IROTT SZÖVEGEK STATISZTIKAI TULAJDONSÁGAINAK MODELLEZÉSE MATEMATIKAI ESZKÖZÖKKEL

szakdolgozat

Készítette: **Bíró Tamás**
Elméleti nyelvészet szak
Eötvös Lóránd Tudományegyetem BTK

Budapest, 2000.

Összefoglaló

Dolgozatom témája az írott szövegek statisztikus tulajdonságai, közelebbről a korrelációk, és azok modellezése nyelvészetileg is releváns matematikai modellekkel. Az írott szövegeket mint egy véges ábécé (lexikon) elemeiből alkotott szimbólumszekvenciát fogom fel, és ilyen szempontból alig különböznek a más ábécék fölött generált szekvenciáktól, például számítógépprogramoktól vagy a genetikai kódtól. Kiderül, hogy létezik ezen szimbólumszekvenciáknak egy széles osztálya, amelyet a hosszú távú korrelációk létével és a hatványfüggvényként lecsengő Zipf-függvénnyel lehet többek között jellemezni. A természetes nyelveken írott szövegek beletartoznak ebbe az osztályba. Ezen osztály modellezéséhez vezetek be sztochasztikus eszközöket, többek között definiálom a sztochasztikus veremautomata fogalmát, amely képes visszaadni ezeket a jellemzőket. A modell nyelvészeti relevanciáját az adja, hogy az X-vonás elmélet mintájára épül fel. De a teljes magyarázóértékhez fel kell tételeznünk egy ,,vakmerő" hipotézist is, miszerint az írott szövegek globális struktúrája – tehát a mondatszint feletti összetevők is – jellemezhetők az X-vonás elmélettel.

Tartalomjegyzék

1. fejezet

Bevezetés

"A clash of doctrines is not a disaster, it is an opportunity. "

Whitehead [1]

A nyelvtudomány hagyományosan a *par excellence* bölcsésztudományok egyike volt, és így a külső szemlélő számára meglehetősen furcsa lehet a ,,matematikai nyelvészet", mint főnévi csoport. Pedig a modern nyelvészet egyik legnagyobb hatású irányzatáról van szó.

Amikor arra kérnek, fejtsem ki részletesebben, mit is fed a ,,matematikai nyelvészet", a következő csoportosítást szoktam elmagyarázni. Először is ,,kvantitatív nyelvészetnek" nevezném a hagyományos nyelvészet azon irányzatát, amely a nyelv kvantitatív tulajdonságaival foglalkozik (Nagy Ferenc [1986]). A hang- és szógyakorisági vizsgálaton túl valószínűleg a sokat vitatott *glottokronológia* ezen iskola legjelentősebb eredménye. A hagyományos nyelvészet másik természettudományokhoz közelálló ága a fonetika fizikai és biológiai vonatkozásai.

Amikor matematikai nyelvészetről beszélünk, inkább a modern nyelvtudomány erösen matematizált irányzataira gondolunk. Ezen belül is három lehetőséget említenék meg. Az ,,algebrai nyelvészet" az ötvenes évektől kezdődően, a generatív nyelvészet megjelenésével kezdte el erőteljesen befolyásolni a nyelvtudományt. A Chomsky és társai által kidolgozott formális nyelvek elmélete után újabb és újabb matematikai – főleg algebrai – fogalmak találtak utat a modern nyelvészet különböző modelljeibe. Egyesek fogékonyabbak voltak a formalizálásra, mint mások, de majdnem minden területen találunk a matematikából

[1] *"Ha a doktrínák ütköznek egymással, az nem katasztrófa, hanem kihasználandó esély.", Whitehead, Science and Modern World, p. 186, idézi Prigogine & Stengers [1986].*

kölcsönzött, azzal néha kevés kapcsolatban lévő, szavakat (pl. jegymátrix, fastruktúra,...). A matematizálhatóbb elméletek a ,,számítógépes nyelvészet" kiinduló pontjaivá váltak, amelyet a modern matematikai nyelvészet második ágának tekinthetünk, és amelynek az a kitűzött célja, hogy algoritmizálja ezen elméleteket.

A harmadik ágat ,,statisztikus nyelvészetnek" nevezném. Szemben az absztrakt algebrai eszköket használó matematikai nyelvészettel, ezen kutatások során a nyelv számokkal mérhető, kvantitatív tulajdonságait vizsgálják, akárcsak a klasszikus kvantitatív nyelvészetben. De egyrészt komplexebb matematikai aparátussal, másrészt – modern nyelvtudományról lévén szó – a cél ezen tulajdonságok modellezése, illetve a létező modellekkel való összevetése.

Dolgozatom témája ezen utolsó csoportba sorolható. A ,,komplexebb matematikai aparátus" jelen esetben a statisztikus fizika által motivált eszközöket jelenti, és sztochasztikus modellekkel kívánom magyarázni a megfigyelt tényeket. Az empírikus adatokat fizikusok kutatásai szolgáltatták, amelyeket az 1990-es évek első felében publikáltak a világ sok pontján, és az ehhez illesztett, nyelvészetileg remélhetőleg motivált modellek saját magam eredményei. Nem tudok hasonló jellegű próbálkozásokról.

A modern fizikai háttér-történetéhez annyit érdemes tudni, hogy a statisztikus fizika a termodinamika és a szilárdtestfizika (anyagtudomány) modern hajtása, az elmúlt néhány évtizedben kezdte önálló életét. Az említett tudományterületek matematikai aparátusát fejlesztette tovább egy absztraktabb szinten, majd rájött a jelenségek univerzális voltára. A dinamikai rendszerek, a káoszelmélet, a fraktálok és a növekedési jelenségek mind-mind összefüggő fogalmak, és az ezekkel leírható jelenségek megtalálhatók a kémiában és az evolúcióbiológiában éppúgy, mind a közgazdaságtanban vagy a közúti közlekedés modellezésében.[1] A fraktálok fogalma elő fog kerülni dolgozatom negyedik fejezetében is.

[1] Közérthető cikkek találhatók magyarul a társadalomtudományokból vett példákkal a *Rend és káosz* [1997] kötetben. Mainzer [1997] tárgyalja többek között az evolúciót, az emberi agyat, a mesterséges intelligenciát és az emberi társadalmat a komplexitás szempontjából. Port & van Gelder [1995] a kognitív tudományok valamennyi területét felöleli a dinamikai rendszerek szemszögéből. Szépfalusy & Tél [1982] tartalmaz cikkeket a kémia és az elektronika területéről is. Az irodalomjegyzékben szerepelnek további interdiszciplináris témájú publikációk is, pl. Bouchaud [1997], Ghashghaie [1997], Liu [1997], Martinás [1995, 1997], Mantegna [1996, 1997], Potters [1997], Stanley [1996a, b] az elmúlt években a fizikusok közgazdaságtan felé való fordulásának a termékei, míg Stanley [1992, 1993a, b], Vicsek [1995], Derényi [1996], Geritz [1997] a fizikusok ,,biológiai korszakának" a példái. Family & Vicsek [1991] és Vicsek [1992] szintén sokféle példát tartalmaznak mindenféle területről.

A fraktálokkal foglalkozó fizikusok az 1990-es évek első felében kezdtek el párhuzamosan a DNS-szekvenciák és az írott szövegek, általánosan fogalmazva a szimbólumszekvenciák, hosszú távú korrelációival foglalkozni, akkor, amikor hirtelen megnőtt a számítógépen könnyen hozzáférhető adatbázisok száma. Az NCBI-genbankban található információ mennyisége 1982 decembere és 1998 tavasza között körülbelül 14 hónaponként kétszereződött meg (Bíró [1998a]). E kutatások hatására a Zipf-függvények vizsgálata is előtérbe került. De 1995-96-ra leült a publikálási kedv, talán azért, mert a témákat kimerítették a fizikusok szemszögéből, és ugyan ezen kutatók jelentős része áttért a tőzsdei folyamatok vagy a vállalatok méret szerinti eloszlásának hasonló elemzésére.

(Mivel a cikkek többsége inkább a DNS-szekvenciákkal foglalkozik, és kevesebb szó esik a velük gyakran analóg tulajdonságokkal rendelkező írott szövegekről, ezért dolgozatom egyes fejezeteiben nagyobb súlyt kapnak a genetikai szekvenciák. De ezen megfigyelések fontosak az írott szövegek szempontjából is.)

A dolgozatomban megpróbálom e a ponton felvenni a fonalat, és a nyelvészet szempontjából diszkutálni ezen eredményeket. Meggyőződésem, hogy a fizikusok eredményeinek a nyelvészek számára is van mondanivalója. De tudtommal, evvel a kérdéssel eddig senki sem foglalkozott.

A második fejezetben a dolgozatom által megkívánt matematikai előismereteket foglalom össze, majd ismertetem azokat a korábbi eredményeket, amelyeket, mint empirikus tényeket, kívánok modellezni. A módszer a tudományosság által szeretnék megfigyelés-modellezés lesz. A második fejezet végén modellezem az előzőekben leírt rövid távú korrelációkat, majd a harmadik fejezetben ismertetem az előttem végzett megfigyeléseket a hosszú távú korrelációk létével kapcsolatosan. A negyedik fejezetben pedig ismertetem az ehhez általam javasolt modelleket, majd néhány szimuláció eredményén mutatom be a hasznosságukat. Végezetül összefoglalom a dolgozatomban leírtakat.

1.1. Excursus: Az X-bar struktúra túlmutat-e a nyelvészeten?

Viszont, még mielőtt belekezdenék a matematikai fejtegetésekbe, egy egyelőre meglehetősen gyenge lábakon álló hipotézist szeretnék felvetni. Ennek indoka a 4. fejezet második felében válik világossá.

Ha a Chomskyánus felfogás szerint a nyelvi modul egyedülálló jelenség a kognitív modulok között, míg más felfogások szerint nincsen éles különbség közöttük, a hipotézisem maximalista felfogása szerint a nyelvi modul struktúráját veszi át a kognitív rendszer nagy része. Szerényebben megfogalmazva, azt szeretném belátni, hogy a nyelvészetben megis-

mert struktúrák dominálnak az emberi (nyelvi?) alkotások egy jelentős halmazában. Olyan területeken is, amelyeket nyelvészeten kívülinek, például társadalmilag meghatározottnak fogunk fel. Legszűkebb értelemben: a szövegek globális struktúráját – amely valószínűleg már nem a nyelvi modulhoz tartozik – a nyelvészetből ismert elvek határozzák meg.

Konkrétabban, az X-bar struktúráról szeretném belátni, hogy egy széles – statisztikai fizikai terminussal élve – univerzalitási osztályban érvényes. A közeljövőben a vallási liturgiák szerkezetén szeretném konkrétan kidolgozni a módszertant, majd más területekre is átvinni az elképzelést.

E ponton a levélformán fogom bemutatni az ötletet. A levélforma országonként, kultúránként kisebb változatosságokat mutat, mégis univerzális jelenség. A módszertan alapját éppen az univerzalitás és a változatok összehasonlítása adja.

A levél, mint az emberi kultúra ősi kommunikációs jelensége, alapvetően a következő elemekből épül fel: feladó, címzett, e kettő társadalmi és szituációs viszonya, a fogalmazás helye és ideje. A közölt információ, amely önmagában nem lenne más, mint egy mondatszint fölötti nyelvi produktum, éppen ezen elemek hangsúlyos, és csak a levélformára jellemző jelenléte miatt válhat levéllé.

Ezért, talán önkényes és a jelenlegi stádiumban nem eléggé megalapozott módon, a levél búcsúformuláját tartom a levél – az X-vonás elméletből ismert fogalommal élve – „fejének". Nevezzük L-nek. A búcsúformula, főleg a francia hivatalos levélformulához hasonló szerkezetekben, tartalmazza a levélformula fentebb felsorolt valamennyi elemét. De magyarul is, a búcsúformulában benne van implicite a feladó, a címzett, és e kettő térbeli, társadalmi, és minden egyéb viszonya.[2] A levél törzse lehet eme L fej XP bővítménye. Bővítmény a sztenderd fogalmak szerint, hiszen a fej szelektálja (például nem akárkinek írunk meg akármit, hanem a levél törzse a két fél közötti viszonytól függ); és maximális projekció: a levél törzse „szöveg", amely a mondat fölötti, nyelvileg meghatározott legnagyobb, legmagasabb egység, amely már nem nyelvi, hanem társadalmi tényezők által meghatározott struktúrákba ágyazódhat bele. (Más kérdés, hogy e társadalmilag meghatározott struktúrákat szintén az ember kognitív rendszere alakította ki.)

A búcsúformula, mint fej, és a levél törzse, mint bővítmény, alkot egy $\overline{L}$ komponenst. A mai levélformulában fej-végű az $\overline{L}$, de az ókori levélformulákban nyitóformula (is?)

[2] A "*Csókollak*" ragjában explicite szerepel mindkét fél, az ige pedig meglehetősen intim viszonyt feltételez kettejük között. A "*Tisztelettel*" jelentése "én tisztelettel vagyok teirántad". A "*Várom leveledet*" vagy az "*[I am] Looking forward to meeting you*" kifejezések nem csupán a két személyt tartalmazzák nyelvtanilag, hanem a két személy térbeli viszonyáról, jövőbeli kapcsolattartásáról is árulkodnak.

található, ami fej-kezdetű $\overline{L}$-t jelenthet. De $\overline{L}$ nem maximális projekció, hiszen legtöbbször van valami más is, rajta kívül a levélformulában. A megszólítás és az aláírás specifierek, és ettől válik teljes projekcióvá, LP-vé, amely önállóan létezhet. E két specifierbe a fej néhány eleme kerül: az aláírásba a feladó, míg a megszólításba a címzett, valamint a feladó és a címzett viszonya ,,mozog fel". Az nyilvánvaló az írásképből, valamint a mondatfűzésből, hogy a megszólítás közelebb van $\overline{L}$-hez, mint az aláírás,[3] és ez a megfigyelés megmenti a modell bináris voltát. Vajon ez azt jelenti, hogy a maximális projekció három vonásszámú? Vagy mondjuk azt, hogy egy újabb X-bar struktúra épül az LP fölé? Például az LP testvére az aláírás, míg e nagyobb struktúra specifier pozíciójába a keltezés kerül? Vagy egy fiktív *Sign* fej emeli ki az L-ből a feladót egy *Spec-Sign*-ba? És vajon hasonlóképpen kerül a (hely-)idő adat a *Spec-Date*-be?

Elvi szempontok mérlegelésén túl, a különböző változatok vizsgálata adhat választ ezen kérdésekre. Például a magyar és a nyugat-európai keltezési hagyományok különbözősége visszaadható azzal a paraméterrel, hogy a *Date* fej az időn túl a hely-információt is magához vonzza-e, vagy sem; valamint milyen ezen konstituens konfigurációja, bal vagy jobb fejű, azaz a levél elejére vagy végére kerül-e a keltezés?

Miért nem szerepel a címzett és a feladó végül is az LP fejében? Mert elmozgattuk őket onnan a megszólítás és az aláírás pozícióba, ezért ott csak a nyomukat (pl. névmások és személyragok formájában, vagy csak *pro*-ként) találjuk őket, mondhatjuk egy generatív szemléletben. Végül, e ,,játék" zárásaként még annyit, hogy a hivatalos levelek további projekciókat tartalmaznak, amiket magánlevelek nem. Ezzel magyarázható a feladó, a címzett, a tárgy (hivatkozási szám),... feltüntetése a levelek elején.

E metodológiát módszeresebben ki kell dolgozni, és sok más jelenségre is alkalmazni. Remélem, hogy erre a közeljövőben lesz módom. E ponton csupán a lehetőséget kívántam bemutatni, mert a 4. fejezet második felében a statisztikai modellünk nyelvészeti relevanciáját éppen az X-bar struktúrával támasztom alá. De ez csakis akkor működik, ha elfogadjuk, hogy a szöveg globális, mondatok fölötti szintjén is működik az X-vonás elmélet, különben abba a csapdába esünk, hogy egy nyelvészeti elméletre hivatkozom, a nyelvészeti elmélet hatókörén kívül.

De ha úgy tetszik, meg is fordíthatjuk a gondolatmenetet: az a tény, hogy a hosszú távú korrelációkat egy globális szinten alkalmazott, az X-vonás elméletre utaló sztochasztikus modellel tudtuk leírni, további érv lehet a globális X-vonlás elmélet mellett.

[3] A megszólítás része a levél mondatstruktúrájának, mert írásjellel kapcsolódik a következő mondathoz, míg az aláírás esetén ez a látszat kevésbé van meg.

2. fejezet

Rövid távú korrelációk és modellezéseik

2.1. Matematikai bevezetés

A dolgozatom matematikai fogalmait szeretném jelen fejezetben összefoglalni. Habár bölcsész szakdolgozatról van szó, bizonyos matematikai előismereteket feltételezek, és csupán a specifikusabb tudnivalókra szorítkozom.

Az általam bemutatandó megközelítésben az írott szöveget, mint szimbólumsorozatot kezelem. Ilyen szempontból érdektelen, hogy adott statisztikai eljárást vagy matematikai modellt írott szövegre, DNS-szekvenciára,[4] vagy mondjuk egy számítógépes programra alkalmazzuk-e. E szimbólumszekvencia statisztikai tulajdonságai a sorozatot generáló nyelvet jellemzik kvantitatív (sztochasztikus) szemszögből, és célom ennek vizsgálata.

Dolgozatom központi fogalma a *korreláció*. Ha A és B két valószínűségi változó (tetszőleges számmal kifejezett mennyiségek), akkor C korrelációjuk

$$C := < AB > - < A >< B >, \tag{2.1}$$

ahol a $< >$ jelek a várható értéket jelentik. Szemléletesen, e két mennyiség egymástól való függését fejezi ki a korreláció, amely csak akkor zérus, ha független mennyiségekről van szó: az egyik változása nem vonja tipikusan maga után a másik változását. Korrelált

[4] A DNS-szekvencia egy négyelemű ábécé, a lehetséges bázispárok (adenin, citozin, guanin és timin, azaz `A`, `C`, `G` és `T`) fölött generált szekvencia.

mennyiségek esetén az egyik növekedése a másik növekedésével (negatív korreláció esetén a csökkenésével) jár legtöbbször együtt.

Egy szimbólumsorozat esetén a szekvencia egyes pozicióit tekinthetjük valószínűségi változóknak (A_i), és ekkor vizsgálhatjuk az i-ik és a j-ik pozíció közötti korrelációt (*autokorrelációnak* is nevezhetjük, mivel a sorozat „önmagával" vett korrelációját tekintjük):

$$C_{i,j} := < A_i A_j > - < A_i >< A_j > . \tag{2.2}$$

Bizonyos tulajdonságokat teljesítő, ún. *ergodikus* rendszerek esetén a várható értékeket képezhetjük úgy is, hogy „a sorozaton megyünk végig", vagyis az összes i-re átlagolunk. Azaz ekkor értelme van definiálni az *autokorrelációs függvényt*:

$$C(k) := \langle X_i \cdot X_{i+k} \rangle - \langle X_i \rangle \langle X_{i+k} \rangle, \tag{2.3}$$

ahol az átlagolást az összes i-re végezzük.

Az autokorrelációs függvény szemléletesen azt fejezi ki, hogy milyen erős az egymástól l távolságra lévő szimbólumok közötti függőség. Rulettszámok sorozata esetén függetlenek a különböző események, a napi átlaghőmérsékletek sorozata viszont sokkal erősebben összefügg: annak a valószínűsége, hogy valamelyik napon 20 fok lesz sokkal nagyobb, ha a megelőző napokon is 20 fok körüli volt a hőmérséklet, mint ha -5 fok volt. Ha például pletykák kialakulását szeretnénk modellezni, akkor a hír továbbterjedésének a láncszemei alkothatják a szekvencia egyes elemeit. Kiderülne, hogy az autokorrelációs függvény meglepően gyorsan lecseng: valamely állapotban valymi kevés köze van a pletykának ahhoz, hogy l állapottal korábban mi volt a pletyka, ha l elegendően nagy szám.

