PINTÉR TIBOR

Amit a modern nemzeti korpuszokról tudni kell[1]

TIBOR PINTÉR
801.8:519.766
WHAT SHOULD BE KNOWN ABOUT THE NATIONAL CORPUSES

Definition of corpus linguistics. Research areas and tools of corpus linguistics researches. The issue of representativeness. Definition of the corpuses' representativeness according to Douglas Biber. Quality and quantity of materials involved in the corpus. Computer processing on the basis of materials from the Internet coded in HTML format. Characteristics of the Hungarian Word-source in Slovakia. The use of corpuses in linguistics and education.

I

A nyelvészettudomány állandó fejlődésében a régitől új szemléletet hozó paradigmaváltások mellett időnként, külső hatásra (új eszközök, lehetőségek: pl. számítógép) a már létező áramlatoktól lényegében független új diszciplínák is létrejönnek. Ilyen új nyelvészeti ág többek között a számítógépes nyelvészet és a *korpusznyelvészet* is. Tanulmányomban kísérletet teszek a korpusznyelvészet rövid bemutatására, valamint áttekintem, hogyan készül a Magyar Nemzeti Szövegtár szlovákiai magyar anyaga, amelynek összeállításán a Gramma Nyelvi Iroda munkatársai dolgoznak.

A latin eredetű *korpusz* (corpus = test, törzs; összesség, gyűjtemény [Gyökösy 1989]) szó a magyar nyelvű terminológiában az angol nyelven keresztül honosodott meg (corpus, tsz. *corpora* vagy *corpuses).* Hagyományos felfogásban (írott) szövegek halmazát jelenti, ám a modern nyelvészetben ehhez az alapjelentéshez sajátos kiegészítő jelentések is kapcsolódnak. A számítógépes adatfeldolgozás elterjedése miatt újabban korpusznak csak az olyan szövegek gyűjteményét nevezik, amely már előzőleg számítógépes feldolgozáson ment keresztül (a számítógépes feldolgozás folyamatára a későbbiekben még kitérek) (vö. pl. Šulc 1999, 9–10; Čermák 1995, 119; Váradi 2000, 263).

II

A korpusznyelvészet az a nyelvészeti diszciplína, amely rendszeresen és rendszerszerűen foglalkozik a nyelvi korpuszokkal, valamint az azokat tároló és feldolgozó eszközökkel, illetve a nyelvi rendszerek és nyelvi funkciók jobb megismerése céljából vizsgálataiban olyan eszközöket használ, amelyekre ez idáig nem volt lehetőség (vö. Čermák 1995, 121). Egy másik megfogalmazás szerint „a korpusz-alapú nyelvészet az empirikus vagy más szóval adat-intenzív nyelvészetnek azon ága, amely számítógépen tárolt, számítógépes kereséseket lehetővé tevő, strukturált szövegegyüttesen alapszik" (Reményi, megjelenés alatt). A két definícióból kitűnik, hogy

olyan nyelvészeti ágról van szó, amely vizsgálati eszközei révén szoros kapcsolatban van a számítógépes nyelvészettel. František Čermák cseh korpusznyelvész szerint a két diszciplína közötti különbségek főleg a módszerekben és az eszközökben vannak, a kutatások kiindulópontja mindkét esetben megegyezik – ez a számítógép (Čermák 1995, 121).

A korpusznyelvészetet adatorientáltsága és adatfeldolgozásának módszerei egyértelműen az empirikus nyelvészetbe sorolják. Az adatok esetlegessége, a kapott eredmények megkérdőjelezhetősége csökken, illetve megszűnik, hiszen a korpusznyelvész eredményeit minden esetben (nagy mennyiségű) adattal tudja alátámasztani. Az ilyen alapon nyugvó kutatások eredményei megbízhatóbbak, hiszen azok minden esetben *konkrét* írott vagy elhangzott (a beszélt szövegek is írott formában kerülnek feldolgozásra) szövegeken alapszanak. Az eredmények adekvátsága természetesen itt is az anyagmennyiség nagyságával azonos arányban növekszik. Teljességgel megbízható eredményt csak nagy korpusz tud felmutatni, viszont azt is érdemes szem előtt tartani, hogy a különböző nyelvészeti kutatásokhoz szükséges korpuszok nagysága különbözhet. Az anyagmennyiség nagysága azonban állandóan növelhető, mivel a mai, nagy teljesítményű, gyors számítógépeknek a több száz millió szavas korpuszok tárolása sem okoz gondot, s a bennük történő keresés is másodpercek, percek alatt elvégezhető. Ilyen háttérrel a leíró nyelvészeti diszciplínák és a szociolingvisztika is nagyobb eredményességgel dolgozhat. Nagy mennyiségű anyagon ugyanis a morfológiai vagy szintaktikai vizsgálatok biztosabban végezhetők el (ezekhez ma már számítógépes programok is készültek), de egy kellőképpen strukturált korpusz a beszéd normáinak vizsgálatában is nagy segítség lehet (vö. Štícha 1994). Az élőnyelvi, illetve nyelvrendszerbeli vizsgálatok mellett az sem lehet mellékes, hogy a jövőben napvilágot látó nyelvtankönyvek példamondatai, nyelvtani szerkezetei nemcsak a gondosan szerkesztett irodalmi művekből, hanem az élő nyelvből is átvehetők lesznek. Nem szabad elfelednünk azonban, hogy „egy ilyen korpusz nem végcél, hanem eszköz, amely adatokat szolgáltat a beszélőközösség szintjén érvényes nyelvi rendszer szabályainak megfogalmazásához. Ez utóbbi, azaz az X nyelv grammatikája vezet el a ténylegesen előfordultakon túl a lehetséges esetekről számot adó leíráshoz" (Váradi 2001, 1286).

A számítógépes feldolgozást igénybe vevő korpusznyelvészet kezdetei az 1960-as évek elejére esnek. Egyes adatok szerint szövegek elektronikus adatbázisának létrehozását Paul Imbs már 1957-ben szorgalmazta (Klímová 1994, 256). Ez persze nem jelenti azt, hogy az 1960-as évek előtt élő nyelvészek nem dolgoztak volna különböző célokra összegyűjtött szövegekkel, korpuszokkal, csupán ezek gyűjtése, feldolgozása kézi erővel, nem pedig számítógépekkel történt. A korpuszok első felhasználói valószínűleg a lexikográfusok voltak (Šulc 1999, 28), akik szótáraik elkészítéséhez nagy mennyiségű „preparált" szöveget használtak fel.[2] Azonban nemcsak a lexikográfusok, hanem a diakronikus nyelvállapottal foglalkozó más nyelvészek munkája sem képzelhető el összegyűjtött szövegek vizsgálata nélkül, így természetes, hogy a „korpuszokkal" dolgozó nyelvészetnek nagy hagyománya van. A korpuszok jelentőségét csak a generatív nyelvelmélet elterjedése után vonták egy időre kétségbe.

