# Szótípusok bevezetésének szabályszerűsége magyar és angol nyelvű nyomtatott szövegekben

Csernoch Mária[1], Hunyadi László[2]

[1]Angol-Amerikai Intézet,
[2]Általános és Alkalmazott Nyelvészeti Tanszék
4010 Debrecen, Hungary
mcsernoch@hotmail.com

**Absztrakt.** Magyar és angol nyelvű irodalmi művekben, valamint angol nyelvkönyvekben vizsgáltuk a szótípusok megjelenésének szabályszerűségét. A szövegek egyenlő hosszúságú, rövid intervallumokra való darabolásával a szövegben bekövetkező apró változások is nyomon követhetővé váltak. Az eredeti mű szótípusainak gyakorisága alapján elkészült egy modell, egy mesterséges szöveg. Az eredeti és a mesterséges szöveget összehasonlítva meg tudjuk határozni azokat a szövegrészeket, amelyek kiugranak a történetből, nem épülnek be szorosan az események folyamatába. A nyelvkönyvek vizsgálata során azt tapasztaltuk, hogy egy nyelvkönyv szókészlete sokkal inkább véletlenszerűen összeválogatott, majd összefűzött novellák szókészletéhez hasonlít. Megfigyeléseink szerint a nyelvkönyvek tervezésénél a szerzők figyelmen kívül hagyják, hogy nemcsak a szótípusok számát kellene meghatározni és magasan tartani, hanem az egyes típusok megfelelő számú ismétléséről is gondoskodni kellene.

## Bevezetés

Szavak gyakoriságán alapuló modellek feltételezik, hogy a szavak véletlenszerűen jelennek meg egy műben. Azonban a véletlenszerű válogatások is több különböző stratégia alapján végezhetőek el [1], [2], [3]. A legjobb modelleket azok a kísérletek adták, amelyek feltételezték, hogy a típusok binomiális eloszlást követnek. A típusok binomiális eloszlását feltételezve Baayen korábbi [3], majd Hoover által megismételt [4] modellje minden m-szer előforduló szótípushoz kiszámított egy konstanst, amely az m-szer előforduló szótípusok várható száma lesz a szöveg tetszőleges pontján. Elvégezve az összegzést a szövegben előforduló valamennyi szótípusra, meghatározható az adott szövegrészben előforduló szótípusok száma. Az így előállított modell, következésképpen statikus. Egy adott M ($M \leq N$, ahol N = a szövegben előforduló szavak száma) esetén mindig azonos. A nagy teljesítményű személyi számítógépek megjelenésével azonban ma már lehetőség van dinamikus modellek építésére is.
A kísérletek fő célja az volt, hogy egy ilyen dinamikus modell megalkotásával meg tudjuk mutatni, milyen okokkal magyarázható egy új szótípus megjelenése, melyek azok a források, amelyek az írót egy új szótípus bevezetésére ösztönzik.

## Módszerek

Baayen modelljében a szövegek 40 egyenlő hosszúságú intervallumra darabolódnak, s így 40 különböző mérési pontban lehet elvégezni a számolást és ábrázolni a kapott eredményeket [2], [3]. Az itt bemutatásra kerülő modell szintén az eredeti szöveg típusainak gyakoriságát használja kiindulási pontként, de folytatásként a típusok relatív gyakorisága és előfordulási valószínűsége kerül kiszámolásra, majd ezen értékek ismeretében előállítunk egy eloszlásfüggvényt. Az eloszlásfüggvény értékkészletének elemeit véletlenszerűen válogatjuk, majd a függvény alapján visszakereshető az értékkészletnek az az eleme, amelyhez a véletlenszerűen választott szám hozzá lett rendelve. A véletlen szám válogatását mindaddig ismételjük, amíg el nem érjük az eredeti szószámot. Ezzel az eljárással elő tudunk állítani mesterséges szövegeket, amelyek az eredeti mű típusainak gyakoriságából származtathatóak.
Modellünk abban is eltér az előzőektől, hogy a szövegeket nem konstans számú darabra osztja függetlenül a szöveg hosszától, hanem a blokkok hossza lesz állandó. Általában 100 szó hosszúságú blokkokat használtunk, és ennek megfelelően a blokkok száma változó volt. Az így elvégzett darabolásnak két előnyét is találtuk a régiekkel szemben. Az egyik, hogy a blokkok hosszúsága független a szöveg hosszától, így a különböző hosszúságú szövegek darabkái sokkal inkább összevethetőek, mint különböző hosszúságú blokkok esetén. A másik előny a rövidre választott blokkhosszúságból ered. Rövid blokkokat használva a szövegben bekövetkező apró változások is nyomon követhetőek (1. ábra).

