REKURZÍVAK-E A TERMÉSZETES NYELVEK?

Kornai András

az MTA doktora, tudományos tanácsadó,
Harvard University, MTA SZTAKI
kornai@sztaki.hu

Nem tudjuk mi történt ezzel a férfiúval, Mózessel, ki minket Egyiptomból kihozott
2 Móz. 32.1.

0. Bevezetés

A *Magyar Tudomány* 2009/9 számában hozza a Noam Chomsky 80. születésnapjára a Nyelvtudományi Intézetben 2008 decemberében rendezett szimpózium anyagát. A „vita Chomsky jelentőségéről" (http://www.nytud.hu/archives/chomskyvita2008.html) annyiban természetesen illuzórikus, hogy Chomsky jelentőségét, érdemeit (el)vitatni nem lehet, nincs a kortárs nyelvészek közt egy sem, akinek ilyen erős, szerteágazó és tartós hatása lett volna. A sok dicsérő, sőt időnként magasztalásba hajló írás közt éppen ezért némileg furcsán hat Kálmán László kijelentése: „A számítógépes nyelvészet főáramában a generativizmuson alapuló modelleket nem találunk. A szabályalapú megközelítések általában is sikertelennek bizonyultak."

Ebben a cikkben (szakmaibb, jegyzetekkel ellátott változatát lásd http://www.szv.hu/cikkek/rekurzivak-e-a-termeszetes-nyelvek) azt próbáljuk meg körüljárni, hogy miként történhetett ez meg. Hogy lehet, hogy Chomsky gondolatait ma éppen a nagyrészt általa elindított formális/számítógépes/matematikai nyelvészetben veszik legkevésbé komolyan? Kálmán szerint ennek egyik alapvető oka Chomsky antiempirista hozzáállása (melyet igen frappáns idézetekkel dokumentál), de szerintünk az igazi ok mélyebben van, és Chomsky munkásságának matematikai tartalmát figyelmen kívül hagyva nem is érthető.

A címben feltett kérdés körüljárását azzal kezdjük, hogy egy kicsit pontosabban megnézzük, mi a rekurzivitás és mi a nyelv. Természetesen ha *rekurzión* csak annyit értünk, hogy valamilyen konfiguráció ismétlődik, ismétlődhet, akkor a válasz triviális. Ilyen ismétlődésre jó példa a koordináció, hiszen a *Láttuk Jánost és Pétert és Zolit és …* konstrukció addig terjeszthető, amíg ki nem fogyunk a lélegzetből. (A leghosszabb ilyen mondat állítólag a második világháború végén az arlingtoni nemzeti temetőben hangzott el, ahol felolvasták a hősi halottak névsorát.) Hogy tovább tudjunk lépni ezen a trivialitáson, a rekurzivitás a matematikában megszokott definíciójánál maradunk: **rekurzíve felsorolható** az, amire Turing-gépet (Turing Machine – TM) be lehet programozni. Azok, akik a Turing-gépeket elsősorban a logikából ismerik (emlékeztetőül: a Turing-gép egy végtelen szalagból és az ennek mozgatását/írását/olvasását szabályozó véges kontroll-automa-

tából áll), gyakran hozzá vannak szokva, hogy ezek (bináris) számokon operálnak. Alan Turing eredeti definíciója ezt a megkötést nem tartalmazza: a szalagra tetszőleges véges szimbólumhalmaz elemeit írhatjuk. Egy rögzített TM által megadott **formális nyelv** azon füzérek halmaza, melyeket a gép szalagjára írva a program véges idő alatt vagy megáll úgy, hogy a szalag üres (üres szalaggal való elfogadás), vagy kitüntetett állapotok valamelyikébe kerül (állapothalmazzal való elfogadás), vagy egy előre rögzített füzér, illetve erre a célra fenntartott OK-szimbólum kiírásával reagál (jelzéssel való elfogadás). Bár egy adott nyelvhez természetesen másféleképpen kell programozni a TM-et, aszerint, hogy melyik elfogadás-definíciót választjuk, összességében a TM-ek által megadható (Chomsky terminológiájával: **nullás típusú**) nyelvek halmazát ez a döntés nem befolyásolja.

A *nyelv* fogalmának alapos, mind filozófiai, mind nyelvészeti szempontból kielégítő definíciója messzire vezetne, de céljainkhoz ez nem is szükséges, hiszen matematikai kérdést csupán matematikai objektumokról lehet feltenni. A formális nyelv a természetes nyelvek Chomsky által bevezetett matematikai modellje: a nyelvészeti alkalmazásokban az ábécé elemeire gyakran mint a természetes nyelv fonémáira (általában néhány tucat elem), illetve szófajaira (gyakran többezer elem) gondolunk. A természetes nyelvről fontos plusz információ, amit most első közelítésben elhanyagolunk, hogy a szavak és nagyobb konstrukciók közt különféle kapcsolatok állhatnak fenn, és hogy a szavaknak/mondatoknak mérhető gyakoriságuk/valószínűségük van. A formális nyelvekről immár formális szigorúsággal felvethető az a kérdés, hogy vajon végesek vagy végtelenek, és ha végtelenek, akkor rekurzívak-e?

A kérdés, ha nem is egészen ebben a formában, egyidős a formális nyelvészeti kutatással, melynek alapjait még Pāṇini (i. e. 520–460) vetette meg, nagyjából két évszázaddal az előtt, hogy Euklidész megvetette a matematika alapjait. A *Mahābhāṣhya (Nagy Kommentár)* az első fennmaradt Pāṇini-magyarázat, i. e. 200 körülről. A bevezető részben a szerző, Patañjali azzal kezdi, hogy egyszerűbb a helyes (grammatikus) alakokat felsorolni, mint a helyteleneket, majd azt a kérdést veti fel, hogy hogyan kell ezt megcsinálni: szedjük listába a helyes alakokat? Nem, ez túlságosan nehéz lenne. *Mert mint tudjuk*, Bṛhaspati (az istenek tanára) ezer égi évig (360 ezer földi évig) tanított Indrának egy olyan munkát, amely felsorolta a helyes szanszkrit kifejezéseket, és még így sem jutott a végére. Akkor hogy lehetne most, amikor az emberek még száz nyarat sem élnek meg, ily módon tanítani?

