Papp Ferenc—Hexendorf Edit

MAGYAR SZÓKINCS A KÖNYVNYOMTATÁSTÓL NAPJAINKIG — SZÁMÍTÓGÉPRE TERVEZVE

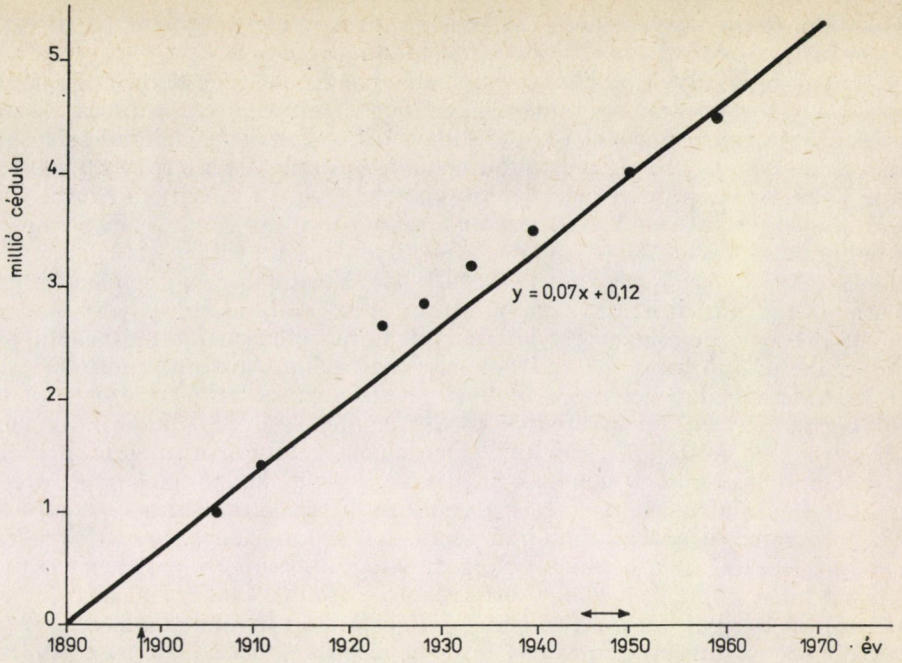
A Nagyszótár gondolata másoknál és nálunk, korábban és ma

Richelieu bíboros 1635-ben egy nagyszótár megalkotása céljából alapította a francia akadémiát, az orosz akadémia 1783-ban ugyanilyen céllal jött létre, s két egymást követő szótári kiadvány megjelentetése után 1841-ben olvadt bele a nála régebbi pétervári tudományos és művészeti akadémiába. A mi tudományos akadémiánk is reformkori létesítésétől kezdve több különféle műfajú szótár készítését szorgalmazta és biztosította, mintegy könnyebben megvalósítható előzményként ahhoz a nagyszótárhoz képest, amelynek munkálatai a múlt század utolsó évtizedében indultak meg. Így amikor Akadémiánk elnöksége 1984. február 28-án határozatot fogadott el egy új koncepciójú nagyszótár összeállításáról, ezzel régi hagyományokat folytat.

Az új nagyszótári tervnek közvetlen előzményét az úgynevezett „régi Nagyszótár" jelenti, amely tulajdonképpen egy gazdag cédulaanyag, egy múlt század végi elgondolás alapján hosszú évtizedekig folyó gyűjtés eredménye. Az elgondolás az volt, hogy a felvilágosodás korával kezdődő időszak irodalmi szókincséből kell nagyszótárt készíteni. Tekintsük meg az 1. ábrát. Ezen, főleg [1], továbbá [2] számadatai alapján azt tüntettük fel, milyen ütemben haladtak a tervezett szótár munkálatai: hány cédula készült el — rajta egy-egy szóval s annak környezetével, a forrás többé-kevésbé pontos megjelölésével — a múlt század utolsó évtizedétől a közelmúltig. Becslés szerint ma mintegy ötmillió kézzel rótt, géppel írt, a napisajtóból egyszerűen kivágott adatunk van feltehetően több százezer magyar szóra, tehát mindegyikükre — rendkívül nagy szórással! — átlagban tízes nagyságrendű kártyánk. Néhány évvel ezelőtt leállítottuk ennek a hatalmas archívumnak a további gyarapítását, most pedig egy részben tartalmilag is új, de arányaiban és technikáját tekintve korábban meg sem álmodhatott munka készítésébe kezdtünk. A régi gyűjtemény mérsékelt ütemű továbbrendezéséről nem mondunk le, de adatait egyelőre semmiképpen sem kívánjuk beolvasztani az új anyag tengerébe. Nem lehet meghatódottság nélkül kézbe venni a mély rekeszű polcokon, nagy dobozokban tárolt, összesen több száz folyómétert kitevő anyag céduláit. Az adatok gyűjtésén több generáció tagjai fáradoztak — akkori s jövőbeli akadémikusok a főtitkártól az egyszerű tagig, jeles nyelvészek és névtelen középiskolai tanárok, szegénysorsú diákok. De részben ezért, részben nagy mennyisége miatt az anyag nehezen kezelhetővé vált, s a feljegyzések pontosságát csak hosszas utánjárással lehet ellenőrizni.