Konkrétabban, az autokorrelációs függvény háromféle tipikus viselkedése figyelhető meg az alkalmazásokban. Amennyiben *korrelálatlan* az adatsor (például egy kockadobások eredményeiből álló sorozat esetén), úgy $C(k) = 0$, $k \neq 0$-ra. Ekkor a sorozat elemei egymástól függetlenek, valószínűségük nem függ a többi pozícióban található elemektől.

Rövid távú korrelációkról akkor beszélhetünk, ha létezik olyan R *karakterisztikus távolság*, amely szerint $C(k)$ exponenciálisan lecseng (például Markov-folyamatok esetén, ld. a 2.3 fejezetben):

$$C(k) = C_0 e^{-k/R}. \tag{2.4}$$

Amennyiben nem létezik ilyen R karaktertisztikus távolság, hanem $C(k)$ hatványfüggvény szerint cseng le,

$$C(k) = C_0 k^{-\gamma}, \tag{2.5}$$

akkor *hosszú távú korreláció* van jelen a szimbólumszekvenciában. Kimutatható, hogy n-ed rendű Markov-folyamatok nem produkálhatnak hosszú távú korrelációkat, azok lecsengenek n nagyságrendjében.

A korrelációk a körülöttünk lévő világ sokféle jelenségének leírásánál játszanak központi szerepet. A statisztikus fizikában egy sor ún. *kritikus jelenséget* – úgy mint a fázisátalakulásokat (pl. olvadás, forrás) – e fogalom nélkül nem érthetünk meg. A tözsdei folyamatokat és a szívritmust vizsgálták többek között az elmúlt évtizedben ilyen szempontból, és az alábbiakban írott szövegekre bemutatandó módszereket DNS-szekvenciákra is alkalmazták.

Az írott szövegek *korrelációmentes* statisztikai tulajdonsága a szekvenciát alkotó alapábécé egyes elemeinek (hangoknak, betűknek vagy szavaknak) a gyakorisága. Ennek vizsgálata sok érdekes eredménnyel szolgált már a hagyományos statisztikai nyelvészetben is (ld. pl. Nagy Ferenc [1986]), de ezen a ponton nem kívánok ezzel hosszabban foglalkozni. Az alábbiakban a szövegek rövid, majd hosszú távú korrelációinak kimutatásával és modellezésével fogok foglalkozni.

2.2. Rövid távú korrelációs jelenségek: A vektor-tér technika

Az elmúlt években egyre több algoritmus születik szövegek automatikus szelektálására: például egy hírügynökség számára nagyon hasznos lenne, ha a befutó híreket emberi beavatkozás nélkül lehetne nyelv és témakör szerint csoportosítani. Damashek [1995] éppen egy ilyen algoritmusra tesz javaslatot, melyet ő *Acquaintance*-nek nevez.

Ennek az algoritmusnak a lényege az, hogy a szövegekből vektorokat készít, majd a vektorok skaláris szorzatát kiszámítva kapjuk meg a szövegek hasonlóságának valamilyen ,,mértékét" (ha nem is a szó egzakt matematikai értelmében). A vektor, mint azt diszkutálni fogom, a szöveg rövid távú korrelációira jellemző. Vagyis a módszer a szövegeket a rövid távú korrelációs tulajdonságok alapján képes szelektálni.

A módszert korábban DNS-szekvenciák csoportosítására sikerrel alkalmaztam a fizikus szakdolgozatomban (Bíró [1998a,b], Bíró et al. [1998]).

2.2.1. A Damashek-módszer ismertetése

Képzeljük el, hogy egy n hosszúságú ablakot (a szekvencia és a számítógépes kapacitás véges volta miatt általában $n = 3...6$) mozgatunk végig a szövegen, karakterről karakterre. A különböző lehetséges karakter-n-eseket indexeljük $i = 1...J$-vel. Jelöljük m_i-vel az i-ik

karakter-n-es előfordulásainak a számát. A szövegből képezett $\mathbf{x}$ vektor i-ik komponensét éppen az m_i-k 1-re való lenormálásából kapott frekvenciaként definiáljuk:

$$x_i := \frac{m_i}{\sum_{j=1}^{J} m_j} \tag{2.6}$$

Ha van két szövegünk, a belőlük ilyen módon képezhető vektorok, $\mathbf{x}$ és $\mathbf{y}$, skaláris szorzata jellemzi a szövegek hasonlóságát:

$$S = \frac{\sum_{j=1}^{J} x_j y_j}{\left(\sum_{j=1}^{J} x_j^2 \sum_{j=1}^{J} y_j^2\right)^{1/2}} = \cos\theta \tag{2.7}$$

A matematikailag műveltek számára fontos megemlíteni, hogy a $d(\mathbf{x},\mathbf{y}) := 1 - S$ távolság nem definiál metrikát a vektorok terében, azaz nem távolságfogalom a szövegek között, a szó matematikai értelmében. Ugyanis igaz, hogy $d(\mathbf{x},\mathbf{y})$ pozitív, szimmetrikus, és $d(\mathbf{x},\mathbf{y}) = 0$ akkor és csak akkor teljesül, ha $\mathbf{x} = \mathbf{y}$, mivel utóbbiak az 1-es norma szerint lenormált vektorok. És attól a valószínűtlen esettől eltekintve, amikor két különböző szöveg ugyanarra a vektorra képezhető le, ez azt is jelenti, hogy a két szöveg azonos. Ugyanakkor belátható, hogy a háromszög-egyenlőtlenség nem teljesül, például egy síkban fekvő, kis szöget bezáró három vektorra.[5] Ennek ellenére, a szó informális értelmében, mondhatjuk, hogy S a szövegek hasonlóságát méri, $1 - S$ pedig a szövegek távolságát.

Ez a szorzat elsősorban a szövegek nyelv szerinti szétválasztására alkalmas, de szerencsés esetben az azonos nyelvű szövegek téma szerinti szétválasztása is lehetséges. Saját magam például öt, fizikai témájú angol e-mailből (Ph1-Ph5), három szintén angol nyelvű, de más témájú e-mailből (E1-E3), két francia levélből (Fr1-Fr2) és három magyar nyelvű, magánjellegű e-mailből (H1-H3) álló korpuszt vizsgáltam. Az egyes szövegek hossza 3400 és 6000 karakter között volt, kivéve a két francia levelet, amelynek hossza csupán 1000 - 1200 karakter. Az angol ábécé 26 betűjét, a pontot, a vesszőt, valamint az üres helyet vettem figyelembe ($r = 29$-féle karakter). A többi karaktert ignoráltam, továbbá kis- és nagybetű között, ill. ékezetes és ékezet nélküli betű között – mint azt az e-mailekben megszoktuk – nem tettem különbséget. A több üres helyből álló szekvenciákat előzőleg törölni kell. Eredményeimet $n = 3$-ra és $n = 4$-re a 2.1, ill. a 2.2 táblázat tartalmazza (Bíró [1997a,b, 1998a,b]).

[5] Legyenek például $\mathbf{x}$, $\mathbf{y}$ és $\mathbf{z}$ egy síkban fekvő vektorok ilyen sorrendben. A szomszédos vektorok zárjanak be 30°-os szöget. Ekkor $d(\mathbf{x},\mathbf{y}) = d(\mathbf{y},\mathbf{z}) = 1 - \frac{\sqrt{3}}{2} \approx 0,134$, míg $d(\mathbf{x},\mathbf{z}) = 0,5$, vagyis $d(\mathbf{x},\mathbf{y}) + d(\mathbf{y},\mathbf{z}) \geq d(\mathbf{x},\mathbf{z})$ nem teljesül.

2.1. táblázat: Angol nyelvű fizikai (Ph), egyéb témájú angol szövegek (E), francia (Fr) és magyar (H) levelek (`.txt`*-fájlok) alapján készített, a karakterhármasok gyakoriságát jellemző ($n = 3$) vektorok szorzata. Látható, hogy az azonos nyelvű szövegek egymás között szignifikánsan magasabb szorzatokat adnak, mint a különböző nyelvű szövegek szorzatai. A módszer téma szerinti válogatásra is alkalmas, hiszen két E-szöveg, ill. két Ph-szöveg hasonlósága 15%-kal magasabb, mint egy E- és egy Ph-szöveg szorzata.*

A két táblázat adatai között szignifikánsan magasabbak az azonos nyelvű szövegek vektoraiból képezett vektorok szorzatai ($n = 3$-ra $0,73 \pm 0,06$), mint a különböző nyelvűek szorzatai ($0,25 \pm 0,024$). A két, különböző témájú angol szövegcsokor is elkülönül egymástól: a Ph-szövegek (pl. $n = 3$-ra a Ph-szövegek egymás közötti szorzata: $0,79 \pm 0,024$) ill. az E-szövegek szorzata ($0,80 \pm 0,015$) magasabb, mint egy E- és egy Ph-szöveg szorzata($0,68 \pm 0,03$).

Damashek a módszert továbbfejleszti. Bevezeti az ún. *centroid vektort*, amely egy-egy szövegcsokor (például az angol nyelvű szövegek, vagy az angol nyelvű, számítástechnikai témájú szövegek) vektorainak az átlaga. Ezt a vektort levonva a szövegek vektoraiból, kitranszformálhatóak a szövegcsokor közös jellemzői, például az adott nyelv funkcionális-grammatikai szavaiból adódó jellegzetességek (a magyar esetén pl. az "␣a␣" hármas vagy az "␣az␣" négyes, ahol '␣' az üres karaktert, a *space*-t jelenti). Ilyen módon, a szövegek

2.2. táblázat: A 2.1 táblázatban használt szövegek hasonlósága, $n = 4$ karakterből álló "ablakot" használva. A szorzatok értéke kisebb, az eltérések még szignifikánsabbak. Viszont nagyobb n csak hosszabb szövegekkel használható: a többi szövegnél jóval rövidebb francia levelek hasonlósága mindkét táblázatban kisebb, mint a többi, azonos nyelvű szövegpárok hasonlósági mértéke.

finomabb csoportosítása is lehetséges, tartalom, esetleg stílus szerint is.

A 2.2 táblázatban látható, hogy a rövidebb francia szövegek esetén kevésbé látványos az eredmény. Ezért a rövid szövegek statisztikai hibáinak a kikerülésére azt javasolja Damashek, hogy rendeljünk egy-egy nulla és egy közötti λ súlyt az egyes vektorokhoz, kisebbet a rövidebbekhez és nagyobbat a hosszabbakhoz, majd a két dokumentum ezen súlyával, mint szorzófaktorral, korrigáljuk a centroid vektorok levonását követően kapott szorzatot.

DNS-szekvenciák esetén az ún. *kódoló* és *nem kódoló* részeket sikerült kettéválasztani, látványos módon (Bíró [1998a], Bíró *et al.* [1998]). A kódoló szakaszok (exonok) közvetlenül kódolnak aminosav-szekvenciákat (fehérje-részleteket), mint ahogy azt minden alapszintű genetikai könyvben olvashatjuk (például Berend [1980]-ban) a genetikai ábécéről. Ezzel szemben, az intronok szerepe általában még nem tisztázott: valószínűleg vagy közvetett szerepük van, vagy csupán ,,evolúciós szemetek", azaz a múltból meg-

maradt, de szerepüket elvesztett szakaszok (Weaver & Hedrick [1992]). E két típust sikerült határozottan különválasztani Damashek módszerével.

2.2.2. A Damashek-módszer továbbfejlesztése, diszkussziója szövegekre

A módszert én inkább más irányba fejlesztettem tovább: összefűztem mind a négy szövegtípusból (E, Ph, Fr, H) néhány szöveget egyetlen fűzérré.[6] Egy $m = 100$ hosszúságú dobozt mozgattam végig ezen a fűzéren, és a doboz éppen aktuális tartalmából képezett vektort ($n = 3$ ablak mellett) összeszoroztam a fűzér elejéből (0-6775. karakterek közötti angol nyelvű Ph-szövegekből) képezett vektorral. Az eredményeket a 2.1. ábrán mutatom be. A fűzér angol és nem angol nyelvű részei élesen különválnak egymástól: ott, ahol a doboz nem angol szöveget tartalmazott, drasztikusan leesett a doboz tartalmának és az angol nyelvű referenciaszövegnek a hasonlósága.

A módszer alkalmasnak tűnik arra, hogy különböző jellegű szövegekből összetett szekvenciákat a komponenseire bontsunk szét, legalábbis olyan pontossággal, amilyet a mozgó doboz mérete lehetővé tesz. A doboz minimális méretét viszont n határozza meg, a kis n-ek viszont kevésbé nyelv- vagy szövegspecifikus vektorokat adnak. Talán e módszert továbbfejlesztve, a jövőben a filológusok kezébe is egzakt módszereket adhatunk?

Az eredményt röviden azzal magyarázhatjuk, hogy különböző nyelvekre különböző karakter-szekvenciák jellemzőek. Például az angol nyelvű szövegekben, az $n = 3$ esetben kiugrik a "the" karakterhármas: gondoljunk a határozott névelőn kívül a névmásokra és egyéb gyakori szavakra ("they", "their", "them", "these", "there", stb.).

Ezt a gondolatot tovább kifejtve, a Damashek-módszer sikerét három tényezőre vezethetjük vissza: két nyelvi és egy nyelven kívüli tényezőre.

Az első tényező lexikai, vagy szintaktikai-szemantikai: az adott nyelv mondattana meghatározza a nyelvre jellemző funkcionális morfémákat (névelőket, kötőszavakat, előljárókat, névmásokat, ragokat,...), míg a szöveg témája a gyakori, tartalommal bíró szavakat. Ezek a szövegre jellemző szavak erősen befolyásolják a domináns karakter-n-eseket. A nyelvészet ezen ágához tudtommal nincsenek olyan sztochasztikus modellek, amelyekkel a Damashek-módszer alaposabban diszkutálható lenne.

A második tényező a nyelv fonológiája, méghozzá a hangtan *fonotaktikának* nevezett ága, amely a markovi tulajdonságokat, vagyis a lehetséges hangkapcsolatokat írja le.[7] Úgy

[6] A file szerkezete: 0-6775. karakter: Ph-típus; 6776-13551. karakter: E-típus; 13552-23219. karakter: Ph-típus; 23220-25862. karakter: H-típus; 25863-32319. karakter: Ph-típus; 32320-35178. karakter: Fr-típus.

[7] Egy *corpus-based* statisztikus megközelítés különbséget tud tenni a gyakori *teremt*,

2.1. ábra: Különböző nyelvű és tartalmú szövegeket fűztem, össze, majd egy $m = 100$ hosszúságú dobozt mozgattam végig ezen a füzéren. Lépésről lépésre kiszámítottam a doboz tartalmának és a sztring első 6775 karakteréből álló, fizikai témájú angol referenciaszövegnek a hasonlóságát, $n = 3$ ablak mellett. Az ábra a hasonlóság mértékét (a Damashek-féle szorzatot) ábrázolja a doboz kezdőpozíciójának a függvényében.

is fogalmazhatunk, hogy a fonotaktika foglalkozik a nyelv legrövidebb távú korrelációival. A fonotaktikai szabályok között találunk (majdnem) univerzálisakat, és ezzel magyarázható jelentős részben a különböző nyelveken írt szövegek nem túlságosan alacsony szorzata.

---

*bara***ck**, *rece***pt** szavak kevésbé agrammatikus volta, a ritka és idegen hangzású **ng***anaszán*, **scs***í*, *fá***jl** fokozottabb agrammatikussága (amit az anyanyelvi beszélő is érez), valamint az egyáltalán nem előforduló hangkapcsolatok nyelvi státusza között.

De a különbségek is jelentősek, amelyek pedig a szétválasztás sikerét eredményezik. A fonotaktikára nézve vannak, vagy könnyen készíthetők statisztikák, amelyeket össze lehetne vetni a Damashek-módszerrel. Továbbá elemezhető lenne hasonlóbb és jelentősen eltérő fonotaktikájú nyelvek (pl. magyar és grúz) viselkedése a Damashek-módszerrel szemben. Damashek [1995] cikke tartalmaz egy ilyen típusú ábrát (Fig. 3.), amely csoportosít 31 nyelvet a szövegminták hasonlósága alapján, de hiányzik a nyelvészeti tárgyalás.

Miután diszkutáltuk a Damashek-módszer sikerének két nyelvészeti tényezőjét, említsük meg a harmadik, nyelvészeten kívüli tényezőt is, a helyesírási hagyományt. Az "sz" pár csak azért jellemző a magyar szövegekre, mert a magyar helyesírás így jelöli a [s] hangot. Ugyan az a [š] hang a magyar nyelvű szövegekben '*s*'-ként, az angolban leggyakrabban '*sh*'-ként, a franciában '*ch*'-ként, míg a németben 'sch'-ként jelenik meg. Eme tényező szerepét sem szabad figyelmen kívül hagyni, hiszen a német szövegekből készült vektorok meghatározó komponensei lesznek az '*sch*'-t tartalmazó komponensek, amelyek értéke a magyarban szinte zérus. A helyesírási hagyomány hatását a módszerre fonetikai-fonológiai átírás alkalmazásával lehetne kiküszöbölni, de egyelőre nem állnak a rendelkezésemre kellő számban elég hosszú átírt szövegek.

(Az a tény, hogy az ékezeteket figyelmen kívül hagytuk, felfogható egy olyan helyesírási hagyományként, amely még a szokásos ortográfiánál is távolabb van a fonológiai vagy fonetikai átírástól.)

2.3. Markov-modellek, a Damashek-módszer magyarázata

Már utaltam néhányszor a Markov-modellekre. Tekintsük át az alábbiakban a legfontosabb tudnivalókat velük kapcsolatban, majd nézzük meg, hogyan lehet a segítségükkel leírni a szövegek rövid távú korrelációit, és hogyan alkalmazhatók a Damashek-módszer diszkutálására.

Sztochasztikus folyamatnak egy olyan fizikai (vagy bármilyen egyéb valóságos) rendszert vagy matematikai modellt nevezünk, amely "*egy valószínűségi sorozat által szabályozott szimbólumsorozatot produkál*" (Shannon & Weaver [1986], p. 55). Ezzel a jelzővel a nem determinisztikus, csupán statisztikai eszközökkel elemezhető idősorokat (szimbólumsorozatokat) szokás jelölni, legyen szó tőzsdei árfolyamok alakulásáról, vagy a szívverések között eltelt időintervallumokról. De a sztochasztikus folyamatok prototípusai kezdettől fogva a karakterszekvenciák, például egy természetes nyelven írott szöveg szimbólumainak sorozata. A szimbólumszekvenciák hagyományos sztochasztikus elemzési eszközei a *Markov-láncok* és *Markov-modellek*. Shannon & Weaver [1986] látványos példákkal il-

lusztrálja mindezt.

Egy n elemű ábécé (szimbólumhalmaz) feletti *Markov-láncot* egy P mátrix definiál, ahol a p_{ij} mátrixelem $(i, j = 1...n)$ annak a feltételes valószínűségét jelenti, hogy a ,,mondat" valamely pozíciójában az ábécé j-ik betűjét találjuk, feltéve, hogy a megelőző pozícióban az i-ik betűt találjuk. Az első pozíció karakterének a valószínűségeit egy külön vektorban kell megadni. Az ún. *ergodikus Markov-láncok* esetén (a lánc elejétől elég távol) az egyes szimbólumok előfordulási valószínűsége független a pozíciótól.

Magasabb rendű (n-ed rendű) Markov-láncot úgy kapunk, ha a következő karakter valószínűsége a megelőző n betűtől függ. Tehát az előző bekezdésben leírt Markov-lánc az $n = 1$ esetnek felel meg, míg a 0-ad rendű Markov-lánc független szimbólumok sorozata.