Az első nagyobb, nem számítógépes korpuszok közé tartozik az Oxford English Dictionary (OED), amelynek 1928-ban megjelent kiadása például 414 825 címszót

tartalmazott, ami 50 milliós szóanyagnak3 felel meg. A számítógépes korszak előtti idők legjelentősebb korpusza a Survey of English Usage (SEU) Corpus, amelyet elsődlegesen az angol nyelv grammatikájának tanulmányozására hoztak létre (természetesen ma már létezik számítógépes formában is).

A korpusznyelvészet átértékelését az 1961-ben megkezdett és 1964-ben publikált Brown Corpus („Brown University Standard Corpus of Present-Day Edited American English") idézte elő. A Brown Corpus volt az első számítógéppel összeállított elsődlegesen nyelvészeti célokra készített korpusz. Végső formája mintegy 1 014 312 szót tartalmaz, amit 500 darab átlagosan 2000 szót tartalmazó amerikai angol nyelven írott összefüggő szöveg alkot. A Brown Corpus a későbbiekben kidolgozott szerkezete, nagysága és anyaga miatt valamilyen formában több korpusz mintájául szolgált (a felsorolástól most eltekintek).

Az 1980–1990-es évekig készült korpuszok a nyelvészeti kutatások számára új lehetőségeket nyitottak. Ekkor a kisebb korpuszok mérete már nem volt elég a különböző kutatások számára, s nyilvánvalóvá vált, hogy megbízható kutatásokat csak nagyobb korpuszokon lehet végezni (Šulc 1999, 35). A korpusz méretét egyszerűen úgy határozhatjuk meg, mint az azt alkotó részek (szavak) összességét (www.ilc.pi.cnr.it/EAGLES96/corpustyp/node11.html). A kisebb korpuszok csupán egyes nyelvi jelenségek vizsgálatára elegendőek. A nagy korpuszok időszakát a John Sinclair által vezetett projekt, a COBUILD Corpus („Collins Birmingham University International Language Database") kezdte el. Ez egy új angol szótár készítése kapcsán készült, amelyet a Collins Kiadó és a birminghami egyetem közösen állított össze. Szóanyaga az 1960-as évektől gyűjtött nem tudományos írott és beszélt nyelvi (a beszélt nyelv a korpusz 25%-át teszi ki) szövegeket tartalmaz. Sinclairék a COBUILD Corpust tovább bővítették, és létrehozták a Bank of English (BoE) korpuszt, az első nem zárt, anyagában állandóan bővülő *(monitor corpora)* korpuszt (az interneten lévő anyag szerint 2002 januárjában 450 millió szót tartalmazott). Nem sokkal a BoE után három kiadó, két egyetem és egy könyvtár támogatásával létrehozták a British National Corpust (BNC): a korpusz 4124 modern brit angol írott és beszélt szöveget tartalmaz, ami hat és negyed millió körüli mondatot, azaz 100 milliónál is több szót tartalmaz.

A korpuszok elkészítését hosszas tervezés folyamata előzi meg. Mielőtt a korpusz struktúrája elkészülne, a szerkesztőknek át kell gondolniuk, hogy a végleges strukturált elektronikus szövegtár milyen célt szolgál majd (például egy nagyszótár alapját képezi-e majd, vagy morfológiai vizsgálatok anyaga lesz). Továbbá még a tervezés első fázisában el kell dönteni, hogy a születendő korpusz milyen mennyiségű anyagot tartalmazzon, illetve hogy a korpusz zárt *(referenciakorpusz)* vagy nyílt, azaz állandóan bővülő *(monitorkorpusz)* legyen-e. A referenciakorpuszok (ilyen például a BoE, BC) általában előre meghatározott nagyságúak és struktúrájúak, tehát általában statikusak. Céljuk, hogy elégséges mennyiségű[4] anyagot tartalmazzanak az alapvető lexikológiai és megbízható grammatikai vizsgálatok számára. Mivel statikus, anyagukban nem változó korpuszokról van szó, ezért megfelelnek a párhuzamos korpuszok követelményeinek. A párhuzamos korpuszok olyan két- vagy többnyelvű korpuszok, amelyben egy mű és annak egy vagy több nyelvre lefordított változatai szerepelnek, így a fordításelméleti munkákban nagy jelentőségűek (a párhuzamos korpuszról lásd pl. Váradi 2002a). A monitorkorpuszok az előzőektől eltérő-

en dinamikusak, folyamatosan bővítettek, így akár több száz millió szót is tartalmazhatnak. Mivel a monitorkorpuszok a referenciakorpuszokhoz viszonyítva nagyobbak, ezért a referenciakorpuszokon elvégezhető vizsgálatok a monitorkorpuszokon megbízhatóbb minőségben vihetők végre (www.ilc.pi.cnr.it/EAGLES96/corpustyp/node1.html).

