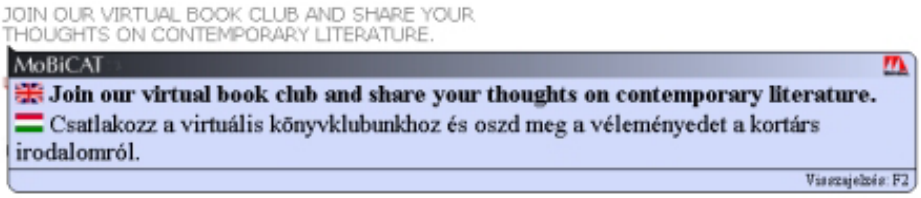
Legígéretesebbnek a MorphoLogic cég által fejlesztett *METAMORPHO* fordítást segítő, három programból álló programcsaládja tűnik. A kezdőknek megértés-támogatási eszközök, a haladóknak fordítóprogram, a hivatásos fordítóknak pedig intelligens fordítómemória támogatás készül.

A MorphoLogic fordítóprogramjai egyesítik a példa-alapú és a szabály-alapú számítógépes fordítók előnyeit, létrehozva ezzel az un. minta-alapú számítógépes fordító rendszert.

A *METAMORPHO* fordítóprogramok az elérhető legjobb nyelvi minőség érdekében mindig két nyelv között, és egy adott nyelvi irányban működnek. Elsőként a legnagyobb érdeklődésre számot tartó angolról magyarra fordítás valósult meg, de a magyar-angol fejlesztések is megkezdődtek. A család első tagja a *MOBICAT* angol-magyar megértés-támogató fordítóprogram 2004 júniusában jelent meg.

I.6.1.1. MoBiCAT

A *MOBICAT* (MorphoLogic Bilingual Computer Assisted Translation) használata igen egyszerű: elég az egérmutatót a szöveg fölé húzni, majd néhány másodperc múlva az adott mondat magyar megfelelője automatikusan megjelenik a képernyőn. A világ első megértés-támogató rendszerétől nem várható el minden angol mondat szabatos és helyes magyar fordítása, ám segítségével a szöveg értelmezhetővé válik, és az információtartalom kinyerhető belőle.



1.1. ábra. Példa mondat a *MOBICAT* működésére. A *MOBICAT* a Microsoft Internet Explorer, a Microsoft Word, Microsoft Outlook és az Adobe Acrobat programokban használható. A program együttműködik a számítógépre telepített *MOBIMOUSE* szótárprogrammal is, melynek eredményét külön buborékban jeleníti meg (http://www.morphologic.hu).

„A *MOBICAT* technológia – bár nagyságrendekkel fejlettebb, mint az eddigi angol-magyar fordítóprogramok, és a többi professzionális gépi fordítórendszerrel is versenyre kelhet – nem készít tökéletes fordításokat. A *MOBICAT* mondatokat elemez, azokat fordítja le, de nem képes szövegösszefüggések felderítésére. Csak korlátozottan képes a nyelvtanilag helytelen mondatok értelmezésére, és a jelentésbeli többértelműségek feloldásában is egyelőre messze az emberi fordítás mögött jár.” (http://www.morphologic.hu)

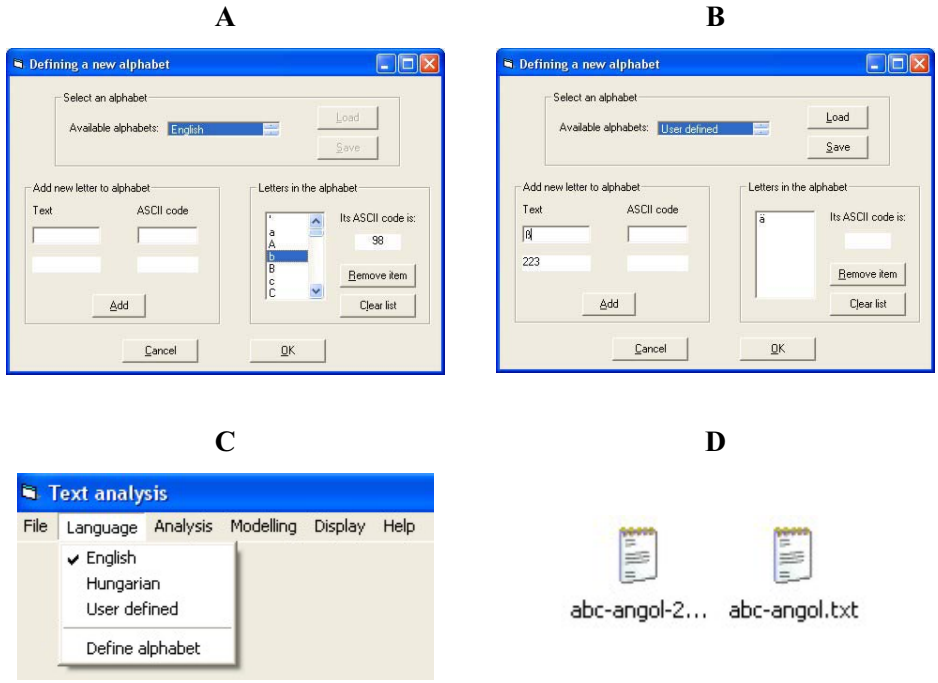
I.6.1.2. morphoWAP

A *MORPHOWAP* a MorphoLogic *METAMORPHO* alapú, mobiltelefonokról elérhető ingyenes WAP-os szolgáltatása. A fordítás szolgáltatóktól függetlenül mindenkinek rendelkezésre áll.

1.2. ábra. Példa a *MORPHOWAP* működésére (http://www.morphologic.hu).

I.6.1.3. METAMORPHO: élő szolgáltatás

„A www.metamorpho.hu szerveren működtetett nyelvi adatbázist az Önök közreműködésével rövid időn belül mérvadó angol-magyar fordítástámogatói oldallá szeretnénk fejleszteni. Javaslataik először még közvetve, hamarosan azonban közvetlenül is beépülhetnek a program nyelvi tudásába, amit a METAMORPHO legfontosabb erénye, könnyű fejleszthetősége tesz lehetővé. Az adatbázist észrevételeik alapján lehetőleg még aznap, de legkésőbb 5 munkanapon belül bővítjük vagy javítjuk. Ez indokolja az internetes kliens-szerver architektúrát is, hiszen így a fejlesztések mindenkinél azonnal érvényesülni tudnak. A programok mindegyike egyszerű és hatékony visszajelzési lehetőségeket kínál.” (http://www.morphologic.hu)



3.1. ábra. Részletek a *DyMoCASAT* karakterkészleteket definiáló párbeszédablakáról (A,B) és menüjéből (C), valamint két minta fájl (D), amelyekben a felhasználó által definiált karakterkészletek vannak tárolva. Ezek a fájlok egyszerű szöveges dokumentumok, amelyek a szöveg feldolgozásához használni kívánt összes karaktert tartalmazzák. Egy User defined karakter készlethez használatához azt előbb be kell tölteni (C, Define alphabet, majd B, Load parancsa), majd kiválasztani, mint aktuális ábécé (C, Language menü User defined).

Az ’a’ karakterrel kezdődő szavak elrendezése

	1	**2**	**3**	**…**	**n**
1	x_{a11}	x_{a12}	x_{a13}		x_{a1n}
2	x_{a21}				
3					
…					
m_a					

A ’b’ karakterrel kezdődő szavak elrendezése

	1	**2**	**3**	**…**	**n**
1	x_{b11}	x_{b12}	x_{b13}		x_{b1n}
2	x_{b21}				
3					
…					
m_b					

3.2. ábra Az ’a’ és ’b’ karakterrel kezdődő szavak elrendezése. A fájlok első bekezdése (a táblázat első sora, ahol a második index 1) az ASCII kódok alapján a legelső ’a’-val, illetve ’b’-vel kezdődő szavakat tartalmazzák, míg az utolsó bekezdések (a táblázat utolsó sora) ezen elrendezés szerinti utolsó szavakat. Az egyes fájlokon belüli bekezdések száma változó, tehát m_a várhatóan nem egyenlő m_b-vel.

Az ’a’ karakterrel kezdődő szavak elrendezése értékes jegyekkel

	1	**2**	**3**	**…**	**n**
1	2	1	0	0	1
2	0	0	1	0	0
3	1	0	0	0	2
…					
$\mathbf{m_a}$	0	1	0	0	0

A ’b’ karakterrel kezdődő szavak elrendezése értékes jegyekkel

	1	**2**	**3**	**…**	**n**
1	0	0	1	1	0
2	2	1	1	2	2
3	0	0	0	0	1
…					
$\mathbf{m_b}$	0	0	1	0	0

3.3. ábra. A szavak előfordulását tároló háromdimenziós tömb ’a’ és ’b’ kétdimenziós lapjai értékes jegyekkel feltöltve egy lehetséges minta alapján.

Az ’a’ karakterrel kezdődő szavak első megjelenése

	1	**2**	**3**	**…**	**n**
1	T				
2			T		
3	T				
…					
$\mathbf{m_a}$		T			

A ’b’ karakterrel kezdődő szavak első megjelenése

	1	**2**	**3**	**…**	**n**
1			T		
2	T				
3					T
…					
$\mathbf{m_b}$			T		

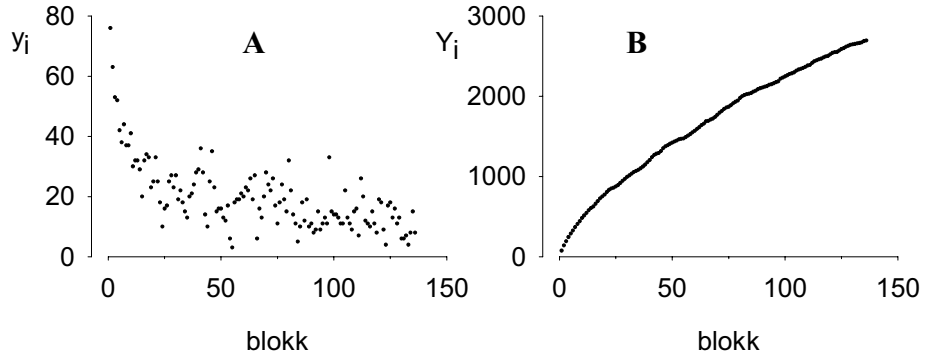
3.4. ábra. Az egyes blokkokban az újonnan megjelenő szóalakok (y_i) megszámlálásához az egyes szavak első előfordulását kell megtalálnunk és az így kapott pozíciók alapján meghatározhatóak ezen értékek.

'em
@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@
@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@B@@
@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@
@@A
'tis
@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@
@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@
@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@
@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@
@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@C@@@@
@@@@@@@@@@@@@@@@@@@@@@@@@@@@A

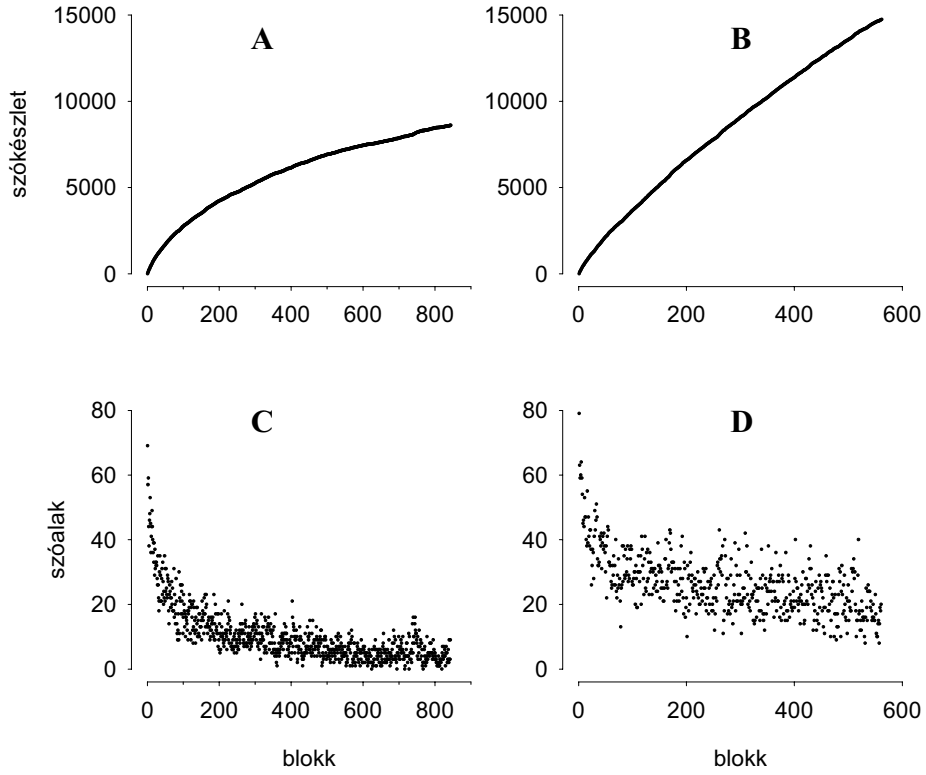
3.5. ábra. Lewis Carroll: ALICE'S ADVENTURES IN WONDERLAND aposztróffal kezdődő szavai. Az első értékes karakter annyiadik blokkban jelenik meg a bekezdésben, amelyikben a program először megtalálta az adott szót. Azokat a blokkokat, amelyekben nem szerepelt az adott szó az @ karakterrel jelöli a program.

