„MAGYAR NYELVJÁRÁSOK"
A KOSSUTH LAJOS TUDOMÁNYEGYETEM
MAGYAR NYELVTUDOMÁNYI INTÉZETÉNEK
ÉVKÖNYVE

XVII, 147—158

DEBRECEN
1971.

A magyar nyelvre irányuló számítógépes vizsgálatok a Kossuth Lajos Tudományegyetemen

Egyetemünkön évek óta folynak részint elektromechanikus, részint elektronikus gépeken a mai magyar nyelv rendszerére irányuló vizsgálatok. Ezekről e sorok írója, közvetlen s távolabbi munkatársai nem egyszer s nem egy helyen írtak már,[1] szükségét érezzük azonban annak, hogy most bizonyos összképet nyújtsunk az eddigi eredményekről, a további távlatokról.

E témának ebben a kötetben való érintése egyik oldaláról nem vált ki semmiféle kérdést: ezek a kutatások mind a magyar nyelvre vonatkoznak. Utalni szeretnék azonban arra, hogy miért érzem kapcsolatosnak a dolog másik, módszertani részét is Debrecennel s jubilánsunkkal. Idekerülvén, Debrecent egyáltalán nem éreztük a maradandóság városának (vagy legalábbis nem a maradiságénak) — éppen az olyan örökifjú, örökké nyugtalan szellemek miatt, mint amilyennek Papp Istvánt is írásaiban, szavaiban megismerhettük. A ma nyugtalanságának, a ma újságának képviselői mindenképpen azok a modern gépek, amelyeket a kutatás szolgálatára állíthatunk. De sajátosabban is: Papp István munkásságában a grammatikai rendszerlátást értékeljük többek között nagyra — s vajon mi segíthet ma jobban az elaprózott jelenségek rendszerben való láttatásának, mint a szorgos s pontos gépek? Ahhoz természetesen, hogy valaki a magyar hangtan egészét, a szóalkotás problémáit — általában a magyar nyelv szerkezetét együtt lássa: ember kell, méghozzá tudós a talpán. Ám ma már e munkája során nem csupán írógép és cédulák állnak a kutató elme segítségére.

1. Ezt távolról sem csak az alkalmi bevezetés okán kellett elmondanom, hanem azért, hogy még a tulajdonképpeni tárgyalás előtt kiemeljem gépi feldolgozásunk leglényegesebb célját, legsajátosabb vonását. Azzal, hogy kezdetben az ÉrtSz. minden egyes önálló szócikkel rendelkező címszavát, összesen tehát mintegy 60 000 magyar szót külön-külön lyukkártyára vittünk úgy, hogy az egyes szavak mellett számos lexikográfiai ismérvet is feltüntettünk, lehetővé tettük, hogy anyagunkat mindeme szempontok szerint külön-külön, valamint

[1] E munkák legteljesebb bibliográfiáját l.: *A magyar szókincs lexikográfiai törzsanyagának lyukkártya rendszerű feldolgozása.* Debrecen, 1968. (Kézirat.)

mindeme szempontok tetszőleges kombinációiban, tételesen felsorolva vagy csupán statisztikai összesítésekben — együtt láthassuk.

Tekintsük itt át némely közvetlen eredményünket — abban a sorrendben, ahogy a szokásos szótárak az egyes információkat közölni szokták.

A legelső információcsoport természetesen — maga a címszó, szokásos helyesírási alakjában leírva. Még címszavakban felsorolni is nehéz, mennyi mindent tud csupán ebből kihozni a mai technika — olyat, amire a lexikográfusnak, a grammatikusnak szüksége lehet. Íme, csupán ízelítőül: *a)* A gépek könyörtelen logikája rákényszerített bennünket még az előkészítő munka során arra, hogy az eddiginél pontosabban megállapítsuk a magyar ábécé betűit — ha betűn minden olyan jelelemet értünk, amely címszóban előfordulhat (hiányoztak: *ë, à,* "-" [kötőjel], " " [spácium] — ezeknek természetesen egyszersmind meg kellett találni a megfelelő betűrendi helyét is); *b)* A címszavakat ezek után tetszőleges sorrendbe rakathattuk — e számos lehetséges sorrend egyike lett a VégSz. alapja: ha csupán a szótesteket nézzük is, ebben a sorrendben önként kínálja magát bizonyos grammatikai, szóképzéstani stb. kutatásokhoz; *c)* A többet tévedő s ilyen munkán rendkívül unatkozó, fáradó ember helyett ezek után a gépek számolták meg címszavaink *n*-ekben mért hosszát; számolták ki, hány 1, 2, 3 stb. *n*-es szó volt anyagunkban, melyek ott a leghosszabb elemek, milyen az átlagos szóhossz (8—9 *n*, vö.: VégSz. 589); *d)* Pontos statisztikákat kaptunk arra vonatkozóan, milyen anyagunk címszói szóalakvégi fonémák, digrammák, trigrammák, tetragrammák szerinti eloszlása (vö. VégSz. Függelék 1, 541—588). Fájdalmas primitívségnek érezheti itt a magyar nyelvész azt, hogy a gép *n*-ekben és nem betűkben gondolkodott — e fokon, elektronika nélkül, ezen még nem tudtunk segíteni; a későbbiekben azonban természetesen megoldottuk ezt a kérdést (vö. alább).

