ÉLŐ NYELV

Gondolatok a Kárpát-medencei magyar nyelvi korpusz bővítéséről*
A magyar nyelv „határtalanításának” egyik újabb eredménye

D) P r o b l é m á k . – Az előző fejezetben felvázolt alapkódolás az egyes régiókban eltérő gyorsasággal, eltérő módszerekkel, illetve eltérő számítógépes programokkal valósult meg. (A végeredmény azonban minden kutatóállomáson azonos volt: ez garantálta az egységes kimenetet.) Az eltérő módszerek természetesen később a munkafolyamatban eltérő problémákat okoztak. Ezek megvitatásával és megoldásával több csatornán próbálkoztunk. Erre szolgáltak a már említett korpusznyelvészeti tréningek, továbbá az irodák közös megbeszélései, az illyefalvi találkozók, illetve tájékoztató céllal jött létre a Kmmnyk. határon túli korpuszának honlapja (http://corpus.nytud.hu/mnszworkshop/index.html), valamint az egymás közti kommunikáció elősegítése végett, az irodák közös ügyeinek megvitatására létrehozott „nyelvészet-levelezőlista” vagy „nyelvésznet” is. A felmerülő kérdések megválaszolásában a közös fórumok mellett elsősorban a Nyelvtudományi Intézet Nyelvtechnológiai Osztályának munkatársai (ORAVECZ CSABA és VÁRADI TAMÁS) segítettek.

A határon túli korpusz sajátos természetű problémája az élőnyelvi alkorpusz. A probléma alapját az élőnyelvi szövegek lejegyzését elősegítő egységesített lejegyzési útmutató elkészítésének csúszása jelentette. A kutatóhálózat megbeszéléseiről készült emlékeztetők tanúsága szerint már 2002 májusában szó esett az élőnyelvi lejegyzés elkészítéséről, az arra szóló megbízásról. Ez kommunikációs és egyéb (szervezési) problémák miatt sajnos csak 2005 decemberében készült el. Az élőnyelvi szövegek lejegyzésének esszenciája az egységes kódolás. Az alkorpusz létrehozásának csak akkor van értelme, ha minden régióban azonos minta alapján történik a lejegyzés. Mivel az összes határon túli régió egy közös szövegtár anyagát bővíti, ezért a régiókban készülő anyagok végső formátumának kivétel nélkül azonosaknak kell lenniük, hogy a szövegekben történő egységes kereshetőséget biztosítsák. Ez azonban csak akkor valósulhat meg, ha előzőleg a szövegek azonos rendszer alapján voltak kódolva. Ilyen megfontolásból tehát különböző kódolási minták használatának nem lett volna értelme: pontosan a határon túli korpusz alapgondolatát, a különböző régiók nyelvi anyagában történő egységes keresést akadályoznák meg. Ez természetesen még nem zárja ki az egyes irodákban felmerülő, az alapkódoláson túli további, speciális kódolást, mivel minden iroda saját akarata szerint tovább kódolhatja a szövegeket. Az alapkódolásnál részletesebb anyag sorsa azonban még nincs tisztázva. Ez vagy a korpusz része lesz, vagy nem kerül a többi, alapkóddal ellátott szöveg közé, és csupán az iroda saját korpuszát fogja gyarapítani.

Az egységes lejegyzési útmutató elkészítésében minden iroda szabad kezet kapott. A lejegyzendő hangtani jelenségek összeállítása feladata lett volna minden irodának: a közös megegyezések értelmében elsődlegesen egy nyers változat készült volna el, amely tartalmazta volna az irodák által fontosnak tartott élőnyelvi jelenségek lejegyzésére vonatkozó javaslatokat. Az irodák által összeállított lejegyzési útmutatót később KASSAI ILONA egységesítette volna. Sajnos félreértések miatt a lejegyzési útmutató összeállításának ez a terve

* L. MNy. 2008: 81–9.

