Irodalmi művek szókészletének statisztikai elemzése és matematikai modellezése

Statistical Analysis of the Introduction of Word types in Literary Works

doktori (PhD) értekezés tézisei

Csernoch László Józsefné

Debreceni Egyetem
Természettudományi Kar
Debrecen, 2005.

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 a pontos ismeretében további olyan formulák határozhatók meg, amelyek képesek a szövegek egy-egy tulajdonságának a jellemzésére. Lehet arról vitázni, hogy a nyers adatok/szóalakok mennyire alkalmasak irodalmi művek stilisztikai elemzéséhez, de úgy tűnik, hogy ezek statisztikai vizsgálatánál mostanáig nem sikerült megbízhatóbb módszert találni az irodalmi művek nyelvi gazdagságának leírására (Holmes, 1994).

A számítógépes nyelvészet mozgatója a kezdetektől a gépi fordítás (machine translation) megvalósítása 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 (IBM, 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 Mercel, 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.

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 (Demetrovics et al., 1985). Ezeknél a vizsgálatoknál az egységnek egy betűt (jelet) tekintenek.

Korszakalkotó jelentőségűnek mondható Markov modellje (Markov, 1916; Mandelbrot, 1962; Arató és Knuth, 1970), 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. A gépi fordítás többek között azért nem valósulhatott meg, mert nem tudjuk megmondani, hogy „mi a jó fordítás”.

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 (Balázs, 1985).

Napjainkra számos olyan eredmény látott napvilágot, amely ezzel az egyszerűsítéssel él (un. lexikai statisztikai modellek; összefoglaló értékelés Baayen 2001-ben 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 (Baayen, 2001). A szavak egymástól független megjelenését feltételezve, a szókészlet méretének és egy mű szógazdagságának jellemzésére zárt, matematikai képletekkel leírható megoldást kerestek. Ilyen képlet felállítása 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 gazdagságára, 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 eseményeket leíró LNRE (Large Number of Rare Events) osztályba tartoznak (Khmaladze, 1987). Mivel az LNRE típusú eloszlások számítógépes modellezésére még kevés a sikeres és gyors algoritmus az elméleti meggondolásokon nyugvó, számokkal kifejezhető 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 pl. arra a következtetésre jutottak (Baayen, 1996a; Baayen, 1996b; Baayen, 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 szövegszinten bekövetkező változások (ezekre viszont nincs matematikai modell).

2. 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.

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 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.

Angol szövegekre azért esett a választás, hogy eredményeinket össze tudjuk hasonlítani korábbi, 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 (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 szövegszinten 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 szövegszinten következik be, az egyszer előforduló szavak (hapax legomena) megjelenésének vizsgálatát is elvégeztük.

3. Módszerek

3.1. Szövegek feldolgozása

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, University of Virginia E-book Library, míg a magyar nyelvűek a Magyar Elektronikus Könyvtár, illetve a Neumann-ház elektronikus könyvtárakból kerültek letöltésre. Az elektronikus formában nem elérhető irodalmi művek elektronikus formára alakítását kézi szkenneléssel sikerült pótolni.

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 (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.

3.1.1. Szövegek blokkokra tördelése

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

Vezessük be a következő jelöléseket:

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

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

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

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

***b**i* blokkra bontjuk a szöveget, ahol minden blokk azonos számú szövegszót (*h*) tartalmaz;

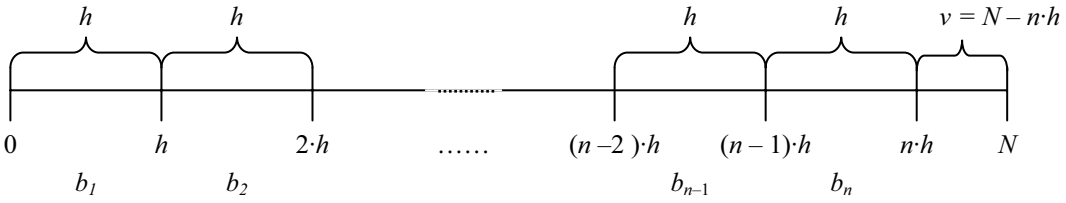
h a blokkok hossza;

n blokkok száma,

$$b_i,\ i = 1, \ldots, n, \text{ ahol } n = \left[\frac{N}{h} \right] . \tag{2}$$

$$N \geq n \cdot h;\ N - n \cdot h = \nu. \tag{3}$$

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.



3.1.2. Szavak tárolása az egyes blokkokban

A blokkok hosszúsága az esetek többségében 100 szövegszó hosszúságúra volt állítva, tehát *h* = 100. A végső cél az volt, hogy minden egyes 100 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

$$0 \leq y_i \leq h . \tag{4}$$

3.2. Modellépítés

Vizsgálataink elvégzéséhez három modellt építettünk. Mindhárom modell dinamikus, hiszen a szavak ténylegesen végrehajtott statisztikailag független válogatásán alapszik. Az így

definiált modellek alapján elő tudunk állítani tetszőleges számú mesterséges szöveget, amelyek a szöveg folyásának menetében összevethetők az eredeti szöveggel. Az első két modell 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 harmadik modellünk az egyes szóalakok (ω_i) visszatevés nélküli válogatásán alapszik, így egy hipergeometrikus eloszlást eredményező dinamikus modell.