A címszót követi a szótárban a homoníma-index. Kiderült, hogy teljes anyagunk mintegy 1650 homonímát és álhomonímát tartalmazott. Ezek nagyobb része természetesen homoníma-pár, ám itt látjuk összegyűjtve a hármas, négyes, ötös csoportokat (ez utóbbiak: *a, c, költés, ss, század, tus*), sőt volt egy hatos homonímabokor is (*haj*: *haj*[1] fn 'szőrszerű képződmény', *haj*[2] fn 'héj' [ez csak utalószó], *haj*[3] fn 'padlás', *haj*[4] isz, *haj*[5] msz 'terelésre használt szó', *haj*[6] msz 'hé!'). Együtt szemlélhetjük immár e homonímákat ábécérendben és a tergo sorrendben, szófajok szerint csoportosítva és külön a tőszavak közül kiválogatva (arányuk ott lényegesen magasabb az össz-tőszókincshez viszonyítva) stb. Látjuk, hogy általában rövidebbek, mint a nem-homonímák; látjuk egyes további közös vonásaikat is.

A címszóból magából derül ki az, hogy hány tőből áll. Említett, hatvanezres mennyiségű anyagunk e szempontból így oszlik meg: egytövű 24, kéttövű 21, háromtövű 1,2, igekötős összetétel 9,4 ezer (ezen kívül van néhányszor tíz négytövű, ezer-egynéhány kétes összetettségű stb.). E puszta statisztikai összesítésnél talán érdekesebb volt az, hogy míg a szokásos szótárak módot nyúj-

tottak az azonos előtagú összetételek vizsgálatára, addig egy a tergo rendezés lehetővé tette az azonos utótagú elemek együttes szemlélését — ez bizonyos gondolatokat szülhetett a szóösszetétel szemantikai természetére vonatkozóan.[2]

A szófajok szerinti csoportosítás érdekes volt például a következő szempontokból. Megláthattuk azokat az arányokat, amelyekben az általunk vizsgált szótár az egyes szófajokat tartalmazta — ez volt a kérdésnek pusztán statisztikai oldala. Már nem statisztikai és sajátosan magyar nyelvi kérdésre kaptunk választ akkor, amikor az egyes szófajkombinációkat együtt találtuk: a kettős, hármas, négyes szófajkombinációkat, amelyek egyáltalán lehetségesek forrásunk állásfoglalása szerint; vagy amikor megvizsgálhattuk azon címszavak összességét, amelyek az ÉrtSz.-ban (nem véletlenül) nem kaptak egyáltalán semmiféle szófaji minősítést. Az is természetes, hogy az egyes szófajok ábécérendes, vagy a tergo listái alkalmasabbaknak bizonyultak a grammatikai kutatásokra, mint például egy egyszerű a tergo lista (amilyen e szempontból a VégSz. szólista-része is): az ikes igék közé nem keveredtek be a *trafik*, *kuvik*-féle szavak; az *-ás* képzősök közé az *ás* ige igekötős származékai stb.

A feljegyzett alaktani információk természetesen már a szófajoktól függőek voltak, szófajonként különbözők. Az ige esetében megkaptuk az azonos ragozási típusokhoz tartozók teljes listáit (ehhez természetesen még a géprevitel előtt fel kellett állítani az egyes igeragozási típusokat; ezt a rendszert mindjárt az első gépi eredmények visszanyerése után tökéletesíteni lehetett és kellett[3]); az azonos vonzatúak listáit stb. A főnév esetében a tőtípus (*ember*, *hajó*; *alma*; *bokor* stb. — a tőtípusokat eléggé szigorúan szinkron szempontból vettük, a toldalék-allomorf előhangzójának tekintve a tő és a toldalék mássalhangzó-eleme között esetlegesen fellépő magánhangzó-elemet), az egyes acc., a többes nom., illetőleg az egyes harmadik birtokos személyragos alak allomorfjai szerint kaptunk különféle listákat és statisztikai összesítéseket. A melléknév esetében természetesen külön csoportban szerepelnek egyik rendezésünk szerint azok az elemek, amelyek határozóragos alakjukat az *-(a)n/-(e)n* formánssal, és külön azok, amelyek az *-(u)l/-(ü)l* formánssal képezik — ismét külön azok, amelyek forrásunk és/vagy kódolóink nyelvérzéke szerint e szempontból ingadoznak.