A Markov-folyamatoknak egy általánosabb osztályát jelentik a *Markov-modellek*. Egy *Markov-modell* esetén adva van egy véges állapothalmaz ($\Omega = \{s_1, ..., s_n\}$), egy véges ábécé ($\Sigma = \{\sigma_1, ...\sigma_m\}$), egy $n \times n$-es P *állapotátmeneti mátrix* és egy $n \times m$-es A *jelkibocsátási mátrix*. A modell úgy ,,működik", hogy ha a modell az i-ik állapotban van, a_{ik} valószínűséggel kibocsátja a k-ik szimbólumot, majd p_{ij} valószínűséggel átmegy a j-ik állapotba, és újból jelet bocsát ki, stb.

Markov-lánc esetén a modell paramétereit, azaz a P mátrix elemeit megbecsülhetjük kellően nagy korpusz feldolgozásával, azaz a megfelelő karakter- és karakterpár-gyakoriságok empirikus megfigyelésével. Markov-modellek esetén viszont, általában nem ismert az, hogy adott korpusz mögött milyen, azt létrehozó állapotsorozat húzódik meg. Ez esetben meg kell találnunk, mely $S \in \Omega^*$ állapotsorozat esetén maximális a $P(S|O)$ feltételes valószínűség, ahol $O \in \Sigma^*$ a megfigyelt jelsorozat. (Az Ω^* és a Σ^* az Ω, ill. a Σ elemeiből képezett véges hosszúságú sztringek halmaza.) Azaz, Bayes törvénye miatt, maximalizálni kell a $P(O|S) \cdot P(S)$ szorzatot. E célra különböző iteratív algoritmusokat dolgoztak ki (Viterbi-algoritmus, stb.), melyek segítségével meg lehet becsülni a Markov-modell paramétereit ("rejtett Markov-modellek"; Rabiner [1989], Krenn & Samuelsson [1996]).

A következőkben próbáljuk meg a Damashek-módszer sikerességét megmagyarázni a Markov-modellekkel. Tekintsünk el a helyesírási hagyománytól, tegyük fel mondjuk, hogy fonológiai átírást használunk. Az alábbi gondolatmenet leginkább az előző fejezetben említett második tényezőre, a fonotaktikára alkalmazható. Ha a fonotaktika Markov-modelljének paramétereit korpusz alapján állapítjuk meg előzetesen, a nyelvre jellemző gyakori szavakat (főleg a grammatikai szavakat) szintén leírtuk. A szöveg témájára jellemző szavakról akkor adunk számot, ha a paraméterbecslés alapjául szolgáló referenciakorpuszt az adott témájú szövegek közül vesszük fel; de erre csak akkor van szükség, ha a finomabb, már nem a nyelv, hanem a téma szerinti szétválasztást akarjuk modellezni.

Tegyük fel, hogy a nyelvünk fonotaktikáját le tudjuk írni egy olyan Markov-modellel,[8] amely N állapotot $(s_1, s_2, ..., s_N)$ és M darab jelet $(\sigma_1, \sigma_2, ..., \sigma_M)$ tartalmaz. Az i-ikből a j-ik állapotba történő átmenet valószínűségét jelöljük p_{ij}-vel, míg a_{ij} annak a valószínűsége, hogy az i-ik állapotban a modell a j-ik jelet bocsátja ki. Tegyük fel, hogy ergodikus a Markov-modellünk; és mivel hosszú láncokról, reguláris nyelv hosszú modatairól van szó (ne felejtsük el, hogy a reguláris nyelvek osztálya zárt a * műveletre), feltehetjük azt is, hogy a modell már ,,egyensúlyivá vált". Azaz annak a v_i valószínűsége, hogy a szekvencia valamely pozíciójában a rendszer az i-ik állapotban lesz, független a pozíciótól. Ekkor a $\sigma_{k_1}\sigma_{k_2}...\sigma_{k_n}$ szimbólum n-es elméletileg jósolt valószínűsége, azaz a Damashek-vektor megfelelő komponense:

$$P(\sigma_{k_1}\sigma_{k_2}...\sigma_{k_n}) = \sum_{i_1=1}^{N} v_{i_1} a_{i_1 k_1} \sum_{i_2=1}^{N} p_{i_1 i_2} a_{i_2 k_2} ... \sum_{i_1=1}^{N} p_{i_{n-1} i_n} a_{i_n k_n}. \tag{2.8}$$

Mivel a p_{ij} és a_{ij} paramétereket nem ismerjük, a jövőbeni kutatások feladata lehet valamely adott korpusz esetén ezen ,,rejtett Markov-modell" (*hidden Markov Model*, Krenn & Samuelsson [1996], Rabiner [1989]) paramétereinek a becslése. A becsléshez rendelkezésre állnak a $P(\sigma_{k_1}\sigma_{k_2}...\sigma_{k_n})$-k empirikus közelítései, valamint felhasználhatjuk a jelek empirikus gyakoriságait. Az i-ik jel ν_i frekvenciájára adható elméleti becslés:

$$\nu_i = \sum_{j=1}^{N} v_j a_{ji}, \tag{2.9}$$

hiszen v_j valószínűséggel van a rendszer az s_j állapotban, és ekkor a_{ji} valószínűséggel bocsátja ki a σ_i jelet.

A Markov-folyamatok ugyanakkor nem elégségesek a természetes nyelvi szekvenciák modellezésére. Chomsky [1957] a *Syntactic Structures*-ben amellett érvel – és a generatív szintaxis azóta is alátámasztotta ezt a gondolatmenenetet –, hogy az angol, és általában

[8] Felmerülhet, hogy miért nem egyszerűsítem le a gondolatmenetemet Markov-láncokra. Két okból ragaszkodtam a Markov-modellhez. Egyrészt, a reguláris grammatikákat – amelyekkel leírhatjuk a fonológiát Kaplan & Kay [1994] szerint – Markov-modellekre, és nem Markov-láncokra vezethetjük vissza. A második ok nyelvészeti: a fonotaktikai szabályok sok esetben felhasználnak metrikus fonológiai információkat is (szótagszerkezet, szószerkezet, stb.). Ezt úgy valósíthatjuk meg, ha a metrikus szerkezet elemeit (pl. szótagkezdet, mag, szótagzárlat) az állapotokkal reprezentáljuk, míg a szegmentumoknak (fonémáknak) felelnek meg a Markov-modell jelei.

a természetes nyelvek szintaxisa nem írható le Markov-folyamatokkal, ill. reguláris grammatikával, csupán környezetfüggetlennel. A harmadik fejezetben olyan eredményeket ismertetek, amelyek a Markov-modellek elégtelen voltát kantitatív (statisztikai) empirikus adatokkal is alátámasztja. De ezt megelőzően, nézzünk további eredményeket, látszólag a rövid távú korrelációk köréből.

2.4 A Zipf-analízis

A következő alfejezetben leírtak eleinte egy ártatlan játéknak tűnnek, a korrelációmentes statisztika világából. Esetleg, ha betű-szinten vizsgáljuk a szöveget, rövid távú korrelációkat idéz. De, mint azt a 4. fejezetben látni fogjuk, a jelenség a hosszú távú korrelációkkal mutat szoros rokonságot.

G. K. Zipf már az 1930-as években végzett szógyakorisági vizsgálatokat természetes nyelven írott szövegekkel (Zipf [1935, 1949]). Összeszámolta nagy korpuszokban az egyes szavak előfordulási gyakoriságát, és sorrendbe helyezte a szavakat gyakoriságuk szerint: $R = 1$ jelentette a leggyakoribb szót, $R = 2$ a második leggyakoribbat, és így tovább. Ezután ábrázolta az ω gyakoriságot (frekvenciát) az R sorrend függvényében. Az eredmény meglepő volt: az $\omega(R)$ függvény nem exponenciálisan csengett le, ahogy várnánk, hanem hatványfüggvényszerűen:

$$\omega(R) \sim R^{-\zeta}, \tag{2.10}$$

ahol a ζ kitevő, nyelvtől és a korpusz tartalmától függetlenül, univerzálisan 1 körüli értéket vett fel: $\zeta \approx 1$. Ez azt jelenti, hogy log-log skálán ábrázolva, az $\omega(R)$ függvény egy $-\zeta$ meredekségű egyenest ad. A Czirók et al. [1995] által idézett irodalom szerint, a Zipf-analízist elvégezték városok, valamint ipari vállalatok méret szerinti eloszlása vizsgálatára is.

A módszer alkalmazása DNS-szekvenciák esetén felveti azt a kérdést, hogy vajon mit tekinthetünk "szónak" a genomban? Mantegna et al. [1994, 1995b] bevezeti a Zipf-analízis egy ésszerű általánosítását, amelyet magyarul ,,szimbólum n-es Zipf-analízisnek" nevezhetünk (*n-tuple Zipf-analysis*). Ennek lényege az, hogy egy n szimbólum hosszúságú ,,ablakot" mozgatunk végig a szekvencián, minden lépésben egyetlen (tehát nem n) karakterrel elmozdítva az ablakot, és az így kapott jel n-esek frekvenciáira számítjuk ki az $\omega(R)$ gyakoriság-sorrendben elfoglalt hely függvényt.

A szimbólum-n-es Zipf-analízis természetes nyelvi szövegek esetében éppúgy hatványfüggvényt eredményez, mint a hagyományos Zipf-analízis, viszont eltérő lesz az expo-

nens. Mantegna et al. [1995b] egy angol nyelvű enciklopédia cikkeit elemezték ilyen eszközökkel: a hagyományos, szavak gyakorisága alapján végzett Zipf-analízis $-0,85$ kitevőt eredményezett, míg a betű-n-esek ($n = 3,4,5,6$) $-0,57$ körüli meredekségekkel jellemezhető ábrákat produkáltak a $10 \leq R \leq 1000$ két nagyságrendet átfogó intervallumban. A két exponens különbségét a szerzők abban látják, hogy míg a hagyományos Zipf-analízisben csupán egy meghatározott szókincs elemei vesznek részt, addig az n-esekkel végzett Zipf-analízis során kibővül ez a halmaz.

Kipróbálták ezt a szimbólum-n-es Zipf-analízist mesterséges nyelvekre, méghozzá a UNIX operációs rendszer kompillált fájljaira is ($n = 6,8,10,12$). A 9000000 bit hosszúságú szöveg egy kételemű ábécéből ($\{0;1\}$) épül fel, és a DNS-adatokkal való könnyű összevethetőség érdekében (négy elemű ábécé) választották meg így az n-eket. Ebben az esetben is lineáris ábrákat kaptak a kétszer logaritmikus skálán, és a linearitási tartomány észrevehetően nőtt n növelésével. Az $n = 12$ mellett $\zeta = 0,77$ exponenst kaptak a szerzők.

Mantegna et al. [1994] és Mantegna et al. [1995b] az akkoriban hozzáférhető leghosszabb emlős, gerinctelen és virális DNS-szekvenciákra alkalmazta a Zipf-analízist. Ahhoz, hogy egy L hosszúságú szakaszt n-es Zipf-analízisnek lehessen alávetni, $L \gg 4^n$-nek teljesülnie kell (a cikkben $L > 10 \cdot 4^n$ teljesül). A cikkben az $50000bp$-nál (bázispárnál) hosszabb, vagy az ezt az értéket megközelítő hosszúságú szekvenciákat használtak fel, így az $n = 3,4,...7$-eseket lehetett csak vizsgálni, de a különböző n-ekre nem kaptak lényegesen eltérő eredményeket. Nyitott kérdés, hogy vajon a ζ értéke konstans vagy monoton nő, ha növeljük n értékét. E kérdés megválaszolásához az 1995-ben hozzáférhető szekvenciáknál jóval hosszabb szekvenciákra lett volna szükség.

A következő lépésben a szerzők az egyes szekvenciákat két-két jelsorozattá bontották szét: a GenBank-beli fájl információi alapján különválasztották az egyes szekvenciák ismert kódoló régióit a többi szakasztól. A nem kódoló szekvenciák a természetes nyelvi szövegekhez hasonló, (2.10) alakú hatványfüggvényt eredményeztek az $R \leq 4^{n-1}$ intervallumban. Ezzel szemben, a kódoló szakaszok Zipf-analízise az $R \leq 4^{n-1}$ tartományban *logaritmikus* függvénnyel közelíthető görbét adott, amely a lin-log skálán lineáris:

$$\omega = b - c \log_{10} R. \tag{2.11}$$

Ezt az eredményt Mantegna et al. [1995b] kellő számú és kellő hosszúságú szekvenciára kimutatta, $n = 3,4,..7$-re. Érdemes megjegyezni, hogy e cikk és Kanter & Kessler [1994] szerint (2.11)-hoz hasonló eloszlást követnek az ábécé betűinek gyakoriságai az emberi nyelveken írt szövegekben is.

Mit jelent ez a megfigyelés? Czirók et al. [1995, 1996] vizsgálja azt a kérdést, hogy

milyen típusú ,,szövegek" milyen Zipf-függvényt eredményeznek. Levezeti, majd szimulációval is igazolja azt, hogy a Markov-modellek Zipf-függvényei lépcsőzetesen futnak le, szemben a különböző módszerekkel előállított, hosszú távú korrelációt mutató szekvenciák hatványfüggvény szerinti lecsengéseivel. Utóbbiak jól közelíthetők Markov-modellekkel a középső tartományban, de a Zipf-plot két széle lényegesen eltér az illesztett Markov-modellből számított görbétől. (Másik magyarázatot javasol a Zipf-törvény és a Markov-folyamatok összekapcsolására Kanter & Kessler [1994].)

Vajon illeszthető-e Markov-modell a karakterszekvenciákból számított $\omega(R)$ függvényekre? Mantegna et al. [1995b] szerint az egyes szekvenciákból empirikusan számítható $p_{\sigma_1\sigma_2}$ átmeneti mátrix és v_σ gyakorisági vektor elemei alapján

$$P(\sigma_1\sigma_2...\sigma_n) = v_{\sigma_1} p_{\sigma_1\sigma_2}...p_{\sigma_{n-1}\sigma_n} \tag{2.12}$$

adja meg a $(\sigma_1\sigma_2...\sigma_n)$ bázis n-es gyakoriságát az illesztett első rendű Markov-láncban (ahol $\sigma_i = \mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}$). Ez a próbálkozás a kódoló DNS-szakaszok esetén elég jó közelítést ad. Viszont a nem kódoló részek esetén – amelyek, mint látni fogjuk, további statisztikai tulajdonságaikban is hasonlítanak az írott szövegekre –, az első másfél nagyságrendben szignifikáns az eltérés a megfigyelés és a modell között.

A további próbálkozások azt mutatják, hogy a Markov-lánc rendjének növelésével egyre jobban lehet ,,utánozni" a megfigyelt statisztikai tulajdonságokat. Mégis, összhangban a következőkben ismertetésre kerülő eredményekkel is, le kell vonnunk azt a következtetést, hogy a *markovi sztochasztikus folyamatok nem írják le elég jól az írott szövegeket, valamint a hozzájuk hasonló egyéb szekvenciákat, pl. a nem kódoló DNS-szakaszokat.*

Azok a szekvenciák – írott szövegek, nem kódoló DNS-szekvenciák, kompilált UNIX-fájlok –, amelyek Zipf-tulajdonságait nem modellezhetjük elég jól Markov-láncokkal, mert hatványfüggvény alakú Zipf-függvényt produkálnak, további tulajdonságok szempontjából is közös csoportba tartoznak, például a kódoló DNS-szakaszokkal szembeállítva. És ezek legfontosabbika az, hogy hosszú távú korrelációkat mutathatunk ki az esetükben. A 4.4 fejezetben bemutatom, hogy a Markov-láncok kudarca után adható olyan modell, amely a hosszú távú korrelációk létén túl, számot ad a Zipf-függvény hatványfüggvény alakjáról is.

3. fejezet

Hosszú távú korrelációk

3.1. Hosszú távú korrelációk kimutatása

Hosszú távú korrelációk létét több módon is kimutathatjuk valamely szekvenciában. Az első megoldás a $C(k)$ függvény kiszámítása közvetlenül, az adatokból, (2.3) alapján. Ennek viszont meglehetősen nagy a komputációs igénye. Viszont az alábbiakban bemutatandó ,,véletlen bolyongás-módszer" egy olyan mennyiséget vezet be, amely szintén kimutatja a hosszú távú korrelációkat, viszont könnyebben számítható.

A *véletlen bolyongás-modellt* először 1992-ben C.-K. Peng és társai, többségük a Bostoni Egyetem munkatársai (Peng et al. [1992], Stanley et al. [1992, 1993a,b], Peng et al. [1993]) alkalmazták, méghozzá DNS-szekvenciákra. Később más fizikusok használták fel a módszert írott szövegek elemzésére (Schenkel et al. [1993], Amit et al. [1994], Dietler & Zhang [1994], Ebeling & Neiman [1995]).

Az eljárás lényege a következő: először is a szöveget kódoljuk át valamilyen szabály szerint egy bináris sorozattá. Írott szövegek esetén a legegyszerűbb az angol ábécé 26 betűjét, a szóközt és az írásjeleket 5 biten kódolni.[9]

Képzeljünk most el egy bolhát, amelyet egy számegyenes origójába helyezünk. Felolvassuk neki a szekvenciát, és amennyiben 1-et hall, úgy előre ugrik egyet, míg 0 esetén hátra ugrik. Amennyiben u_i-vel jelöljük az i-ik lépést ($u_i = \pm 1$), úgy a bolha helyzete az

[9] DNS-szekvenciák esetén például információ-vesztő kódolást alkalmaztak, amennyiben például `C` és `T` helyett 1-et, míg `A` és `G` helyet 0-t írtak.

i-ik lépés után nyilván

$$y(l) = \sum_{i=1}^{l} u_i. \tag{3.1}$$

Az $y(l)$ mennyiség várható értéke kevés információt nyújt, hiszen azt az egyes betűk relatív gyakorisága határozza meg. Viszont annál érdekesebb mennyiség az $y(l)$-nek az elmozdulás átlaga körül vett $F(l)$ átlagos fluktuációja (*root mean square fluctuation*):

$$F^2(l) := \left\langle \left(\Delta y(l) - \langle \Delta y(l) \rangle\right)^2 \right\rangle = \left\langle \Delta y(l)^2 \right\rangle - \langle \Delta y(l) \rangle^2 , \tag{3.2}$$

ahol

$$\Delta y(l) = y(l_0 + l) - y(l_0), \tag{3.3}$$

és az összes lehetséges l_0 pozícióra kell az átlagolást képezni.

Az $F^2(l)$ szoros összefüggésben áll a (2.3)-ben definiált $C(k)$ autokorrelációs függvénnyel (Peng *et al.* [1992]):

$$F^2(l) = \sum_{i=1}^{l} \sum_{j=1}^{l} C(j-i). \tag{3.4}$$

Ebből következik, hogy – $C(k)$ viselkedésétől függően – $F(l)$ háromféle viselkedést mutathat. Mindhárom esetben $F(l)$ hatványfüggvény-szerű viselkedést követ,

$$F(l) \sim l^{-\alpha}, \tag{3.5}$$

és az eltérés az α-exponensben nyilvánul meg:

• Amennyiben tisztán véletlen sorozattal állunk szemben, azaz $C(k) = 0$ átlagban (kivéve a $k = 0$ esetet, hiszen $C(0) = 1$), akkor klasszikus, korrelálatlan bolyongásról van szó: $\alpha = 0,5$.

• Rövid távú korrelációk esetén az autokorrelációs függvény (2.4)-szerinti exponenciális lecsengése miatt az $F(l)$ – egy átmeneti állapot után – aszimptótikusan szintén $\alpha = 0,5$ kitevőjű hatványfüggvényt követ.

• Hosszú távú korrelációk esetén viszont, $C(l)$ nem jellemezhető R karakterisztikus hosszal, (2.5) szerint hatványfüggvényként viselkedik, és ezért $F(l)$ α-kitevőjének az értéke eltér a $0,5$ értéktől. Az eltérés mértéke jellemzi a hosszútávú korrelációk erősségét.