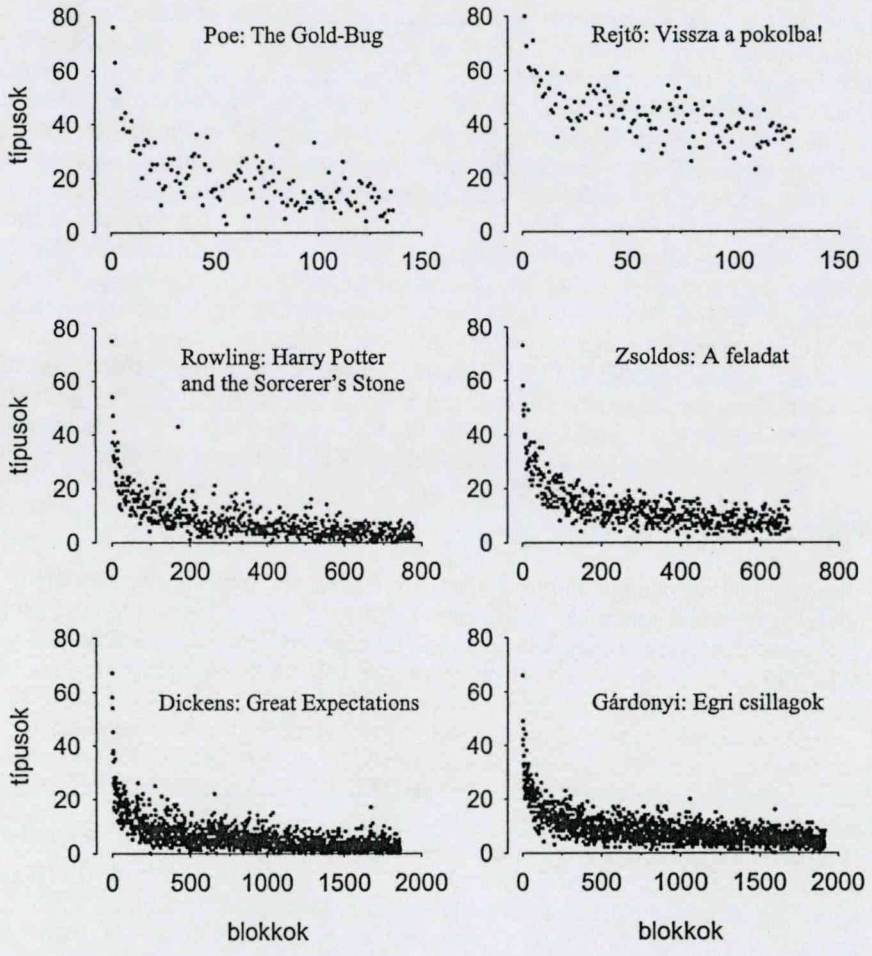
## Felhasznált anyagok

A rövid blokkra bontott szöveg ismét felveti azt a kérdést, hogy a típusok gyakorisága, mint paraméter alkalmas-e a szerző azonosításra. Ahhoz, hogy összehasonlítható eredményeket kapjunk, a szövegek kiválasztása az alábbi stratégia alapján történt: egy szerző több műve, egyszerzős sorozat kötetei, egyazon műfajhoz sorolható művek különböző szerzőktől, összefűzött novellák egy illetve több szerzőtől, egynyelvű nyelvkönyvek. A program alapértelmezés szerint angol és magyar nyelvű szövegek feldolgozására alkalmas, de a felhasználónak lehetősége van saját karakterkészletének beállítására, így további, más nyelvű művek feldolgozására is alkalmassá tehető.
A szövegek feldolgozásához az eredeti, nyomtatott szövegek elektronikus verziójára volt szükség. Az elektronikus szövegek fő forrása az Internet volt, az Interneten ingyenesen nem elérhető szövegek pótlása kézi szkenneléssel történt. A különböző forrásból származó szövegek egységesítését, szabványosítását a szövegek feldolgozása előtt meg kellett oldani. Itt szeretnénk megjegyezni, hogy a feldolgozandó szövegek elérhetősége is nagyban befolyásolta, hogy melyeken végeztük el kísérleteinket.