3.2.1. Visszatevéses válogatás (P1)

Ha $f(i,N)$ az ω_i gyakorisága az N szövegszó hosszúságú szövegben, akkor a szóalakok megjelenése modellezhető egy polinomiális eloszlással (Meszéna és Ziermann, 1981) 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 ($\sum_{i=1}^{V(N)} p_i = 1$) 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:

$$\omega_1 = k_1, \omega_2 = k_2, \ldots, \omega_{V(N)} = k_{V(N)}, \quad k_1 + k_2 + \ldots + k_{V(N)} = N, \tag{5}$$

$$P\left\{\omega_1 = k_1, \omega_2 = k_2, \ldots, \omega_{V(N)-1} = k_{V(N)-1}, \omega_{V(N)} = k_{N-(k_1+\ldots+k_{V(N)-1})}\right\} = \tag{6}$$

$$= \frac{N!}{k_1!\cdots k_{V(N)-1}!\left(N-k_{V(N)}\right)!} p_1^{k_1}\cdots p_{V(N)-1}^{k_{V(N)-1}} p_{V(N)}^{N-\left(k_1+\cdots k_{V(N)-1}\right)},$$

$$\sum \frac{N!}{k_1!\cdots k_{V(N)-1}!\left(N-k_{V(N)}\right)!} p_1^{k_1}\cdots p_{V(N)-1}^{k_{V(N)-1}} p_{V(N)}^{N-\left(k_1+\cdots k_{V(N)-1}\right)} = 1. \tag{7}$$

Esetünkben természetesen a kísérlet egy tetszőleges szó kiválasztása a szövegből. Ha egy szót megkülönböztetünk a többitől speciálisan a p_{i_1} paraméterű binomiális eloszlást (Meszéna és Ziermann, 1981) kapjuk:

$$P\left\{\omega_{i_1} = k_{i_1}, \omega_{i_2} + \cdots + \omega_{i_{V(N)-1}} = k_{N-(k_{i_2}+k_{i_3}+\cdots+k_{i_{V(N)-1}})}\right\} = \binom{N}{k_1} p_{i_1}^{i_1} (1-p_{i_1})^{N-(k_{i_2}+\cdots+k_{i_{V(N)-1}})}. \tag{8}$$

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.

$$frel(j,N) = \frac{f(j,N)}{N}. \tag{9}$$

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

$$Femp(j) = \sum_{i=1}^{j} frel(i,N). \tag{10}$$

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.

3.2.2. 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 együttes relatív gyakorisága ne változzék.

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

$$\mathrm{rel}\big(V(1,N)\big)=\frac{V(1,N)}{N}, \tag{11}$$

addig a módosított modellben minden egyes egyszer előforduló szó relatív gyakorisága

$$\frac{1}{N\cdot\left(1+\dfrac{V2}{V(1,N)}\right)}=\frac{V(1,N)}{N\cdot\big(V(1,N)+V2\big)}, \tag{12}$$

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

Az eltérés az eredeti és a mesterséges szöveg között azonban nem lényegesen kisebb, mint a korábban használt statikus modellek esetén (Baayen, 1993; 1996a; 2001). Az eredeti és a mesterséges szöveg közötti különbség csökkentésére ezért egy újabb modellt építettünk.

3.2.3. 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 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.

Ha egy olyan urnát feltételezünk, amelyben N golyó (a szóalakok száma) – köztük M egyszínű (egy szóalak) – van, annak a valószínűségét, hogy n-et találomra kihúzva (n elemű mintát véve) éppen k adott színűt találunk azok közt a

$$P_k=\frac{\binom{n}{k}}{\binom{N}{n}\binom{N-M}{k}}=\frac{(n!)^2}{(n-k)!}\frac{(N-n)!}{N!}\frac{(N-M-k)!}{(N-M)!} \tag{13}$$

szolgáltatja (Meszéna és Ziermann, 1981).

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.

3.3. Szezonalitások meghatározása

Vizsgálatainkhoz tehát az eredeti művekben újonnan megjelenő szóalakok számát használtuk kiindulásként. Megszámoltuk, hogy 100 szövegszó hosszúságú blokkokban hány új szóalak (y_i, $i = 1, \ldots, n$) jelenik meg az előzőekhez képest és az így kapott értékeket ábrázoltuk. 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. 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 100 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.

Ezt a simított görbét hasonlítottuk a modell által előállított mesterséges szöveg szóalakjait leíró görbék sorozatához (f_k, $k = 1, \ldots, 100$), ahol f_{ki} jelöli a k. függvény i. blokkjában megjelenő szóalakok számát. A modell alapján előállítottunk 100 mesterséges szöveget, megszámoltuk ezen szövegekben az újonnan megjelenő szavak számát a 100 szövegszó hosszúságú blokkokban és vettük az így kapott függvények átlagát (F).

A következő lépésben vettük a simított függvény és az átlag függvény különbségét

$$fp - F\,, \tag{14}$$

$$\Delta y_i = fp_i - F_i,\ i = 1, \ldots, n, \tag{15}$$

majd a különbségek átlagát (M) és szórását (σ) (Hajtman, 1971; Nemetz és Kusolitsch, 1999; Solt, 1971; Yule, 1950).

Azokat az eltéréseket tekintettük szignifikánsnak, amelyek az átlagtól 2σ-val térnek el, tehát az $M \pm 2\sigma$ tartományon kívül esnek.

4. Eredmények és megbeszélés

4.1. Angol és magyar nyelvű irodalmi művek elemzése

A fentebb ismertetett módon megalkotott modellek alkalmasnak bizonyultak különböző hosszúságú angol és magyar nyelvű szövegek modellezésére. Ennek ismeretében a magyar és az angol nyelvű szövegek további feldolgozásánál nem volt szükség egyéni, csak az adott nyelv sajátosságait figyelembe vevő módszerek bevezetésére. Ez a megfigyelés nagyban megkönnyítette a különböző nyelveken írt szövegek összehasonlítását.

4.2. Különböző zsánerű művek elemzése

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) és egy magyar (Kertész Imre: SORSTALANSÁG) regényt, egy angol novellás kötetet, amelyben minden mű ugyanattól a szerzőtől származik (Rudyard

Kipling: THE JUNGLE BOOK) és egy olyan gyűjteményt, amelyben hasonló zsánerű művek szerepelnek, de különböző szerzőktől (AMERICAN MYSTERY STORIES).

A nagy kiugrás THE ADVENTURES OF TOM SAWYER-ben akkor jelenik meg, amikor az iskolaév végén a gyerekeknek egy házi dolgozatot kell írni és azt felolvasni, tehát egy olyan szövegrész jelenik meg a műben, amelyik nem tartozik szervesen a történethez és stílusában is eltér a könyv egészének stílusától.