Az alábbi esetekben $0,5$ és 1 közötti értékekkel fogunk találkozni.

A statisztkus fizikában, ha különböző mennyiségek között hatványfüggvény alakú összefüggések vannak, az exponensek között gyakran sikerül univerzális *skálatörvényeket* felállítani. Például a (2.5)-beli γ és a (3.5)-beli α között érvényesül a következő összefüggés:

$$\alpha = \frac{1+\gamma}{2} \tag{3.6}$$

3.2. Kísérleti eredmények

Az eljárás alkalmazható a természetes nyelven írott szövegekre. Mint említettem, ebben az esetben egy-az-egyhez kódolást alkalmaztak, azaz átírták a szöveget egy bináris ábécé segítségével (szemben a "DNA-walk"-kal, ahol két-két bázist a legtöbbször ugyanúgy kódoltak). Például öt biten kitűnően lehet kódolni az angol ábécé 26 betűjét, a szóközt, és a legfontosabb írásjeleket.

Az eredmények nagyon érdekesek írott szövegek esetében. A vizsgálathoz használt szövegek között megtaláljuk a Biblia eredeti, héber nyelvű szövegét, valamint modern fordításait, továbbá Shakespeare-drámákat, regényeket, sőt egy szótárat is. Néhány érdekesebb eredmény:

- A szövegek sok nagyságrenden keresztül állandó α-exponenssel jellemezhetőek, melynek értéke általában $0,6 - 0,7$ közé esik (Schenkel et al. [1993]).
- A kitevő értéke nem jellemző a szerzőre, hiszen például a *Hamlet* exponense $0,56$, míg a *Rómeó és Júliáé* $0,60$ (Schenkel et al. [1993]).
- A fordítások során, úgy tűnik, a korreláció mértéke csökken. Habár a Biblia exponense kimagaslóan magas ($\sim 0,75$), a fordításai során ezen érték szisztematikusan csökken.[10] (Amit et al. [1994])
- A szótár a látszólag független szócikkek hosszánál jóval hosszabb korrelációkat mutat, amely jelenség nem magyarázható pusztán a tartalom korreláltságával. Ezek szerint a szótár szócikkeinek elrendezése "nem véletlenszerű", előzetes sejtésünkkel ellentétben, írja Schenkel et al. [1993].
- A szövegeket kisebb részletekre, például szavakra, mondatokra vagy $l = 10 - 10000$ karakter hosszúságú, azonos darabokra szabdalva és ezeket a darabkákat szabadon permutálva, a szabdalás hosszának nagyságrendjéig megmaradnak a korrelációk, azon túl

[10] Nem találtam a szakirodalomban világos választ arra a kérdésre, hogy a magas hatványkitevő miért nem lehet a héber nyelvnek vagy ortográfiának a jellegzetes tulajdonsága.

pedig eltűnnek (Schenkel et al. [1993]).

Lefordított számítógépprogramokat (`.exe`-file-okat) is megvizsgáltak, ezek esetében $0,9$-et is meghaladó kitevőt találtak (Schenkel et al. [1993]). Érdekes még a ,,humán véletlenszámgenerátor" vizsgálata is: az egyik szerző 0 és 9 között írt le számokat, "véletlenszerűen". A véletlenszerű eloszlásra való törekvés eredményeképpen, rövid távon ($l \sim 10$) anti-korrelációt fedezhetünk fel, de ennél hosszabb távon – a számokat generáló személy minden igyekezete ellenére – megjelentek a korrelációk. Ráadásul, a hatványkitevő értéke nem is állandó, tart az 1-hez! A szerzők ezt úgy magyarázzák, hogy – ellentétben a szövegek és a számítógépes programok írásával –, a véletlenszámok generálásának nincsen értelmes célja, vagyis a tudattalan tényezők, amelyeket felelőssé tesznek a hosszútávú korrelációkért, nagyobb szerepet játszhatnak (Schenkel et al. [1993]).

DNS-szekvenciák esetében, a csupán kódoló részeket tartalmazó szekvenciák $\alpha = 0,5$ eredményt adtak, azaz nem mutattak hosszú távú korrelációt. Ezzel szemben, a nemkódoló szakaszok a természetes nyelven írt szövegekkel és a számítógépprogramokkal mutattak hasonlóságot ebből a szempontból is, akárcsak a Zipf-függvénynél, mint láttuk.

Az $F(l)$-függvény szoros kapcsolatban áll a szimbólumszekvencia, mint jelsorozat Fourier-transzformáltjával is. Hosszú távú korrelációk esetén az $S(f)$ teljesítményspektrum szintén hatványfüggvény:

$$S(f) \sim f^{-\beta}. \tag{3.7}$$

A két exponens közötti elméleti összefüggés (skálatörvény) (Schenkel et al. [1993]):

$$\alpha = \frac{\beta + 1}{2}. \tag{3.8}$$

A DNS-szekvenciák diszkrét Fourier-analíziséből kapott eredmények jól visszaadják ezt az elméleti összefüggést is.

Hogyan használható ez a módszer különböző típusú szekvenciák szétválasztására? A 2.1 ábra esetében használt, mozgó dobozos módszert alkalmazta Stanley *et al.* [1993a] az élesztő III. kromoszómáján. Minden lépésben kiszámolták a dobozon belüli szekvenciának a fentiekben bevezetett α exponensét, és ábrázolták ezt az értéket a doboz pozíciójának a függvényében. Az idézett cikk 9. ábrája illusztrálja ezen új algoritmus által produkált függvényt: ahol a pásztázó doboz kódoló szakaszba lép, leesik a függvényérték (az exponens, a pozíció függvényében) $0,5$-re, ahol pedig kilép a kódoló szakaszból, ott megnő az exponens $0,6$ fölé.

3.3. A Zipf-függvény és a hosszú távú korrelációk kapcsolata

A 2.4 fejezetben bevezetett Zipf-függvény valamint a most tárgyalt hosszú távú korrelációk között szoros összefüggés látszik. A tapasztalatokat úgy összegezhetjük, hogy létezik a szimbólumszekvenciáknak egy olyan osztálya, amely hosszú távú korrelációkat, azaz hatványfüggvény szerint lecsengő autokorrelációs függvényt, valamint szintén hatványfüggvény alakú Zipf-függvényt produkál. Ebbe az osztályba tartoznak például az írott szövegek, a nem kódoló DNS-szakaszok, valamint a számítógépes programok. Ezzel szemben, a kódoló DNS-szakaszok például nem mutatnak hosszútávú korrelációkat, és a Zipf-függvényük is exponenciálisan cseng le, akárcsak például a betűk eloszlása a szövegekben.

Az alábbiakban két további módszert szeretnék felvázolni, amelyek szintén alkalmasak ezen speciális szekvencia-osztály körülhatárolására. (Részletesebb tárgyalásuk megtalálható Bíró [1998a]-ban.)

Mindkét fogalom az információelméletből meríti az ihletétést. Az *Ivo Große* és *Hanspeter Herzel* által bevezetett *kölcsönös információ-függvény* (Herzel *et al.* [1995, 1997a, b], Große *et al.* [1997]) a szimbólumok statisztikai függetlenségét, ill. a korrelációk erősségét méri, és a következőképpen definiálhatjuk:

$$I(k) := \sum_{i,j=1}^{\lambda} p_{ij}(k) \cdot \log \frac{p_{ij}(k)}{p_i \cdot p_j}, \tag{3.9}$$

ahol a λ elemű ábécénk i-ik elemének az előfordulási valószínűségét p_i-vel jelöljük, míg $p_{ij}(k)$ annak a valószínűsége, hogy valamely pozícióban az i-ik betű fordul elő, és tőle k távolságra (előre) a j-ik.

Große és Herzel a kódoló és a nem kódoló DNS-szakaszokat élesen el tudta egymástól választani a kölcsönös információ-függvény viselkedése alapján. Nem tudok arról, hogy ezt a mennyiséget írott szövegek esetén is használták volna, de valószínűnek tűnik, hogy ebből a szempontból is a nem kódoló szakaszokkal ,,tartanának" az írott szövegek, valamint a többi hasonló tulajdonságokkal rendelkező szimbólumszekvencia-típusok.

A következő fogalmat, a *redundanciát*, *Mantegna* [1995b] vezeti be konkrétan ezen szekvencia-osztály vizsgálatára, az általunk kitűzött szempontokból. Az ő jelöléseit használva, az *n-gramok entrópiája* (λ elemű ábécére)

$$H(n) := -\sum_{i=1}^{\lambda^n} p_i \log_2 p_i, \tag{3.10}$$

ahol a p_i-k az egyes bázis n-esek valószínűségét jelölik. Ekkor a redundancia:

$$R := 1 - \lim_{n\to\infty} \frac{H(n)}{kn}, \tag{3.11}$$

ahol $k := \log_2 \lambda$. Mantegna megfogalmazása szerint a redundancia a nyelv *flexibilitásának* a megnyilatkozása (Mantegna et al [1994, 1995b]).

Nagy Ferenc [1986] becslése szerint 68% egy átlagos nyelv redundanciája (32 betű, 10 betűs szavak esetén). Shanon [1986] a hétköznapi angol nyelv redundanciáját durván 50%-nak becsüli. Ezt az eredményt adja mind a nyelv Markov-láncokkal történő approximációján alapuló entrópiaszámítás, mind a titkosírások elméletének néhány eredménye. Ezt támasztja alá az is, hogy a kísérletek szerint egy angol szöveg általában rekonstruálható, ha a karakterek felét véletlenszerűen töröljük. Említésre méltó még Shanon eszmefuttatása arról, hogy milyen mértékű redundancia esetén lehetséges keresztrejvényt készíteni: ha túl sok megszorítás vonatkozik a nyelvre, nyilván nehéz nagy keresztrejtvényt szerkeszteni. Viszont zérus redundancia esetén bármely betűsorozat létező szót ad, vagyis a karakterek tetszőleges két dimenziós elrendezése keresztrejtvény. Shanon szerint 50%-os redundanciáig lehetséges két dimenziós, míg 33%-os redundanciáig három dimenziós rejtvényt készíteni, ha a nyelvre vonatkozó megszorítások véletlen természetűek.

Mantegna *et al.* megfigyelései szerint a nem kódoló DNS-darabok redundanciája szignifikánsan magasabb a kódoló szekvenciák redundanciájánál, míg az összetett genomdaraboké a kódoló és a nem kódoló szakaszok arányától függően magasabb vagy alacsonyabb. Az írott szöveg redundanciája – a nem kódoló szakaszokéhoz hasonlóan – magas.

Annak a jelenségnek, miszerint létezik egy ennyi szempont szerint elkülöníthető szekvenciatípus-osztály, egyik magyarázata az lehetne, hogy az említett szekvenciaosztályban egységes struktúráló elvek működnek. A redundancia a DNS-szakaszok esetében jó példát szolgáltat: a kódoló DNS-sorozatok minden részlete sok információt hordoz – azaz kicsi a redundancia –, mivel egymástól független (gyengék a rövid távú korrelációk is, ld. Bíró [1998a], 4.8. ábráját) önálló egységekből áll: a kódolt fehérje egy-egy részletét tartalmazza minden karakter-hármas. Ezzel szemben, a nem kódoló szekvenciák esetén létezhetnek olyan ,,globális" struktúrák, amelyek csökkentik az egyes szimbólumok vagy szimbólum n-esek információtartalmát, és globális korrelációkat hoznak be.

A 4. fejezetben az X-vonás elméletről mutatom be, hogy potenciálisan alkalmas ilyen jelenségeket produkáló szervezőelvként funkcionálni. De ennek csak akkor lesz magyarázó értéke, ha helyesnek bizonyul az 1.1 fejezetben felállított hipotézis, amely szerint tényleg kimutatható az X-bar struktúra a szövegek magasabb szintjein is.

4. fejezet

A sztochasztikus veremautomata és alkalmazása

4.1. Környezetfüggetlen sztochasztikus eszközök

Ha a Markov-folyamatokat a reguláris nyelvek és a véges állapotú automaták sztochasztikus kiterjesztéseinek tekintjük, amennyiben hasonló elven működnek, de valószínűségi fogalmakkal bővülnek, akkor kereshetjük a környezetfüggetlen eszközök sztochasztikus megfelelőit is.

A formális nyelvek elméletében és az automataelméletben egy szöveghalmaz akkor tartozik valamely nyelvosztályhoz, ha az adott nyelvosztály eszközeivel generálható. Az általam javasolt, statisztikai nyelvészeti megközelítésben viszont, a szöveghalmaz statisztikai tulajdonságairól kell eldönteni, hogy az adott nyelvosztály eszközeivel modellezhetők-e. A 2. fejezetben arra láttunk példákat, hogy az empirikus adatok nem adhatók vissza Markov-folyamatokkal, vagyis reguláris nyelvekkel (véges állapotú automatákkal).

Jelen fejezet célja az, hogy kitalálja azokat a sztochasztikus eszközöket, amelyekkel vissza lehet adni ezeket a jelenségeket. Méghozzá úgy, hogy az eszközök a modern kvalitatív nyelvészet eszközeinek plauzibilis kvantitatív megfelelői legyenek, azaz a sztochasztikus modellt összhangba lehessen hozni a megszokott nyelvészeti modellekkel.

Ha a reguláris eszközök nem megfelelőek, térjünk át a környezetfüggetlen eszközökre. Ezt tette Chomsky is a *Syntactic Structures*-ben, és ezt tesszük mi is.

A környezetfüggetlen nyelvek sztochasztikus megfelelői a SCFG-k (*Stochastic Context Free Grammar*), amelyet Krenn & Samuelsson [1996] alapján foglalok röviden össze.

4.1.1. Sztochasztikus környezetfüggetlen automaták

A generatív nyelvészet által használt formális nyelvek elméletének klasszikusan ismert formája nem teszi lehetővé, hogy a korrelációk statisztikus jelenségének nyelvészeti vonatkozásait megvizsgálhassuk, vagy a korrelációk létére nyelvészeti magyarázatot találjunk. Ehhez az szükséges, hogy a formális nyelvek elméletét kiegészítsük sztochasztikus eszközökkel. A környezetfüggetlen grammatikák ilyen általánosításait *sztochasztikus környezetfüggetlen grammatikák* (*SCFGs*) néven ismerik. A reguláris nyelvek osztályának sztochasztikus általánosításai lényegében a Markov-modellek, tágabb nyelvosztály általánosításairól pedig nincs tudomásom. Utóbbiakra viszont egyelőre nem is lesz szükség, mivel a formális nyelvek elméletének legfontosabb felhasználói (nyelvészet, számítástudomány) leggyakrabban az $\mathcal{L}_2$ nyelvosztállyal dolgoznak.

Ha a generatív grammatikák négyesét egy valószínűségi függvénnyel egészítjük ki, sztochasztikus grammatikákat kapunk. Ilyen módon nem csak azt mondhatjuk meg, hogy egy adott mondat vajon eleme-e a grammatika által generált nyelvnek, hanem azt is, hogy milyen valószínűséggel vezethetjük le az S mondatszimbólumból az adott szekvenciát. Ezt a valószínűséget úgy értem, hogy amennyiben összefűzzük a ,,végtelen" sok mondatát, azaz képezünk egy *ideális szövegkorpuszt*, akkor az adott mondat milyen gyakorisággal fordul elő a korpuszban. Az ideális korpusz jó közelítését adhatja természetes nyelvek esetén egy könyv (pl. egy regény), a ,,DNS-nyelv" esetén pedig a DNS-szekvenciákat tartalmazó adatbázisokat tekinthetjük ilyen korpusznak.

A nyelvészet hagyományos szemléletmódja két ponton is eltér ettől a megközelítéstől. Először is, a nyelvi adatok helyes voltát binárisan kezeli: egy alak vagy grammatikus vagy agrammatikus, vagy eleme a nyelvnek, vagy nem. Ezen megközelítés hátrányait ecseteli meglehetősen irónikus hangnemben, és a statisztikus-sztochasztikus szemlélet térnyeréséért tőr lándzsát Abney [1996]. A másik eltérés pedig abban van, hogy a fonológiai grammatikalitást nem *korpuszra* (*corpus-based*), hanem a szókincsre, azaz a *lexikonra* (*lexicon-based*) vizsgálják, vagyis nem tesznek különbséget aközött, hogy például egy rendhagyó hangkapcsolat egy gyakori vagy egy ritka szóban fordul-e elő. (Ennek az az oka, hogy a hagyományos generatív nyelvészet mind a ritka, mind a gyakori szót a nyelv egyenrangú elemének tekinti, és az előfordulásban mutatkozó különbségeket a performanciához, azaz a nyelvészet vizsgálati körén kívülre sorolja.)

A környezetfüggetlen grammatikák sztochasztikus általánosítását a következőképpen definiálhatjuk Krenn & Samuelsson alapján:

4.1. Definíció: Sztochasztikus környezetfüggetlen grammatikának (Stochastic Context-free Grammar, SCFG) nevezzük az (V_N, V_T, S, R, P) *ötöst, ahol:*

V_N a nemterminálisok véges halmaza,

V_T a terminálisok véges halmaza (legyen $V := V_N \cup V_T$),

$S \in V_N$ a nemterminálisok közül a kitüntetett mondatszimbólum,

R az $X \to Q$ alakú levezetési szabályok egy véges halmaza, ahol $X \in V_N$ és $Q \in V^*$,

P pedig egy $R \to [0,1]$ függvény oly módon, hogy

$$\forall X \in V_N : \sum_{Q \in V^*} P(X \to Q) = 1.$$

A $P(X \to Q)$ valószínűség azt jelenti, hogy ha egy levezetési láncban megjelenik egy X nemterminális, azt milyen valószínűséggel fogjuk Q-ként újraírni.

Ezek után, egy levezetési lánc, ill. levezetési fa valószínűségét úgy definiálhatjuk, mint az alkalmazott helyettesítési (újraíró) szabályokhoz rendelt valószínűségek szorzatát. Valamely mondat valószínűsége pedig a hozzátartozó levezetési fák (*parse trees*) valószínűségeinek az összege lesz. A módszer a számítógépes nyelvészetben hasznos elemző automaták (*parser*-ek) futási idejének optimalizálására. Ha a korpuszt úgy tekintjük, mint nagy számú, egymás mellé írt mondatszimbólumból („poli-S láncból") levezetett sztringet, úgy látható a kapcsolat a korpuszból számított empirikus gyakoriság és a fentebb bevezetett elméleti valószínűség között.

4.1.2 Becslés a Zipf-függvényre

Ahhoz, hogy a modell paramétereiből eljuthassunk a Zipf-törvényben szereplő $\omega(R)$ függvényhez, az alábbi gondolatmenetet javasolom: számítsuk ki az egyes terminálisok előfordulási valószínűségét egy „korpuszban", azaz mondatok végtelen láncában. (Ezt a számítást tudtommal még nem végezte el senki, mivel a sztochasztikus környezetfüggetlen grammatikákat jelenleg más célokra, például valószínűségeket is felhasználó parser-ek készítésére használják.)

$v_S(a)$-val jelölöm azt, hogy az $a \in V_T$ terminális várhatóan hányszor fordul elő egy S-ből levezethető mondatban:[11]

$$v_S(a) := \sum_{\sigma \in V_T^*} p_S(\sigma) \left(\frac{\sigma}{a} \right),$$

[11] Az alábbiakban a kisbetűk mindig terminálisra, a nagybetűk nemterminálisra, míg a görög betűk terminálisokból álló sztringre fognak utalni.

ahol $p_S(\sigma)$ jelöli a $\sigma \in V_T^*$ mondat fentebb bevezetett valószínűségét, $\binom{\sigma}{a}$ pedig az a terminálisok számát a σ mondatban. (Megjegyzem, hogy habár formálisan nem látom be, de ezen végtelen szummának konvergensnek kell lennie. Hiszen felülről becsülhető a mondatok hosszának várható értékével, amely viszont véges, hiszen enélkül a kommunikáció nem lenne elképzelhető.)