III

Mivel a korpuszok a nyelv egészére érvényes vizsgálatokat tesznek lehetővé (ez az elsődleges céljuk), ezért a velük szemben elsődlegesen elvárható tartalmi és formai követelmény a *reprezentativitás.* A korpusznyelvészet fejlődésével a reprezentativitás fogalma is változik, módosul (Čermák–Králik–Kučera 1997, 117). A kezdetleges, mai mércével nézve kisebb korpuszoknál a reprezentativitás fogalmát „bizonyos optimális változattal (csakis a megfelelő, sőt ideális változattal) hozták összefüggésbe" (Čermák–Králik–Kučera 1997, 117). Ez azt jelentette, hogy azok a korpuszok számítottak reprezentatívnak, amelyek a lehető legtöbb szót tartalmazták, és struktúrájuk a lehető legtöbb regisztert tartalmazta. Ma az élő nyelvvel foglalkozó nyelvészek szemében az „ideális" jelző negatív jelentéstartalmú, egy olyan állapot jelzője, amelyet a változó nyelv soha nem tud elérni, csak megközelíteni. Mivel a nyelv állandóan változik, ezért korpuszokkal soha nem leszünk képesek lefedni az „ideális nyelvi nagyságot" (még a monitorkorpuszokkal sem). Ezért ma már általánosan elfogadott tény, hogy a korpuszok nem lehetnek abszolút értelemben reprezentatívak, így esetükben a reprezentativitás statisztikai értelemben vett reprezentativitást jelent, azaz a reprezentativitásnak az adott közösség, populáció összetettségét, annak elvárásait kell tükröznie (vö. Reményi, megjelenés alatt; Bieber 1993; Čermák 1995, 124–125; Váradi 2000, 266, 2001, 1286). A korpusz egyes részeinek olyan arányban kell szerepelnie, ahogy az a valóságban létezik, illetve ha ez nem lehetséges, akkor legalább ennek az állapotnak az elérésére kell törekedni. A demográfiai statisztikák mellett a reprezentativitást a szövegek recepciója (kiadói oldal: kiadási lista, Books in print, kurrensperiodika-lista, tehát egy szűkebb nyelvi közösség produktumai) és percepciója (befogadói oldal: bestseller listák könyvtári kölcsönzési statisztikák, periodikák olvasottsági statisztikái) is befolyásolja. A reprezentativitás megközelítésénél mindkét oldalt egyaránt figyelembe kell venni, s a korpuszok kialakításánál meg kell keresni a két oldal közötti megfelelő arányt. Biber 1993-as cikkében a receptív és perceptív oldal mellett külső *(external criteria)* és belső *(internal criteria)* kritériumokról is beszél. A belső kritériumokat nyelvészeti (nyelvi szempont, a szöveg formalitása stb.), a külső kritériumokat nem nyelvészeti kritériumokként (nem nyelvi szempont, a szöveg tipológiáját érintő szempontok: pl. eredet, műfaj, szituáció, idő stb.) határozza meg (Biber 1993, 245).

A korpusz reprezentativitását érintő nézetek nagyon változatosak. Bizonyos nézetek szerint a reprezentativitás rétegzett mintavétellel biztosítható, megközelíthető. Ezt a mintavételt választották például a budapesti szociolingvisztikai interjú készítői is, azaz esetükben is a minta a valóságot tükröző arányokban szerepelt. Biber, a reprezentativitás egyik nagy szakértője 1993-as cikkében éppen ennek ellenkezőjét hangsúlyozza, amikor azt mondja, hogy „az arányos minták csak abban az értelemben reprezentatívak, hogy hűen tükrözik a nyelv regiszterei közötti gyakorisági arányokat – nem reprezentálnak azonban számokban nem kifejezhető relatív

fontosságot" (Biber 1993, 247–248[5]). Biber elveti a rétegzett mintavételen alapuló korpuszokat, mivel az ilyen korpuszok szerinte nem tükrözik „reprezentatívan" a nyelvi változatokat, mert így a korpuszba számos olyan szövegtípus nem kerülne be, amelyeknek a mindennapi életben fontos szerepük van (pl. államszerződések, törvények, biztosítási kötvények vagy bármilyen ritkán olvasott könyv). Biber a rétegzett mintavétel helyett a mintavétel alábbi hierarchikus rendszerezését ajánlja (Biber 1993, 245[6]):

1. Közeg	írott/beszélt/felolvasott
2. Közreadás formája	kiadott/nem kiadott
3. Beszédhelyzet	intézményes/egyéb nyilvános/ magán/személyes
4. Címzett	
a) száma	tömeges/többes/egyéni/saját
b) jelenléte, azaz idő és hely	jelen van/nincs jelen
c) részvétel	nincs/kicsi/intenzív
d) közös tudás	általános/szakmai/egyéni
5. Közlő	
a) demográfiai változók	nem/kor/foglalkozás stb.
b) elismertség	elismert egyén/intézmény
6. Tényszerűség	tényszerű/informatív/köztes/fikció
7. A közlés célja	meggyőzés, szórakoztatás, tájékoztatás, irányítás, magyarázás, elbeszélés, leírás, feljegyzés, önkifejezés stb.
8. Téma	…

A korpuszok tervezésénél Biber azért sem tartja elfogadhatónak a reprezentativitás arányosságra épülő fogalmát, mivel szerinte az ilyen, a valóságot mintázó reprezentatív korpusz durván 90 százaléka konverzáció lenne, 3 százaléka levél és feljegyzés, míg a fennmaradó 7 százalék tartalmazná a többi beszédstílust (beleértve a különféle újságokat, cikkeket, akadémiai székfoglalót, kiadatlan írásokat stb.) (Biber 1993, 247), az ilyen korpusz pedig nem biztosít a különböző nyelvészeti vizsgálatok számára elegendő nyelvi változatosságot. Szerinte a fent ismertetett összetételből a konverzáción kívüli 10 százaléknyi szöveg az érdekes, mivel ez tartalmazza a nyelvi változatok széles skáláját. Biber értelmezésében tehát a korpusz reprezentativitása megváltozik: nem az a cél, hogy a minta visszaadja a valóságban észlelt arányokat, hanem hogy a korpusz minél szélesebb rétegben tartalmazza (reprezentálja) a nyelvi változatokat. Ez esetben a korpusz célja, hogy minél több nyelvi változatot gyűjtsön össze, így azonban az összegyűjtött anyagon belüli strukturálás kérdése nincs megoldva.