A THE JUNGLE BOOK (Book 1) hét mesét és hét verset tartalmaz. Ezzel szemben öt olyan csúcsot találtunk, amely egyértelműen szignifikáns eltérésre utal és ebből az ötből is csak három esik egybe egy új mese kezdésével. Ez a három mese a White Seal, Rikki-Tikki-Tavi és Toomai of the Elephants. Mindhárom érdekessége, hogy új helyszínt vezet be a szerző, és ezzel magyarázható az újonnan bevezetésre kerülő szóalakok magas száma. A legelső kiugrás még a dzsungelben történik, de hasonlóan a már említett háromhoz, itt is új helyszínt vezet be az író, a királyi palotát írja le. Sem a többi mese kezdetén, sem a verseknél nem találtunk kiugrást, tehát nem jellemző, hogy hasonló zsánerű műveknél, egy új mű kezdeténél megemelkedne az újonnan bevezetett szóalakok száma.

Ezt támasztja alá a különböző szerzőktől származó AMERICAN MYSTERY STORIES gyűjtemény is. Vizsgálatunk három jól megkülönböztethető csúcsot eredményezett, lényegesen kevesebbet, mint amennyi a történetek száma. Ezen három csúcs közül is csak egyetlen egy esett össze egy történet kezdetével. Ez a csúcs Edgar Allan Poe THE GOLD-BUG című történetének kezdeténél jelent meg. Ez csúcs annyiban is érdekes, hogy az ezt megelőző történetnek is Poe a szerzője, tehát itt is látszik, hogy egy váltás a zsánerben még a szerzőt is felülmúlhatja 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, hirtelen valami új kerül ismertetésre, bevezetésre. Sem a szerző, sem az új fejezetek nem eredményeznek olyan látványos emelkedést, mint a zsáner, vagy a regiszter váltása. Ezek az eredmények már arra engedtek következtetni, hogy az új szavak nem-véletlenszerű bevezetése a szövegszinten bekövetkező változásokkal magyarázható. További vizsgálatokat tartottunk azonban fontosnak ahhoz, hogy ezen állításunk bizonyítást nyerjen.

4.3. Mű és fordításainak összehasonlítása

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. Így esett a választás

Kertész Imre SORSTALANSÁG című művére, amely angolul (FATELESS) is és németül is (ROMAN EINES SCHICKSALLOSEN) hozzáférhető, Rudyard Kipling THE JUNGLE BOOKS (Book 1 és Book 2) és Lewis Caroll ALICE'S ADVENTURES IN WONDERLAND és THROUGH THE LOOKING GLASS (ALICE) című műveire és ezek magyar fordítására (A DZSUNGEL KÖNYVE, ALICE CSODAORSZÁGBAN, ALICE TÜKÖRORSZÁBAN). A választás azért esett ezekre a művekre és fordításaikra, mert szerkezetükben lényegesen eltérő szövegekről és nyelvekről van szó. A nyelvek csoportosítását aszerint végezve, hogy a morfémákból a nyelv a szavakat hogyan képzi a kiválasztott három nyelv három különböző kategóriába sorolható. A német a flektáló, a magyar az agglutináló nyelvek csoportjába tartozik, míg az angol több különböző kategória eszközeit is felhasználja, így igazán egyikbe sem illik bele, de leginkább az izoláló nyelvekhez hasonlít (O'Grady, 1993; Prószéky, 1989; Quirk et al., 1995; Uzonyi, 1996; É. Kiss, 1998; Kiefer, 1998; Kugler, 2000; Laczkó, 2000).

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.

A 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. A fentebb ismertetett módszert alkalmazva az eredeti szövegek fordításaira megtalálhatjuk az eredeti szövegnek azokat az intervallumait, amelyekben az újonnan megjelenő szavak száma lényegesen magasabb, mint az a modell alapján várható lenne. A kérdés az volt, hogy mivel magyarázhatóak ezek a kiugrások, tehát a mi indokolja a különböző szóalakok szokatlanul magas számát és találunk-e olyan jellemzőjét a szövegnek, amellyel leírhatóak ezek a hirtelen változások. 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.

A kiugrások pontos helyének, a blokk sorszámának meghatározása után a *DYMOCASAT* (Csernoch, 2003; Csernoch és Hunyadi, 2003) segítségével megkaphatjuk azokat a k*100 szövegszó hosszúságú szövegrészeket, amelyekben ezek a kiugrások megjelentek. A szövegrészt ismerve vissza tudjuk azt keresni az eredeti műben, és magyarázatot tudunk adni arra, hogy miért növekedett meg hirtelen az újonnan bevezetett szavak száma.

Eredményeink azokkal az előzetes várakozásokkal, leginkább szubjektív véleményekkel egyeztek, amelyek a hosszabb lélegzetű, a műhöz szervesen nem kapcsolódó szövegrészeknél érzékelték a szóalakok számának emelkedését (Genette, 1980), szemben azokkal, akik fejezet határokra várták ezeket (Balázs, 1985). Különös tekintettel arra, hogy a fordításokban nem feltétlenül ugyanott vannak a fejezet határok, mint az eredeti szövegben vagy egy másik fordításban.

Hasonló eredményeket kaptunk a THE JUNGLE BOOKS (Book 1 és Book 2 együtt) és az ALICE történetek és ezen művek magyar fordításának elemzésénél is. Nem feltétlenül az újabb mese kezdetekor növekedett meg az újonnan bevezetett szóalakok száma, hanem sokkal inkább akkor, amikor egy hosszabb lélegzetű leírás jelent meg a műben. Ennek megfelelően egyes, nem a dzsungelben játszódó történetben (The White Seal, Rikki-Tikki-Tavi, Toomai of the Elephants, The Miracle of Purun Bhagat, Quiquern), mivel színhelyük és témájuk rendkívül változatos. A kiugrások minden esetben egy-egy részletes leírás eredményei. A dzsungelről szóló történetekben is találtunk két lényeges kiugrást, de egyiket sem az adott mese kezdeténél, hanem egyszer a királyi palota, míg a másik alkalommal a kincstár leírása okozta a szóalakok számának hirtelen emelkedését.

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.

A 3-4. ábrákon jól látható kiugrásokon túl 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.