A képző megléte vagy hiánya is leolvasható volt a címszó hangtestéből (ha nem is egészen egyszerű módon: ezt a munkát, mint ismeretes, nem különben igen becsült hallgató-munkatársainkra, hanem Jakab László kollégánkra bíztuk, az etimológia kódolásával együtt). Egy puszta a tergo rendezés is összehozza természetesen az azonos képzővel ellátott elemeket, legalábbis nagyjából — ám a képző kódolása finomabb, pontosabb ilyen szempontú rendezések létrehozását is lehetővé tette. Egyébként itt is el kellett végeznünk, mintegy

[2] Vö.: Hunyadi László, Az ÉrtSz. „új" stiláris minősítésű szavai. Nyr. XCV — sajtó alatt.

[3] Jánoska Sándor, A magyar ige automatikus toldalékolásának egy modellje. NytudÉrt. 58. sz. 464—468 (1967)

mellékesen, egy előkészítő munkát a szóképzéstan köréből — ki kellett emelnünk a „gyakori idegen szóvég"-eket egyrészt a puszta, kétségtelen magyar tővégek, másrészt a képzők tengeréből (vö. *-um*, *-ans/-áns*, *-(o)id*, *-ing*, *-(c)ia* stb.).

A címszó egészére vonatkozik annak stiláris minősítése (mi legalábbis ebben a feldolgozási szakaszban csak a címszó egészére vonatkozó stiláris minősítéseket vettük figyelembe). A legkülönfélébb szempontokból érdekesnek mutatkozott a forrásunkban azonos stílusminősítéssel ellátott elemek együttes szemlélése (vö. pl. a 2. lábjegyzetben említett munkával: az „új" minősítésű szavaknak a puszta, rendszerezés nélküli olvasása is egy korszak levegőjét, annak minden jó és kevésbé jó illatát közvetíti felénk, amikor ilyen töménységben, csak őket kapjuk együtt: *pártiroda*, *tejcsárda*, *újgazda*, *úttörőruha*, *pártdemokrácia*, *pártkritika*, *rohammunka*, *kuka*, *érdemtábla*, *szégyentábla*...: l. VégSz. 593—594). Megjegyzendő, mi az ÉrtSz. szerkesztői szellemének megfelelően stílusminősítésnek tekintettük a hangutánzó-hangulatfestő minősítést, ezen szavak tehát különféle kombinációkban egy-egy csoportot képeznek nálunk; vagy a „helytelen" minősítést (illetőleg az azzal egyenértékű emelt csillagot a címszó után) — tehát például az újabb szótárak készítői vagy a nyelvművelők egy csokorba gyűjtve találhatják nálunk mindazokat az elemeket, amelyeket jeles forrásunk elítélően minősített.

A címszótól s a rá vonatkozó szócikktől búcsúzva állapíthatjuk meg, hány jelentésre tagolva tárgyalta őt az ÉrtSz. E kérdésnek is volt egy statisztikai oldala: kiderült, hogy az egy, két, három stb. jelentésű címszavak bizonyos törvényszerűség szerint oszlottak meg forrásunkban (vö. VégSz. 590—591). Ugyanakkor volt több nem statisztikai vonatkozása is: ezek közül a legegyszerűbb a legtöbb jelentésű szavak elemzése.

Végül az egytövű címszavak mellett rögzítettünk még egy, nem szinkrón információt: a szó etimológiáját a SzófSz. alapján. Itt is érdekes volt mind a statisztikai összkép, mind az egyes etimológiai listák; az előbbi, mint látni fogjuk, még tovább színesedett, amikor elektronikára tértünk át.

Az eddig itt felsoroltak főként az egy-egy szempont szerinti rendezések lényegét érzékeltették, abba esetenként csak a szokásos ábécérend, vagy az a tergo rendezés szempontját vegyítve. Ám gondoljuk meg, mennyi további kombináció lehet releváns a nyelvész számára — az egyáltalán felkínálkozó számtalan lehetséges kombináció közül. Ha például érdekes volt az összetettség szerinti rendezés külön, meg a képző szerinti is külön — mennyivel érdekesebb lehet ezek kombinációja: a tőszavak (=egy tő és 0 képző), az egytövű egy képzősök, a kéttövű képzőtlenek, a kéttövű képzősök (alcsoportokkal: képző csupán az utótagon, képző csupán az előtagon, képző mind az elő-, mind az utótagon) stb. A tulajdonképpeni tőszavak hatezres csoportja egyáltalán, önmagában is igen lényeges volt: szinte mindent, amit az egész anyaggal elvégeztettünk, érdemesnek mutatkozott megvizsgálni a tőszavakon is — hosszúság, je-

