Majomnyelv

Szavak előfordulási gyakoriságának modellezése nyelvi statisztikák alapján

Témavezető:
Dr. Járai-Szabó Ferenc, egyetemi adjunktus
Babeş-Bolyai Tudományegyetem
Fizika kar
Elméleti és számítógépes fizika tanszék

Szerző:
Balázs Melinda
Babeş-Bolyai Tudományegyetem
Fizika kar
Fizika-informatika szak, 3 év

XI. Erdélyi Tudományos Diákköri Konferencia – Kolozsvár, 2008. május 23–24.

Kivonat

A szövegek statisztikai elemzése arra az érdekes következtetésre vezetett, hogy a társadalmi vagyoneloszlásra megállapított Pareto-törvény igaz a szavak előfordulási gyakoriságára is. Ez azt jelenti, hogy van néhány szó, ami nagyon gyakran fordul elő a szövegekben, míg a szavak többségét ritkábban használjuk. Matematikailag ezt úgy mondhatjuk, hogy a szavak előfordulási gyakorisága hatványfüggvénnyel írható le. Ennek okát vizsgáltuk véletlenszerűen generált szöveg, vagyis a „majomnyelv" segítségével. A szövegek létrehozásakor rendre figyelembe vettük az eredeti, angol nyelvű szöveg betű-, betűpár- illetve szótageloszlását. Azt tapasztaltuk, hogy már az eredeti betűeloszlás alkalmazásakor visszakaptuk az angol szöveg szavainak hatványfüggvény eloszlását. A szavak Top 10 listájának elemzése viszont azt mutatta, hogy a szótageloszlás alapján generált szöveg közelítette meg legjobban az angol szöveget.

Tartalomjegyzék

Bevezető

A nyelvstatisztika nem tekinthető a nyelvtudomány önálló ágának, de eredményei sok területen hasznosíthatók [1]. Az írógép billentyűzetének vagy a nyomdai szedőgép betűállományának megtervezésében fontos szerepet játszottak az első betűstatisztikák. A gyorsírás kifejlesztésénél (amelyben a szótagoknak, szavaknak van külön jelük) szükséges volt arra, hogy tudjuk, mik a leggyakoribb szótagok, hangkapcsolatok. A Morse-abc is alkalmazza a statisztikát: az 'e' betű az angolban gyakori, ezért a jele rövid, egyszerű: '.'; míg a ritkább betűk kódjelei hosszabbak, bonyolultabbak.

A nyelvre jellemző statisztikák felállításhoz gyakoriságelemzéseket használunk [2], amelyeknek alapja az, hogy bármely szövegrészletben bizonyos betűk vagy betűkombinációk különböző gyakorisággal fordulnak elő. Mindezek mellett felhasználjuk azt a tényt, miszerint egy adott nyelvben a betűknek rájuk jellemző az eloszlása, amely minden szövegrészletben nagyjából azonos. Például az angol nyelvben az *E* betű nagyon gyakori, szemben az igen ritka *X*-szel. Az ún. bigrammák vagy digráfiák, azaz a két egymást követő betűből álló betűkombinációk (betűpárok) terén az angol nyelvben azt tapasztaljuk, hogy a *TH* és a *HE* fordul elő a leggyakrabban. Ez nem meglepő, hiszen a *THE* szó az angol nyelv határozott névelője és egyben a leggyakrabban használt szava. A legritkábban előforduló bigrammák között említhetjük az *NZ* és a *QJ* betűpárokat.

A szövegek statisztikai elemzése arra az érdekes következtetésre vezetett, hogy a társadalmi vagyoneloszlásra megállapított Pareto-törvény igaz a szavak előfordulási gyakoriságára is. Ezeket az eredményeket reprodukáljuk angol nyelvű szövegek elemzésével, majd a törvény okait próbáljuk kideríteni. Ezért úgy generálunk véletlenszerű szövegeket, hogy a szimuláció során rendre figyelembe vesszük az angol nyelv betűinek nullad-, első- és másodrendű korrelációit.

A következő fejezetben röviden ismertetjük a Pareto-törvényt és ennek néhány megnyilvánulási formáját. Ezt követően, saját eredményeink alapján megmutatjuk, hogy ez a törvény igaz az agol nyelvű szavak előfordulási gyakoriságára is. A dolgozat harmadik részében pedig azokat az eredményeinket mutatjuk be, melyek mesterségesen generált szövegekre alapozott nyelvi statisztikákkal próbálják megmagyarázni a szavak gyakoriság-eloszlását.