A bennünket érdeklő $v_S(a)$ várható értéknél többet fogunk kiszámítani a SCFG paramétereiből, azaz a P valószínűségfüggvényből. Jelölje $p_A(\sigma)$ bármely $A \in V_N$ nemterminálisra annak a valószínűségét, hogy A-ból a $\sigma \in V_T^*$ sztringet vezetjük le: az A gyökerű, σ-t eredményező levezetési fák valószínűségeinek[12] az összegét. és jelölje $v_A(a)$ annak a várható értékét, hogy az A-ból sztringekben hányszor szerepel az a terminális:

$$v_A(a) := \sum_{\sigma \in V_T^*} p_A(\sigma) \binom{\sigma}{a}. \tag{4.1}$$

Ezek a $v_A(a)$-k egy $\mathbf{v}(a)$ vektornak a komponensei, amely egy, a V_N elemeinek a számával megegyező dimenziójú térben van.

A $p_a(\sigma)$ átírásához be kell vezetni a *Chomsky-féle normálalak* fogalmát. Belátható (pl. ld. Révész [1979], SCFG-re Krenn & Samuelsson [1996]), hogy bármely környezetfüggetlen grammatikához létezik olyan Chomsky-féle normálalakban megadott grammatika (*Chomsky Normal Form*, *CNF*), amely ugyanazt a nyelvet generálja, és amelynek a szabályai

$$A \to BC$$

vagy

$$A \to a$$

alakúak, ahol $A, B, C \in V_N$ és $a \in V_T$. Tetszőleges SCFG-hez is adható meg olyan SCFG, amelynek R-je CNF-alakú levezetési szabályokat tartalmaz, és amely minden sztringet ugyanolyan valószínűséggel generál, mint az eredeti SCFG.

Definiáljuk még a $\pi(a)$ vektort és a μ mátrixot:

$$\pi_A(a) := P(A \to a)$$
$$\mu_{AB} := \sum_{C \in V_N} \Big[P(A \to BC) + P(A \to CB) \Big]$$

[12] Az ilyen fák valószínűségét szintén az alkalmazott levezetési szabályok valószínűségeinek a szorzataként számíthatjuk ki, akárcsak az S gyökerű fák esetén.

Amikor az A-ból akarom levezetni a σ-t egy CNF alakú grammatikában, két lehetőségem van az elindulásra: amennyiben a σ egyetlen a terminálisból áll, úgy az $A \to a$ szabályt kell alkalmaznom, egyébként pedig az A-t egy $A \to BC$ szabállyal írom újra, és B-ből levezetem σ_1-et, C-ből σ_2-t, ahol $\sigma = \sigma_1\sigma_2$ alakú (e kettő konkatenációja). Ennek megfelelően:

$$p_A(\sigma) = \sum_{b \in V_T} P(A \to a)\delta_{\sigma,b} + \sum_{B,C \in V_N} \sum_{\substack{\sigma_1,\sigma_2 \in V_T^* \\ \sigma = \sigma_1\sigma_2}} P(A \to BC)p_B(\sigma_1)p_C(\sigma_2) \qquad (4.2)$$

(A $\delta_{\sigma,b}$ szimbólum a Kronecker-deltát jelöli: értéke akkor és csak akkor 1, ha a két indexe azonos, egyébként zérus.) Ha beírjuk a (4.2) egyenletet (4.1)-be, úgy bizonyos egyszerűsítésekre lesz lehetőségünk, a következő három összefüggés felhasználásával:

$$\left(\frac{b}{a}\right) = \delta_{b,a}$$

$$\sum_{\sigma \in V_T^*} p_C(\sigma) = 1$$

$$\left(\frac{\sigma}{a}\right) = \left(\frac{\sigma_1}{a}\right) + \left(\frac{\sigma_2}{a}\right)$$

Az eredmény a következő lesz:

$$v_A(a) = \pi_A(a) + \sum_{B,C \in V_N} P(A \to BC)(v_B(a) + v_C(a)),$$

amelyből:

$$v_A(a) = \pi_A(a) + \sum_{B \in V_N} v_B(a)\mu_{AB},$$

vagyis:

$$\mathbf{v}(a) = \pi(a) + \mu\mathbf{v}(a). \qquad (4.2b)$$

A (4.2b) összefüggés átrendezésével a SCFG paramétereiből ki tudjuk számítani az egyes terminálisok frekvenciáit. Ehhez az szükséges, hogy az $1 - \mu$ mátrix invertálható legyen, de ezzel azért nincs baj, mert a μ elemeire nincsen semmiféle megkötés ($\pi(\mathbf{a})$ kis megváltoztatására μ is megváltozik), vagyis kis perturbációval megszüntethető egy esetleges szingularitás.

4.1.3. A sztochasztikus veremautomata

Az alábbiakban definiálom a veremautomaták sztochasztikus megfelelőjét. Ha a sztochasztikus környezetfüggetlen grammatikákat szimbólumszekvenciák generálására akarom használni, a számítógépes implementáció igényli az átfogalmazást az automata-elmélet nyelvére. Az e fejezetben leírtakat tudtommal más nem fogalmazta meg, mivel a SCFG-eket más területeken alkalmazzák.

4.2. Definíció: Sztochasztikus veremautomatának (Stochastic Push-Down Automaton, SPDA) nevezzük a $(\Sigma, A, V, a_0, v_0, P)$ *hatost, ahol:*

Σ *véges halmaz a kimenő ábécé,*

A *az állapotok véges halmaza,*

V *a véges veremábécé,*

$a_0 \in A$ *a kezdő állapot,*

$v_0 \in V$ *a veremfenékjel,*

$P : (A \times V) \times ((\Sigma \cup \{\lambda\}) \times A \times V^*) \to [0,1]$ *az állapotátmenet-függvény,*[13] *amelyre igaz az, hogy*

$$\forall a \in A \forall v \in V : \sum_{s \in S, b \in A, w \in V^*} P(a, v; s, b, w) = 1.$$

Jelen pillanatban nem tudom eldönteni, vajon szükséges-e kikötni, hogy a P függvény csak véges sok $(a, v; s, b, w) \in (A \times V) \times (\Sigma \times A \times V^*)$-re vegyen fel zérustól különböző értéket, mindenesetre az alábbiakban ez teljesülni fog.

Az SPDA a hagyományos veremautomatákhoz hasonlóan működik. Induláskor az automata állapota a_0, míg a veremben csupán a v_0 veremfenékjel szerepel. Minden lépésben, annak a valószínűsége, hogy az $a \in A$ állapotból $b \in A$ állapotba kerül az automata, miközben a verem tetején található $v \in V$ jelet felülírja a $w \in V^*$ szekvenciával, és az $s \in \Sigma$ karakter kerül kiírásra, adott a-ra és v-re, a különböző s-ekre, b-kre és w-kre éppen $P(a, v; s, b, w)$.

Az automata működése akkor áll le, ha a létrejött állapothoz és a verem tetején található szimbólumhoz nem tartozik zérustól különböző valószínűségű átmenet.

Felvehetnénk egy végállapothalmazt, de meg fogunk elégedni az üres veremmel való ,,elfogadással" is: ha az automata megáll, akkor fogadjuk csak el a keletkezett sorozatot, ha a végén csupán a veremfenékjel található a veremben. Különben ,,hibaüzenetet" kapunk.

[13] *A* λ *jel az üres sztringet jelöli.*

Ha a verem kiürül, általában újabb szekvencia generálása kezdődhet a P állapotátmenet-függvénytől függően. Ha ezen a ponton le akarjuk állítani az automatát, mert mondjuk eddig tart az egy mondatszimbólumból levezetett szöveg, P-t úgy adjuk meg, hogy ha P második argumentuma a veremfenékjel, akkor csak az a_0 állapot esetén létezzék nem zérus valószínűségű átmenet. Ugyanakkor, ezzel az első átmenettel, kikerül az automata az a_0 állapotból, és már soha nem fog ide visszatérni.

Egy alaposabb tárgyalás esetén formalizálni lehetne az SPDA-k működését, és bebizonyítani az SPDA-k és az SCFG-k ekvivalenciáját. Ez az ekvivalencia azt jelentené, hogy azonos sorozatot azonos valószínűséggel generálnak, és a bizonyítás váza hasonlítana a nem-sztochasztikus megfelelők ekvivalenciájának a bizonyításához. Informálisan, ahogy a nem-sztochasztikus megfelelők esetében az állapotátmeneteket le lehetett fordítani újraíró szabályokra, és fordítva, ugyanúgy meg lehet ezt tenni most is, és a valószínűségeket is meg lehet egymásnak feleltetni.

Az A Függelékben belátom, hogy minden SCFG-hez konstruálható vele ekvivalens SCFG: egy levezetési fa valószínűsége megegyezik az SPDA egy működésének a valószínűségével, azaz az SCFG levezetési fái és az SPDA levezetési láncai között ekvivalenciareláció létesíthető (azonos szekvenciát generálnak, azonos valószínűséggel). Majd belátom, hogy valamely karaktersorozathoz egymással ,,ekvivalens" levezetési fák és SPDA-működések tartoznak. Precízebben az A függelékben dolgozom mindezt ki.

Az alábbiakban az SPDA-k, mint modellek lehetőségeit vizsgáljuk meg, azt, hogy az írott szövegek korábbi fejezetekben említett statisztikai tulajdonságait visszaadják-e. És mindeközben a nyelvészeti modellekkel való összevethetőséget is szem előtt akarjuk tartani.

4.2. Hosszú távú korrelációk generálása SPDA-val

4.2.1 A véletlen bolyongó matematikai tulajdonságai

Ebben a fejezetben bemutatjuk, hogy SPDA-k képesek hosszú távú korrelációt tartalmazó szekvenciát generálni. Egy konkrét modellen, amit ,,véletlen bolyongónak" fogunk nevezni, megvizsgáljuk, hogy milyen matematikai (statisztikus fizikai) tulajdonságok jellemzők az íly módon generált szekvenciára, amelyek az írott szövegek megfigyelt tulajdonságaival mutatnak rokonságot.

Vegyük elő a 3.1 fejezetben munkára fogott bolhánkat, és helyezzük újból a számegyenesre. Tegyük fel, hogy p valószínűséggel ugrik egyet jobbra, $+1$ irányba, míg q valószínűséggel ugrik egyet balra, azaz -1-et. $(p+q=1)$ Közben, ha a bolha az origóban

van, 1-et írunk ki a kimenő szalagra, egyébként 0-t.

Hogyan fogalmazhatjuk ezt át sztochasztikus veremautomatává? A kimenő ábécé legyen $\Sigma = \{0, 1\}$, az állapothalmaz $A = \{a_0, a, b\}$, és a veremábécé $V = \{v_0, v\}$. a_0 a kezdőállapot és v_0 a veremfenékjel. A P átmeneti valószínűségfüggvény pedig legyen a következő helyeken zérustól különböző:

$$P(a_0, v_0; \lambda, a, v_0) = 1$$

$$P(a, v_0; 1, a, v_0 v) = p$$

$$P(a, v_0; 1, b, v_0 v) = q$$

$$P(b, v_0; 1, a, v_0 v) = p$$

$$P(b, v_0; 1, b, v_0 v) = q$$

$$P(a, v; 0, a, vv) = p$$

$$P(a, v; 0, a, \lambda) = q$$

$$P(b, v; 0, b, \lambda) = p$$

$$P(b, v; 0, b, vv) = q,$$

ahol $p + q = 1$. Micsoda P szemléletes jelentése? Az első sor kimozdítja a rendszert a kezdőállapotból, de ennek most nincs jelentősége, mert a verem kiürülése esetén is működik tovább a rendszer. A további szabályok a „bolha mozgását" írják le: az a állapot a számegyenes pozitív felét, míg a b állapot a negatív felét jelenti, a veremben szereplő v-k száma a bolha pozíciójának abszolút értékét jelenti. Ha a bolha az origóban van, a verem üres (csak v_0 van benne), és az automata bármely állapotban lehet. Ezt az automatát mostantól fogva *véletlen bolyongónak* nevezzük.

Világos, hogy ha a bolha az origóban van, csak páros számú lépés után érhet újból az origóba. A páratlanadik kimenő jelek szükségszerűen 0. Ezért úgy módosítjuk az automatánkat – további két, a-val és b-vel ekvivalens, de λ-t kiíró állapot felvételével – hogy csak minden második lépés után írjon ki a kimenő szalagra.

Annak a valószínűsége, hogy a *módosított véletlen bolyongó* által kiírt szalagon egy 1-es után l karakterrel újból 1-es következik,

$$f(l) = \binom{2l}{l} p^l q^l, \tag{4.3}$$

mivel $2l$ lépés – azaz l kiírt karakter – után pontosan akkor kerül újból a bolyongó az origóba, ha l-szer lépett jobbra, melynek p a valószínűsége, és l-szer lépett balra, amely esemény q valószínűséggel következik be. A binomiális eloszlás képletét alkalmaztam.

A közismert Stirling-formula szerint[13] :

$$n! = \left(\frac{n}{e}\right)^n \sqrt{2\pi n} \cdot e^{\frac{\Theta_n}{12n}},$$

ahol $0 < \Theta_n < 1$, vagyis az utolsó tényező, főleg nagy n-ek esetén, jó közelítéssel elhagyható. E képletet beírva a binomiális együttható képletébe:

$$\binom{2l}{l} = \frac{2l!}{(l!)^2} \approx \frac{\left(\frac{2l}{e}\right)^{2l}\sqrt{4\pi l}}{\left[\left(\frac{l}{e}\right)^l \sqrt{2\pi l}\right]^2} = 4^l \frac{1}{\sqrt{\pi}} l^{-1/2}.$$

Innen, jó közelítéssel:

$$f(l) = \frac{1}{\sqrt{\pi}} l^{-1/2} (4pq)^l. \tag{4.4}$$

E képletből jól látható, hogy $p = q = 0,5$ esetén $f(l)$ hatványfüggvényt követ, míg egyéb esetekben gyorsan, gyakorlatilag exponenciálisan lecseng. Ez nem meglepő, hisszen $p \neq q$ esetén a bolhánk hosszú távon eltávolodik az origótól, és nem tér vissza. Viszont a $p = q$ esetén akármilyen távolra távolodik is el az origótól, a mozgás helyzetének várható értéke az origó marad, vagyis statisztikailag az várható, hogy vissza fog térni oda. Az origótól való eltávolodás várható értéke k lépés után (vagyis a mozgás szórása) a fizikában közismert: $\sqrt{k}$ (mint például a Brown-mozgás esetén), vagyis itt is megjelenik ez az $-1/2$ exponens.

Mostantól kezdve csak a $p = q = 0,5$ esettel foglalkozunk.

Informálisan, e legutóbbi tag adja az autokorrelációs függvényt, mert az autokorrelációs függvény (2.3) képletében azok az $X_i \cdot X_{i+l}$ típusú tagok nem zérusok, ahol mind X_i, mind X_{i+l} értéke 1. Precízebben:

$$C(l) = \langle X_i \cdot X_{i+k} \rangle - \langle X_i \rangle \langle X_{i+k} \rangle = \langle X_i \rangle f(l) - \langle X_i \rangle^2. \tag{4.5}$$

Az alábbiakban bemutatjuk, hogy $\lim \langle X_i \rangle = 0$. végtelen hosszú sorozat esetén ezért $C(l) = 0$; de véges sorozatra mondhatjuk, hogy hatványfüggvény szerint cseng le, hiszen a második tag hosszú sorozat esetén elhanyagolható az első taghoz képest. Ezért állítjuk, hogy a véletlen bolyongó hosszú távú korrelációt produkál, $\gamma = 0,5$ exponenssel:

[13] Megtalálható például Rényi Alfréd: *Valószínűségszámítás*, Tankönyvkiadó, Budapest, 1973. 141. oldalán

$$C(l) = C \cdot l^{-0,5}. \tag{4.6}$$

Számítsuk ki most $\langle X_i \rangle$-t. A bolyongót az origóból indítjuk, vagyis a sorozat első tagja mindig 1 lesz. Vagyis ekkor, egy N hosszúságú sorozatban

$$E(N) := \langle X_i \rangle = \frac{1}{N} \sum_{i=1}^{N} f(i) = \frac{1}{\sqrt{\pi}} \frac{1}{N} \sum_{i=1}^{N} \frac{1}{\sqrt{i}}. \tag{4.7}$$

Könnyen belátható, hogy[14]

$$\lim_{N \to \infty} E(N) = 0. \tag{4.8}$$

Ennek a szemléletes jelentése az, hogy ha elég sokáig generáljuk a sort, akkor nagyon ritkán lesznek az 1-esek a 0-sok között, tetszőlegesen nagy „üregek" várhatók két 1-es között.

Ezek a tetszőlegesen nagy üregek azt sugalják, hogy a keletkezett sorozatot érdemes fraktálként felfogni. A *fraktálok* az elmúlt két évtizedben hihetetlen népszerűségre tettek szert sok tudományterületen (és még a művészetekben is), és tipikusan ún. „skálázási tulajdonságokkal", pl. tört dimenzióval és hatványfüggvény alakú autokorrelációs függvénnyel jellemezhetők. Mit jelent a tört dimenzió? Szemléletesen, ha egy L oldalú „dobozt" helyezünk a fraktálra, mint geometriai alakzatra, és V-vel jelöljük a fraktál dobozba eső részének „térfogatát" (hosszát, területét, súlyát, ahogy tetszik), akkor egy

$$V \sim L^D$$

összefüggés érvényesül. Ezt a D exponenst nevezzük fraktáldimenziónak. Egy klasszikus kétdimenziós alakzat esetén a felület a hossz négyzetével arányos, míg a háromdimenziós térfogat a hossz köbével. Fraktálok esetén, ez a D fraktáldimenzió tört is lehet, vagyis a fraktálok átmeneti alakzatok például az egy és a két dimenziós geometria között.

Esetünkben, a generált szekvenciát felfoghatjuk úgy, mint a természetes számok egydimenziós diszkrét számegyenesén elszórt, betöltött pontok halmazát: az i-ik pont pontosan akkor van betöltve, ha a szekvencia i-ik eleme 1-es. Ha a fraktáldimenziót akarjuk

[14] Mivel a számtani és négyzetes közepek közötti összefüggés alapján

$$0 < \frac{\sum_{i=1}^{N} \frac{1}{\sqrt{i}}}{N} < \sqrt{\frac{\sum_{i=1}^{N} \frac{1}{i}}{N}} < \sqrt{\frac{1 + \log_2 N}{N}}$$

és $\lim_{N \to \infty} \frac{\log N}{N} = 0$.

meghatározni, egy N hosszúságú dobozban nagyságrendileg $V(N) := N \cdot E(N)$ darab 1-es található. A (4.7) egyenlet alapján

$$V(2N) = \frac{1}{\sqrt{\pi}} \sum_{i=1}^{2N} \frac{1}{\sqrt{i}} = \frac{1}{\sqrt{\pi}} \left(\sum_{i=1}^{N} \frac{1}{\sqrt{2i-1}} + \sum_{i=1}^{N} \frac{1}{\sqrt{2i}} \right) \approx \frac{2}{\sqrt{2\pi}} \sum_{i=1}^{N} \frac{1}{\sqrt{i}}.$$

Vagyis közelítőleg

$$V(2N) = \sqrt{2} V(N). \tag{4.9}$$

Az ilyen alakú összefüggéseket nevezik a fizikában skálázási tulajdonságnak, és e függvényegyenletek egyedüli ,,ésszerű" megoldása:

$$V(N) = \Gamma N^{1/2} \tag{4.10}$$

alakú. A közelítések miatt nem állítom, hogy $V(N)$ egzaktul (4.10), de megbecsülhető a viselkedése ilyen módon. Vagyis a véletlen bolyongó által generált szimbólumszekvencia felfogható egy $0,5$ dimenziójú fraktálként is.

