REMÉNYI ANDREA ÁGNES

Tervezési megfontolások a Magyar Nemzeti Szövegtár számára

Az alábbi dolgozat két részből áll. Az első részben felvázolom a számítógépes nyelvi nagykorpuszok összeállításánál felmerülő legfontosabb elméleti kérdéseket — referencia-korpuszt és optimális kutatási helyzetet feltételezve. A második részben a Magyar Nemzeti Szövegtár (MNSz) számára — moduláris korpuszt feltételezve és a MNSz jelenlegi anyagi korlátait figyelembe véve — az adott helyzetben alkalmazható megoldásokat vázolom, melyek e dolgozat megírásával egy időben testet is öltöttek a MNSz első változatában.

A Magyar Nemzeti Szövegtár[1] (MNSz) alapvető, hosszú távú célja az, hogy olyan nagyméretű, szinkrón, általános számítógépes szövegkorpuszt hozzon létre, amely a lehetőségekhez mérten minél megbízhatóbban „reprezentálja" a mai magyar nyelv teljességét, hogy ezzel általános alapot, mintegy „nyersanyagot" teremtsen a korpusz-alapú[2] magyar nyelvészeti munkálatok számára.

A hosszú távú tervezés kedvéért a dolgozat első felében a számítógépes korpusz tervezésekor felmerülő általános, elméleti és gyakorlati kérdéseket veszem sorra. Érintem a szövegpopuláció, a mintavétel, a minta struktúrájának és méreteinek problémáját. Ebben a részben új javaslatokkal állok elő a populáció felmérésének, illetve a korpusz felépítésének egy-egy kérdésében.

1 ÁLTALÁNOS KÉRDÉSEK

Egy számítógépes nyelvi korpusz körültekintő megtervezése azért fontos, mert ezektől az összeállítás korai szakaszában meghozott döntésektől függ, hogy később a komoly emberi és pénzügyi ráfordításokat igénylő szöveg-begyűjtésen és elő-elemzésen (például morfológiai, szintaktikai, intonációs, stb. kódolás) alapuló elemzések milyen szintű általánosításokat engednek majd meg a korpuszt kutató nyelvészek, például leíró nyelvészek, szociolingvisták, diskurzuselemzők, finnugristák, lexikográfusok, a nyelvoktatás vagy a nyelvtörténet szakemberei számára.

[1] Az MTA Nyelvtudományi Intézet korpusznyelvészeti osztályán fejlesztett MNSz-at — és így ennek a dolgozatnak a megírását is — az OTKA támogatja (T 026091). Köszönöm Váradi Tamásnak e dolgozat megírásában nyújtott segítségét.

[2] A korpusz-alapú nyelvészet az empirikus, vagy más szóval adat-intenzív nyelvészetnek azon ága, amely számítógépen tárolt, számítógépes kereséseket lehetővé tevő, strukturált szövegegyüttesen alapszik.

Egy nyelvi korpusz tervezésének legelső feladata annak eldöntése, hogy a korpuszt általános vagy specifikus célokra kívánjuk felhasználni. *Specifikus* célok lehetnek például az idegen nyelvet tanulók szövegeinek vizsgálata a nyelvoktatás jobb megtervezése, vagy különböző nyelvű jogi szövegek összehasonlítása a különböző országok jogrendszereinek jobb megértése kedvéért. Az *általános* korpuszok célja a nyelv önmagáért való jobb megismerése („language [...] for its own sake", Sinclair 1992: 380), tehát például korpusz-alapú szinkron vagy diakrón szótárak és nyelvtanok létrehozása, ill. a már létező nyelvtanok tesztelése, továbbfejlesztése, de e korpuszok felhasználhatók alkalmazott célokra is, mint a gépi fordítás vagy az ember és számítógép közti kommunikáció (beszédfelismerés és -értelmezés) fejlesztésének elősegítése. Az általános *referenciakorpuszok* statikusak, tehát előre meghatározott nagyságúak és többnyire előre meghatározott struktúrájúak, míg a *monitorkorpuszok* dinamikusak, bővítésük folyamatos. Összeállításuknál a tervezés, strukturálás, arányosság kritériuma többnyire háttérbe szorul, és a minél nagyobb szövegmennyiség begyűjtése válik a vezérlőelvvé.

Az általános szinkrón nagykorpuszok valamilyen értelemben egy nyelv teljességét kívánják megragadni. Mivel a korpuszok a korábbinál megbízhatóbb kvantitatív vizsgálatokat tesznek lehetővé, alapvető fontosságú, hogy a nyelv „egészéből" kiválasztott minta a lehető legteljesebb mértékben reprezentálja ezt az egészet. A korpusz *reprezentativitásának* kívánalma azt jelenti, hogy a minta a lehető legnagyobb mértékben közelítsen a populáció összetettségéhez

- mind szituációs tekintetben: a szövegtípusok eloszlását illetően (vagyis a minta szövegtípusok szerinti összetettsége a lehető legteljesebb mértékben képezze le a populáció szövegtípusok szerinti összetettségét),
- mind nyelvészeti tekintetben: a nyelvi eloszlásokat illetően (vagyis a minta nyelvi jellemzők szerinti összetettsége a lehető legteljesebb mértékben képezze le a populáció nyelvi jellemzők szerinti összetettségét),
- mind pedig demográfiai tekintetben: a szövegeket alkotók/befogadók eloszlásának tekintetében (vagyis a minta demográfiai összetettsége a lehető legteljesebb mértékben képezze le a populáció demográfiai összetettségét).

Abban egyetérthetünk Biberrel (1993: 243), hogy a nyelvi eloszlások összetettsége valószínűleg csak annyiban fogja megközelíteni a populáció összetettségét, amennyiben a szövegtípusok összetettsége megközelíti azt.

A reprezentativitás igénye azonban komoly nehézségeket jelent a tervezésben. Sokak számára elfogadott tény, hogy egy nyelv teljes szinkrón szövegállományát nem lehet feltérképezni, hiszen mindannyian folyamatosan állítunk elő és fogadunk be beszélt- és írottnyelvi szövegeket. Meg kell találnunk azokat az eszközöket, amelyekkel egyrészt fogódzkodókat kapunk a „mai magyar nyelv minden szövege" körülhatárolásának tekintetében (ez a körülhatárolt összesség a *célpopuláció*), másrészt azokat, amelyekkel ebből a lehető legmegbízhatóbb módon tudunk mintát venni.

1.1 A célpopuláció

A célpopulációról való döntéssel jelöljük ki tehát a magyar nyelv korpusz-specifikus határait. A szövegpopuláció határainak kijelölésekor ugyanis bizonyos szövegek bekerülnek a populációba, más szövegeket kizárunk. Ha a mai magyar nyelv teljességét kívánjuk megragadni, meg kell

fontolnunk például a következőket: a Magyarországon elhangzott/leírt/kiadott (esetleg meghallott/olvasott) magyar nyelvű szövegeket tekintjük-e a célpopulációnak, vagy a bárhol magyar nyelven elhangzott/leírt/kiadott (esetleg meghallott/olvasott) szövegeket? Az utóbbi esetben csak a magyar anyanyelvűek szövegeit vesszük figyelembe, vagy nem alkalmazzuk ezt a szűkítést? Teszünk-e megkötéseket a szövegek létrehozóinak életkora tekintetében, például csak felnőtt beszélők szövegeit gyűjtjük, vagy gyermekekéit is?

Jelentős probléma annak eldöntése is, hogy milyen időhatárokat szabjunk. Dönthetünk úgy, hogy csak egy bizonyos évben vagy néhány évben létrejött szövegeket emelünk be a korpuszba. Ez helyes korlátnak tűnik, még akkor is, ha ebben az esetben esetleg kizárunk olyan korábban létrejött, de a befogadói oldalon a gyűjtés időszakában is gyakran hallott/olvasott szövegeket, mint a *Biblia*, a *Gőgös Gúnár Gedeon* vagy a gyűjtés időszakában a televízióban vetített *A tizedes meg a többiek* szövege („az oroszok már a spájzban vannak") — ha pontosan ezek a szövegek nem is feltétlenül kerülnének a véletlen választásos korpuszba, a hozzájuk hasonlóak bekerülését ezzel a korlátozással kizárjuk.

A célpopulációról való döntéskor kell azt a nehéz feladatot is megoldanunk, hogy szövegtípusok előzetes megállapításával valamilyen módon strukturáljuk-e a célpopulációt, amely struktúra a mintában is érvényesül majd.

A statisztikai mintavételt és következtetést gyakran hasznosító társadalomtudományi ágakban leggyakrabban a *véletlenválasztásos* és a *rétegzett véletlenválasztásos* mintavételi eljárásokat alkalmazzák. Az első módszer szerint a célpopuláció minden egyes tagjának teljesen azonos esélye van a mintába való bekerülésre. A második módszer szerint a kutatók előre megállapítanak néhány, a célpopulációt szerintük jelentős mértékben strukturáló alaposztályt (például nem, életkor, iskolai végzettség, településtípus; csatorna[3], regiszter, műfaj, diskurzus-funkció stb.), ezen 'rétegek' populációbeli arányait figyelembe véve állapítják meg a 'rétegek' mintabeli arányait, és ezek altípusain belül vesznek véletlen mintát. Ezért például amennyivel a nők száma nagyobb egy populációban, minden egyes nőnek annyival nagyobb esélye lesz bekerülni a mintába.

A korpusznyelvészeti vizsgálatokban is célszerű eldönteni, hogy a kétféle mintavétel közül melyiket alkalmazzuk, illetve a populáció pre-strukturálása esetén milyen 'rétegek' (pl. szövegtípusok vagy nyelvhasználók) mentén osztályozzuk a populációt, mert ezt a struktúrát örökli a minta is. Vegyük észre, hogy ez nem csupán mintavételi, hanem konceptuális kérdés is, hiszen a 'rétegeket' a kutató vagy a kutatóközösség definiálja.

1.1.1 Demográfiai vagy szövegkritériumok alapján strukturáljunk?

A 'rétegek' definiálása a *megfigyelési egységek* kettős természete miatt okoz nehézségeket a korpusznyelvészetben. Míg a statisztikai alapú társadalomtudományi vizsgálatokban általában egyértelmű, hogy kik/mik a *megfigyelési egységek* (az emberek), addig a korpusznyelvészetben ez nem így van. A nyelvhasználók (a szöveget alkotó — és esetleg befogadó — emberek) vagy a szöveg (a nyelvhasználat produktuma) a megfigyelési egység? Míg egy ország nyelvhasználói demográfiai összetételének felmérésére jól bevált módszerekkel rendelkezünk, és a felmérések adatai hozzáférhetőek (ld. népszámlálási jelentések, statisztikai évkönyvek), a nyelvhasználók által létrehozott szövegek teljességének, illetve szövegtípus szerinti összetételének felmérésére

[3] Az írott nyelv - beszélt nyelv szerinti megkülönböztetés.

nem rendelkezünk hasonló módszerrel[4]. Ez azt jelenti, hogy az utóbbi esetében kétségek merülhetnek fel a ‘rétegek’, vagyis a populáció struktúrájának előzetes megállapításával kapcsolatban, hiszen a szöveg-alapú mintavételben a ‘rétegek’ összetételének és arányainak nincs statisztikai alapja.

Ennek ellenére tudomásom szerint valamennyi létező nagykorpusz a *szövegtípus* alapján rétegzett, többnyire véletlenválasztásos mintavételi technikát alkalmazza úgy, hogy a kommunikációs helyzet nem-nyelvi sajátosságai és témája alapján meghatározott nyelvi változatok szerint csoportosít.[5] A Brown- és a LOB-korpuszban a műfaj, a Longman-korpuszban a téma, a Cobuild-korpuszban az elsődleges diskurzus-funkció (például áttekintés, folyamat, érvelés, narratíva, történeti regény, levél, hétköznapi beszélgetés) a csoportosítás elsődleges alapja, mégpedig úgy, hogy az egyes szövegkategóriák mérete — gyakorlati megfontolásokból — egymáshoz képest kiegyensúlyozott legyen: az alkategóriákon belül összehasonlítható mennyiségű szöveg gyűjtésére törekedtek. Megjegyzem, a Brit Nemzeti Korpuszban (*British National Corpus*, BNC) részben véletlenválasztásos, részben rétegzett véletlenválasztásos mintavételt alkalmaztak, az utóbbin belül többféleképpen dimenzionálták a ‘rétegeket’. (A BNC-ről részletesebben lejjebb írok.)