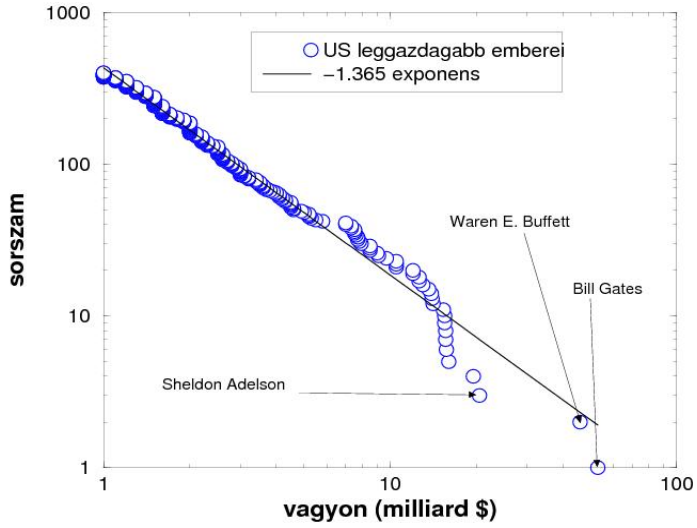
Pareto-törvény

1906-ban Vilfredo Pareto, itáliai közgazdász, felállított egy matematikai képletet, hogy segítségével leírja az országában megfigyelt vagyoni egyenlőtlenségek jellegzetességeit. Megfigyelte, hogy az emberek 20%-a rendelkezett az összvagyon 80%-a fölött. A Pareto-törvényt később 80/20-as szabályként kezdték emlegetni [3], és sokan mások is hasonló megfigyeléseket tettek a saját szakterületükön.

A 80/20-as szabály jelentése, hogy bármely vizsgálandó dologban 20% részarányt képviselnek a lényeges elemek, míg 80%-ban lényegtelen részek alkotják. Juran kezdeti munkájában megfigyelte, hogy a hibák 20%-a okozza a problémák 80%-át. Project managerek tisztában vannak vele, hogy a munka 20%-a fogyasztja el a rendelkezésre álló idő és erőforrások 80%-át. A tárolt készletek 20%-a foglalja el a raktárterületek 80%-át, és a készletek 80%-a a szállítók 20%-ától érkezik. Az értékesítés 80%-át az értékesítők 20%-a valósítja meg. Az alkalmazottak 20%-a okozza a problémák 80%-át, de egy másik 20% állítja elő a hozzáadott érték 80%-át.

Hatványfüggvény eloszlások

A Pareto-törvényt úgy is megfogalmazhatjuk [4], hogy ha sorba rakjuk a társadalom tagjait vagyonaik szerint (első a leggazdagabb, második a következő gazdagságú és így tovább), a sorindexet a vagyon függvényében ábrázolva egy hatványfüggvényt (x^n) kapunk.



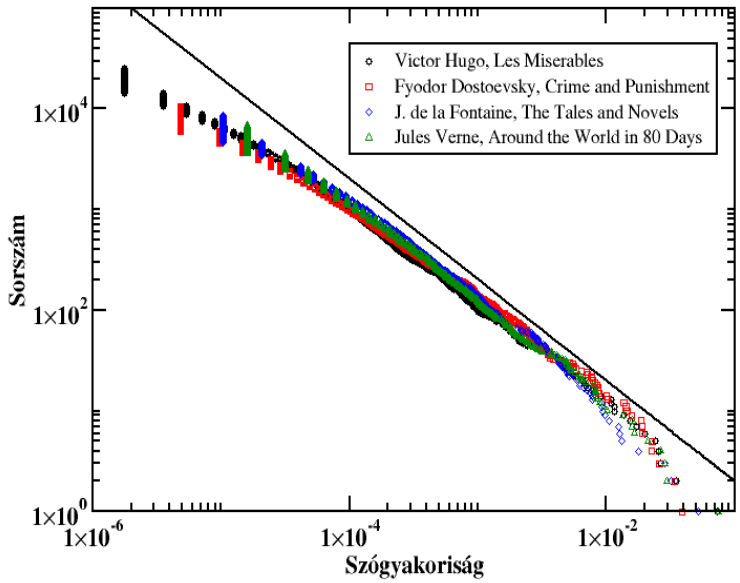
1. ábra. Vagyoneloszlás az Amerikai Egyesült államokban [4].

