A nyelvtudomány műhelyéből

Megszólaltak a gépek

Nincs messzé az az idő, amikor, legalábbis írásban — m a g y a r u l is meg fognak szólalni a gépek (vö. Magyar Tudomány 1965. 710). Egyelőre azonban még nem erről van szó: mindössze csak arról, hogy a m a g y a r n y e l v r ő l kezdtek nekünk vallani az általunk alkalmazott okos elektromechanikus berendezések. Folyóiratunk hasábjain már beszámoltunk kísérleti mennyiségen nyert eredményeinkről (vö. Nyr. 88: 457—64). Most hírt adhatunk az első végleges eredményekről, melyek az ÉrtSz. önálló szócikkel rendelkező címszavaira (a továbbiakban: címszavak) vonatkoznak.

1. Lássunk előbb néhány összefoglaló adatot szókincsünk e lexikográfiai törzsanyagára vonatkozóan.

a) Mindenekelőtt: h á n y c í m s z ó van az ÉrtSz.-ban s így a mi lyukkártyarendszerű feldolgozásunkban? Az ÉrtSz. szerkesztői a 7. kötet végén (671) 58 023-ban állapítják meg az „önálló szócikkek", vagyis az önálló szócikkel rendelkező címszavak számát. Mi meg, miután pontosan kimásoltuk ezeket a címszavakat, s a gépekkel elölről-hátulról jól megszámoltattuk a kártyamennyiséget — 58 323-at kaptunk eredményül, vagyis pontosan 300-al többet. Kiderült az is (és ezen már nem tudunk változtatni), hogy feldolgozásunkból kimaradt a *színielőadás* címszó (megvan a *színelőadás)*, tehát eddigi ismereteink szerint 301 szóval többet dolgoztunk fel, mint amennyit a szerkesztők címszóként „bevallottak". Hol az igazság, hány szó van hát végül is megmagyarázva az ÉrtSz.-ban? Az igazság ezúttal valószínűleg valahol középen van: minden bizonnyal nincs több címszó forrásunkban 58 323-nál, de nagyon valószínű, hogy több van benne, mint 58 023. Ne tessék csodálkozni következő állításomon: n i n c s e m b e r, a k i p o n t o s a n m e g t u d n á m o n d a n i, h á n y c í m s z ó v a n a z É r t S z.-b a n. Ahány ember, és ahányszor megszámlálná, annyiféle, kissé eltérő eredményt kapna (pontosabban: egyes eredmények többször, mások kevesebbszer fordulnának elő; még azt is előre megmondhatjuk, hogy az eredményeknek ez a megoszlása egy matematikailag jól meghatározható görbe, a normális eloszlás vagy más néven a hibagörbe szerint alakulna). Az embernek mint automatának a belső struktúrája olyan, hogy lenullázható számlálóberendezésként nemcsak sebesség, hanem pontosság dolgában is jelentősen elmarad a mai hasonló rendeltetésű berendezések mögött.[1]

[1] Az ember lassabban fut, mint a strucc, aránylag kisebbet ugrik, mint a bolha, nem tud repülni, nem tud huzamos ideig víz alatt lenni. És így tovább. De m e g t u d j a s z e r v e z n i az őt környező természetet úgy, hogy mindezt mégis elérje, ő fusson a leggyorsabban, repüljön, ússzon stb. Tudomásul kell vennünk, hogy bizonyos szellemi (gondolati) adottságokkal is így van az ember: hiszen azok is anyagi adottságok, ha más mozgásformákhoz vannak is kötve. Egyes gépek sokkalta gyorsabban és pontosabban tudnak számlálni, bizonyos egyéb algebrikus műveleteket elvégezni, logikai műveleteket elvégezni, mint az ember. Aki ezt degradálónak tartja az emberre nézve, az próbálja meg puszta lábbal átugrani akár csak az Eiffel-tornyot.

Egészen pontosan szólva tehát minden további eredményünk már nem magára az ÉrtSz.-ra, hanem saját anyagunkra támaszkodik. Ez az anyag hajszálra megközelíti a forrást, de nem azonos vele; az ezen az anyagon gépi úton kapott eredmények egymáshoz képest igen pontosak (bár távolról sem abszolút pontosak, vö. **2.**) Itt már bármely mérést akárhányszor elvégezhetnénk, és mindig ugyanarra az eredményre jutnánk. (A gépi munkának ezt a pontosságát egyes esetekben kontrollcélokra külön is felhasználtuk, amint arról alább szó lesz.)