Tehát az általános nagykorpuszok struktúráját többnyire az összehasonlíthatóság gyakorlati követelménye és/vagy intuitív arányok diktálják, részben a reprezentativitás kizárólag szövegszempontú felfogása, részben a szövegkorpuszok tervezőinek abbéli hitetlensége miatt, hogy van-e lehetőség a szövegek csatorna-műfaj-regiszter szerinti ill. demográfiai eloszlásának felmérésére.

Ha a mintavételkor a nyelvhasználók demográfiai jellemzőiből indulunk ki, a minta nyelvi variabilitása a populációéhoz képest igen csekély lesz. Bibernek (1993: 247) *talán* igaza van, *talán* nem, amikor az alapeloszlásokat így becsüli: „az így [vagyis demográfiai alapon] tervezett korpusz durván 90% beszélgetést, 3% levelezést és jegyzeteket fog esetleg tartalmazni, a maradék 7%-on kell osztoznia például a sajtóhírek, a népszerű magazinok, a tudományos próza, a szépirodalom, az előadások, TV-műsorok és a nyomtatásban meg nem jelenő írott szövegek regisztereinek. [...] Az ilyen korpusz lehetővé tenné a korpusz által képviselt nyelv egészére vonatkozó összesítő statisztikák készítését. Az ilyen általánosítások azonban általában nem tartanak számot a nyelvészeti kutatások érdeklődésére.” [6]

Ha viszont a nyelvi variabilitást helyezzük előtérbe (ld. néhány nagykorpusz összetételét az 1. Mellékletben), többnyire háttérbe szorul a szövegek előállítóinak demográfiai összetétele — nem beszélve a befogadói oldalról. Mint azt Váradi (2001: 1289) megjegyzi, a korpuszokkal szemben igenis elvárható, „hogy mutassa fel a gyakorit, a jellemzőt, a tipikusat, a természeteset.” Összefoglalva tehát úgy tűnik, hogy a kétszintű megfigyelésiegység-

[4] Ez sem teljesen reménytelen vállalkozás. Ld. a naplómódszert az 1.2.1-2. alfejezetekben.

[5] A rétegzett mintavétel előnye, hogy kisebb mintát igényel, mint az egyszerű véletlen választás. Az elektronikus korpuszgyűjtés hagyományában a rétegzett mintavétel talán éppen ezért van a kezdetektől jelen. Johansson például kijelenti, hogy „a LOB-korpusz valódi ‘reprezentativitása’ abból a határozott erőfeszítésből ered, hogy szövegek releváns kategóriáit és alkategóriáit tartalmazza a vak statisztikai választás helyett. A véletlen választás csak azt biztosította, hogy a megállapított szempontokon belül az egyes szövegek kiválasztása mentes legyen a személyes ízlés és preferenciák tudatos vagy tudatalatti befolyásától.” (1980: 26, idézi Kennedy 1998: 28)

[6] Pedig a gyakoriság vizsgálata nemcsak összesítő statisztikák készítését jelenti. Egy kvantitatív szociolingvisztikai elemzés számára a gyakorisági összefüggések elemzése elsődleges jelentőségű. Egy, a magyart mint idegennyelvet oktató tanár vagy nyelvkönyvíró számára is fontos lehet, hogy a hétköznapi beszélt nyelvben, a spontán beszélgetésben milyen gyakorisággal fordulnak elő egyes szavak vagy nyelvtani szerkezetek.

problematika miatt vagy a szó eredeti értelmében vett reprezentativitás, vagy a diverzitás/variabilitás kritériuma szenved csorbát.

1.1.2 BNC: Demográfiai és szövegrétegzés

A szövegalapú és a demográfiai mintavétel együttes alkalmazásával a jelenleg legkorszerűbbnek tartott általános nagykorpusz, a BNC egyszerre próbálta megülni a két lovat. Az elektronikus nagykorpuszok első generációjánál szinte etalonnak tekintett egymillió szövegszavas nagysághoz képest a BNC 100 millió szövegszavas nagysága lehetővé tette, hogy számos változó mentén rétegezzék a mintát.

A BNC Burnard (1995) által ismertetett, az 1. Mellékletben táblázatos formában összefoglalt szerkezetére tekintve láthatjuk, hogy a csatorna alapvető megkülönböztetésén túl (írott-beszélt) az írott kategóriában (90 millió szövegszó) részben a véletlenválasztásos módszert alkalmazták. Itt nemcsak kiadási listákat vettek figyelembe (*Books in print*, kurrensperiodika-lista), hanem a befogadói oldalt is (bestseller listák, díjnyertes könyvek, könyvtári kölcsönzési statisztikák, periodikák olvasottsági statisztikái). Az írott szövegek másik felében szelekciós jegyek (közeg, idő, téma) szerint rétegezték a mintát.

A beszéltnyelvi minta (10 millió szövegszó) nagyobbik felét a *kontextus* alapján választották ki (informatív, hír-, üzleti, hivatalos, szabadidős kontextusokat különítettek el), melyeket 12 földrajzi régióban gyűjtöttek. A monologikus és dialogikus *műfajok* arányát is kontrollálták ebben az alkorpuszban. A beszéltnyelvi korpusz kisebbik felét pedig úgy nyerték, hogy elvileg az Egyesült Királyság népességének *demográfiai* összetételét figyelembe véve, véletlenszerű területi mintavétellel kiválasztottak beszélőket (n = 124), akik 2-7 napon keresztül hordozható magnetofonnal jártak, és lehetőség szerint minden interakciójukat felvették; a résztvevőket arra is megkérték, hogy az interakciók részleteiről készítsenek jegyzeteket. Összesen mintegy 700 órányi beszéltnyelvi szöveghez jutottak így.

Ugyanakkor „[a] résztvevőket úgy választották ki, hogy biztosítsák a férfiak és nők egyenlő arányát, az egyes korcsoportok nagyjából azonos nagyságát, és az egyes társadalmi csoportok azonos nagyságát." (BNC Online 1997) Vagyis a rétegzett véletlenválasztásos mintavétel csak a terület változójának tekintetében valósult meg, a többi alapvető demográfiai változó (nem, életkor, szocio-ökonómiai státusz) tekintetében viszont nem[7], vagyis a mintabeli korcsoportok vagy szocio-ökonómiai csoportok arányai valójában nem követték az Egyesült Királyság népességbeli arányait. A Longman Beszélt- és Írottnyelvi Korpusz (*Longman Spoken and Written Corpus*, LSWE) beszélgetési alkorpuszát szintén demográfiai alapon gyűjtötték: „az adatközlők csoportját úgy jelölték ki, hogy életkor, nem, társadalmi csoport és régió szerint az ország (az Egyesült Királyság vagy az Egyesült Államok) angol beszélőinek széles skáláját képviseljék" (Biber et al. 1999:29), vagyis a rétegzettség, de talán a véletlenszerű kiválasztás is érvényesült ebben a korpuszban (a leírás szűkszavúsága miatt ez nem nyilvánvaló).

Visszatérve a BNC-re, ennek szerkezete a fentiektől eltekintve igen meggyőzőnek tűnik, főként a korábbi megközelítésekhez képest, de nem szabad elfelejtenünk, hogy a fent ismertetett, demográfiai alapú almintát leszámítva, a korpusz többi részében itt is a kizárólag szövegszempontú összetettségi kritériumnak megfelelő, szövegtípusok szerinti kiegyensúlyozottság diktálta a szerkezetet. Hogy csak két példát említsek: az írottnyelvi

[7] Erre a problémára Váradi Tamás hívta fel a figyelmemet.

alkorpuszban a társadalomtudományi szövegek ugyanolyan arányban szerepelnek, mint a szabadidős szövegek (az írott korpusz-rész 10-10%-a); a beszéltnyelvi rész kontextus alapján kiválasztott almintájában a monológok aránya 40%, a dialógusoké 60%. Miért pont annyi? — kérdezhetjük, miközben magunk sem tudjuk, mi a valódi arány a populációban.

A Longman Korpusz-Hálózat (*Longman Corpus Network*) egyik vezetője, Summers amellett érvel, hogy a mintavétel kezdetén „alkalmazzuk az *objektív* eszközökkel definiált dokumentum- vagy szövegtípusok széles skálájának gondolatát mint alapvető rendszerezési elvet" (1991: 5, idézi Kennedy 1998: 63, kiemelés tőlem), melyet aztán a korpusz elemzésekor módosíthatunk vagy finomíthatunk. A finomhangolás gondolata egyébként máshol is felbukkan (Biber 1993, Váradi 1998), és teljesen elfogadható monitorkorpuszok tervezése esetében, referenciakorpuszok esetében azonban a korpusz arányait már nem érinti — legfeljebb a meglévő szövegmennyiségen belül, csökkentéssel változtathat az arányokon a kutató. A legfőbb probléma Summers érvelésével kapcsolatban azonban az, hogy ezek a kategóriák nem objektívek, ez idáig ugyanis a korpusztervezés nem talált ilyen objektív eszközöket!

Summers (1991) gondolatmenete egyébként azért érdekes, mert csoportokba sorolja az írottnyelvi szövegek jellegzetes gyűjtési kritériumait (idézi Kennedy 1998: 63-4):

- 'elitista': az irodalmi/tudományos érdem vagy a 'fontosság' szerinti gyűjtés
- véletlen választás
- 'kurrensség', olvasottság (ez a módszer előnyben részesíti az újságnyelvet és a best-sellereket)
- a 'tipikusság' szubjektív megítélése
- hozzáférhetőség (ezt alább részletesen tárgyalom)
- az olvasási szokások demográfiai megközelítése
- a szövegkiválasztás empirikus módosítása a nyelvészeti kívánalmaknak megfelelően
- pragmatikus: a fentiek kombinációja

Summers természetesen az utóbbit részesíti előnyben, de szerintem ezzel sem jut túl az intuitív megközelítés problémáján. Az alábbiakban olyan módszert ajánlok, amely véleményem szerint megoldja ezt az egyébként ismeretelméleti problémát.

1.2 Az arányok problémája

A korpuszok tervezői általában *elvileg is* elvetik annak lehetőségét, hogy a beszélt- és írottnyelvi regiszter- és műfajvariabilitás eloszlása felmérhető lenne (például Biber 1993:247, Burnard 1995: 20, Kennedy 1998:63). Valóban, közvetlen módszerek idáig csak a nyomtatásban megjelent szövegek és az internetre feltett oldalak felmérésére születtek (nemzeti könyv és periodika-publikációs listák, illetve internetes keresőmotorok), ezen szövegek befogadói oldalának felmérésére csak közvetett módszerek állnak rendelkezésre (bestseller-listák, könyvtári kölcsönzési és folyóirat-kiadási statisztikák, illetve egy adott internet-oldalra való rákattintások mérése). Nincsenek közvetlen kvantitatív módszereink a nyomtatásban megjelent szövegek befogadásának, illetve a nyomtatásban meg nem jelent szövegek előállításának és befogadásának mérésére, sem a beszélt nyelven létrehozott illetve befogadott szövegek objektív becslésére.