lentésszám, szófaj, homonímia stb. szerinti eloszlásukat és ezek listáit. — Vagy, ragadjunk ki egy másik szempontot, a jelentésszámét, és kombináljuk azt egy vagy több további szemponttal. Immár pontos adataink vannak arra nézve, milyen összefüggés van a jelentésszám és a hossz között, szófaj és jelentésszám között (amint az várható volt: a legpoliszémebbek az igék, őket követik a melléknevek, majd a legoligoszémebbek a főnevek). Vizsgáltuk és részeredményeit tekintve a közeljövőben kívánjuk publikálni az összefüggést a jelentésszám és a szó strukturáltsága (motiváltsági foka) között. Eszerint minél strukturáltabb egy szó, statisztikailag szólva annál oligoszémebb (a tőszavak összességének átlagban több jelentése van, mint az egytövű-egyképzősöknek; ez utóbbiaknak átlagban viszont még mindig több jelentése van, mint a kéttövűeknek stb.), egyetlen csoport kivételével: az igekötős összetételekével. Az igekötős szavak összességének ugyanis átlagban több jelentése van, mint az egytövű, igekötőtlen szavaknak. Ezen anomália oka jelenlegi feltevésünk szerint azonban egy újabb szempont, a szófaj bevonásával megadható: az „igekötő nélküli szavak" zöme főnév, ezek pedig, mint mondottuk, a legoligoszémebbek; a „puszta igekötős szavak" csoportja mögött viszont zömmel igék bújnak meg.

2. A közvetlen eredmények továbbfejlesztésének, finomításának egyik lehetséges útja — egy-egy részkérdés tüzetes, mintegy monografikus, feldolgozása. Itt már látszólag messzebb kerülnek a gépek és közelebb a kutató ember — a valóságban az szokott a helyzet lenni, hogy itt sem tudunk megválni a gépektől: a probléma megoldása során egyre újabb és újabb, pontosabb és finomabb listákat, statisztikákat kérünk tőlük, mintegy állandó és intenzív párbeszédet folytatva velük, a problémáknak legalább részleges megoldásáig. Csupán két példát erre.

Vajon különbözik-e a mai „keletmagyar" fiatal értelmiség nyelve az ÉrtSz.-ban tükrözött normától, és ha igen — miben? Ezt a kérdést még éppen csak feltettük és néhány alaktani példával szemléltettük.[4] A kérdés helyes megértéséhez a következőket kell tudnunk. Az ÉrtSz.-t a gépi feldolgozás számára előkészítő hallgatók (meg az egyik tanár-résztvevő is: Jakab László) Hajdú-Biharból, Szabolcs-Szatmárból, Borsodból valók voltak. Azt az utasítást kapták, hogy az egytövűek nyelvtani szerelését vegyék át szigorúan az ÉrtSz.-ból, az összetetteket pedig lássák el hasonló szereléssel saját nyelvérzékük szerint. Tehát, például, állapítsák meg maguk az összetett főnevek tőtípusát, egyes acc., többes nom., egyes harmadik birtokos személyragos allomorfját; állapítsák meg maguk, *-an* vagy *-ul* toldalékot kap-e az összetett melléknév, és így tovább. Tekintettel arra, hogy a VégSz.-ban, az a tergo sorrendnek megfelelően, előbb valamely (képzős vagy képző nélküli) egytövű áll, majd a vele mint utótaggal alkotott összetételek, a VégSz.-ból közvetlenül leolvasható előbb (az egytövű kódjaiból) az ÉrtSz.-norma, majd a rögtön alatta következő össze-

[4] Papp Ferenc—Jánoska Sándor, A Magyar Szóvétmutató Szótár és az alaktani ingadozások vizsgálata. MNy. LXIII, 138—148 (1967)

tételekből az ifjabb s egységesen „keletmagyar" (jobb közelítést nem lehet adni rájuk) norma, ingadozások stb. Íme, egyetlen példa, némi rövidítéssel: az ÉrtSz.-ban a *tár*[2] fn jellemzése: *-t (-at), -ak* v. *-ok, -a*. Tekintsünk el az összes *-tár*[2] utótagú összetétel vizsgálatától és vegyük csupán a *szótár* összetételt s az utóbbival mint utótaggal alkotottakat. A következő képet kapjuk:

szótár	*-t, -ok, -a*
zsebszótár	*-at* v. *-t, -ak, -a* v. *-ja*
kéziszótár	*-at, -ak, -a*
tájszótár	*-t, -ak, -a*
szakszótár	*-t, -ok, -a*
rímszótár	*-t, -ok, -ja*
műszótár	*-t, -ak* v. *-ok, -a* v. *-ja*
nagyszótár	*-t, -ok, -a.*