b) Hogyan alakul szótári szókincsünk hosszúság szerinti megoszlása? Kísérleti anyagunk alapján erről egy grafikont is rajzoltunk annak idején (460). Az egész anyag feldolgozása lényegében ugyanezt az eredményt hozta, csak még szabályosabban (az 5—6 betűs szavaknál észlelt „behorpadás" nélkül) alakul a görbe, mely egyébként ugyancsak az épp előbb említett normális eloszlásgörbe. Eszerint tehát legtöbb szavunk szótári alakjában 8 betűs (8659), illetőleg 9 betűs (8267), jelentős mennyiségű még 7 betűs (7705), illetőleg 10 betűs (6877) szavaink száma is. 7—10 betűs szavaink az egész anyagnak több mint a felét alkotják (31 508 egyeddel).[1]

Nem kell immár törnünk a fejünket azon, melyik a leghosszabb magyar szó az ÉrtSz. címszavai között. Ha az *n*-eket vesszük, a leghosszabb szó: *kereszténységszocializmus* (22 n), csak eggyel marad el mögötte a *kereszténységszocialista* (21 n) és még további 15 21 *n*-es szavunk. Ám, ha figyelembe vesszük, hogy egyes *n*-ek együttesen fejeznek ki egy fonémát (a szokásos kifejezésekkel élve: ha a két- és háromjegyű betűket egy-egy betűnek számítjuk), akkor kiderül, hogy leghosszabb egybeírt szavunk a *lelkiismeretvizsgálat* és a *rabszolgakereskedelem* 20—20 fonémával (betűvel a szó hagyományos értelmében): a *keresztényczocializmus* ezzel szemben mindössze 19 fonémát tartalmaz. A korábban említett *megfellebezhetetlenség* (23 betű) nem szerepelhet anyagunkban, mert csak származékként van feltüntetve. Most már azonban elég könnyűszerrel tudunk nála hosszabb szavakat is találni: *összeegyeztethetetlenség* (24 *n*), *összehasonlíthatatlanság* (24 *n*) — egyszerűen leghosszabb, még szereplő szavaink egyes származékait képezzük. Különféle célokra egyébként nyilván hol az *n*-ekben, hol a betűkben (fonémákban) mért hosszúság fog majd kelleni; ezért természetesen továbbra sem ejtettük el azt a tervünket, hogy most lyukkártyákon levő anyagunkat elektronikus gép memóriájába konvertáljuk, és ott megfelelő program alapján immár nem *n*-eket, hanem betűket (fonémákat) számoltatunk.

Mindez az egybeírt szavakra vonatkozott. Különírt vagy kötőjellel írt címszavunk van hosszabb is. Így anyagunk leghosszabb egysége tulajdonképpen három szó, egy kötőjellel és egy különírással: *illegeti-billegeti magát* (a kötőjelet meg a spáciumot is számítva, amint mi tesszük, ez 24 *n*), rögtön utána következik a *magánhangzó-illeszkedés* 23 *n*-nel.

A szóköz meg a kötőjel betűként való értelmezése egyébként felvet már néhány nem statisztikai jellegű kérdést is. Így mindjárt azt, hogy egyáltalán lehet-e ezeket a jeleket betűként értelmeznünk. Egyes nyelvészek számára egészen természetes, hogy ezek a „valamik" is betűk (így vélekedik például Országh László). Mások — mint Kelemen József — határozottan tiltakoznak ez ellen. Mi, amint ezt már többször szóban és írásban kifejtettük, és amint ezt most a lyukkártya-rendszerű feldolgozásban érvényre is juttattuk, természetesen az előbbi nézet mellett vagyunk. A nyomdász, a híradástechnikus, az elektronikus gépen dolgozó szakember, amikor a magyar nyelv írott változatával dolgozik, bizonyos számítások, tervezések elvégzéséhez egyáltalán nem fog az iránt érdeklődni, milyen jelnek milyen a hangértéke, van-e egyáltalán hangértéke stb. Amikor ezek a szakemberek azt kérdezik tőlünk: hány betűből állhat a leghosszabb magyar szó, milyen a magyar szavak hosszúság szerinti megoszlása stb. — akkor éppen az érdekli őket, mennyi

[1] Az itt közölt és további számadatokból a végső ellenőrzések után egy ezrelékesnél nem nagyobb eltérések még lehetségesek.

helyet hagyjanak fel egy-egy szónak, és egyáltalán nem érdekli őket, milyen hangértékű jelekkel vannak ezek a helyek betöltve. Természetes módon és nem véletlenül erre az álláspontra szorított bennünket az általunk alkalmazott lyukkártyatechnika is. A kártya nem gumiból van, nem nyújtható: mindig 80 pozíció (oszlop) van rajta. E 80 oszlopból valamennyit a szöveges résznek (az egyes címszavaknak) kell fenntartanunk, és az édes mindegy, milyen jelek vannak éppen a betűk számára fenntartott oszlopokon — esetleg nulla betű, esetleg kötőjel. Ha lemarad egy kötőjel, vagy egybeíródik egy szó különírás helyett — az bizony éppen olyan nagy baj, mint ha — mondjuk — *ly* helyett *j*-t írtunk volna.