nem valósult meg. A kutatóhálózatból – LANSTYÁK ISTVÁN munkájának köszönhetően – csupán a Gramma Nyelvi Iroda tette meg javaslatát. Mivel a LANSTYÁK által összeállított kódolási útmutató (ennek egy korábbi változatát l. LANSTYÁK 2004: 181–5) – idő hiányában – hosszúnak és bonyolultnak bizonyult, ezért a Gramma Nyelvi Iroda előállt egy rövidebb és számítógépes szempontokat is figyelembe vevő javaslattal. A többi iroda közül később csupán a vajdaságiak tettek javaslatot (RAJSLI 2004: 65), azonban ez nem felelt meg az előzőleg meghatározott követelményeknek. (Az általuk készített útmutató inkább dialektológiai leírást, a vajdasági nyelvváltozatok sajátos elemeinek leírását, és nem egy általános élőnyelvi lejegyzést takar: ezt mutatja az is, hogy helyspecifikus és nem általános jelenségeket tartalmaz.) Mivel így a szövegtárral foglalkozó négy régióból csupán egyikük javaslata volt használható, a szervezők KASSAI ILONÁt kérték fel egy alkalmazható lejegyzési útmutató elkészítésére. KASSAI 2006 elejére készítette el az útmutatót, mely nagy részben a fent említett LANSTYÁK által készített lejegyzési útmutatón alapszik.

Az élőnyelvi szövegek lejegyzésének problémája napirenden volt az irodák találkozóin; 2004 júliusában Illyefalván is felvetődött. Az irodák és az MTA Nyelvtudományi Intézetét képviselő ORAVECZ CSABA akkor abban egyeztek meg, hogy amíg a lejegyzést végzők nem kapnak közös lejegyzési útmutatót, elegendő lesz, ha a meglévő szövegeket valamilyen editorban (.txt-fájlként) standard helyesírással lejegyezik, s így – ideiglenesen – ez képezné a későbbi feldolgozás alapját (a standard helyesírást annak egységes jellege miatt választottuk). A kódolás formája mellett egyezség született a lejegyzendő szöveg típusait illetően is. Az egyezség szépséghibája, hogy a 2004-es illyefalvi találkozón a négy iroda közül csupán a szervezők (Szabó T. Attila Nyelvi Intézet) és a Gramma Nyelvi Iroda képviseltette magát. Örvendetes azonban, hogy a nyelvi irodák (kutatóállomások) mellett képviseltette magát az őrvidéki (Ausztria) és a muravidéki (Szlovénia) kutatóhely is. (Sajnálatos módon az illyefalvi egyezmények korpusznyelvészeti teendői csupán két iroda megbeszélései után jöttek létre, a kárpátaljai – Hodinka Antal Intézet – és a vajdasági – Vajdasági Magyar Nyelvi Korpusz – kutatóállomások később hagyták jóvá azokat.)

A beszélt nyelvi korpusszal kapcsolatosan az irodák munkatársai 2004-ben a következőkben egyeztek meg:

1. A lejegyzendő hangfelvételek nem lehetnek az 1990-es éveknél korábbiak.

2. A standard mellett dialektusoknak is helyet kell adni a hangfelvételek között, ezek a dialektusok azonban csupán a főbb nyelvjárási területeket képviselhetik. A korpuszba kerülő egyes dialektusok arányát az azokat beszélők arányából kell kiszámolni. A nyelvjárási hanganyagnak nemcsak informális beszélgetéseket, hanem formális regisztereket is kell tartalmaznia (pl. ritualizált szövegek, élettörténetek). A nyelvjárási hanganyag az egész anyag 40–50%-át teheti ki.

3. A felvételek között formális (pl. műszaki, orvosi, humán szövegek; konferenciák, prédikáció, tanári magyarázat, politikai nyilatkozat, önkormányzati ülés) és informális (különféle beszélgetések, pl. bolti) regiszterekhez tartozó standard szövegek is legyenek. A dialogikus és informális regisztereknek kell többségben lenniük, az összes 70–80%-át kell alkotniuk.

4. Kétnyelvűségi típusok: a magyardomináns kétnyelvű beszélőktől származó hangfelvételek az anyag 40–50%-át, az államnyelvi domináns beszélőktől származó felvételek az anyag 35%-át, egynyelvű beszélők hanganyagának az egész 15%-át kell alkotnia.