A hatványfüggvény eloszlást tipikusan log-log grafikonon ábrázoljuk (mindkét tengelyen logaritmikus skálát alkalmazva), ahol a függvény egy egyenes (általában jobbra lejtő) vonalként jelenik meg. Például, az 1. ábrán látható egyenes egy ilyen hatványfüggvény eloszlást ábrázol, nevezetesen az Amerikai Egyesült Államokban a vagyoneloszlást.

Pareto-törvény a szavak gyakoriság-eloszlására

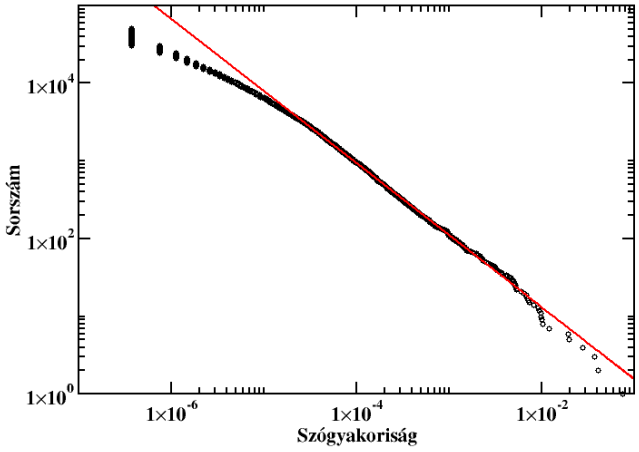
A szavak gyakoriság-eloszlását tanulmányoztuk az angol nyelvben különböző stílusú szövegek alapján, mint például: szépirodalom, Biblia, természettudomány, történelem, földrajz és közgazdaságtan.

Az 2. ábrán négy különböző irodalmi mű szógyakoriságát ábrázoltuk. Az ábrán látható, hogy a különböző szövegek szógyakoriságának eloszlása ugyanolyan típusú hatványfüggvénnyel írható le, melynek hatványkitevője közel *-1*.



2. ábra. Szavak gyakoriságának eloszlása különböző szépirodalmi művekben.

Az angol nyelvre jellemző átlagos szógyakoriság eloszlást úgy határoztuk meg, hogy különböző stílusú szövegeket vágtunk össze és az így kapott több mint 2,5 millió szóból álló szöveget vizsgáltuk. Ezt a listát csökkenő sorrendbe rendeztük a gyakoriságok szerint és log-log skálán ábrázoltuk.



3. ábra. Szavak gyakoriságának eloszlása az angol nyelvben.

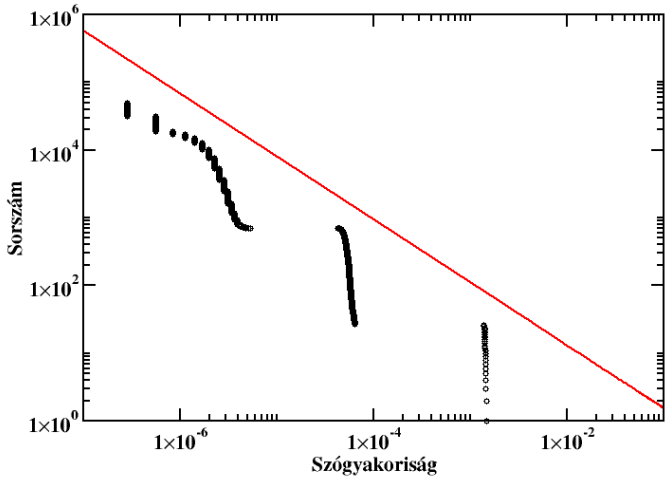
Az eloszlásfüggvényünkre ezek után egy x^{-1} alakú hatványfüggvényt fitteltünk. Azt tapasztaltuk, hogy a már jól ismert és sikeresen alkalmazott 80/20 szabály igaz a szavak gyakoriság eloszlására is. Ez azt jelenti, hogy a szavak kis részét nagyon gyakran használjuk, míg a szavak többségét csak igen ritkán. Vagyis a szavak előfordulási gyakorisága hatványfüggvény eloszlást mutat.

A szógyakoriság korrelációs okainak vizsgálata

A szavak előfordulási gyakoriságának modellezése érdekében számítógépes szimulációkat végeztünk. Először teljesen véletlenszerű szövegeket generáltunk, majd sorra megnéztük a valós angol nyelvű szövegek különböző rendű korrelációinak hatásait a kapott szövegek szógyakoriság eloszlásaira.