(A (4.10) összefüggésből is következik (4.8), de ez a levezetés nem volt eléggé egzakt.)

Zárásként, a 3.1 fejezetben bevezetett $F(l)$ függvény α exponensét a (3.6) skálatörvény alapján számíthatjuk ki. Mivel (4.6) szerint az autokorrelációs függvény exponense $\gamma = 0,5$, így a skálatörvény $\alpha = 0,75$ értéket jósol.

4.2.2 Empírikus eredmények a véletlen bolyongóval

Most vizsgáljuk meg, hogy a véletlen bolyongó tényleg mutatja-e ezeket a kiszámított tulajdonságokat. Készítettem egy számítógépes programot, amely megvalósít egy sztochasztikus veremautomatát, és a véletlen bolyongót adva be neki, generáltam egy 0-kból és 1-ekből álló szekvenciát.

Ha a p és q paraméterek $0,5$-től különböztek, a sorozat elején néhányszor megjelent egy egyes, majd a bolyongó végleg eltávolodott az origótól, és a szekvencia már csak zérusokat tartalmazott. Íme egy példa egy ilyen sorozatra, $p = 0,55$ és $q = 0,45$ paraméterek mellett, amely aránylag ,,sokáig tartja magát az origó körül", hiszen a 106-ik pozícióig találunk 1-est:

```
11000000000000000000000000000000000000000000000000000000000010000000001010010
10011001111111010000111100000001000000000000000000000000000000000000000000000
00000000000000000000000000000000000000000000000000000000000000000000000000000
```

Ezzel szemben, $p = q = 0,5$ esetén tetszőlegesen messze találunk 1-eseket. A B. Függelékben, példaként hozok egy 1800 karakter hosszúságú sorozatot, amely esetén jól megfigyelhető, hogy akármilyen nagy „üregek" lehetségesek két 1-es között. A véletlen bolyongó „nem távolodik el túlságosan" az origótól, ill. visszatér hozzá, ezzel alakítva ki a sorozat említett fraktálszerkezetét.

Ez a sorozat eredetileg 32000 karakter hosszú volt, és egy második program készítette el az eredeti szekvencia (3.2) szerint definiált $F(l)$ függvényét. A 4.2 ábrán az $F(l)$ függvény látható, kétszer logaritmikus skálán.

4.1. ábra: A véletlen bolyongó által generált sorozat $F(l)$ függvénye, kétszer logaritmikus skálán. Kétszer logaritmikus skálán a hatványfüggvények egyenesként jelennek meg, amelyek meredeksége az exponenst adják meg. A folyamatos egyenes a $p = q = 0,5$ paraméterek mellett kapott sorozatra illesztett, $0,748$ meredekségű hatványfüggvény.

Kétszer logaritmikus skálán a hatványfüggvény egyenesként jelenik meg, a meredeksége pedig az exponenssel egyenlő. Így az ábrán látható illesztett hatványfüggvény meredeksége $\alpha = 0,748$. Ez az eredmény jól egyezik az elméleti számításból – a (4.6) autokorrelációs függvényből és a (3.6) skálatörvényből – kapott $\alpha = 0,75$ jóslattal.

Az illesztést az $1 \leq l \leq 4000$ intervallumban végeztem, mivel a tartomány felső felénél további effektusok lépnek fel a 32000 karakter hosszúságú szekvencia véges hossza miatt. Az illesztési tartományt tovább csökkentve, nő az α értéke, egészen $0,77$-ig.

4.3. Az X-SPDA és a szövegek hosszú távú korrelációi

A véletlen bolyongó bebizonyította, hogy a sztochasztikus veremautomata potenciálisan alkalmas arra, hogy az írott szövegek körében megfigyelt statisztikai tulajdonságok egy részét visszaadjuk. A veremautomata rokonságban áll a környezetfüggetlen grammatikákkal, ami a nyelvészeti interpretálhatóság szempontjából fontos, ugyan akkor matematikailag is számítható, azaz az empíriát összevethetjük egy elméleti jóslattal.

Jelen fejezetben úgy általánosítjuk a veremautomata fogalmát, hogy a nyelvészeti elméletekhez közelebb álljon. Habár veszítünk a matematikai egzaktságból – a jövőben szeretném ezt is jobban kidolgozni –, de még közeleb jutunk a szövegek megfigyelt tulajdonságaihoz.

Az SPDA-k, pontosabban a SCFG-k fogalmát azzal bővítem ki, hogy az egyes nemterminálisok ,,örökölnek" egy hozzájuk csatolt, terminálisokból álló sztringet is. Azaz a nemterminálisok szerepét $V_N \times V_T^*$ veszi át, ahol V_N a nemterminálisok, és V_T a terminálisok halmaza a klasszikus környezetfüggetlen grammatikákban. Veremautomata esetén ez azt jelenti, hogy egy veremszimbólum helyett egy veremszimbólumot, és a kimenő ábécé egy sztringjét kell beolvasnia egyszerre a verem tetejéről. (Ez végtelen nagy veremábécét vagy egy második, speciálisan működő vermet jelent.)

Mit szeretnék ezzel leírni? Az az *általánosított veremautomata* – így nevezzük mostantól fogva ezt az automata-típust –, amit megvalósítottam a modellemben, az X-bar struktúrát modellezi. Egy sztochasztikus X-bar fa a következőképpen nézhetne ki:

$$P(\mathrm{XP} \rightarrow \mathrm{ZP}\ \overline{\mathrm{X}}) = p \tag{4.11a}$$

$$P(\mathrm{XP} \rightarrow \overline{\mathrm{X}}) = 1 - p \tag{4.11b}$$

$$P(\overline{\mathrm{X}} \rightarrow \mathrm{X}\ \mathrm{YP}) = q \tag{4.11c}$$

$$P(\overline{\mathrm{X}} \rightarrow \mathrm{X}) = 1 - q \tag{4.11d}$$

Azaz p annak a valószínűsége, hogy specifier ágazik el a maximális projekcióból, míg a kötelező bővítmény valószínűsége legyen q. A szabad bővítmények lehetőségétől az egyszerűség kedvéért eltekintünk. Az adjunct-ok szerepét további vizsgálódásokkal lehetne elemezni. Ezt az SCFG-t az A Függelékben leírt módon alakíthatjuk át SPDA-vá.

Az általánosított (,,öröklő") SCFG esetén az egyes kategóriacimkékhez egy σ terminális-szekvenciát is csatolunk. Az általam megvalósított esetben például

$$P(\mathrm{XP}[\sigma] \rightarrow \mathrm{XP}[0\sigma]\ \overline{\mathrm{X}}[\sigma]) = p \tag{4.12a}$$

$$P(\mathrm{XP}[\sigma] \to \overline{\mathrm{X}}[\sigma]) = 1 - p \tag{4.12b}$$

$$P(\overline{\mathrm{X}}[\sigma] \to \mathrm{X}[\sigma]\ \mathrm{XP}[1]) = q \tag{4.12c}$$

$$P(\overline{\mathrm{X}}[\sigma] \to \mathrm{X}[\sigma]) = 1 - q \tag{4.12d}$$

szerepelt. Az első sorban szereplő 0σ jelentése a 0 karakter és a σ karakterfűzér konkatenációja. A generálás kiindulópontjaként szolgáló mondatszimbólum XP[1] volt, és felvettem még egy

$$P(\mathrm{X}[\sigma] \to \sigma) = 1 \tag{4.12e}$$

kiíró szabályt. Az XP, YP és ZP közötti különbségeket tudatosan tüntettem el, mivel a különbséget ebben a megközelítésben csak a hordozott szekvencia tartalmazza.

Mit jelent a (4.12a-e) alatti szabályrendszer? Az X fej kiírja az XP által hordozott sztringet. A specifier, mint maximális projekció, megörökli a karakterláncot, és az elejéhez fűz még egy nullát; míg a bővítmény ,,tiszta lappal indul", egy egyelemű szekvenciával. A bővítmény tehát azonos a mondatszimbólummal, így a bővítményből induló részfa hasonló lesz az egészhez. Ezt a jelenséget nevezzük *önhasonlóságnak*, mely a fraktálok egy további híres tulajdonsága, és ez biztosítja a fraktál-szerű statisztikai jelenségeket. Éppen ezért, a létrejövő szekvenciát, amelyet a (4.12a-e) szabályrendszer generál, *X-vonás fraktálnak* nevezem.

A kétszintű X-vonás struktúrát az egyszerűség kedvéért egyszintűre redukálhatjuk az egyszerűbb íráskép kedvéért. A p és q valószínűségek, illetve az inverz valószínűségek szorzata adja az egyes lehetőségekhez tartozó P valószínűségeket. Ekkor:

$$P(XP[i] \to XP[i0]X[i]XP[1]) = p \cdot q \tag{4.13a}$$

$$P(XP[i] \to X[i]XP[1]) = (1-p) \cdot q \tag{4.13b}$$

$$P(XP[i] \to XP[i0]X[i]) = p \cdot (1-q) \tag{4.13c}$$

$$P(XP[i] \to X[i]) = (1-p) \cdot (1-q) \tag{4.13d}$$

Ezekkel a formájú szabályokkal generáltam szekvenciákat. Egy szekvenciát akkor fogadok el, ha üres veremmel leáll a működés, vagyis egy mondatszimbólumból kell a véges mondatot generálni. Ha túl kicsik a p és q paraméterek, nem ,,indul be" a generálás, néhány karakterből álló mondatot kapunk. Nagy paraméterek esetén ,,elszáll" a generálás, túlságosan nagy fa generálódik, és nem áll le a működés. Ezért közepes, $0,5$ körüli

értékekkel kellett próbálkoznom. Előfordult, hogy mindjárt leállt, és volt, hogy elszállt. Ezért addig kellett futtatnom a programot, amíg értékelhető méretű szekvenciát nem kaptam. Ha a generálás alatt álló sorozat hossza túllépte a 32000-t, újraindítottam automatikusan a futást.

Meglehetősen ritkán és kevés paraméterpár mellett kaptam értékelhető hosszúságú sorozatot. A 4.2 ábrán a $p = q = 0,5$ paraméterek mellett kapott X-vonás fraktál $F(l)$ függvénye látható, kétszer logaritmikus skálán. Ilyen tengelyek esetén a hatványfüggvény képe egyenes, és a meredeksége adja meg az exponenst.

4.2. ábra: A $p = q = 0,5$ valószínűségek mellett generált X-vonás fraktálból készített $F(l)$ függvény. Az illesztett hatványfüggvény $\alpha = 0,60$ meredekségű.

Az illesztett hatványfüggvény exponense $\alpha = 0,60$. A szekvencia hossza 26435 karakter volt, de az ábrán látható, hogy a néhány ezernél nagyobb l-ekre már jelentős hatással van a szekvencia véges hossza, vagyis az a tanulság, hogy az utolsó nagyságrendre nem lehet illeszteni. A következő ábrán (4.3) a $p = q = 0,5$ paraméterek mellett generált szekvenciát hasonlítjuk össze a hasonló hosszúságú, de $p = 0,4$ és $q = 0,6$ mellett generált sorozattal.

Minden esetben a kétszer logaritmikus skálán illesztett egyenesek meredeksége adja az illesztett hatványfüggvények kitevőjét. A $p = q = 0,5$ paraméterek mellett $\alpha = 0,59$ körüli kitevőket kaptam többször megismételt szimuláció esetén is. Hasonlóképpen, $p = 0,4$, $q = 0,6$ esetén, minden esetben $\alpha = 0,56$ meredekségű egyenes illeszkedett a legjobban a kapott

4.3. ábra: A $p = q = 0,5$ valószínűségű szekvenciához tartozó $F(l)$ függvény (kockák) $\alpha = 0,60$ meredekségű, míg a $p = 0,4$, $q = 0,6$ paraméterekhez (keresztek) az $\alpha = 0,56$ meredekségű egyenes tartozik. Az ábrát a $10 \leq l \leq 100$ tartományban nagyítottam ki, hogy a különbség jobban látható legyen.

adatokhoz. Látható, hogy az exponens csak a p és q paraméterektől függ, megismételhető, de különböző paraméterekhez különböző exponensek tartozhatnak. Vagyis – szemben a véletlen bolyongóval – ezzel a modellel ,,hangolhatjuk" az exponenseket. Idő, türelem (vagy gyors számítógép) hiányában, nem tudok további értékelhető eredményeket prezentálni lényegesen más paraméterek mellett.

A 4.4 ábrán azt az esetet vizsgáltam, amikor megtiltom a bővítmények beillesztését ($q = 0$ eset). Az így generált szöveg `1101001000100001` és így tovább alakú, vagyis az egyesek között egyesével nő a nullák száma. Az ábrázolt szekvencia hossza közel 26800.

A 4.4 ábrán érdekes viselkedés észlelhető: rövid távon ($1 \leq l \leq 300$) a korrelációk teljes hiánya, esetleg kis mértékű antikorreláció ($l < 200$ esetben $\alpha < 0,48$ egyértelműen) figyelhető meg. Ennek oka világos, a lokális önhasonlóságot előidéző bővítményeket kivettem a modellből. Lokálisan, teljesen véletlennek tűnik a sorozat. Ezzel szemben, a globális tulajdonságok a logaritmikus skála felső felében ($300 \leq l \leq 6000$), a nagyobb nagyságrendeknél tűnnek elő. Itt már határozott korreláció, $\alpha = 0,69$ figyelhető meg. A legnagyobb nagyságrendek esetében a megszokott, véges-szekvencia-effektusok kerülnek elő. Az ábra szerkezete a 3.2 fejezetben említett, emberi véletlenszám-generátorra emlékeztet (Schenkel et al. [1993]).

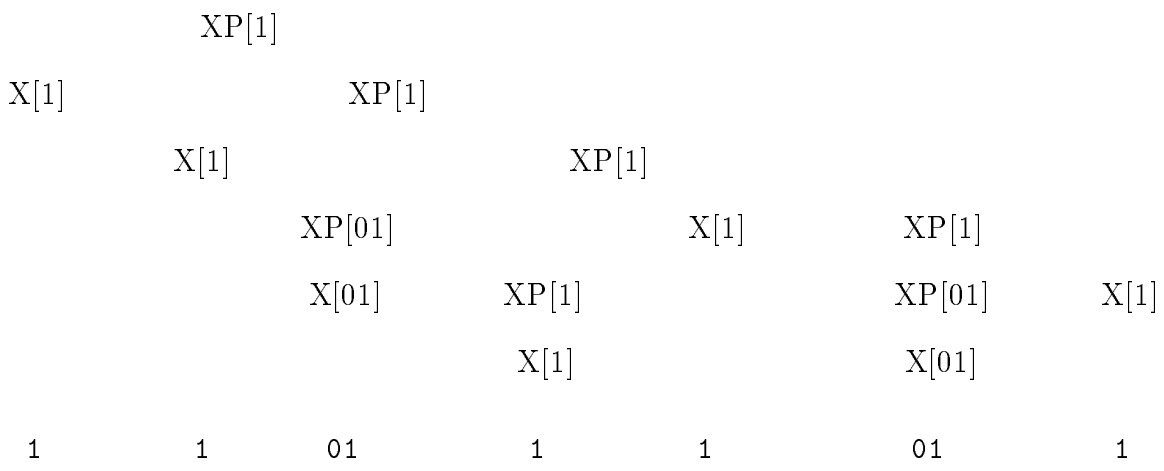
4.4. ábra: Ha $p = 0,99$ és $q = 0$, akkor a keletkezett szekvencia determinisztikus, és kettős szerkezetű (l) függvényt kapunk: az alsó nagyságrendekben $\alpha \leq 0,50$, míg a felső nagyságrendekben $\alpha = 0,69$.

További kutatások során fel kellene térképezni a teljes (p, q) tartományt. Esetleg a fizikai fázisdiagramokhoz[15] hasonló ábrát kapunk, ha a különböző (p, q) párok mellett, adott k számszor lefuttatjuk a programunkat: kis (p, q)-kra túlságosan rövid sorozatokat kapunk, nagy (p, q)-kra pedig állandóan elszáll a sorozat hossza. Közepes (p, q)-kra különböző exponenseket kapunk, ha az értékelhető sorozatok exponenseit átlagoljuk. Nagy p-re és kis q-ra a 4.3 ábrához hasonló, ,,törött" függvény várható. Nagy q-kra és kis p-kre a sorozat csupa egyesből fog állni, hiszen a fában csak bővítmények szerepelnek, amelyek mondatszimbólumból vezetendők le. Ezen ábrában nyelvészetileg az lesz érdekes, hogy melyik az a (p, q) tartomány, amelyik a nyelvek modellezésénél releváns, azaz az a fázis (vagy fázisok?), amelyek a szövegeket jellemzik, hogyan helyezkednek el a többi fázishoz képest.

Az X-bar fraktál által generált szekvenciát visszafele is lehet parszolni. Az írott szövegekre jellemző, hogy nem minden esetben egyértelmű a parszolás. Jelen esetben úgy tűnik nekem, hogy a visszafejtés is egyértelmű, de ezt nem tudom formálisan bebi-

[15] Két paraméter függvényében ábrázoljuk a rendszer lehetséges, minőségileg különböző állapotait, ún. *fázisait*, például a halmazállapotokat a hőmérséklet és a nyomás, mint paraméterek, függvényében.

zonyítani. Vegyük példaként az `110111011` szekvenciát, amelyet a számítógépes program $p = q = 0,5$ mellett, egy rövidre sikeredett generálás során írt ki. Ennek a szekvenciának a lehetséges parszolása látható a 4.5 ábrán.[16]



4.5. ábra: Az `110111011` *szekvencia parszolása a (4.13) szabályok szerint.*

Más struktúra nem lehetséges, ha módszeresen végiggondoljuk a lehetőségeket: alulról kell „építkezni", a két darab `01` részlet egy-egy XP[01] feje, amelyek pedig egy-egy XP[1] specifier-jeinek kell lenniük. Ezen kívül több specifier-pozíció nem lehetséges, mert különben további nulláknak is kellene szerepelniük a mondatban. Ezek után, a maradék struktúrát csak úgy lehet összekötni, ahogy az a 4.5 ábrán látható.

Összefoglalásként megállapíthatjuk, hogy a jelen fejezetben bevezetett modell több ponton is relevánsabb nyelvészeti szempontból, mint a (4.3) fejezetben definiált véletlen bolyongó. Egyrészt, megfeleltethető az X-vonás elméletnek. Habár a modell csak akkor releváns az írott szövegekre, mint globális egészre, ha az 1.1 fejezetben feltételezett hipotézisemet – amely szerint az X-vonás elmélet a mondat szintje fölött is érvényes valamilyen formában – tekintetbe vesszük. Másodszor, a véletlen bolyongó $\alpha = 0,75$ exponensével ellentétben, jelen modell az írott szövegek exponenseihez hasonló, és állítható-változtatható exponenseket produkál. Végezetül, a következő alfejezetben bemutatandó módon, a Zipf-függvény is reprodukálható vele.

[16] Egy szintre vetített struktúrákat használok, (4.13) szerint.

4.4. Az X-bar grafok Zipf-tulajdonságai

Utolsó modellünkben úgy módosítjuk a *módosított sztochasztikus veremautomatát*, hogy megengedjük állandóan új terminális bevezetését. Újraíró szabály alakjában például:

$$X[\sigma] \to Y[n\sigma] \tag{4.14}$$

jelentse azt, hogy az újraírt kategóriához kapcsolódó sztring bővül egy n elemmel, amely egy ,,új" eleme a ,,nagy" lexikonnak (terminálisok halmazának). A ,,nagy" azt jelenti, hogy gyakorlatilag – nem matematikai értelemben véve – végtelen, míg az ,,új" arra utal, hogy az adott terminálist még nem használtuk fel az automata működése során.