Igaz ugyan, hogy egy adott nyelv eddig elhangzott/leírt mondatai véges halmazt alkotnak, de a nyelvészek körében teljes az egyetértés (és mindig is az volt), hogy az ezen a tényen alapuló naiv modell érdektelen, hiszen minket nemcsak egy létező korpusz leírása érdekel, hanem az is, hogy predikciókat tegyünk a még el nem hangzott (vagy le nem írt) mondatok halmazára nézve is. Ha megadjuk a koordináció szabályait, például a perl, python és más programnyelvekből ismert szabályos kifejezésekkel (regular expressions), akkor máris egy olyan nyelvtanunk van, amely végtelen sok, eddig még nem hallott/látott mondat elfogadhatóságára tesz *tesztelhető* jóslatot. 1956 előtt a matematikusok a végtelen nyelvek algoritmikus megadására csupán két módszert tartottak számon: a véges automatákon (vagy ami ugyanaz, szabályos kifejezéseken) keresztüli, illetve a Turing-gépes

definíciót. A véges automaták által elfogadott (szakszóval: **hármas típusú**) nyelvekre lehet úgy is gondolni, mint az olyan TM-ek által elfogadott nyelvekre, amelyek csak olvasni tudnak, de a szalagot nem írhatják (természetesen ehhez az állapothalmazzal való elfogadás-definíciót kell használni). Az ilyen TM csak véges sokféle részeredmény megjegyzésére képes, hiszen a memóriakapacitását behatárolja a kontroll-automata (véges) állapottere.

Az **egyes típust** Chomsky a XIX. századi neogrammatikus hangtörvények formalizálására általa bevezetett **környezetfüggő nyelvtan** (context sensitive grammar – CSG) segítségével írta le: ezekben a füzérek egyes elemeit a szabályok át tudják alakítani akkor, ha az elemek környezetére bizonyos feltételek teljesülnek. Például a magyar szavak végén a zöngétlen mássalhangzók zöngésülnek, ha zöngés mássalhangzóval kezdődő rag (vagy szóösszetétel második eleme) követi őket: vaskalap, vazsgolyó (ejtésben), vassal, de vazsból. A környezetfüggő nyelvtanok ezt a tényt egy s→zs/_Z szabállyal ragadják meg, melynek jelentése *cseréld ki s-t zs-re ha jobboldali kontextusa Zöngés*. A Turing-gépek perspektívájából nézve az egyes típust úgy nyerjük, hogy engedélyezzük az írást (a részeredmények tárolását a gép szalagján) de csak bizonyos korlátok közt: a TM-nek csak akkora memóriát teszünk írhatóvá, amekkora a bemenő füzér.

A **kettes típust** Chomsky a közvetlen összetevős elemzés formalizálására szintén általa bevezetett **környezetfüggetlen nyelvtan** (context free grammar – CFG) segítségével definiálta. Ezekben a nyelvtanokban szintén x→y alakú szabályok vannak, de most mindenféle megszorítás nélkül: egy ilyen szabály mindig alkalmazható, *függetlenül* attól, hogy *x* előtt és után milyen szimbólumok állnak. (Szigorú értelemben itt is és az egyes típusnál is meg kell különböztetni az ún. terminális és nemterminális szimbólumokat, ennek részleteit most figyelmen kívül hagyjuk.) Megemlítjük, hogy ez az osztály nem zárt komplementációra: például (legalább kételemű ábécé fölött) az az *N* nyelv, amely a nem négyzetes füzérekből áll (tehát elemei nem állnak elő *xx* formában, ahol *x* tetszőleges füzér) kettes típusú, míg komplementuma, tehát az az *I* nyelv, ami pontosan a négyzetes (*xx* alakú) füzérekből áll, nem lesz kettes típusú. Az eddigieket összefoglalva már készen is áll az eredeti **Chomsky-hierarchia,** melyet itt bővített formában hozunk (az eredeti 0-3-hoz itt hozzátett nyelv-, illetve nyelvtanosztályokról később lesz szó). *(1. táblázat)*

E tipológia annyiban hierarchikus, hogy a csökkenő számoknak egyre bővülő eszköztár felel meg: minden nyelv, amit le tudunk írni 3. típusú nyelvtannal, az leírható 2. típusúval is, amit le lehet írni 2. típusúval, az le-

típus	nyelvosztály	definíciós eszköz	nyelvtan
0	rekurzíve felsorolható (r. e.)	TM (egyoldalú)	tetszőleges
0.5	rekurzív	TM (kétoldalú)	
1	környezetfüggő (CSL)	lin. korl. aut (LBA)	környezetfüggő (CSG)
1.5	enyhén környezetfüggő (MCS)	beágyazott veremautomata	linear indexed, CCG, LTAG
2	környezetfüggetlen (CSL)	veremautomata (PDA)	környezetfüggetlen (CFG)
3	véges állapotú (regular)	véges automata (FSA)	FSG, szabályos kifejezések

1. táblázat

írható 1. típusúval is, és persze minden, amit egyáltalán le lehet írni nyelvtannal, az leírható Turing-géppel is. Chomsky érdeme, hogy a címben felvetett triviális kérdést egy sokkal izgalmasabbra cserélte fel: *hova esnek a nyelvek a Chomsky-hierarchiában?* (amit ő természetesen még nem hívott így). De ha egyszer ilyen jó, tartalmas kérdést tett fel, olyan formai eszközöket kínálva, melyek egyben a mesterséges (programozási) nyelvek elméletét is forradalmasították, akkor végül is miért vesztette el hitelét pont a legbelsőbb szakmai körökben? Felfogásunk szerint ez csak úgy történhetett meg, hogy a kérdésre nemcsak rossz választ adott, hanem ahhoz kitartóan, egyre nagyobb retorikai vehemenciával ragaszkodott akkor is, amikor a tények ennek minden irányból ellentmondottak. Történetileg Chomsky radikális antiempiricizmusa nem ok, hanem okozat: ha nem kvadrálnak az elmélettel, hát antul rosszabb a tényeknek.