5. Az adatközlők kiválasztásának szempontjait hierarchizálni kell.

6. Korcsoportok: gyerekekre és idős adatközlőkre is szükség van. A gyerekek képviselhetik az informális, egynyelvű, az idősek a nyelvjárási beszélőket.

7. Az egyes digitalizált hangfájlokhoz és a hozzájuk tartozó lejegyzett szöveghez csatolni kell fejlécet is, amit célszerű lenne külön fájlban tárolni. Ennek a fejlécnek a következő adatokat kellene tartalmaznia: a felvétel időpontja, a felvételt készítő személy neve; az adatközlő neve, neme, életkora, foglalkozása, születési helye, lakóhelye, hol élt többet: városban/faluban, családi állapota; az általa elsajátított nyelvek, a családjában használt nyelvek; téma, szituáció, a jelen levő személyek száma, azok és az adatközlő közti viszony jellege; rádióban elhangzott felvételek esetében: élő műsor vagy felvett műsor, nyers vagy javított felvétel; a hangfájl helye a számítógépen (annak elérési mutatója), a fájl formátuma, a fájl száma.

Ott, ahol lehetett, igyekeztük az egyes szövegtípusok százalékos arányát is meghatározni. Mivel tisztában voltunk vele, hogy az arányok betartása nehéz feladat, ezért úgy határoztunk, hogy a megállapított arányoktól minden iroda 10%-kal eltérhet.

Bár az anyaggyűjtéshez tartozik, mégis itt szólnék a hivatali nyelvet és a személyes közlést (amely magában foglalja a beszélt nyelvi szövegeket) bemutató alkorpuszról. A két alkorpusz gyűjtése két különböző problémát vet fel. A határon túli magyar hivatali nyelvvel kapcsolatban két kérdés merül fel. A hivatali írásbeliség leggyakrabban formanyomtatványok formájában van jelen, ezek pedig leggyakrabban a magyarországi nyomtatványok formahű átvételei. Ezek esetében tehát nem beszélhetünk szlovákiai magyar vagy romániai magyar hivatali nyelvről. A magyarországi minták követését illetően jó lenne különbséget tenni a beszélt és írott nyelvváltozatok között, hiszen nyilvánvaló, hogy az írott nyelvváltozat jobban közelít majd a standard formákhoz, illetve a magyarországi mintákhoz, míg a beszélt változat erősebben tükrözi a kétnyelvű beszédkörnyezetben élő kontaktusváltozatokat. (Egy későbbi változatban talán jó lenne megkülönböztetni egy írott és egy beszélt hivatali nyelvet bemutató alkorpuszt.) A kisebbségi régiók hivatali nyelvének egy másik sajátossága a megvalósulásuk sokfélesége. Mivel a hivatalos dokumentumok (legyen az fordítás vagy eredeti szöveg) kiadása nem centralizált, így gyakori jelenség egy régión belül is, hogy ugyanannak a dokumentumnak különböző településeken eltérő formája van. A kutatóhálózat egyik szerepe éppen a hivatalos dokumentumok, formanyomtatványok központosítása, a jogi-közigazgatási terminológia egységesítése és az adott régió magyar nyelvű hivatalos írásbeliségének kialakítása.

A beszélt nyelvi alkorpusz elkészítése szintén két alapvető kérdést vet fel. A Magyar nemzeti szövegtár anyagaiból és elveiből kiindulva, ennek az alkorpusznak tartalmaznia kellene egy élőnyelvi lejegyzéseket magában foglaló beszélt nyelvi részt, illetve a beszélt nyelvhez közelítő, gyors beszédfordulókból álló csetfórumok anyagát (ezt nevezhetjük személyes közlésnek is). Mivel az élőnyelvi anyagok problémájáról már szóltam, most csak a személyes közlésekkel foglalkozom. Sajnos egyik régióban sem találtunk megfelelő fórumot, ezért a határon túli alkorpusz „személyes közléseket” magában foglaló része tartalmában eltér majd a magyarországitól (pl. emlékezések, magánlevelek). A beszélt nyelvet és a személyes közlést bemutató korpusz esetében előre meg kellett volna határozni a belső struktúrát és arányokat, azonban erre nem került sor. A két alkorpuszról összegezve elmondható, hogy egyik esetben sem teljesítik majd a szerkesztők által meghatározott legalább 10%-os arányt. Ennek okai összetettek: kereshetjük a nyelvi valóságban és az irodákban is.