western
@AA
westward
@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@
@@@@@@@@@@@@@@@@@@@@@@@@@@@@A
whar
@@@@@@@@@@@@@@@@@@@A
wharf
@@@@@@@@@@@@@@@@@@@@@@@@@@@A
what
B@@@A@@@A@AA@@@@@@BBBBDAB@@DC@BAA@@@@@@A@@A
AB@@ABAB@DAA@@@@@@@@@@@@A@A@@@AA@@@@@@@@@
@@@@@A@@@@@@@@@A@@@@@@AA@@@@@@@@A@@@@@@
@@A@@@AAA@@@A
when
@@A@@A@@@@AA@@@A@@A@@A@@@@@@@A@@@A@@A@A@A
@B@@@@@@@@@@@@@@@@@@A@A@A@@B@A@@@A@@@@@@
BBB@AAB@@AA@AB@AA@@@@@@@A@@@@@@@@@@@A@@AA@
AA@@@@@A@A

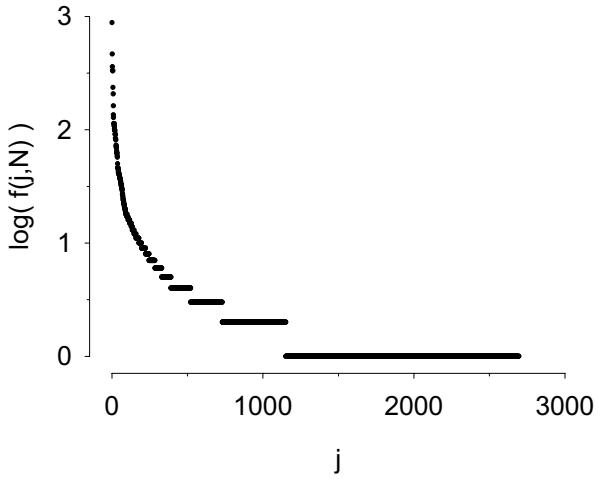
3.6. ábra. Edgar Allan Poe: THE GOLD-BUG ’w’-vel kezdődő szavaiból részlet. A *western* szó a mű második és harmadik blokkjában szerepel egy-egy alkalommal, a *what* már az első blokkban megjelenik és kétszer is előfordul, majd legközelebb az ötödikben, stb.



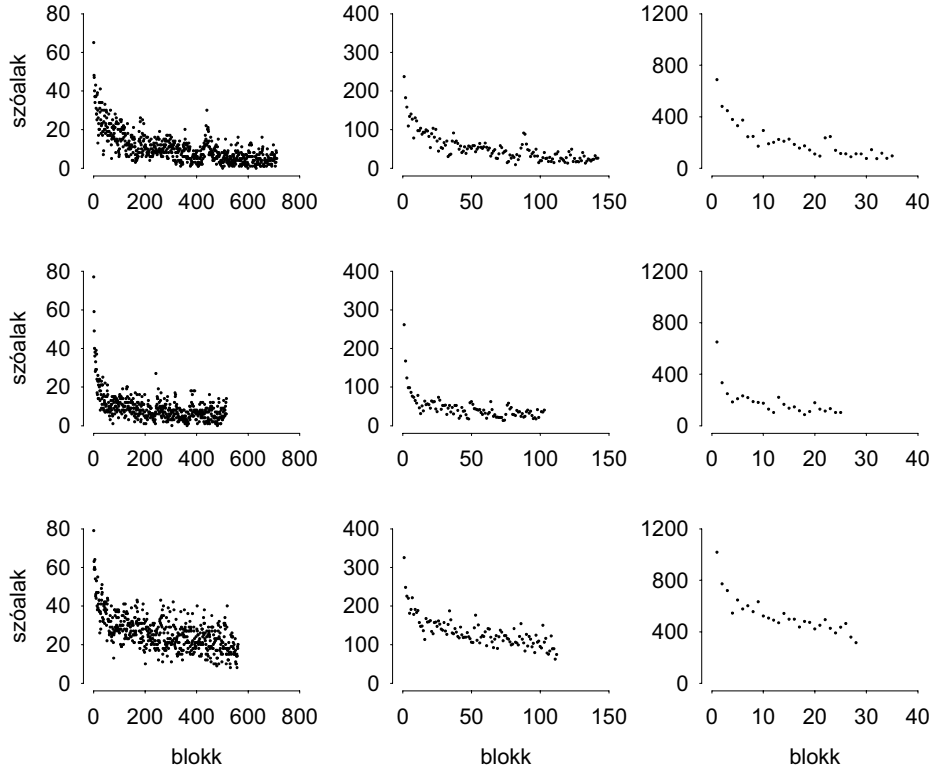
3.7. ábra. Edgar Allan Poe: THE GOLD-BUG. A műben megjelenő különböző szóalakok száma száz-szavas blokkokra bontás esetén. Az első blokkhoz tartozó érték megadja, hogy hány különböző szóalak található a műnek ebben az intervallumában. Minden más blokkhoz tartozó érték azt mutatja, hogy az azt megelőző blokkokhoz képest hány új szóalak jelent meg (A). A szóalakok számának összegzésének eredménye egy monoton növekvő függvénnyel ábrázolható (B). Az első blokkhoz tartozó függvényérték megegyezik az (A) függvény függvényértékével ebben a pontban, minden egyes további érték az azt megelőző függvényértékek összege.



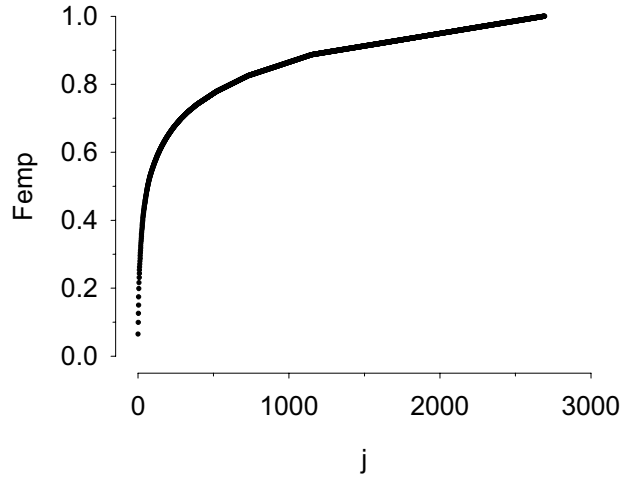
3.8. ábra. Szóalakok megjelenése angol (Hawthorne: THE SCARLET LETTER, A) és magyar (Kertész Imre: SORSTALANSÁG, B) nyelvű szépirodalmi művekben. Az alsó ábrák (C és D) az újonnan bevezetett szóalakok számát mutatják az egyes blokkokban az A és a B ábrán mutatott művekhez. Megfigyelhető ezeken az ábrákon, hogy a magyar nyelvű szövegben a különböző szóalakok száma és a szóalakok megjelenésének zaja lényeges nagyobb, mint egy hasonló hosszúságú angol szövegben.



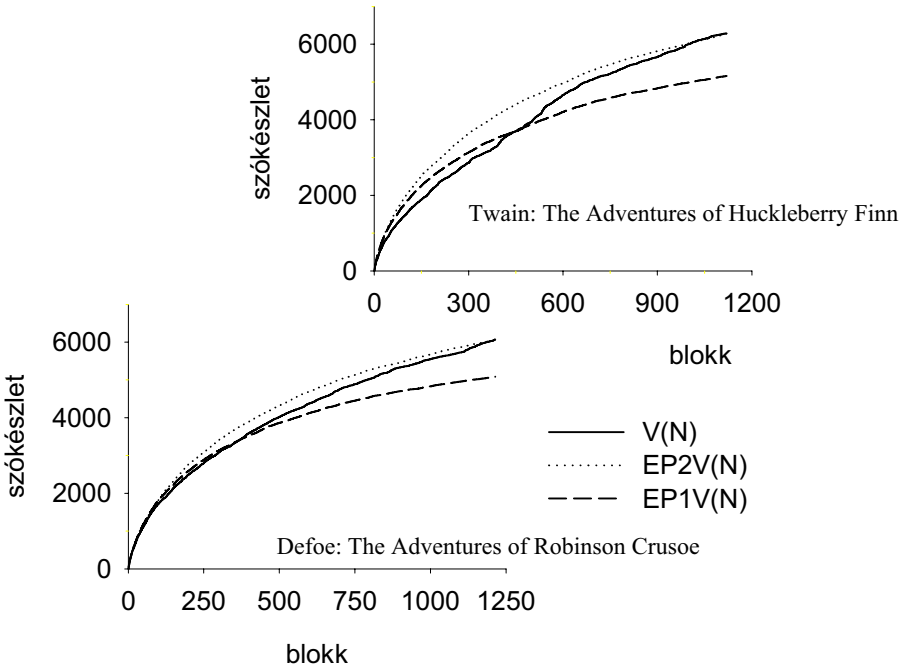
3.9. ábra Szavak gyakorisági görbéje Edgar Allan Poe THE GOLD-BUG című művében. A görbe vízszintes szakaszai jobbról balra haladva az egyszer, kétszer, stb. előforduló szavak számát mutatják.



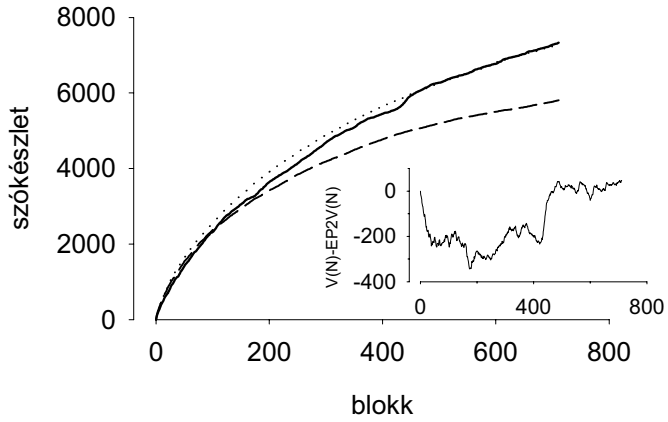
3.10. ábra. Az újonnan bevezetett szóalakok számának módosulása az egyes blokkokban a blokkok hosszának változtatásával. A különböző szóalakok számának meghatározása, $h = 100$ (balra), $h = 200$ (középen) és $h = 500$ blokkhosszúságok felhasználásával három irodalmi műben: Mark Twain: THE ADVENTURES OF TOM SAWYER (felső sor), Rudyard Kipling: THE JUNGLE BOOK (középső sor), és Kertész Imre: SORSTALANSÁG (alsó sor). A blokkok hosszának növekedésével nemcsak a zaj lett kisebb a függvényeken, de ezzel együtt kisimult a görbe és megszűntek azok a kiugrások, amely változások másodlagosnak tekinthetők a művekben, de még mindig jól érzékelhetőek voltak $h = 100$ esetén.



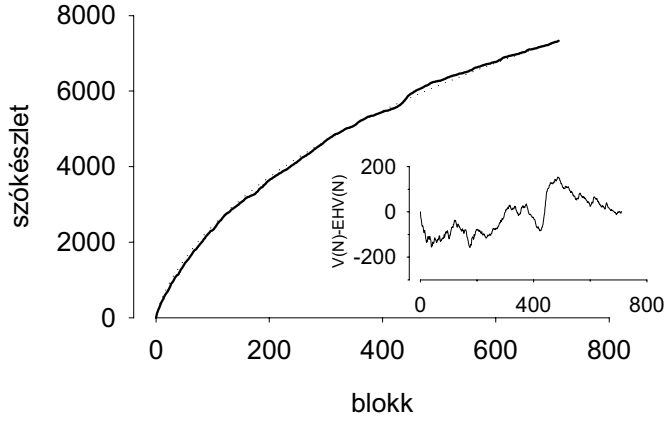
3.11. ábra. Edgar Allan Poe THE GOLD-BUG című művéből a szóalakok előfordulási gyakorisága alapján előállított empirikus eloszlás függvény.