5. Összefoglalás

Megalkottunk három dinamikus modellt, melyek alkalmasak irodalmi művekben és nyelvkönyvekben megjelenő szóalakok bevezetésének leírására. A három modellt összehasonlítva a szavak hipergeometrikus eloszlását feltételező és használó modellel előállított mesterséges szövegek adták az eredeti mű legjobb közelítését. Készítettünk egy Windows operációs rendszer alatt futó programot (*DyMoCASAT*), amely alkalmas ennek a speciális problémának az automatizált feldolgozására és kiértékelésére. A program az eredeti szöveg szóalakjainak gyakorisága ismeretében képes a megfelelő modell létrehozására, a modell alapján mesterséges szövegek előállítására, majd ezeket felhasználva az eredeti szövegek analizálására.

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.

A szövegek különböző nyelvű fordításait összehasonlítva az eredeti szöveggel, valamint az egyszer előforduló szavak megjelenését vizsgálva bizonyítást nyert, hogy ezek a hirtelen változások az újonnan bevezetésre kerülő szóalakok számában valóban nem szintaktikai, illetve szemantikai, hanem szöveg szinten jelennek meg.

1. Introduction

The statistical analysis of a literary text can be justified by the need to apply an objective methodology to works that for a long time have received only impressionistic and subjective treatment. Hesitation by literary scholars and mistrust of such a blatantly quantitative approach may be alleviated by choosing the least contestable mode of analysis, namely that of counting. The stylometrist therefore looks for a unit of counting that translates accurately the style of the text. The advent of computer has meant that data for this purpose are now readily available in the form of a concordance or word-index to a literary work. The choice of the number of different words (types) in a text as a counting unit allows the stylometric analyst the freedom of working on the raw data and of operating a lemmatization according to norms that he can define himself. This choice may run the risk of treating the individual written or printed word as unduly sacrosanct, yet, to date, no stylometrist has managed to establish a methodology that is better able to capture the style of a text than that based on lexical items (Holmes, 1994).

The machine translation and the need for some tool that would take over the tiresome job of translation was the starting point of computational linguistics. To the contrary of the original expectations it became clear to the late 1950s (IBM, 1959) that a word by word transcription cannot be acceptable as a translation, to the 1960s it was already seen that computers without human interaction would not be able to produce good translation for a long time to go (Prószéky, 1989; Church and Mercel, 1994; Prószéky and Kis, 1999). Not just because computers are not good enough but there is (are) no definition to state what is considered a good translation.

Reaching the millennium, computational linguistics became more and more multilingual and was not restricted to the English language any more and the need for machine translation was revived with the new generations of computers and the huge amount of official texts to translate. The fully automatic machine translation is still in waiting, but many problems arising, while searching for the ever mighty, were solved with the help of computers. The developments achieved during the search for fully automated machine translation also gave rise to modelling languages and texts. At the beginning these results were applied to cryptography and to coding and decoding messages. The theoretical background of this discipline is marked by C. Shannon's works (Demetrovics et al., 1985). Markov's (1916) revolutionary stochastic model was rediscovered and many computer aided applications of it

have been born. Nowadays Markov's model is mainly used in statistical Part of Speech (POS) taggers.

Since, the fully automatic processing of texts written in any natural language is not solved one of the possible approaches to the problems is to reduce the complexity of them. A kind of the simplification is to pick a feature of the texts and give explanation for this special question, problem. One of these reductions is to use the obvious simplification that words occur randomly in texts. Up till now several promising results have come to life which all used this simplification. These are mainly focusing on vocabulary size and richness and try to find formulae which are able to give reliable pieces of information about these characteristics of the texts.

Along with the randomness assumption another simplification had to be applied. The distribution of the words in a text belongs to the Large Number of Rare Events (LNRE) zone, but until now no really good and fast algorithm has been found to model this kind of distribution of words, so usually it is assumed that the words are multinomially, or, as a special case, binomially distributed.

Applying one of these models to real texts Baayen (Baayen, 1996a; Baayen, 1996b; Baayen, 2001) came to the conclusion that the randomness assumption is violated not on sentence level, but either on paragraph or discourse level. He also had the feeling that the constraints on discourse level might be responsible for the differences between the original and the expected vocabulary size and he also gave a vague explanation.

2. The aims of this study

The primary aim of this study was to gather information about the introduction of word types in literary works, novels and short stories. All this was carried out by using the theories and methods of computer-aided lexical statistical analyses. We mainly worked with English and Hungarian texts searching for explanations, reasons why, when, how many etc. relatively new words are introduced into the texts.

It was known from previously published studies that the analysis of word types on its own is not perfunctory for identifying the authors. Aware of this fact we were looking for parameters which can be gained from the appearance, not from the number of the word types.

Most of us have the sensation that reading a book becomes easier and easier as the story goes on, as we are heading towards its end, especially in the case of texts written in foreign languages. All this was proved by counting the words and following the changes in

vocabulary size. It also became evident after some polls that although a single measure of vocabulary richness that can characterize an author or a text as an attractive idea, reader's perceptions about vocabulary richness are not necessarily accurate (Hoover, 2003).

The number of the newly introduced word types in a text, as it was expected from the ever slowing rate of vocabulary increase, shows, in general, a monotonic decay. On the other hand, in most of the texts we can find intervals, parts of the texts where this monotonic decay is reversed and a sudden increase in the number of the newly appearing words can be detected.

In our study we wanted to give explanations to these sudden increases, aimed to find reasons why the authors use more words than previously to these slices of texts. We also wished to see whether these changes are predictable, if there is any regularity in their appearance or not.

To carry out our experiments we had to build a dynamic model which is able to give a good approximation of the original texts in its progress. The other constrain on the model was that it should be language free, that is it should be able to work with texts written in different languages. Our main goal was to analyze both English and Hungarian texts. English texts were chosen to obtain results that are directly comparable to previously published works while Hungarian to see how an agglutinating language can be modelled and get comparable, if there are any, results to texts in English and in any other languages. The emphasis was put on the dynamic characteristic of the model, which should be able to reproduce at least the trends but most preferably also the seasonalities of the original texts.

Previously published works showed clearly that even experts of the field do not share their opinions on when, at which place of the text new words show up, which is somewhat understandable since their opinions are mostly impressionistic. Some of them thought that the boundary of the chapters is where sudden increase in the number of the word types can be detected, while in other's opinion a change occurs when there is an interruption in the flow of the text, a text slice is inserted which differs in style from the text as a whole, e.g. a longish description appears unexpectedly.