## Eredmények

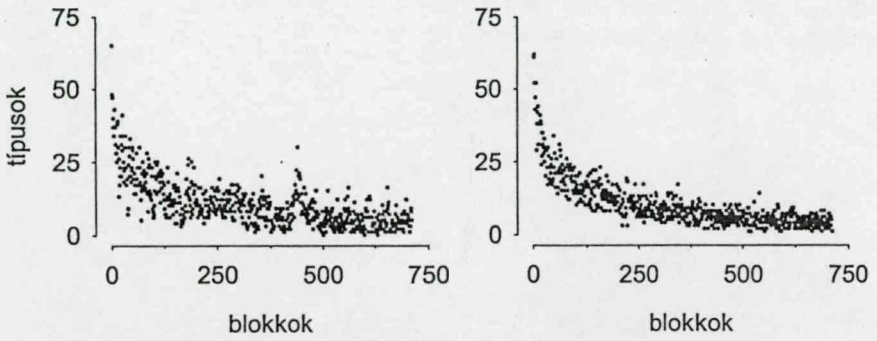
A program a vizsgált szövegeket többféle szempont alapján is elemzi. A gyakoriságok szolgáltak a modell kiindulási értékeiként. A gyakoriságok ezen túlmenően

alkalmasak arra, hogy összevessük őket nagy korpuszon végzett gyakorisági vizsgálatokkal. Ennek az összehasonlításnak igazán nagy jelentősége a nyelvkönyvek szókészletének vizsgálatánál van. A nyelvkönyvek szókészletéről valóban reális képet azonban akkor kapnánk, ha rendelkezésünkre állna egy szókészlet minimum és ezzel lehetne összehasonlítani a program által kiszámolt értékeket.




**1. ábra.** Angol (balra) és magyar (jobbra) szövegek szótípusainak megjelenése különböző hosszúságú szövegek esetén. A szövegeket 100 szó hosszúságú blokkokra daraboltuk. Az ábrák az újonnan megjelenő típusok számát mutatják. Fent „rövid", kb. 15000; középen „közepes", kb. 80000; lent „hosszú", kb. 200000 szó hosszú szövegek elemzésének eredménye látható

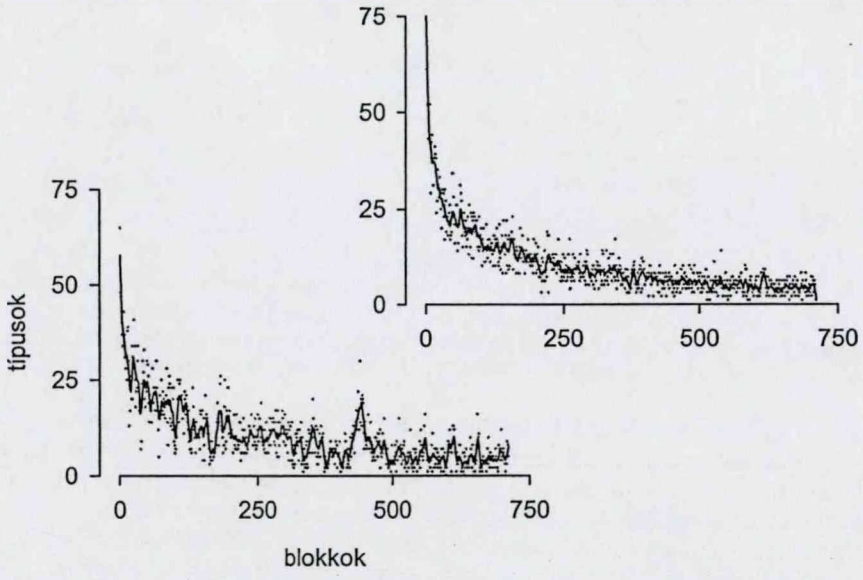
A program ábrázolja a blokkban megjelenő új szavak számát. Látványosan olyan pontok ugranak ki a függvény menetében, amelyek a történethez szervesen nem kötődő esemény bekövetkezésére utalnak.