Kennedy azt írja: „a korpusz egyensúlya nem úgy érhető el, hogy azonos mennyiségű szöveget emelünk be különböző forrásokból, például beszélt és írottnyelvi angolból. Senki sem tudja, hogy egy adott napon egy bizonyos nyelven használt szavak között milyen a kimondottak illetve a leírtak aránya. Az egyén szintjén, legtöbbünk esetében az elmondott vagy meghallott beszéd aránya nagyobb az írottnál egy tipikus napot tekintve. Ugyanakkor egy írott szöveget (mondjuk egy újságcikket) akár 10 millió ember is elolvashat, míg egy cipővásárlással kapcsolatos párbeszédet esetleg az eredeti beszélgetőpartnereken kívül senki nem hallgatja. [...] Az írott korpuszon belül az arányok épp ilyen kezelhetetlenek. [...] Nem egyszerű megtalálni az arányokat egyrészt azon kevés magas presztízsű író és beszélő, másrészt a szövegek létrehozóinak és befogadóinak hatalmas tábora között, akik nem támaszthatnak igényt hírnévre." (1998: 63)

Biber szintén a fontosság faktorát hangsúlyozza: „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" (1993: 247-8[8]). Biber ebben a cikkében Sinclairhez (1991: 13) hasonlóan a kultúra szociológiájának keretébe utalja ennek a kérdésnek a megoldását.

Úgy tűnik, ezeknek a szerzőknek a gondolatmenetében a 'reprezentativitás' és az 'arányosság', illetve a 'diverzitás/variabilitás' fogalmaival még egy, ezekhez képest külsődleges (tehát nem a fentiek értelmében vett strukturáló) fogalom, a 'fontosság' (*influentialness*) is összekeveredik.

Mint korábban leszögeztem, egy általános korpusznak a lehető legösszetettebbnek kell lennie, mert csak így felel meg annak a statisztikai axiómának, mely szerint minden erőfeszítést meg kell tennünk annak érdekében, hogy a minta reprodukálja a populáció varianciáját. A fontosság kontroll nélküli, korai bevezetése azonban kizárja annak lehetőségét, hogy később a szövegek fontosságát mint változót vizsgálhassuk az elemzés során.

Egy szociológiatörténeti párhuzammal szeretném megvilágítani ezt a gondolatot. Egy demográfiai vagy szociológiai elemzésnél is feltehető, hogy a populáció egyes tagjai, például az iskolázottabbak vagy vagyonosabbak nagyobb hatásúak másoknál, és a minták lehetővé is teszik ennek mint változónak a vizsgálatát. A fentihez hasonló hibás lépés lett volna, ha a demográfia vagy a szociológia hajnalán csak az iskolázott, felső- és középosztálybeli, idősebb férfiakból, vagyis a számarányuknál köztudottan nagyobb befolyású, 'nagyobb hatású' tagjaiból vettek volna mintát az akkori társadalomtudósok. Szerencsére ez a hagyomány *abban* a tudományágban nem alakult ki.

1.2.1 A napló-módszer[9]

Hogyan állapíthatjuk meg a magyar nyelv teljességének csatorna-, téma-, műfaj- és regiszterspecifikus eloszlását? Ez lehetővé válik, ha a magyar (felnőtt) lakosságra nézve reprezentatív mintával, táblázatos formában kitölthető *naplóval/logbookkal* gyűjtünk adatokat. Az adatközlőket megkérjük, hogy néhány napig hordjanak magukkal füzetként összefűzött táblázatokat, és töltsenek ki egy-egy sort, ahányszor csak használják a magyar nyelvet.

[8] Váradi fordítása (2001:1289).
[9] Részletesebben ld. Reményi 2001.

(Kizárhatjuk azokat az eseteket, amikor az adatközlő magában beszél/gondolkodik, illetve magának a táblázatnak a töltögetését.)

A táblázatok oszlopaiban a következő információkat csoportosíthatjuk (a gyorsabb kitölthetőség kedvéért rövidített formában, vagy kódokkal):

- a kommunikációs tevékenység időtartama
- az adatközlő: szövegalkotó/szövegbefogadó/harmadik fél
- a résztvevők hozzávetőleges száma
- a többi résztvevő néhány demográfiai jellemzője (ha ismert)
- csatorna
- keret
- műfaj (bővíthető lista alapján)

Más változókra (pl. közeg, téma, a nyelvhasználat célja) is rákérdezhetünk, ezek körét szintén a naplóhoz rögzített rövid ismertetőben megadott, bővíthető listával kontrollálhatjuk. Az adatközlőket részletes ismertetővel is el kell látni, illetve részletesen be is kell tanítani (mi a vizsgálat célja, hogy kell a táblázatokat kitölteni, mi minden tekinthető nyelvhasználatnak). Az adatgyűjtés adatközlőnként 2-7 napig tart; a minimális két napot is úgy kell kiválasztani, hogy az adatközlőre jellemző, eltérő tevékenységekből álljon (például hétköznap - hétvége).

A napló-módszer fejlesztésére két próbavizsgálatot végeztem.

1.2.2 A próbavizsgálatok

Az első próbavizsgálat célja annak tesztelése volt, hogy a napló-módszerrel történő adatfelvétel valóban véghezvihető-e. A következő szövegműfajokra derült fény két adatközlő utazás közben leírt jegyzetei alapján: beszélgetés ismerőssel/ismeretlennel, futó megjegyzés ismeretlennek („*Elnézést!*...”), mások beszélgetésének hallgatása, hangosbeszélő hallgatása, újságolvasás, könyvolvasás, plakátnézegetés, bevásárlólista olvasása, stb.

A második próbavizsgálat során hét adatközlő[10] (öt nő és két férfi) jegyzetelt a táblázatos naplókba egy teljes napig minden alkalommal, amikor a magyar nyelvet használta. Három adatközlő esetében ez hétköznap, három esetében hétvégi nap volt (a hetedik adatközlő vasárnap kezdte a töltögetést, és másnap, hétfőn fejezte be).

Az 1. táblázatban összefoglaltam a számszerűsíthető eredményeket: a hét adatközlő összes nyelvi aktivitásának a hosszát percekben (átlag: 622,86 perc, szórás: 240,2), valamint ugyanezt csatorna szerinti bontásban is. Az elvárásnak megfelelően mindannyian többet használták a beszélt, mint az írott nyelvet, de az arányok igen eltérőek voltak. (Természetesen a résztvevők számának csekély és nem strukturált volta miatt nem vonhatunk le általánosítható következtetéseket.)

[10] Hálásan köszönöm adatközlőim segítőkészségét és kooperativitását, akik számos felvetéssel gazdagították is a napló-módszert.

kor	nem	a héten belül	össz (perc)	írott (perc)*		beszélt (perc)*		HSSzB**	
40	nő	hétvége	739	95	13%	644	87%	503	78%
39	nő	hétköznap	1058	509	48%	549	52%	451	82%
24	férfi	hétvége	736	175	24%	561	76%	540	96%
46	nő	hétvége	653	196	30%	457	70%	455	99%
12	nő	hétvége-hétköznap	502	114	23%	388	77%	125	32%
31	nő	hétköznap	400	192	48%	208	52%	141	68%
36	férfi	hétköznap	272	32	12%	240	82%	227	95%

1. táblázat A magyar nyelvi aktivitás hossza adatközlőnként a 2. próbavizsgálatban (perc);
* = százalék, az összes arányában; ** = a hosszabb spontán személyes beszélgetés (HSSzB) hossza és aránya a beszédtevékenységben

Az 1. táblázatban feltüntettem a családban, ismerősökkel vagy ismeretlenekkel zajló hosszabb spontán személyes beszélgetések (HSSzB) arányát. A beszélt nyelvi műfajok ezen kívül többek között a következők voltak: telefonbeszélgetések, futó interakciók (köszönés, elnézéskérés, bolti fizetés), mozifilmek vagy tévéműsorok nézése[11], rádióműsorok, hangosbeszélő vagy mások beszélgetéseinek a hallgatása és vegyes osztálytermi interakciók. Az írott nyelvhasználatként az adatközlők a következő műfajokat jelölték meg: újságolvasás (különböző sajtótermékek és azok rovatai), könyvolvasás (szépirodalom, népszerű tudomány, szakszöveg, tankönyv), óriásplakátok nézegetése, térképböngészés, e-mailek írása és olvasása, saját jegyzetek olvasása, könyvtári katalógus böngészése, könyvtári és könyvesbolti újságok és könyvek lapozgatása, plakátok és utcai hirdetések olvasgatása, turistáknak szóló információk, buszmegállóban kihelyezett menetrendek és információk, utcanevek és utcai feliratok olvasása, esti mese felolvasása, mobiltelefonos adatrögzítés, iskolai dolgozat és házi feladat írása, könyvtári kérőlap kitöltése, magyar nyelvű számítógépes szövegszerkesztő használata angol szöveg javításakor.

Bár a napló-módszerrel valódi szöveget nem rögzítünk, de megfelelően kiválasztott mintán statisztikailag megbízható képet kapunk a szövegtípusok eloszlásáról a demográfiai populációban, melyet a korpusz demográfiai szempontból is reprezentatív összeállításakor felhasználhatunk. Így megoldódik a korpuszban szerepeltetendő csatornák, témák, műfajok vagy regiszterek arányainak megbízható felmérése. Mindezeken túl új műfajokat találhatunk, valamint lehetővé válik a befogadók száma és a 'relatív fontosság' közti összefüggés megragadása, méghozzá sokkal közvetlenebb módon, mint az indirekt befogadás-specifikus módszerek (könyvtári statisztikák, bestseller-listák stb.) esetében. Mivel a logbook-módszer egyszerű és viszonylag olcsó, ezért viszonylag nagy mintán is lebonyolítható.

Az elektronikus korpuszok első generációjában természetesen a korpusznagyság szigorú limitáló tényező volt, ezért esetleg luxus lett volna a nyelvileg kevésbé gyakori műfajokat a demográfiai populációnak megfelelő arányban szerepeltetni. A nagykorpuszok (≥ 100 millió szövegszó) világában kevésbé kell számot vetni a korpusznagyság limitáló hatásával, ezért — ha valaki a 'reprezentatív' szót akarja a zászlajára (szótára, nyelvtankönyve, tankönyve, fordítóprogramja stb. címoldalára) tűzni — elvárható, hogy az adott nyelv teljes populációjának műfaji stb. eloszlásával legalábbis tisztában legyen. Annyi gyakorlati engedményt tennünk kell, hogy a beszéltnyelvi transzkripció 'drágasága' miatt a megfelelő arányokat nem feltétlenül lehet képviseltetni a mintában, de a valós arányokat mindenképpen tisztázni kell, túljutva a Biber-féle, „90% + 3% + 7%" becslésen alapuló elutasításon, hiszen az is a sokat idézett szerző intuícióján alapul.

[11] A műsorokban gyakran számos műfaj keveredik.

1.3 A mintavételi keret

A statisztikailag kívánatos véletlenszerű mintavételhez mintavételi keretet (*sampling frame*) kell kijelölnünk. A mintavételi keret a (rétegzett) populáció tételes felsorolását tartalmazza, melyből valamilyen véletlen kiválasztásos módszerrel megkapjuk a szövegek mintáját.

A magyar nyelv referencia-korpuszánál a tételes felsorolás kritériuma a publikált és az interneten hozzáférhető írottnyelvi szövegek esetében a szöveg-előállítás dimenziójában gyakorlatilag akadálymentes, az időbeli korlátozás figyelembevételével. A nem publikált írottnyelvi, a beszéltnyelvi és az előre megírt, felolvasásra szánt beszéltnyelvi szövegek esetében a tételes felsorolás megoldhatatlan (mind a szöveg-előállítás, mind a szövegbefogadás dimenziójában), ezeknél véleményem szerint az optimális módszer az, ha a szövegeket alkotó ill. befogadó személyek demográfiai populációján alapuló napló-módszer megszabta célpopuláció arányait követjük a mintában. Ez a módszer a publikált írottnyelvi szövegek befogadói dimenziójának megismerésére is célravezető lehet.

A publikált szövegek szöveg-előállítási dimenziójának mintavételi kerete a magyar ISBN- és ISSN-hivatal hivatalos éves jegyzéke lehet, melyben az összes publikált kiadvány megtalálható, vagy — ha szűkebb populációs határokat szabunk meg — egy adott könyvtár adott évi gyarapodási jegyzéke (ez az Országos Széchényi Könyvtár esetében egyébként tágabb keret lenne, mint az előző kritérium szerint felvett keret). Az elsőhöz hasonló módszert használt a Lancaster-Oslo-Bergen (LOB) korpusz, ahol az 1961. évi brit nemzeti bibliográfia kumulált tárgyszójegyzékét vették alapul a könyvek, a *Willing's press guide*-ot a periodikák esetében, és rétegzett véletlen választással jelölték ki ezekből a mintába kerülő szövegeket. A LOB-korpuszhoz hasonló amerikai Brown-korpusz összeállításakor a második módszert választották: ott a Brown-egyetem katalógusának 1961. évi tételei adják a mintavételi keretet.

