Alkalmazott Matematikai Lapok **24** (2007), 57-77.

DINAMIKUSAN KEZELHETŐ STATISZTIKAI MODELLEK IRODALMI MŰVEK SZÓALAKJAINAK VIZSGÁLATÁRA

CSERNOCH MÁRIA

Munkánk során arra vállalkoztunk, hogy a szavak véletlenszerű válogatásával egy olyan dinamikusan kezelhető statisztikai modellt építsünk, amely jó közelítéssel képes az újonnan megjelenő szóalakok természetes nyelvi szövegekben megfigyelt viselkedését visszaadni. Modellünk építéséhez az eredeti mű szóalakjainak gyakoriságát használtuk, tehát az így felállított modell segítségével előállított mesterséges szövegek szóalakjai ugyanolyan gyakoriságokkal rendelkeztek, mint értelmes megfelelőjük az eredeti szövegben. Három modellt is építettünk, amelyek közül az első a korábban ismertetett és statikus modellek megépítéséhez használt, a szavak polinomiális eloszlását feltételező elképzeléseket követte. Bár ezzel a modellel a korábbi vizsgálatokban elért pontosságot nem tudtuk javítani, sikerült azonban az újonnan megjelenő szóalakok számát leíró görbékre jellemző trendeket visszaadni. A második modellel, még mindig ezt az eloszlást feltételezve, az előzőnél már jobb közelítést sikerült elérni. A harmadik módszer, amely az eredeti szövegek legjobb közelítését adta, a szavak hipergeometrikus eloszlását feltételező modell volt. Ez utóbbi modell alkalmasnak bizonyult mind angol, mind magyar nyelvű szövegek modellezésére, amely mutatja, hogy az újonnan bevezetett szavak megjelenését nem befolyásolják egy nyelv grammatikai eszközei, a szintaktikai és szemantikai megkötések.

1. Bevezetés

A korábban szinte kizárólagosan alkalmazott szubjektív megítéléssel szemben, a statisztikai módszerek alkalmazása lehetővé teszi irodalmi művek számszerűsített (objektívebb) feldolgozását. A számítógép, illetve a számítógéppel segített szövegelemzés jelenti, ahogy sok más probléma esetén is, a szövegek korábban megoldhatatlannak tűnő vizsgálatát. A szóalakok, mint egy lehetséges minimális egység számának pontos ismeretében további olyan formulák határozhatók meg, amelyek képesek a szövegek egy-egy tulajdonságának a jellemzésére. 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 nyelvi gazdagságának leírására [11].

A számítógépes nyelvészet mozgatója a kezdetektől 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 a szavak szószerinti átírása nem adhat megfelelő kimenetet egy fordítási problémára [12]. 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 [8], [20], [21].

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.

Nyelvek és szövegek matematikai modellezéséhez is a gépi fordítások vizsgálata adott nagy lendületet. Kezdetben ezeket az eredményeket a titkosításban és a titkosítás megfejtésében (kódolás feltalálása), különösen a számítógépek biztosításánál, széles körben alkalmazták. Ennek elméleti kidolgozását C. Shannon amerikai matematikus végezte el [9]. Ezeknél a vizsgálatoknál az egységnek egy betűt (jelet) tekintenek.

Korszakalkotó jelentőségűnek mondható Markov modellje [1], [16], amely szintén egymást követő szimbólumok nem függetlenül történő kiválasztására adott algoritmust. Ezt az eljárást tovább módosítva napjainkban a Markov modell leginkább statisztikai alapon működő szófaj meghatározások (Part of Speech, POS) algoritmusaként használatos.

Szövegek teljes számítógépes feldolgozása egyelőre nem megoldott. A szövegek bizonyos tulajdonságait leírni képes részeredményekhez jutunk, ha egyszerűsítjük modelljeinket, pl. az általunk választott jellemző (paraméter) kiszámolásával. A szövegre jellemző bizonyos számszerű paraméterek vizsgálatára példa az a nyilvánvaló egyszerűsítés, hogy – szemben egy értelmes nyelvi szöveggel – a modellben a szavak egymástól függetlenül jelenjenek meg (randomness assumption). Ez annyit jelent, hogy figyelmen kívül hagyunk mindenféle szintaktikai, szemantikai és szövegszerkezeti megkötést [7].

Napjainkra számos olyan eredmény látott napvilágot, amely ezzel az egyszerűsítéssel él (ún. lexikai statisztikai modellek; összefoglaló értékelés [6]-ban található). Nyilvánvaló, hogy a szöveg visszaállítására a szavakat véletlen módon válogató modellek nem lehetnek alkalmasak, de nem is ez a céljuk. A véletlen válogatás természetes következménye ugyanis, hogy az említett vizsgálatoknál különbség van az eredeti „értelmes” szöveg és a modell között.

A korábban megjelent lexikai statisztikai modellek valamennyien statikus modellek voltak [6]. A szavak egymástól független megjelenését feltételezve, a szókészlet méretének és egy mű szógazdagságának jellemzésére zárt, matematikai képletekkel leírható megoldást kerestek. Ilyen képlet azt jelentette, hogy sikerült egy, a szöveg egészére jellemző, annak egy bizonyos tulajdonságát leíró paramétert (vagy paramétereket) találni. Ezek a modellek, következésképpen, nem adják vissza sem az eredeti szövegben jelenlévő trendeket, sem a szezonalitásokat.

A lexikai statisztikai modellek elsősorban a szókészlet nagyságára és gazdag-

ságára, valamint a szóalakok előfordulási gyakoriságára próbáltak meg összefüggéseket találni. A 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 ritka eseményeket leíró (Large Number of Rare Events (LNRE)) osztályba tartoznak [13]. Mivel az LNRE típusú eloszlások számítógépes modellezésére még kevés a sikeres és gyors algoritmus az elméletileg megszámolható eredményekkel végezhetünk összehasonlítást. A korábbi statikus modellek közül azok adták a legjobb közelítéseket, amelyek azt feltételezték, hogy egy szöveg szavai polinomiális eloszlást követnek. Ezek a modellek alkalmasnak bizonyultak arra, hogy vizsgálják a szavak nem-független megjelenésének forrásait. Segítségükkel, többek között, arra a következtetésre jutottak [4], [5], [6], hogy bár a mondaton belüli kötöttségek a legnyilvánvalóbbak, mégsem ezek a legfőbb forrásai a teljes szöveg szavai nem-véletlenszerű megjelenésének. Sokkal inkább meghatározóak a bekezdés vagy szövegszinten bekövetkező változások (ezekre viszont nincs matematikai modell).