Véletlenszerű szövegek

Első lépésként reprodukáltuk a véletlenszerűen generált (vagyis teljesen korrelálatlan) szövegek esetében jól ismert eloszlást. Ennek érdekében írtunk egy programot, amely teljesen véletlenszerűen betűket állít elő (a-tól z-ig). A betűk mellett, velük azonos előfordulási valószínűséggel, szóköz karaktert is generált. Ez jelentette a szavak kezdetét illetve végét. Az így kapott szövegben, amely 100 millió karakterből állt, megszámoltuk a különböző szavak előfordulásait. A szavakat előfordulási gyakoriság szerint sorba rendezve a 3. ábrán látható gyakoriság-diagrammot kaptuk.



4. ábra. Szavak gyakoriságának előfordulása teljesen véletlenszerűen generált szövegben.

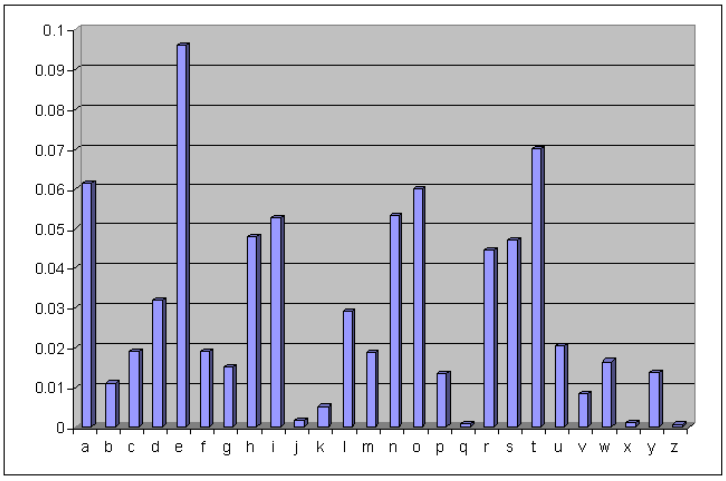
A 4. ábrán a pontok jelölik a véletlenszerű szöveg szógyakoriságait, míg az egyenes az előző fejezetből ismert, az angol nyelvre jellemző -1 kitevőjű hatványfüggvény. A pontokból kialakuló három, többé-kevésbé függőleges egyenes, az egy-, két- illetve hárombetűs szavak valószínűségeit mutatja. Ez könnyen belátható, ha figyelembe vesszük, hogy az egybetűs szavak kialakulásának valószínűsége $1/N^2$, a kétbetűs szavaké $1/N^3$ és így tovább, ahol N a

rendelkezésre álló karakterek száma. Tehát, általánosan, egy n betűből álló szó előfordulási valószínűsége $1/N^n$ lesz.

Betűkorrelált szövegek

Vizsgálatunk második lépéseként azt terveztük, hogy az angol nyelvű szöveg betűinek jellegzetes eloszlását figyelembe véve generálunk szöveget és ebben vizsgájuk a szavak előfordulásának gyakoriságát.

Az előző fejezetben használt, 16 millió karakterből álló szöveg elemzése alapján a 4. ábrán látható betűgyakoriság hisztogramot készítettük el. Az ábra alapján megjegyezzük, hogy a mi számításaink szerint is az angol nyelv 5 leggyakoribb betűje az E, T, A, O, N. A 4. ábra azért tér el a szakirodalomban ismert betűgyakoriság diagramoktól [5], mert a gyakoriságok számolásánál figyelembe vettük a szóközöket is. Ez nem befolyásolja a betűk 'ranglistáját', csak az egyes előfordulási valószínűségi értékeket.

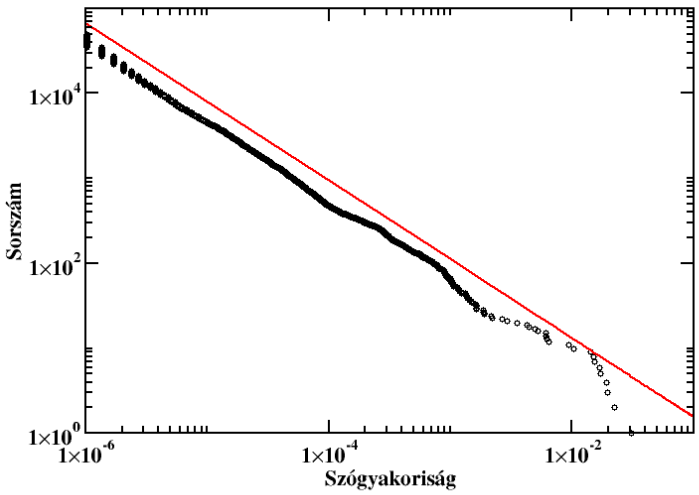


5. ábra. Az angol nyelvű szövegekben a betűk jellemző eloszlása.