Miért kell a nyelvésznek több millió adat a szavakra?

Jól sejtették hazai és külföldi elődeink, hogy egy nép nyelvének szóanyaga felbecsülhetetlen értékű nemzeti kincs: őrzi múltunkat, jellemzi jelenünket, bizonyos mértékű kipillantást enged a jövőbe. Igen logikus volt a gondolat,



1. A Nagyszótár cédulaállományának gyarapodása. — Megjegyzések: *1.* Az MTA csak 1898-ban szervezett állandó Szótári Bizottságot (erre a dátumra nyíl mutat), egy 1891-i, ill. 1893-i *Nyelvőr*-cikk hatására azonban a gyűjtés már ezekben az években spontán [!] megindult (vö. [1], 447, 450—1). — *2.* „1945 és 1950 között a nagyszótáron nem dolgoztak" (uo., 463). Ettől az egyetlen periódustól eltekintve háború, forradalom, infláció, válság stb. csak alig-alig érezteti hatását a növekedésen, mely regressziószámítással lineáris megközelítés esetén egy $y = 0{,}07x + 0{,}12$ összefüggéssel írható le, ahol y a cédulamennyiség, x az 1892 óta eltelt évek száma. Az anyag évi gyarapodása tehát — úgy látszik — független volt a cédulaírásra kifizethető pénzösszegtől, másrészt az irányító akadémiai szervek lelkesedésétől is: „1900 és 1906 között már az érdektelenség légkörében zajlottak a Szótári Bizottság ülései" ([1], 455), vö. ezzel az ezt követő normális gyarapodást

hogy ezzel a kinccsel sáfárkodnunk kell — legelőször leltárba kell vennünk. Méghozzá nem elszigetelten az egyes szavakat, hanem szövegkörnyezetükkel, legalább azzal a mondattal együtt, ahol előfordultak. A szavak csak így élnek, így mutatják meg igazi természetüket, jelentésüket, használati köreiket. Ha számba vesszük, hogy egy-egy mai nyelv szókészlete alkalmasint több százezer, talán az egymilliót is meghaladó különböző szót tartalmaz, s ezeknek mindegyikét többször, mindig az adott környezettel és a forrás megjelölésével kell rögzítenünk, akkor elég könnyen belátható, milyen hatalmas feladat csupán ennek a leltárnak az elkészítése is, hány mega- vagy gigabájtnyi memóriára van szükség, s ennek rugalmas, sokoldalú hozzáférhetőségét is biztosítani kell. Ezt meggondolva különösképpen tisztelettel nézünk elődeinkre: nem ők voltak a törpék — ők jól látták a feladatot; a technika volt törpe, s az ember nagy. *Babbage* gépének esetében az eszköz teljesítőképessége és a megoldani kívánt feladat között csupán kis távolság volt ahhoz képest, amekkora filológus elődeink hagyományos munkaeszközeit a megálmodott cél elérésétől, a nyelvi tények tökéletes leltározásának lehetőségétől elválasztotta.

Hozzunk fel néhány példát arra, mit kíván a nyelvész (történész, néprajzkutató és így tovább) ebből a leltárból kibányászni!