Vizsgálataink hosszú távú célja főként angol és magyar nyelvű szövegek egy olyan speciális tulajdonságának kiszámítása, amelynek segítségével részben választ kaphatunk a fenti problémára. Arra keressük tehát a választ, hogy az írók mikor, a szöveg mely pontján találják indokoltnak olyan szavak bevezetését, amelyek korábban nem szerepeltek az adott műben. Ehhez szükség van egy olyan dinamikus vizsgálati modell kidolgozására, amely mind az angol, mind a magyar szövegekben az újonnan megjelenő szóalakok számának viselkedését a lehető legjobb közelítéssel képes visszaadni. Tekintettel arra, hogy a szavak számának pontos meghatározása nem volt célunk – azt vizsgáltuk csak, hogy mikor és mennyi új szó jelenik meg –, egyszerűsítésként megengedhető volt a statikus modellek megépítésénél is felhasznált randomness assumption alkalmazása.

Ugyanezen elméleti alapon olyan dinamikus modell megépítését tűztük ki célul, amely szemben a statikus modellekkel, az eredeti szövegben meglévő trendek és szezonalitások leírására is alkalmas lehet meghagyva a függetlenség feltételét. Angol szövegekre azért esett a választás, hogy eredményeinket össze tudjuk hasonlítani a korábban kapott, a szókészlet méretére vonatkozó, statikus modellek alapján kapott eredményekkel. Magyar szövegek ilyen jellegű számítógépes feldolgozására, tudomásunk szerint, ez idáig nem történtek kísérletek. Érdemesnek tűnt tehát megvizsgálni, hogy egy agglutináló nyelv [19], [15], [14], [20] esetén hogyan alkalmazhatóak a szavak függetlenségét feltételező modellek.

2. Módszerek

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 defi-

niálása, a szöveg szavakra bontása képezte. A feldolgozás első lépéseként definiálni kellett azt a karakterkészletet (ábécét), amellyel a program dolgozni fog, amely alapján el fogja dönteni, hogy a szöveg mely karaktersorozata tekinthető szónak. Mivel a szövegeken előfeldolgozást nem végeztünk, ezért vizsgálataink alapegysége a szóalak (két elválasztó karakter közötti összefüggő karakter sorozat) lesz.

2.1. Szövegek blokkokra tördelése

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

Vezessük be a következő jelöléseket:

$\boldsymbol{N}$	a szöveg (mű) hosszúsága; szavainak, a **szövegszó**knak a száma;
$\boldsymbol{V(N)}$	az N szövegszó hosszúságú szöveg különböző szavainak, a **szóalakok**nak a száma ($V(N) \leq N$);
$\boldsymbol{\omega_i}$	N szövegszó hosszúságú szöveg i-edik (leggyakoribb) szava;
$\boldsymbol{f(i,N)}$	N hosszúság esetén az ω_i szó gyakorisága;

az i-dik leggyakoribb ω_i szó $\{P(\omega_i) = p_i\}$ valószínűség eloszlása teljes, ha

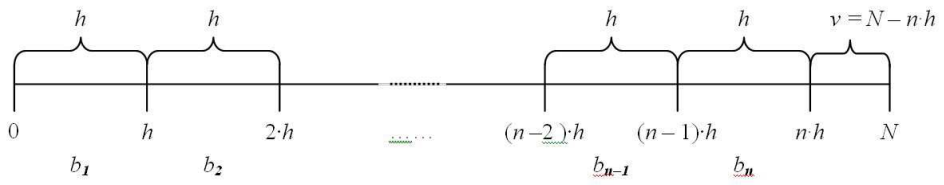
$$\sum_{i=1}^{V(N)} p_i = 1.$$

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

$\boldsymbol{b_i}$	blokkra bontjuk a szöveget, ahol minden blokk azonos számú szövegszót ($\boldsymbol{h}$) tartalmaz
$\boldsymbol{h}$	a blokkok hossza
$\boldsymbol{n}$	blokkok száma

$$b_i, i = 1, \ldots, n, \quad \text{ahol } n = \left[\frac{N}{h}\right]. \tag{1}$$

$$N \geq n \cdot h;\ N - n \cdot h = \nu. \tag{2}$$



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 (ν) nem kerülnek feldolgozásra.

$$\nu = N - h \cdot \left[\frac{N}{h}\right]. \tag{3}$$

$\nu = 0, 1, \ldots, h-1, \Pr(\nu_j = i) = \dfrac{1}{h}$, a „veszteség".

Az így bevezetett ν egy egyenletes eloszlású véletlen szám (valószínűségű változó) a $0 \leq \nu < h$ intervallumon [10], [17], [23]. Ennek megfelelően a feldolgozásra nem kerülő szavak száma, a szóveszteség várható értéke a lehetséges értékek számtani közepe: $\overline{\nu} = \dfrac{h}{2}$.

Regények esetén, ahol N, a szövegszók száma általában meghaladja 40 000-et és nem több, mint 400 000 (a feldolgozott művek közül egyedül Tolsztoj: Háború és Béke című műve tartalmazott több, mint 400 000 szövegszót) az átlagos relatív veszteség (ν_r)

$$\frac{\overline{\nu}}{400\,000} < \nu_r < \frac{\overline{\nu}}{40\,000},$$

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

2.2. Szavak tárolása 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 szóalakok számát y_i ($y_i, i = 1, \ldots, n$). Az y_i definíciójából következik, hogy bármely i-re

$$0 \leq y_i \leq h.$$

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, $k =$ '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 lettek azonosítva.