A beszélt nyelvi szövegek esetében a magyar nyelvű vagy magyar anyanyelvű beszélők alkothatják a demográfiai populációt, akiknek egy időben behatárolt időszak alatt elhangzott beszédprodukciója az a nyelvi populáció, melynek mintája feldolgozható lenne.

Ameddig a napló-módszeren alapuló reprezentatív mintavétel nem valósul meg, egy meglévő, demográfiai szempontból Budapest lakosságára nézve nem, életkor, iskolai végzettség tekintetében 1989-re vetítve reprezentatív vizsgálat, a Budapesti Szociolingvisztikai Interjú 3. változata (BUSZI3) átiratainak felhasználása a járható út. A BUSZI3 200 beszélővel felvett, egyenként 2-2,5 órás interjú, amely demográfiai szempontból a fenti változók tekintetében reprezentatív, nyelvi szempontból azonban nem törekedhetett erre, mivel a kvantitatív szociolingvisztika hagyományait követve kizárólag interjú segítségével szimulált különböző beszédműfajokat/regisztereket (felolvasott szöveg, interjú, beszélgetés). Mivel csak budapesti beszélőkre nézve volt reprezentatív, ezért általános korpusz kialakításakor ki kell egészíteni például a BNC-nek megfelelő gyűjtési módszerrel különféle beszédhelyzetekben begyűjtött szövegekkel, illetve más településtípusokkal is.

Referencia-korpusz esetén mind az írottnyelvi, mind a beszéltnyelvi szövegek populációjának kijelölésekor időbeli határokat kell szabnunk, meg kell adnunk, hogy például kizárólag az 1998-ban kiadott, vagy a 2001 májusában interneten hozzáférhető magyar nyelvű szövegek alkotják mintavételi keretünket, vagy például a 2001 decemberében rádióban és televízióban elhangzott szövegekből veszünk mintát.

1.4 Méretezési kérdések

A méretezéssel kapcsolatos problémák megfontolása — melyre elsősorban referencia-korpusz esetén van szükségünk — csak a célpopuláció és a mintavételi keret problémájának megfontolása után következhet (ld. Biber 1993:243).

Egy korpusznak elég nagynak kell lennie ahhoz, hogy eltérő típusú szövegekből a statisztikai elemzés és az erre épülő általánosítható következtetések számára megfelelő számú nyelvi adatot lehessen nyerni egy bizonyos nyelvi jelenségre vonatkozóan. Nyilvánvaló, hogy a szövegek elsősorban nem abban fognak különbözni egymástól, hogy egy bizonyos lexémát, nyelvtani jelenséget stb. használnak-e vagy sem, hanem inkább abban, hogy milyen *gyakorisággal* fordulnak elő bennük ezek a jelenségek.

Az elektronikus nagykorpuszok első generációjának sztenderd mérete egymillió szövegszó volt. A második generációs nagykorpuszok mérete egy vagy két nagyságrenddel nagyobb (a Cobuild-korpuszok kb. 21 millió, a Bank of English monitorkorpusz >300 millió, a BNC 100 millió, az International Corpus of English 20 millió szövegszót tartalmaz).

Az általános korpuszokat éppen azért kell nagyobbra tervezni, mint a specifikus célokra szánt korpuszokat, mert sokféle elemzést kívánunk általuk lehetővé tenni, beleértve például a kollokációs vizsgálatokat[12] is.

Fel kell tennünk a kérdést, hogy egy általános referenciakorpusznak mekkora az optimális mérete, hiszen nemcsak a túlságosan kis korpusz okozhat gondokat (nem kapunk elemezhető számú nyelvi adatot egy-egy jelenség vizsgálatára), hanem a túlságosan nagy méret is: például a szövegek besorolásának nehézsége vagy a morfológiai elemzésben (*tagging*) a hibák száma a korpusz méretével arányosan nő. Nagy korpuszok esetén ezeknél jelentősebb problémák a copyright megszerzéséből vagy a futtatások megnövekedett időigényéből, a nagy tömegű adatok kezelésének korlátaiból adódnak.

Biber (1993) a korpusz szövegszavai és szövegei számának tekintetében is megvizsgálta ezt a kérdést, de véleménye szerint — legalábbis a nyelvtani jelenségek tekintetében — nem adható meg abszolút szám, hiszen a nyelvtani jelenségek gyakorisága igen eltérő lehet (például a főnevek jóval gyakoribbak, mint a feltételes alárendelő tagmondatok (1993: 253); kollokációs vizsgálatokhoz pedig igen nagy korpuszra van szükség). Kennedy (1998: 68) szerint prozódiai információt tartalmazó 100.000 szövegszavas spontán beszéltnyelvi korpusz elegendő a prozódiai jelenségek tanulmányozására. Az angol ige-morfológia esetében 500.000 szövegszavas korpusz elegendő — azonban egy magyar nyelvi vizsgálathoz ez az információ nem irányadó. Kennedy szerint a gyakori szintaktikai jelenségek és a szókincs leggyakoribb szavainak tanulmányozására a 0,5-1 millió szövegszavas állomány elegendő.

Németh és Zainkó (2000)[13] magyar nyelvű korpuszokon alapuló szövegszó-elemzésében arra a megállapításra jutott, hogy „az 1 %-os változási sebesség — ami azt jelenti, hogy átlagosan minden egyes új különböző szóhoz a korpusz méretét 100 szóval kell növelnünk —

[12] A kollokációs vizsgálatokban újabban nemcsak a közvetlenül egymás mellett álló szópárok/szócsoportok vizsgálatára van egyébként lehetőség, hanem a szövegbéli szóláncban egymástól távolabb állókéra is.

[13] E tanulmány eredményeinek validitását csökkenti az a tény, hogy az elemzett korpuszok a *Magyar Elektronikus Könyvtár* kiválasztott dokumentumain kívül a *Magyar Nemzet* 2000. április-október, a *Magyar Hírlap* 2000. január-április közötti, valamint a *Heti Világgazdaság*nak „az elmúlt 3,5 évben" (Németh-Zainkó 2000:159) megjelent szövegeit tartalmazzák, és ezekhez illesztik a MNSZ időben közelebbről nem specifikált állományát is, miközben a MNSZ a fenti újságszámokat maga is tartalmazza.

31 millió szónál található. Az 1 ezrelékes határ viszont már 310 millió szónál van" (2000: 161). Ez utóbbi méret véleményem szerint a MNSZ mint általános referencia-korpusz esetében nem elérhetetlen vagy a méretet tekintve megvalósíthatatlan vágyálom.

A teljes korpusz mérete azonban csak az egyik eldöntendő kérdés. Referencia-korpusz esetében arról is előzetes döntést kell hoznunk, hogy mekkora legyen a szövegek száma és a mintában lévő egyes szövegek hossza, teljes szövegeket emeljünk-e be a korpuszba vagy azonos hosszúságú szövegeket/szövegrészleteket. Ez utóbbi esetében előre meg kell határozni, hogy hosszabb szövegek esetén mely szövegrészletet emeljük be.

Az azonos hosszúságú szövegminták mellett szól, hogy így a különböző szövegtípusok hordozta nyelvi jellegzetességek összehasonlíthatóvá válnak, míg a hosszabb szövegek esetében a rájuk jellemző nyelvi jellegzetességek felülreprezentálódnak. Hosszú szövegek beemelése sérti a minél nagyobb variabilitás elvét is, hiszen a fenti mintavételi elvek alapján kiválasztott több szöveg nyilvánvalóan nagyobb variabilitást eredményez. A teljes szövegek beemelése mellett szól viszont az, hogy csak így válnak vizsgálhatóvá a diskurzus-jelenségek (pl. szövegkohézió, anafora, szövegstruktúra).

Az egyes szövegek hosszának tekintetében Biber (1990) arra a következtetésre jutott (az angol nyelvre vonatkoztatva), hogy a 2000-5000 szövegszó hosszúságú szövegek nyelvtani szempontból megfelelően reprezentálják szövegkategóriájukat, és a LOB strukturálásában használt műfajokat tekintve műfajonként 20-80 szöveg elegendő a leggyakoribb grammatikai jelenségek korrelációs vizsgálatára.

Schlüter[14] szerint nem elsősorban a korpusz nagysága vagy a szövegek hossza számít az elemzésben, hanem a vizsgálandó jelenség előfordulási gyakorisága. Kétértékű változó (például aktív-passzív igealak az angolban, határozott-határozatlan igeragozás a magyarban) esetén, ha megelégszünk a 0,05 szignifikanciaszinttel, a különböző szövegekben elegendő a változó 400 előfordulása.[15]

1.5 Másutt is most szerveznek nagykorpuszt: Az Amerikai Nemzeti Korpusz (ANC)[16]

Annak ellenére, hogy az első, egymillió szövegszavas nagykorpusz, a Brown-korpusz a hatvanas években az Egyesült Államokban jött létre, valamint a Pennsylvaniai Egyetemen működik az egyik legnagyobb korpusz-gyűjtemény és -szolgáltatás, a NyelviAdat-Konzorcium (*Linguistic Data Consortium*, LDC), nemzeti összefogással épített, heterogén, strukturált nagykorpusz az USÁ-ban még csak most van születőben: 1998-ban indult be a nemzetközi ipari és könyvkiadói támogatás megszervezése. A munkálatokat az LDC irányítja, a korpusz összeállítása és kódolása is ott folyik.

[14] Ld. *Corpora*-lista (2001. június 3.).
[15] Ez az eredmény az $n = (4*p*(1-p))/ alfa^2$ egyenlet megoldásából adódik, ahol p = 0,5 és alfa = 0,05. Ez utóbbi egyenlet az alfa = t * √ ((p(1-p))/n) egyenlet átalakításából adódik. (Az alfa = 0,05-höz tartozó t értéket felkerekítettük 2-re.) Részletesebben ld. Holm (1975). — Köszönöm Norbert Schlüternek, hogy az egyenletet levezette és a szakirodalmi hivatkozást megadta számomra.
[16] Ide - Macleod (2001) alapján.

Az ANC két részből áll majd: a 100 millió szövegszavas, jobbára a BNC szerkezetére (csatorna, közeg) épülő referencia-korpuszba — a BNC-től eltérően — csak 1990 után keletkezett szövegek kerülnek, míg az öt évenként tíz százalékkal bővítendő monitor-korpusz esetében az egyik gyűjtési kritérium a hozzáférés egyszerűsége lesz, míg a másik a referencia-korpusz kiegyensúlyozott összetétele helyett a változatosság (például elektronikus levelek, rap-szövegek, történeti szempontból fontos regények és más írások gyűjtését is tervezik). Jelenleg mind írott, mind beszéltnyelvi szövegeket gyűjtenek; a beszéltnyelvi szövegek egyelőre egyszerű transzkripció formájában kerülnek a korpuszba, az annotációra később kerül sor. Tervezik amerikai angol nyelvjárási szövegek (pl. kanadai angol), valamint spanyol és kanadai francia szövegek gyűjtését is (párhuzamos angol fordítással), de ezek megvalósítását egyelőre elhalasztották. Jelenleg a gyűjtés mellett a szövegek „alap-kódolása" folyik: XML-kódolás (XCES) a cím, fejezet, bekezdés, stb. jelzésére, illetve automatikus morfológiai elemzés és kódolás. Az annotált referencia-korpusz elkészültét három éven belül ígérik.

Az ANC-t kutatási célokra kezdettől fogva az egész világon ingyenesen (illetve névleges összegért) hozzáférhetővé teszik — a BNC-től eltérően, amely a közelmúltig csak az Európai Unió országaiban volt hozzáférhető.