Valódi problémát jelent a százalékos arányok betartása is, hiszen ez nem minden alkorpusz esetében kivitelezhető. Az előzetes megállapodások értelmében az egyes határon túli alkorpuszok szerkezeti egységei (szépirodalom, tudományos próza, sajtó, hivatalos nyelv, személyes közlés) azok legalább 10%-át kellett, hogy alkossák. Ez a 10%-os határ azonban nem minden alkorpusz esetében volt megvalósítható; leginkább a hivatalos nyelvváltozatot és a személyes közlést tartalmazó alkorpuszok esetében nem. Ennek oka, hogy a hivatalos nyelvet bemutató alkorpusz esetében nem találtunk megfelelő mennyiségű anyagot. Ebben a pontban a valóság „nem felelt meg az eredeti elképzeléseknek", hiszen a kisebbség nem „termel" akkora mennyiségű hivatalos iratot, mint az elvárható lenne, illetve ennek összetétele is – a tudományos prózához hasonlóan – kevésbé hivatalos anyagokkal van vegyítve. Átmenetileg problémát jelent a személyes közlés alkorpusz is: ennek legalább két részből kellene állnia – egyik része a gyors beszédfordulókból álló csetfórumok szövege, a másik a beszélt nyelvi szövegek lejegyzett változata. A határon túli magyar csetfórumok a magyarországiakhoz képest alulreprezentáltak, így nehezebb a kellő (arányaiban megfelelő) mennyiségű szöveget összegyűjteni. A beszélt nyelvi szövegek folyamatosan bővíthetőek, de csupán azután, hogy az irodák kellő gyakorlatot szereztek a lejegyzési útmutató használatában. Így a 10% elméletileg elérhető (vagy inkább csak e l k é p z e l h e t ő), ám mivel a többi alkorpusz is gyarapszik, ennek esélye egyre kevesebb (a hivatalos nyelvi szövegek esetében inkább elképzelhetetlen).

E) W o r d j e c t . – Végül szólnék még a kutatóhálózat legfrissebb vállalkozásáról, a MorphoLogic Kft. által gyártott magyar nyelvű helyesírás-ellenőrző és nyelvhelyesség-ellenőrző (a továbbiakban csak: helyesírás-ellenőrző) programcsomag határon túli magyar anyagának összeállításáról (gyűjtés és kódolás). Ez a program a Microsoft Office termékcsomagban használatos Windows Word, illetve Quark XPress helyesírás-ellenőrzőjeként ismeretes, de korpuszelemzőként is működik. A program fő célja, hogy jelezze a szövegben előforduló elütéseket és hibás szavakat. A termék felhasználhatósága azonban ezen túlmutat, hiszen rendelkezik egy, a nagyközönség által kevésbé ismert funkcióval is: a nyelvhelyesség-ellenőrzés alapja egy magyar nyelvre alkalmazott morfológiai generáló–elemző motor (HUMOR), amely számítógépen tárolt korpuszok nyelvi elemzésére is alkalmazható. Mivel ezeket a műveleteket nem ember, hanem gép végzi, ezért „taníthatósága" eléggé korlátozott: csak meglévő nyelvtani szabályok és kész szótár alapján tud generálni, illetve elemezni. Ez azt jelenti, hogy csak azokat a szavakat fogadja el helyesnek, amelyek az ellenőrző szótárában megtalálhatók (amelyeket a morfológiai elemzőprogram generál); ez lehet vagy az alapcsomag szótára, vagy a felhasználó által összeállított ún. sajátszótár. Az alapcsomag szótárát a MorphoLogic Kft. állítja össze, így ezt minden általuk terjesztett helyesírás-ellenőrző tartalmazza. Ez akár több millió felhasználót is jelenthet, ha figyelembe vesszük a számítógépen magyar nyelven írók számát. A leírtakból következik, hogy feltehetően ma ez a Magyarországon leggyakrabban használt szótár (bár a felhasználók valószínűleg nem tudnak erről). Az alapszótár csak Magyarországon készített szótárakból áll, így érthető, hogy nem tartalmaz anyagot a magyar nyelv határon túli változataiból. (Bár az elemző legújabb változata tartalmazza az „Értelmező kéziszótár" második kiadását és az Osiris Kiadó Helyesírását.)