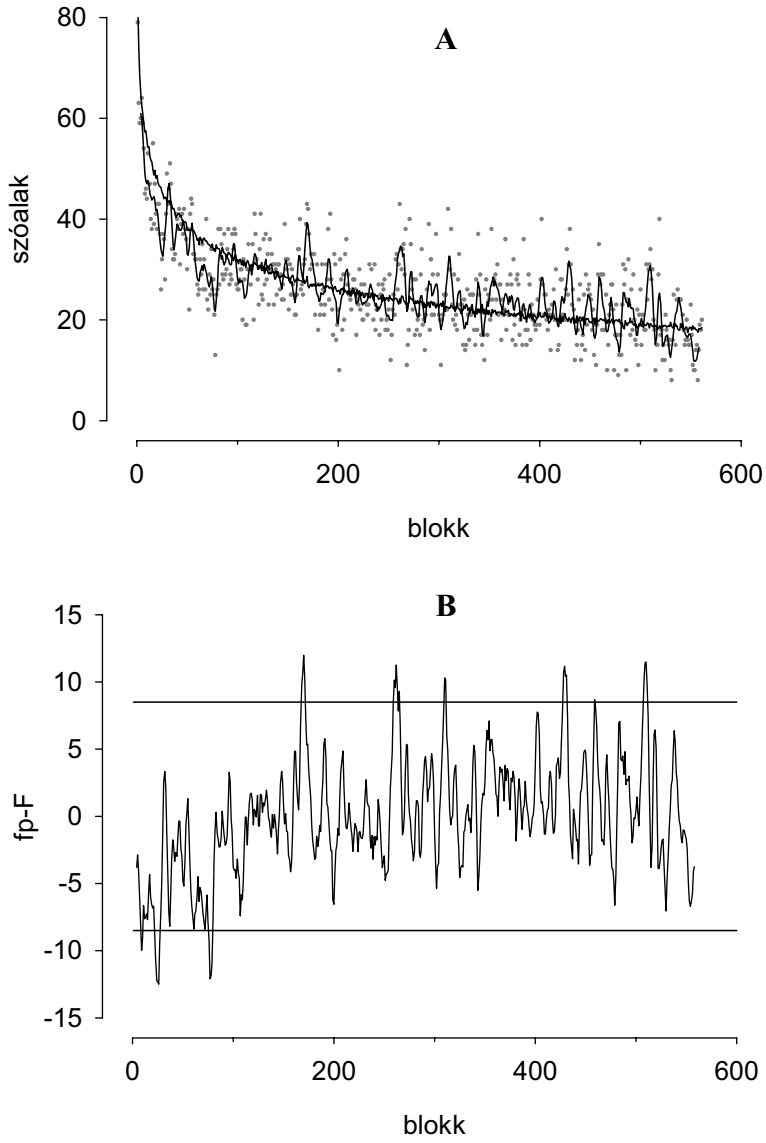
3.12. ábra Két közepes méretű – Mark Twain: THE ADVENTURES OF HUCKLEBERRY FINN (felső) és Daniel Defoe: THE ADVENTURES OF ROBINSON CRUSOE (alsó) – angol nyelvű regény szóalakjainak összehasonlítása. A folyamatos vonal az eredeti mű szóalakjait mutatja, a szaggatott vonal a *P1* modell alapján számolt szóalakokat, míg a pontozott vonal a módosított, *P2* modell alapján számolt szóalakokat mutatja.



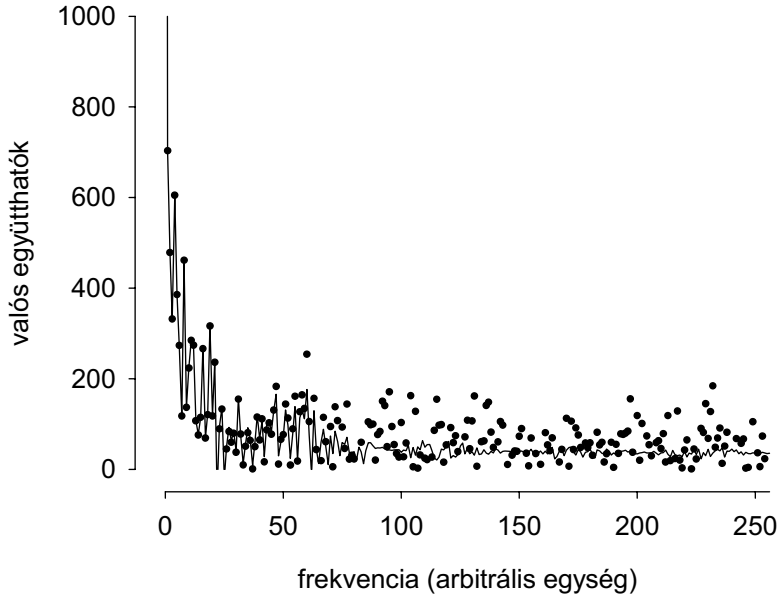
3.13. ábra. Szóalakok megjelenése Mark Twain THE ADVENTURES OF TOM SAWYER című művében és a mű alapján, polinomiális eloszlást feltételező modellel, előállított mesterséges szövegekben. Az eloszlás függvény alapján létrehozott modelleket használva a szaggatott vonal az eredeti modell alapján kapott mesterséges szöveg szóalakjainak számát (*EP1V*(*N*)), a pontozott vonal pedig a módosított modellel kapott értékeket adja (*EP2V*(*N*)). A belső ábra az eredeti és a mesterséges szöveg szókészletének nagysága közötti eltérést mutatja.



3.14. ábra. Mark Twain THE ADVENTURES OF TOM SAWYER. Hipergeometrikus eloszlást feltételező modellel előállított a mesterséges szöveg szókészletének nagyságát mutatja (*EHV*(*N*)). Ebben a modellben minden szóalak pontosan annyiszor lett kiválasztva ahányszor az eredeti szövegben szerepelt. A belső ábra az eredeti és a mesterséges szöveg szókészletének nagysága közötti eltérést mutatja.



3.15. ábra. Kertész Imre SORSTALANSÁG című művének elemzése. Az újonnan megjelenő szóalakok száma (pontok), a simított görbe (folyamatos, zajos függvény) és száz modell átlaga (felső). A simított görbe és az átlag függvény közötti eltérés (alsó). Az alsó ábrarészen a vízszintes vonalak az $M \pm 2\sigma$-t jelölik.



3.16. ábra. Kipling THE JUNGLE BOOK című művében az újonnan bevezetett szóalakok Fourier-spektruma. A pontok a mért adatokból, a folyamatos vonal az ezek simítása (7-pontos másodfokú polinomiális) után kapott értékekből meghatározott valós Fourier-együtthatókat mutatják.

3.1. táblázat. Példák aposztróffal kezdődő és végződő szavakra az eredeti elektronikus szövegekben.

Aposztróf helyzete	**Példák**
Tévesen aposztróffal kezdődő, végződő szavak	*The remarkable poem of 'The Raven' was* *'secret' drawer* *"How? – in what way?'** *" 'We will suppose,' said the miser, 'that his symptoms are such and such; now, doctor, what would you have directed him to take?'*
„Helyesen” aposztróffal kezdődő, végződő szavak	*'cause, 'ting*, *thinkin'*, *likin' flyin', an' didn', 'bout*
Szó belsejében „helyesen” használt aposztrófok	*G'night, sumf'n, more'n a minute longer, per'aps*

* Ez típusú hiba valószínűleg a digitalizálás során került a szövegbe.

3.2. táblázat. Példa hibásan használt szavakra, kifejezésekre Charles Dickens: DAVID COPPERFIELD és Winston Groom: FORREST GUMP című művekből. Nagyban megkönnyíti a szereplők azonosítását, ha ezeket a kifejezéseket eredeti alakjukban tartjuk meg és nem próbálunk a szövegeken lemmatizálást végezni.
Szövegben használt kifejezés → nyelvtanilag helyes kifejezés párok adják az egymásnak megfelelő szavakat, szóösszetételeket.

Charles Dickens: David Copperfield	**Winston Groom: Forrest Gump**
anywheres → anywhere	dunno → don’t know
heerd → heard	I says, she say→ I say, she says
fur → for	I be → I was
'twixt, betwixt → between	po ole → poor old
furdest → farthest	srimp bidness → shrimp business

3.3. táblázat. Angol és magyar szövegekben a szóalakok számának összehasonlítása. Egy angol nyelvű szövegben, ha egy segédige, egy helyhatározó affixum stb. bevezetésre kerül, akkor ezek ismételt előfordulása nem fogja emelni a szóalakok számát, szemben a magyar nyelvű szövegekkel, amelyekben ugyanezek a morfémák hozzácsatlakoznak a szótőhöz és folyamatosan emelik a szóalakok számát. A negyedik oszlopban az első összeadandó a független funkcionális morfémák számát, míg a második összeadandó a tartalommal bíró szavak számát mutatja.

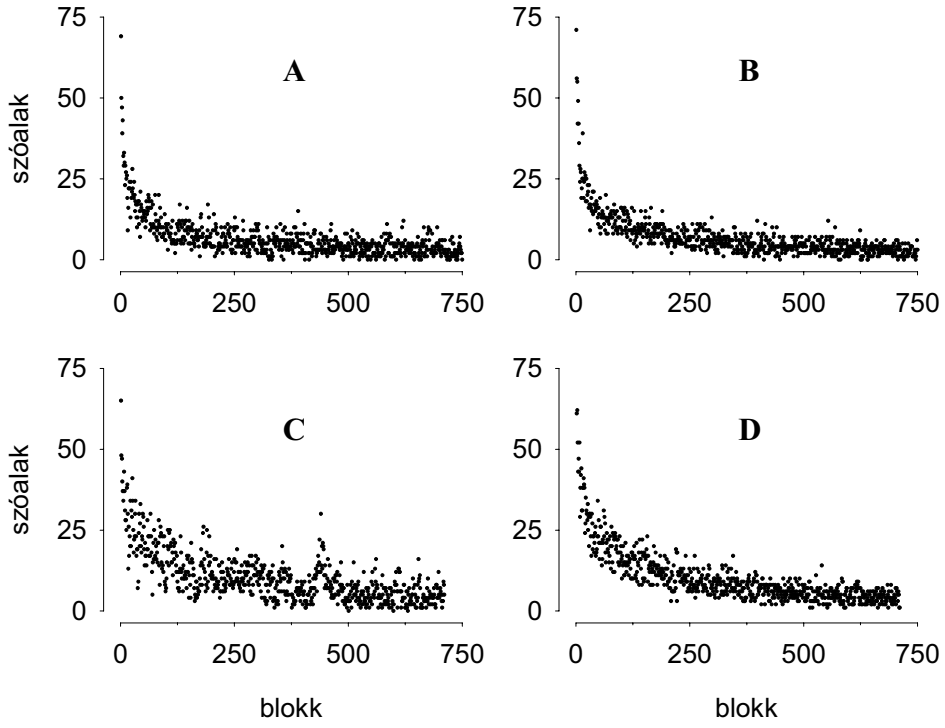
Magyar szavak	**Angol szavak**	**Szóalakok számának alakulása**	
		magyar szövegben	**angol szövegben**
ház	house	1	1
ház**ak**	houses	2	2
házak**ban**	**in** the houses	3	1+2
házak**ból**	**from** the houses	4	2+2
háza**ink**ból	from **our** houses	5	3+2
ablak	window	6	3+3
ablak**ok**	window**s**	7	3+4
ablakok**ban**	**in** the windows	8	3+4
ablakok**ból**	**from** the windows	9	3+4
ablaka**ink**ból	from **our** windows	10	3+4

*Az újonnan hozzáadott morfémák félkövérrel szedve.
**Névelők nincsenek számolva.

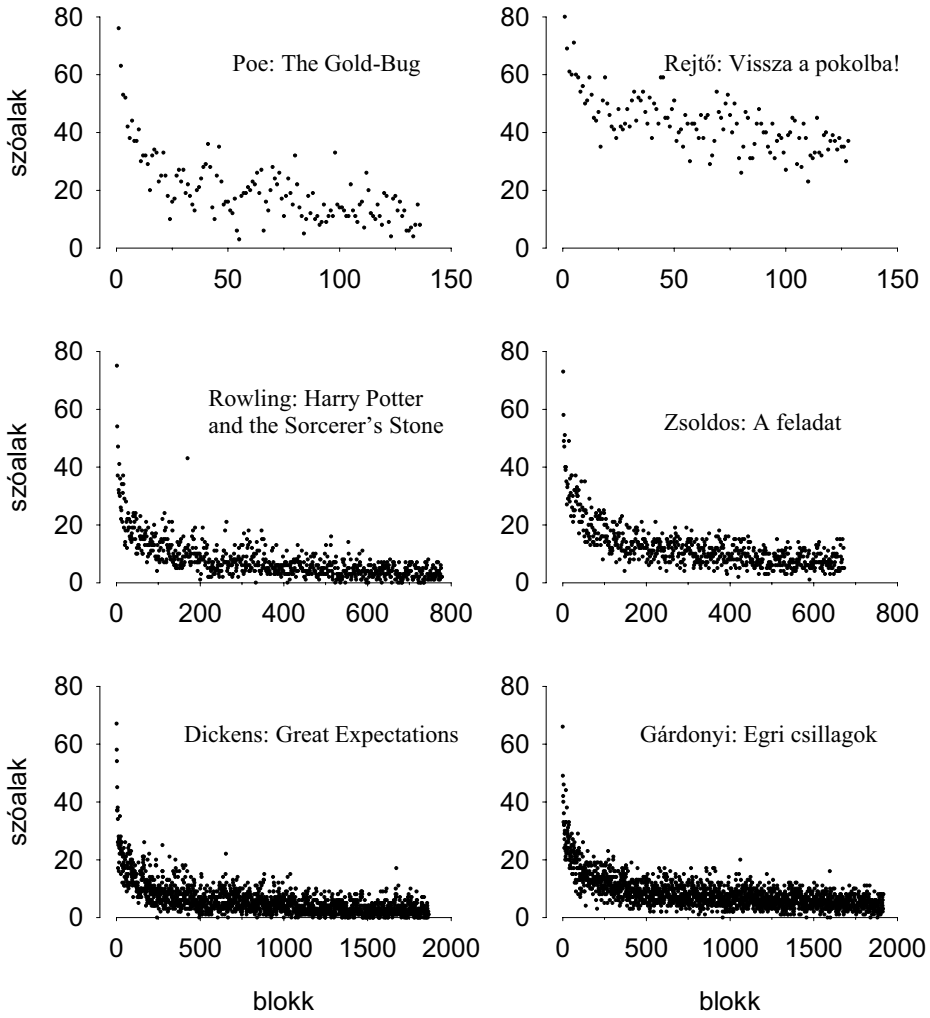
3.4. táblázat. Magyar mondatok és ezek angol fordítása. A magyar mondatok jellegzetessége, hogy ugyanazt négy* szót használják különböző sorrendben, és a szavak sorrendjének megváltoztatása elegendő ahhoz, hogy a mondat új jelentést kapjon.