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

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

Az egyes bekezdések pedig legfeljebb n számú karakterből állhatnak (1-3). A bekezdéseken belül az egyes pozíciókon vagy a szóalak előfordulásának számát vagy annak hiányát jelöltük az adott sorszámú blokkban.

A különböző szóalakok tárolására egy hármas indexű elem ($X = \{x_{ksi}\}, k, s, i$) alkalmas, 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 (1. és 2. táblázat):

$$N = \sum_{k='\mathrm{a}'}^{'\mathrm{z}'} \sum_{s=1}^{m_k} \sum_{i=1}^{n} x_{ksi}.$$

1. táblázat. Az 'a' és 'b' karakterrel kezdődő szavak elrendezése. A fájlok első bekezdése (a tömb első sora) 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					
	1	**2**	**3**	**...**	**n**
1	x_{a11}	x_{a12}	x_{a13}		x_{a1n}
2	x_{a21}				
3					
...					
$\boldsymbol{m_a}$					

A 'b' karakterrel kezdődő szavak elrendezése					
	1	**2**	**3**	**...**	**n**
1	x_{b11}	x_{b12}	x_{b13}		x_{b1n}
2	x_{b21}				
3					
...					
$\boldsymbol{m_b}$					

2. táblázat. A szavak előfordulását tároló háromdimenziós tömb 'a' és 'b' kétdimenziós lapjai értékes jegyekkel feltöltve egy lehetséges minta alapján.

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
...					
$\boldsymbol{m_a}$	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
...					
$\boldsymbol{m_b}$	0	0	1	0	0

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. táblázat) a blokkonkénti (a táblázat oszlopai) T-k száma adja, ha összegezzük ezeket valamennyi karakterre.

$$N = \sum_{k='\mathrm{a}'}^{'\mathrm{z}'} \sum_{s=1}^{m_k} xt_{ksi}. \tag{4}$$

3. táblázat. 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 y_i értékek.

Az 'a' karakterrel kezdődő szavak első megjelenése					
	1	**2**	**3**	**...**	**n**
1	T				
2			T		
3	T				
...					
$\boldsymbol{m_a}$		T			

A 'b' karakterrel kezdődő szavak első megjelenése					
	1	**2**	**3**	**...**	**n**
1			T		
2	T				
3					T
...					
$\boldsymbol{m_b}$			T		

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: $1 \to$ A; $2 \to$ B; stb (1. ábra).

3. Eredmények

3.1. Az újonnan megjelenő szóalakok ábrázolása DyMoCASAT-tel

Kutatásaink elsődleges célja az volt, hogy angol és magyar nyelvű 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 éppen soron következő blokk, amit az addig vizsgált blokkok összességéhez hasonlítunk. Két ábrázolási módot is használtunk:

- az újonnan bevezetett szóalakok száma az adott blokkban (y_i) (2. A, 3. C, 3. D és 4. ábra),
- az addigi szóalakok száma (kumulatív szókészlet), a teljes szókészlet nagysága (Y_i) (2. B, 3. A, 3. B, 6., 7. és 8. ábra).

A szövegek újonnan bevezetésre kerülő szóalakjainak számát ábrázoló görbék (2. A, 3. C és a 4. ábra bal oldali görbéi) jól szemléltetik a tendenciát, miszerint a szövegben előre haladva csökken azoknak a szavaknak a száma, amelyek a szöveg egy későbbi pontján kerülnek bevezetésre. Az ábrák azonban azt is mutatják, hogy vannak a szövegnek olyan szeletei, amelyekben ez a csökkenő tendencia visszafordul, és váratlanul megnő az addig nem használt szavak száma. Az ábrákról az is leolvasható, hogy nem a szöveg hossza az, amely befolyásolja, hogy mennyi az újonnan bevezetett szavak száma, hanem az, hogy a szövegnek mely pontján járunk. A váratlan kiugrásoktól eltekintve igaz, hogy ha i, j a blokkok sorszámát jelöli és $i < j$, akkor $f(i) > f(j)$, valamint az is, hogy $f1(i) \sim f2(i)$ és $f1(j) \sim f2(j)$, ha $f1$ és $f2$ két azonos nyelven írott szöveg újonnan bevezetett szóalakjainak a számát mutatja.

Magyar nyelvű szövegek vizsgálatánál azt találtuk, hogy az újonnan bevezetett szóalakok száma magasabb az egyes blokkokban, mint azt angol szövegek esetén tapasztaltuk (3. D és a 4. ábra jobb oldali görbéi). Ez az eltérés a két nyelv sajátosságaiból következik. A magyar az agglutináló nyelvek csoportjába tartozik, míg az angolt, ha nem is egyértelműen, de leginkább az izoláló nyelvek csoportjába lehet sorolni. A blokkonkénti magasabb szóalakszámoknak egyenes következménye, hogy azonos hosszúságú angol és magyar szövegek esetén a magyar szövegek szókészlete, a szóalakok össz-száma magasabb, mint angol szövegek esetén (3. B, 8. A, 8. B ábra).

3.2. A szavak további feldolgozása

A fent ismertetett módszer – a szavak szövegfájlokban történő tárolása – további feldolgozásra is alkalmassá teszi a kapott értékeket. Ezen lehetőségek közül a következők a leggyakrabban használtak:

- a szóalakok számából meghatározható, 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 ábécé rendben,
- az egyes blokkok szövege,
- az egyes blokkokban újonnan megjelenő szóalakok, illetve
- az egyszer előforduló szavak (hapax legomena) listája blokkonként.

3.3. Dinamikusan kezelhető statisztikai 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.

Éppen ezért 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.

Egy, a szavak előfordulási gyakoriságán alapuló mesterséges szöveg létrehozásánál elsőként a szókészlet nagyságát célszerű meghatározni. Ez egy természetes elvárás, mivel az írók is ezt teszik, 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 szókészlete [18], [22]. 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 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 az urna modellt alapul vevő statikus modell [3], [4], [6] 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 vonatkozóan visszatevés nélküli válogatáson alapszik, így egy hipergeometrikus eloszlást eredményező dinamikus modell.