Egy másik kérdés ezzel kapcsolatosan a sorrendbe állításé. Hiszen, ha az AkH.-tól eltérően két új betűt bevezettünk, annak helyet is kell találni. Ezt a kérdést úgy oldottuk meg, hogy a nulla betű (szóköz, spácium) az ábécé legelső betűje, utána következik a kötőjel betű, aztán az *a, á*... stb. (Megjegyzendő, a gépen nem tudtuk megoldani, hogy szóközben kötőjelet írjon: így a kötőjel a spáciumtól csak szó végén — pl.: *leg-, legesleg-* —, illetőleg szó elején — pl.: *-e, -szerű, -féle* stb. — különbözik.) Az így kapott sorrend nemegyszer megegyezik a hagyományos lexikográfiai sorrenddel, nemegyszer különbözik tőle. Példa az eltérésre: a *ki-be* a *kiabál* előtt van (mert a *ki* kezdetet benne az *a*-t megelőző nulla betű, illetőleg kötőjel követi). — Itt jegyzem meg, hogy az AkH. 17. pontjától és a lexikográfiai gyakorlattól eltérően számunkra nem léteznek „egyenlő értékű betűpárok": mind szókezdő helyzetben, mind szó közben előbb jönnek az *a*-t tartalmazó szavak valamennyien, aztán az *-á*-t tartalmazók; előbb jönnek az *-i*-sek, aztán az *í*-sek stb. A *bank, barát* szavak tehát előbb állnak, mint a *bán, bánat* stb. szavak, mert az előbbiek kezdete *ba-*, az utóbbiaké — *bá-*. Ugyanez áll az „a-tergo"-sorrend esetén is, csak természetesen megfordítva.

Egy, véleményem szerint igen fontos elvi kérdésre szeretnék kitérni az itt tárgyalt apróságokkal kapcsolatosan: a nyelvi leírások sokféleségének és ebből fakadó viszonylagosságának kérdésére. (Sokkal lényegesebb probléma kapcsán ezt nemrég Melcsuk is érintette nálunk, vö. MNy. 1965. 3. 268.) Hogy mit veszünk betűnek és mit nem — ízlés, módszer kérdése. Tekintettel arra, hogy a nyelvészetben nem dolgozunk ilyenkor deduktív módszerekkel, e l v i l e g b e b i z o n y í t h a t a t l a n, hogy „kinek van igaza", hogy melyik leírás a helyes. Ugyanígy állunk a sorrendbe állítás kérdésével is: ízlés kérdése, hogy legyenek-e „egyenrangú betűpárok" vagy ne, és ha vannak, melyek legyenek azok. Csak két dolog lényeges: *a)* az elfogadott módszer meg kell hogy feleljen a szem előtt tartott alkalmazásoknak; *b)* az elfogadott módszert következetesen kell alkalmazni. Nagyon is ésszerű például, hogy az egyenrangú betűpárokat alkalmazzuk, vagy hogy a kötőjeles szavakat úgy tegyük sorrendbe, mintha kötőjel bennük nem volna, ha az e m b e r t tartjuk szem előtt felhasználóként. Mégpedig nem is a nyelvészt, hanem a közönséges embert, aki éppen azért fordul a szótárhoz, mert helyesírásilag ingadozik: nem tudja, hosszú vagy rövid *ü*-t kell-e írnia, egybe vagy külön kell-e valamit írnia. A szót kereső fejében tehát a probléma cseppfolyós vagy ömlesztett állapotban van, alkalmasint sokszor meg sem találná a szót, ha az általa feltételezett alak alatt keresné, és ott nem találná. Az említett gépi alkalmazások (melyeknek nem véletlenül megfelel a helyesírásunkat, sőt annak lehetséges buktatóit jól ismerő nyelvész álláspontja) viszont bizonyos szigorúbb rendet követelnek, ott meg éppen az anyag ömlesztett tálalása okozna nehézségeket. A gépnek éppúgy nincsenek helyesírási problémái, mint a nyelvésznek, illetőleg, amennyiben ilyenek felmerülnek, éppúgy oldja meg őket a maga számára, mint a nyelvész. Ha, teszem azt, a **papir* szót nem találta így meg, az *-ír* végűek között, ebből rögtön teljes biztonsággal veszi azt az információt, hogy e szót hosszú *-í*-vel kell írni. A mi feldolgozásunk pedig elsősorban éppen nyelvészemberek és nyelvészgépek számára készült.