Magyar mondatok	**Angol mondatok**
Tegnap mindenki eljött hozzánk.	Yesterday everyone came to our place.
Tegnap jött el hozzánk mindenki.	It was yesterday that everyone came to our place.
Tegnap hozzánk jött el mindenki.	It was our place that everyone came to yesterday.
Mindenki eljött hozzánk tegnap.	Everyone came to our place yesterday.
Hozzánk jött el mindenki tegnap.	It was our place that everyone came to yesterday.
Eljött hozzánk mindenki tegnap.	Everyone did come to our place yesterday.
Eljött hozzánk mindenki tegnap?	Did everyone come to our place yesterday?
stb.	

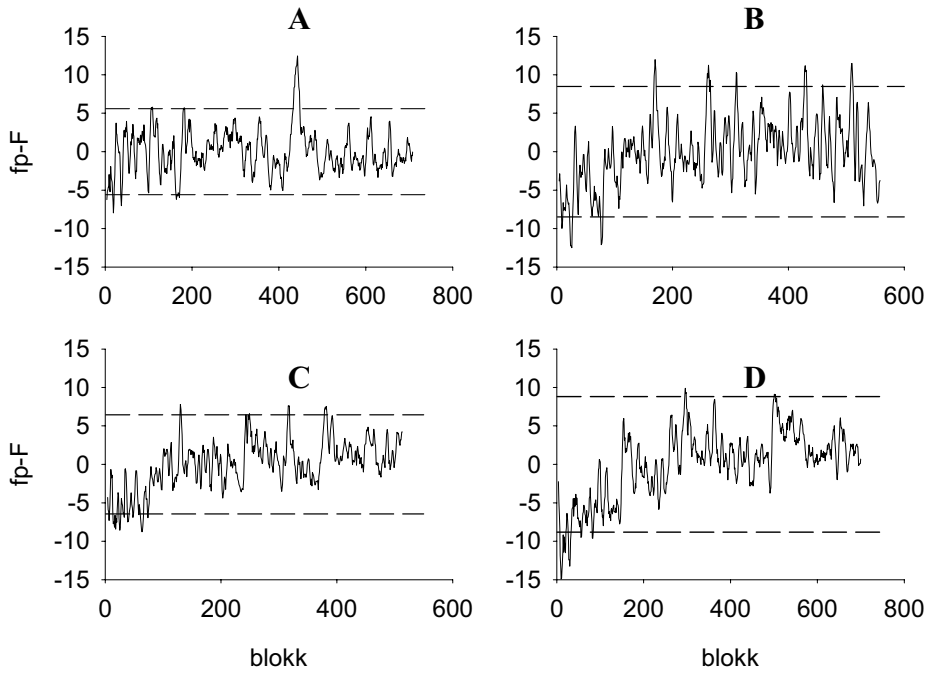
*Az igekötős igéket egy szónak vettük abban az esetben is, ha az igekötő az ige után áll.



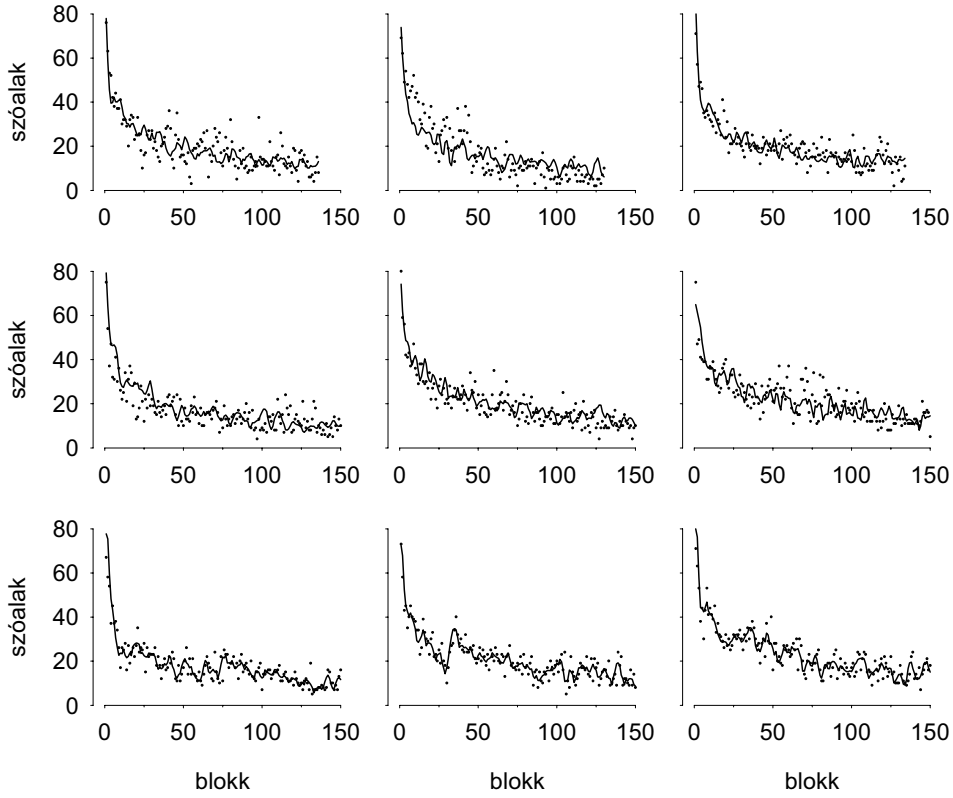
4.1. ábra. Daniel Defoe: ROBINSON CRUSOE (A, B) és Mark Twain: THE ADVENTURES OF TOM SAWYER (C, D). Balra (A, C) az eredeti művek újonnan bevezetésre kerülő szóalakjainak számát ábrázoltuk száz szövegszó hosszúságú blokkok esetén, jobbra (B, D) a mesterséges szövegek szóalakjait. A modell a szövegben jelenlévő trendeket vissza tudta adni. Ezeket nevezzük a görbe elsődleges kiugrásainak. Ezzel szemben a szezonalitásokat követő másodlagos kiugrások nem jelentek meg a mesterséges szöveget leíró görbén (D).



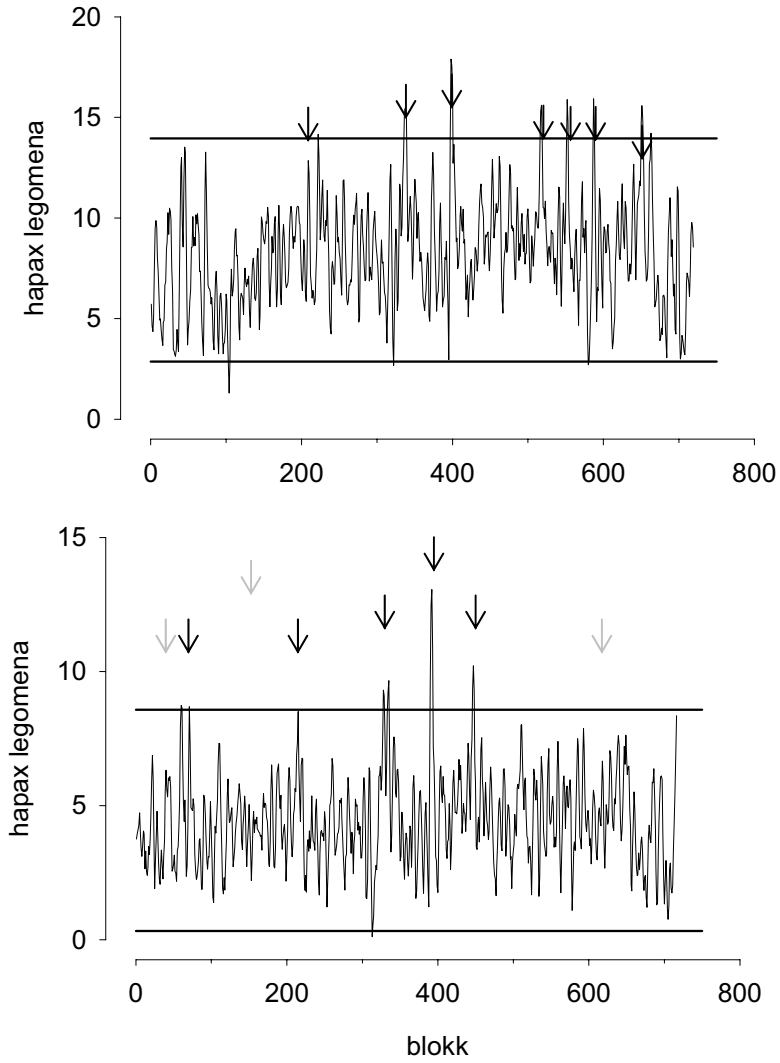
4.2. ábra. Szóalakok bevezetése angol (bal) és magyar (jobb) nyelvű irodalmi művekben. A szövegeket száz szövegszó hosszúságú blokkokra osztottuk. A grafikonok az egyes blokkokban újonnan bevezetett szóalakok számát mutatják különböző hosszúságú szövegek esetén. A felső sorban „rövid", kb. 15000, a középső sorban „közepes" hosszúságú, kb. 80000, míg az alsó sor hosszú, kb. 150000 szövegszót tartalmazó művek újonnan bevezetett szóalakjainak száma látható.



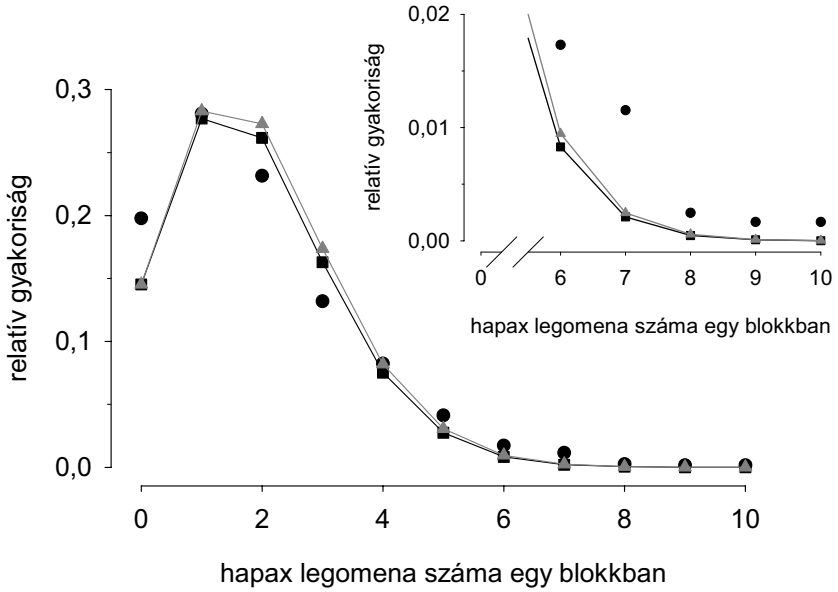
4.3. ábra. Az eredeti és a mesterséges szöveg közötti különbség. Az eredeti szöveg szóalakjainak leírásához használt görbe 7-pontos simítását, valamint a modellek szóalakjainak átlagát (100 modell átlaga) hasonlítottuk össze. A szaggatott vonal az $M \pm 2\sigma$ értéket mutatja. A feldolgozott művek: Mark Twain: THE ADVENTURES OF TOM SAWYER (A), Kertész Imre: SORSTALANSÁG (B), Rudyard Kipling: THE JUNGLE BOOK (C) és AMERICAN MYSTERY STORIES (D) különböző szerzőktől.