1. A korai szakasz: 1956–1982

Chomsky nemcsak felvetette a problémát, de úgy vélte, hogy kielégítően meg is oldotta. Azt az állítást, hogy a harmadik típus nem elégséges a természetes nyelvek leírásához, az ún. középponti beágyazás (center-embedding) jelenségével indokolta: matematikailag bebizonyította, hogy az olyan CF-nyelvtanok, amelyek megengednek X→aXb alakú levezetést (ahol tehát a végeredményben a kiinduló *X a* és *b* közé beágyazva jelenik meg) szükségképpen túllépnek a 3. típuson (e kikötés nélkül ez nem igaz, CF, azaz 2. típusú nyelvtan is generálhat olyan nyelvet, amely szabályos kifejezésekkel, azaz 3. típusú nyelvtannal is megadhatók) majd rámutatott, hogy az angolban a vonatkozói mellékmondatok középponti beágyazott helyzetben is megjelenhetnek: *a rat that stole the cheese, a cat a woman loves, the cheese that a rat (that a cat (that a woman loves) chased) stole*. A *Mondattani szerkezetek* (1957, magyarul 1999) ezért írja, hogy „Nemcsak nehéz, de *lehetetlen* olyan [véges automatát] létrehozni, amely az angol nyelv valamennyi nyelvtanilag helyes mondatát létrehozná, és csak azokat. […] E tétel azt állítja, hogy a nyelv […] Markov-folyamat koncepciója elfogadhatatlan, legalábbis a nyelvtan céljaira." (Chomsky, 1957, 24.)

Az érvelés nyelvtani része, különösen a zárójelezés nélkül gyakorlatilag érthetetlen: *the cheese that a rat that a cat that a woman loves chased stole* már annak idején is sok vitát váltott ki, erre a kérdésre majd a 2.1 szakaszban térünk vissza. Chomsky (1957) nem sok kétséget hagyott a felől sem, hogy szerinte a CF-nyelvtanok sem elégségesek a feladathoz: „[A CF-nyelvtanok] angol nyelvre történő alkalmazásának korlátait tovább vizsgálva, meggyőzően igazolható, hogy ezek a nyelvtanok olyan reménytelenül bonyolultak, hogy teljesen érdektelenné válnak, hacsak nem építünk beléjük [transzformációkat]." (Chomsky, 1957, 50.)

A korai szakaszban ezt az érvelést szinte mindenki elfogadta, sőt nem egyszerűen elfogadta, hanem mint a XX. századi nyelvtudomány legnagyobb felfedezését ünnepelte: „The single most important contribution to the development of linguistic theory in the [20th] century is [the demonstration of] the inadequacy of CFGs as a model of linguistic structure."[1] (Selkirk, 1977)

A tét nagy: ha sikerül általános matematikai formulákkal leírni a nyelvtanilag helyes mondatok generálási szabályait, akkor hatal-

[1] A huszadik század legeslegfontosabb hozzájárulása a nyelvtudomány fejlődéséhez annak a bebizonyítása, hogy a környezetfüggetlen nyelvtanok alkalmatlanok a nyelvi szerkezetek modellálására.

mas lépést tettünk a gépi fordítás, a géppel történő dialógus, az automatikus szövegkezelés felé. Patañjali teljes joggal elvárhatta olvasóitól a védikus bölcsesség ismeretét és feltétel nélküli elfogadását, de a modern nyelvészektől már kicsit furcsábbnak tűnik a *mert mint tudjuk* érv használata. E korszak végét Geoffrey Pullum és Gerald Gazdar (1982) ma már klasszikus „meztelen a király" cikke jelzi (Pullum – Gazdar, 1982, 471–504), melyben sorra vették az irodalomban fellelhető érveket, és egyenként kimutatták róluk, hogy tarthatatlanok, méghozzá három egymással gyakran összefüggő hiba miatt. Ezek közül az első és legfontosabb az, hogy időről időre *1. az eredeti érvelés matematikailag hibás.* Erre jó példa Chomsky saját érvelése, ami azon a jelenségen alapul, hogy az angolban a középfokú összehasonlításban *nem szeretjük*, ha ugyanazzal hasonlítunk: *This desk is wider than that chair is tall* de **This desk is wider than that chair is wide*. Ez utóbb esetben inkább az összehasonlítás alapját képező NP törlésével dolgozunk: *This desk is wider than that one*. Hogy ez a „nem szeretjük" mit jelent, arra majd később visszatérünk (Pullum és Gazdar igen szórakoztatóan írnak arról, ahogy Chomsky később megváltoztatta az itt még csillaggal hozott mondatok grammatikalitásáról való véleményét), most fogadjuk el, hogy a jelenség valóban így igaz. A baj az, hogy az így kijelölt *N* nyelv nem ellenpélda CF-nyelvre, csak a komplementuma, *I* lenne az, de a CF-család nem zárt komplementumra! *Quandoque bonus dormitat Homerus*.