c) De térjünk vissza anyagunkhoz: hogyan oszlanak meg benne a szavak j e l e n-

t é s ü k s z á m a szerint. Amikor a jelentésszámot mint mutatót felvettük, egyesek éppen az ÉrtSz. szerkesztői közül figyelmeztettek bennünket, hogy ezt ők esetenként igen szubjektíven határozták meg. Valóban eléggé önkényesnek tetszhet, még ha jól kidolgozott alapelvekkel rendelkezünk is, mit vegyünk egy jelentésnek és mit kettőnek, mit vegyünk egy jelentés aljelentéseinek, és mit számozzunk már külön. (Feldolgozásunk során, mint ismeretes, az arab számmal számozott jelentéseket vettük egy-egy jelentésnek, egyszerű esetben tehát a jelentésszám annyi volt, amekkora a legnagyobb sorszám a jelentések felsorolásában.) Mi mégis bíztunk az ÉrtSz. összeállítóinak egyöntetű eljárásmódjában (és nem elhanyagolható módon: a nagy számok törvényében) — ezért bevettük ezt a mutatót is. Nem csalatkoztunk, sőt eredményünk messze meghaladja várakozásunkat. A legtöbb szó — az egésznek több, mint a fele — egyjelentésű, ennél kevesebb a kétjelentésű, ennél is kevesebb a háromjelentésű — és így tovább az *is* szóig, melynek 101 jelentését (!) sorolják fel az ÉrtSz.-ban. A várakozást meghaladta itt az a törvényszerűség, amely szerint a több s több jelentésű szavak száma csökken. Ez a törvényszerűség, pillanatnyilag úgy tetszik, leírható egy igen egyszerű exponenciális függvénnyel. A függvényben fontos szerepet játszik az ö s s z e s s z ó m e n n y i s é g (vagyis a közel 60 000 egyed). Joggal feltételezhető, hogy a tényleges megoszlás azért közelíti meg a vártnál sokkal jobban az említett matematikai formulát (vagy megfordítva), mert az összeállítók valóban egységesen jártak el óriási munkájuk során. Görbénk szabályossága tehát valószínűleg nem kis mértékben a karmester — Országh László — munkáját dicséri, meg munkatársainak fegyelmezettségét.

Többet erről most nem kívánok mondani. Egyrészt: a végleges matematikai értékelés még folyamatban van. Másrészt: egy évvel ezelőtt örömmel közöltük olvasóinkkal, hogy akkora anyag fog rendelkezésünkre állni nyelvstatisztikai célokból, amekkorával maga Zipf sem rendelkezett ismert törvényeinek megalkotásakor (vö. i. h. 459). Akkor még nem éreztem e körülmény rossz oldalát. Nevezetesen azt, hogy eredményünkkel egészen egyedül állunk, nem tudjuk azt pillanatnyilag mihez hasonlítani; így például nem tudjuk, a Websterben, vagy az Usakovban milyen törvényszerűségek uralkodnak, mi függ hát a nyelvtől és mi az emberektől. De egy bizonyos: szigorú törvényszerűségek uralkodnak nem csupán a szövegekben, hanem az olyan mesterséges alkotmányokban is, mint amilyen egy szótár. Még olyan szubjektívnek felfogható kérdés tekintetében is, mint amilyen a jelentések száma. Az épp előbb említett körülménynél fogva egyelőre nem látjuk tisztán e megállapításunk jelentőségét. Lehet, hogy egészen vagy majdnem egészen triviálisnak fog bizonyulni — bizonyos matematikai meggondolások alapján. Akkor érdemünk csupán e triviális igazság kísérleti igazolása (ez nem valami nagy érdem a matematikában).

d) Néhány hónappal ezelőtt Honvéd Katalin (Mátraalmás) azt kérdezte Lőrincze Lajostól: van-e az *ország*-on és a *jószág*-on kívül más *-ság*-ra végződő magyar szó. Ha kissé késve is, de immár teljes biztonsággal válaszolhatunk: nincs. (Tájszavaink, a nyelvtörténet során kihalt szavaink körében esetleg még bukkanhatunk ilyen elemre, de az is kevéssé valószínű.) Az olvasó e válaszból láthatja, hogy elkészült, egyebek között, a teljes a-tergo (szóvégmutató) lista is, s így már elég pontos képünk van szavaink szótári alakjának végződés szerinti megoszlásáról. A leggyakoribb végződések: *-t* (7150), *-s* (6730), *-l* (5005), *-a* (4196); az e betűk valamelyikére végződő szavak alkotják összesen az anyag 39,6 százalékát. (Megjegyzendő, a kísérleti mennyiségtől eltérően, itt már a kétjegyű betűket a végződésekben külön számítottuk, tehát a *-s* végűek nem tartalmazzák sem a *-cs*, sem a *-zs* végűeket. A hagyományos nyelvész számára ez egészen magától értetődőnek tetszik; számunkra nem az, és ismét csak bizonyos célszerűségi-alkalmazási meggondolások késztettek bennünket erre a döntésre.) Már a kísérleti mennyiségben is meglepően sok volt a *-ly*-végű: azt hittük, csak a kis anyagban véletlenül jöttek így össze