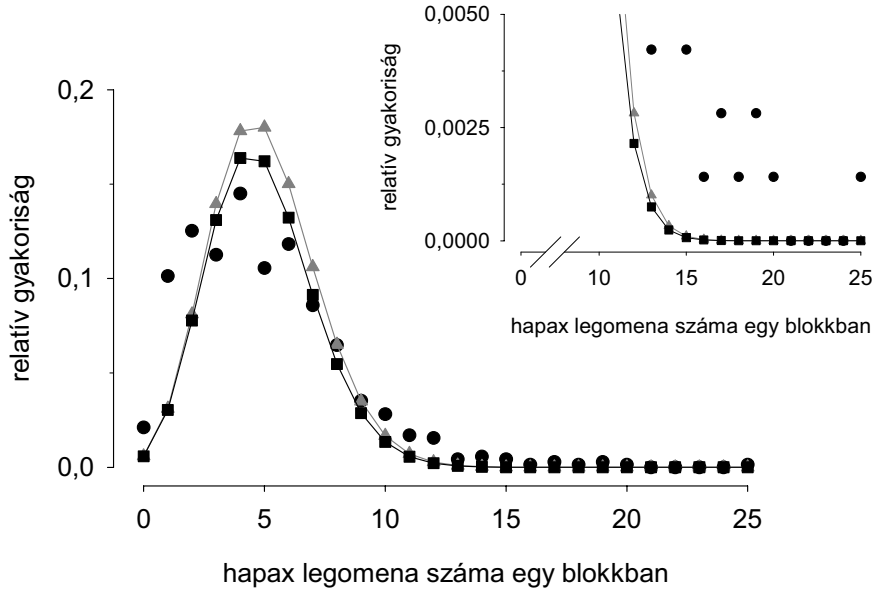
4.4. ábra. 150 blokk hosszúságúra csonkított szövegek összehasonlítása. A felső sorban rövid történetek, novellák, a középső sorban közepes méretű regények, míg az alsó sorban hosszabb lélegzetű művek első 150 blokkjában újonnan bevezetett szóalakok számát ábrázoltuk. A középső oszlopban Jack London három műve: AN ODYSSEY OF THE NORTH, CALL OF THE WILD, SEAWOLF, a jobb szélsőben Joseph Conrad három műve: YOUTH, THE END OF THE TETHER, és LORD JIM, míg a bal szélsőben három különböző szerző különböző hosszúságú műve: Edgar Allan Poe: THE GOLD-BUG, J. K. Rowling: HARRY POTTER AND THE SORCERER/(PHILOSOPHER)'S STONE, és végül Charles Dickens: GREAT EXPECTATIONS szerepel. A folyamatos vonal a simítás eredményét mutatja.



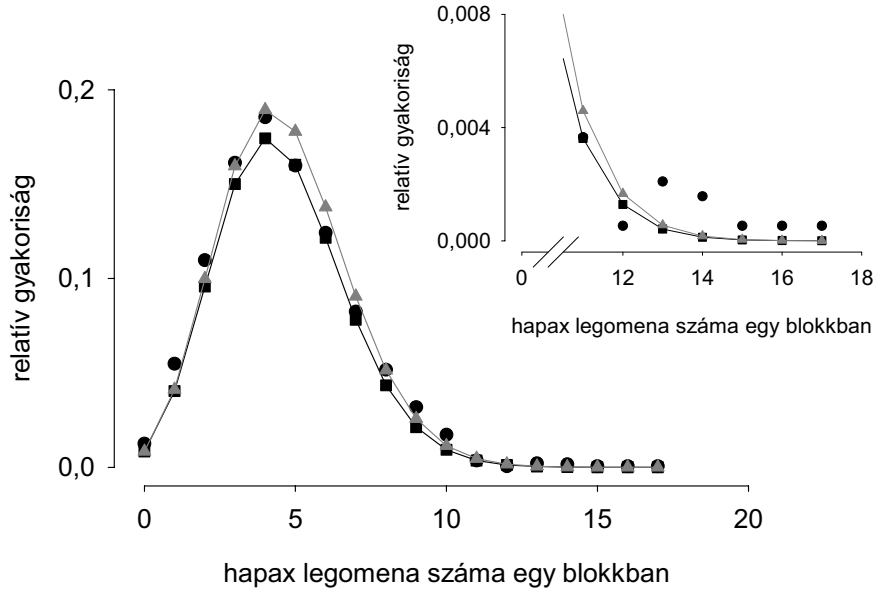
4.5. ábra. Az egyszer előforduló szavak megjelenése a ROMAN EINES SCHICKSALLOSEN (fent) és a FATELESS (lent) művekben. Az ábrán az $M \pm 2 \cdot \sigma$ jelző vonalakat a hapax legomena hipergeometrikus eloszlását feltételezve húztuk meg. A kiugrások az esetek többségében azokon a helyeken jelentek meg, ahol az eredeti szövegben a megnövekedett az újonnan bevezetett szóalakok száma. A nyilak azokat a helyeket mutatják, ahol az újonnan bevezetett szóalakok száma szignifikáns eltérést eredményezett. A szürke nyilak arra utalnak, hogy a hapax legomena száma nem haladta meg abban a pontban szignifikancia küszöböt, míg az újonnan bevezetett szóalakok száma igen.



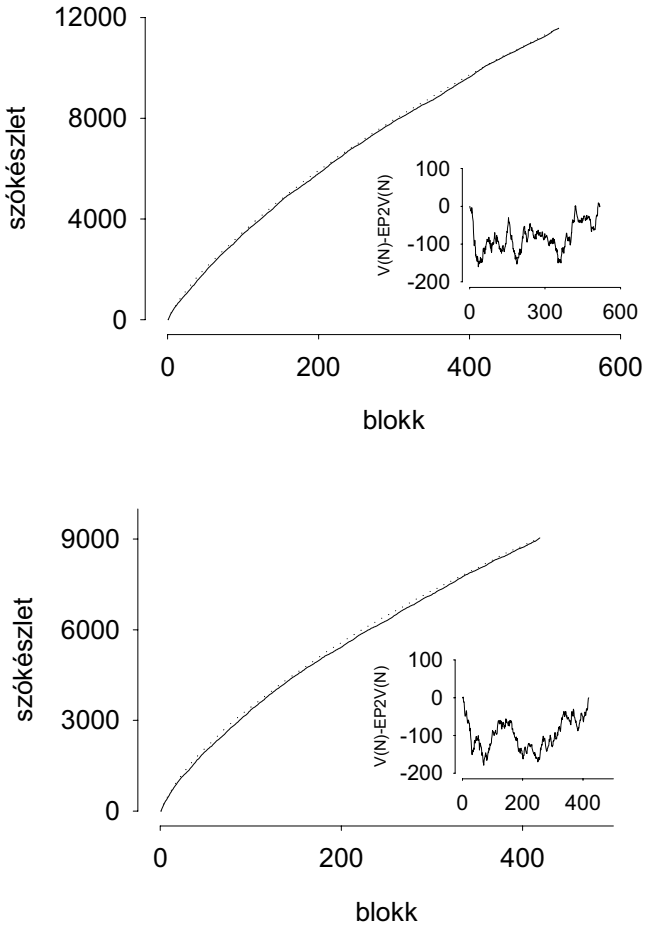
4.6. ábra. Defoe: THE ADVENTURES OF ROBISON CRUSOE. Ebben a műben viszonylag alacsony a hapax legomena száma ($V(1,N) = 2319$), így a blokkonkénti hapax legomena száma is ($n = 1214$, $V(N) = 6073$, $rep = 19{,}99$). Ennek következtében viszonylag alacsony azoknak a blokknak a száma, amelyekben a mért és a modell alapján számított értékek eltérnek.



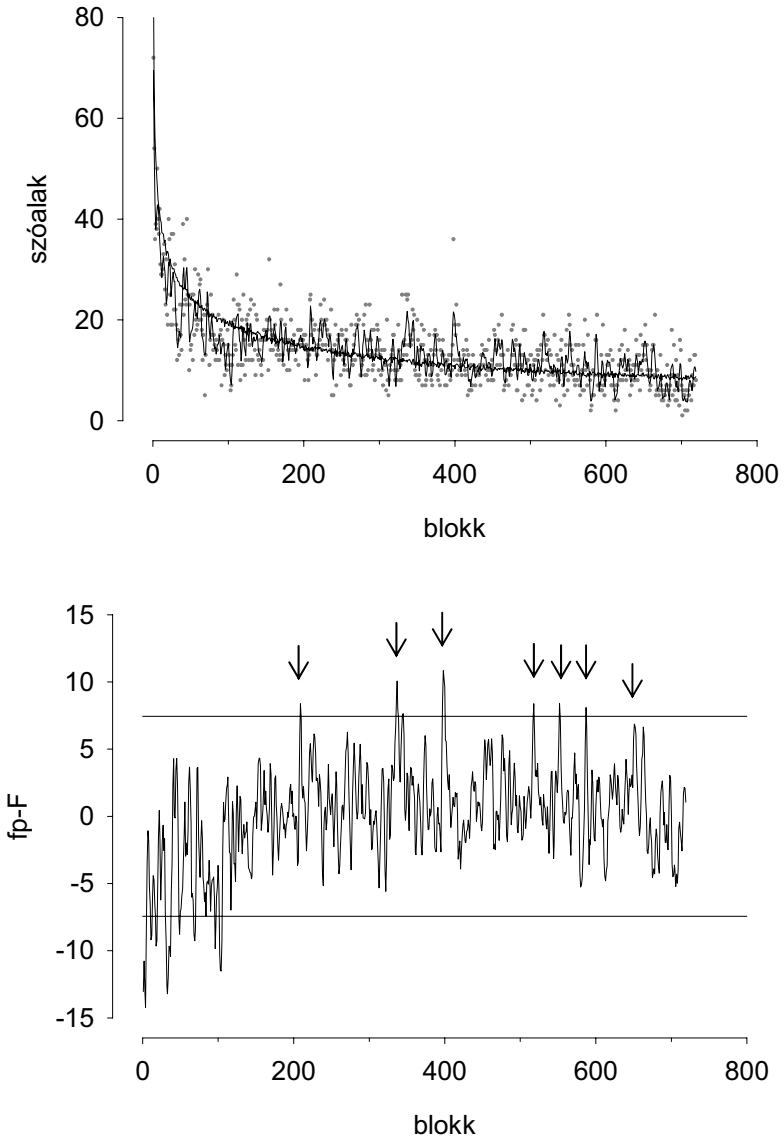
4.7. ábra. Twain: THE ADVENTURES OF TOM SAWYER. Ebben a műben viszonylag magas a hapax legomena száma ($V(1,N) = 3556$), így a blokkonkénti hapax legomena száma is ($n = 711$, $V(N) = 7331$, $rep = 9{,}7$). Azoknak a blokkoknak a száma, amelyekben magas a hapax legomena száma jóval a becsült érték fölött van. Ezek azok a blokkok, ahol hirtelen megemelkedik az újonnan bevezetésre kerülő szóalakok száma.



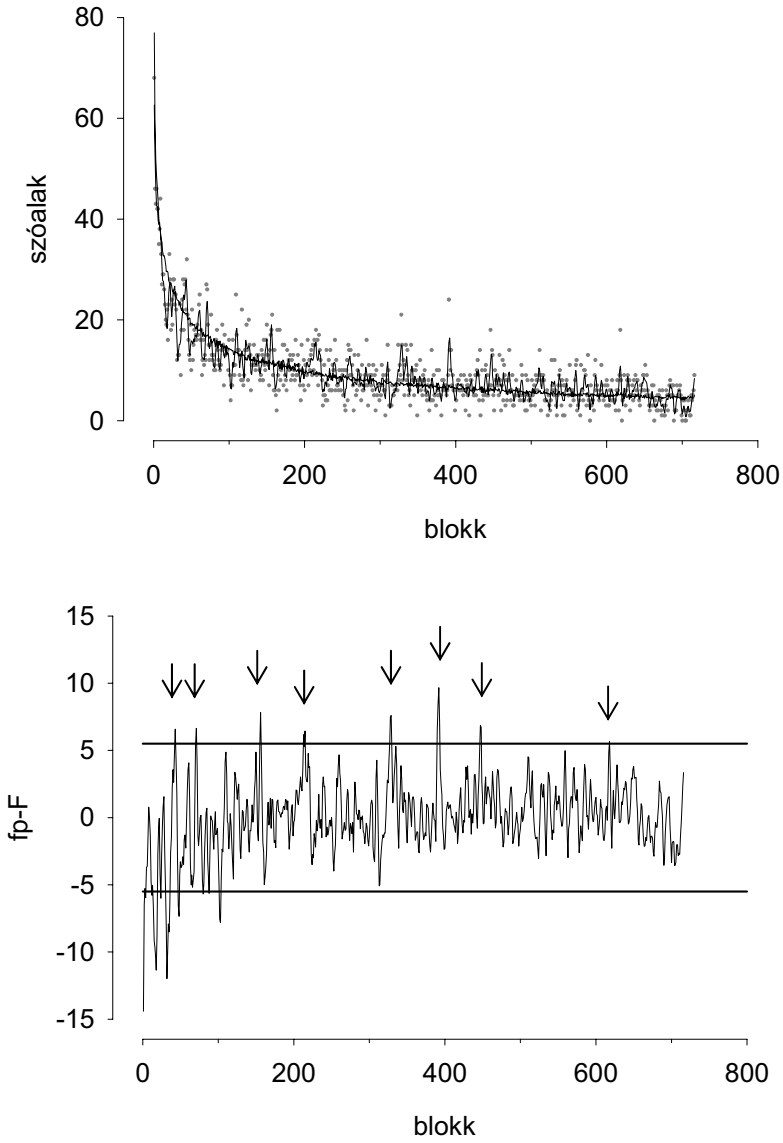
4.8. ábra. Gárdonyi: EGRI CSILLAGOK. Ahogy azt magyar nyelvű műveknél láttuk a különböző szóalakok és a hapax legomena ($n = 1916$, $V(N) = 16237$, $rep = 11{,}8$, $V(1,N) = 2319$) száma magasabb, mint hasonló hosszúságú angol nyelvű műveknél. Vizsgálva azonban a hapax legomena blokkonkénti eloszlását és ezeket az értékeket összehasonlítva a könyvhöz tartozó modellel magyar nyelvű szövegek esetén is megkapjuk azokat a blokkokat, amelyekben a vártnál magasabb ezeknek a száma.