A második, hasonlóan gyilkos ellenérv az, hogy *2. Az eredeti érvelés összekeveri a szintaxist a szemantikával*. Ezt most Zwicky (1963) példáján illusztráljuk, amely a *trillió, kvadrillió, kvintillió (trilliárd, kvadrilliárd, kvintilliárd)* és hasonló nagy számok nyelvi kifejezésén alapul. Nem tudjuk, mi a legnagyobb ilyen, de nem is fontos, hogy elkötelezzük magunkat egy konkrét *-illió* (vagy *-illiárd*) mellett, legyen a **zillió** a legnagyobb szótári szó, ami 1000^n-t fejez ki. Ennek a négyzete *egyzillió zillió*. Még ennél is nagyobb szám az *egyzillió zillió egyzillió egy*. De az **egyzillió egyzillió zillió* nem legális számnév, mert a nagyobb zillió-hatványokat kell előbb mondani.

$$\{p_1 z^{n_1} p_2 z^{n_2} \ldots p_r z^{n_r} | n_j > n_{j+1}\} \notin \mathrm{CF}$$

A probléma az, hogy ez nem nyelvtani, hanem matematikai tudás. Későn sajátítjuk el, és nem is mindenki tudja, aki egyébként kompetens anyanyelvi beszélő. Ugyanez a baj a híres *respectively* konstrukción alapuló érveléssel is, mely szerint a *John, Mary, and Bill are a widower, widow, and widower respectively* típusú mondatokban, ha csupán a nem szerint egyértelmű keresztnevekre szorítkozunk, és elvárjuk hogy *widower* csak hímnemű, a *widow* csak nőnemű legyen, akkor a grammatikus mondatok halmazát az *xx* halmazba tudjuk képezni, ahol *x* tetszőleges füzér a kételemű *hímnem, nőnem* halmaz felett (tehát a nem-CF *I* nyelvet nyerjük).

Külön hangsúlyozzuk, hogy a nyelvtan nem törődik a tényekkel; az a mondat, hogy *Einstein was a great physician* grammatikailag ugyanolyan helyes mint az, hogy *Einstein was a great physicist* bár tényszerűleg az egyik igaz, a másik hamis. Az *Anna özvegyember* mondat valóban nehezen értelmezhető (hacsak nem Boris Viannál találjuk) de ebben a nehézséget nem a mondatszerkezet, hanem a világról való ismereteinkkel való összeférhetetlenség okozza. Igen, de nem lenne elképzelhető olyan nyelv, ahol a nem szerinti egyeztetés nem szemantikai, hanem grammatikai kérdés? Miután pontosan tudjuk, hogy számtalan ilyen nyelv van, a *respectively*-n alapuló érvelés esetleg az angolban nem, de mondjuk,

a spanyolban tarthatónak tűnik. A probléma az, hogy nyelvtani alapon már a két felsorolás hosszúságának megegyezése sem garantálható, hiszen a *Going left to right, the last two people in the line are John and Bill respectively* mondat helyes, szemantikailag is és grammatikailag is, pedig a *respectively*-vel összekapcsolt felsorolások nem tartalmaznak ugyanannyi elemet, hiszen a baloldalt egyetlen NP, *the last two people*, áll szemben a jobboldalt két NP-vel, *John and Bill*.

A harmadik ellenérv annyiban hasonló az elsőhöz, hogy ez is egy matematikai hibát pécéz ki: *3. az eredeti érvelés empirikusan lyukas*. Általában ahhoz, hogy egy nyelv nem-CF voltát igazoljuk, nem elég rámutatni egy nem-CF résznyelvre, mert a Chomsky-hierarchia nem zárt tartalmazásra, egy nem-CF nyelv résznyelve is lehet CF, és egy CF-nyelv résznyelve is lehet nem-CF (és hasonlóan a hierarchia többi tagjára, a véges nyelvek családjának kivételével). A problémát az egyik legkorábban felfedezett és legizgalmasabb jelenségkör, a mohawk főnév-inkorporáció (Postal, 1964) erősen egyszerűsített változatán illusztráljuk. A nyelvészetben szokatlan módon elhagyjuk az eredeti mohawk példamondatokat és csupán magyarított glosszákat adunk (az eredeti mondatok megtalálhatók Paul Postalnál és kritikusainál). A mohawk nyelv a tárgyas ige tárgyát gyakran megismétli az igei csoportba beépítve: *Nekem ház-tetszik a ház* „Tetszik a ház". Az inkorporált elem lehet pronominalizált formában is: *Nekem ideatetszik ez* „Egyetértek ezzel". Postal azt állította, hogy az inkorporált főnév megegyezik az inkorporálatlan (külső) tárggyal, a mohawk tehát *I* nyelv. Igen ám, de az általa vizsgált nem az egyetlen inkorporatív konstrukció, be lehet építeni teljes birtokos szerkezeteket is: *Nekem János-ház-tetszik János ház* 'Tetszik János háza'. Ez még nem lenne baj, de az ilyen szerkezetekből a birtokos elhagyható: *Nekem ház-tetszik János* 'Tetszik János háza', és ez betölti a lyukakat, a nyelv tehát végső soron nem *I* jellegű. Már itt megjegyezzük, hogy a mohawk egyik legalaposabb leíró nyelvésze, Floyd Lounsbury szerint az érvelés eleve fiktív annyiban, hogy az inkorporáció nem iterálható, a kétszeres inkorporálásnál az egyik tő mindig egy idióma része, de ez most a birtokos szerkezet által felszínre hozott probléma szempontjából közömbös, a jelenségre később térünk majd vissza.

2. Az elszakadás időszaka: 1982–2000

Geofrey Pullum és Gerald Gazdar cikke csupán negatív érveket hozott, és retorikailag nyitva is hagyta a kérdést, hogy vajon a második Chomsky-típusba beleférnek-e a természetes nyelvek. Sokkal fontosabb volt, hogy ezek a szerzők megalapozták az általánosított frázis-struktúra nyelvtan (generalized phrase structure grammar – GPSG) elméletét, amelyben a nehéz, mindaddig a természetes nyelvek nem-CFL voltának igazolására használt nyelvi problémákat, mint például a hosszú távú függőség *(unbounded dependecy),* sorra oldották meg. De nem tartott sokáig, amíg megjelentek az új érvek, elsősorban Stuart Shieber (1985) a svájci némettel foglalkozó, Christopher Culy (1985) a bambara nyelvvel foglalkozó, és Kenneth Beesley és Lauri Karttunen (2000) a malájjal foglalkozó cikkei – ez utóbbi érdekessége, hogy nem a szintaxisban, hanem már egy lépéssel előbb, a morfológiában (ahol z füzérek a szavak, az ábécé pedig a morfémák) mutat nem-CF konstrukciót.