Ez az eszköz arra használható, hogy a Zipf-függvényt próbáljuk meg modellezni. A felhasznált terminálisok statisztikáját vizsgálva, a különböző típusú Zipf-függvényeket kaphatjuk vissza.

Czirók *et al.* [1995] számítógépes modellezéssel bemutatja, hogy a hosszú távú korrelációkat tartalmazó szekvenciák hatványfüggvénnyel közelíthető Zipf-függvényei hogyan állnak szemben a Markov-láncok lépcső alakú Zipf-függvényeivel. A különbség észlelhető, még ha a Markov-láncokkal az eredeti szzekvenciákat akarjuk is közelíteni. Ugyanakkor bizonyos esetekben exponenciális Zipf-függvényt kaphatunk, például csupán kódoló szakaszokat tartalmazó DNS-szekvenciák esetében, vagy írott szöveg esetén az ábécé betűinek az eloszlása.

Modelljeink a (4.12a-e) egyenletek által megadott veremautomata (SCFG) analógjai, csupán a nemterminálisokhoz asszociált sztringekben különböznek. Az egyszerűbb írásmód kedvéért, (4.13)-hoz hasonlóan egy szintre egyszerűsítjük a modellt, sőt csak a (4.13a) sort írjuk le, mivel abból a többi értelemszerű: a fejet megelőző specifier valószínűsége mindig p, míg a fejet követő bővítményé q. A mondatszimbólum mindig $XP[1]$, ahol 1 a lexikon első eleme.

Az első modellünk alapképlete ezen leegyszerűsítéseket követően:

$$XP[s] \to XP[s]X[s]XP[ns], \tag{4.15}$$

ahol s egy, terminálisokból (lexikonbeli elemekből) alkotott sztring, míg n az ,,újonnan bevezetett" terminális. Ez a modell lépcsőfüggvényszerű Zipf-fügvényt produkál ($p = q = 0,5$ mellett kapott 4.6 ábra).

Látható, hogy se nem hatványfüggvénnyel, se nem exponenciálissal nem közelíthető a Zipf-függvény. Ugyanis a kétszer logaritmikus skálán a hatványüggvény képe lineáris, míg az exponenciális függvény a lin-log skálán jelenne meg egyenesként, de egyik ábrán sem közelíthező jól a függvénykép egyenessel.

4.6. ábra: A (4.15) egyenlet által megadott modell lépcsőszerű Zipf-függvénye ($p = q = 0,5$). Látható, hogy se nem hatványfüggvény (egyenes a log-log skálán), se nem exponenciális (egyenes a lin-log skálán) nem illeszthető a lépcsős függvényre.

Hasonló ábrákat kaptunk az alábbi modellel ($p = q = 0,5$, 4.7 ábra):

$$XP[s] \to XP[s]X[s]XP[n]. \tag{4.16}$$

Ez a modell hasonlít a (4.13) modellhez abban, hogy a bővítmény egy új mondatszimbólumként egy olyan részfa kiindulásának lesz az alapja, amely hasonló az egészhez. Az egyedüli különbség az, hogy az új mondatszimbólumhoz más, de egy szóból álló sztring van asszociálva. (Egyébként a hatványfüggvény illesztése nem tragikusan rossz, csak az $R = 1$-hez nem illeszkedik, tehát lehet, hogy felhasználható (4.16) az írott szövegek modellezésére?)

E két modellel szemben, a következőek exponenciálist produkálnak (4.8 ábra):

$$XP[is] \to XP[is]X[is]XP[ni] \tag{4.17}$$

$$XP[is] \to XP[i]X[is]XP[nis], \tag{4.18}$$

ahol i a baloldali nemterminális sztring első eleme, míg s a maradék alsztring. A (4.17) modell eredménye a 4.6 ábrán látható, $p = q = 0,5$ mellett. E két modell (4.16), ill. (4.15)

4.7. ábra: A (4.16) modell lépcsőszerű Zipf-függvénye ($p = q = 0,5$) log-log és lin-log skálán.

megfelelői, azzal a különbséggel, hogy az X-vonás struktúra tetején lévő maximális projekció (a képlet bal oldalán lévő XP) sztringjének első eleme különleges módon öröklődik: a specifier vagy a bővítmény lényegében csak ezt kapja meg. Ez az az elem, amely a ,,legjellemzőbb" az XP-re, mivel ha a projekció bővítményi szerepet tölt be abban a nagyobb struktúrában, amelybe be van ágyazva, akkor ez az elem a ,,nóvum" az adott XP sztringjében.

Hatványfüggvényt akkor kaptam, amikor a sztring *elejét* örökítettem át a specifierre, és ezzel megnöveltem a Zipf-függvény értékét kis R-ekre:

$$XP[is] \to XP[i]X[is]XP[isn]. \tag{4.19}$$

Ezt a modellt módosíthatjuk, ha több terminálist is örökítünk a specifierbe, például:

$$XP[i_1 i_2 s] \to XP[i_1 i_2]X[i_1 i_2 s]XP[i_1 i_2 sn]. \tag{4.20}$$

A (4.19)-re $-\xi = -0,73$ exponensű hatványfüggvényt, míg (4.20)-ra $-\xi = -0,88$ exponensű hatványfüggvényt illesztettem ($p = q = 0,5$, 4.9 ábra). A természetes nyelven írt szövegek hasonló, $\xi \approx 0,85$ értéket adnak, míg a 2.4 fejezetben leírt szimbólum n-es Zipf-analízis csupán $0,57$-t eredményezett.

A 4.9 ábrán látható a (4.19) és a (4.20) modell Zipf-függvénye kétszer logaritmikus

4.8. ábra: A (4.17) modell Zipf-függvénye az illesztett exponenciálisokkal. A log-log skálán az ívelt vonal az exponenciális illesztés. Az exponenciális illesztés egyedül a kis R-ekre pontatlan. ($p = q = 0,5$)

4.9. ábra: A (4.19), ill. a (4.20) modellek Zipf-függvénye. A hatványfüggvény láthatóan jobban írja le a viselkedését, mint az ívelt exponenciális. Mindkét esetben $p = q = 0,5$ volt.

skálán. A (4.19) ábráján az $R = 1$ elem emelkedik ki, az, amelyik öröklődik a specifier-re is. A (4.20) modell ábráján mindkét öröklődő elem kiemelkedik. Mivel a specifierre is öröklődő elem(ek) megdobják a Zipf-függvény elejét, azt állíthatjuk, hogy a hatványfüggvány közelíti legjobban $\omega(R)$-t. (Az illesztést alig befolyásolja a függvény kiugró eleje, mert az illesztési pontok nagy többsége a függvény végén van.)

Az összes többi terminális előfordulási gyakorisága attól függ, hány bővítményi szinten keresztül öröklődik a levezetési fában. Vagyis annak a valószínűsége, hogy k-szor fog előfordulni, éppen $q^k(1-q)$. Ez, mint a (4.9) ábrán látható, szintén hatványfüggvénnyel közelíthető inkább, mintsem exponenciálissal.

4.10. ábra: A (4.19), ill. a (4.20) modellek Zipf-függvénye, a specifier-re is öröklődő elemek levágása után, lin-log skálán. Az egyenes exponenciálisnál most is jobb illesztés az ívelt hatványfüggvény.

Milyen szemléletes jelentést adhatunk a (4.19) modellnek (ill. (4.20)-nak), ha azt állítjuk, hogy ez közelíti meg a természetes nyelvek Zipf-függvényét? Fogjuk fel úgy a cimkékhez asszociált sztringet, mint kulcsszavak listáját, amely az adott projekció ,,értelmét" írja le. A fej mindig kiírja a maximális projekció ,,értelmét", míg a bővítmény egy olyan maximális projekció, amely egy újabb ,,kulcsszót" ad hozzá az előző kulcsszavaihoz. Ezzel szemben, a specifierben az egész szöveg kulcsszava(i) jelen(nek) meg. Igazság szerint (4.18) állna legközelebb a nyelvészetben megszokott X-vonás elmélethez, amennyiben ott éppen az adott maximális projekcióra jellemző kulcsszó (azaz az, amivel bővítmény

esetén éppen most bővült a sztring) jelenik meg a maximális projekció specifier-jében.

A (4.19) modell rokona igazából a (4.13) általánosított SCFG-nek, és ezért nem meglepő, hogy e két modell egyszerre adja vissza a természetes nyelvi szövegek statisztikai tulajdonságait. Hogyan értelmezzük ugyanis (4.13)-t a (4.19) fényében? Tegyük fel, hogy a szöveget (4.19) szerint generálom, és kíváncsi vagyok, hol, hogyan fordul elő a mondatszimbólum sztringjének egyetlen eleme. Ekkor ezen elemet 1-essel helyettesítem, és minden más helyére 0-t írok. Ebben az esetben (4.19) éppen (4.13)-ba megy át (a bővítmény és a specifier szerepet cserél, de jelen esetben ez érdektelen). Vagyis a hosszú távú korrelációkat az egész szövegre globálisan jellemző, a mondatszimbólum sztringjében már előforduló, és a specifier-ekre is átöröklődő szó eloszlása adja. Minden más szó legfeljebb lokális korrelációkat okoz.

Végezetül hadd vessem fel, hogy meg kell vizsgálni azt az esetet, hogy mi történik, ha ,,kicsi", ,,véges" a terminálisok halmaza. Vagyis ha egy pont után, amikor ,,új" terminálist akarok bevezetni az n-et tartalmazó szabályommal, akkor olyan terminálishoz folyamodok, ami valahol már korábban, véletlenszerűen, a jelen előfordulástól függetlenül, előfordult. Ez egy tisztán véletlen faktort fog behozni a modellbe, és valószínűleg ez felelős például az ábécé betűinek eloszlásánál megfigyelhető exponenciális Zipf-függvényért.

5. fejezet

Összefoglalás

Dolgozatomban az írott szövegek statisztikai tulajdonságaiból néhányat figyeltem meg, majd igyekeztem modellezni olyan módon, hogy az nyelvészetileg is minél relevánsabb legyen. Nem tudok arról, hogy ezzel ilyen módon mások is foglalkoztak volna ezt megelőzően.

Az írott szövegeket, mint szimbólumsorozatokat vizsgáltam, és ebből a szempont akár más szimbólumsorozatokkal, például DNS-szekvenciákkal (kódoló, ill. nem kódoló szakaszokkal), vagy számítógépes programokkal is összevethetők. A második fejezet elején bevezettem a $C(k)$ autokorrelációs függvény fogalmát, a (2.3) képlet szerint, amely a szimbólumsorozatok jellegére jellemző. Eszerint három típusú sorozatot különböztethetünk meg: korrelációmenteset, csupán rövid távú korrelációval jellemezhetőt (exponenciális $C(k)$) és hosszú távú korrelációval jellemezhetőt ($C(k)$ hatványfüggvénnyel közelíthető).

A második fejezet második részében a szöveg rövid távú korrelációit használtuk fel a szövegek nyelv és téma szerinti szétválasztására. A Damashek-féle *vektortértechnika* a fonotaktika, a lexikon és a helyesírási konvenciók által meghatározott korrelációkra épít. Korábbi munkáimban (Bíró [1998a] és Bíró *et al.* [1998]) bemutatom a DNS-szekvenciákon, hogy a szekvenciákhoz rendelt vektorok hogyan hasonlítanak a korrelációk figyelmen kívül hagyásával készített vektorokhoz. Ez a módszer a rövid távú korrelációk erősségét méri.

A rövid távú korrelációk modellezésére szolgálnak a Markov-folyamatok. De Czirók András munkái bebizonyították, hogy nem csupán Chomsky érvei a *Syntactic Structures*-ben, vagy a harmadik fejezetben bemutatott eredmények – hosszú távú korrelációk léte – miatt inadekvát a Markov-folyamatok használata írott szövegek esetén, hanem a Zipf-függvény sem adható vissza ilyen módon.

A harmadik fejezetben bevezettem az $F(l)$ függvényt, amely α exponensének

vizsgálata egyenértékű az autokorrelációs függvény vizsgálatával. Ezen eszköz, valamint további eszközök segítségével, amelyekre csupán utaltam a 3.3 fejezetben, kétségtelenné vált, hogy a szimbólumszekvenciák egy széles osztálya – amelyhez az írott szövegek is tartoznak – nem elhanyagolható hosszú távú korrelációkkal, és egyéb különlegességekkel jellemezhető. Emiatt elkerülhetetlenné vált, hogy a negyedik fejezetben átlépjünk a környezetfüggetlen sztochasztikus eszközök területére.

A sztochasztikus környezetüggetlen grammatikák (SCFG-k) ismertetését követően felhasználtam a grammatika paramétereit a Zipf-függvény előállításához szükséges adatok megbecsülésére, majd bevezettem a sztochasztikus veremautomata (SPDA) fogalmát. Bemutattam, hogy a sztochasztikus veremautomatával létrehozható olyan szimbólumszekvencia, amely rendelkezik hosszú távú korrelációval, és az elméleti számítások jó összhangban vannak a számítógépes modellezéssel kapott eredményekkel. Végezetül továbbfejlesztettem ezen sztochasztikus környezetfüggetlen eszközöket (veszélyeztetve ezáltal a környezetfüggetlenségüket is), és íly módon nyelvészetileg relevánsabb modelleket kaptunk, amelyek az írott szövegekéhez hasonló exponensű $F(l)$-függvényt (azaz autokorrelációs függvényt) és Zipf-függvényt eredményeztek.

A nyelvészeti relevancia ezen a ponton az X-vonás elmélethez hasonló struktúrákat jelent. Az egyszerűség kedvéért, elhagytam a szabad bővítményeket, ezek szerepének a vizsgálata további kutatások tárgyának kellene lennie. Hasonlóképpen, a kötelező bővítmények és a specifier-pozíció mibenlétét sem tisztáztam elméletileg, így – az adott pozíciók létén túl –, meglehetősen *ad hoc* módon szerepelnek ezek a fogalmak.

De a legproblémásabb pont a nyelvészeti relevancia szemszögéből, az az, hogy *globális* X-bar struktúrát feltételeztem az egész szövegre, azaz egységes szerkezetet a szöveg legnagyobb szintjeitől kezdve, a mondat legapróbb összetevőjéig. (Ez utóbbi jelenthet szavakat, morfémákat, az X-vonás struktúrával rendelkező szótagszerkezetbeli fonémákat, vagy akár a fonológiai jegygeometria elemeit is. A vizsgálat módszerétől függ.) Ezt a hipotézist, amely nélkül az egész modellezés elveszíti kapcsolatát a nyelvleírással, az 1.1 fejezetben vetettem fel, és szándékomban van a jövőben sokkal alaposabban kidolgozni. Ezen munka eredményétől az X-vonás struktúra pozíciói szerepének jobb megértését is várom, amely feladat, mint írtam, elkerülhetetlen a 4.3 és 4.4 fejezetben leírt modellek (az *X-bar fraktálok*) magyarázó értékének a növeléséhez.

A. Függelék

Az SPDA-k és a SCFG-k ekvivalenciája

E függelék célja matematikailag precízebben kidolgozni a sztochasztikus veremautomaták (SPDA-k) elméletét. Először is, ismételjük meg a 4.2. definíciót:

A.1. Definíció: Sztochasztikus veremautomatának (Stochastic Push-Down Automaton, SPDA) nevezzük a $(\Sigma, A, V, a_0, v_0, P)$ *hatost, ahol:*

Σ *véges halmaz a kimenő ábécé,*

A *az állapotok véges halmaza,*

V *a véges veremábécé,*

$a_0 \in A$ *a kezdő állapot,*

$v_0 \in V$ *a veremfenékjel,*

$P : (A \times V) \times ((\Sigma \cup \{\lambda\}) \times A \times V^*) \to [0,1]$ *az állapotátmenet-függvény, amelyre igaz az, hogy*

$$\forall a \in A \forall v \in V : \sum_{s \in S, b \in A, w \in V^*} P(a, v; s, b, w) = 1.$$

A veremautomata működését *konfigurációk sorozataként* írhatjuk le. A konfiguráció tartalmazza a már generált szöveget, az éppen aktuális állapotot, valamint a vermet, mint egy sztringet a veremábécé felett.

A.2. Definíció: A konfigurációk C *halmaza rendezett hármasokat tartalmaz:*

$$\mathcal{C} := \Sigma^* \times A \times (v_0 V^*),$$

ahol $v_0 V^*$ *a* V^+*-nak a* v_0*-val kezdődő elemeit jelenti.*

Az utolsó kitétel jelentése az, hogy a verem alján mindig a veremfenék jelnek kell szerepelnie. A következő fogalom, amit definiálunk az, hogy mikor lehetséges két konfiguráció között az átmenet (egy lépésben):

A.3. Definíció: Az átmenet lehetséges a $c_1 = (\sigma_1, a_1, w_1) \in \mathcal{C}$ *konfigurációból a* $c_2 = (\sigma_2, a_2, w_2) \in \mathcal{C}$ *konfigurációba,* **ha** *teljesül:*

1. $\exists s \in \Sigma$, *hogy* $\sigma_2 = \sigma_1 s$, *és*

2. $\exists x, y \in V^*$ *és* $\exists v \in V$, *hogy* $w_1 = xv$ *és* $w_2 = xy$ *(értsd: ezek konkatenációja).*

Ekkor az átmenet valószínűsége $\P(c_1 \to c_2) := P(a_1, v; s, a_2, y)$. [17]

Az automata olyan konfigurációk sorozatán halad keresztül, amelyek szomszédos elemei között lehetséges az átmenet:

A.4. Definíció: A $\gamma = (c_1, c_2, ..., c_n) \in \mathcal{C}^*$ *konfigurációszekvencia egy levezetési láncot alkot, ha lehetséges az átmenet* c_i*-ből* c_{i+1}*-be,* $\forall i = 1, ..., n-1$*-re.*

Ha $c_n = (\sigma, a, w)$, *a* γ *levezetési lánc által generált szekvencia (szöveg, szó, mondat):* $\mathrm{Gen}(\gamma) := \sigma$. *A* γ *levezetési lánc valószínűsége pedig:*

$$\P(\gamma) := \prod_{i=1}^{n-1} \P(c_i \to c_{i+1}).$$

A levezetési lánc teljes, ha $c_1 = (\lambda, a_0, v_0)$, *és* $c_n = (\sigma, a_F, v_0)$ *alakú.*

Vagyis teljes levezetési lánc esetén a kezdő állapotból és üres veremmel indulunk, és üres veremmel fejezzük be a működést. Az íly módon generált szekvenciákat fogadjuk el ,,helyes" szövegnek (mondatnak, szónak). Ez utóbbiak valószínűsége:

A.5. Definíció: Egy $\sigma \in \Sigma^*$ *szöveg valószínűsége:*

$$\P[\sigma] = \sum_{\substack{\gamma \in \mathcal{C}^*,\ gen(\gamma)=\sigma \\ \gamma \text{ teljes levezetesi lanc}}} \P(\gamma).$$

A következőkben megmutatom, hogy a SCFG-k és a SPDA-k ekvivalensek. Mint azt a 4.1.2 alfejezet végén informálisan kifejtettem, egy levezetési fa valószínűsége valamely SCFG esetén az alkalmazott újraíró szabályok valószínűségeinek szorzata. A bizonyítás

[17] A következő néhány definicióban bevezetek különböző valószínűségeket. Engedtessék meg nekem, hogy a ¶ jelet használjam valamennyire, ahelyett, hogy állandóan szabatosan definiáljam az újabb és újabb valószínűség-függvényeket, és újabb és újabb jelöléseket vezessek be rájuk. Az eljárás teljesen automatikus és triviális.

két lépésből áll. Először bebizonyítom, hogy adott G SCFG-hez adható olyan A SPDA, hogy a két modellbeli levezetési láncok között kölcsönösen egyértelmű megfeleltetés tehető: azonos szekvenciát generálnak, azonos valószínűséggel.