Familiar with these opinions and also with Baayen's results and expectations we wanted to prove the hypothesis that the number of newly introduced word types increases when a sudden change can be detected in the flow of the text. These changes are relatively short, compared to the length of the whole text, but clearly separable. This statement can be rephrased: the differences between the original and the artificial texts, created by the dynamic

model, are due to changes at discourse level of the original text, neither the changes on sentence nor paragraph level cause measurable changes in the number of word types.

In our works beyond building the dynamic model we also applied a method hardly ever used in lexical statistical studies, the comparison of the original text and its translations. This method did not seem applicable in earlier works because they focused on the overall number of words, and texts with different vocabulary size cannot be, or are not easily, compared. The difference in vocabulary size is due to both the characteristics of the languages and the translator's freedom. Since our aim was not to count the words but to follow the changes of the words in progress, the previously considered problems did not cause difficulties or meant any obstacles to carry out our experiments.

To give further proof of our ideas two other methods were applied. First the appearance of the hapax legomena was also examined, modelled, and then their behaviour compared to the changes in the original text. Finally, the condensed or somehow flattened, shortened versions of the original works were examined and compared to the corresponding original text.

3. Methods

3.1. Data retrieval from texts

Our main concern was to analyze English and Hungarian literary works, restricted to novels and short stories. For the analyses we needed the electronic versions of the original, printed texts. The main source for these electronic versions was the Internet. The texts that were not available free through the Internet were scanned manually. It should be noted here that the availability of electronic versions greatly influenced the selection of works that were finally included into the present study.

To carry out the experiments a software, *DyMoCASAT* (Dynamic Models for Computer Aided Statistical Analysis of Texts) was developed. *DyMoCASAT* carries out the data retrieval from the original text, the building of the model, and based on the model the generation of the artificial texts.

DyMoCASAT has two character sets by default: English and Hungarian. (Any other character sets can be set up within the program offering access to texts written in other languages.)

Since our final goal was to gather information about the appearance and the behaviour of the word types in literary texts, the starting point of our experiments had to be the definition of words (word types). First, the character set, the alphabet, was determined upon which the program is able to decide which string is a word (type) and, based on this crucial information, were all the experiments carried out. Since pre/processing was not applied to any of the texts, the word type, a string of characters between two separator characters, was declared as the basic unit of the analysis.

3.2. Storing data

The analysis of the texts had to be preceded by saving all the available information about the number and the exact place of the word types. This all was carried out automatically by *DyMoCASAT*. In contrast to previously published works, we were to examine the appearance of the word types in progress. Since the number of the newly introduced word types is greatly influenced by the length of the intervals in question, intervals of different lengths could not be used. Considering all these, our model differs from those presented earlier since the texts are not divided into equal to same number of intervals independent of the length of the given text. Instead, we kept the lengths of the blocks constant (h). To use this novel approach a suitable constant for the length of the intervals had to be chosen.

Usually blocks containing one hundred tokens (h = 100) were chosen. Therefore, the number of blocks varies from text to text. Two advantages of these short blocks of constant length were found over the previously used method. First, since the length of a block is independent of the length of the original text, the slices from different texts can be readily compared. A shorter and a longer text divided into 20 or 40 equally spaced intervals – suggested and used in earlier published works – are not comparable considering either the number of tokens or the word types.

The second advantage of using hundred-token-long blocks comes from the relatively short length of these blocks. Using these short blocks subtle changes, couple of hundred-token-long text slices, in the narrative can also be traced.

The following variables are used by the program:

$\boldsymbol{N}$ the number of tokens in a text, the sample size,

$\boldsymbol{V(N)}$ the size of the vocabulary in an N-token-long text, the number of the different word types,

$\boldsymbol{\omega_i}$ the i^{th} word in a list of word types ordered by frequency,

f(i,N) the frequency of ω_i in a sample size of N token,

h the length of the intervals (blocks) into which the text is divided,

b_i the i^{th} block, by dividing the text into h-token-long blocks,

n the number of the blocks, using h-token-long blocks.

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

The method of dividing the texts into h-token-long slices always produce some loss, since the text at the end is truncated to $n \cdot h$ words.

The loss is minor, $\nu = N - h \cdot \left[\frac{N}{h}\right]$, compared to the size of the texts, so it will not influence the results of our experiments.

3.3. Building the models

Models based on the frequency of words assume that the words appear randomly within texts. There are, however, a number of strategies how random selections can be carried out (for review see Baayen, 2001). The best results were obtained with models that assume that word types follow the multinomial distribution, since multinomial distribution arises when each trial has k possible outcome. Selecting word types from a set of tokens is exactly the same problem, where the number of the possible outcome is $V(N)$, the number of the different word types in an N token long text.

If ω_i (i = 1, … , $V(N)$ mark the frequency of f(i,N), the i^{th} word type in the frequency order of an N token long text, then the appearances of the word types can be modelled with the multinomial distribution in the following way.

Let $A_1,\ldots,A_{V(N)}$ a random vector, a set of random variables, with $p_i = P(A_i) > 0,\quad i = 1,\ldots,V(N)$. If we assume that we have N independent trials ($\sum_{i=1}^{V(N)} p_i = 1$), and ω_i marks the number of the outcomes of the A_i event, then the $(\omega_1,\ldots,\omega_{V(N)})$ joint distribution is an N and $(p_1,\ldots,p_{V(N)})$ parametric multinomial distribution:

$$\omega_1 = k_1,\ \omega_2 = k_2,\ldots,\ \omega_{V(N)} = k_{V(N)},\quad k_1 + k_2 + \ldots + k_{V(N)} = N, \tag{2}$$

$$P\left\{\omega_1 = k_1,\ \omega_2 = k_2,\ldots,\ \omega_{V(N)-1} = k_{V(N)-1},\ \omega_{V(N)} = k_{N-(k_1+\ldots+k_{V(N)-1})}\right\} = \tag{3}$$

$$= \frac{N!}{k_1!\cdots k_{V(N)-1}!\left(N-k_{V(N)}\right)!} p_1^{k_1}\cdots p_{V(N)-1}^{k_{V(N)-1}} p_{V(N)}^{N-\left(k_1+\cdots k_{V(N)-1}\right)},$$

$$\sum \frac{N!}{k_1!\cdots k_{V(N)-1}!\left(N-k_{V(N)}\right)!} p_1^{k_1}\cdots p_{V(N)-1}^{k_{V(N)-1}} p_{V(N)}^{N-\left(k_1+\cdots k_{V(N)-1}\right)} = 1. \qquad (4)$$