Elődeikkel ellentétben ezek a munkák már matematikailag hibátlanok, tisztán nyelvtani (nem pedig szemantikai) tényeken alapulnak, és empirikusan sem lyukasak. Ez azonban

nem jelenti azt, hogy a kérdést végképp eldöntik, hiszen másfajta gyengeségeik azért még lehetnek, és mint látni fogjuk, vannak is. A modern ellenérvek két nagy csoportra oszthatók, egyrészt a megfigyelhető bizonytalan grammatikai státus, a „nem szeretjük" körüli problémák, ezekről *korlátozott iterativitás* néven beszélek a 2.1 részben, másrészt a *nagyon kis gyakoriság* okozta problémák, lásd 2.2. Egy kicsit előreugorva megjegyezzük, hogy ezek az ellenérvek egyben a klasszikus középponti beágyazási példákat is kilövik, így nemcsak a 2. osztály elégtelensége, hanem az ennél jóval kisebb 3. osztály elégtelensége (és ezzel Chomsky eredeti, a Markov-modellezéssel szembeni dörgedelmei) is kérdésessé válnak. De mielőtt erre rátérnénk (lásd 3.), lássuk a modern ellenérveket részletesebben.

2.1 Bizonytalan grammaticitás, korlátozott iterativitás

A klasszikus generatív felfogásban éles dichotómia van a grammatikus (OK) és az agrammatikus (*) mondatok közt. Hogy egy konkrét kifejezés hova esik, azt a nyelvész intuíciója (illetve az anyanyelvi informáns) dönti el. Sajnos a Shieber, Culy és mások által vizsgált szerkezetek mindegyike nagyon hamar olyan kifejezésekhez vezet, ahol a nyelvész/informáns intuíciója elbizonytalanodik. Ezt az önmagában érdekes tényt Chomsky (1965) a performancia és a kompetencia közti megkülönböztetéssel próbálta magyarázni, de nyitva hagyta azt a kérdést, hogy ha a beszélők fejében lévő grammatikai apparátus olyan nagyon komplex, akkor miért pont ezek a kifejezések okoznak nehézséget, míg egyéb tetszőlegesen nagyra növelhető konstrukciók (mint a koordináció) nem.

Az általános performancia-probléma fontos speciális esete az, amit itt *korlátozott iterativitásnak* fogunk nevezni, lássuk ezt egy egyszerű beágyazási példán. Tekintsük először elemi kijelentések valamilyen *S* halmazát: *Meleg van, esik az eső, kigyulladt a ház…*, majd kezdjük el bővíteni ezt attitűdöt kifejező kijelentésekkel: *Az hogy S (az) hazugság/egy nagy hülyeség/biztos/kétségbeejtő/…* Az első iterációban egészen rendes, értelmes magyar mondatokat nyerünk: *Az hogy esik az eső az kétségbeejtő, az hogy kigyulladt a ház az hazugság,...* Mindez valamiféle S→Th S (D) Att szabály felvételét indokolja, ahol *Th* az „Az hogy" formatíva, *D* az „Az" formatíva, *Att* pedig az attitüdinális kifejezések „kacsa, hétszentség, elszomorító, …" gyűjteménye. A második iterációban ezek a szabályok már különös eredményeket hoznak: *??Az hogy az hogy meleg van az kacsa az elszomorító* – mit is jelent ez? Hát, vidámabbak lennénk, ha a hír nem lenne kacsa (hanem tényleg meleg lenne). Ez még talán rendben is van, bár a kognitív folyamat már inkább a rejtvényfejtésre, mint a szokásos nyelvi megértésre emlékeztet. De ha még egyszer-kétszer iterálunk, az amúgy olyan remekül működő mondatelemzőnk végképp fejreáll: *????Az hogy az hogy az hogy esik az eső az bizonytalan az hétszentség az hazugság*, és csak a rejtvényfejtés marad.

A középponti beágyazás hamar kivezet az emberi ésszel felfogható (és előállítható) mondatok köréből: ezt találjuk más nyelveknél és más konstrukcióknál is. Fred Karlsson (2007) tizenhat nyelvre kiterjedő vizsgálatai szerint az írott nyelvben maximum háromszoros, a beszélt nyelvben maximum kétszeres beágyazást találunk. Ez hát egy erős, jól replikálható nyelvi jelenség, és ha ezt tudjuk, mindegy is, hogy a kompetencia vagy a performancia részének tekintjük. Chomsky (1965) még elsősorban azért különítette el a kompetenciát a performanciától (ezzel nagy, évtizedekig

nem csillapuló módszertani vihart kavarva) hogy a középponti beágyazások korlátozott iterativitását átsorolhassa a performanciába, és ezáltal (hiszen minket mint nyelvészeket a kompetencia modellezése jobban érdekel) fenntarthasson egy olyan idealizációt, ami kivezet a szabályos kifejezések közül. De ebben a formában az érvelés már nem meggyőző: ha egyszer a naiv matematikai modell, ami az iterálást egyáltalán nem korlátozza, a tényektől épp egy ilyen kritikus ponton tér el, akkor célszerűbbnek tűnik a modellt finomítani, például ellátni egy olyan számlálóval, ami legfeljebb egyszeres vagy kétszeres iterációt engedélyez. Tulajdonképpen mindegy is, hogy hánynak választjuk ezt a *d* iterációs korlátot, kettőnek vagy ötnek, hiszen a kétszer és az ötször iterált konstrukciók közötti különbséghalmazban már csak marginális (grammatikailag kétes és szemantikailag csak igen nehezen értelmezhető) füzérek lesznek.