Mint látjuk, van itt minden: ingadozás olyan helyen, ahol a tőszó maga nem is ingadozik; előhangzótlan tárgyeset és nyíltabb előhangzós többes (általában a zártabb előhangzót várnánk a ∅ előhangzó mellé egy paradigmába); ingadozás az előhangzóban; ingadozás a birtokos személyragos alakban. Egyelőre még nem tudtuk közös nevezőre hozni az efféle mikroeltérések tömegét, egy alább említendő részkérdést kivéve nem tudtunk felderíteni valamely általános tendenciát — ha egyáltalán felderíthető lesz ilyen. Hangsúlyozni szeretnénk, hogy itt egyáltalán nem tájnyelvi kutatásról van szó, legalábbis annak szokásos értelmében. Az összetételek utótagjában saját döntésükre bízott fiatalemberek nem valamely magyar nyelvjárás képviselői: készülő (ma aktív) magyar tanárok, a magyar irodalmi nyelv hordozói, sőt terjesztői. Ha tehát lesz valamely eltérés az ÉrtSz.-norma és az ő normájuk között — akkor ez tulajdonképpen a mai irodalmi nyelven, művelt köznyelven belüli normapárhuzam.

Másik példaként egy már — érzésünk szerint — megoldottnak tekinthető kérdést veszünk elő, szintén a grammatika szövevényéből. Régóta foglalkoztatott bennünket a *lába—combja*-féle alakok kérdése, az a szakirodalmunkban igen sokszor felvetett probléma, hogy az egyes harmadik birtokos személyragos alakok és a velük szoros kapcsolatban levő több birtokra mutatók (*lábaim—combjaim)* mikor képezik *j*-vel és mikor *j* nélkül jelzett alakjaikat. Példapárunk már mutatja, hogy az elsődlegesen kínálkozó hangtani magyarázat elesik: a *-b* vég után (és így még jó néhány mássalhangzó-vég után) állhat is *j*, meg nem is. Sőt, elesik egy további, logikusnak látszó magyarázat, amely szerint esetleg egy már eleve meglévő, esetleg önmagában is kellemetlen mássalhangzókapcsolat nem bővül még egy *j*-vel is, míg a magában álló mássalhangzó esetleg jobban vonzza a *j*-t: ez már példapárunkban sincs így, de igen-igen sok további esetben sem. (Vö., pl.: *ablaka—barackja* — a kiemelt *-ckj-* — mássalhangzókapcsolat még valamely szláv nyelvnek is becsületére válnék, nemhogy a mi,

mássalhangzótorlódást köztudomásúan nem nagyon kedvelő rendszerünknek). Részint (nem is kis mértékben) egyes korábbi kutatásokra, részint a külön e célból készült számos gépi rendezés eredményére támaszkodva végülis a következő hipotézisre jutottunk ezzel kapcsolatosan. Tegyük fel, hogy nyelvünkben működik egy bázismegkülönböztető tendencia, mely a kérdéses esetben a puszta (nem birtokos) bázisvéget állítja szembe a birtokos bázisvéggel. Ha „szokásos" (tehát gyakori) tővéggel van dolgunk — nincs szükség külön eszközzel való kiemelésre, a birtokos toldalék lehet egyszerűen *-a/-e*. Ha viszont „szokatlan" (ritka) tővéggel van dolgunk — bármennyire ellenkeznénk is nyelvünk hangtani rendszere, győz az alaktan parancsoló követelménye: ezt a szokatlan véget megtoldjuk még egy *-j*-vel is, éppen, hogy felhívjuk a figyelmet: bármilyen szokatlan, de itt a tővég. Így jönnek létre véleményünk szerint a *-mbj-*, *-ckj-* stb. hangkapcsolatok a már önmagukban is eléggé szokatlan *-mb*, *-ck* stb. tővégek után. Ezért lép fel csak igen gyéren a *j* produktív képzőink után — ezért léphet fel nem produktív vagy nem kimondottan főnévképzőinket követően; ezért veszhet el egy korábbi *j* a nyelvtörténet során újból produktívvá vált képző mögül *(magzatja>magzata)*, és így tovább.

Nem kívánjuk itt konkrétabbá tenni ezt a példát, hiszen jelen összefüggésben csupán példa volt a közvetlen gépi feldolgozás egy továbbvitelére az emberi gondolat segítségével. (Egy gép egyelőre nem tud hipotézist felállítani — bár éppenséggel erre is programozható lenne alkalmasint.) Ugyanakkor jól látható a gépi munka elengedhetetlen volta a kiinduló pontnál: a „szokásos", „gyakori" és megfordítva — „nem szokásos", „ritka" szóvégeket kellő mennyiségű anyagon kézzel csak igen fáradságosan lehetne megállapítani, vagy egyáltalán nem is lehetne. Hadd tegyük hozzá, hogy épp a *j*—nem *j* kérdését alaposan megvizsgáltuk az összetételekben, tehát volt munkatársaink nyelvérzéke szempontjából is. Úgy találtuk, hogy ők egyáltalán nem törekednek a *j* szaporítására, holott a *j* fiatalabb és máig is élő volta miatt eleve talán ezt feltételeznők. Sőt, az összetételekben érezhetően kisebb volt a *j*-vel bővülők aránya, mint az ugyanolyan végű egytövűek között. Ez az eredmény természetesen egy újabb kérdést vetett fel: vajon nem kezeli-e a nyelvérzék a gyakori utótagot mintegy képzőszerűen, tipikus szóvég-szerűen és legalábbis részben nem ez a körülmény okozza-e az összetételek viszonylagos *j*-tlenségét?