1.6 Moduláris korpusz

A MNSz első lépésben referencia-korpusznak épül, s e korpusz felépítése után a későbbiekben monitor-korpusszá fog válni (Váradi 1998). Az ilyen összetett szerkezetű általános korpusz esetén azonban további kérdéseket is érdemes megfontolni.

Először is felmerül a kérdés: kell-e egyáltalán reprezentativitás? Nem elegendő-e inkább csak mindenből minél többet gyűjteni, a szövegek forrását pontosan dokumentálni, majd a felhasználóra bízni a megfelelő szövegstruktúra kiválasztását?

A pre-strukturálásnak ez az elhárítása referencia-korpusz esetében nemigen tartható, hiszen az általános felhasználónak (lexikográfusnak, nyelvtanírónak, nyelvkönyvírónak) nincsenek meg az eszközei a populáció felmérésére, a kész korpuszt legfeljebb intuitív módon arányíthatja az általa elképzelt populációhoz. Monitor-korpusz esetén a pre-strukturálás kívánalmán lazíthatunk, de akkor sem érdemes teljességgel elvetni, hiszen a demográfiai és szövegpopuláció felmérése leginkább a korpusznyelvész (és a szociolingvista) lehetősége és feladata, mind referencia-, mind monitorkorpusz esetében. A korpusznak a nyelvészeti felhasználás kívánalmainak megfelelő „ciklikus finomhangolása" (Biber 1993, Váradi 1998) is csak monitorkorpusz esetében lehetséges.

A felhasználó nyelvésznek azonban nem feltétlenül van szüksége egy nagy méretű korpusz teljes állományára. Megfelelő előkészítés esetén a korpusz *moduláris* felhasználása is lehetővé válik. Ha ugyanis mind a referencia-korpusz, mind a monitor-korpusz szövegei jól dokumentáltak a lényeges kritériumok („rétegek") tekintetében, egy, megadott kritériumok szerint válogató felhasználói interfész segítségével bármilyen összetételű alkorpusz létrehozható, az egyéb kritériumok szerinti kontroll lehetőségével.

Ilyen interfésszel létrehozható akár *kiegyensúlyozott* szövegösszetétel, például egy szintaktikai jelenség műfajonkénti eloszlásának tanulmányozására, akár a *demográfiai eloszlást követő* szövegösszetétel (ahol az eloszlás a napló-módszerrel gyűjtött eredményeket követi), például éppen egy, a szöveg 'fontosságát', más szövegek jellemzőire való hatását vizsgáló

szociolingvisztikai elemzés, vagy egy, a politikailag korrekt kifejezések terjedését vizsgáló elemzés számára. Ezzel az interfésszel a felhasználó számára fontos más összetétel is beállítható lenne, például egy prozódiai elemzés számára csak a prozódiai kódolással ellátott beszéltnyelvi szövegek kiszűrése, vagy egy új szavak/kifejezések terjedését vizsgáló kutatás számára a csak egy adott időponttól kezdve létrehozott szövegeké. (Természetesen a korpusz teljessége is választható lenne.) A feladat típusától függően azt is lehetővé kellene tenni, hogy a felhasználó beállíthassa, teljes szövegeket vagy egyforma hosszú szövegeket emeljen a válogató program a felhasználói alkorpuszba. Az interfésznek képesnek kell lennie arra, hogy a kiválasztott szövegek összetételéről megadja az alapstatisztikákat, hogy az elemző változtathasson a kritériumrendszeren, ha az eredetileg kiválasztott alkorpusz túl kicsi vagy aránytalan lenne. Ez könnyen előfordulhat, ha az eredeti monitor-korpusz aránytalan; ezen a problémán a felhasználó nem tud segíteni, a hiányzó szövegtípusok szerinti ciklikus finomhangolással, tehát újabb szövegek beemelésével a moduláris korpusz tervezője-fenntartója viszont igen.

A moduláris felépítés esetén szükség van egy néhány millió szövegszavas, a legalaposabban feldolgozott, bekódolt „belső magot” létrehozni a tanuló-korpusz, a tesztvizsgálatok és a robosztusabb eredményeket követelő futtatások számára. A MNSz esetében ez a négymillió szövegszavas, a Humor morfológiai elemzővel elemzett, majd egyértelműsített tanuló-korpusz elkészült. A későbbiekben érdemes lenne összetételét a MNSz struktúrájához hangolni.

A fentiekben azokat az kérdéseket feszegettem, amelyek minden általános számítógépes nagykorpusz tervezőjét foglalkoztatják. A továbbiakban a jelenleg korlátozott anyagi lehetőségekre való tekintettel azokat a kérdéseket veszem sorra, amelyek a MNSz-at összeállító munkacsoportot[17] a fentieken túl a Szövegtár első változatának összeállításakor foglalkoztatták.

2 A MNSz ELSŐ VÁLTOZATA

A Magyar Nemzeti Szövegtár összeállításában 2000 végéig a tanuló-korpusz kialakítását célzó újságnyelvi és szépirodalmi szövegek gyűjtésén túl az erőfeszítések elsősorban az automatikus feldolgozás fejlesztésére irányultak. A 2001. év célja a korpusz szélesebb alapokra helyezése, a strukturált általános nagykorpusz első változatának létrehozása volt. Ennek jelenleg legköltségkímélőbb bővítési módja az egyszerű és ingyenes hozzáférés okán az elektronikus formában, részben az interneten hozzáférhető szövegek gyűjtése. (A továbbiakban kizárólag az interneten keresztül hozzáférhető szövegekből felépülő korpusz tervezési problémáival foglalkozom.)

2.1 Az internetes szövegek populációja

Az internetes szövegek populációja a MNSz esetében minden, az interneten hozzáférhető magyar nyelvű szöveg. A populáció értelmezése során a következő problémákkal kellett megbirkóznia a MNSz tervezőinek:

[17] Váradi Tamás (projektvezető), Kiss Gábor, Reményi A. Á.

- A világhálón elérhető dokumentumok nem mindegyike szöveg. Léteznek képi, adatbázis- és program-típusú dokumentumok is, sőt, egy-egy oldal szöveget, képeket, adatlistákat stb. vegyesen is tartalmazhat. A szövegek formátuma is eltérő lehet: a szövegstruktúráról (pl. cím, fejezet, alfejezet, bekezdés, stb.) legtöbbet eláruló, ezért korpusznyelvészeti szempontból legkedvezőbb kódolású HTML-formátum mellett egyszerű szövegfájlok (.txt), különböző szövegszerkesztők alatt hozzáférhető fájlok (.doc, .pdf) és képként elmentett szövegek (.jpg) is nagy számban találhatók a világhálón.

- A magyarországi szervereken (többnyire — de nem kizárólag — a *.hu* tartományon) található szövegek nem mindegyike magyar nyelvű, bár ez a probléma az automatikus nyelvfelismerő módszerekkel megoldottnak tekinthető. Ugyanakkor a nyelvek keveredhetnek is egy-egy szövegen belül.

- Egy-egy szöveg (jogszabály, vers, reklám, újságcikk stb.) különböző helyeken azonos vagy hasonló formában több, akár sok-sok példányban is létezhet.

- Az internet állandóan változik, egy-egy adott cím alatti dokumentum a feltérképezés-gyűjtés ideje alatt is eltűnhet, részben vagy teljesen megváltozhat.

- A szövegek egy részének szerzőjét nem lehet felderíteni. Az olvasók száma és összetétele a MNSz számára kideríthetetlen.

Mekkora a magyar nyelvű internetes szövegoldalak populációja? Hogyan állítható össze ezek mintavételi kerete? 2000 őszén keresőprogramokra bíztuk a becslést úgy, hogy egy gyakorisági lista, a Füredi-Kelemen (1989) szótár szerint leggyakoribb magyar szavakat (*az, és, hogy*) kerestettük le néhány ismert keresőprogrammal. A magyar nyelvű szövegekben való keresés az alábbi találatszámot adta (ezek a listák nagyságrendileg megfeleltethetők a mintavételi keretünknek):

	altavista.hu, .com (okt.)	altavizsla (okt.)	google (dec.)[18]
„ és”	6.080.969 talált/ 1.290.280 dokumentumban	2.306.632	kb. 815e (.hu)
„ az”	11.701.551/1.549.210	2.243.081	kb. 569e (.hu)
„ hogy”	2.805.574/460.940	1.019.936	kb. 601e (.hu)

2. táblázat A keresőszavak találatszámai

A 2. táblázatról leolvasható, hogy az Altavista keresők októberben körülbelül 1,3 millió magyar nyelvű dokumentumban összesen 6 millió *és* előfordulást számláltak össze. Az Altavizsla a .hu tartományban 2,3 millió olyan dokumentumot talált, amely az *és* szót tartalmazza. A Google decemberben ugyanebben a tartományban 815 ezer dokumentumban

[18] A www.altavista.com, www.altavista.hu ill. www.altavizsla.hu 200 találat, a www.google.com 999 találat megtekintését engedélyezi. Az Altavista keresőkben nyelv szerint, a Google-ban nyelv vagy tartomány (.hu) szerint szűrhetjük a keresést; az Altavizsla csak a .hu tartományban keres. Decemberben teszteltem a www.northernlight.com keresőt is, de az nem jól kezeli az ékezeteket (*és* keresésekor *es* eredményeket is kapunk), és sem a magyar nyelv, sem a .hu domain szűrését nem ajánlja fel (bár más nyelv- ill. tartomány szerinti szűkítés lehetséges). Egyedi szolgáltatása, hogy az eredmények csoportosíthatók, de a csoportosítás módszere nem világos (pl. az *az* 3.777.145 találata közül 327.531 „Hungarian sites”, 3.805 „in Hungary”, de hogy a kettő megfelel-e a nyelvek ill. tartományok szerinti szűrésnek, az nem derül ki). Nincs információ arról sem, hogy hány eredmény megtekintését engedélyezi ez a kereső.

találta meg ezt a szót. A másik két keresett szó esetén a táblázat számai hasonlóképpen értelmezhetők.

Az adatok közti eltéréseknek számunkra nincs jelentősége, hiszen csak a nagyságrendeket akartuk tisztázni. Az eltérések részben az eltérő időpontokkal, részben az eltérő keresési kritériumokkal ('tartomány' vagy 'nyelv'), részben a keresőprogramok adatbázisainak és keresőmotorjainak eltéréseivel magyarázhatók. Nem tisztázott ugyanis, hogy melyek azok az internetes oldalak, melyeken ezek a keresőprogramok nem keresnek. Az Altavizsla ismertető oldala szerint „Az AltaVizsla maga indexeli a World Wide Web magyar tartományának („.hu domain") teljes egészét", s ezt az Altavizsla kezelői kérdésemre meg is erősítették, miközben az Altavizsla még saját híroldalairól (*Főbb híreink; A nap sztorija*) ill. fórumairól sem jelez találatokat, ill. más web-fórumokról sem. Mindez felveti annak a kérdését, hogy a keresőkön keresztül valóban hozzáférhetnénk-e a teljes magyar nyelvű (vagy a '.hu' tartományban létező) internet-világhoz. Nyilvánvaló ugyanakkor, hogy a keresőkkel becsült 1,5-2 millió dokumentumnál nagyobb a .hu tartomány — bár ezeknek csak egy része szöveg.