3.3.1. Visszatevéses válogatás ($P1$)

Ha $f(i, N)$ az ω_i gyakorisága az N szövegszó hosszúságú szövegben, akkor a szóalakok megjelenése modellezhető egy polinomiális eloszlással [17] a következőképpen.

Legyen $A_1, \ldots, A_{V(N)}$ egy teljes eseményrendszer, és

$$p_i = P(A_i) > 0, \quad i = 1, \ldots, V(N),$$

továbbá ismételjünk egy kísérletet N-szer $\left(\sum_{i=1}^{V(N)} p_i = 1\right)$ egymástól függetlenül. Jelölje ω_i az A_i esemény bekövetkezéseinek a számát. Ekkor $\omega_1, \ldots, \omega_{V(N)}$ együttes eloszlása N és $p_1, \ldots, p_{V(N)}$ paraméterű polinomiális eloszlás:

$$\omega_1 = k_1, \omega_2 = k_2, \ldots, \omega_{V(N)} = k_{V(N)} \quad k_1 + k_2 + \cdots + k_{V(N)} = N,$$

$$P\left\{\omega_1 = k_1, \omega_2 = k_2, \ldots, \omega_{V(N)-1} = k_{V(N)-1}, \omega_{V(N)} = k_{N-(k_1+\cdots+k_{M-1})}\right\} =$$
$$= \frac{N!}{k_1! \ldots k_{V(N)-1}!(N-k_{V(N)})!} p_1^{k_1} \ldots p_{V(N)-1}^{k_{V(N)-1}} p_{V(N)}^{N-\left(k_1+\cdots+k_{V(N)-1}\right)},$$
$$\sum \frac{N!}{k_1! \ldots k_{V(N)-1}! \left(N-k_{V(N)}\right)!} p_1^{k_1} \ldots p_{V(N)-1}^{k_{V(N)-1}} p_{V(N)}^{N-\left(k_1+\cdots+k_{V(N)-1}\right)} = 1$$

Esetünkben természetesen a kísérlet egy tetszőleges szó kiválasztása a szövegből. Ha egy szót megkülönböztetünk a többitől speciálisan a p_{i_1} paraméterű binomiális eloszlást [17] kapjuk:

$$P\left\{\omega_{i_1} = k_{i_1}, \omega_{i_2} + \cdots + \omega_{i_{V(N)-1}} = k_{N-\left(k_{i_2}+k_{i_3}+\cdots+k_{i_{V(N)-1}}\right)}\right\} =$$
$$= \binom{N}{k_i} p_{i_1}^{i_1} (1-p_{i_1})^{N-\left(k_{i_2}+\cdots+k_{i_{V(N)-1}}\right)}.$$

A modell megépítéséhez az eredeti mű szóalakjainak gyakoriságát használtuk fel. Ennek megfelelően először az egyes szavak gyakoriságát ($f(j,N)$; a j-edik szóalak gyakorisága az N szövegszót tartalmazó szövegben), majd a relatív gyakoriságát ($frel(j,N)$) határoztuk meg:

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

A szóalakok relatív gyakoriságának ismeretében meg tudtuk határozni az adott eloszláshoz tartozó empirikus eloszlásfüggvényt ($Femp$, szokás kumulatív empirikus eloszlásfüggvénynek is nevezni), ahol minden egyes szóalaknál a relatív gyakoriságok összege szerepel:

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

Ezen relatív gyakoriságok és a hozzájuk tartozó empirikus eloszlás függvény (5. ábra) 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 [2]. 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 6-8. ábrákon az eredeti szöveg szókészletének nagyságát ($V(N)$) a folyamatos, míg a polinomiális eloszlást feltételező

modellel előállított szöveg szókészletének nagyságát ($EP1V(N)$; 6. és 7. A ábrák) a szaggatott vonal jelöli.

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 szavak relatív gyakorisága ne változzék (6. és 7. A ábra).

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

$$\mathrm{rel}\left(V\left(1,N\right)\right)=\frac{V\left(1,N\right)}{N},$$

addig a módosított modellben az egyszer előforduló szavak relatív gyakorisága

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

kifejezéssel adható meg, ahol $V2$ a hozzáadott szóalakok száma.

A módosított modell alapján előállított szöveg szókészletének nagyságát ($EP2V(N)$) a 6. és 7. A 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 ([3], [4], [6]; 6. ábra). Az eredeti és a mesterséges szöveg közötti különbség csökkentésére ezért egy újabb modellt építettünk.

3.3.3. Visszatevés nélküli válogatás (H)

Ebben a modellben a szövegszókat egy vektor komponenseiként tároltuk, majd az így tárolt elemeket véletlenszerűen válogattuk, de ebben az esetben visszatevés nélkül. A már felhasznált szövegszó nem került vissza a vektorba miután 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.

Ha egy olyan urnát feltételezünk, amelyben N golyó (a szóalakok száma) – köztük M egyszínű (egy szóalak) – van, n-et találomra kihúzva (n elemű mintát

véve) éppen k adott színűt találunk azok közt [17]. Ezeket a valószínűségeket

$$P_k = \frac{\binom{n}{k}}{\binom{N}{n}\binom{N-M}{k}} = \frac{(n!)^2}{(n-k)\,!}\frac{(N-n)\,!}{N!}\frac{(N-M-k)\,!}{(N-M)\,!}$$

szolgáltatja.

A visszatevéses és visszatevés nélküli válogatással készült modellek alapján előállított mesterséges szövegek és az eredeti szöveg közötti eltéréseket a 7. és 8. ábrákon mutatjuk be. Figyeljük meg, hogy a visszatevés nélküli válogatás még a módosított ($P2$) polinomiális eloszláson alapuló modellnél is jobb közelítését adta az eredeti szövegeknek. Különösen szembetűnő ez a különbségi görbéken ($V(N) - EP2V(N)$, illetve $V(N) - EHV(N)$; 7. és 8. ábra belső görbék).

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. A 7. B és a 8. ábrák belső görbéi mutatják, hogy 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.