A szövegek proporcionális reprezentáltsága mellett tehát – ahogy azt már az előző bekezdésben említettem – fontos kérdés a tematikus reprezentáltság is, azaz nemcsak az fontos, hogy *mekkora*[7] legyen a korpuszba kerülő minta, hanem hogy *mi* kerüljön a korpuszba. Ez esetben elsőrangú feladat eldönteni, hogy milyen célt szolgál majd a korpusz, hiszen a vizsgálat milyensége meghatározhatja a korpuszba kerülő anyagokat. Így például a publicisztikai nyelvet vizsgáló korpuszba eleve

nem kerül bele például a helyi pékség alkalmazottai között folyó vita szövege, míg a nagyszótári korpuszban, amelynek célja egy nyelv szótári anyagának összeállítása, ilyen minta is elfogadható. A „pékség dolgozóinak vitája" felvet egy további kérdést, mégpedig azt, hogy a korpuszokban szereplő anyagban a beszélt és írott nyelv milyen arányban legyen képviselve. A korpusznyelvészek általában elvetik annak lehetőségét, hogy a „beszélt és írott nyelvi regiszter- és műfajvariabilitás eloszlása felmérhető lenne (pl. Biber 1993, 247; Reményi, megjelenés alatt). A beszélt nyelvi korpuszok elkészítése jelenleg feldolgozásuk miatt nagyon költséges, ezért az írott nyelvet rögzítő korpuszokhoz képest jóval kevesebb van belőlük, illetve a nyelv mindkét formáját rögzítő korpuszokban az írott változathoz képest jóval kisebb arányban szerepelnek (a beszélt nyelvet feldolgozó korpuszok is természetesen megfelelő módon és technikával lejegyzett írott korpuszok). Az írott és beszélt nyelvet egyaránt tartalmazó korpuszokban a beszélt nyelv mennyisége a valósághoz viszonyítva jóval alulreprezentált (egyesek szerint a mindennapi életben létrejövő szövegek 90-95 százaléka beszélt nyelvű, és csupán mintegy 5 százaléka írott nyelvű szöveg [Šulc 2001, 53]), illetve azok a korpuszok, ahol ezek az arányok megfelelnének a valóságnak, a kevés anyag miatt még sokáig nem lesznek felhasználhatók az alapvető nyelvészeti vizsgálatok számára (Šulc 2001, 53). A korpuszok proporcionalitásáról befejezésképpen még annyit, hogy jelenleg még nem létezik olyan általánosan elfogadott belső struktúra, amelyet a korpuszok összeállításánál megnyugtatóan követni lehetne (vö. Šulc 1999, 20).

Az elektronikus rendszerekben tárolt korpuszoknak csak akkor van jelentőségük, ha felhasználásuk is elektronikus úton történik. A felhasználást segítő szoftverekhez ma már nem nehéz hozzáférni (ahogy a különböző nagykorpuszokhoz sem, mivel ezek – még ha nem egész terjedelmükben is, de – megtalálhatók az interneten). Mivel a korpuszok eleve számítógépes feldolgozáson mennek keresztül (annotáció), s a szövegek minden esetben preparáltak (kódokkal ellátottak – *tagging),* ezért számítógépes keresőprogramok, valamint más (nyelvi) elemzőprogramok számára könnyen kezelhetők, a különböző munkálatokat bennük megfelelő programokkal mindenki problémamentesen elvégezheti. A számítógépes felhasználás eszközeivel, azok működési elveivel, illetve az ilyen programok megalkotásának nehézségeivel most nem foglalkozom, mindössze annyit említek meg, hogy Magyarországon ilyen jellegű angol és magyar nyelvű programok készítésével a MorphoLogic Kft. foglalkozik (ők szerkesztették többek között a Windows Word magyar nyelvű helyesírás-ellenőrző programját).

IV

A korpuszok gyakorlati jelentőségét felismerve (különböző nyelvészeti és nem nyelvészeti kutatások anyagaként egyaránt használatosak) az 1990-es években az angol nyelvű korpuszokon kívül más nemzetek is megalkották saját nemzeti korpuszaikat. A szlovák, cseh és magyar korpusz összeállítása is az 1990-es évek elején, közepén kezdődött el. Jelenleg annak ellenére, hogy sorra jönnek létre az egyes nemzeti korpuszok, még mindig az angol nyelv rendelkezik a legtöbb, leggazdagabb és legjobban strukturált korpuszokkal, ismereteim szerint több mint hússzal.

A Magyar Nemzeti Szövegtár[8] (MNSZ) munkálatai 1998 elején kezdődtek el a Magyar Tudományos Akadémia Nyelvtudományi Intézetének Korpusznyelvészeti Osz-

tályán, amely 1997 elején alakult meg. A nagyszabású munkálatokat Váradi Tamás vezeti. A korpusznyelvészeti osztály célja létrehozni egy reprezentatív korpuszt, amely legalább 400 millió szót tartalmazna, s amivel az MNSZ felzárkózna a jelenlegi nyugat-európai szintre. A kezdeti tervek alapján az MNSZ 100 millió szót tartalmazott volna, ám a későbbiek folyamán ez a mennyiség szerencsére jóval felülmúlhatónak bizonyult (vö. Váradi 2000, 266). Az MNSZ jelenleg mintegy 152 millió szót tartalmaz, amelynek Magyarországon kívüli anyaga csupán elenyésző mennyiségű (mintegy 1,5 millió szó, így a jelenlegi korpusz „nemzeti" megnevezése nem teljesen adekvát). Ennek forrása a szlovákiai *Új Szó* és a *Romániai Magyar Szó* internetes anyaga volt. Mivel a szövegtár a jelenlegi élő nyelv tára kíván lenni, ezért az alkotók igyekeztek az 1980-as évek végétől, 1990-es évek elejétől megjelent anyagokat összegyűjteni. Ez természetesen nem volt minden kategóriában lehetséges (az MNSZ felépítését lásd az első táblázatban), ezért a korpusz egyes szerkezeti egységei, alkorpuszai, a szépirodalom és kisebb mennyiségben a tudományos próza tartalmaz régebbi keletkezésű anyagokat is.

A korpusz tervezésekor a nyelvészek számára nagy kérdést jelentett, hogy a készülő szövegtárban a beszélt és írott szövegek aránya milyen legyen. Mivel a hangzó anyag lejegyzése nagyon hosszadalmas és költséges feladat, ezért a tervezők úgy döntöttek, hogy a beszélt nyelvi szövegek felvételétől eltekintenek (Váradi 2002b, 385), illetve ezt a kategóriát a már nagy részében lejegyzett Budapesti Szociolingvisztikai Interjú (BUSZI) fogja képviselni. Budapest lakosságával életkor, nem, iskolázottság és foglalkozás szerinti (lásd Kontra 1990, 7) reprezentatív mintavétellel készült BUSZI 250 adatközlőjének mintegy 600 órányi beszélt nyelvi anyaga alkotja jelenleg az MNSZ beszélt nyelvi részét (Váradi 2002b, 385).