A szövegszerkesztőkbe épített helyesírás-ellenőrző aláhúzással jelzi, hogy a felhasználó „valószínűleg" hibás szót írt le, vagy egyéb nyelvhelyességi hibát vétett. A zöld hullámvonallal történő aláhúzás általában nyelvhelyességi vagy szövegszerkezeti hibát jelöl: pél-

dául szóközök *(felesleges szóköz)*, mondathatár ellenőrzése *(! ez egy új mondat.)* vagy trágár kifejezések megjelölése *(szar)*. Ez valójában érdektelen a magyar nyelv állami vagy határon túli változatainak megítélése szempontjából, hiszen a szövegszerkezeti sajátosságok és az elemző által kezelt stilisztikai apróságok minden magyar nyelvváltozatra egyformán érvényesek. A piros hullámvonallal történő aláhúzás a helyesírás-ellenőrző által nem ismert szavak megjelölését jelenti. Minden olyan szót aláhúz, amelyet sem az alapszótárban, sem a sajátszótárban nem talál meg. Mivel a határon túli magyar nyelvváltozatok nem részei a szótárnak, így minden határon túli magyar közszót és a helységnevek túlnyomó többségét aláhúzza, azaz hibás szónak minősíti. Az már tudományos közhelynek számít, hogy a magyar nyelvközösség normatív beállítottságú, azaz a nyelvészektől, szótáraktól kapott információt általában mérlegelés nélkül elfogadja – mivel az úgyis szakemberektől származik. Ebben a folyamatban nagy szerepet játszik a helyesírás-ellenőrző is, hiszen egy ilyen széles körben használt termék (szótár) nem hibázhat. Tehát a nyelvhelyesség-ellenőrző minősít: a Magyarország határain kívüli magyar településnevek esetében gyakori, hogy a szótár nem ismeri a helységnevet, ezért hibának minősíti azt. Ez azonban régi és/vagy széles körben ismert magyar településnevek esetében kétszeresen is bántóan hathat, hiszen ilyenkor az elemző akaratlanul is a magyar nyelv olyan elemeit stigmatizálja, amelyek annak „teljes jogú" és gyakran használt részei és a magyar kultúra alapelemei, például *Huszt*, *Ilosva* stb.

Nyilvánvaló, hogy a magyar nyelv ellenőrzésére legszélesebb körben használt nyelvhelyesség-ellenőrző alapszótára kiegészítésekre szorul. Az azonban nem várható el a magyarországi nyelvészektől, hogy többletenergiát belefektetve felgyűjtsék termékeikbe a magyar nyelv határon túli elemeit, valamint megfelelően kódolják is azokat.

Azon kívül, hogy az alapszótár bővítése árnyaltabbá tenné a helyesírás-ellenőrző munkáját, teljes mértékben elemezhetővé tehetné a Kárpát-medencei magyar nyelvi korpusz határon túli alkorpuszát is, amely a határon túli magyar nyelvváltozatok sajátos lexikai elemei miatt jelenleg csak részben elemezhető.

A szótár bővítése az MTA Határon túli irodáinak munkatársaitól két munkafolyamatot követel meg:

1. Az alapszótárba bekerülő szavak kiválasztása. – A válogatás közben mindvégig szem előtt kell tartani, hogy a szövegszerkesztőt használók legnagyobb része magyarországi magyar beszélő, illetve hogy az elemzőt – írott szövegek elemzése miatt – magasabb fokú normativitással rendelkező nyelvváltozatok (szövegek) elemzésére tervezték (nem pedig nyelvjárási vagy regionális köznyelvi szövegekre). Ebből az következik, hogy a felgyűjtött szavaknak túl kell mutatniuk a regionalitáson (ideális esetben az egész magyar beszélőközösségben azonosan használt szavaknak kellene lenniük) és – legalább az állami változatok szintjén – normatívnak kell lenniük. Ezeknek a követelményeknek leginkább a tulajdonnevek, illetve a közvetlen kölcsönszók (idegen nyelvből átvett idegen szavak: *cujka, zmizik* stb.) felelnek meg. Az utóbbiaknak nagy szerepük van az összetett szavak elemzésében, mivel csak azt az összetett szót fogadja el helyesnek a program, amelyet vagy tartalmaz a szótár, vagy össze tudja azt rakni a meglévő elemekből. Terveinkben a következő típusú szavak gyűjtését kívánjuk megvalósítani: a) földrajzi nevek, b) vezetéknevek, c) keresztnevek, d) közvetlen kölcsönszavak, e) magyar eredetű közvetett kölcsönszavak.

2. Az összegyűjtött anyag előkódolása. – A gondosan megfogalmazott követelmények szerinti gyűjtés utáni következő lépés a kész szólisták kódolása. Ez alapján később minden szó hovatartozása egyértelműsíthetővé válik, valamint a morfológiai

kódok alapján a szavak az elemzőbe is beépíthetőek lesznek. Annak illusztrációjaként, hogy hogyan épül fel a szótár, vegyük az őrvidéki Sopronkeresztúr példáját (ezt egyébként értelemszerűen az elemző pirossal aláhúzza, hiszen az adott toponimát a szótár nem ismeri): Sopron+kereszt+úr[FN|pse];nyv:őv;rp.

Jelölni kell tehát az összetételi határt (a + jel jelöli), mivel a szó végi toldalékoláskor módosulhat a szótest (a szó elejére kerülő elemek esetében természetesen nem); hogy milyen szófajú az elem (*FN*, azaz főnév); hogy a szófajon belül milyen altípusba tartozik (*pse*, azaz helynév); hogy melyik állami változat eleme (*nyv:őv*, azaz őrvidéki nyelvváltozat); hogy szótő-e vagy toldalék (*rp*, azaz jobbra bővülő, tehát szótő), illetve főnevek esetében az egyes szám harmadik személyű birtokos személyjeles alakját is (a példában nincs semmi, azaz *Sopronkeresztúrja* a kívánt alak). A melléknevek esetében többletként jelölni kell a melléknév essivusi alakját (ESS_Ul, azaz *sopronkeresztúriul*): sopron+kereszt+úr@i[MN|pse];nyv:őv;rp:Ess_Ul.

A munka első fázisában a helységneveket és az egyéb földrajzi neveket (folyók, térségek stb. nevei) gyűjtjük össze, s a gyűjtés, illetve kódolás tapasztalataiból kiindulva folytatjuk majd a személynevekkel és a köznevekkel. A köznevekre vonatkozóan már vannak tapasztalataink, amelyet az ún. ht-adatbázis (azaz „a határon túli vonatkozású magyar szókészleti elemek listája") összeállításával szereztünk és szerzünk folyamatosan (az adatbázis bárki számára – regisztráció után – elérhető a http://nytud.hu címen). Furcsa helyzet, de ez esetben nem is a gyűjtés, hanem a válogatás jelent majd problémát. Bár a MorphoLogic Kft.-től szabad kezet kaptunk az anyag mennyiségi és minőségi kritériumainak meghatározására, mégsem vehetünk fel minden szót, hiszen egyebek mellett azt is figyelembe kell vennünk, hogy az egyes határon túli szócsoportok a magyarországiakhoz viszonyítva ne legyenek túlreprezentálva – az például nagyon furcsa lenne, ha a program szótára több határon túli helységnevet tartalmazna, mint magyarországit.

5. Összefoglalás. – Háromévnyi munka után elkészült a Kárpát-medencei magyar nyelvi korpusz határon túli alkorpusza. Annak ellenére, hogy az anyag csupán töredéke a magyarországinak, mégis jelentős előrelépés a magyar nyelvű korpuszok terén, hiszen ezzel a Nyelvtudományi Intézetben olyan korpuszt alkottak, amely már a határon túli magyar nyelvváltozatokat is magába foglalja, lehetővé téve ezzel akár az összehasonlító kutatásokat is.