In our case the trial is the selection of a word type from the text. If a word type selected and marked as different from the others the multinomial distribution is reduced to the binomial distribution. Each of the *k* components separately has a binomial distribution with parameters *N* and p_i, for the appropriate value of the subscript *i*:

$$P\left\{\omega_{i_1} = k_{i_1},\, \omega_{i_2} + \cdots + \omega_{i_{V(N)-1}} = k_{N-(k_{i_2}+k_{i_3}+\cdots+k_{i_{V(N)-1}})}\right\} = \binom{N}{k_1} p_{i_1}^{i_1} (1-p_{i_1})^{N-(k_{i_2}+\cdots+k_{i_{V(N)-1}})}. \qquad (5)$$

3.3.1. Selecting words with replacement (P1)

The model presented here also uses the frequencies of the word types ($f(i,N)$) of the original text, and their relative frequencies

$$frel(i,N) = \frac{f(i,N)}{N} \in \left]0;1\right[, \qquad (6)$$

thus the probability of occurrences (p_i). While previous works focused on the overall vocabulary size ($V(N)$) and richness, the given formulae were able to produce reliable pieces of information (for review see Baayen, 2001). However, our aim was not the determination of the vocabulary size, rather to find trends or trace seasonalities, if there are any, in the text flow. The previously given formulae are not able to provide information about a text in progress. Given these constraints new methods with new theoretical background had to be found.

The essence of our method is to create artificial texts using the frequencies and relative frequencies of the word types of the original text. Based on the relative frequencies of the word types a distribution function (*Femp*) is generated to each original text where each word type (ω_i) is represented with its own relative frequency ($frel(i,N)$).

$$Femp(j) = \sum_{i=1}^{j} frel(i,N), \quad j = 1,\ldots,S \qquad (7)$$

Randomly selecting numbers from the $]0,1[$ interval and mapping them to the word types through the distribution function allows the generation of randomly selected words which have the same probability of occurrence as in the original text. This random selection is repeated until the number of words in the model text reaches that of the original. With this simple method model texts can be generated in which the probability of a given word type equals that of the original text.

3.3.2. Selecting words with replacement, modified version (P2)

There is, however, a slight problem with the above algorithm. Since the word types are selected randomly, that is only their frequency is set, there is no guarantee that each and every word type will actually appear in the generated text. Indeed, running the program repeatedly gave, as expected, consistently smaller number of word types in the generated than in the original text. The discrepancy was the largest for words that appear only once (hapax legomena) in the original text. In order to correct this slight difference between the original and the generated text the algorithm was modified by artificially increasing the number of word types from which the random selection was carried out. In order not to change the frequency of all word types the following strategy was implemented. The number of hapax legomena was increased so that the relative frequency and the probability of each of them was decreased. This was carried out with the constraint that the overall relative frequency of hapax legomena should not be changed.

The relative frequency of all hapax legomena together in an N token long text is

$$\mathrm{rel}\left(V\left(1,N\right)\right)=\frac{V\left(1,N\right)}{N}. \tag{8}$$

Using this equation the relative frequency of a new word type becomes

$$x=\frac{V\left(1,N\right)}{N\left(V\left(1,N\right)+V2\right)}\ , \tag{9}$$

where $V2$ is the number of the newly added word types.

Applying the dynamic model the difference between the vocabulary size of the original text ($V(N)$) and the artificial text ($EP2V(N)$) is hardly smaller than it was measured using the static models. To further reduce the differences between the original and the artificial texts a third model was created.

3.3.3. Selecting the words without replacement (H)

For this model the tokens of the texts were stored in a one dimensional array. The tokens were randomly picked from this array, but after checking and saving their types they were not put back. Using this method for picking the words solved the previously present problem, namely, that not all of the types had been chosen.

If we consider an urn of N marbles which stand for the tokens of a text, among them are M of the same color, then the probability of selecting n marbles from this urn in a way that k share the same color is (Meszéna and Ziermann, 1981):

$$P_k = \frac{\binom{n}{k}}{\binom{N}{n}\binom{N-M}{k}} = \frac{(n!)^2}{(n-k)!}\frac{(N-n)!}{N!}\frac{(N-M-k)!}{(N-M)!} \tag{10}$$

To compare the two models based on the multinomial distribution of the words to the hipergeometric distribution we found that the artificial texts based on the latest gave the best approximation of the original text.

3.3.4. Trends in the appearance of the word types

After counting and storing all the occurrences of the words the program plots the number of newly introduced word types in each block ($f(b_i) = y_i$, $i = 1, \ldots, n$). The number of the newly introduced word types, in general, follows a decaying tendency. There are, however, parts of the texts where their number is greater than what is expected from this general trend. A point or a group of points that fall significantly outside of the general trend and form a local maximum within the neighbouring blocks is referred to as a protuberance. As mentioned earlier, the protuberances on the graphs of the newly introduced word types are visible only if h was defined appropriately.

It is clear that the number of the newly introduced word types follows a generally decaying tendency (as mentioned by e.g. Muller, 1964; Holmes, 1994, Baayen, 1996a, 1996b; for review see Baayen, 2001) with an appreciable amount of noise. For detailed comparisons it was necessary to reduce this noise. To this end a 7-point smoothing with a second order polynomial (Scarborough, 1966) and a Gaussian weighting function was used (*SIGMAPLOT*, SPSS Inc.).

Filtering the graph of the original text gave rise to a decaying function on which the smaller and larger secondary humps were now clearly visible. To decide which of these peaks

stand for significant changes and which are due only to the noise of how the author selected the word types the smoothed original graph (*fp*) and the average function of the artificial texts ($F(b_i) = Y_i$) were compared. The difference of the smoothed original and the average artificial text was determined and plotted (*fp*-*F*). To decide which values mark significant differences the mean (M) and the standard deviation (σ) of the difference of the two functions were calculated. Those differences are considered as significant which reach the $M \pm 2\sigma$ values.