Az MNSZ „átvette a 40 millió szövegszavas Longman Beszélt- és Írottnyelvi Korpusz (LSWE) szerkezetét, egy változással: még egy regiszter beemelésével" (Reményi megjelenés alatt). Ötödik kategóriaként a Biber által is kiemelt „hivatali nyelvet" is bevették a korpusz struktúrájába. Az MNSZ interneten található anyagában a kívánt szót az egyes kategóriákban külön, illetve az öt kategóriában egyszerre is kereshetjük.

Az MNSZ jelenlegi összetétele:

	Személyes közlés	Szépirodalom	Sajtó	Tudományos próza	Hivatali nyelv
Források	Online interaktív internetes fórumok	Digitális Irodalmi Akadémia + meglévő állomány	A korábban begyűjtött állomány	Magyar Elektronikus Könyvtár + internetes szakfolyóiratok	Minisztériumok, önkormányzatok stb. internetes portáljai
Interaktivitás	igen	csak szépirodalmi párbeszédekben	nem	nem	nem
Közös szituáció	van	nincs	nincs	nincs	nincs
Fő kommunikációs cél/tartalom	személyes	szórakozás, műélvezet	tájékoztatás, értékelés	tájékoztatás, érvelés, magyarázat	utasítás, magyarázat, tájékoztatás
Közönség	egyéni	széles körű	széles körű	szakközönség	szakközönség
Közönség az interneten	bárki	bárki	bárki	bárki	bárki
Nyelvváltozat	helyi	többnyire sztenderd	helyi vagy sztenderd	sztenderd	sztenderd

Forrás: Reményi, megjelenés alatt

Az MNSZ gyűjtési módszer sajátossága, hogy a tárolt anyag nagy része az internetről származik. Ez a későbbi számítógépes feldolgozásban nagy segítséget és egyben problémát is jelent, mivel az így begyűjtött anyag már HTML[9]-formában kódolva van.

Példa HTML-formátumú szövegre:

```
<html>
<head>
      <meta http-equiv="Content-Type" content="text/html; charset=windows-
1250">
      <link rel="stylesheet" href="/styles/main.css">
      <meta http-equiv=Refresh content=900>
<meta http-equiv=Expire content=now>
<meta http-equiv=Pragma content=no-cache>
<meta http-equiv=Content-Type content="text/html; charset=windows-1250">
<meta name=description content="Hungarian daily Uj Szo in Slovakia:
international, national and local news coverage from the newspaper, nonstop
updates, technology news, sports, reviews.">
<meta name=keywords content="daily Új Szó, international news, minorities,
newspaper, national politics, science, business, breaking news, technology,
sports, weather, film, forums, archive, tudósítás, hazai, külföldi, folya-
matos aktualizació, kultúra, chat – online interjúk, ingyenes, számítógé-
pek, internet, Szlovákia, Slowakei, Slovakia, Slovaquie, jégkorong, labda-
rúgás, vélemények, Central Europe">
<meta name=author content="Petit Press, a.s. - e-publishing division - Bit-
Media, s.r.o.">
<meta name=classification content=Media>
<meta name=distribution content=Global>
<meta name=rating content=General>
<meta name=copyright content="Petit Press, a.s.">
<meta name=language content=HU>
<meta name=doc-type content="Web Page">
<meta name=doc-class content=Published>
<meta name=doc-rights content="Copywritten Work">
<meta name=doc-publisher content="Petit Press, a.s.">
<SCRIPT src="/js/time.js" language="JavaScript1.2"></script>
      <title>ÚJ SZÓ online</title>
<!-- BS START //-->
<SCRIPT language="JavaScript">
<!-- var bsstyles, bsafterbody;
bsstyles = ""; bsafterbody = "";
//-->
</script>
<script                                                language="JavaScript"
src="http://ads.reklama.sk/ads/ads.asp?pl=168"></script>
<SCRIPT language="JavaScript">
<!-- if (bsstyles!="") { document.write(bsstyles);}
//-->
</script>
<!-- BS STOP //-->
</head>
<body    bgcolor="#FFFFFF"    leftmargin=0    topmargin=0    marginwidth="0"
marginheight="0" onLoad="WriteTime()">
<!-- BS START //-->
<SCRIPT language="JavaScript">
<!--
      if (bsafterbody!="") { document.write(bsafterbody); }
//-->
</script>
```

V

Miért jelent segítséget, ugyanakkor problémát az interneten hozzáférhető anyagok felhasználása? Mivel az internetről letölthető anyagok gyűjtése jelenleg mindenkép-

pen az anyaggyűjtés legegyszerűbb és legköltségkímélőbb módszere, nagy mennyiségű anyagok gyűjtésekor mindenképpen ez kívánkozik a legkedvezőbb lehetőségnek. Ugyanakkor az interneten található anyagok feldolgozása olyan problémákat gördít a nyelvészek elé, amelyek más források felhasználásakor valószínűleg nem jelentkeznének:

– Az MNSZ számára csak a szövegobjektumok fontosak. A világhálón természetesen nemcsak szövegek, hanem képek, különböző adatlisták stb. is szerepelnek, amelyekre a korpusz elkészítéséhez nincs szükség, tehát a letöltött anyagból ezeket el kell távolítani. Ebben a HTML kódnyelv van a segítségünkre, mivel itt a különböző szövegstruktúrák egyéni kóddal vannak ellátva. A HTML-formátumban előforduló szövegek mellett más szövegfájlok (.txt, .doc, .pdf) és képként (.jpg) elmentett szövegek is találhatók az interneten, amelyek számunkra szintén fölöslegesek.

– A *.hu* tartományú, magyarországi szervereken nem minden szöveg magyar nyelvű, valamint a magyar nyelvű oldalak szövegeiben a nyelvek keveredhetnek is.

– Ugyanaz a szöveg több helyen is (esetleg más formátumban) előfordulhat.

– Az egyes honlapok a gyűjtés alatt megszűnhetnek, illetve az újságok honlapjain ugyanaz a lapszám több napon keresztül is megjelenhet, mert csak az újság fejlécét frissítik, tartalmát változatlan formában közlik.

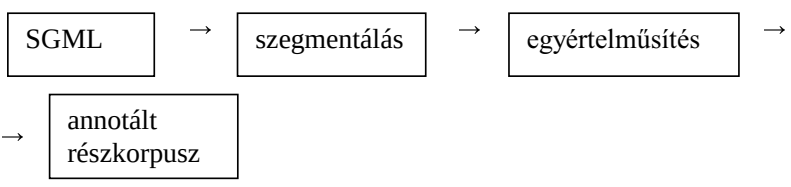
– A szövegek szerzőinek egy részét nem lehet megállapítani.