A kapott betűgyakoriság alapján a szöveget a rulettkerék algoritmus segítségével generáltuk. Ennek lényege, hogy minden karakter az előfordulási valószínűségével arányos nagyságú helyet foglal el a rulettkeréken. A programunkban ezt úgy valósítottuk meg, hogy a [0,1] intervallumon szétosztottuk az előfordulási valószínűségek értékeit, majd egy véletlenszámot generáltunk ezen az intervallumon. Azt a betűt tekintettük a következőnek, amelyiknek az intervallumába esett a véletlenszám.

A fent leírt algoritmussal generált szöveg a következőképpen alakult: „ ... *ro h n b m g rristl ns ne rieiah eloftt ue o oi drr sse lye ataoh euaedf i onhheoibiwb wneainw o hnft t l oanaoes trd r f l os ain kia n gda o amtioom oacrw w omed nitn avay hm dooif tay ssm boyxrau a lt oaogr hnfgponlidrnninmdtie l mrdea ips ampu ...* ".

Első meglepetésünkre a szavak gyakoriságának eloszlásfüggvénye drasztikusan megváltozott a teljesen véletlenszerűen generált szöveghez képest. Amint az a 6. ábrán látható a betűk nulladrendű korrelációjának figyelembe vétele a valódi angol nyelvű szógyakoriság eloszláshoz nagyon közeli eloszlást eredményez.



6. ábra. Szavak gyakoriságának előfordulása a nulladrendű korrelációkat figyelembe vevő véletlenszerűen generált szövegben. Az egyenes az angol nyelvre jellemző -1 kitevőjű hatványfüggvény.

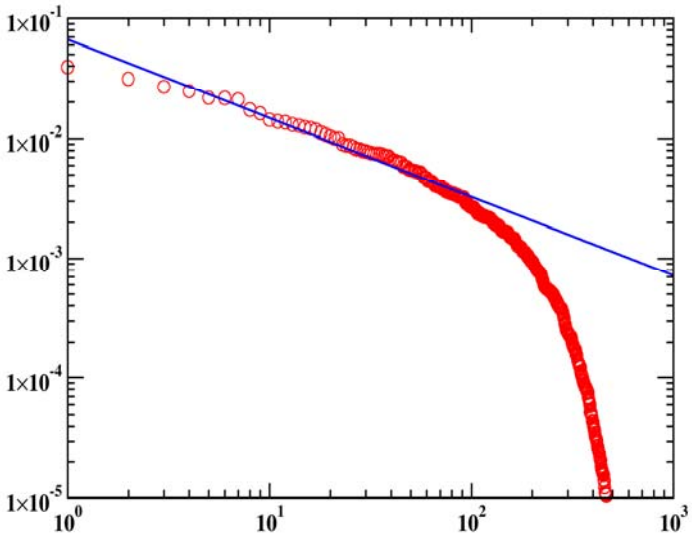
Azonban, ha a leggyakrabban előforduló szavakat megnézzük, azok mind egybetűs szavak lesznek, az adott betűk előfordulási valószínűségével arányos gyakoriságokkal.

Az eddigi eredményeink alapján azt mondhatjuk, hogy a ritka szavak statisztikája magyarázható az adott nyelv betűeloszlása alapján.

Betűpár korrelált szövegek

Ezek után tovább léptünk és a betűk közti elsőrendű korrelációt tekintettük a betűgenerálásunk alapjául. Ez azt jelenti, hogy megvizsgáltuk az angol szövegben a különböző betűpárok gyakoriságait. Itt is meg kell jegyeznünk, hogy a szóköz karaktert is a betűk közé soroltuk. Számításaink alapján itt is az irodalomból ismert tipikus betűpár eloszlást [5] kaptuk. A 7. ábrán ennek az eloszlásnak a sorszám-gyakoriság eloszlása látható.