4. Results

4.1. Literary works in English and Hungarian

The dynamic model created with the above detailed method was able to give account for the appearance of the word types not only in English but in Hungarian (and also in German) texts. It was found that the number of word types is indeed higher in Hungarian texts due to the morphologically productive characteristic of the language. The monotonic decay of the graphs of the newly introduced word types and the noise on the graph, however, follow the same pattern as found in English texts. Caused by the same feature the noise, as it was expected, was greater in the Hungarian texts. Furthermore, the randomness assumption was equally effective in describing the introduction of word types both in Hungarian (and German) texts as it was found for English texts. Texts of similar lengths regardless of the differences between the two languages did not show greater differences between the original and the artificial text than with the English text.

This result meant that for further investigation of this type there is no need for different models which have to be taught. Texts written in different languages can be analyzed with the same method, which greatly simplifies the comparison of these texts.

4.2. Comparison of the original and the artificial texts

The number of newly introduced word types has, as expected, a general tendency to decrease along the course of the narrative. On top of this decaying tendency the number of newly introduced types can, in many cases, be more at a later point in the discourse than in a previous section. These sudden changes cause smaller or greater protuberances on the graphs of the newly introduced word types.

In the observed works not only the intensity but the length of the protuberances are different, so shorter or longer rising phases were observed that interrupted the otherwise

declining function of the newly introduced word types. The subtle changes in the discourse were visible only if the intervals into which the texts were divided were short enough. The graphs, of course, show not only those subtle changes, marked by primary peaks, that were coming from the logical flow of the story but also those, where the text contains parts which are only slightly related to the events, causing secondary peaks. In both cases the monotonic decay of the graphs of the newly introduced word types was somewhat reversed.

The question arises whether the randomness assumption (Mandelbrot, 1962; Carroll, 1967; Sichel, 1986; Baayen, 1996a) is able to account for these changes in the course of graphs or not. As it was found, our dynamic model was able to simulate the feature marked by the primary peaks remarkably well. On the other hand, the model was incapable of reproducing the secondary peaks, which are, as mentioned earlier, totally unpredictable. The question was to give an explanation for the emergence of these secondary peaks. Fortunately, however, events for which the model was not able to give clear account can be traced back in the original text with the same model.

Relying on the readers' intuition the changes in the texts which produce the protuberances on the graphs of the newly introduced word types should mainly coincide with the launch of a new chapter. On contrary to these subjective opinions Genette (1980) gives a detailed, but still subjective analysis of several literary works concerning the changes and the results of these changes in the flow of the texts. The above listed reasons provided, in most cases, more significant changes in the number of the newly introduced word types than the introduction of a new chapter. His findings, concerning changes in lexis and grammatical markers, were supported by applying our method.

We found the model used on literary works, was able to follow the overall monotonic decay of the newly introduced word types. The size of the noise on the graphs of the artificial texts was also similar to that of the original text, if only the general noise is considered. These changes which were due to the flow of the story gave rise only to small peaks in the otherwise decaying graph. However, the trace of those surprising, unrelated events, understandably, never occurred in the model. We were interested whether, by comparing the original text and the related model, we can pinpoint parts which are only loosely related to the story, the pure narrative of the story.

Examining several English literary works with our method we were able to distinguish those places where significant differences are found between the original graph and the graph based on the model. These differences are due to differences between the original and the artificial text. The text slices corresponding to points above the threshold of significance were

searched back and saved by *DyMoCASAT*. Using these text slices we were able to tell precisely the reasons for the increase in the number of the newly introduced word types.

We selected texts of different genre, length, author, language to show that none of them influences the comparison of the original and the artificial texts. We have chosen for this comparison an English and a Hungarian novel (Mark Twain: THE ADVENTURES OF TOM SAWYER (from now on TOM SAWYER) and Kertész Imre: SORSTALANSÁG – FATELESS for which Kertész Imre was awarded the Nobel Prize in literature in 2002), and two collections of short stories. One of them is Rudyard Kipling: THE JUNGLE BOOK, and the other is a collection of AMERICAN MYSTERY STORIES from different authors.

Analyzing the protuberances of TOM SAWYER we found two which hardly reached and one that exceeded the threshold of significance. The first stands for a prayer, the second is a daydreaming, while the third one is large and corresponds to a section where a school year examination is detailed, for which each student had to write a short story or a poem. The first of these little writings raised the number of newly introduced word types significantly, and was kept high until all the little stories had been read. The style and vocabulary of this little piece of writing is clearly different from the rest of the novel, not Tom-sawyerish at all, while the first two is not as far off from the style of the whole text.

The JUNGLE BOOK contains seven tales and poems but only five peaks were identified on the difference trace. From these five three corresponded to the start of a new tale. These tales, White Seal, Rikki-Tikki-Tavi and Toomai of the Elephants introduce new sceneries, especially evident for the White Seal, explaining the sudden increase in the newly introduced word types. It should be noted that neither the other four tales nor the songs appeared as peaks on the trace. The first peak, which is the most easily separable from the start of a tale, is the most characteristic. It is a long description of the Kings' Palace, the setting did not change, we are still in the jungle but there is a change in the register, which produces this huge increase in the number of word types.

The concatenated AMERICAN MYSTERY STORIES gave three distinguishable peaks from which only one coincided with the start of a new story. All three correspond to a relatively long description. Similarly to the JUNGLE BOOK, it was found that the beginning of a new story does not necessarily give rise to the number of the word types, only in case when the story starts with a longish description of a setting. What was, however, remarkable that even the change in author did not produce protuberances. The one peak which coincides with the beginning of a story marks a long description at the beginning of E. A. Poe's THE GOLD-BUG. The interesting feature of this peak is that the story previous to this one is also from Poe,

which further strengthens our hypothesis that the genre overrides the author if the behavior of word types is in question.

These observations clearly establish that the secondary peaks observed on the graphs of newly introduced word types correspond to events, descriptions only loosely related to the discourse and, furthermore, even a text from a new author does not necessarily increases the number of newly introduced word types.