A szöveg automatikus letöltése még nem jelenti a letöltött anyag korpuszba való azonnali bekerülését. Az előbb felsorolt okok miatt az ilyen úton szerzett szövegeket egy program segítségével ellenőrizni kell, s ehhez már nem kevés emberi erőforrásra is szükség van. A HTML-formátumban begyűjtött anyagok feldolgozása a végső formáig a következőképpen alakul:

1. HTML → SGML[10]

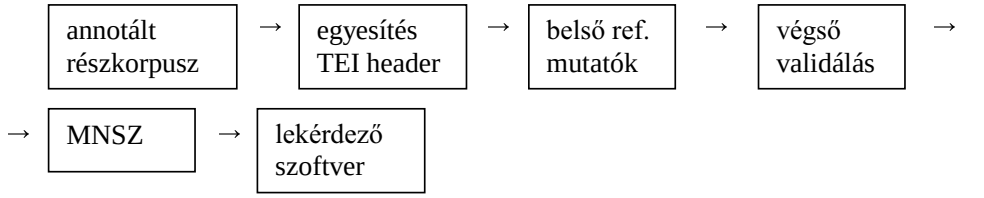
Ebben a folyamatban az internetről letöltött HTML-formátumú szövegekből el kell távolítani mindent, ami nem szöveg. Ebben segítenek a HTML-kódok, mivel azok ismeretében csak a felesleges HTML-kódokat kell a kiválasztott anyagból eltávolítani. Az így kapott HTML-formájú szöveget át kell alakítani SGML-formába, majd a nyers SGML szöveget ellenőrizni kell, hogy a szöveg szerkezete (szintaxisa) megfelel-e az előre megalkotott, definiált szerkezetnek (DTD). A validálás folyamán a már meghatározott szövegstruktúrát egyeztetik a kész SGML-formátumú szöveggel, s a még felmerülő hibákat itt kijavíthatják.

2. SGML → annotált korpusz



Ebben a fázisban az ellenőrzött (validált) SGML-formátumú szövegeket mondatokra, szavakra kell bontani, majd egy elemzőprogram[12] segítségével a morfológiai elemzést a szövegen végre kell hajtani (szegmentálás). Mivel az elemzőprogram az egyes szóalakoknak (szótőnek, lemmának, amelyet az ún. lemmatizáció során kapunk) gyakran többféle felbontását is felkínálja (pl. *szemetekkel=szemét*+PL+INS, illetve *szemetekkel=szem*+PERS-PL-2+INS), az egyértelműsítés folyamán a program kiválasztja a szövegkörnyezetnek megfelelő alakot (az egyértelműsítés folyamatára lásd Prószéky 2001, 992). Mindezen folyamatok után megkapjuk a megfelelő morfológiai kódokkal ellátott részkorpuszt. Az egész folyamat talán legnehezebb része a morfológiai elemzés, hiszen a bonyolult morfológiai rendszerrel rendelkező magyar nyelv számára egy olyan kódrendszert kell megalkotni, amelynek tartalmaznia kell az összes magyar szó morfológiai információját.

3. Annotált korpusz → MNSZ



A munkálat utolsó fázisában a kódokkal ellátott korpuszt véglegesítik, a már meglévő kódolást utoljára ellenőrzik. A kódolás folyamán a szövegek saját „fejlécet" kapnak, melyből a kódolás segítségével leolvasható a szöveg típusa, szerzője, keletkezésének időpontja, megjelenési helye stb. A szöveg minden szavát szintén saját kódokkal látják el, melyből kiolvashatók az adott szó morfológiai kategóriái.

Táblázat: Minta a Magyar Nemzeti Szövegtárból

```
<!-- HVG ./0116/0116009.htm --> <div type="article" column="unspec">
<opener> <dateline> <w lemma="HVG" msd="N.NOM" ctag="NS3NN">HVG</w>
<w lemma="2001/16" msd="DIG" ctag="Q">2001/16</w> <c lemma="."
Msd="SPUNCT" ctag="SPUNCT">.</c> <w lemma="szám" msd="N.NOM"
Ctag="NS3NN">szám</w> <date iso8601="04-21-2001"> <w
lemma="2001._április_21." msd="DATUM"
ctag="DATUM">2001._április_21.</w> </date> </dateline> </opener>
<head rend="IT" type="unspec"> <s> <w lemma="egészségügyi"
msd="A.NOM" ctag="AS_A">Egészségügyi</w> <w lemma="szigorítás"
msd="N.PL.NOM" ctag="NP3NN">szigorítások</w> </s> </head> <head> <s>
<w lemma="sok" msd="Num.NOM" ctag="Q">Sok</w> <w lemma="zseb"
msd="N.ELA" ctag="NS3NE">zsebb½ ol</w> <w lemma="vérzik" msd="V.e3"
ctag="VS3RI">vérzik</w> </s> </head> <head rend="BO" type="display">
<s> <w lemma="Alaposan" msd="Adv" ctag="R">Alaposan</w> <w
lemma="felkavar" msd="Pre.V.TMe3" ctag="@VS3PD">felkavarta</w> <w
lemma="a" msd="Det" ctag="D">a</w> <w lemma="kedély" msd="N.PL.ACC"
ctag="NP3NA">kedélyeket</w>
```

VI

A Magyarország határain kívül megjelent írásokat magyarországi nyelvészek lassan és nehezen tudnák összegyűjteni, illetve ez a feladat számukra nem kívánt munkatöbbletet jelentene, ezért az látszott célszerűnek, ha a korpuszba kerülő anyagokat a Magyarország határain kívül élő nyelvészek gyűjtik össze. Mivel az MTA tervezetében szerepelt egy-egy kutatóállomás létrehozása Szlovákiában, Ukrajnában, Romániában, Szerbiában és Horvátországban, ezért a Magyarország határain kívül megjelent szövegek összegyűjtése könnyebben megvalósítható. Az említett kutatóállomások faladatai közé bekerült az MNSZ anyagának bővítésében való segítség, ami egyrészt anyagok gyűjtésében, másrészt pedig az összegyűjtött szövegek előzetes feldolgozásában merül ki. A létrehozandó korpusz – noha az internetes korpuszban külön is kereshető, önálló nevét is megtartó alkorpusz lesz – azonban csak akkor kivitelezhető, ha mennyiségileg, szerkezetileg és formailag valamennyire igazodik az MNSZ-hez. A határon túli korpusz teljes mérete a tervek szerint legalább 15 millió szövegszó lenne, és struktúrájának valamelyest tükröznie kellene a magyar közösségek eltérő nagyságát is (a feltételes módot a határon túli MNSZ korpusz kezdeti jellege indokolja). Ennek mennyiségi vonzata a következőképpen alakul: Románia: 6 millió szövegszó, Szlovákia: 4 millió szövegszó, Ukrajna: 3 millió szövegszó, Szerbia és Horvátország: 2 millió szövegszó. Az MTA Nyelvtudományi Intézetének Korpusznyelvészeti Osztályán meghatározott szövegmennyiség természetesen csak alsó határt jelent, ennél több szövegszó összegyűjtése természetesen lehetséges.