**2. ábra.** Twain: The Adventures of Tom Sawyer. A bal oldali ábra az eredeti mű alapján készült és a megfelelő szótípusok számát mutatja 100 szó hosszúságú blokkok esetén. A jobb oldali ábrán az eredeti mű típusainak gyakorisága alapján készült modell látható. A modell nem követi azokat a változásokat, amelyek a történethez szervesen nem kapcsolódnak

A rövid, egyenlő hosszúságú intervallumok alkalmasnak bizonyultak arra, hogy a szöveg apró változásainak a jelei is megfigyelhetőek legyenek a grafikonon. Ilyen jellegű változások következhetnek be, amikor például egy új szereplő, helyszín, esemény hosszas bemutatása szakítja meg a történet folyamát, egy olyan új szereplőt beszéltet az író, akinek a stílusa nagyban eltér a korábban megszólaltatott szereplőkétől, idegen kifejezések, mondatok keverednek az egynyelvű szövegbe.
A program által generált modell vissza tudta adni azokat az apró változásokat, amelyek a történet logikus következményei voltak. Ezzel szemben azok az események, amelyek nem illettek a szövegbe, vagy nem tartoztak szervesen a történethez, nem jelentek meg a modellben (3. ábra).
Azt is sikerült megmutatni, hogy a szöveg hosszának fontos szerepe van a típusok megjelenésének szabályszerűségében. Korábbi írások már említették a szövegek hosszának meghatározó jellegét [5]. A megrajzolt függvények alapján állíthatjuk, hogy a típusok megjelenése egyenletesebb novellák esetén, mint regényekben, valamint azt, hogy a bevezetésre kerülő új típusok száma novellák esetén még a szöveg végén is magas (1. ábra). Ezzel magyarázható, hogy hasonló műfajú összefűzött novellák esetén sem találtunk nagy ugrásokat az összefűzési pontoknál még különböző szerzők esetén sem. Az összefűzött novellák viselkedéséből az is látható, hogy a típusok bevezetésének szabályszerűsége nem annyira a szerző, mint inkább a műfaj jellemzője, hasonló eredmények más modellek alkalmazása során is születtek már [6].
Nemcsak irodalmi művek, hanem egynyelvű nyelvkönyvek is feldolgozásra kerültek. A nyelvkönyvek megválasztásának szempontjai már tekintélyes részét képezik a nyelvoktatás módszertani irodalmának [7], de sok-sok szempont között az utolsók között kullognak azok, amelyek a szókészlet tudatos megválasztásáról szólnak. Azt vallják, hogy tanítsanak a nyelvtanárok annyi szót, amennyit csak lehet, de kurzusonként (120-150 óra) legalább 1000 szót, és amennyiben ez lehetséges, a leggyakoribb és „leghasznosabb" szavakat. Az általunk választott nyelvkönyvsorozat elemzése azt mutatta, hogy egy nyelvkönyv szótípusainak megjelenése nem tér el

lényegesen az összefűzött novelláknál megfigyeltektől (1. táblázat). Másik, figyelemre méltó eredmény pedig, hogy rendkívül magas azoknak a típusoknak a száma, amelyek csak egyszer szerepelnek a nyelvkönyvben (2. táblázat).




**3. ábra.** Twain: The Adventures of Tom Sawyer. A bal oldali ábra pontokkal az eredeti könyv típusainak megjelenését mutatja. Folyamatos vonallal egy 5-pontos simítás eredményét jelöltük. A jobb oldali ábra a modell típusait és a simítás eredményeit szemlélteti. A simított függvényen jól láthatóak azok a szövegrészek, amelyek kiugranak a történet logikus menetéből

**1. táblázat.** Összefűzött novellák és egy nyelvkönyv szókészletére vonatkozó adatok

| Szerző, cím | Blokkok | Típusok | Hapax legomena |
|---|---|---|---|
| Kipling: The Jungle Books | 516 | 4688 | 2067 |
| Soars: Headway Intermediate | 500 | 4803 | 2072 |