3. A közvetlen eredmények továbbfejlesztésének egy másik útja: visszatérés a gépekhez, de immár, az elektromechanikusak helyett — az elektronikusokhoz. Bennünket, nyelvészeket, nem túlságosan érdekelnek a technikai részletek: elég számunkra annyi, hogy az elektronikus gépek hajlékonyabb logikájúak, azokkal már nem csupán nagy tömegű cédulák rendezését, hanem ennél valamivel magasabb szellemi munkát is el lehet végeztetni.

A legfontosabb, amit ezen a téren az elmúlt évek alatt elértünk, az, hogy megtanítottuk egyetemünk ODRA-1013 nevű számológépét — magyarul olvasni, legalábbis elemista fokon. Ezen azt értjük, hogy ha beadunk neki

szövegeket, akkor ő azt a magyar helyesírás szerinti betűkké (tehát tulajdonképpen: fonéma-jelekké) tudja átalakítani, ilyenekként tudja értelmezni. Bármely magyar szöveg ugyanis elsődlegesen *n*-ekből és nem betűkből áll, ha a *sz, zs, ly* stb. *n*-kapcsolatokat tekintjük betűknek, amint ez a magyar szakirodalomban eléggé hagyományos. Ezt az *n*-ekből álló (*n*-ekkel leírt) szöveget át kell valahogy alakítanunk, hogy nyelvészetileg érdekes dolgokat kaphassunk, bennünket ugyanis csupán a legritkább esetben szoktak az *n*-ek érdekelni, sokkal érdekesebbek az egy- és többjegyű betűk, a mögöttük meghúzódó fonéma-realitással. (Az „elemista fokon" megszorítás azt jelenti, hogy csupán tőszavak között mozog teljes biztonsággal a gép. Ezeken túlmenően is sokszor jól „olvas", így például az *arcszín* összetett szót helyesen tagolná: *arc/szín*, ám olykor téved — így a *község* szóban kétjegyű *zs* betűt vélne felfedezni.)

Érdekességként megjegyzem a következőt. Világos az *n*-eknek betűkké való átalakításában az alapprobléma: bizonyos *n*-ek teljesen egyértelműek, mindig önmagukban alkotnak betűt (ilyen például a *w*, az *x*); mások viszont nem. Ha következetesek vagyunk, ez utóbbi, tehát többjegyű betűk csoportjába kell sorolnunk nem csupán a *cs, gy* stb. kapcsolatokat, hanem a *bb, cc, dd* stb. kapcsolatokat is. Ha ugyanis egy-egy betűnek azt tartjuk, ami az írásban egy-egy fonémának felel meg — akkor, elismerve, hogy a hosszú mássalhangzók is külön fonémák, az őket jelölő *n*-kapcsolatokat is egységes betűknek kell tekintenünk. Mármost hogyan lehet automatikusan felismertetni ezeket a többjegyű betűket? Kiderült a következő: Egyszerűbb lesz a felismerő algoritmus, ha a szót a végétől (jobb szélétől) kezdjük olvasni. A gép tehát a következő lépésekben olvas: *a)* bevesz egy szót helyköztől—helyközig és elhelyezi azt *n*-enként a memóriájában; *b)* az így elhelyezett szót végétől kezdve alakítja át betűkké. Vegyünk egy rövid példát: *lyuk*, ezt *a)* elteszi a szokásos sorrenben: *l-y-u-k*, majd *b)* kezdi: *k*, megvizsgálja, nincs-e ez előtt még egy *k* — ha, mint itt, ilyen nincs, azonosítja: „ez a *k* betű", az *u* — „*u* betű"; az *y*-t észlelve — és megállapítva, hogy ez nem a *brandy, yard* szavak valamelyike — azonnal előrébb ugrik és azonosítja: „ez az *ly*!". Hasonlóan jár el természetesen a *hattyú, briddzsel* [foglalkozik] stb. stb. szavak esetében is: beolvassa *n*-enként mindegyiket szokásos módon, balról jobbra, ám betűkké göngyölíti — jobbról balra haladva.

Ha a gép ismeri a fonémákat — ezen ismeretét hasznosítani tudjuk. Bemutatunk erre is két példát.