A szavak nem kövek, abban az értelemben semmiképpen nem, hogy simára kopnának a sok használattól, hiszen közülük éppen a leggyakrabban használtak legtöbbje egyre ágasabbá-bogasabbá válik, egyre több jelentéssel gazdagodik. Íme, nagy akadémiai értelmező szótárunknak [3] mintegy 60 000 címszava közül azok, amelyeknek a legtöbb jelentése van nyilvántartva, zárójelben a jelentés-számmal (NB.: ezt az értelmező szótár számítógépes — elektromechanikus! — feldolgozásából tudjuk ilyen pontosan): *is* (101); *van* (65); *úgy, hogy*(2) (55); *csak* (48); *jár* (47); *áll*1 (45); *megy* (43) . . . Igen hasonlít ez a lista a leggyakoribb szavakéhoz. Mármost számoljunk úgy, hogy az *is*-re kapott 101 különféle jelentés, jól disztingválható használat reális. Ha ezek mindegyikére több különféle szövegkörnyezetű példát kívánunk fellelni s rögzíteni, akkor világos, hogy csak erre az egyetlen szóra tíz- és tízezer adat kell a feldolgozni kívánt fél évezredből. S akkor még csupán ellenőriztük, hogy az ehhez képest nyilván kisebb anyagon dolgozó nyelvész intuíciója megfelelő volt-e. Gondolható, hogy nemcsak a „nagy anyagot" kell majd a maga egészében gépen mozgatnunk, hanem az ilyen kisebb részleltárakat is érdemes lesz külön programmal áttekinthetőbbé tenni. Az is kiderülhet, hogy lényegesen egyszerűbben kell megfognunk a dolgot, átfogóbb csoportokra bontva: de ez akkor derülhet ki, ha ez a több tízezer rekord előttünk van, ezt a nagy anyagot a kutató kívánságainak megfelelően többféle módon kombináltuk. S egy más, az értelmező szótár lapjain szinte fel sem merült kérdés: vajon az eltelt fél évezred alatt nem változott-e szócskánk használata? Már előre is sejthető, hogy változott, tehát valószínűleg még pontosabb és bonyolultabb képet kapunk a valóságról, mint amit az értelmező szótár adhatott.

Ez a példa öncélúan nyelvészetinek tűnhet. (Pedig valójában nem az. A logikai nyelvészetnek egy nemrég nálunk járt világhírű szovjet művelője terveinket s a régi gyűjtésű anyag céduláit szemlélve felsóhajtott: „Ó, bárcsak együtt láthatnám a *vagy* választó kötőszónak [or. ili] a Háború és békében előforduló összes adatát!" Épp a nyelvészet bizonyos legmodernebb — a logikához közeledő — ágai igen sokat foglalkoznak az ilyen „semmi" szavakkal.) Álljon itt tehát egy másfajta kutatási tárgy is példaként. Egy történészt, jogászt, irodalomtudományi vagy néprajzi kutatót nyilván érdekel az, hogy a történelmi, jogi, irodalmi stb. terminológiának az őt éppen érintő része hogyan is alakult az évszázadok folyamán. Joggal kérheti akkor tőlünk csak az ilyen jellegű dokumentumokat századonként, tárgynak, fogalomkörnek, korszaknak, szerzőnek, vagy egyéb szempontnak megfelelően. Szókincsünk tanúja múltunknak, de túl sokat tudó tanú. Épp szerteágazósága, gazdagsága miatt eddig ö s s z e g e z ő vallomását nem is igen lehetett számba venni. (Az összegező-t azért hangsúlyozom, mert e g y e s e s e t e k b e n eddig is vallott nekünk, nem is keveset s nem is érdektelent. Így például tudjuk, hogy a *tárgy* szó, melynek eddigi első magyar előfordulását 1495-ből ismerjük, ófrancia eredetű, s az akkori lovagi világ fontos, 'pajzs' jelentésű kifejezése volt. Vagy hogy ha valaki ma divatos kifejezéssel szólva „Franciába megy" üdülni, ezzel csak ő hiszi magát modern előkelőnek: a „ffrancyab[a] menny" kifejezés — a későbbi köznyelvi „Franciaországba . . ." helyett — eddigi tudomásunk szerint a Jókai Kódexben fordul elő először, mintegy hatszáz évvel ezelőtt.)