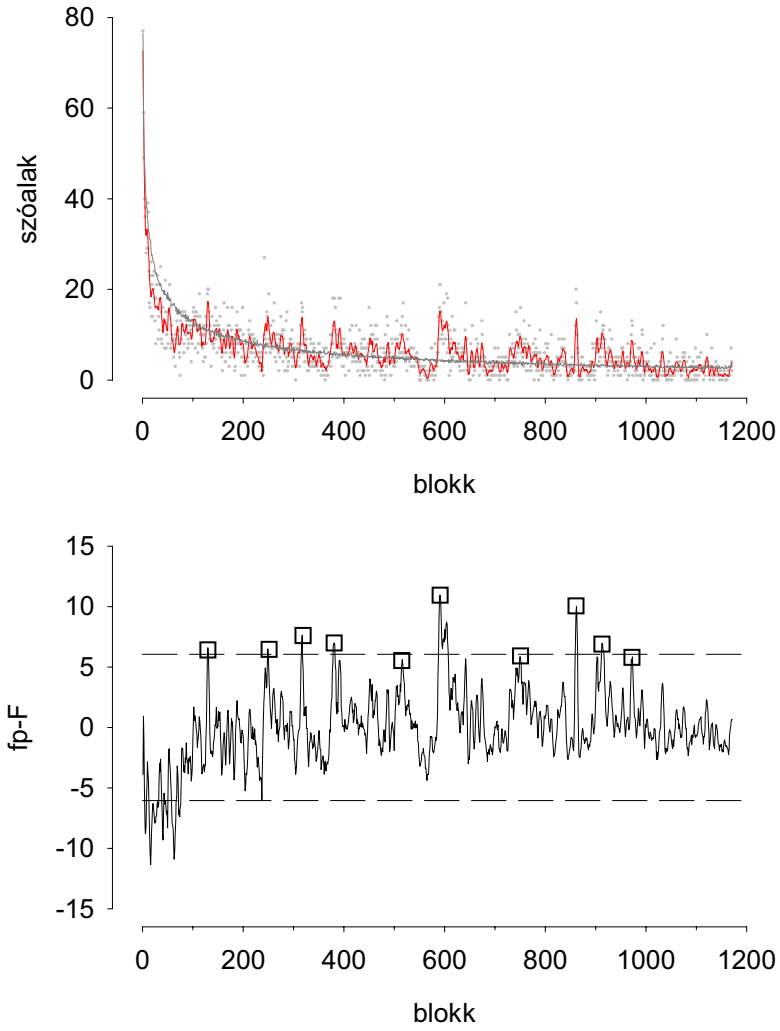
4.9. ábra. Szóalakok bevezetése két magyar nyelvű szövegben. Az eredeti szöveg szóalakjait a pontozott görbével ábrázoltuk, míg a modell szóalakjait folyamatos vonallal. A belső grafikonok az eredeti és a mesterséges szöveg szóalakjainak száma közötti eltérést mutatják. A két feldolgozott mű: Tamási Áron: ÁBEL A RENGETEGBEN (felső), és Molnár Ferenc: A PÁL UTCAI FIÚK (alsó).



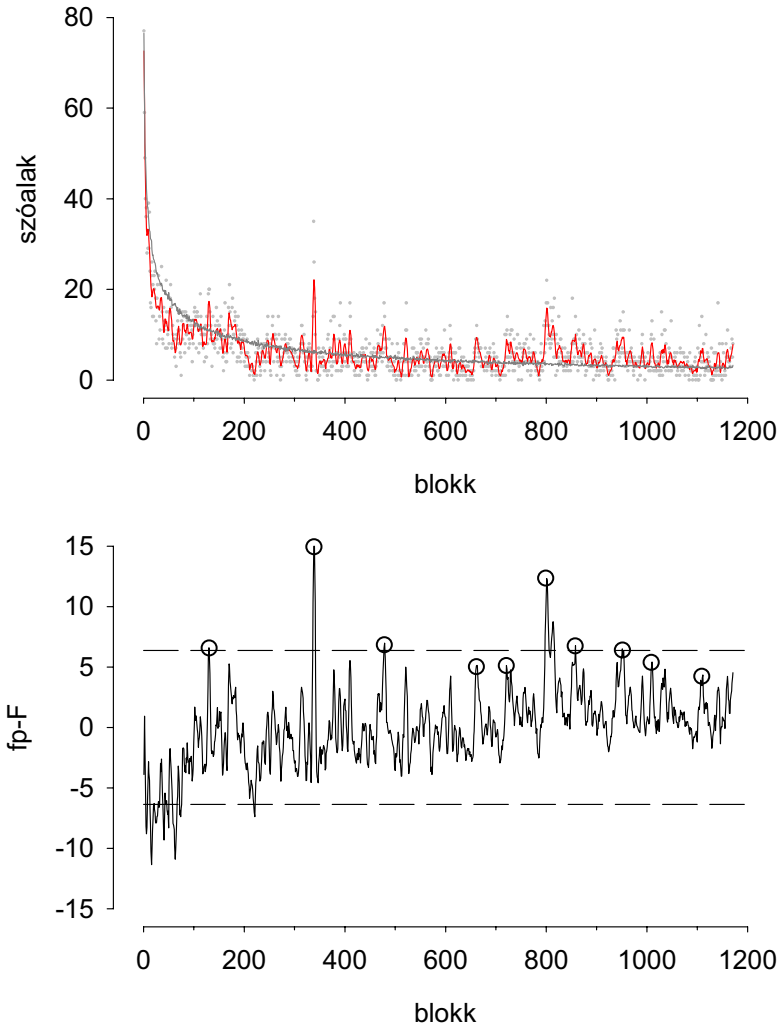
4.10. ábra. Kertész Imre SORSTALANSÁG című művének német fordítása: ROMAN EINES SCHICKSALLOSEN. A német és a magyar nyelvű szövegben apró eltérésektől eltekintve a szövegnek ugyanazon a pontján emelkedett meg az újonnan bevezetett szóalakok száma. A nyilak a német nyelvű szöveg azon pontjait mutatják, amelyekben az újonnan megjelenő szóalakok száma szignifikáns eltérést eredményezett.



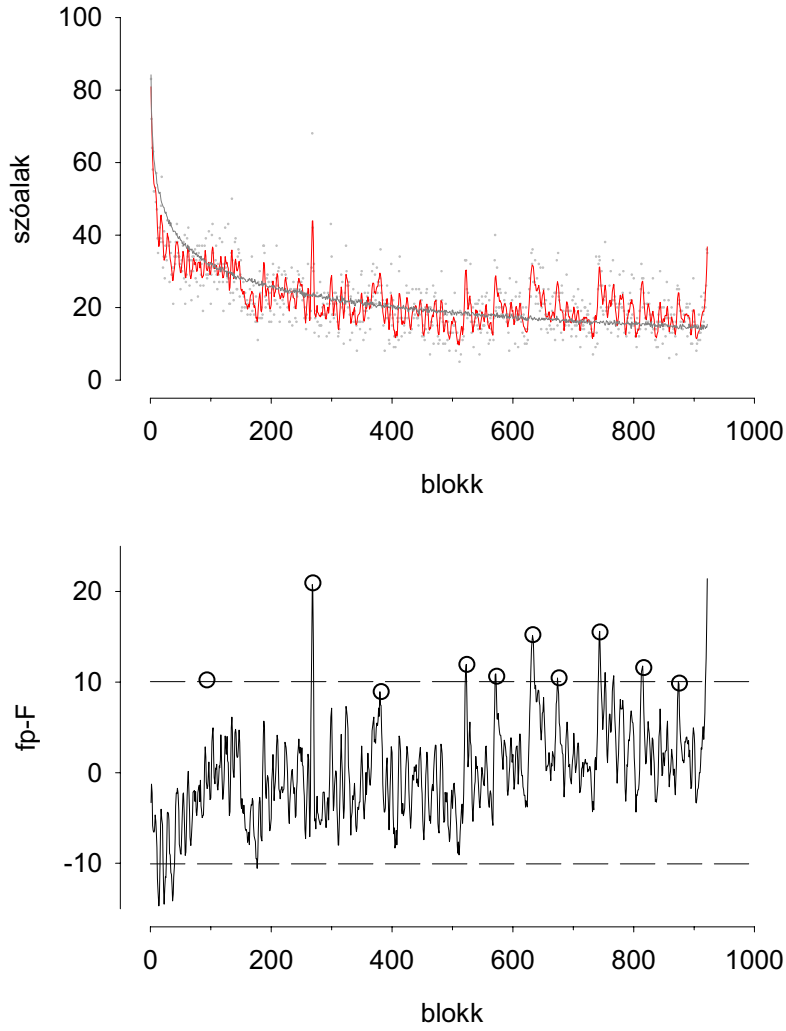
4.11. ábra. Kertész Imre SORSTALANSÁG című művének angol fordítása: FATELESS. A magyar és a német nyelvű szöveghez hasonlóan olyan eseményeknél jelentek meg a kiugrások, amelyek nem képezik szerves részét a szövegnek, nem logikus következményei az előzményeknek, és a folytatáshoz sem kötődnek.



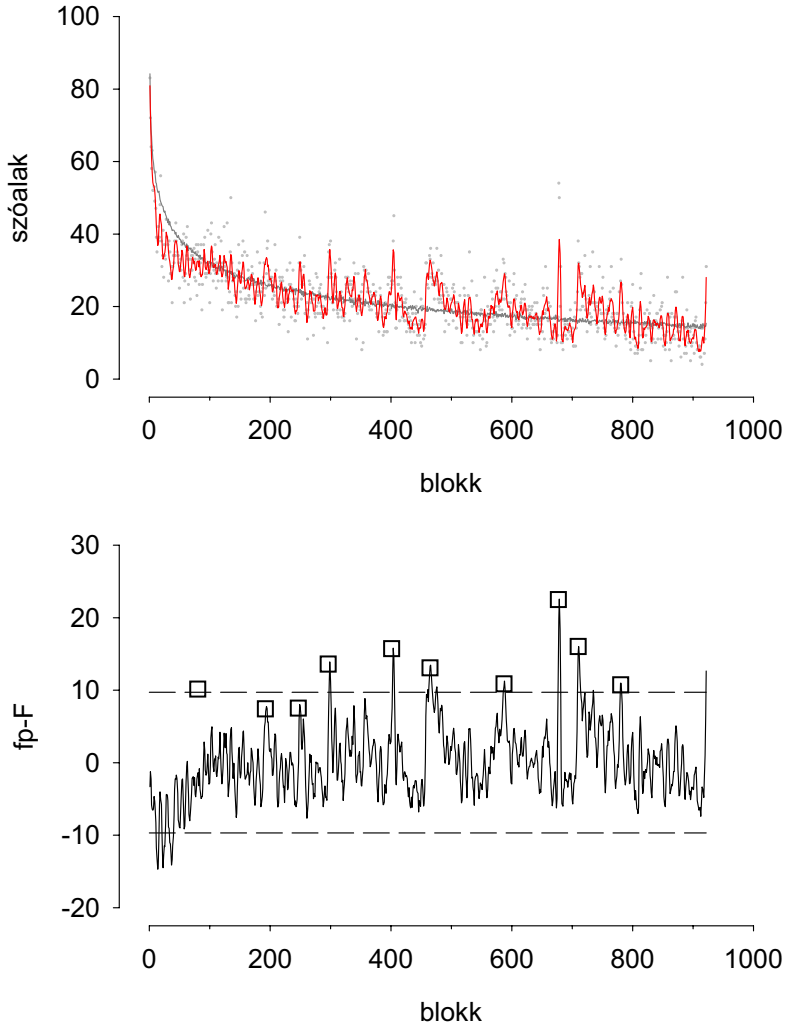
4.12. ábra. Kipling THE JUNGLE BOOKS című művében (eredeti sorrend, 4.9. táblázat) az újonnan megjelenő szóalakok száma $h = 100$ esetén (fent). A blokkonkénti szóalakok és a modell által számolt értékek közötti különbség. (A négyzetek értelmezését lsd. a szövegben.)