Az internetes hozzáférhetőség egyébként sem dichotóm kategória (hozzáférhető/nem férhető hozzá), mivel az egyes internetes típusokhoz való hozzáférés szintje más és más. A számítógépen őrzött információ tárolásának és elérhetőségének fokozatai Szakadát (2000: 4-5) szerint:

1	a gépemen van	csak én érhetem el	magánszféra
2	email-ben küldöm el	csak a címzett és én érhetem el	magánszféra
3	zárt newsgroupban, zárt intraneten olvasható	csak a közösség tagjai érhetik el	zárt közösségi nyilvánosság
4	interneten van, link nélkül	véletlenül bárki elérheti, gyakorlatilag senki nem érheti el	nyilvánosság
5	interneten van, linkkel, promóció nélkül	kevesen érhetik el	nyilvánosság
6	interneten van, linkkel, promócióval	sokan érhetik el	nyilvánosság

3. táblázat A számítógépes információhoz való hozzáférés fokozatai (Szakadát 2000: 4-5)

A fórumok és a társalgószobák a 3. táblázatban sincsenek feltüntetve. Ezeken kívül a MNSz-ban feltétlenül gyűjtésre érdemes a későbbiekben nemcsak a 4-6., hanem lehetőség szerint a 2-3. sorban feltüntetett típus is. Az internetes korpusz összeállításakor is figyelembe veendők a hagyományos szövegek gyűjtésétől részben eltérő jogi, etikai problémák, melyek megoldása egyelőre nem tisztázott.[19]

2.2 Mintavétel az internetről

Bár a korpusznyelvészek szerte a világon hamar felismerték a világhálóban rejlő lehetőségeket, jelenleg még nincs bevált módszer a mintavételre. Maga a világháló az általam használt értelemben nem tekinthető korpusznak, mert egyrészt automatizált módszerekkel körülményes annak kiderítése, hogy egy kizárólag vagy nagyrészt szöveges, egynyelvű dokumentummal

[19] A vitáról ld. a Corpora-listát: www.comp.lancs.ac.uk/computing/research/ucrel/public/ 1581.html-től .../1586.html-ig tartó hozzászólásokat.

állunk-e szemben, másrészt a szövegekről szóló meta-információhoz (például szerző(k), műfaj) való hozzáférés esetleges. [20]

A gépi, automatikus letöltés gyors ugyan, tehát lehetővé teszi a hozzáférést akár több millió szöveghez is, azonban az automatikus szövegtipizálás jelenleg még megoldatlan. Emiatt két lehetőség kínálkozik. Az egyik lehetőség az, ha „ömlesztve" mentjük le az internetes oldalakat, így tárolási kapacitásunk függvényében a .hu tartományt mintegy „megduplázva" alakítjuk ki a mintavételi keretet, amelyből azután rétegzetlen véletlenválasztásos mintavétellel veszünk mintát. Ebben az esetben le kell mondanunk a 'rétegek' kialakításáról, tehát a korpusz strukturálásáról. A másik lehetőség az, hogy — lemondva a dolgozat első részében vázolt reprezentativitásról — csupán annyi szöveg letöltésére vállalkozunk, amennyinek az osztályozását a rendelkezésre álló emberi erőforrások lehetővé teszik. Az utóbbi tényezőre való tekintettel a MNSz első változatának kialakításakor a második utat választottuk.

Mivel egyelőre nem volt lehetőség műfajok és regiszterek széles skálájának gyűjtésére, ezért a MNSz jelenlegi fejlesztési periódusában átvette a 40 millió szövegszavas Longman Beszélt- és Írottnyelvi Korpusz (LSWE) szerkezetét, egy változtatással: még egy regiszter beemelésével.

2.2.1 A LSWE-korpusz

A LSWE-korpusz[21] négy regiszteren alapul: beszélgetés, szépirodalom, újságnyelv és tudományos próza. A korpusz összeállítói éppen ezeket a regisztereket egyrészt fontosságuk és igen jelentős mennyiségű szöveget produkáló voltuk miatt választották, másrészt azért, mert számos szituációs szempont (csatorna, interaktivitás, közös szituáció, fő kommunikációs cél, közönség, nyelvváltozat) szerint különböznek egymástól. Ugyanakkor az egyes kategóriák csak viszonylag homogének, mivel jelentős a belső variabilitás, például az újságnyelven belül az egyes újságok között és az újságokon belül is (a rovatoknak megfelelően) különféle műfajok léteznek.

A Berlin-Brandenburgi Tudományos Akadémia által fejlesztett, a Huszadi Századi Német Nyelv Digitális Szótára (DWDS) számára összeállított korpusz felépítése hasonló: ez 25 százalék szépirodalmat, 25% újságnyelvet, 20% tudományos szöveget, 20% „szakszöveget" (például hirdetéseket, kezelési útmutatókat, stb.) és 10% beszélt nyelvet fog tartalmazni (Cavar et al. 2000:113).

2.2.2 A MNSz első változatának kategóriái

A MNSz első változatában a LWSE-féle négy regiszter mellett ötödikként a „hivatali nyelv" kategóriája szerepel. Az utóbbi gyűjtésében az a megfontolás vezette a tervezőket, hogy a jogi-

[20] A világháló nagy segítség lehet a strukturált szövegtárat nem igénylő feladatoknál, így a gépi szövegfeldolgozásban (például Radev és McKeown 1997, Volk 2001), az enciklopédia-építésben (Fujii és Ishikawa 2000), sőt, lexikográfiai vizsgálatokban (Varantola 2000) is. A kifejezetten lexikográfiai céllal fejlesztett keresők (így a liverpooli egyetemen fejlesztett WebCorp: http://www.webcorp.org.uk) és egyes általános keresők ugyanis szövegkontextusba ágyazva jelenítik meg az előfordulásokat. Kilgarriff (2001a és személyes közlés) a nyelvi eloszlásokból kiindulva szeretné megoldani a világháló szövegeinek automatikus tipizálását virtuális korpusza számára.

[21] Ismertetése Biber et al. (1999: 15-17) alapján

hivatali nyelv ritka és ugyanakkor „fontos”, társadalmilag nagy hatású, amire Biber (1993: 245) is felhívta a figyelmet: „kevesen írnak valaha is törvényt vagy államszerződést, biztosítási kötvényt, vagy egyáltalán bármilyen könyvet, és e szövegfajták némelyikét kevesen is olvassák. […] Márpedig e kategóriák közül sok nagyon fontos részét képezi az adott kultúrának.”[22]

A 4. táblázatból leolvasható a 2001. decemberéig begyűjtött és elemzett szövegek szószáma, valamint a fő gyűjtési források kategóriánként. Ugyanitt összefoglalom az öt szövegkategória közti különbségeket (Biber et al. (1999:16) sémáját követve). A MNSz 2001 végén összesen mintegy 180 millió szövegszót tartalmaz.

	személyes közlés	szépirodalom	sajtó	tudományos próza	hivatali nyelv
szövegszó (cca.)	20 millió	40 millió	80 millió	20 millió	20 millió
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	széleskörű	széleskörű	szakközönség	szakközönség
közönség az interneten	bárki	bárki	bárki	bárki	bárki
nyelvváltozat	helyi	többnyire sztenderd	helyi vagy sztenderd	sztenderd	sztenderd

4. táblázat. A MNSz jelenlegi kategóriáinak nagysága, forrásai és jellemzői (utóbbi Biber et al. (1999:16) sémája szerint)

A MNSz 2001 előtt a *Népszabadságot* (45 millió), a *Magyar Hírlapot* (17m), a *Népszavát* (22m), a *Magyar Nemzetet* (25m), az *Országgyűlési Naplót* (11m), a *Magyar Narancsot* (1,3m), a *HVG-t* (0,9m), valamint a *Kortárs* szépirodalmi folyóiratot (0,5m) gyűjtötte.[23] A sajtónyelvi gyűjtemény mérete akkora, hogy a fenti arányok kialakításához nincs szükség további újságnyelvi gyűjtésre. (Bár felmerül a helyi lapok hiányának a problémája, ezen a későbbi finomhangolással lesz érdemes javítani.)

A fentiek miatt az újságnyelv aránya nagyobb, mint a másik négy regiszteré. Sok általános korpusz legterjedelmesebb komponensét az újságnyelv adja (például a Linguistic Data Consortium korpuszaiban, ld. Ide - Macleod 2001: 274), melynek oka részben a szerzői jogból eredő problémák könnyebb kezelése (egységnyi szerzői jogi megállapodásra sok szöveg esik), részben az újságnyelvi szövegek adott újságon belül egységes formátuma.

A szépirodalmi kategóriában a kortárs irodalmat a *Kortárs* képviseli. A 20. századi klasszikusok közti válogatás nehéz feladatát a Neumann-ház égisze alatt működő Digitális Irodalmi Akadémia (DIA) végzi: itt 2001 végén a 20. századi magyar szépirodalom 52 jeles képviselőjének teljes életművéből 40 millió szövegszónyi korrektúrázott, egységesített HTML-

[22] Váradi (2001: 1287-1288) fordítása. — A fontosság problémájának tárgyalását ld. a jelen dolgozat 1.2 alfejezetében.
[23] Zárójelben a 2001. februári szövegszómennyiség.

formátumú szöveghez lehet hozzáférni, melyet a fenntartók engedélye alapján a MNSz átemelhet.[24]

A tudományos próza kategóriájában a BNC arányait követve kétharmad rész (13 millió szövegszó) monografikus terjedelmű művet és egyharmad rész (7 millió) folyóirat-cikket gyűjtöttünk össze. A folyóirat-cikkeket az interneten hozzáférhető szakfolyóiratokból szereztük be, míg a monográfiákat a Magyar Elektronikus Könyvtár[25] (MEK) gyűjteményéből vettük át, a MEK tudományági struktúrájával együtt. A MEK korrektúrázott, homogenizált formájú, text- ill. HTML-formátumú szövegeket tartalmaz, melyeket a fenntartók engedélyével mentettünk a MNSz számára.

Hivatali nyelvként definiáltunk minden olyan szöveget, melyet állami intézmények (parlament, minisztériumok, bíróságok), az állam által fenntartott egyéb intézmények (ORFK, VPOP, APEH, Állami Számvevőszék, Gazdasági Versenyhivatal, Országos Szabadalmi Hivatal, egyetemek, MTA, stb.) és önkormányzatok honlapjain az intézmény működése során jogi szövegként, utasításként vagy tájékoztatásként keletkezett. Az interaktívan kiválasztott oldalak között szerepelt az Alkotmány, az 1990 és 2001 között létrejött törvények, országgyűlési határozatok és egyéb irományok, törvénytervezetek, az Országgyűlés házszabálya, a kormányprogram, alkotmánybírósági határozatok, a Legfelsőbb Bíróság jogegységi döntései, kormány-, miniszteri és önkormányzati rendeletek, határozatok és ajánlások, pályázati felhívások, közlemények, tájékoztatók, háttéranyagok, szabályzatok, rendszabályok (a KRESZ is), koncepciók (például a vízgazdálkodási országos koncepció és az ISM nemzeti sportkoncepciója), stb. Amennyire emberileg lehetséges volt, kiszűrtük a fenti intézmények azon internetes oldalait, amelyeket statisztikák és adatbázisok (például képviselők, tisztviselők életrajzai, tagintézmények telefonszámai, címei, nyitvatartásai), tanulmányok, ülések jegyzőkönyvei, az adott intézmény története, sajtószemléi, linkjei, sajtóanyagok, hírek, faliújság, programok vagy eseménynaptárak töltöttek meg. A letöltött oldalak formátuma természetesen nem egységes, mint például a sajtó-kategória oldalain, ahol egy adott újságon belül akár sok ezer, a dátum, rovat, szerző, cím, alcím, bekezdés, stb. tekintetében azonos formátumú szöveget találunk. A legtöbb letöltött szöveg HTML-formátumú, de nem vetettük el a text-formátumú, vagy különböző szövegszerkesztőkkel olvasható (Word, Acrobat) szövegeket sem.