2.2 Gyakoriság

A klasszikus érvelés (Chomsky, 1957, 2.4) szerint a nyelvtan világában a gyakoriság nem számít, hiszen *colorless green ideas sleep furiously* és *furiously sleep ideas green colorless* egyaránt nulla gyakoriságúak, de előbbi grammatikus, utóbbi pedig nem. Ha ez igaz, a grammaticitás nem jellemezhető valószínűséggel, hiszen itt mindkét példa gyakorisága nulla. A tudomány történetének különös fintora (bővebben lásd Pereira, 2000), hogy ezt a minden matematikusnak azonnal láthatóan hibás érvelést a szakma évtizedekig nem tudta, nem merte megkérdőjelezni. Hol a hiba? Ott, hogy a nulla empirikus frekvenciából *nem következik* nulla valószínűség.

Természetesen mindkét mondatnak nagyon kicsi a valószínűsége. Ez már abból is kiderül, ha a mért szógyakoriságokat egymástól függetlennek tekintő (unigram) modellt vesszük, hiszen ekkor a mért szógyakoriságokat összeszorozva $2{,}14\times10^{-25}$ körüli értéket nyerünk – ebből már látható, hogy mindenképpen nagyon nagy mintára lenne szükség ahhoz, hogy az ilyen jellegű mondatok előbukkanjanak. Ha most a nyilván túlságos egyszerűsítést jelentő függetlenségi feltevést elhagyjuk (annál is inkább, hiszen az unigram modellek még nem különítik el a szavak permutálásával nyert füzérekre jósolt valószínűségeket), és szópárokon, szóhármasokon alapuló (bigram, trigram) modelleket veszünk, akkor a két mondat valószínűségére egyre inkább eltérő értékeket kapunk. A híres példában a két valószínűség hányadosa mintegy 2×10^{5}, tehát a Chomsky által grammatikusnak ítélt változat mintegy kétszázezerszer valószínűbb agrammatikus társánál. Ezen az intervallumon belül bárhol (tehát meglepően robosztusan) meghúzhatjuk a határt úgy, hogy a *colorless green ideas sleep furiously* grammatikusnak, a *furiously sleep ideas green colorless* pedig agrammatikusnak minősüljön, pusztán valószínűsége alapján. Igaz ugyan, hogy ezt a valószínűséget matematikai modelljeink csupán becsülni tudják, direkt méréséhez nem áll rendelkezésünkre elégséges minta, de ez módszertanilag épp oly kevéssé zavar minket, mint az, hogy a nap belsejének a hőmérsékletét sem tudjuk hőmérővel megmérni.

Gyakran találkozunk a fenti hibás érvelés konverzével is, mely szerint „a bizonyíték hiánya nem a hiány bizonyítéka” – abból, hogy egy kifejezést a korpuszban nem találunk meg, még nem tudjuk megmondani, hogy a kifejezés csak ritka vagy tényleg agrammatikus. Ha ez igaz, akkor az intuícióra (akár a nyelvészére, akár az informánséra) való hivatkozás a nyelvészet kikerülhetetlen része. Természetesen ez az érv ugyanúgy nem állja meg

a helyét, mint az előző. Hol a hiba? Vegyük például azt az érdekes jelenséget, hogy az angol *cost* igének nincs passzívuma: *The book cost thirty dollars. *Thirty dollars were cost(ed) by the book.* Való igaz, hogy a passzívum hiányát nyelvi intuíciónk világosan jelzi – a fentebb tárgyalt példákkal ellentétben itt senki nem fog a csillagok elhelyezésén vitatkozni. De tényleg csak az jelzi? Anatol Stefanowitsch (2006) az alábbi kétszer kettes kontingenciatáblát közli:

	Passive	Active	Total
cost	0	63	63
-cost	13,861	122,627	136,488
Total	13,861	122,690	136,551

Ebből bármilyen megszokott statisztikai teszttel (például Fisher–Yates) kiszámolható, hogy a bal felső sarokban álló nulla nem véletlen nulla, az a tény, hogy a *cost* esetén nem találunk passzív alakot szignifikáns ($p<0.01$). Külön figyelmet érdemel az, hogy a statisztikai és a performancia-alapú megfontolások igen hasonló eredményre vezetnek: ha csak annyit teszünk fel, hogy az S→Th S (D) Att szabály mondjuk 1/1000 valószínűséggel működik, akkor iterációjának már csak egy a millióhoz, kétszeri iterációjának már csak egy a milliárdhoz az esélye.

2.3 A fennmaradó esetek

Bár a CFG-ellenpéldák eredeti bestiáriumából nem sok maradt, van mégis egy olyan konstrukció a hollandban, amelyre már Rini Huybregts (1976) felhívta a figyelmet (ez mind szinkron nyelvtanát, mind történeti kialakulását tekintve közeli rokona a Stuart Shieber (1985) tárgyalta svájci német példának), és amely változatlanul sok fejtörést okoz, annak ellenére, hogy mint füzérhalmaz (stringset) környezetfüggetlen. A holland *hogy*-os mellékmondatok szórendjét, beágyazott infinitivális tárgyak esetén, kereszteződő szerkezet jellemzi:

… dat Jan de kinderen zag zwemmen
hogy Jan a gyerek.PL lát.PAST úszik.INF
hogy Jan látta a gyerekeket úszni

… dat Piet de kinderen hielp zwemmen
hogy Piet a gyerek.PL segít.PAST úszik.INF
hogy Piet segítette a gyerekeket úszni

… dat Marie de kinderen liet zwemmen
hogy Marie a gyerek.PL küld.PAST úszik.INF
hogy Marie elküldte a gyerekeket úszni

A kereszteződés (crossed dependency) azt jelenti, hogy a dependenst a fejjel összekötő gráf élek (például *Jan* és *lát* illetve *gyerek* és *úszik* közt) keresztezik egymást, hiszen nem a gyerek lát és Jan úszik hanem épp fordítva. Az ilyen szerkezeteket rekurzíve egymásba is lehet helyettesíteni *(2. táblázat).*