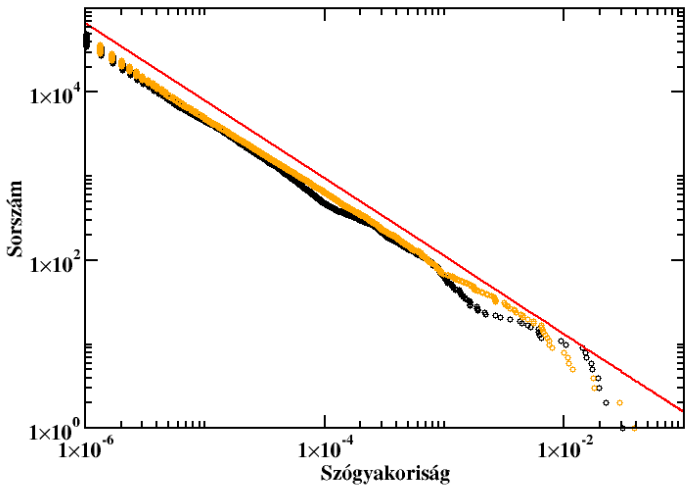
A kapott betűpár-gyakoriság alapján a szöveget itt is a rulettkerék algoritmus segítségével generáltuk.



7. ábra. Az angol nyelvű szövegekben a betűpárok jellemző eloszlása.

A betűpárok korrelációira alapuló véletlenszerűen generált szöveg a következőképpen néz ki: „ ... o *s risomo n hat ly s anachet he avirin jopl thahe the the hino noueches tharokl orkn n then isharrg big theninof m fum rtis ind f ane ing andisilts gut sme by iberelawe yotond indeamuban fangend gibow ists wele as bund wisan otthethirerscans me onguris beir ofeds bie* ... ”.

Azt tapasztaltuk, hogy az eloszlásfüggvény alakja az előző esethez viszonyítva nem sokat változott. Ez látható a 8. ábrán, amelyen a nullad- és elsőrendű korrelációkkal kapott sorszám-gyakoriság diagramokat ábrázoltuk.

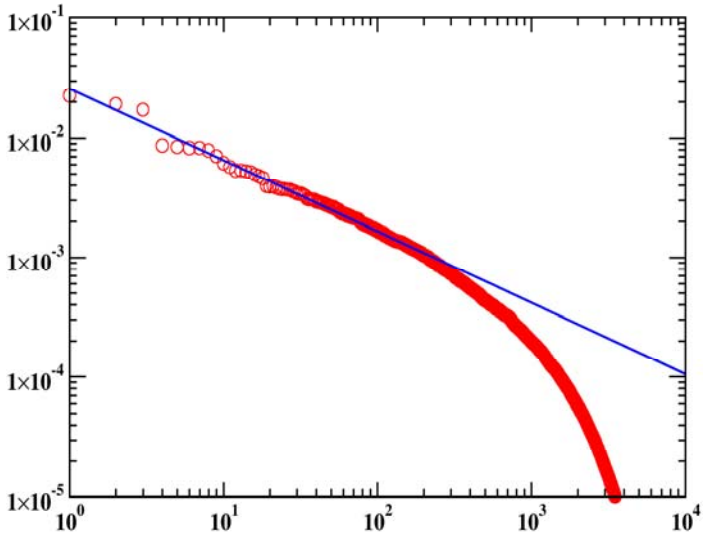


8. ábra. Szavak gyakoriságának előfordulása a elsőrendű korrelációkat figyelembe vevő véletlenszerűen generált szövegben. Az egyenes az angol nyelvre jellemző -1 kitevőjű hatványfüggvény.

Az ábrát vizsgálva megállapíthatjuk, hogy a szavak előfordulási gyakorisága csak a gyakori szavak esetén változik látványosan, a ritka szavak esetében nem.

Szótagkorrelált szövegek

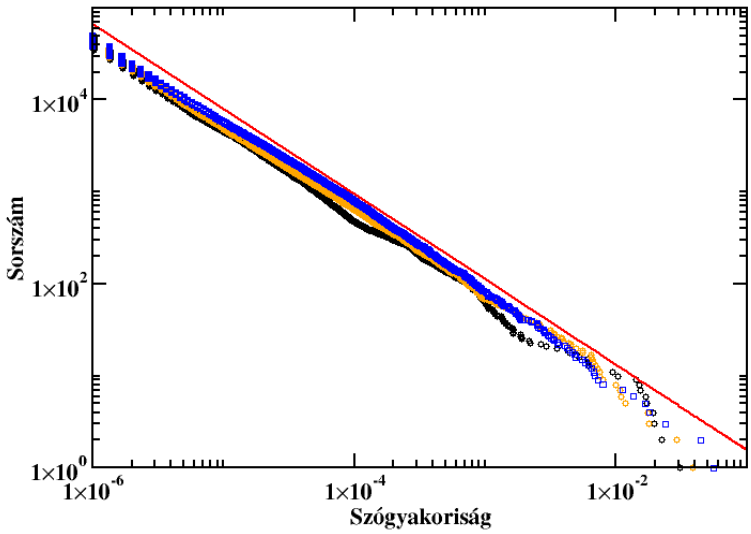
A nullad- illetve elsőrendű korrelációk után a betűk másodrendű korrelációját vizsgáltuk és tekintettük a szöveggenerálásunk alapjául. Ebben az esetben a hármas betűcsoportok (’szótagok’) gyakoriságait számoltuk. A szóköz karaktert itt is a betűk közé soroltuk. Vizsgálatunk most is a várt eredményeket mutatta [5]. A 9. ábrán a szótagok sorszám-gyakoriság eloszlása látható.