A LSWE-korpusz „beszélgetési" kategóriájához képest a MNSz a „személyes közlés" kategórianevet használja, hiszen a jelenlegi gyűjtési kör az interneten hozzáférhető szövegeket jelenti. Így valójában Az MNSz minden szövege írott szöveg, nincs csatorna szerinti megkülönböztetés az egyes regiszterek között. Közismert ugyanakkor, hogy az internetes műfajok között a fórumok (valamint a társalgószobák) szövegei állnak legközelebb az élőbeszédhez, számos jellemzőjüket tekintve a beszélt nyelv szabályait követik: mindkettőre jellemző az interaktivitás, az utóbbira a szinkronitás is.[26]

[24] A DIA gyűjtési szempontjairól, az ezeket kísérő vitákról ld. www.irodalmiakademia.hu, ill. http://www.neumann-haz.hu/scripts/DIATxcgi?infile=diat_vm_main_menu.html

[25] http://www.mek.iif.hu/

[26] A társalgószobák (*chatrooms*) beszélgetéseiben a köszönés, a megszólítás, a névmáshasználat vagy a modális segédigék gyakorisága, a cselekvések, gesztusok jelzése, stb. szintén a beszélt nyelv szabályait követik. (Ld. Baym 1996, Herring 1996, Collot-Belmore 1996, Yates 1996, Werry 1996, Miller 1996, Bays 1998, Hentschel 1998, Harrison 1998, Gousseva 1998, Cherny 1999, Paolillo 1999, magyarul Bakonyi-Drótos 1995 írásait.)

A hozzászólások helyesírási sztenderdtől való eltéréseit azonban a szövegek MNSz-ba való beemelésekor kezelni kell. A sztenderd helyesírástól való eltérés[27] egyrészt plusz terhet ró az elemzésre, másrészt jelentős információforrás lehet bizonyos, például szociolingvisztikai elemzések számára.

A sztenderd helyesírástól való eltérés nagyságrendje az írott kommunikációban elsősorban a szöveg előállítójának szocio-ökonómiai státuszáról árulkodik.[28] Az interaktív internetes műfajokban (ahol feltételezhetően nem előre elkészített, így esetleg helyesírás-ellenőrzővel átnézett szövegekről van szó) az eltérésnek azonban ennél több-, legalább háromféle oka lehet. Az *elütések* feltehetőleg véletlenszerűek. A *direkt elírások* használata az általánosan elterjedt rövidítéseket (pl. *asszem, nemtom, léci*) leszámítva egyéntől vagy a csoporttól függ. Kiszűrhetők, mert rendszeresek ugyan, de egyes szavakhoz/kifejezésekhez kötődnek. A *helyesírási sztenderdtől való egyéb eltérések* részben valamilyen — gyakran kiejtésbeli — szociolingvisztikai változóhoz kötődnek. Rendszeresek, vagyis nemcsak bizonyos szavakban, kifejezésekben jelennek meg, hanem az adott változó lehetséges előfordulásaiban az átlagosnál statisztikailag szignifikánsan gyakrabban bukkannak fel. A morfológiai elemzőprogram szótárát az említett szokásos rövidítések mellett ki kell egészíteni az internetes zsargon szókincsével is: ezek részben angolból való átvételek (angol ill. magyar helyesírással, valamint ezek változatai, pl. *csati*), részben magyar szakkifejezések, ezek változatai (pl. *progi*), illetve rövidítések, mozaikszavak és mosolykódok (*emoticons*).

Egyes piaci helyesírás-elemzők auto-correct üzemmódban is működnek. Off-line felhasználás esetén lehetőség van a talált eltérések kigyűjtésére (javítás nélkül), ezek részben automatikus utólagos javítására, a fennmaradó példányokat azonban interaktívan kell egyértelműsíteni. A Morphologic angol nyelven készített ilyen programot, melynek a magyar nyelvre írt béta-változatát Prószéky Gábor a MNSz rendelkezésére bocsájtotta, tesztelésre.[29]

2.3 Az internetes mintavétel létjogosultsága

Felmerül a kérdés, hogy az internetes mintavétel mennyiben reprezentálja demográfiai szempontból a magyar beszélők/szövegalkotók illetve szövegbefogadók populációját. Ennek a kritériumnak az internetről építkező korpuszok egyelőre nem felelnek meg, de még kevésbé felelnek meg a hagyományos formában publikált, írott szövegekből építkező korpuszok. Szélesebb körű ugyanis a hozzáférés az interneten lévő bármely szöveghez, mint a nyomtatásban hozzáférhető szövegekhez, és a hozzáférők és felhasználók száma nő.

Egy 1997-ben és 1998-ban végzett, Magyarország felnőtt (18+) lakosságára nézve reprezentatív (n = 800) felmérés szerint internethez való *hozzáférése* 1997-ben a lakosság 10 százalékának, 1998-ban 15 százalékának volt (Angelusz-Tardos 1999). Aktív internet*használó* volt 1997-ben a lakosság 4 százaléka, 1998-ban 9 százaléka. A vizsgálat szerint az internet aktív használatát 1998-ban elsősorban az apa iskolai végzettségével mért társadalmi háttér, másodsorban az életkor, harmadsorban az iskolai végzettség befolyásolta. A háztartás vagyoni helyzete 1998-ra elveszteni látszott befolyását, míg a településtípus és a nem egyik vizsgált

[27] Hasonló problémák merülnek fel a Magyar Irodalmi és Köznyelv Nagyszótára korpuszában található történeti szövegek helyesírási sztenderdizálásakor, ld. például Kiss-Pajzs 2001.
[28] A *Journal of Sociolinguistics* 2000/4-es számát a nem-sztenderd ortográfia szociolingvisztikai vonatkozásainak szentelték (*Non-standard orthography and non-standard speech*, szerkesztő: Alexandra Jaffe).
[29] Köszönjük a támogatást!

évben sem tűnt befolyásoló tényezőnek. A szerzők megjegyzik (1999: 41), hogy a településtípus szerinti diszkrimináció meglepő hiánya mögött elsősorban a vidéki városokban élőknek a budapestiekhez való felzárkózása rejlik.

Egy másik, 1999 második felében végzett reprezentatív felmérés szerint[30] az internet*hozzáférés*sel rendelkező 14 évesnél idősebb magyarországi lakos jellemzően fiatal, közép- vagy felsőfokú végzettségű, diák vagy alkalmazott városlakó. Bővebben: 58 százalékuk harminc év alatti (e csoport aránya a lakosságban 27%), 26 százalékuk felsőfokú, 37 százalékuk középfokú végzettségű (ezen csoportok aránya a lakosságban 10, ill. 23%). A foglalkoztatás tekintetében az internethez hozzáférők 39 százaléka diák, 26 százaléka fehérgalléros alkalmazott, és 15 százaléka vállalkozó (ezek aránya a lakosságban 9, 11 ill. 7 százalék) — az internetet a megkérdezettek többsége (75%) a munkahelyéről éri el.[31] Településtípus szerint 55 százalékuk budapesti vagy nagyvárosi (arányuk a lakosságban 37%). A nemek között e felmérés szerint sincs jelentős különbség.

Optimista becslés esetén sem számíthatunk arra, hogy 2001-ben a lakosság több, mint 25-30 százaléka rendelkezik internethozzáféréssel, illetve több, mint 15-20 százaléka aktív internethasználattal.

Ugyanakkor egyrészt az internetfelhasználók száma folyamatosan nő, másrészt — ahogy a 4. táblázatban jeleztem is — mindannyian hozzáférhetnek mindazon szövegekhez, melyek a MNSz mintájában is szerepelnek. A szövegek átlagosan szélesebb olvasóközönségre számíthatnak, mint a hagyományos formában publikált szövegek, gondoljunk akár a személyes közlés (fórumok, társalgószobák), akár a hivatali nyelv (pl. pályázati kiírások, jogszabályok) csoportjában szereplő szövegekre. Ami a szövegek előállítóit illeti, az interneten kisebb a különbség a szövegalkotók és a szövegbefogadók halmaza között, mint a hagyományos formában publikált szövegek esetében; egyáltalán, a publikált műfajokat gyakran jellemző elitizmusnak nincs helye az interneten. Például bárki megjelentethet tudományos szöveget, annak értékéről nem a folyóiratok szerkesztői, hanem az olvasók döntenek. A Magyar Elektronikus Könyvtár számára például bárki felajánlhatja tudományos munkáját. Azt azonban figyelembe kell vennünk, hogy a MNSz jelenlegi gyűjtőkörében a sajtónyelvi, a szépirodalmi és a hivatali nyelvi csoportban az intézményes szűrés megvalósult, amikor ezek a szövegek az internetre felkerültek: az újságcikkeket a szerkesztők, a szépirodalmi műveket a Digitális Irodalmi Akadémia, a hivatali szövegeket az adott intézmény szűri. A MNSz második változatában ezt a problémát a személyes honlapokról, illetve a társalgószobákból letölthető szövegek beemelésével lehet kiküszöbölni.

Összefoglalva, a MNSz első változatában 2001. végén öt, egymástól igen eltérő regiszterbe tartozó, eltérő hosszúságú, összesen 180 millió szövegszónyi elő-elemzett szöveg szerepel. Ez a strukturált szövegkorpusz megfelel a csoportok közti heterogenitás és a csoporton belüli viszonylagos homogenitás elvének.[32] A MNSz első változata megfelel az általános szinkron nagykorpuszokkal szemben támasztott számos kívánalomnak, és lehetővé teszi a korpusz-alapú, más néven adat-intenzív magyar szinkrón nyelvészeti munkálatok lehetőségét.

[30] A kereskedelmi célú felmérés eredményeihez informális úton jutottam, ezért a forrást nem áll módomban megnevezni.

[31] Ezért nem mérvadó a KSH 2001. márciusi felmérése, mely szerint a magyarországi háztartások számítógépes ellátottsága 14,4%, az internethozzáféréssel rendelkezők aránya 2,3%. (Forrás: *Népszabadság,* 2001. március 3.)

[32] A homogenitás méréséről ld. Kilgarriff (2001b) cikkét.

HIVATKOZÁSOK

Angelusz Róbert - Tardos Róbert (1999): Útban az internet-galaxis felé? Tájkép az új technikák hazai expanziójáról. *Jel-Kép* 1999/2: 33-43.

Bakonyi Géza - Drótos László (1995): Netszleng. *Magyar Tudomány* 40/11: 1381-3.

Baym, N. (1996): The emergence of community in computer-mediated communication. In S. Jones (szerk.) *Cybersociety: Computer mediated communication and community*. Thousand Oaks.

Bays, Hillary (1998): Framing and face in internet exchanges: A socio-cognitive approach. *Linguistik online* 1. http://viadrina.euv-frankfurt-o.de/~wjournal/bays.html

Biber, Douglas (1990): Methodological issues regarding corpus-based analyses of linguistic variation. *Literary and Linguistic Computing* 5: 257-269.

Biber, Douglas (1993): Representativeness in corpus design. *Literary and Linguistic Computing* 8: 243-257.

Biber, Douglas - Stig Johansson - Geoffrey Leech - Susan Conrad - Edward Finegan (1999): *Longman grammar of spoken and written English*. Harlow: Longman.

BNC Online (1997) http://info.ox.ac.uk/bnc/what/spok_design.html

Burnard L (1995) *The BNC handbook*. Oxford, Oxford University Press.

Cavar, Damir - Alexander Geyken - Gerald Neumann (2000): Digital Dictionary of the 20th Century German Language. In: Tomaz Erjavec - Jerjena Gros (szerk.) *Jezikovne Tehnologije - Language technologies. Proceedings of the Information Society 2000 multi-conference*. Ljubljana, 112-115.

Cherny, Lynn (1999): *Conversation and community: Chat in a virtual world*. Stanford CA: Center for the Study of Language and Information.

Collot, Milena - Nancy Belmore (1996): Electronic language: A new variety of English. In Susan Herring (szerk.), 13-28.

Francis, W. Nelson - H. Kucera (1964/1979): *Manual of information to accompany a standard corpus of present-day edited American English, for use with digital computers.* Department of English, Brown University.

Fujii, Atsushi - Tetsuya Ishikawa (2000): Utilizing the world wide web as an encyclopaedia: Extracting term descriptions from semi-structured text. In: *Proceedings of the 38th meeting of the ACL*. Hong Kong, 488-495.

Füredi Mihály - Kelemen József (1989): *A mai magyar nyelv szépprózai gyakorisági szótára*. Budapest: Akadémiai.