4. Eredmények összefoglalása

Kutatásaink során főként 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 szóalakok a gyakoriak. Ezzel szemben angol nyelvben a morfémák jelentős hányada önálló egységként, szóalakként jelenik meg. Ennek következménye, hogy angol és magyar nyelven írott szövegek szövegszóinak és szóalakjainak száma eltérő. Megegyező hosszúságú (N) angol és magyar nyelvű szövegeket összehasonlítva a felhasznált különböző szóalakok száma ($V(N)$), az egyszer előforduló szavak száma ($V(1, N)$) magyar nyelvű szövegekben nagyobb, aminek következménye, hogy az egyes szavak relatív gyakorisága kisebb a magyar nyelvű szövegekben.

Az angol és a magyar nyelv közötti nyilvánvaló eltérések ellenére egy olyan dinamikus modell létrehozását tűztük ki célul, amely alkalmas lehet bármelyik nyelven írt szöveg újonnan bevezetett szóalakjai viselkedésének leírására.

A modell megépítésénél azzal, a korábbi statikus modelleknél is használt feltételezéssel éltünk, hogy a szavak polinomiális eloszlást követnek egy szövegen belül. Azt találtuk, hogy az így létrehozott dinamikus modell közel olyan hibával dolgozik, mint a statikus modellek, de ugyanakkor képes visszaadni a szövegben jelenlévő trendeket is, amit a statikus modellek zárt formulái nem tudtak visszaadni.

A modell további fejlesztése során, amelyben a szavak a polinomiális eloszlása helyett azok hipergeometrikus eloszlását feltételeztük, sikerült előállítani egy olyan

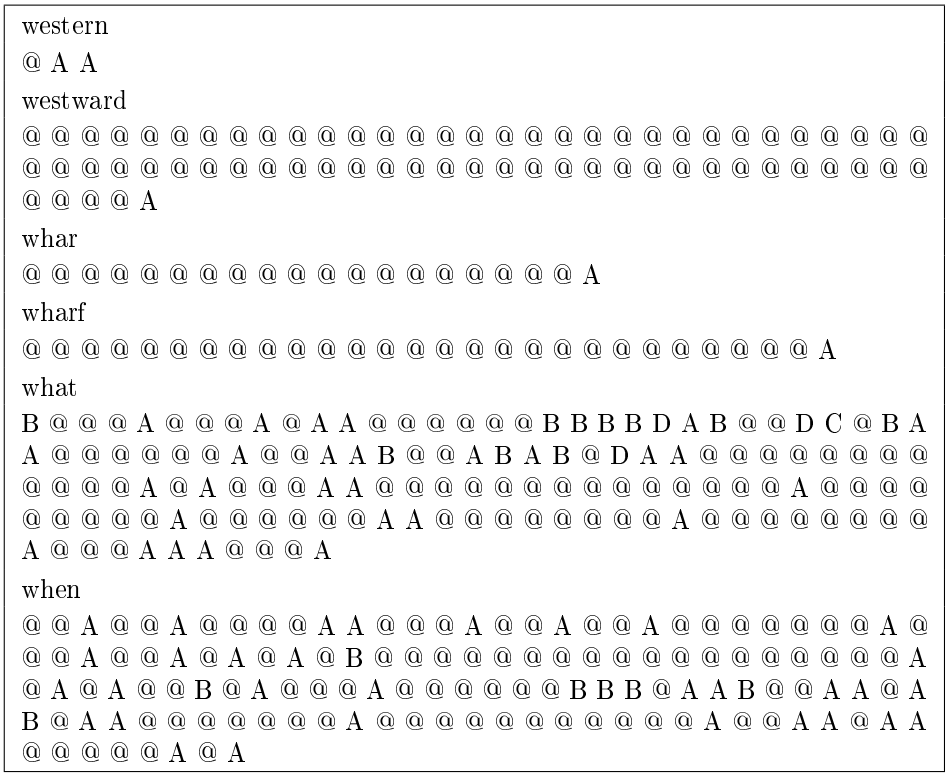
modellt, amely megőrizte az előzőnek azt a tulajdonságát, hogy a szöveg trendjeit visszaadja, ugyanakkor az eredeti szövegnek egy jobb közelítését kaptuk.

Az eredeti és a modell által generált mesterséges szöveget összehasonlítva azt találtuk, hogy az újonnan bevezetésre kerülő szóalakok viselkedésében nincs eltérés magyar és angol nyelvű szövegek esetén. Ez a megfigyelés nem mond ellent annak a hipotézisnek, hogy az eredeti és a mesterséges szöveg közötti eltérés nem mondat és bekezdés szintű, tehát nem szintaktikai és szemantikai kötöttségek miatt következik be, hanem szövegszerkezeti meggondolások következménye lehet.

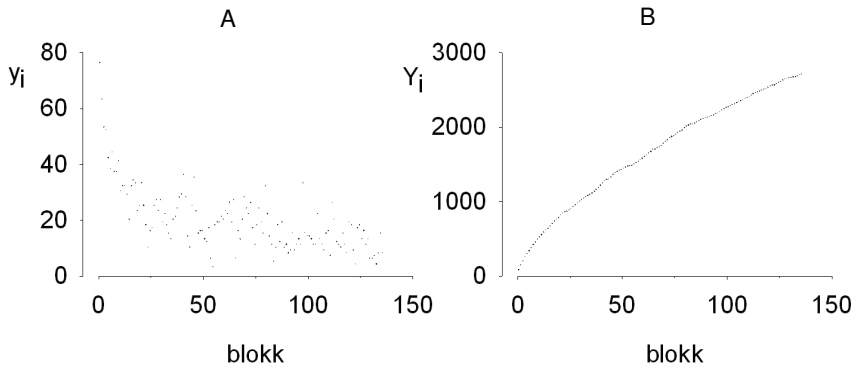
Az újonnan bevezetett szóalakok számának a modell alapján nem megjósolható hirtelen növekedése olyan szövegszerkezeti változásokra utal, ahol a szerző váratlanul szakít a szöveg addig megszokott folyásával. Ilyen jellegű szakadást, törést okozhat a szóalakok számának várható alakulásában egy-egy helyszín, szereplő, esemény részletes leírása, egy az eredeti történethez szervesen nem kapcsolódó szövegrész megjelenése, egy-egy, az előzőekhez képest új stílusú, esetleg idegen anyanyelvű szereplő megjelenése, hosszas beszéltetése.