9. ábra. Az angol nyelvű szövegekben a hármas betűpárok jellemző eloszlása.

A már ismert eloszlást felhasználva ebben az esetben is a rulettkerék algoritmus használatával állítottuk elő a szöveget. A szimuláció a következőket eredményezte: „ ... *they drigivicen raw anthat ned the a momard fiet relly tromen have chille youstry id fume logg froxerned boodis ray smen hat of nothe fave pince come tooten socculd the med he noted of strucks be gre darlor te in suffew ...* ”.

A szavak előfordulási gyakoriságának képét vizsgálva ismét nem tapasztaltunk feltűnő változást az előzőekhez képest. A 10. ábrán. a nullad-, első- és másodrendű betűkorrelációk figyelembe vételével készült szövegek szógyakoriság eloszlása van ábrázolva. Megfigyelhető, hogy a grafikonok egymástól csupán a nagyon gyakori szavak tartományában különböznek. A szótaggyakoriság figyelembe vétele esetén (üres négyzetek a 10. ábrán) azt tapasztaltuk, hogy a gyakori szavak eloszlása is követi a hatványfüggvény eloszlást. A ritka szavak tartományában egységesen követik az eredeti angol szövegek statisztikáiból megállapított x^{-1} alakú hatványfüggvényt.



10. ábra. Szavak gyakoriságának előfordulása a másodrendű korrelációkat figyelembe vevő véletlenszerűen generált szövegben. Az egyenes az angol nyelvre jellemző -1 kitevőjű hatványfüggvény.

Ebben a szövegben a leggyakoribb szavak már megegyeznek az angol nyelvben leggyakrabban használt szavakkal. Ez azt jelenti, hogy minimálisan a betűk másodrendű korrelációját kell figyelembe venni ahhoz, hogy az eredeti nyelv statisztikait kapjuk.

Top 10 elemzés

A generált szövegeknek az angol nyelvtől való eltérésének mértékét tanulmányoztuk a Top 10 elemzések során. Ehhez az 1. táblázatot használtuk, melyben az angol nyelv 10 leggyakoribb szava található, megjelölve azok helyezését a különböző típusú, véletlenszerűen generált szövegek szó-toplistájában is.

	angol	betű	betűpár	szótag
the	1	445	6	1
of	2	177	25	2
and	3	1433	32	4
to	4	43	39	3
in	5	78	22	8
a	6	3	11	5
he	7	35	5	6
was	8	2810	136	15
his	9	1280	140	21
that	10	11664	288	58

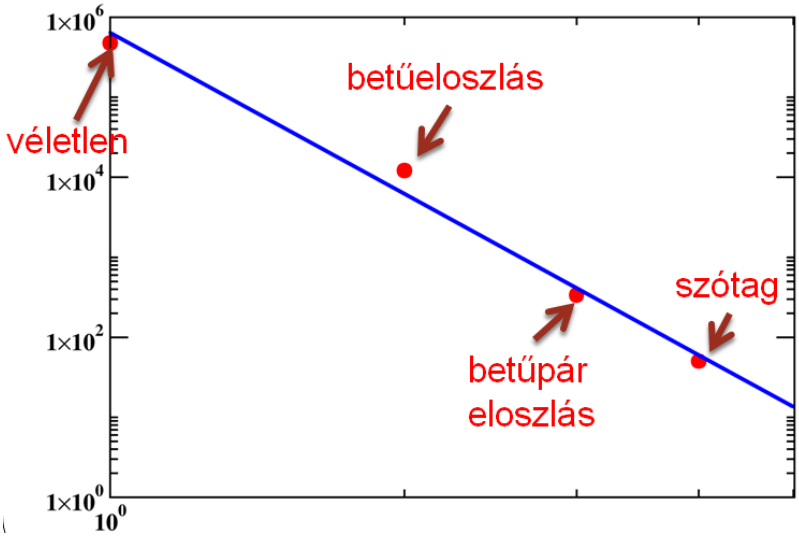
1. táblázat. A különböző módon generált szövegek Top 10 elemzése.