A gép fonémaismeretét fel tudjuk használni tisztán kutatási célokra. Például: megfelelő program alapján (melyet Jékel Pál tudományos kutató készített) utasítottuk a gépet: készítsen nekünk fonémastatisztikát a beadott magyar szövegekről. Adja ki az eredményt kétféleképpen: *a)* a fonémák ábécérendjében, közölve az abszolút értékeket és a százalékokat, *b)* az előfordulások csökkenő gyakorisága sorrendjében, közölve a sorszámot, az abszolút értéket, a százalékot és a kumulált százalékot (erről l. alább). Az első

részfeladat, amellyel ennek kapcsán megbíztuk a gépet, az volt: készítsen tőszókincsünk egyes etimológiai rétegeiről fonémastatisztikát. Nem bízván a facsimilék nyomdai minőségében, pontosan kimásolom ide a gép írásos válaszának első néhány sorát, majd némi magyarázatot fűzök hozzá (l. 1. táblázat).

1. táblázat

a ad agg
oesszes szoo: 611
oesszes hang: 2137
aatlagos szoohossz: 3.50 hang/szoo
a) betueerendben

hang	db	%
a	175	8.19
aa	72	3.37
b	24	1.12
bb	0	0.00
c	2	0.09
cc	0	0.00
cs	25	1.17
ccs	0	0.00

stb.

Az „a ad agg" a táblázat elején a következőt jelentette: A gép nem tudta magától, hogy első etimológiai csoportunkat mi úgy hívjuk — „finnugor eredetű tőszavaink". Ránk bízva ezen elnevezés adását, kiírta a táblázat élére az e csoportban előforduló legelső egy-két szót, a finnugorok között ez éppen az *a*, *ad*, *agg* volt; ezek az első szavak tehát mintegy azonosították ezt a csoportot, megkülönböztették a többi etimológiai csoporttól, melyek mindegyike úgyszintén az első egy-két szóval volt jellemezve. Kiderül tehát, hogy összes, a gép által vizsgált (az ÉrtSz.-ben szereplő) fgr eredetű tőszavunk 611 volt, ezek összesen 2137 betűt (nem *n*-et!) tartalmaztak, e szavak átlagos hossza ugyanezen mértékegységben: 3,50. Az ezt követő részeredmények talán önmagukért beszélnek (*aa*=*á* stb., a szokásos távirat-kód szerint). Az „*a)* betűrendben" azt jelentette, hogy ezek után jött egy ilyen felirat: „*b)* csökkenő gyakoriság szerint:". Az alábbiakban épp ez utóbbi módon, a csökkenő gyakoriság szerint közlöm néhány fontosabb etimológiai rétegünk százalékos fonémastatisztikáját (l. 2—5. táblázat).

2. táblázat
Fgr eredetű tőszavak

Átlagos szóhossz: 3,50

1.	a	8	8
2.	e	8	16
3.	k	6	22
4.	l	6	28
5.	r	5	33

3. táblázat
Török eredetű tőszavak

Átlagos szóhossz: 4,43

1.	r	9	9
2.	k	9	18
3.	a	7	25
4.	e	5	30
5.	o	5	35

4. táblázat
Szláv eredetű tőszavak
Átlagos szóhossz: 5,21

1.	a	15	15
2.	r	7	22
3.	k	7	29
4.	o	6	35
5.	e	6	41

5. táblázat
Német ereredetű tőszavaink
Átlagos szóhossz: 5,17

1.	r	10	10
2.	a	8	18
3.	á	6	25
4.	t	6	30
5.	o	5	35

E táblázatokhoz a következő megjegyzéseket kell fűznünk. Az eredmények nek itt csupán a kerekített százalékos értékeit közöljük, mégpedig az első oszlopban — magának az adott fonémának a százalékos arányát az adott eti mológiai rétegbe tartozó összes tőszó valamennyi fonémája között; a második ban — az úgynevezett kumulált százalékot. Ez utóbbi szám megmutatja, hog y az addigi fonémák összesen hány százalékát teszik ki az adott etimológiai réteg által képviselt valamennyi fonémának. Például: fgr. tőszavaink között 1. h elyen (leggyakoribbként) az *a* áll, 8%-kal, a 2. az *e* ugyancsak — kerekítve — 8%-al, ők ketten együtt alkotják a fgr. tőszavainkban tartalmazott összes f onéma 16%-át; 3. a *k* önmagában 6%-kal — e három leggyakoribb együtt 22%; és így tovább. Lehet vizsgálni az így nyert eredményeket a magánhangzók— mássalhangzók aránya szempontjából; a magánhangzókon belül lehet vizsgálni a mélyek és magasak arányát; mind a magánhangzókon, mind a mássalhangzókon belül a hosszúak- rövidekét — és így tovább. De szembe lehet állítani ezeket az eredményeket egészen más jellegű magyar eredményekkel is. Végeztettünk például fonémastatisztikát nem olyan kissé mesterséges „szövegen", mint amilyen egy szótár (annak bizonyos részei), hanem valódi, mintegy 10 000 szövegszót felölelő, összefüggő magyar szakszövegen, úgyszintén Gulyás Pál „Misztikus ünnepi asztal" c. kötetén. E vizsgálatok eredményeinek egy részét közli a 6—7. táblázat.