a szavak. Azonban a teljes anyag is a vártnál talán több, 435 ilyen szót tartalmaz. (Ez a lista az egyszerű szóvégmutató lista, tehát ha valamely *-ly*-ra végződő szó összetétel utótagjaként többször is előfordul, akkor többször is számításba jön. Abból a rendezésből, melynek fő szempontja az összetettség — nem összetettség lesz és ezen belül az a-tergo sorrend, maguktól fognak adódni az *-ly*-ra végződő tőszavak. A 435 nem túlságosan nagy szám, abból persze még kézzel is könnyűszerrel ki lehetne válogatni a kérdéses tőszavakat. De minek, amikor úgyis meglesz, külön listán, magától is — azaz nem egészen magától, hanem a géptől.)[1]

e) A szótári szókincs szófajonkénti megoszlása a legfontosabbak tekintetében így alakult: főnév 30 574 (tehát az egész anyagnak több mint a fele), ige 14 269, melléknév-főnév: 604, főnév-melléknév: 3703, melléknév: 5902.

A főnév-melléknevek és a melléknév-főnevek egyébként felvetik a kérdést: mi van azokkal a címszavakkal, amelyek nem homonimák, de különféle szófajokként szerepelhetnek a mondatban? Az ilyen több szófajú szavak legtömegesebb képviselői az említett főnév-melléknevek és melléknév-főnevek. Úgy találtuk, összesen 34 szófaji osztály állítható fel úgy, hogy egy-egy osztályba csak azonos szófaj-kombinációkban és azonos fontossági sorrenddel fellépő szavak legyenek. Például: a főnév egy osztály, a melléknév egy másik osztály, a főnév-melléknév egy harmadik, a melléknév-főnév egy negyedik és így tovább. (A két utóbbi osztály a fontossági sorrend tekintetében tér el egymástól.) Íme, néhány ritkább, bonyolultabb összetételű osztály és képviselői: határozószó-igekötő-mondatszó-főnév kettő van: *újra, vissza*, egy-egy szó képviseli a következő osztályokat: számnév-melléknév-főnév-névmás: *egy*[1], melléknév-főnév-határozószó-mondatszó: *jelen*, határozószó-főnév-melléknév-mondatszó: *kontra*, határozószó-igekötő-melléknév-főnév: *teli*, határozószó-igekötő-névutó-melléknév: *szerte*, határozószó-igekötő-névutó-főnév: *túl*, határozószó-igekötő-kötőszó-mondatszó: *viszont*. Több olyan osztály nincs is, ahol négy egyszerű szófaj kombinálódhat egyetlen címszó mellett.

Az igék fent megadott számához hozzá kell még vennünk néhány száz hiányos igét, melyek rendszerünkben a hiányzó alakoknak megfelelően más-más igei alosztályba kerültek (és ezzel egyidejűleg elkerültek a teljes paradigmájú igék mellől).

Szótárunk szófaji megoszlását szerencsére már tudjuk mihez hasonlítani. Josselson professzor, ha egyelőre kézirat gyanánt is csupán, egyik New York-i előadásában említi egy, a mi ÉrtSz.-unkhoz hasonló orosz értelmező szótár anyagának szófaji megoszlását. Eszerint 81 523 címszóból ott 28 114 (34,7%) az ige és 34 435 (42,2%) a főnév, 15 923 a melléknév. Az oroszban alig van melléknév-főnév, illetőleg főnév-melléknév minősítésű szó (és általában: az olyan ragozó nyelvekben, mint az orosz, a szó rendszerint csak egy bizonyos szófajhoz tartozik egyszer s mindenkorra; végződése lehetetlenné teszi számára, hogy más szófaj képviselőjeként is fellépjen a mondatban). De ami ennél is érdekesebb: az orosz szótár, mint látjuk, viszonylag több igét tartalmaz, mint hasonló magyar társa. Ez nem lehet véletlen: az orosz folyamatos és befejezett igék általában külön-külön címszóként szerepelnek, nyilván ez növeli meg e szófaj szótári arányát. Vagyis: újból, egészen más oldalról azt kellett megállapítanunk, hogy a szótár, minden mesterkéltsége ellenére, bizonyos objektív, az adott nyelv jellegétől is függő törvényszerűségeknek van alávetve. Nem csupán az író, költő stb. kezét kötik gúzsba