Az eltérések mértékének a kiszámítására a következő összefüggést értelmeztük:

$$\langle \Delta p \rangle_{10} = \sqrt{\frac{1}{10} \sum_{i=1}^{10} \left(p_{en}^{(i)} - p^{(i)} \right)^2}$$

ahol $p_{en}^{(i)} = i$ a i-edik szó helyzete az angol toplistában, míg a $p^{(i)}$ ugyanannak a szónak a helyzete az egyik szimulált toplistán. Ennek segítségével egyfajta szórást számolunk az angol és a különböző módon generált szövegek 10 leggyakoribb szavára. A szórásnak *0*-hoz közeli az értéke, ha az eredeti angol és a vizsgált szimulációs szöveg szó-toplistái kevésben különböznek. Ellenkező esetben nagy szórás értéket kapunk.

A generált szövegek szórásait a 11. ábrán tüntettük fel log-log skálán.



11. ábra. A különböző módon generált szövegek szórása.

A grafikon vízszintes tengelyén a betűk között tekintett korreláció fokát ábrázoltuk (valójában a korreláció fok - 2 értéket), a függőleges tengelyen pedig a szórást. Látható, hogy minél magasabb rendű korrelációt veszünk figyelembe, a generált szöveg statisztikái egyre inkább megközelítik az angol nyelvét. Sőt, az ábra alapján még azt is elmondhatjuk, hogy a korreláció fokának növekedésével a szórás hatványfüggvényszerűen tart a *0* értékhez – *6,7* hatványkitevővel.

Következtetések

Ebben a dolgozatban a Pareto-törvény érvényességének okait próbáltuk kideríteni az angol nyelv szógyakoriság eloszlására vonatkozóan. Ennek érdekében úgy generáltunk véletlenszerű szövegeket, hogy a szimuláció során figyelembe vettük az angol nyelv betűinek nullad-, első- és másodrendű korrelációit.

Kezdetben teljesen véletlenszerű szövegeket generáltunk és azt tapasztaltuk, hogy a kapott szöveg szavainak statisztikája nagyban eltér az angol nyelvű statisztikáktól. Ezután az angol nyelvű szöveg betűinek jellegzetes eloszlását figyelembe véve generáltunk szöveget és ebben vizsgáltuk a szavak előfordulásának gyakoriságát. Ekkor a szavak gyakoriságának eloszlásfüggvénye megváltozott a teljesen véletlenszerűen generált szöveghez képest és a ritka szavak esetében nagyon közel került az angol nyelv szógyakoriságának eloszlásfüggvényéhez. Ugyanezt az elemzést elvégeztük betűpár és szótag előfordulások figyelembe vételével generált szövegekre is. A szótaggyakoriság figyelembe vételével a gyakori szavak eloszlása is követte az eredeti hatványfüggvény eloszlást.

Vizsgálataink alapján azt mondhatjuk, hogy a Pareto-törvény érvényességének elsődleges oka az adott nyelvre (a mi esetünkben az angol) jellemző tipikus betűeloszlásban keresendő. Ezt a következtetést megmagyarázhatjuk, ha figyelembe vesszük azt a tényt, hogy a betűgyakorisággal egy időben a szóközgyakoriságot is figyelembe vettük, ami implicit módon magába foglalja az adott nyelvre jellemző átlagos szóhosszúságot is. Ennek az általános kijelentésnek a teljes igazolásához azonban még szükség lesz az általunk végzett elemzések/szimulációk kiterjesztésére más nyelvek esetére is.

Könyvészet

[1] Zsobrák Róbert, Isz siveöl tégaz hazuj
http://www.sulinet.hu/tart/fcikk/Kcl/0/21182/1

[2] http://hu.wikipedia.org/wiki/Gyakoris%C3%A1gelemz%C3%A9s

[3] http://www.businessmix.hu/index2.php?option=com_content&do_pdf=1&id=33

[4] Néda Zoltán, Szociális hálózatok és vagyoneloszlása társadalmakban
(Korunk Június 2005 A háló tudománya – művészetek hálója)

[5] Bill Cherowitzo, Modern Kriptográfia, online kurzus,
http://www-math.cudenver.edu/~wcherowi/courses/m5410/engstat.html