Még egy összegző példa. Más irányú vizsgálataink igen valószínűvé teszik (történeti-etimológiai szótárunk [4] egy harmadának számítógépes feldolgozása

eredményeként), hogy az 1945 utáni nagy történelmi változást szókincsünk elsősorban nem az orosz kölcsönzések nagyobb tömegével tanúsítja. Hanem azzal, hogy a nemzetközi szavaknak, közös európaizmusoknak egy elég jelentős része (*agresszor, aktivizál, aspirantúra, centrizmus, centrista, diszpécser, diverzáns, fasizál, fesztivál, fetisizál* stb.) került betöbbé-kevésbé valószínűleg o r o s z k ö z v e t í t é s s e l. Hogy így számos Nyugatról kiinduló hatás az elmúlt négy évtizedben Keletről érkezett hozzánk, azt mások talán a szókincs számítógépes feldolgozása nélkül is tudják. De mi nyelvészek vagyunk, s ezért nekünk ezt ezen az anyagon kellett konkrétan megnéznünk, kimutatnunk.

Az összegező példák egyébként messzire vinnének. Hangalaki kérdések köréből hozunk példát. Mintha az utóbbi időben újra tért hódítana az a régebbi tévhit, hogy a beáramló nagy számú jövevényszó a mássalhangzók és a magánhangzók előnytelen arányával rontja nyelvünket, túl sok bennük a mássalhangzó. Ebben az esetben az összegezésnek immár nem csupán a szavakat, azok egyes csoportjait kell érintenie, hanem azok összetevőit, a hangokét. Nos, az ilyen részletekig is elmehető leltár mást, lényegesen bonyolultabb képet mutat. A hangalak nem lévén írásunk tárgya, csupán két feladványt tűzünk ki az olvasó elé ezzel kapcsolatban: *a)* Hol jobb a mássalhangzó-magánhangzó arány: a *szinkronciklofazotron, ribonukleinsav* stb. szavainkban, vagy az ősi *térd, kard* stb. elemekben? *b)* Ősi *térd* szavunk egyik birtokos személyragos alakja: *térde* (és csak népiesen, esetleg: *térdje, térgye).* De akkor a számítástechnikai kifejezésként 'adatállomány' jelentésben ismert *fájl* (más helyesírásban *file)* ragozva miért *fájlja (file-ja)?* Miért követel meg nyelvünk az úgyis elég sok mássalhangzó után még egy *-j*-t is, miért nem lehet így: *fájla (file-a)?*

Nos, a több millió adat azért i s kell, hogy az efféle kérdésekre választ tudjunk adni.

A Nagyszótár új koncepciója: szavak számítógép memóriájában

Az új koncepció igen könnyen vázolható. Feladatunk, hogy a könyvnyomtatás korától, pontosabban: *Komjáti Benedek* bibliafordítás-részletének megjelenésétől (1533) napjainkig számítógép memóriájába vigyünk mintegy tízmillió „cédulát" (rekordot: tízmillió szót, amilyen alakban éppen előfordul, a helyes értelmezést biztosító szövegkörnyezettel és pontos forrásmegjelöléssel). Hogy abból azután a föntebbi példák némelyikében már elő-előbukkanó, a hangoktól a szószerkezetekig különféle szinteket érintő, különféle kombinációjú rendezéseket kapjunk vissza. Hogy abból, annak tetszőleges — például korok, műfajok, szerzők stb. szerinti — részleteiből vagy egészéből rögtön kész fényszedő szalagot nyerjünk, s így hagyományos könyvet (szótárt) készíthessünk. Hogy az terminálon egyelőre a fontosabb munkahelyeken, később egyes kutatók otthonában is rendelkezésre álljon. Hogy a szókincsünkkel, annak történetével, mai kérdéseivel foglalkozó cikkét munkahelyi vagy otthoni számítógépe segítségével készítő szakember, e szolgáltatás anyagait közvetlenül felhasználva, maga is fényszedő szalagjával kopogtasson szerkesztősége, kiadója ajtaján, gépelt kézirata mellett. — Hogy ez a gépi szótár ne romoljon, hanem állandóan javuljon: egy dolgozónk a használóktól érkező kritikai megjegyzéseket este beviszi az anyagba, s ez másnap már ennyivel jobb lehet. Hogy ez a szótár ne avuljon el, hanem mindig napra kész legyen: az esti vizsgálatkor belekerülnek az új szavak, új jelentést sejtető környezetek.