További vizsgálatainkban elsődleges célként tűztük ki ezen, szövegszerkezeti szinten megjelenő, változások pontos leírását. Éppen a modellek és a természetes nyelvi szöveg közötti eltéréseket tudjuk arra felhasználni, hogy megtaláljuk az eredeti szöveg azon pontjait, intervallumait, amelyek szakítva a szóalakok megjelenésének várható alakulásával szeleteket emelnek ki az addig megszokott logikus szövegfolyamból.

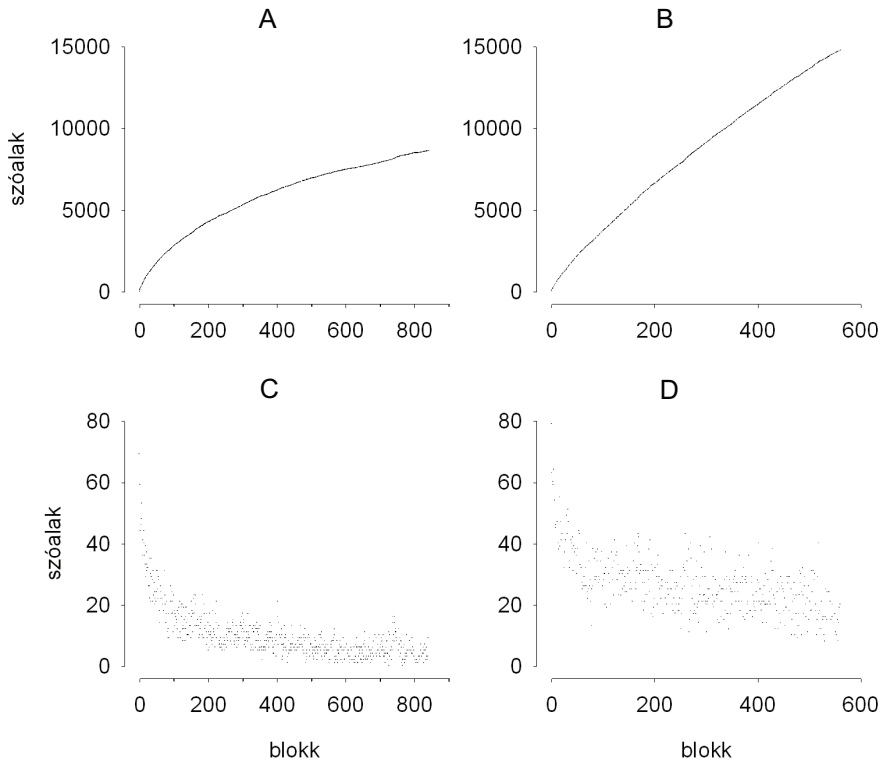
Ábrák gyűjteménye



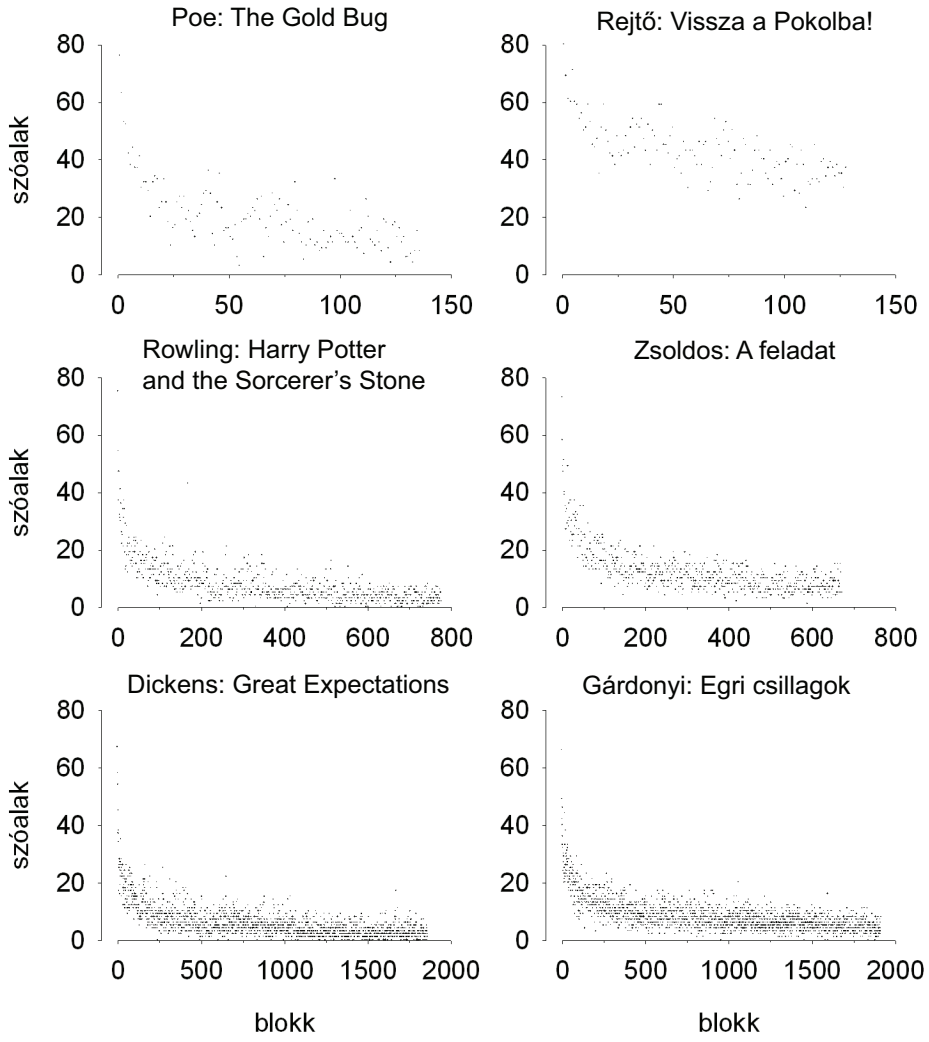
1. ábra. Edgar Allan Poe: The Gold Bug 'w'-vel kezdődő szavaiból részlet. A *western* szó a mű második és harmadik blokkjában szerepel egy-egy alkalommal, a *what* már az első blokkban megjelenik és kétszer is előfordul, majd legközelebb az ötödikben stb.