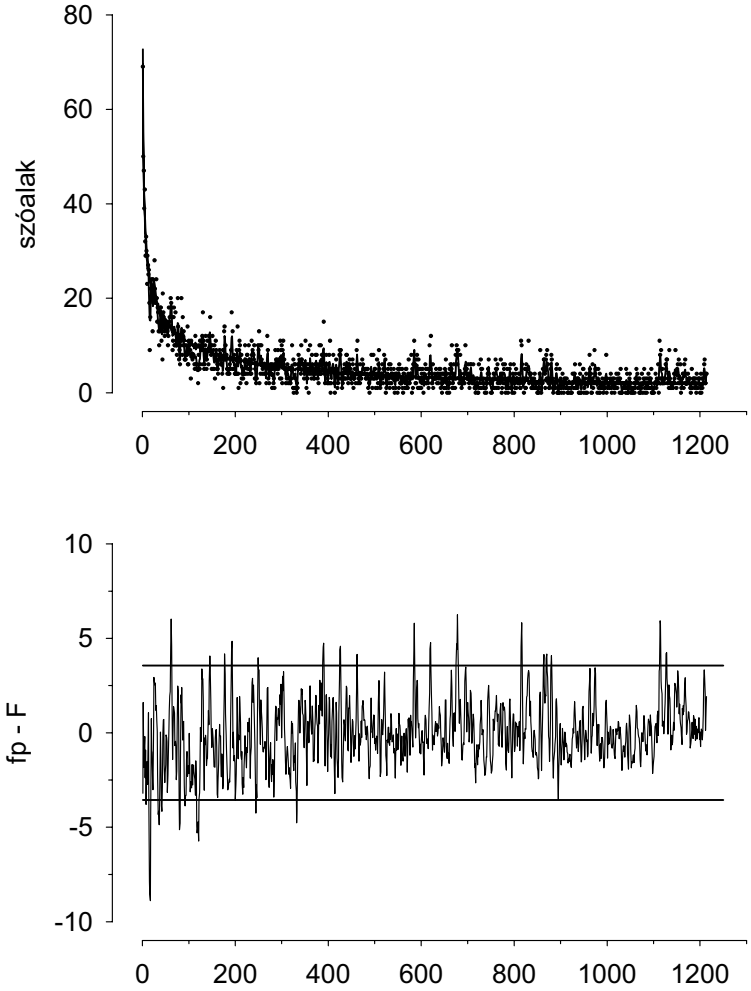
4.13. ábra. Kipling THE JUNGLE BOOKS című művének egy olyan verziója, ahol a novellákat a magyar sorrend szerint rendeztük (4.8. és 4. 10. táblázat). Az újonnan megjelenő szóalakok száma (fent) és a mesterséges szövegtől való eltérése (lent) $h = 100$ esetén. (A körök értelmezését lsd. a szövegben.)



4.14. ábra. A DZSUNGEL KÖNYVE magyar nyelvű mű elemzése (4.11. táblázat). A magyar fordításban megváltozott a novellák eredeti sorrendje. Ez a sorrend is alkalmas arra, hogy megtaláljuk a műben azokat a helyeket, ahol megváltozott a felhasznált szókészlet. Ezzel szemben, az eredeti mű és fordításának közvetlen összehasonlítása a novellák eltérő sorrendje mellett nem lehetséges. Akkor tudjuk csak összehasonlítani ezeket a könyveket, ha az egyik könyvben a novellák sorrendjét hozzá igazítjuk a másik sorrendhez.

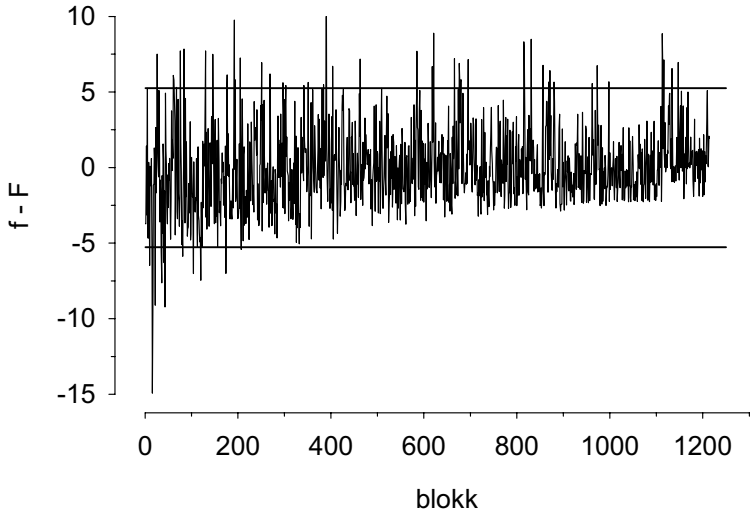


4.15. ábra. A DZSUNGEL KÖNYVE magyar nyelvű mű, amelyben a novellák sorrendje az eredeti, angol sorrendet követi (4.12. táblázat).

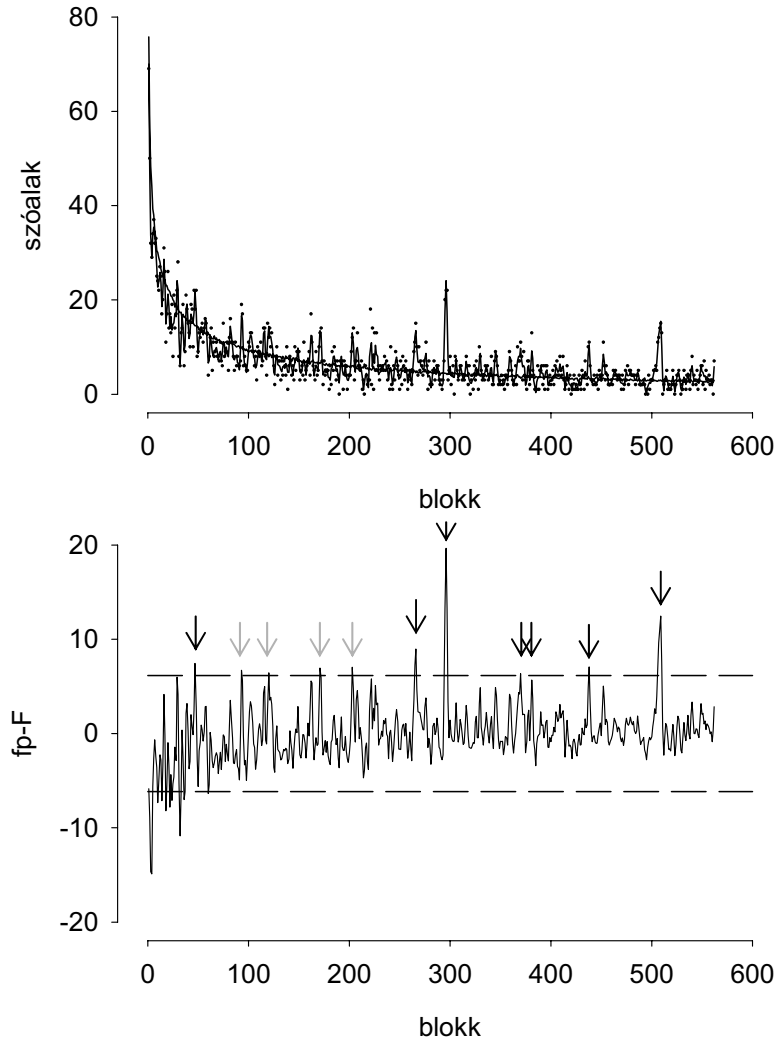


a

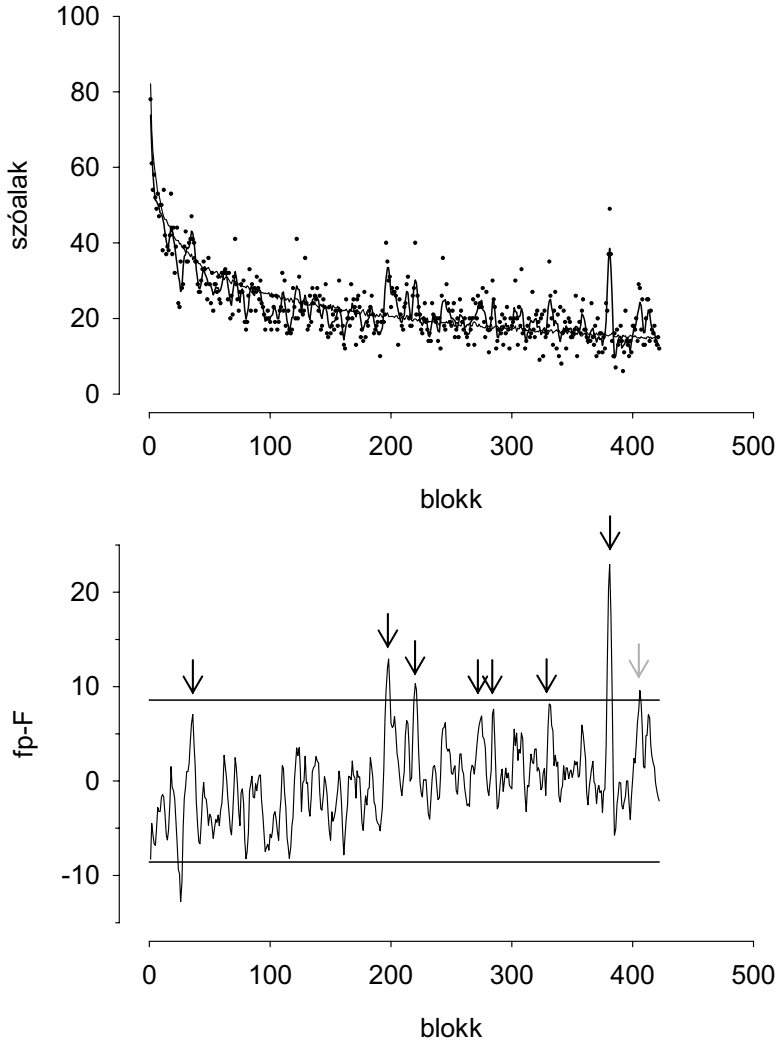
4.16. ábra. Defoe: THE ADVENTURES OF ROBINSON CRUSOE. A megjelenő szóalakok egyenletesebb eloszlást mutatnak, mint az eddig vizsgált művekben. Ennek következménye, hogy $2\sigma = 3{,}559$. Az egyenletesebb eloszlásból az is következik, hogy kisebb változások is érzékelhetőek, míg egy olyan műben, ahol vannak durva, a könyv stílusától nagy eltérések, a kisebbek nem fognak szignifikáns eltérésként megjelenni.



4.17. ábra. Defoe: THE ADVENTURES OF ROBINSON CRUSOE. Az eddigiektől eltérően nem a simított görbét hasonlítottuk a mesterséges szövegek átlagához, hanem az eredeti könyv alapján mért adatokat.



4.18. ábra. Az ALICE történetek eredeti, angol nyelvű szövegének elemzése. A fekete nyilak azokat a pontokat mutatják, ahol a magyar nyelvű szövegben is kaptunk kiugrást, míg a szürkék azokat, amelyeknél a magyar szövegben nincs kiugró szóalakszám emelkedés.



4.19. ábra. Az ALICE történetek magyar nyelvű fordításának elemzése. A magyar nyelvű szövegben egyetlen olyan szignifikáns kiugrás kaptunk, amely az angol szövegben nem volt jelen. Ez a könyv végén egy ünnepség leírása.

4.1. táblázat. Mark Twain THE ADVENTURES OF TOM SAWYER című művében a mesterséges szövegek átlagától legalább 2σ-val eltérő helyek és az ottani események (h = 100). A blokk annak az intervallumnak a sorszámát adja, amelyben a függvény felveszi a helyi maximumát.

esemény	**blokk**
részlet a vasárnapi miséből	109
álmodozás, mi lenne ha…	182
év végi záróvizsga verse (A Missouri Maiden's Farewell to Alabama)	443

4.2. táblázat. Rudyard Kipling THE JUNGLE BOOK című művében a mesterséges szövegek átlagától legalább 2σ-val eltérő helyek és az ottani események (h = 100).

esemény	**blokk**
királyi palota leírása	130
a fehér fóka leírása a több hónapos helykeresés után	249
Rikki-Tikki-Tavi mese kezdete	316
az elefánt életének leírása	382

4.3. táblázat. AMERICAN MYSTERY STORIES című novellás kötetben a mesterséges szövegek átlagától legalább 2σ-val eltérő helyek és az ottani események (h = 100).

cím	**esemény**	**blokk**
The Man and the Snake	az épület egyik szárnyának leírása	296
The Gold-Bug	a sziget leírása	364
Wolfert Webber or Golden Dreams	a Webber család és lakóhelyük környékének leírása	501

4.4. táblázat. Edgar Allan Poe két novellás kötete összehasonlítva különböző szerzők hasonló zsánerű műveivel (ebben a válogatásban is szerepel néhány történet, amelynek a szerzője Poe), valamint egy eltérő zsánerű novellás kötettel: Kipling: THE JUNGLE BOOK.

	blokk	**szóalak**	**hapax legomena**	$\frac{N}{V(N)}$	$\frac{V(1,N)}{V(N)}$
The Works of Edgar Allan Poe Volume 1 of the Raven Edition	880	9186	4324	9,58	0,471
The Works of Edgar Allan Poe Volume 2 of the Raven Edition	945	10087	4691	9,37	0,465
American Mystery Stories (rövid)	703	8231	4107	8,54	0,5
American Mystery Stories	1236	11170	4874	11,07	0,436
Rudyard Kipling: The Jungle Book	516	4688	2064	11,01	0,440
Rudyard Kipling: The Jungle Books	1171	7452	3124	15,71	0,419