[1] Azon a napon, amikor meghoztam Pestről a teljes a-tergo listát, ötödikes leányom házi feladata az volt, hogy írjon minél több *-ly*-ra végződő szót. Megragadtam az írógépet és kiválogattam neki az összes tőszót a listáról. Reggel leányom, aki különben is elég szigorú hozzám, minden meghatottság nélkül végignézte a listát, és azt mondta: „Kihagytad azt, hogy *gerely*." Még egy tanulság: az embert mechanikus válogató munkára sem szabad befogni, mert eközben is gondolkodik, és ezzel elrontja a dolgot: a *gerely* szót véletlenül valahogy valóban kihagytam, pedig listánkon persze rajta van.

a statisztika szigorú szabályai alkotása során. A látszólag náluk önkényesebben dolgozó („tetszése szerinti szavakat kiválogató") nyelvészét is.

Az szintén világos előttünk (részint ismét Josselson, részint Steinfeldt gyakorisági szótárából), hogy a szövegekben szófaji megoszlás szerint más törvényszerűségek uralkodnak, mint a mesterségesen összeállított szótárban. Hogy csak a legszembeötlőbbre utaljunk: mindössze néhány száz egyedet kitevő névmások rendkívüli gyakorisággal fordulnak elő a szövegekben, a szövegeknek nem csupán minden századik vagy ezredik szava névmás (körülbelül ez az arány a szótárban), hanem minden tizedik-tizenötödik az, legalábbis az oroszban! (Vö. NyK. 1965: 166.)

f) Egy francia kolléga keresztrejtvényfejtők számára kiadta a francia nyelv lyukkártya-rendszerrel készült olyan szótárát, amely néhány tízezer francia szót hosszúsága és ezen belül ábécérend szerint tartalmaz. Nálunk is készült ilyen lista, ábécérendben is, meg a-tergó rendben is. E listák alapján rögtön lehet válaszolni az olyan kérdésekre, mint „Mi lehet az ... Hat betűből áll és annyi van meg belőle, hogy *ál-*?" vagy: „Mi lehet az ... Tizenhárom betűből áll, és a végén megvan annyi, hogy — *-ba..*" — legtöbbször még az illető sor kérdését sem kell tudni, adott hosszúság mellett két-három azonos kezdő vagy végző betűvel nem túlságosan sok szavunk van. Hogy pontosan mennyi, arra is külön kimutatás készült. A továbbiakban tehát egész pontosan tudni fogjuk, mi a valószínűsége annak, hogy egy magyar ember *a*-t mond (ti. *a*-val kezdődő szót mond), és, ha *a*-t mondott, akkor mi a valószínűsége annak, hogy *b*-t is mond (vagyis: hogy *ab-* kezdetű a szó). Sőt, e közmondásnál mélyebben is bontunk, a harmadik betűig. Visszafelé pedig, jobbról számítva, a negyedik betűig. Vagyis meg tudjuk határozni, mi a valószínűsége annak — fenti példánknál maradva —, hogy valamely szavunk *-a*-ra végződjék; ha így végződött, mi a valószínűsége annak, hogy ez előtt a jel előtt, mondjuk *-b-* állt; s ha ily módon a szóvég *-ba* volt, mi a valószínűsége annak, hogy ezt megelőzően, mondjuk, *-r-* álljon *(-rba)*, ezt megelőzően meg *-u- (-urba:* pl. *girbe-gurba)*. Ezeket az eredményeket részben már az elektromechanikus gépeken is rendezni fogjuk, részben végleges feldolgozásuk ismét csak az elektronikus gépektől várható e lyukkártyák alapján. Természetesen nem a keresztrejtvényfejtők és -készítők számára végezzük ezeket a kutatásokat, hanem abból a célból, hogy jobban megismerjük szavaink morfonológiai felépítését, azokat az átmeneti valószínűségeket, amelyek az egyes fonémákat szavainkon belül egymáshoz fűzik szókezdő és szóvégző helyzetben. Melcsuk javaslatára lyukkártyákra előkészítettük a magyar főnévi és igei toldalékokat is; a toldalékok hasonló adatainak birtokában immár nem csupán a szótári alakok, hanem általában a magyar szóalakok vonatkozásában ismerni fogjuk ezeket a törvényszerűségeket.

A keresztrejtvényfejtők kézikönyvétől az említett átmeneti valószínűségek számításáig még sok mindenre jó a különféle szempontú rendezés — hogy ne szóljak most többről közülük. Már csak azért se, mert e kérdések egészének egy nagyobb, összefoglaló munkát szeretnék szentelni.