A távlatok mellett jól látjuk nehézségeinket is. (Igaz, hogy ha ezeket leküzdhetetleneknek ítélnénk, bele sem fogtunk volna a dologba.) Íme, ezekből is csak néhányat.

A problémáknak igen széles körét átlátó, azt megoldani tudó szakemberekre lenne szükségünk. A problémák két távoli pontja: *a)* Most készít részünkre tervtanulmányt az OSZK-nak egy tudományos dolgozója a magyar nyelvű nyomtatványoknak történetileg teljes betűkészletéről. Csak egészen halvány elképzelésünk van arról, hány különféle karakterre lesz szükségünk 1533-tól napjainkig. Vajon nem fogja-e ezek száma végül is meghaladni a 256-ot? Rokonszenves megoldásnak látszanék a szabad karaktergenerálás, tehát hogy a teljes betűkészletet kiadványonként, esetleg nyomdánként teremtsük újra adatrögzítéskor, de akkor e jelek későbbi megjelenítése az összesített anyagban, betűrendbe rendezése stb. (szinte?) lehetetlen volna. *b)* A hatalmas anyag fogadására, rugalmas kezelésére, de még egyszerűen az adminisztrálására is egy rendszerszervező programozó kellene, aki eddig pl. a lakosságnyilvántartásban dolgozott (vö. fentebb emlegetett 10 milliós rekordmennyiségünkkel). És ami az *a)* és a *b)* között van? És ami alatta van: valamennyire is megbízható és állandó adatrögzítő stáb?

A távoli célokon és a jelenlegi nehézségeken kívül szólhatunk olyan pozitívumokról is, amelyek a munkálat megindítása nyomában már a közeljövőben is biztos és hasznos eredményekként jelentkeznének. Két példát erre: *a)* Fentebb említést tettünk a számítógép → fényszedő szalag kapcsolatáról. Ez természetesen fordított irányban is áll. Mai publicisztikai anyagot, mai és korábbi klasszikus műveket alig akarunk külön a magunk számára rögzíteni. Egyszerűen át kívánjuk venni az egyébként úgyis kidobásra kerülő fényszedő szalagokat és azokat konvertálni saját céljainkra, ezzel enyhítve az adatrögzítéssel kapcsolatos munkaerőgondokat, növelni az egész munkálat sebességét. *b)* Ahogy az anyag gyűlik, annak egy-egy valamilyen szempontból koherens része alapján (például: mai publicisztika, XVIII. századi szépirodalom stb.) megfelelő program segítségével kevés emberi munkaráfordítással elkészíthető e részlet gyakorisági szótára. Vagyis egy olyan szótár, amelyben a szavak előfordulásuk gyakoriságának feltüntetésével vannak felsorolva, esetleges egyéb — szükségesnek ítélt — információkkal együtt. Az ilyen szótáraknak mind elméleti, mind gyakorlati jelentősége kétségtelen; ezek mint a „nagy mű", a számítógépes adattár előzetes hasznos melléktermékei jelentkeznek.

*

Nyelvtudományunk az eltelt évtizedek alatt szép, nagy munkákkal szolgálta a magyar és a nemzetközi tudományt, a hazai közművelődést, s készít újakat: szótárakat, grammatikákat, kézikönyveket. És újszerűeket, amilyeneket eddig a nyelvészektől el sem vártak: a magyar és az orosz beszédszintetizátort. Nem szeretnénk méltatlanok lenni ezekhez a hagyományokhoz, ehhez a jelenhez.

IRODALOM

[1] R. HUTÁS MAGDOLNA: Az Akadémiai Nagyszótár történetének vázlata (1898—1952). — *Nyelvtudományi Közlemények.* 75. 2. 447—485 l. (1974)
[2] GÁLDI LÁSZLÓ: Mutatvány A Magyar Irodalmi Nyelv Nagyszótárából. — *Magyar Nyelvőr.* 84. 2. 182—96 l. (1960)
[3] *A magyar nyelv értelmező szótára.* I—VII. Budapest; 1959—62.
[4] BENKŐ LORÁND (főszerk.): *A magyar nyelv történeti-etimológiai szótára.* I—III. Budapest, 1967—76.