Igaz, hogy a nyelv CF ($a^n b^n$), de a struktúra nyilván nem az, mert az *i*-edik *a* az *i*-edik *b*-hez kapcsolódik, nem pedig az *n–i*-edikhez, míg egy CF-nyelvtan, például S→aSb;S→ab ez utóbbi struktúrát állítaná elő. Ezeket a tényeket Chomsky és tanítványai a mozga-

… dat Jan Piet de kinderen zag helpen zwemmen
hogy Jan Piet a gyerek.PL lát.PAST segít.INF úszik.INF
hogy Jan látta Piet-et (amint) segíti a gyerekeket úszni

… dat Jan Piet Marie de kinderen zag helpen laten zwemmen
hogy Jan Piet Marie a gyerek.PL lát PAST segít.INF elküld.INF
hogy Jan látta Piet-et Marie-nak segíteni elküldeni úszni a gyerekeket

2. táblázat

tószabályok (transzformációk) cáfolhatatlan bizonyítékának tekintették, de már csak ők tekintették annak, mert a más forrásból (elsősorban a kategoriális grammatika elméletéből) merítő modern matematikai nyelvészet számos alternatív eljárást dolgozott ki az ilyen esetek kezelésére: itt csak a beillesztés (wrap), a fa-adjunkció (tree adjunction), és a kombinátoros kategoriális nyelvtan (combinatory categorial grammar) módszereit említem. Külön érdekesség, hogy ezeknek az egymástól gyökeresen eltérő eljárásoknak mindnek van olyan variánsa, amelyik ugyanahhoz az *enyhén környezetfüggő* (mildly context sensitive) nyelvosztályhoz vezet, melynek a fenti táblázatban a másfeles típusszámot adtuk.

3. Nébó hegyén: 2000–

Az enyhe környezetfüggés fogalmával a kiinduló kérdésünk körüli vita annyiban nyugvópontra jutott, hogy ennél bővebbet ma senki nem javasol a természetes nyelvek kezelésére, maga Chomsky sem, akinek „minimalista" elmélete ugyancsak egy enyhén környezetfüggő osztályra mutat. Tudományszociológiailag azonban nem elhanyagolható az a tény, hogy a Chomsky-hierarchiában a CFG-nél bővebb, de a CSG-nél szűkebb nyelv- és nyelvtanosztályok szisztematikus vizsgálatát nem Chomsky, hanem a kortárs matematikai nyelvészet legnagyobb alakjának tartott Aravind Joshi kezdeményezte, és a legfontosabb előzmény, a lineáris indexált nyelvtanok, sem a nyelvészetből, hanem a számítógéptudományból indult, abból a formális programelemzésből (compiler design), melynek alapjait indirekte még Chomsky vetette meg. A minimalizmus a Chomsky-tanítványok körében sem talált egyértelműen lelkes fogadtatásra, sőt vannnak, akik egyenesen miszticizmussal vádolják Chomskyt az elmélet alapját adó tökély-hipotézist (perfection) ilyeténképp jellemezve:

Imagine a biologist specializing in human physiology announcing that (…) his work is motivated by two related questions: (1) what are the general conditions that the human urinary tract should be expected to satisfy? , and (2) to what extent is the urinary tract determined by these conditions, without special structure that lies beyond them? The first question in turn has two aspects: what conditions are imposed on the urinary tract system by virtue of (A) its place within the array of physiological systems of the body and (B) general considerations of conceptual naturalness that have some independent plausibility, namely simplicity, economy, symmetry, non-redundancy, and the like?

It seems to us, and we suspect would to the great majority of working physiologists, that to ask what conditions the human urinary tract should be expected to satisfy makes no sense whatsoever. (…) Why then would one expect that it makes any more sense with 'language faculty' substituted for 'urinary tract'?[2] (Lappin et al., 2000)

[2] Képzeljünk el egy, az emberi fiziológiára szakosodó biológust, amint kijelenti, hogy […] munkáját két, egymással összefüggő kérdés vezérli: (1) mik azok az általános feltételek, amelyek teljesítését elvárhatjuk az emberi húgyúttól? és (2) milyen mértékben határozzák meg ezek a feltételek az emberi húgyutat, figyelmen kívül hagyva a mögöttes speciális struktúrát? Az első kérdésnek két aspektusa is van: milyen feltételeknek van alávetve a húgyút (A) az emberi test fiziológiás rendszerei közt betöltött helye által és (B) olyan általános fogalmi megfontolások alapján, mint egyszerűség, gazdaságosság, szimmetria, irredundancia és hasonlók? Nekünk (és gyanítjuk, a fiziológiával foglalkozók nagy többségének is) úgy tűnik, hogy semmi értelme nincs azt kérdezni, hogy a húgyútra vonatkozóan milyen feltételek teljesülése *várható el*. […] Ha ez így van, nem remélhetjük, hogy a kérdésnek több értelme lesz akkor, ha a kérdések tárgya a *húgyút* helyett a *nyelvi készség*.

Messzire vinne annak vizsgálata, hogy Chomskynak ma mekkora hatása van az elméleti nyelvtudományon belül a szintaxis kutatóira, de azt gondoljuk, e hatás máig jelentős (az idézet szerzői szerint jóval nagyobb, mint azt a nyelvtan tényei indokolnák). Bennünket most az a kérdés érdekel, hogy az elméleti nyelvtudománytól távolabb álló, a nyelvtan számítógépes modellezésére törekvő kutatók miért szakadtak el a Chomsky által kijelölt kutatási iránytól, hisz az új elmélet, a generatív grammatika a kezdeti időszakban elsősorban az ő körükben hódított.