Gousseva, Julia (1998): An experience in cyberspace communication: Listserv interaction in a freshman composition class. *Linguistik online* 1. http://viadrina.euv-frankfurt-o.de/~wjournal/gousseva.html

Harrison, Sandra (1998): E-mail discussions as conversation: Moves and acts in a sample from a listserv discussion. *Linguistik online* 1. http://viadrina.euv-frankfurt-o.de/~wjournal/harrison.html

Hentschel, Elke (1998): Communication on IRC. *Linguistik online* 1. http://viadrina.euv-frankfurt-o.de/~wjournal/irc.html

Herring, Susan (1996) (szerk.): *Computer-mediated communication: Linguistic, social and cross-cultural perspectives*. Amsterdam: Benjamins.

Holm, Kurt. (szerk.).(1975): *Die Befragung 1*. München: A. Franke Verlag.

Ide, Nancy - Catherine Macleod (2001): The American National Corpus: A standardized resource for American English. In: Rayson et al. (szerk.), 274-280.

Ilson, Robert (1991): *Assembling, analysing and using a corpus of authentic language (A lecture given on the Survey of English Usage at the Linguistics Institute of the Hungarian Academy of Sciences on 2 September, 1988)*. Budapest: MTA Nyelvtudományi Intézet.

Johansson, Stig (1980): The LOB Corpus of British English texts: Presentation and comments. *ALLC Journal* 1: 25-36.

Johansson, Stig - Geoffrey Leech - H. Goodluck (1978): *Manual of information to accompany the Lancaster-Oslo/Bergen corpus of British English, for use with digital computers.* Oslo: Department of English, University of Oslo.

Kennedy, Graeme (1998): *An introduction to corpus linguistics*. London-New York: Longman.

Kilgarriff, Adam (2001a): The Web as corpus. In Rayson et al. (szerk.), 342-344.

Kilgarriff, Adam (2001b): Comparing corpora. *Computational Linguistics*. Megjelenés elõtt.

Kiss Gabriella - Pajzs Júlia (2001): An attempt to develop a lemmatiser for the Historical Corpus of Hungarian. In Rayson et al. (szerk.), 443-451.

Miller, Hugh (1996): *The presentation of self in electronic life: Goffman on the internet*. www.ntu.ac.uk/soc/psych/miller/goffman.html

Németh Géza - Zainkó Csaba (2000): Statisztikai szövegelemzés automatikus felolvasáshoz. In Gósy Mária (szerk.): *Beszédkutatás 2000: Beszéd és társadalom. A „Beszédkutatás 2000" tudományos ülésszakon*

elhangzott előadások válogatott és átdolgozott tanulmányai. Budapest: MTA Nyelvtudományi Intézet, 156-166.

Paollilo, John (1999): The virtual speech community: Social network and language variation on IRC. *Journal of Computer-Mediated Communication* 4: 1-20. (www.ascusc.org/jcmc/vol4/issue4/paolillo.html)

Radev, Dragomir - Kathleen McKeown (1997): Building a generation knowledge source using internet-accessible newswire. In *Proceedings of the 5th applied natural language processing conference.* Washington, 527-534.

Rayson, Paul - Andrew Wilson - Tony McEnery - Andrew Hardie - Shereen Khoja (szerk.) (2001): *Proceedings of the Corpus Linguistics 2001 conference.* Lancaster: Lancaster University.

Reményi Andrea Ágnes (2001): Use logbooks and find the original meaning of „representativeness". In Rayson et al. (szerk.), 485-491.

Sinclair, John (1991): *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Sinclair, John (1992): The automatic analysis of corpora. In Jan Svartvik (szerk.): *Directions in corpus linguistics*. The Hague: Mouton de Gruyter, 145-55.

Summers, Della (1991): *Longman/Lancaster English Language Corpus: Criteria and design*. Harlow: Longman.

Svartvik, Jan - Randolph Quirk (szerk.) (1980): *A corpus of English conversation.* Lund: Lund University Press.

Szakadát István (2000): *A digitális kultúra és világháló mint alternatív nyilvánosság, ennek folyamatai, hatásai, szabályozási módjai, anticipálható politikai következményei*. Kézirat, 10 oldal.

Váradi Tamás (1998): *Nyelv és korpusz: A reprezentativitás a korpusznyelvészetben*. Kézirat, 14 oldal.

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

Varantola, Krista (2000): Translators and disposable corpora. In *Proceedings of CULT (Corpus Use and Learning to Translate)*. Bertinoro.

Volk, Martin (2001): Exploiting the WWW as a corpus to resolve PP attachment ambiguities. In Rayson et al. (szerk.), 601-606.

Werry, Christopher (1996) Linguistics and interactional features of Internet Relay Chat. In Susan Herring (szerk.), 47-64.

Yates, Simeon (1996): Oral and written linguistic aspects of computer conferencing. In Susan Herring (szerk.), 29-46.

1. MELLÉKLET: Néhány nagykorpusz kategóriarendszere

A Brown- és a LOB-korpusz szövegkategóriái

		szövegkategória	szövegek száma: Brown	LOB
I	A	sajtó (riport)	44	44
	B	sajtó (szerkesztőségi)	27	27
	C	sajtó (recenzió)	17	17
	D	vallás	17	17
	E	szakma, hobbi, szabadidő	36	38
	F	általános ismeretek	48	44
	G	szépirodalom, életrajz, esszék	75	77
	H	vegyes (kormányiratok, hivatalos jelentések, egyetemi katalógusok)	30	30
	J	tudományos	80	80
II	K	általános szépirodalom	29	29
	L	krimi	24	24
	M	sci-fi	6	6
	N	kaland, western	29	29
	P	szerelmes	29	29
	R	humor	9	9
		összesen	500	500

Kennedy (1998: 24-26) és Váradi (1998: 11) alapján[33]

[33] A fordításban többnyire Váradit (1998) követtem.

A Survey of English Usage szövegkategóriái

	előre megírt, felolvasott szövegek	18
W.1	előadások	6
W.2	hírek	3
W.3	rádiós szépirodalom	2
W.4	politikai beszédek	3
W.5	drámák	4
	nem publikált írottnyelvi	36
W.6	vizsgadolgozatok, kézírásos jegyzetek	11
W.7	levelek (társas és üzleti)	21
W.8	naplók	4
	nyomtatásban megjelent	46
W.9	tudomány	13
W.10	tankönyvek, kezelési útmutatók	6
W.11	általános ismeretek	
W.12	sajtóhírek	8
W.13	helyi sajtó; szerkesztőségi anyagok	
W.14	jogi szövegek	3
W.15	meggyőzési célú írások (írott prédikáció, kiáltványok, hirdetések)	5
W.16	szépirodalom	7
	beszéltnyelvi szövegek	100
S.1-2	rejtett mikrofonnal felvett beszélgetések egyenrangúak között	
S.3	rejtett mikrofonnal felvett beszélgetések nem egyenrangúak között	34
S.4-5	nem rejtett mikrofonnal felvett beszélgetések egyenrangúak között	
S.6	nem rejtett mikrofonnal felvett beszélgetések nem egyenrangúak között	26
S.7	telefonbeszélgetések közeli kapcsolatban lévő felek között	
S.8	telefonbeszélgetések egyenrangúak között	16
S.9	telefonbeszélgetések nem egyenrangúak között	
S.10	rádió- és TV-közvetítések	8
S.11	spontán monológok, beszédek	10
S.12	előkészített monológok (előadások, beszédek ...)	6

Ilson (1991: 9-14) és Kennedy (1998: 18) alapján; a beszéltnyelvi szövegeket átvette a London-Lund korpusz

A Longman Corpus Network szövegkategóriái

1	természettudományok	6%
2	alkalmazott természettudományok	4.3%
3	társadalomtudományok	14.2%
4	nemzetközi ügyek (történelem, politika, gazdaság...)	10.4%
5	kereskedelem, pénzügy	4.4%
6	művészetek	7.9%
7	hitek, gondolkodás (filozófia, vallás, mitológia...)	4.7%
8	szabadidő	5.7%
9	szépirodalom	40%
10	költészet, dráma, humor	2.3%

Kennedy (1998: 49) alapján

A British National Corpus szövegkategóriái

		informatív	természettudományok	a publikált szövegek	4%
Í	T	(az írott	alkalmazott tudomány	felét véletlenszerűen	11%
		korpusz	társadalomtudományok	választották ki:	15%
R	É	78%-a;	nemzetközi ügyek	produkció/recepció;	15%
		mind	kereskedelem, pénzügy	a másik felét a	9%
O	M	1975 utáni)	művészetek	szelekciós jegyek	8%
			hit, gondolkodás	(közeg, idő, téma)	5%
T	A		szabadidő	szerinti rétegzés	11%
		szépirodalom	(mind 1960 utáni)	után	20%
T	KÖZEG		könyv		60%
			folyóirat		30%
			egyéb publikált		5%
			nem publikált		5%
90%			előre megírt, felolvasásra szánt (beszédek, drámák)		2%
B	kontextus		pedagógiai és informatív (iskolai, egyetemi)		
E	alapján kivál.		üzleti (bemutatók, tárgyalások, interjúk)		
S	(12 régió);		hivatalos,közéleti események (vallási,politikai)		60%
Z	40% monológ,		szabadidős (sportközvetítés, betelefonálós műsor)		
É	60% dialógus				
L	demográfiai		nem	2-7 nap magnóval	
T	minta		életkor (15 - 60+)	jár; az interakciók	
	n = 124		szocio-ökonómiai csoport	részleteit	40%
10%	700 óra		38 régió	is feljegyezték	

Burnard (1995: 11-25) és Kennedy (1998: 50-53) alapján

2. MELLÉKLET

Néhány angol nyelvű elektronikus nagykorpusz

A korpusz neve	típus	csatorna (í/b)	nyelv	év*	mintavétel	össz. szövegszó	alapirodalom	hozzáférés
Brown Corpus	referencia	írott	amerikai angol	1961	rétegzett (műfaj) véletlen	1m (500x2.000)	Francis-Kucera 1964/1979	ICAME
Lancaster-Oslo/Bergen Corpus	referencia	írott	brit angol	1961	rétegzett (műfaj) véletlen	1m (500x2.000)	Johansson et al. 1978	icame@hd.uib.no
London-Lund Corpus	referencia	beszélt	brit angol	1953-87	nem véletlen (sok tudományos szöveg)	500.000 (100x5.000)	Svartvik-Quirk 1980	www.hd.uib.no
Cobuild Corpus (Birmingham) Reserve TEFL Corpus folyt.: The Bank of English	referencia monitor	75% í -25% b	főleg brit angol a. nyelvkönyvek	1980-82 -1987 1990-	diskurzus-funkció témaváltozatosság	7,3m 13m 1m 300m (1997)	Sinclair 1991	titania.cobuild.collins.co.uk/boe_info.html
Longman Corpus Network - Longman/Lancaster EL Corp. - Longman Spoken Corp. - Longman Corp. of Learners' E		írott-beszélt írott beszélt írott	angol mint első nyelv angol mint id. ny.	1980-as évek vége	tematikus műfaj,tudásszint, anyany.	50m 10m	Summers 1991	www.longman-elt.com/dictionaries/corpus/lccont.html
British National Corpus	referencia	90% í -10% b	egynyelvű brit angol	informatív 1975-, szépirod. 1960-	írott: véletlen (kiadói-befogadói oldal) vagy szelekciós jegyek; beszélt: demográfiai minta vagy kontextus alapján	100m (4m spontán beszélt: 4.124 szöveg, legtöbb < 40.000); 6m kontextus alapján)	Burnard 1995	natcorp@oucs.ox.ac.uk http://info.ox.ac.uk/bnc
International Corpus of English	referencia	40% í- 60% b	angol mint 1-2. ny	1990-96		20 alkorpusz x 1m, mind 500x2.000		www.ucl.ac.uk/english-usage/ice/index.htm
Longman Spoken & Written EC	referencia	írott-beszélt	amerikai & brit a.		regiszter, dialektus	40 m	Biber et al.1999	
American National Corpus	referencia monitor	írott-beszélt	amerikai angol	1990-	mint BNC változatosság, hozzáférés	100m + 10% 5 évente	Ide-Macleod 2001	www.cs.vassar.edu/~ide/anc/

* A mintavétel éve(i)