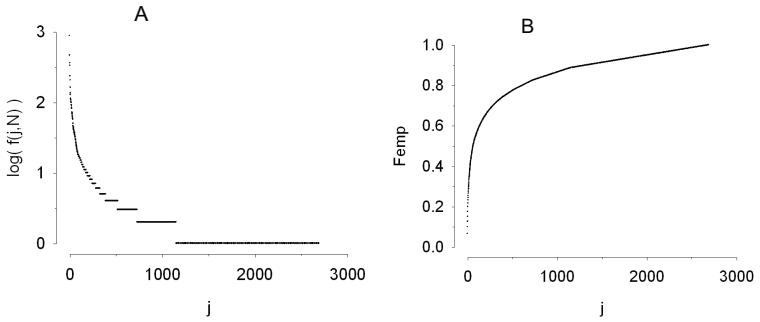
2. ábra. Edgar Allan Poe: The Gold Bug. A műben megjelenő különböző szóalakok száma a szöveg száz szavas blokkokra történő bontása esetén. Az első blokkhoz tartozó érték megadja, hogy hány különböző szóalak található a műnek ebben az intervallumában. Minden más blokkhoz tartozó érték azt mutatja, hogy az azt megelőző blokkokhoz képest hány új szóalak jelent meg (A). A szóalakok száma összegzésének eredménye egy monoton növekvő függvénnyel ábrázolható, mely megadja a mű szókészletének alakulását (B). Az első blokkhoz tartozó függvényérték megegyezik az A részen bemutatott függvény függvényértékével ebben a pontban, minden egyes további érték az azt megelőző függvényértékek összege.



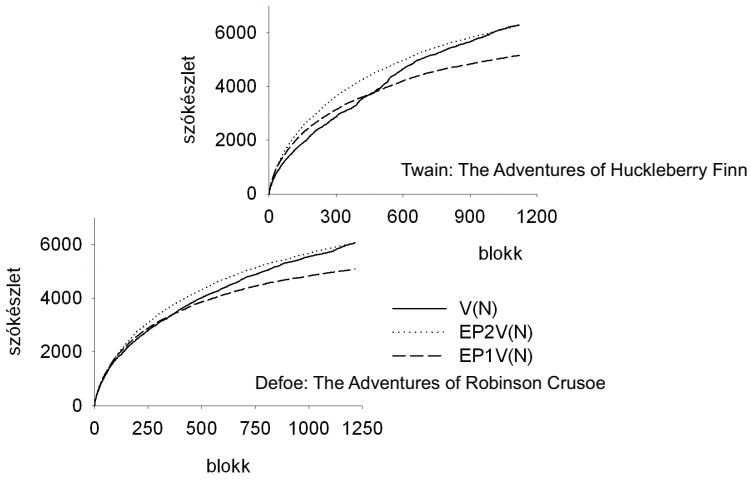
3. ábra. Szóalakok megjelenése és a szókészlet alakulása hasonló hosszúságú angol (Hawthorne: The Scarlet Letter; A és C) és magyar (Kertész Imre: Sorstalanság; B és D) nyelvű szépirodalmi művekben. Az alsó függvények (C és D) az újonnan bevezetett szóalakok számát mutatják az egyes blokkokban, míg az A és a B függvények ugyanezen művek szókészletének változását szemléltetik. 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.



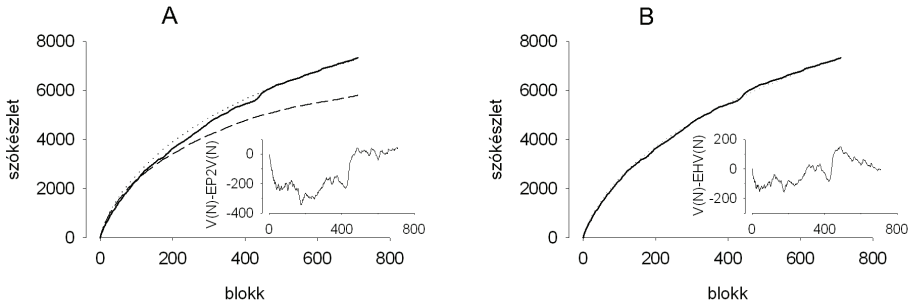
4. ábra. Szóalakok bevezetése angol (bal) és magyar (jobb) nyelvű irodalmi művekben. A szövegeket száz szövegszó hosszúságú blokkokra osztottuk. A grafikonok az egyes blokkokban újonnan bevezetett szóalakok számát mutatják különböző hosszúságú szövegek esetén. A felső sorban „rövid", kb. 15 000, a középső sorban „közepes" hosszúságú, kb. 80 000, míg az alsó sorban hosszú, kb. 150 000 szövegszót tartalmazó művek újonnan bevezetett szóalakjainak száma látható.