4.5. táblázat. Hasonló hosszúságú angol és magyar nyelvű szövegek szókészletének összehasonlítása

	blokk	**szóalak**	**hapax legomena**	$\frac{N}{V(N)}$
Great Expectations	1865	11022	4751	16,92
Egri csillagok	1916	16237	8938	11,8
End of the Tether	530	6785	3468	7,81
The Jungle Book	516	4688	2064	11,01
Ábel a rengetegben	517	11571	7856	4,47
The Gold-Bug	136	2694	1543	5,05
Vissza a pokolba	128	5520	4163	2,32

4.6. táblázat. Kertész Imre SORSTALANSÁG és a mű angol és német nyelvű fordítása, Rudyard Kipling THE JUNGLE BOOKS és a magyar fordítása. Az angol szövegekben fordul elő a legkevesebb különböző szóalak és ezzel párhuzamosan a legkevesebb hapax legomena.

	blokk	**szóalak**	**hapax legomena**
Sorstalanság	561	14740	10253
Fateless	716	6710	3186
Roman eines Schicksallosen	719	9992	6043
The Jungle Books	1171	7452	3124
A dzsungel könyve	922	20362	13372

4.7. táblázat. Kertész Imre SORSTALANSÁG című műve és annak német és angol fordítása. A számok azoknak a blokkoknak a sorszámát jelölik, amelyekben az újonnan bevezetett szóalakok száma magasabb, mint az a modell alapján várható volt.

	magyar	**német**	**angol**
Vili bácsi			43
csepeli üzem			71
indulás a vonattal 1-2		209	156
Auschwitzba érkezés	170		215
Buchenwaldba érkezés	262	337	329
reggeli készülődés, üzem	310	398	392
főszereplő testi állapota			447
főszereplő lelki állapota		518	
kórház	429	552	
Pjetyka főz	459	587	
napi menetrend			618
haza indulás	510	651	

4.8. táblázat. Kipling THE JUNGLE BOOKS és magyar fordítása, A DZSUNGEL KÖNYVE című művekben szereplő novellák a megjelenés sorrendjében

eredeti, angol sorrend	**magyar sorrend**
Book I	
Mowgli's Brothers	Maugli testvérei
Kaa's Hunting	Ká vadászata
„Tiger-Tiger"	Hogyan született a félelem
The White Seal	„Tigris! Tigris!"
„Rikki-Tikki-Tavi"	A király ankusa
Toomai of the Elephants	Rátok szabadítom a dzsungelt
Servants of the Queen	A vörös kutyák
Book II	Tavaszi futás
How Fear Came	
The Miracle of Purun Bhagat	„Riki-Tiki-Tévi"
Letting in the Jungle	A fehér fóka
The Undertakers	Purun Bagát csodája
The King's Ankus	A krokodilus története
Quiquern	Kvikvern
Red Dog	Kis Tumáj és az elefántok tánca
The Spring Running	A királynő szolgái

4.9. táblázat. THE JUNGLE BOOKS című műben, megtartva a novellák eredeti, angol sorrendjét (4.12. ábra), meghatároztuk a mesterséges szövegek átlagától legalább 2σ-val eltérő helyeket (h = 100) és az ottani eseményeket.

ssz.	**cím**	**esemény**	**blokk**
1.	Kaa's Hunting	királyi palota leírása	130
2.	The White Seal	a fehér fóka leírása a több hónapos helykeresés után	249
3.	Rikki-Tikki-Tavi	Rikki-Tikki-Tavi mese kezdete	316
4.	Toomai of the Elephants	az elefánt életének leírása	381
6.	The Miracle of Purun Bhagat	barátok felsorolása foglalkozásuk szerint	591
		zarándokok felsorolása	604
8.	The King's Ankus	kincsek felsorolása	862
9.	Quiquern	Kadlu	913

4.10. táblázat. THE JUNGLE BOOKS (magyar sorrend, 4.13. ábra) című novellás kötetben a mesterséges szövegek átlagától legalább 2σ-val eltérő helyek (h = 100) és az ottani események.

ssz.	cím	esemény	blokk
1.	Kaa's Hunting	királyi palota leírása	130
2.	The King's Ankus	kincsek felsorolása	339
3.	Letting in the Jungle	Hathi lerombolja a falut	479
6.	The Miracle of Purun Bhagat	barátok felsorolása foglalkozásuk szerint	801
		zarándokok felsorolása és a hegy leírása	816
7.	Undertakers	madarak felsorolása	858
8.	Quiquern	Kadlu	952

4.11. táblázat. A DZSUNGEL KÖNYVE című novellás kötetben (magyar sorrend, 4.14. ábra) a mesterséges szövegek átlagától legalább 2σ-val eltérő helyek (h = 100) és az ottani események.

ssz.	cím	esemény	blokk
2.	The King's Ankus	kincsek felsorolása	268
4.	Rikki-Tikki-Tavi	Rikki-Tikki-Tavi mese kezdete	523
5.	The White Seal	a fehér fóka	572
6.	The Miracle of Purun Bhagat	Purun Bhagat jellemzése	633
7.	Undertakers	madarak felsorolása	674
9.	Quiquern	kezdete	744
		Kadlu	753
		tél	767
	Quiquern + Toomai of the Elephants	záró vers + kezdő vers	814
10.	The Servants of the Queen	afganisztáni emír látogatása	873

4.12. táblázat. A DZSUNGEL KÖNYVE (angol sorrend, 4.15. ábra) című novellás kötetben a mesterséges szövegek átlagától legalább 2σ-val eltérő helyek (h = 100) és az ottani események.

	cím	esemény	blokk
4.	Toomai of the Elephants	eleje	299
5.	Parade-Song of Camp Animals		404
6.	The Miracle of Purun Bhagat		466
7.	Undertakers	eleje	588
8.	The King's Ankus	kincsek	679
9.	Quiquern	eleje	711
		Angutivum Taina +	781

4.13. táblázat. Lewis Caroll ALICE történeteinek összehasonlítása

	angol	**magyar**
Hódító Vilmos száraz története	47	36
Advice from a Catterpillar vers	93	
Duchess versikéje	120	
nehézségek a flamingóval	171	
teknős tantárgyai	203	
Alice Adventure's in Wonderland végén Alice féláomban	266	198
Ykcowrebbaj és Jabberwocky (Róscaffurg a és Gruffacsór) vers	296	220
Tweedledum (Subidu) verse	370-381	275-285
Humpty Dumpty (Dingidungi) vers magyarázata	438	331
Lovag éneke	509	381
Alice királynő ünneplés, ebéd		406

4.14. táblázat. Lewis Caroll ALICE történeteiből azon blokkok sorszáma, amelyekben az újonnan bevezetett szóalakok száma szignifikáns eltérést mutatott. A rövid blokkok az angol szövegben is csak éppen elérték a szignifikancia küszöböt, míg a magyar szövegben ugyanezeken a helyeken nem vagy csak nagyon kicsi kiugrásokat találtunk ($2\sigma = 6{,}1486$).

blokk	**új szóalak**	**különbség**	**blokk**	**új szóalak**	**különbség**
47	18	**7,43**	295	20	**13,24**
			296	22	**19,64**
93	19	**6,7**	296	22	**13,3**
120	12	**6,43**	438	11	**7,06**
171	13	**6,95**	506	11	**6,83**
			507	12	**9,63**
203	12	**7,03**	508	14	**11,12**
			509	15	**12,45**
265	11	**7,26**	510	13	**7,53**
266	15	**8,95**			

Kifejezések és fogalmak jegyzéke

Az értekezésben előforduló leggyakrabban használt kifejezések, rövidítések és fogalmak jegyzéke

Fogalmak

érvényes karakterek	az érvényes karakterkészlet (felhasználó által definiált) tetszőleges karaktere
elválasztó karakterek	minden olyan karakter, amely nem eleme az érvényes karakterkészletnek
szövegszó	két elválasztó karakter közötti karaktersorozat
szóalak	egyedi szövegszó
szótári alak, lexéma	a szónak az a formája, amely szótári cím vagy alcímként jelenik meg
szótő	a szónak az inflexiók eltávolítása után kapott formája
lemmatizálás (szótövezés)	az a folyamat, amely során a szóalakokról eltávolítjuk az inflexiókat – ragokat, képzőket – és végeredményként a szótőt kapjuk
hapax legomena	egyszer előforduló szavak

Kifejezések

N	szövegszók száma a szövegben
h	a vizsgált intervallumok, blokkok hossza
$n = \left\lceil \frac{N}{h} \right\rceil$	a vizsgált intervallumok, blokkok száma
b_i, $i = 1, \ldots, n$	blokkok sorszáma

$f(b_i) = y_i$, $i = 1, \ldots, n$	az újonnan bevezetett szóalakok száma az *i*-edik blokkban
k, k = 'a', …, 'z'	karakterek száma a felhasználó által definiált karakter készletben
m_k = max('k…')	a szövegben a karakter-készlet *k*-adik karakterével kezdődő szóalakok száma
x_{ksi}	egy szóalak megjelenése egy adott számú blokkban *k*: a kezdőbetű sorszáma, *s*: a szó ábécébeli sorrendjének a száma az adott betűn belül, *i*: a blokk sorszáma, amelyikben megtalálható az adott szó
$f(j,N)$	*N* szövegszó hosszúságú szövegben a *j*-edik leggyakoribb szóalak számát jelöli
$f(1,N)$	*N* szövegszó hosszúságú szövegben a leggyakoribb szóalak számát jelöli
$frel(j,N) = \frac{f(j,N)}{N}$	*N* szövegszó hosszúságú szövegben, a *j*-edik szóalak relatív gyakorisága
$V(N) = \sum_{k='a'}^{'z'} m_k$	különböző szóalakok száma
E[$V(N)$]	a modell(ek) alapján előállított várható értéke a különböző szóalakok számának
$rep = \frac{N}{V(N)}$	szóalakok ismétlődési hányadosa
$V(r,N) = \sum_{i=1}^{V(N)} I_{[f(i,N)=r]}$ $I_{[\alpha]} = \begin{cases} 1, & ha\ \alpha \quad igaz \\ 0, & ha\ \alpha \quad hamis \end{cases}$	N szövegszó hosszúságú szövegben az *r*-szer előforduló szavak száma
$V(1,N)$	*N* szövegszó hosszúságú szövegben az egyszer előforduló szavak (hapax legomena) száma

E[$V(r,N)$]	a modell(ek) alapján előállított várható értéke az r-szer előforduló szóalakok számának
$Femp(j) = \sum_{i=1}^{j} frel(i,N)$	empirikus eloszlás függvény
$fp(b_i) = yp_i$	$f(b_i)$ simított görbéje
$fm(b_i) = ym_i$	a modell alapján előállított mesterséges szöveg újonnan bevezetett szóalakjait ábrázoló függvény
$F(b_i)=Y_i$	fm_k, k = 1,…, 100 függvények átlaga
$\Delta y_i = yp_i - Y_i$, $i = 1, \ldots, n$ $M = \frac{1}{n}\sum_{i=1}^{n} \Delta y_i$	a simított és az átlag függvény különbségének az átlaga
$\sigma = \sqrt{\frac{(y_i - M)^2}{n}}$	a különbségek szórása
yh_i	hapax legomena száma az i-edik blokkban
Mh	hipergeometrikus eloszlású hapax legomena várható értéke
SDh	hipergeometrikus eloszlású hapax legomena szórása
$hap(r,h) = \sum_{i=1}^{n} I_{[yh_i=r]}$	azoknak a blokkoknak a száma, amelyekben r darab hapax legomena van (h hosszúságú blokkok esetén)
$mh = \max(yh_i)$	blokkokban előforduló hapax legomena számának maximuma
$V(1,N) = \sum_{j=1}^{mh} hap(j,h)$	az egyszer előforduló szavak száma, azok blokkonkénti előfordulásai alapján

Tipográfiai konvenciók

Az egyszerűbb tájékozódás érdekében mellékelem a szövegben használt tipográfiai megoldások listáját, amelyeket az egyes szövegelemek kiemelésére használok.

Fejezetek: A fejezetek számozása római számokkal.

Alfejezetek: Minden alfejezet arab számot kap, de minden esetben öröklődik a magasabb szintű számozás is.

Ábrák: Két szám ponttal elválasztva. Első szám: a fejezet száma arab számmal; második szám: az ábra sorszáma a fejezeten belül.

Táblázatok: Két szám ponttal elválasztva. Első szám: a fejezet száma arab számmal; második szám: a táblázat sorszáma a fejezeten belül.

Képletek: Zárójelben két szám ponttal elválasztva. Első szám: a fejezet száma arab számmal; második szám: a képlet sorszáma a fejezeten belül.

További, a folyó szövegben megtalálható tipográfiai eszközök

Dőlt: példa szövegek (szavak, kifejezések, mondatok)

Kiskapitális: a feldolgozásra került irodalmi művek címe

Dőlt és kiskapitális: szoftverek neve