A.1. Állítás: Bármely G sztochasztikus környezetüggetlen grammatikához létezik olyan A sztochasztikus veremautomata, hogy $\forall\sigma \in \Sigma^*$ két modell szerinti valószínűsége egyenlő, azaz $\P_G[\sigma] = \P_A[\sigma]$.

Bizonyítás: Legyen az adott SCFG a 4.1. definíció szerint (V_N, V_T, S, R, P). Ekkor konstruáljuk meg a következő SPDA-t: $(\Sigma, A, V, a_0, v_0, P^1)$, ahol $\Sigma = V_T$, $A = \{a_0, a\}$ és $V = V_T \cup V_N \cup \{v_0\}$. A P^1 átmenet-valószínűség függvényt a következőképpen adjuk meg:

$$P^1(a_0, v_0; \lambda, a, v_0 S) := 1,$$
$$\forall N \in V_N, M \in (V_N \cup V_T)^*\text{-re } P^1(a, N; \lambda, a, \overline{M}) := P(N \to M),^{18}$$
$$\forall T \in V_T\text{-re } P^1(a, T; T, a, \lambda) := 1.$$

Minden más helyen P^1 zérus értéket vegyen fel. Most azt akarjuk bebizonyítani, hogy ezen két sztochasztikus modell ugyan azt a valószínűséget rendeli bármely $\sigma \in \Sigma^*$ szöveghez.

1. lépés: Ezt úgy fogjuk bebizonyítani, hogy előbb megmutatjuk, az A SPDA bármely teljes levezetési láncához tartozik egy vele ekvivalens levezetési fa, azaz legbaloldalibb levezetési lánc, a G SCFG-ben, és fordítva. Az ekvivalencián azt értem, hogy azonos szöveget generál, azonos valószínűséggel. Ezen segédállítás két oldalát egyszerre mutatjuk be, mivel a bizonyítás párhuzamos. Méghozzá teljes indukció-szerűen bizonyítjuk be, a lánc mentén.

Az állítás az, hogy a levezetési láncok – a SCFG esetén a legbaloldalibb levezetési lánc – párhuzamosan haladnak. Azaz bármely lánchoz készíthető lánc a másik modellben, úgy hogy ha a megfelelő állapotok $< \xi >$ a G SCFG-ben, és $(\sigma, a, v_0 wv)$ az A SPDA-ban, akkor $\xi = \sigma v \overline{w}$ – ezt nevezzük mostantól fogva a két konfiguráció, vagy állapot, *ekvivalenciájának* –, és az eddig vezető levezetési részláncok valószínűsége egyenlő.

A ,,teljes indukció" indítása: az A SPDA a (λ, a_0, v_0) konfigurációval indul, amely 1 valószínűséggel a $(\lambda, a, v_0 S)$ állapotba megy át. Tehát eddig a lánc valószínűsége $\P_A = 1$. Ez ekvivalens a G SCFG levezetési láncainak első elemével, $< S >$-sel, amely szintén 1 valószínűségű láncot képvisel.

Tegyük fel, hogy a G egyik részláncához készítettünk már egy olyan részláncot A-ban, amely azonos p valószínűsségű. A részláncok végkonfigurációi $< \xi >$, illetve $(\sigma, a, v_0 wv)$, és teljesül $\xi = \sigma v \overline{w}$.

Ekkor két lehetőségünk van: ha v nemterminális, akkor ez a legbaloldalibb eleme a G-beli levezetési láncnak, vagyis ezt fogja átírni a következő lépés egy $v \to M$ újraíró

szabállyal, amelynek a valószínűsége $P(v \to M)$. A következő lépéssel bővített lánc valószínűsége így $pP(v \to M)$. Ennek megfelelően, az általunk konstruált A-beli lánchoz, amelynek az utolsó eleme $(\sigma, a, v_0 w v)$, mi a $(\sigma, a, v_0 w M)$ elemet fűzzük hozzá. Eme lépés valószínűsége $P^1(a, v; \lambda, a, \overline{M}) = P(v \to M)$, vagyis az A-beli bővített lánc valószínűsége szintén $pP(v \to M)$. A konfigurációk is ekvivalensek, hiszen a G-beli lánc új eleme a ξ-beli v újraírása után $\sigma M \overline{w}$, ami ekvivalens az A-beli lánc új konfigurációjával. Tehát a G-beli lánc következő eleméhez is generáltunk egy vele ekvivalens és azonos valószínűségű konfigurációt egy A-beli láncban.

A másik lehetőség az, ha v terminális. Ekkor a G-beli láncban nem haladunk előre, hanem csak az A-beli láncot folytatjuk egy 1 valószínűségű átmenettel: a $(\sigma, a, v_0 w v)$ konfigurációból a $(\sigma v, a, v_0 w)$ konfifgurációba megyünk át, mivel $P^1(a, v; v, a, \lambda) := 1$ terminális v-re. A $\xi = \sigma v \overline{w}$ ekvivalencia-feltétel továbbra is teljesül, csak v-t ,,csoportosítottuk át", és a valószínűségek egyenlősége is megmaradt, mindkét lánc p valószínűségű továbbra is. Ha w nem üres sztring, most w utolsó eleme veszi át v szerepét, és lehet vagy terminális, vagy nemterminális, és tovább léphetünk a teljes indukció következő fokára.

Ha pedig w üres sztring, a G-beli lánc végére értünk, a verem kiüresedése miatt leáll a generálás. De hasonlóan, ez azt jelenti, hogy az A-beli lánc is csupa terminálisból áll, vagyis a G-beli lánc is a végére ért. Összefoglalva: mindkét lánc ugyan azt a sztringet generálta, σv-t, és azonos p valószínűséggel.

A gondolatmenet megfordítható: ha adott egy A-beli részlánc, hozzá készíthető ugyan ezzel az eljárással egy ekvivalens G-beli lánc. Láttuk, hogy a két lánc egyenlő – 1 – valószínűségű, ekvivalens – S – konfigurációkkal indul. Ha pedig adott már egy G-beli részlánc, ami egyenlő valószínűségű és ekvivalens végállapotú az A-beli részlánccal, akkor a fenti módon – v terminális vagy nemterminális voltától függően – készíthető el a G-beli lánc következő eleme. (Konkrétan, ha v terminális, akkor a G-beli lánc következő eleme azonos a jelenlegivel.)

2. lépés: Eddig beláttuk, hogy az A és a G teljes levezetési láncai között kölcsönösen egyértelmű megfeleltetést, bijekciót létesíthetünk, úgy hogy a megfelelő láncok ugyan azt a $\sigma \in \Sigma^*$ szekvenciát generálják, azonos valószínűséggel. A SCFG-k legbaloldalibb levezetési láncai ekvivalensek a levezetési fákkal.

Ezek után könnyen belátható, hogy adott $\sigma \in \Sigma^*$ szöveg $\P_A[\sigma]$ valószínűsége az A SPDA-ban egyenlő a G-beli $\P_G[\sigma]$ valószínűséggel. Ugyanis $\P_A[\sigma]$ az A-beli, σ-t generáló teljes levezetési láncok valószínűségeinek az összege. De ezen láncok G-beli megfelelői lesznek azok, és csak azok a G-beli levezetési fák, amelyek σ-t generálják. Vagyis a G-beli, σ-hoz tartozó levezetési fák valószínűségeinek az összege, amely adja $\P_G[\sigma]$-t, egyenlő lesz az A-beli levezetések valószínűségeinek az összegével, $\P_A[\sigma]$-val. *Q.e.d.*

Az előző állítás megfordításaként, a következőkben be kellene látnunk, hogy minden sztochasztikus veremautomatához is található vele egyenlő valószínűséggel generáló sztochasztikus környezetfüggetlen grammatika. De ezt egyelőre nem sikerült belátnom, további munkát igényel.

B. Függelék

Példa véletlen bolyongóval generált szekvenciára

Az alábbi szekvenciát a 4.2 fejezetben leírt, módosított véletlen bolyongóval generáltam, $p = q = 0,5$ mellett. Jól megfigyelhető, hogy az origóból indított bolyongó – ezért kezdődik egyessel a szekvencia – tetszőlegesen sokáig lehet távol az origótól, majd visszatér. Ezért tetszőlegesen nagy „üregek" keletkezhetnek az egyesek között.

```
10000111110000000000000000110000000000001100011000000000000000000000000000
00000000000000000000000000000000000000000000000000000000000000000000000000
00000000000000000000011110100001110000000011011011110000101100100000000000
00000000000000000000000000000000000000000011110100000000000000000000000000
01001000000000000000000000000000000000000000000000000000000000000000000000
00000000000000000000000000000000000000000000000000000000000000000000000000
00000000000000000000000000000000000000000000000000000000000000000000000000
00000000000000000000011110010110000000000000000000000000000000000000000000
00000000000000000000011100000000000001110000000000000000000000000000000000
00000111100100000000000000000000000000000000000000000000000000000000000000
00000000000000000000000000000000000000000000000000000000000000000000000000
00000000000000000000000000000000000000000000000000000000000000000000000000
00000000000000000000000000000000000000000000000000000000000000000000000000
00001010000000000000000000000000000000000010101101000100010000000000000000
00000000000000000000000000000000000000000000000000000000000000000000000000
00000000000000000000000000000000000000000000000000000000000000000000000000
00000000000000000000000000000000000000000000000000000000000000000000000000
00000000000000000000000000000000000000000000000000000000000000000000000000
00000000000000010000001111001101000011100000000000000000000000000000000000
00110011000011100001101111001000000000000000000000000000000010000111011101
11110000000000000111100000101000000000000000000000000000000011011000000000
00000000000000000000000000000000000000000000000000000000000000000000000000
00000000000000000000000000000000000000000000000000000000000000000000000000

00000000000000000000000000000000000000000000000000000000000000000000000000
```

Irodalomjegyzék

S. Abney [1996]: Statistical Methods and Linguistics

M. Amit, Y. Shmerler, E. Eisenberg, M. Abraham, N. Shnerb [1994]: Language and Codification Dependence of Long-Range Correlations in Texts, Fractals, **2**, 1, pp. 7-13.

Berend Mihály [1980]: Genetikai ábécé, Gondolat Kiadé, Budapest.

Bíró T. [1997a]: Physics and Linguistics - What is common?, to be appear in the proceedings of the International Conference of Physics Students, Vienna, 1997.

Bíró T. [1997b]: szimbólumsorozatok elemzése statisztikai módszerekkel, TDK-dolgozat az ELTE TTK fizikus szakán.

Bíró T. [1998a]: DNS szekvenciák elemzése szövegelemzési módszerekkel, diplomamunka, ELTE TTK fizikus szak.

Bíró T. [1998b]: Some Statistical Games with Written Texts, Doximp 3, Graduate Students' Third Linguistics Symposium, June 5, 1998, Budapest.

Bíró T., Czirók A., Vicsek T., Major Á [1998]: Application of Vector Space Techniques to DNA, Fractals **6**, p. 205.

J-Ph. Bouchaud, M. Potters [1997]: Théorie des Risques Financiers. Portefeuilles, options et risques majeurs, Collection Aléa Saclay.

N. Chomsky [1957]: Syntactic Structres, The Hague: Mouton. Magyarul megjelent: Mondattani szerkezetek, Osiris-Századvég, Budapest, 1995.

A. Czirók, R. N. Mantegna, S. Havlin, H. E. Stanley [1995]: Correlations in binary sequences and a generalized Zipf analysis, Physical Review E, **52**, 1 , pp. 446-452.

A. Czirók, H. E. Stanley, T. Vicsek [1996]: Possible origin of power-law behavior in n-tuple Zipf analysis, Physical Review E **53**, 6371.

M. Damashek [1995]: Gauging Similarity with n-Grams: Language-Independent Categorization of Text, Science, **267**, pp. 843-848.

I. Derényi, T. Vicsek [1996]: The kinesin walk: A dynamic model with elastically coupled heads, Proc. Natl. Acad. Sci. USA, **93**, pp. 6775-6779.

G. Dietler, Y.-C. Zhang [1994]: Crossover from White Noise to Long Range Correlated Noise in DNA Sequences and Writings, Fractals, **2**, 4, pp. 473-479.

W. Ebeling, A. Neiman [1995]: Long-range correlations between letters and sentences in texts, Physica A 215, pp. 233-241.

F. Family, T. Vicsek [1991] (szerk.): Dynamics of Fractal Surfaces, World Scientific.

S. A. H. Geritz, J. A. J. Metz, É. Kisdi, G. Meszéna [1997]: Dynamics of Adaptation and Evolutionary Branching, Physical Review Letters, **78**, 10, pp. 2024-2027.

S. Ghashghaie, W. Breymann, J. Peinke, P. Talkner, Y. Dodge [1996]: Turbulent cascades in foreign exchange markets, Nature, **381**, 27 June 1996, pp. 767-770.

S. Ghashghaie, W. Breymann, J. Peinke, P. Talkner [1997]: Turbulence and Financial Market, A Transaction Cascade in Foreign Exchange Markets, az 1997. júliusában, Budapesten tartott Econophyiscs Workshop során osztott preprint.

I. Große, H. Herzel, S. V. Buldyrev, H. E. Stanley [1997]: Mutual information of coding and noncoding DNA (preprint).

H. Herzel, W. Ebeling, I. Große, A. O. Schmitt [1997a]: Statistical Analysis of DNA sequences (preprint).

H. Herzel, I. Große [1995]: Measuring Correlations in Symbol Sequences, Physica A, **216**, 518.

H. Herzel, I. Große [1997b]: Correlations in DNA sequences: The role of protein coding segments, Physical Review E, **55**, 1, pp. 800-810.

I. Kanter, D. A. Kessler [1994]: Markov Process: Linguistics, Zipf's Law and Long-Range Correlations. (Submitted to PRL, Oct. 20, 1994).

R. M. Kaplan, M. Kay [1994]: Regular Models of Phonological Rule Systems, Computational Linguistics, **20**, 3, pp. 331-378.

B. Krenn, Ch. Samuelsson [1996]: The Linguist's Guide to Statistics,
forrás: `http://coli.uni-sb.de/~christer`.

Y. Liu, P. Cizeau, M. Meyer, C.-K. Peng, H. E. Stanley [1997]: Correlations in Economic Time Series (preprint submitted to Elsevier Science).

K. Mainzer [1997]: Thinking in Complexity, The Complex Dynamics of Matter, Mind, and Mankind, Springer, Berlin, etc.

Martinás K. [1997]: A fenntartható fejlődés termodinamikája (személyes példány)

K. Martinás, M. Moreau [1995] (szerk.): Complex Systems in Natural and Economic Sciences, Proceedings of the Workshop Methods of Non-Equilibrium Processes...

R. N. Mantegna, S. V. Buldyrev, A. L. Goldberger, S. Havlin, C.-K. Peng, M. Simons, H. E. Stanley [1994]: Linguistic Features of Noncoding Sequences, Physical Review Letters, **73**, 23, pp. 3169-3172.

R. N. Mantegna, H. E. Stanley [1995a]: Scaling behaviour in the dynamics of an economic index, Nature, **376**, 6 July 1995, pp. 46-49.

R. N. Mantegna, S. V. Buldyrev, A. L. Goldberger, S. Havlin, C.-K. Peng, M. Simons, H. E. Stanley [1995b]: Systematic analysis of coding and noncoding DNA sequences using methods of statistical linguistics, Phys. Rev. E, **52**, 3, pp. 2939-2950.

R. N. Mantegna, H. E. Stanley [1996]: Turbulence and financial markets, Nature, **383**, 17. October 1996, pp. 587-588.

R. N. Mantegna, H. E. Stanley [1997]: Stock market dynamics and turbulence, Parallel analysis of fluctuation phenomena, Physica A, 3357.

Nagy Ferenc [1986]: Kvantitatív nyelvészet, Tankönyvkiadó, Budapest.

C.-K. Peng, S. V. Buldyrev, A. L. Goldberger, S. Havlin, F. Sciortino, M. Simons, H. E. Stanley [1992]: Long-range correlations in nucleotide sequences, Nature, **356**, pp. 168-170.

C.-K. Peng, S. V. Buldyrev, A. L. Goldberger, S. Havlin, M. Simons, H. E. Stanley [1993]: Finite-size effects on long-range correlations: Implications for analyzing DNA sequences, Physical Review E, **47**, 5, pp. 3730-3733.

R. F. Port, T. van Gelder [1995]: Mind as Motion, Explorations is the Dynamics of Cognition, MIT Press, Cambridge, Massachusetts, London, England.

M. Potters, R. Cont, J-Ph. Bouchaud [1997]: Financial markets as adaptive systems (preprint).

W. H. Press, B. P. Flannery, S. A. Teukolsky, W. T. Vetterling [1989]: Numerical Recipes in C, The Art of Scientific Computing, Cambridge University Press.

I. Prigogine, I. Stengers [1986]: La nouvelle alliance, Métamorphose de la science, Gallimard, Paris, 1986, *magyarul*: Az új szövetség, A tudomány metamorfózisa, Akadémiai Kiadó, Budapest, 1995.

L. R. Rabiner [1989]: A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition, in: Readings in Speech Recognition, ed.: A. Waibel & Kai-Fu Lee, Morgan Kaufmann, 1991.

Rend és Káosz [1997], Fraktálok és káoszelmélet a társadalomkutatásban, szerk.: Fokasz Nikosz, Replika Kör, Budapest

Révész Gy. [1979]: Bevezetés a formális nyelvek elméletébe, Akadémiai Kiadó, Budapest.

A. Schenkel, J. Zhang, Y.-C. Zhang [1993]: Long Range Correlation in Human Writings, Fractals, **1**, 1, pp. 47-57.

C. E. Shannon, W. Weaver [1986]: A Kommunikáció matematikai elmélete, az információelmélet születése és távlatai, OMIKK, Budapest.

M. H. R. Stanley, L. A. N. Amaral, S. V. Buldyrev, S. Havlin, H. Leschhorn, Ph. Maass, M. A. Salinger, H. E. Stanley [1996a]: Scaling behaviour in the growth of companies, Nature, **379**, 29 Febr. 1996, pp. 804-806.

M. H. R. Stanley, L. A. N. Amaral, S. V. Buldyrev, S. Havlin, H. Leschhorn, Ph. Maass, M. A. Salinger, H. E. Stanley [1996b]: Can Statistical Physics Contribute to the Science of Economics?, Fractals, **4**, 3 (1996), pp. 415-425.

H. E. Stanley, S. V. Buldyrev, A. L. Goldberger, J. M. Hausdorff, S. Havlin, J. Mietus, C.-K. Peng, F. Sciortino, M. Simons [1992]:Fractal landscapes in biological systems: Long-range correlations in DNA and interbeat heart intervals, Physica A, 191, 1-12.

H. E. Stanley, S. V. Buldyrev, A. L. Goldberger, S. Havlin, C.-K. Peng, M. Simons [1993a]: Long-range power-law correlations in condensed matter physics and biophysics, Physica A 200, 4-24.

H. E. Stanley, S. V. Buldyrev, A. L. Goldberger, S. Havlin, S. M. Ossadnik, C.-K. Peng, M. Simons [1993b]: Fractal Landscapes in Biological Systems, Fractals, **1**, 3, pp. 283-301.

Szépfalusy P., Tél T. [1982]: A káosz, Akadémiai Kiadó, Budapest

T. Vicsek [1992]: Fractal Growth Phenomena (second edition), World Scientific.

T. Vicsek, A. Czirók, E. Ben-Jacob, I. Cohen, O. Shochet [1995]: Novel Type of Phase Transition in a System of Self-Driven Particles, Physical Review Letters, vol. 75, 6, 1226-1229.

R. F. Weaver, Ph. W. Hedrick [1992]: Genetics, second edition, Wm. C. Brown Publishers

G. K. Zipf [1935]: The Psychobiology of Language, Houghton Mifflin, Boston.

G. K. Zipf [1949]: Human Behavior and the Principle of Least Effort, Addison-Wesley Press.