A Kmmnyk. létrejöttével azonban még nem zárultak le a munkálatok. Egyelőre két kérdés maradt megválaszolatlanul. Az élőnyelvi szövegek átírása és annotálása még mindig nem zárult le; hátra van még a munka összehangolása, azaz a már elkészített lejegyzések egységesítése, illetve annotálása. Ez azt is jelenti, hogy a korpuszépítés folytatódik, viszont a további lépések egyelőre nem egészen világosak. Kérdéses, hogy a közeljövőben határon túli magyar nyelvváltozatokat tartalmazó Kmmnyk. határon túli anyagát érintő munkálatok folytatódnak-e. Ennek eldöntése főként VÁRADI TAMÁSon és az MTA Nyelvtudományi Intézetének Nyelvtechnológiai Osztályán múlik, hiszen a projektet szakmailag ők irányítják. Bárhogy alakuljon is a pályázat jövője, a kutatóállomások továbbra is folytatják az anyagok gyűjtését, mivel mind a négy kutatóállomás a saját régiójában elindította regionális korpuszának építését, illetve pályázott a Wordject-projekt elkészítésére. Ha azonban az MTA Nyelvtudományi Intézetének felügyeletében nem valósul meg egy újabb közös projektum, akkor elképzelhető, hogy a kutatóállomásokon folyamatosan gyűlő anyag egymástól eltérő

formájú lesz (Bár egyelőre az sincs kizárva, hogy a későbbiekben más szakmai felügyelet alatt egy másik projektet hozzanak létre.)

A határon túli magyar korpusz megvalósulása a kezdeti elképzelésekhez képest módosult. A változás két alkorpuszt: a hivatali nyelvet és a személyes közlést tartalmazót érintette. Bár a hivatali szövegek gyűjtése eddig is folyamatos volt, ám mivel a magyar nyelv kisebbségi helyzetben csak másodlagos szerepű, használata a hivatalos szférában pedig – nyelvtörvények által – korlátozott, nem valószínű, hogy a határon túli magyar alkorpuszban valaha is elérik a kívánt arányokat. (Már csak azért sem, mert a tudományos, szépirodalmi és publicisztikai alkorpusz nagyobb mértékben bővül, így az abszolút számok is folyamatosan növekszenek, s egyben elérhetetlenné válnak.)

Az NKFP által támogatott pályázat 2005 októberének végén járt le. A korpusz első nyilvános bemutatójára 2005. november 22-én a Magyar Tudomány Napja alkalmából rendezett előadássorozat keretén belül került sor. Személy szerint csak remélni tudom, hogy minél szélesebb körben ismertté válik, s minél többen kihasználják majd az általa nyújtott kutatási és oktatási lehetőségeket.

A hivatkozott irodalom

BEREGSZÁSZI ANIKÓ – CSERNICSKÓ ISTVÁN 2004. Magyar értelmező kéziszótár: (majdnem) minden magyar szótára. In: BEREGSZÁSZI ANIKÓ – CSERNICSKÓ ISTVÁN, ...itt mennyit ér a szó? Írások a kárpátaljai magyar nyelvhasználatról. PoliPrint, Ungvár. 127–36.

BIBER, DOUGLAS 1993. Representativeness in corpus design. Literary and Linguistic Computing 8: 243–57.

CSERNICSKÓ ISTVÁN 2004. A magyar nemzeti nyelvstratégiáról, mulasztásainkról, feladatainkról és vágyainkról. In: BEREGSZÁSZI ANIKÓ – CSERNICSKÓ ISTVÁN szerk., Tanulmányok a kárpátaljai magyar nyelvhasználatról. PoliPrint – Kárpátaljai Magyar Tanárképző Főiskola, Ungvár. 106–16.

CSERNICSKÓ ISTVÁN – PAPP GYÖRGY – PÉNTEK JÁNOS – SZABÓMIHÁLY GIZELLA 2005. A szomszédos országok magyarnyelvi kutatóállomásairól. Magyar Nyelv 105–13.