6. táblázat
Technikai szöveg
Átlagos szóhossz: 7,11

1.	e	11	11
2.	a	9	21
3.	t	7	27
4.	l	5	33
5.	k	5	38

7. táblázat
Gulyás Pál
Átlagos szóhossz: 4,12

1.	a	10	10
2.	e	10	20
3.	t	6	26
4.	n	6	32
5.	l	5	37

Ne siessünk konklúziókat levonni; csak egy lehetséges irányt szerettünk volna itt bemutatni, nem többet.

S egy másik, tisztán gyakorlati felhasználási irány. Az Alföldi Nyomda megkeresett bennünket: üzemük s az egész magyar nyomdaipar teljes automatizálását (a számítógépes fényszedésre való áttérést) nem tudnánk-e azzal segíteni, hogy a magyar elválasztás szabályait számítógép számára feldolgozható formába öntjük? (Ez esetben ui. a szedőnek csak folyamatosan szednie kell — a számítógép gondoskodik a sorzárásról, a tördelésről stb.) Az egy- és

többjegyű betűk ismerete elengedhetetlen feltétele a magyar elválasztás szabályainak (vö. *-ccs- ... cs-cs, -ddzs- ... dzs-dzs;* nem beszélve a *dz* okozta felesleges nehézségekről: *brin-dza*, de *mad-zag*, vö. HSzab.[10] 321.) — persze, csak egyik feltétele. Hozzávettük a többi szükséges feltételt is és ma már mondhatjuk, hogy tőszóállományunkon belül az ODRA-1013 hibátlanul „szótagol" (elválaszt); elő van készítve arra is, hogy azon túllépve, képzett és összetett szavaink tartományában is elvégezze ezt a munkát. Csak memória kérdése, hogy ezek után áttérjen, ha szükséges, a tulajdonnevek tartományába is. Mindezzel végső fokon elérhetjük, hogy könyveink gyorsabban, olcsóbban, szebben készülnek; hogy újságjaink lapzártája 2—3 órával későbbre tolódik — tehát ennyivel frissebb híreket kapunk bennük[5] és így tovább. (Persze a számítógépes fényszedésnek is vannak az egyszerű elválasztáson túlmenő feltételei — reméljük, megoldják azok, akikre tartozik. Mi részben már megtettük azt, ami rajtunk állott.)

4. Így volt eddig a magyar nyelv tudományos vizsgálat tárgya s feldolgozás anyaga számítógépeinkben. A jövőben hasonló, de immár formáján kívül a tartalmát is érintő, jelenén kívül a múltját is feldolgozó munkákat kívánunk gépeinkre bízni — erről azonban ma még korai lenne beszélni.

PAPP FERENC

Исследование венгерского языка с помощью вычислительных машин в университете им. Л. Кошута

В университете им. Л. Кошута уже несколько лет ведется исследование венгерского языка с помощью вычислительных машин.

Первым этапом в этом деле явилась обработка Толкового словаря венгерского языка на перфокартных машинах. В результате этой обработки мы обладаем сводными картинами относительно буквенного состава заглавных слов этого словаря, относительно омоним, содержащихся в нем, относительно сложных слов, суффиксальных образований, распределения слов по частям речи, морфологических признаков обработанных слов, сильного управления глаголов, этимологии однокорневых слов, количества значений и длины слов, а также распределения их по стилистическим пластам.

Можно и нужно более основательно обработать первичные машинные данные. Обрабатывается материал с целью выявить микрорасхождения между нормой Толкового словаря и живой нормой молодых студентов — специалистов по венгерскому языку, подготовивших этот словарь к машинной обработке. На основе машинных данных выдвинута гипотеза о тенденции к различению именных базисов в венгерском языке.

[5] Erről bővebben l. TAKÁCS LAJOS, A nyomdai elválasztás automatizálásáról. Nyr. XCV — sajtó alatt.

Можно пойти и в другом направлении: по пути изучения структуры венгерского языка с помощью электронных вычислительных машин. Важнейший шаг до сих пор в этом направлении заключается в том, что к ОДРЕ 1013 университета составлена программа, с помощью которой она автоматически превращает тексты, поданные в орфографической форме, в фонемную форму — безошибочно в области корневых слов и с ошибками среди других. На базе этой программы составлена статистика фонем в разных этимологических пластах (ср. табл. 1—5), а также в текстах (табл. 6—7). В результате дальнейшего развития этой же программы решен вопрос автоматического переноса корневых слов, важный с практической точки зрения полной автоматизации нашей типографской промышленности.

Ф. Папп