4.3. Analyzing the protuberances

As we have seen, protuberances on the graphs of the newly introduced word types are found at places where longish description of sceneries, events, characters, etc., or text-slices different in style from the main course of the text are found. Rereading these slices of texts, we can see that they do not carry any vital, non-omissible information, which would be absolutely necessary to understand the story. Such events occurred when a relatively short description was inserted into the text, when a new character with a new style (different from the style of the other characters) was introduced, and when not too long foreign expressions, sentences popped up.

These are "details functionally useless in the story" (Genette, 1980). Genette came to his conclusion by comparing Homer's Iliad with Plato's translation transforming it into pure narrative. He did not use any computers, he simple used the method of reading the two works side by side.

All these findings supported our hypothesis that the difference between the original and the artificial text is due to changes on discourse level in the original text. To give further proof of our hypothesis two more methods were applied.

One of them was the comparison of original texts and their translations, while the other the examination of the hapax legomena.

4.4. Comparison of the original text and its translations

To get comparable results we chose works whose translations were also available (at least in printed form). So, the following works and their translations were included in the forthcoming experiments. The original Hungarian work of Kertész Imre (SORSTALANSÁG, mentioned above), its English (FATELESS) and German (ROMAN EINES SCHICKSALLOSEN) translations were chosen. Two other works of English origin (Rudyard Kipling: THE JUNGLE BOOKS (Book 1 and Book 2) and Lewis Caroll: ALICE'S ADVENTURES IN WONDERLAND and

THROUGH THE LOOKING GLASS) and their Hungarian translations (A DZSUNGLE KÖNYVE and ALICE CSODAORSZÁGBAN és ALICE TÜKÖRORSZÁGBAN) were analyzed. The choice fell upon these works because they differ in their original language and also in their style and structure. The question was again whether the difference between the original and the artificial texts are due to constrains on discourse level or below it, either on sentence or paragraph level.

If the difference between the original and the artificial text had been caused by syntactic or semantic constraints the translations would have presented protuberances on the graphs of the newly introduced word types on totally arbitrary places. But this was not the case. The protuberances appeared almost exactly at the same parts of the stories regardless of the differences between the languages. Again, we could see that these protuberances mark changes on discourse level. Only loosely connecting details were found at these places.

The data gained by analyzing the texts show that there are great differences between the number of the word types, the number of the hapax legomena of the works and their translation, which are due to the differences between the languages and the translators' freedom. But, regardless of these differences, as it is shown, the protuberances appeared almost exactly at the same part of the texts, proving that applying the randomness assumption only those changes were not reproduced by the model which occur on discourse level.

4.5. The behaviour of the hapax legomena

If the places marked by the secondary peaks do not carry any vital information and they do not have further consequences in the story, the analysis of the hapax legomena has to cause protuberances at the same places.

Assuming that they are hipergeometrically distributed the theoretical distribution of hapax legomena was compared to their real distribution. Again, it was found that the protuberances occurred at those places which have been found rich in new word types by applying the two methods described above.

5. Summary

Based on a previously developed theoretical background, namely, that the vocabulary size and richness of literary works can be modelled using the randomness assumption, several models have been brought to life. The best results were obtained assuming that the selection of the words of the texts follow the hipergeometric distribution.

Applying this method we built a dynamic model which is able to imitate the text in progress, to give details about the appearance of the word types from the beginning to the last words of the texts, unlike the methods mentioned earlier, which try to describe the overall vocabulary size and the vocabulary richness.

Using our model, based on the frequency and the relative frequency of the word types, artificial texts can be created. To follow the narrative and to trace the behavior of the appearance of words, the number of the newly introduced word types were counted and plotted in both the original and artificial texts. As it was shown these artificial texts were able to follow the general trends of the original texts but not the seasonalities which produced protuberances on the graphs of the newly introduced word types. Analyzing the original texts, these protuberances were found to occur when the narrative is interrupted by a longish text slice which is different in style from the main stream. In previously published but much less objective works one can find indications which are in accordance with our findings but can also find merely subjective opinions which state that the number of the newly introduced word types rises at the beginning of a new chapter, or in case of concatenated short stories at the beginning of a new story.

As we have seen by comparing the original and the corresponding artificial texts those opinions were confirmed which state that the authors can deliberately change the flow of the narrative and then switch back to the original stream.

To give further proof of our hypothesis we applied two new methods, the comparison of the original works to their translations and the analysis of the appearance of the hapax legomena. Both additional methods were able to strengthen our results gained from the comparison of the original texts and their corresponding artificial texts that models assuming the independent usage of words in a text differ from the original text not because of the syntactic and semantic constrain but changes that occur on discourse level.

Hivatkozások jegyzéke

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

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

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, pp.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, H., Halteren, H. and 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

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.

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

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

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

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

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

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

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

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

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

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

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.

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

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

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

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

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

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

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

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

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

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

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

Tweedie, F. J. and 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. (1996) Rendszeres német nyelvtan AULA Kiadó Budapest, Hungary

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

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

Függelék

A tézisekben előforduló fontosabb fogalmak definíciói, az elnevezés után zárójelben az irodalomban használatos angol megfelelőt adtuk meg.

Érvényes karakterek: az érvényes karakterek készletét az adott nyelv ábécéje alapján hoztuk létre és alakítottuk úgy, hogy az alkalmas legyen a szövegek számítógépes feldolgozására. (Ennek megfelelően az angol karakterkészletnek része az aposztróf, míg a magyar karakterkészletben a többjegyű mássalhangzók nem szerepelnek külön karakterként.)
Elválasztó karakterek: minden olyan karakter, ami nem érvényes karakter. (Ezek a szóköz, az arab számok, a mondatvégi írásjelek, a vessző, az idézőjel, kötő- és gondolatjel, kettőspont, pontosvessző, a zárójelek és minden egyéb speciális karakter.)
Szövegszó (token): két elválasztó karakter közötti karaktersorozat. (A program ennek megfelelően két elválasztó karakter közötti karaktersorozatot tekint szövegszónak, nem téve különbséget a kis- és nagybetűk között.)
Szóalak (type, word type): egyedi szövegszó.