Mivel a kutatóállomások által megszerkesztett korpusz is az MNSZ szerves része lesz, ezért annak nemcsak szerkezetében (személyes közlés, szépirodalom, sajtó, tudományos próza, hivatali nyelv), de elkészítésének módjában (kódolásában) is követnie kell az MNSZ-t, tehát a kódolás a határon túli korpuszokban is egységes. A fő struktúrán belüli belső tagolás, valamint az egyes szavak „státusa” (pl. szlovakizmus) kutatóállomásonként változhat. A gyakorlatban ez azt jelenti, hogy az egyes szerkezeti egységekben azzal a megkötéssel létrehozhatók kisebb alegységek (például a sajtón belül elkülöníthetőek az egyes regionális sajtók korpuszai), hogy a legkisebb alegység mennyiségének az egész korpusz legkevesebb 10 százalékát kell kitennie. Az egyes szavak megjelölése, „státusa” is különbözhet, hiszen pl. szlovakizmusok valószínűleg csak a szlovákiai magyar nyelvváltozatban szerepelnek, s ezeket, ha kódoljuk, külön jellel kell megjelölni. Az írott korpuszoknak kutatóállomásonként legalább 50 órányi átírt beszélt nyelvi szöveget is kell tartalmaznia. A beszélt nyelvi szövegek gyűjtéséhez és lejegyzéséhez szükséges digitális diktafont, illetve a számítógépes adatként tárolt élőnyelvi szövegek lejegyzését segítő berendezést az MTA biztosította minden kutatóállomás részére.

Az MNSZ-ben szereplő Magyarországon kívüli korpuszok elvileg tartalmazhatnak szlovák, román stb. nyelvű szavakat és szövegeket is, amennyiben ez is a határon túli magyar nyelvváltozat része, esetleg a begyűjtött sajtótermékekben a két nyelv keverve szerepel. Ilyen problémával az MNSZ készítői nem találkoztak, ezért ez a kérdés még nem megoldott, ez majd a gyakorlat folyamán kristályosodik ki (természetesen ebbe a magyarországi oldalnak is lehet még beleszólása). A másik megoldatlan kérdés a párhuzamos korpuszok kérdése: ez szintén a sajtó kapcsán merülhet fel,[13] ott, ahol egy újságban ugyanaz a szöveg két nyelven is előfordul. Az ilyen

korpuszoknál az összevethetőség kedvéért még a bekezdéseknek is egyezniük kellene, mivel a párhuzamos korpuszok felhasználhatóságának csak így van értelme. Ha ilyen jellegű korpusz ki is alakulna, mindenképpen külön kategóriaként kellene kezelni.

A feldolgozás nem magyarországi nyelvészekre háruló része az internetről letöltött HTML-kódokkal ellátott szövegek (nyers HTML) validált SGML-kódú szöveggé történő átalakítása. Ha a szöveg forrása nem az internet, akkor a leírt szöveget a megfelelő kódokkal nekünk kell ellátnunk. A munkához szükséges felkészítést és szoftvereket az MTA Nyelvtudományi Intézetének Korpusznyelvészeti Osztálya a kutatóállomások részére bocsátotta.

VII

Az 1990-es évek végén megtervezett, a Kárpát-medencei magyarság nyelvét felölelő magyar nagykorpusz megvalósulása egyre reálisabbá válik. Az egyes kutatóállomásoknak a korpusz végső formáját 2005 végére kell elkészíteni, s további feldolgozásra az MTA Nyelvtudományi Intézetének leadni. A munkálatok már elkezdődtek, s remélem a Gramma Nyelvi Iroda beváltja a hozzá fűzött reményeket. Emellett abban is bízom, hogy az idővel a szlovákiai magyar nyelvésztársadalomban a korpuszlingvisztika is megerősödik, s elismertségben, fontosságban felzárkózik a szociolingvisztika mellé.

Jegyzetek

1. Ez a tanulmány a Domus Hungarica Scientarium et Artium Ösztöndíj támogatásával készült.
2. A *preparált* jelzővel Šulc arra kíván utalni, hogy a korpusz a szövegeknek nem csak egyszerű gyűjteménye.
3. Szóanyag alatt a korpuszban előforduló lexikai elemek összessége értendő.
4. Azt, hogy ez a mennyiség mekkora legyen, mindig a kutatás céljától függ. Természetesen egy szótár alapját képező korpusz több szót tartalmaz és más struktúrájú lesz, mint a szintaktikai vizsgálatok céljából létrehozott korpusz. A korpuszok kezdeti fázisában a legkisebb és specifikus korpuszok lehettek csupán 100 ezer szavasak is. A kezdetekkor 100 ezer szavas korpusz elegendő volt a prozódiai jelenségek vizsgálatára, 500 ezer szavas korpusz az angol nyelv morfológiájának vizsgálatára és 1-2 milliós korpusz az alapvető szintaktikai vizsgálatok elvégzésére, valamint ekkora mennyiség elegendő volt a frekvenciaszótárak elkészítésére is (Šulc 1999, 14). Természetesen ezek a korpuszok a mai nyelvészetben már nem állnák meg a helyüket.
5. Váradi Tamás fordítása (Váradi 2001, 1289).
6. Vö. magyar nyelvű fordítása: Váradi 2001, 1288–1289, cseh nyelven Čermák 1995, 124–125.
7. Manapság az nem is igazán kérdéses, hogy *mekkora* legyen a korpusz, hiszen az ezt leginkább befolyásoló tényező a korpuszt tároló számítógép(ek) kapacitása már lassan a végletekig bővíthető, így a készülő korpuszokat leggyakrabban monitorkorpuszoknak tervezik.
8. Az MNSZ megindítása előtti előzmények közül mindenképpen említésre méltó a Papp Ferenc vezetése alatt az 1960-as években működő debreceni iskola tevékenysége, to-

vábbá az 1980-as évek végén kiadott *A magyar nyelv szépprózai gyakorisági szótára* (Füredi–Kelemen 1989), valamint az ún. akadémiai nagyszótár. A magyar irodalmi és köznyelv nagyszótárának munkálatai 1984 végén indultak meg, s az első mintegy tíz év fő feladatául a már meglévő anyagok számítógépes feldolgozását tűzték ki (Pajzs 1997, 289).