**2. táblázat.** A New Headway nyelvkönyvsorozat köteteinek összehasonlító elemzése

| New Headway | Blokkok | Típusok | Hapax legomena |
|---|---|---|---|
| Beginner | 163 | 1539 | 501 |
| Beginner→Elementary | 402 | 2943 | 962 |
| Beginner→Pre-Intermediate | 719 | 4550 | 1628 |
| Beginner→Intermediate | 1220 | 6760 | 2607 |
| Beginner→Upper-Intermediate | 1731 | 8989 | 3458 |

A program magyar nyelvű szövegek elemzését is elvégzi. Angol nyelvű szövegekhez hasonlóan, a típusok megjelenésének szabályszerűsége nem alkalmas arra, hogy azt

szerző azonosításhoz használjuk. Az 1. ábra grafikonjai azonban egyértelműen szemléltetik, hogy a magyar szövegekben a típusok száma magasabb, mint angol szövegekben. Ez azonban nem feltétlenül a magasabb számú szókészlettel magyarázható, hanem a magyar nyelv szóképzési és ragozási szabályaiból következik. Ahhoz, hogy a két nyelv szókészletét, az eredeti és a lefordított mű fordításánál felhasznált szókészletét érdemben össze tudjuk hasonlítani, morfológiai elemzőre és egyértelműsítő programokra lenne szükség.

## Összegzés

Az ismertetett modellt és az eredeti szöveget összehasonlítva megtalálhatók a szövegnek azon pontjai, amelyek nem következnek logikusan az előzményekből. A program segítségével azt is szemléltetni tudjuk, hogy egy egynyelvű nyelvkönyv típusai mennyiben térnek el, vagy mennyiben hasonlítanak egy regényhez.
Azt találtuk, hogy az új típusok megjelenésének szabályszerűsége sokkal inkább a szöveg hosszától, a szöveg műfajától függ, mintsem a szerzőtől. A rövid, egyenlő hosszúságú intervallumokra bontott szövegek sem adnak vissza a szerzőről olyan információt, amely alapján azonosítani lehet a szerzőt. További összehasonlítható eredményeket kaphatunk, ha olyan sorozatok elemzését végezzük el, amelyek kötetei bizonyítottan különböző szerzőktől származnak. Vizsgálataink során szeretnénk összehasonlítani korábbi századokból származó, valamint XX. századi műveket, annak érdekében, hogy megvizsgáljuk, találunk-e kimutatható eltérést a típusok megjelenésének szabályszerűségében az idők folyamán. Terveink között szerepel még nyelvkönyvek további feldolgozása is. A nyelvkönyvek megválasztásánál az elsődleges szempont a kiadó lesz. Ennek megfelelően olyan további sorozatokat szeretnénk elemezni, amelyek ugyanattól a kiadótól származnak, illetve olyanokat, amelyek más kiadó művei. A nyelvkönyvek elemzése a most ismertetett módszerrel további szempontokat adhat a nyelvkönyvek szóanyagának megválasztásához.

## Irodalomjegyzék

1. Yule G. U.: The Statistical Study of Literary Vocabulary. Cambridge University Press (1944)
2. Baayen R. H.: The Randomness Assumption in Word Frequency Statistics, Research in Humanities Computing 5 Selected Papers from the ACH/ALLC Conference, University of California, Santa Barbara, August 1995 (1996) 17-31.
3. Baayen R. H.: The Effect of Lexical Specialization on the Growth Curve of the Vocabulary. Computational Linguistics 22. (1996) 455-480.
4. Hoover D. L.: Another Perspective on Vocabulary Richness. Computers and the Humanities 37 (2003) 151-178.
5. Baayen R. H.: Statistical Models for Word Frequency Distributions: A Linguistic Evaluation. Computers and the Humanities 26. (1993) 347-363.
6. Baayen H., Halteren H., Tweedie F.: Outside the Cave of Shadows: Using Syntactic Annotation to Enhance Authorship Attribution. Literary and Linguistic Computing, Vol. 11, No. 3. (1996) 121-131
7. Cunningsworth A.: Choosing your Coursebook. Heinemann (1995)