2. Quandoque dormitat et bona machina — néha a jó számítógép is aluszik. Tévedés azt hinni, hogy a gépek abszolút hiba nélkül dolgoznak (mint én magam is hittem korábban). Csak lényegesen kevesebbet hibáznak, mint az ember. És persze — másféleképpen. Lássunk ezekből néhányat.

Az „atergósítás" az elektromechanikus gépeken a következő lépésekben megy végbe: *a)* a gép megszámolja, hány betűből áll az adott szó és ezt az eredményt — például: **1** betű, **23** betű stb. — beüti a megfelelő oszlopba (nálunk ez a 48—49-es oszloppár volt: két oszlop, mert van 9-nél, illetőleg 10-nél több *n*-ből álló szavunk is); *b)* a gép a 48—49-es oszlop alapján hosszúságuk szerint csoportosítja a szavakat: 1 betűs szavak, 2 betűs szavak .... 24 betűs szavak; *c)* a gépkezelő az 1 betűs szavakat kiveszi

a dobozukból és a másológépbe teszi azzal az utasítással, hogy minden egyes kártya szövegtartalma másoltassék át a 80. oszlopba; előveszi a 2 betűs szavakat és a másológépbe teszi azzal az utasítással, hogy a kártyák szövegtartalma másoltassék át a 79. oszloptól kezdődően, a 3 betűsöké — a 78. oszloptól kezdődően . . . a 24 betűsöké az 57. oszloptól kezdődően; *d)* a másológépen ennek megfelelően megtörténik az átmásolás, melynek eredményeképpen minden egyes szó a 80. oszlopban v é g z ő d i k; *e)* egy újabb gépen megkezdik az ábécé szerinti rendezést a 80. oszlopon; *f)* az így csoportosított kártyák tartalmát egy további gép kiírja rendes papírra: megvan az a-tergo lista. Mi történik, ha egy porszem becsúszik a mechanizmus legelső láncszemébe, és a gép — mondjuk — eggyel kevesebb betűt számol le a szóban (l. *a)* pont)? Az történik, hogy a *terefere* szó 49-es oszlopába a helyes 8 helyett a 7-es szám kerül. Tovább már minden jól működhet, mégis kész a baj: a *terefere* szó a *b)* eredményeképpen a hétbetűsök közé kerül, a *c)—d)* műveletel az átmásolás a 74. oszlopban kezdődik meg és természetesen megáll a *terefer* alaknál (az utolsó *-e* már a 81. oszlopba kerülne, de ilyen oszlop nincs). Világos, hogy az *f)* értelmében e szó végül is az *-r*-re végződőek közé kerül és a listán, más *-r* (sőt: *-er*, *-fer* stb.) végűek között egyszerre csak elénkpattan: TEREFER. Mindez egy porszemnyi hiba miatt az *a)*-ban. Az egész anyagban mindössze néhány tucat olyan szó volt, melynek ilyen módon lenyeződött az utolsó betűje, mely ezért nem a maga helyére került az ábécébe stb. Így jártunk még többek között a *meghala(d)*, *meghasa(d)*, *megalva(d)*, *megolva(d)*, *décbund(a)* stb. szavakkal.

Vannak természetesen másfajta hibák is. Az a-tergo lista végén találtuk a MUFUR szót, melyről azt írta számunkra a gép: 6 betűs. Nem volt nehéz kitalálni, hogy a *mufurc* szóról van szó — hogy hova lett a végső *-c*-je, jóllehet beszámoltatott, nem tudjuk; hogy miért került az ábécé végére — azt sem, éppúgy, mint ahogy például az is titok marad előttünk, miért kapott a *vetkezik* szó első *e*-je az átmásolás során egy ékezetet s lett így: VÉTKEZIK, besorolva a rendes *vétkezik* szó után. (Itt segített a gép abszolút pontossága számlálás tekintetében: már kihúztuk volna a második *vétkezik*-et, amikor kiderült, hogy ezt nem tehetjük, mert akkor egy másik szót valahol elvesztettünk, végső soron ugyanis egészen biztosan 58 323 egyeddel van dolgunk és nem 58 322-vel. Kelemen József, aki, úgy látszik, kezdi „betéve" tudni az ÉrtSz. egész címszó-anyagát, tőlünk függetlenül rögtön rájött arra, hogy néhány lappal feljebb a VETKEZIK hiányzik s hogy eszerint az egyik VÉTKEZIK nyilván = VETKEZIK.) Rá kellett jönnünk arra, hogy a MRLETNYITÁS helyesen ÜZLETNYITÁS, hogy a HALNDZS=HALANDZSA stb.