9. HTML: Hyper Text Markup Language. Az interneten található fájlok formanyelve.
10. SGML: Standard Generalized Markup Language. 1986-tól a korpuszok szintaktikai formanyelve (ISO 8879).
11. A feldolgozást érintő ábrák forrása Váradi 2003.
12. Az morfológiai elemzés a MorphoLogic Kft. által tervezett HUMOR program segítségével történik.
13. Ez felmerül a szépirodalmi művek esetében is, de jelenleg ilyen párhuzamos korpuszt még nem tervezünk.

Irodalom

Biber, Douglas 1993. Representativeness in corpus design. *Literary and Linguistic Computing,* 8, 243–257.

Čermák, František 1995. Jazykový korpus: Prostředek a zdroj poznání. *Slovo a slovesnost,* 56, 119–140.

Čermák, F.–Králik, J.–Kučera, K. 1997. Recepce současné češtiny a reprezentativnost korpusu. *Slovo a slovesnost,* 2, 117–124.

Füredi Mihály–Kelemen József 1989. *A mai magyar nyelv szépprózai gyakorisági szótára 1965–1977.* Budapest, Akadémiai Kiadó.

Gyökösy Alajos (főszerk.) 1989. *Latin–magyar szótár.* Budapest, Akadémiai Kiadó.

Klímová, Jana 1994. Francouzský textový korpus a systém elektronických slovníků. *Slovo a slovesnost,* 55, 295–300.

Kontra Miklós 1990. A budapesti köznyelvi vizsgálatokról. In: Balogh Lajos–Kontra Miklós (szerk.): *Élőnyelvi tanulmányok.* Budapest, Magyar Tudományos Akadémia Nyelvtudományi Intézete, 3–9. /Linguistica, Series A, Studia et dissertationes 3./

Pajzs Júlia 1997. Milyen szótár készíthető a nagyszótári korpuszból? In: *Szavak – nevek – szótárak. Írások Kiss Lajos 75. születésnapjára.* Budapest, A Magyar Tudományos Akadémia Nyelvtudományi Intézete.

Prószéky Gábor 2001. A nyelvtechnológia és a modern nyelvészet viszonyáról. In: *Szavak – nevek – szótárak.* I. m.

Reményi Andrea Ágnes (megjelenés alatt). *Tervezési megfontolások a Magyar Nemzeti Szövegtár számára.*

Šulc, Michal 1999. *Korpusová lingvistika. První vstup.* Univerzita Karlova v Praze. Praha, Nakladatelství Karolinum.

Šulc, Michal 2001. Tematická reprezentativnost korpusů. *Slovo a slovesnost,* 62, 53. skk.

Štícha, František 1994. Čas korpusové lingvistiky. *Slovo a slovesnost,* 55, 141–145.

Váradi Tamás 2000. Szótár, korpusz – magyar nemzeti szövegtár. In: Gecső Tamás (szerk.): *Lexikális jelentés, aktuális jelentés. Segédkönyvek a nyelvészet tanulmányozásához IV.* Budapest, Tinta Kiadó, 2000.

Váradi Tamás 2001. A nyelvhasználat empirikus vizsgálatáról. In: Andor József–Szűcs Tibor–Terts István (szerk.): *Színes eszmék nem alszanak... Szépe György 70. születésnapjára.* Pécs, Lingua Franca Csoport.

Váradi Tamás 2002a. Kontrasztív szemantikai kutatások párhuzamos korpusz segítségével. In: Gecső Tamás (szerk.): *Kontrasztív szemantikai kutatások. Segédkönyvek a nyelvészet tanulmányozásához XI.* Budapest, Tinta Kiadó, 2002.

Váradi Tamás 2002b. *The Hungarian National Corpus. LREC 2002. Third International Conference on Language Resources and Evaluation.* Las Palmas de Gran Canaria, Spain.
Váradi Tamás 2003. (Előadás.) *Kárpát-medencei szövegtár.*
Váradi Tamás (Kézirat.) *A Magyar Nemzeti Szövegtár munkálatairól.* Budapest, Magyar Tudományos Akadémia Nyelvtudományi Intézete.
www.ilc.pi.cnr.it/EAGLES96/corpustyp/node1.html
www.ilc.pi.cnr.it/EAGLES96/corpustyp/node11.html

Tibor Pintér

What should be known about the national corpuses

The corpus linguistics systematically and regularly deals with linguistic corpuses and with the tools that store and process them, as well, and during the examinations in order to recognise linguistic systems and linguistic functions better, and it also uses such tools that have been impossible before because of the underdevelopment of computing technology. Computational linguistics is the closest to corpus linguistics, we can say that corpus linguistics forms a boundary to computational linguistics and description linguistics, or social-linguistics.

The principal role of corpuses is to be a sample for descriptive and living language researches, thus the most important requirement towards their content and structure is to be representative, i.e. from the contextual and structural point of view the corpuses have to be as real as possible. Beside the quality of the material the quantity of materials involved in the corpus is also an important issue. This can vary according to the goal of corpuses, although the thesis that the corpuses should include the possibly highest amount of materials is very frequent.

Designers of corpuses provide processing of more hundred millions of words with the help of computers. This is made possible with the Internet, since there the materials are already in HTML format. The processors of the Hungarian Word-source in Slovakia also chose this format.

The corpuses can be used not only in linguistics, but also in a number of other scientific fields (according to some of the linguistics, everywhere where there is a work with words), like in education. The author hopes that corpus-oriented linguistics will be applied in Hungarian science in Slovakia, too, and that the opportunities given by the corpuses will be more widely used in the future. The most contributing would be using it in education.