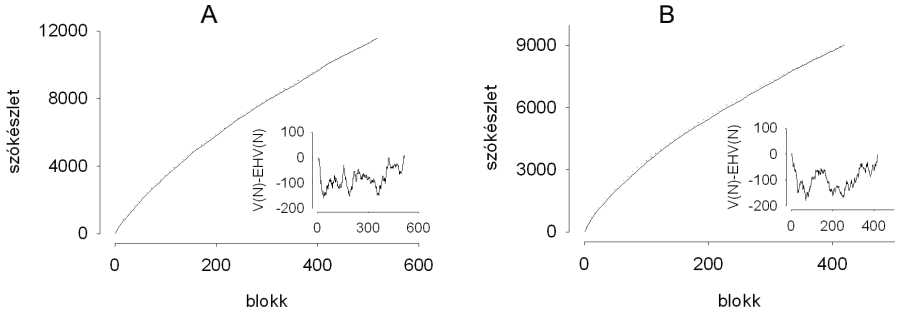
5. ábra. Szavak gyakorisági görbéje (A) és a szóalakok előfordulási gyakorisága alapján előállított empirikus eloszlás függvény (B) Edgar Allan Poe The Gold Bug című műve alapján. $f(j,N)$ a j-edik legnagyobb gyakoriságú szót jelöli, ahol N a szöveg hossza, $V(N)$ a szöveg különböző szóalakjainak a száma, és fennáll az $f(j,N) \geq f(j+1,N)$ egyenlőtlenség, bármely $j = 1,\ldots,V(N)$. Az (A) görbe vízszintes szakaszai jobbról balra haladva az egyszer, kétszer, stb. előforduló szavak számát mutatják logaritmikus skálán. Mint látható, az egyszer előforduló szóalakok száma igen magas ($V(1,N) = 1\,543$), ami az összes szóalak 57%-a ($V(1,N)/V(N) = 0,57$); a szavak eloszlása tehát az LNRE típusú eloszlások közé tartozik.



6. ábra. Két közepes méretű – Mark Twain: The Adventures of Huckleberry Finn (felső) és Daniel Defoe: The Adventures of Robinson Crusoe (alsó) – angol nyelvű regény szókészletének összehasonlítása. A folyamatos vonal az eredeti mű szókészletét ($V(N)$), a szaggatott vonal az eredeti visszatevéses modell alapján számolt szókészlet nagyságát ($EP1V(N)$), míg a pontozott vonal a módosított visszatevéses modell alapján számolt értékeket ($EP2V(N)$) mutatja.



7. ábra A szókészlet nagyságának alakulása Mark Twain The Adventures of Tom Sawyer című művében és a mű alapján polinomiális (A), illetve hipergeometrikus (B) eloszlást feltételező modellekkel számolt mesterséges szövegekben. Az eredeti mű szókészletét a folytonos vonal mutatja. Az eloszlásfüggvény alapján előállított modelleket használva az A ábrarészen a szaggatott vonal az első visszatevéses modell alapján ($EP1V(N)$), míg a pontozott vonal a módosított visszatevéses modellel ($EP2V(N)$) kapott mesterséges szöveg szóalakjainak számát mutatja. A B ábra pontozott vonala a visszatevés nélküli modellel számolt mesterséges szöveg szókészletét ($EHV(N)$) adja. A belső ábrák az eredeti és a mesterséges szöveg szókészletének nagysága közötti eltérést mutatják.



8. ábra. A szókészlet nagyságának alakulása két magyar nyelvű szövegben (Tamási Áron: Ábel a rengetegben; A és Molnár Ferenc: A Pál utcai fiúk; B). Az eredeti szöveg szókészletét folyamatos vonallal, míg a modell szókészletét pontozott görbével ábrázoltuk. A belső grafikonok az eredeti és a mesterséges szöveg szókészlete közötti eltérést mutatják.

Hivatkozások

[1] Arató, M. – Knuth, E.: *Sztochasztikus folyamatok elemei.* Tankönyvkiadó, Budapest (1970)

[2] Ashby, W. R.: *Bevezetés a kibernetikába.* Akadémiai Kiadó, Budapest, (1972)

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

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

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

[6] Baayen, R. H.: *Word Frequency Distributions.* Kluwer Academic Publishers, Dordrecht, Netherlands (2001)

[7] Balázs, J.: *A szöveg* Gondolat, Budapest (1985)

[8] Church, K. W. – Mercer, R. L.: *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 (1994)

[9] Demetrovics, J. – Denev, J. – Pavlov, R.: *A számítástudomány matematikai alapjai.* Nemzeti Tankönyvkiadó, Budapest (1985)

[10] Hajtman, B.: *Bevezetés a matematikai statisztikába.* Akadémiai Kiadó Budapest (1971)

[11] Holmes, D. I.: *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 (1994)

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

[13] Khmaladze, E. V.: *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 (1987)

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

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

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

[17] Meszéna, Gy. – Ziermann, M.: *Valószínűség elmélet és matematikai statisztika.* Közgazdasági és Jogi Könyvkiadó, Budapest (1981)

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

[19] O'Grady, W. – Dobrovolsky, M. – Aronoff, M.: *Contemporary Linguistics.* An Introduction: St. Martin's Press, New York (1993)

[20] Prószéky, G.: *Számítógépes nyelvészet.* Számítástechnika-Alkalmazási Vállalat, Budapest (1989)

[21] Prószéky, G. – Kis, B.: *Számítógéppel – emberi nyelven. Intelligens szövegkezelés számítógéppel.* SZAK Kiadó, Budapest (1999)

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

[23] Solt, Gy.: *Valószínűségszámítás.* Műszaki Könyvkiadó, Budapest (1971)

(Beérkezett: 2005. június 28.)

CSERNOCH MÁRIA
DEBRECENI EGYETEM
INFORMATIKAI KAR
4010 DEBRECEN, PF.: 12.
mcsernoch@hotmail.com

DYNAMIC MODELS FOR THE ANALYSIS OF THE INTRODUCTION OF WORD-TYPES IN LITERARY WORKS

Mária Csernoch

The aim of this work was to build a dynamic model which is able to reproduce the course of newly introduced word-types in literary works. Unlike previously published static models which provided constants at each running, this dynamic model created artificial texts, each of which is an approximation of the original. At each run, however, due to the random selection of words, these artificial texts were different. When building the model the frequency of the word-types in the original text was used, therefore, the frequency of the words in the artificial text was equal to that of the original. All together, three different models were built. The first was based on the same theoretical background as the static models, where the polynomial distribution of word was assumed. Though the accuracy of this dynamic model was not any better than that of the static models, it was able to reproduce the trends in the introduction of word-types in the given text. The second model was a minor modification of the first, with a better approximation of the total number of tokens in the original text. The third model, which gave the best approximation, used the assumption that the words follow a hypergeometric distribution in texts. This model proved to be language independent, that is, it was able to reproduce text written in English or in Hungarian regardless of their morphological productivity.