Egy, véleményem szerint bravúros nyomozást Jakab László végzett számunkra. Az a-tergo lista végén, még a MUFUR után ott állt ennyi: PLUSO. Amennyiben feltételeztük, hogy a többi (szám) kód már jó, akkor annyit mondhattunk erről a szóról e kódokból könnyen kiolvashatóan, hogy nem összetett, melléknév-főnév, 5 jelentése van, stiláris minősítése a fejben nincs, töve a ragozás során nem változik, ha főnévként ragozzuk, a következő ragokat kapja: *-t*, *-ok*, *-a*. Nincs benne a SzófSz.-ban, nincs rajta képző. Bár a MRLETNYITÁS kellő óvatosságra intett bennünket a szókezdő betűk helyességének vélelmezése tekintetében, Jakab László abból indult ki: ha Bárcziban nincs benne, akkor ezzel összhangban állhat az, hogy *pl*-kezdetű; ha valóban mély hangrendű, akkor miért ne lenne az *-u-* is jó benne? Keressük tehát az ÉrtSz. *plu-* kezdetű szavai között, a a többi jellemző alapján. Itt már nem volt nehéz dolgunk: PLUSO = *plusz*. Ellenőriztük a listán: e szó valóban hiányzott a maga helyéről, tehát valóban róla volt szó. Ezzel az azonosítás (javítás) befejeződött. — Más esetben az jött a segítségünkre, hogy csak az *f)* stádiumban, a kiírásában torzult el a szó, mely ilyenkor bekerül pontos betűrendi helyére, ha tehát nem a rosszul kinyomtatott betűt tekintjük, hanem azt, hogy milyen szavak vannak előtte és utána — a szót el tudjuk olvasni helyesen.

Előfordultak (különösen a rendes ábécélistán) sorrendi hibák is; ezek vagy ezek

nagy része feltehetően az egyébként rendkívül pontosan és jól dolgozó gépkezelők ember voltából, nem pedig közvetlenül a gépekből fakadnak.

Ne akarjunk technikusok lenni, csak a hibák elvi okára mutassunk rá. Az igaz, hogy a gépek sokkal pontosabbak és gyorsabbak, mint az ember. De sokkal több műveletet is kell elvégezniük ugyanazon munka folyamán. Így például egy-egy betű leolvasása kétszeres tapogatást jelent a megfelelő gép számára — míg tehát az átlag 9 betű hosszúságú 60 000 szót leolvassa a gép, $2 \times 9 \times 60\,000 =$ több mint egy millió tapogatást kell elvégeznie! És ez csak a leolvasás volt: ugyanennyi tapogatás esik a rendezés során, ugyanennyi a kiíráskor. Amíg tehát egy lyukkártya tartalma valamilyen csoportosításban a kártyától eljutott a tábláig (a kinyomtatott listáig), a berendezés különféle részeinek legalább 3 milliószor kellett egy-egy megadott helyen vezetést (illetőleg nem-vezetést) létrehoznia. Próbáljunk meg egy rendes villanykapcsolót hárommilliószor ki-be kapcsolni: ki sem bírja a rugója, pedig annak nem kell olyan pontosan és finoman működnie, mint a lyukkártya-berendezés érintkezőinek.

3. Befejezésül arról akartam írni, hol koppintott a gép a mi körmünkre: milyen, általunk elkövetett hibákat „hozott ki". Így például egyik munkatársam — még csak végzi a magyar szakot, nem tudhat éppenséggel mindent — az *értékpapír* szót akarta bizonyos szempontból korrigálni, és egész természetességgel a rövid *-ir* végűek között kereste. Ott azonban csak a *kartonpapír*-t találta — gyorsan bedugtuk ezt az egy szót a többi *-papír* utótagú közé, fátylat borítva arra, hogy a mi gépírónk, a lyukasztók (vagy esetleg mégis a gép, vö. *vetkezik — vétkezik..*) követték-e el ezt a hibát. Egyes kontrollokat éppen úgy szerkesztettem meg, hogy bizonyos gyakori helyesírási hibákat kihozzanak: a lyukasztók becsületére legyen mondva, csak egy-egy esetben tudtam rajtakapni őket azon, hogy szó végén rövid *-ü*-t írtak hosszú *ű* helyett, s hogy *-it* képzőt írtak *-ít* helyett! A KSH Számítástechnikai Igazgatósága, azt hiszem, joggal lehet büszke lyukasztói magyar helyesírási készségére (estleg magyar nyelvpótlékot is kellene adni nekik, ha ilyen volna).

Végül csak ennyit (Gáldi Lászlónak, aki efelelől élőszóban nyugtalanságát fejezte ki, és nyilván másoknak is): További munkánk e g y i k (bár tán nem legfontosabb) ága az lesz, hogy feldolgozzuk az ÉrtSz. e g é s z anyagát, tehát a származékokat is. De azt már teljesen automatikusan, elektronikus gépeken.

Papp Ferenc