Emlékeztető az MTA kutatóállomásainak megbeszéléséről. MTA Etnikai-nemzeti Kisebbségkutató Intézet, Bp., 2002. 05. 29. Kézirat.

Emlékeztető a nyelvi irodák műhelytalálkozójáról. Illyefalva, 2004. július 12–17.

KIEFER FERENC 2005. Lehetőség és szükségszerűség. Tanulmányok a nyelvi modalitás köréből. Tinta Könyvkiadó, Bp.

KOLLÁTH ANNA 2005a. Első fejezet a kisebbségi magyar nyelvhasználat összehasonlító vizsgálatából. Határtalanítás: előzmények és eredmények – szándék és megvalósulás. In: LANSTYÁK ISTVÁN – MENYHÁRT JÓZSEF szerk., Tanulmányok a kétnyelvűségről III. Kalligram Könyvkiadó, Pozsony. 15–31.

KOLLÁTH ANNA 2005b. Fejezetek a kisebbségi magyar nyelvhasználat összehasonlító vizsgálatából. Magyar Tudomány 156–64.

KOLLÁTH ANNA – SZOTÁK SZILVIA – ŽAGAR-SZENTESI ORSOLYA 2005. Kiegészítés „A szomszédos országok magyarnyelvi kutatóállomásai" című beszámolóhoz. Magyar Nyelv 371–7.

LANSTYÁK ISTVÁN 2004. Élőnyelvi szövegek fonematikai elvű átírása. In: BEREGSZÁSZI ANIKÓ – CSERNICSKÓ ISTVÁN: „...itt mennyit ér a szó? Írások a kárpátaljai magyar nyelvhasználatról". PoliPrint, Ungvár. 181–5.

LANSTYÁK ISTVÁN 2006. Határtalanítás (a Magyar értelmező kéziszótár 2. kiadása után, 3. kiadása előtt). In: MÁRTONFI ATTILA – PAPP KORNÉLIA – SLÍZ MARIANN szerk., 101 írás Pusztai Ferenc tiszteletére. Argumentum Kiadó, Bp. 179–86.

LANSTYÁK ISTVÁN – MENYHÁRT JÓZSEF 2001. A Gramma Nyelvi Iroda (avagy: Lesz-e álomból valóság?). Fórum Társadalomtudományi Szemle 189–203.

NOVÁK ATTILA 2003. Milyen a jó humor? In: ALEXIN ZOLTÁN – CSENDES DÓRA szerk., Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2003). Szegedi Tudományegyetem, Szeged. 138–45.

NOVÁK ATTILA – M. PINTÉR TIBOR 2006. Milyen a még jobb Humor? In: ALEXIN ZOLTÁN – CSENDES DÓRA szerk., IV. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2006). Szegedi Tudományegyetem, Szeged. 60–9.

PINTÉR TIBOR 2003. Amit a modern nemzeti korpuszokról tudni kell. Fórum Társadalomtudományi Szemle 71–85.

PÉNTEK JÁNOS 2004. A magyar nyelv szótárai, nyelvtanai, kézikönyvei és a határon túli magyar nyelvváltozatok. Az MTA határon túli kutatóállomásainak feladatait is ellátó nyelvi irodák állásfoglalása. Magyar Tudomány 724–7.

RAJSLI ILONA 2004. Útmutató a korpuszba építendő élőnyelvi szövegek lejegyzéséhez. In: PAPP GYÖRGY szerk., Mi ilyen nyelvben élünk. Nyelvszociológiai és korpuszvizsgálati tanulmányok. Magyarságkutató Tudományos Társaság, Szabadka. 65–79.

SZOTÁK SZILVIA 2005. Fejezetek a kisebbségi magyar nyelvhasználat összehasonlító vizsgálatából. Határtalanítás; őrvidéki szavak magyarországi szótárakban. In: KEMÉNYFI RÓBERT szerk., Osztrák források – magyar kutatók, Österreichische Quellen – Ungarische Forscher. Debreceni Egyetem Néprajzi Tanszéke – Collegium Hungaricum, Debrecen–Bécs.

PINTÉR TIBOR