A legfontosabb tényező kétségkívül az, hogy eltelt negyven év, és a sok bolyongás után a csapat, vagy legalábbis az előőrse, megérkezett az ígéret földjére. A beszédmegértés és -szintézis technológiája különösebb csinnadratta nélkül a mindennapi élet részévé vált: ma már gyakran emberi beavatkozás nélkül kapunk a telefonba feltett kérdésre választ, és a szakértők sem tudják megkülönböztetni, még műszeres elemzéssel sem, a mesterséges és a természetes beszédet. Minden szoftverboltban kapható olyan program, ami a PC-ből beszédbemenetű írógépet csinál – a tudományos-fantasztikus jóslatok csak azt nem látták előre, hogy ezek nem válnak közkeletűvé, hanem elsősorban a gépelni nem tudó csökkent mozgásképességűek számára jelentenek majd fontos segítséget. Ma már nem ritka, hogy az ilyesfajta 'voice command' rendszerek jobban értik a súlyosan torzult beszédű beteget közvetlen (emberi) környezeténél; nemcsak az ápolójánál, de még az édesanyjánál is.

Különösen fontos tudni Chomsky jelenlegi visszhangtalanságának megértéséhez, hogy ezek a számítógépes programok éppen azokon a Markov-modelleken (tehát a legegyszerűbb, hármas osztályba tartozó rendszereken) alapulnak, amelyektől Chomsky és George Miller (Miller – Chomsky, 1963, 419–491.) kivont karddal védték az elméleti nyelvészeket. A történet nem lenne teljes annak említése nélkül, hogy a mindehhez a statisztikai hátteret adó George Miller (a Princeton Egyetem nagyszerű pszichológusa, aki a klasszikus Zipf-törvényt Benoît Mandelbrotot megelőzve vezette le egy egyszerű 'majmok és írógépek' modellből), végül is nem ezzel, hanem egy tudományos szempontból ultrakonzervatívnak nevezhető elmélettel, az Arisztotelész eszméit a számítógépes szótárszerkesztésbe átültető WordNet rendszerrel vált a számítógépes munka egyik szellemi vezéralakjává.

Nem tudjuk teljesen elfogadni Kálmán László fentebb idézett megjegyzését, hogy a szabályalapú megközelítések *általában* is sikertelennek bizonyultak, hiszen maradt egy terület, a szótan (morfológia) ahol a mai számítógépes nyelvészetet domináló tanuló-algoritmusok még messze nem olyan sikeresek, mint a képzett fonológus/morfológus által kézzel írt szabályrendszerek. A helyzet külön érdekessége, hogy ezek a szabályrendszerek remekül együttműködnek a statisztikai alapú beszédfelismerő és szintetizáló-rendszerekkel, sőt azok ma még nélkülözhetetlen részei. De ez a fejlődés is lényegében a Chomsky által határozottan kijelölt iránnyal ellentétes vonalú volt: míg Chomsky és Morris Halle (1968) a környezetfüggő (egyes típusú) nyelvtanokat és a szekvenciális szabályalkalmazást szorgalmazták, addig C. Douglas Johnson, Kimmo Koskenniemi, Ronald M. Kaplan, Martin Kay, Lauri Karttunen, és társaik épp a véges automaták (hármas típusú rendszerek) hatékony technikai általánosításával, párhuzamos szabályalkalmazással értek el eredményeket.

A történet még távolról sem ért véget, jól látjuk ezt a gépi fordítás jelenlegi állapotán: e rendszerek jónak semmiképp sem nevezhető, de ma már használható eredményeket hoznak. Úgy gondoljuk, hogy itt is lassú, de feltartóztathatatlan minőségi javulás várható, s az áhított cél, a magas színvonalú, emberi beavatkozás nélküli szövegmegértés és -fordítás még Chomsky életében elérhető lesz.

Kulcsszavak: *Chomsky-hierarchia, formális nyelvek, nyelvtanok*

IRODALOM

Beesley, Kenneth – Karttunen, Lauri (2000): Finite-state Non-concatenative Morphotactics. In: Proceedings of the 5th SIGPHON Workshop. 1–12.

Chomsky, Noam (1956): *Three Models for the Description of Language*. I.R.E. Transactions on Information Theory IT-2.

Chomsky, Noam (1957): *Syntactic Structures*. Mouton, The Hague

Chomsky, Noam (1965): *Aspects of the Theory of Syntax*. MIT Press

Chomsky, Noam and Morris Halle (1968): *The Sound Pattern of English*. Harper and Row

Culy, Christopher (1985): The Complexity of the Vocabulary of Bambara. Linguistics and Philosophy. 345–351.

Huybregts, Rini (1976): *Overlapping Dependencies in Dutch*. Utrecht Working Papers in Linguistics 1. 24–65.

Joshi, Aravind (2003): Tree Adjoining Grammars. In: Mitkov, Ruslan (ed.): *Handbook of Computational Linguistics*. Oxford University Press, 483–500.

Karlsson, Fred (2007): Constraints on Multiple Center-embedding of Clauses. Journal of Linguistics. 43, 2, 365–392.

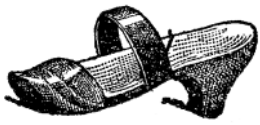
Miller, George – Chomsky, Noam (1963): Finitary Models of Language Users. In: Luce, Duncan – Bush, R. R. – Galanter, E. (eds.): *Handbook of Mathematical Psychology.* **II**. Wiley, New York, 419–491.

Pereira, Fernando (2000): Formal Grammar and Information Theory: Together Again? Philosophical Transactions of the Royal Society, series A. 358, 1239–1253.

Postal, Paul (1964): *Constituent Structure*. Mouton, The Hague

Pullum, Geoffrey – Gazdar, Gerald (1982): Natural Languages and Context Free Languages. Linguistics and Philosophy. 4, 471–504.

Selkirk, Elizabeth (1977): Some Remarks on Noun Phrase Structure. In: Culicover, Peter W. – Wasow, T. – Akmajian, A. (eds.): *Formal Syntax*. Academic Press

Shieber, Stuart (1985): Evidence Against the Context-Freeness of Natural Language. Linguistics and Philosophy. 8, 333–343.

Stefanowitsch, Anatol (2006): Negative Evidence and the Raw Frequency Fallacy. Corpus Linguistics and Linguistic Theory. 2, 1, 61–77