M. PINTÉR TIBOR

„Határtalan” magyar nyelv – az első, határon túli magyar nyelvváltozatokat tartalmazó strukturált magyar nyelvi korpuszról

TIBOR M. PINTÉR
„BORDERLESS” HUNGARIAN LANGUAGE – THE FIRST STRUCTURISED HUNGARIAN LANGUAGE CORPUS COMPRISING OF CROSS-BORDER HUNGARIAN LANGUAGE VARIATIONS IS READY

811.511.141`374:81`322
81`322:811.511.141`374
801.8:811.511.141:81`322

Corpus Linguistics. Hungarian Language Corpus of Carpathian Basin. Computer-based data processing. Wordject.

1. Bevezetés

A mai nyelvészeti kutatások módszertani alapelve az adatorientáltság, a kutatás mélységének és milyenségének megfelelő adatmennyiség biztosítása. A „megfelelő mennyiség” a kutatás céljától, illetve a kutatást végző nyelvészeti diszciplína milyenségétől függően változhat. A kutatás eredményeinek pontossága azonban általában növelhető a feldolgozandó anyag mennyiségének növelésével. Ennek megfelelően a nyelvészetben egyre inkább felértékelődik az adatbázisok szerepe.[1] A különféle kutatásokhoz szükséges adatgyűjtés általában elvégezhető az adott diszciplína területén belül is, azonban az összegyűjtött adatok feldolgozása így általában esetleges, minimális marad, hiszen nem biztos, hogy az adatbázist – az egyféle megközelítésmód miatt – más diszciplína is fel tudja használni. Az ideális állapot valószínűleg az lenne, ha olyan, különböző módon strukturált adatbázisok készül(het)nének, amelyek a legtöbb tudományterület számára felhasználás és feldolgozás céljából elérhetők lennének, és egy teljes beszélő- vagy nyelvközösséget reprezentálnának. Mindkét cél elérése jelenleg szinte megvalósíthatatlannak tűnik, főként két okból kifolyólag. Egyrészt azért, mert a nyelvészet egyes ágai oly mértékben differenciálódtak, hogy szinte lehetetlen valamennyit kielégíteni (nehéz lenne olyan adattárat készíteni, amelyet például a kísérleti fonetika és a nyelvtörténet ugyanolyan mértékben használna), másrészt egy nagy létszámú beszélőközösség, sőt nyelvközösség reprezentatív mintavételen alapuló adattárának összeállítása szinte kivitelezhetetlen (az adattárak reprezentativitásáról lásd Biber 1993; Pintér 2003, 74–76).

Az adatbázisok feldolgozásának esetlegessége, azaz a feldolgozás részletessége és szélessége a széleskörű kívánalmak miatt szinte áthidalhatatlan feladat. Ez azonban nem jelenti azt, hogy nem lennének rá kísérletek – akár a magyar

nyelv(terület)en belül is. Az adattárak kezelésében, szerkesztésében, feldolgozásában legnagyobb szerepet jelenleg a korpusznyelvészet (és a tőle szinte elválaszthatatlan számítógépes nyelvészet) játssza. A korpusznyelvészet elterjedésével módosultak az adattárak feldolgozásának módjai, illetve részben módosult azok besorolása, megnevezése is. Bár a szakirodalom nem egységes a *korpusz* (vagy számítógépes szövegtár) definiálásában, mégis úgy tűnik, módosulnak a korpuszok meghatározásának követelményei. A korpusznyelvészet térnyerésével egyre inkább a számítógépes *feldolgozottság*ot (nem beszélhetünk tehát korpuszról akkor, ha az adattár például újságok vagy hangfelvételek gyűjteménye: ez adattár, de nem korpusz), illetve a *strukturáltság*ot (tehát a számítógépen tárolt szövegek önmagukban még nem korpuszok) tekinthetjük a legfontosabb szempontnak a korpuszok meghatározásában.

A magyar nyelven készült korpuszok közül a legnagyobb a ma már több mint 187 millió szavas *Kárpát-medencei magyar nyelvi korpusz* (*Kmmnyk*). Ennek elődje, a *Magyar nemzeti szövegtár* az NKFP 5/044/2002 pályázatának segítségével kiegészült egy 15 millió szóból álló, a határon túli magyar nyelvváltozatokat bemutató alkorpusszal. Az így összeállított korpusz valóban „nemzeti" lett, mivel nemcsak a magyarországi magyar nyelvváltozatokból merít, hanem a Magyarországgal szomszédos államokban beszélt magyar nyelvváltozatokból is (szervezett gyűjtés és feldolgozás eddig a szlovákiai, a romániai, az ukrajnai és a szerbiai magyar nyelvváltozatokból történt).

2. A kivitelezők – Az MTA határon túli kutatóállomásainak hálózata

A *Kárpát-medencei magyar nyelvi korpusz* határon túli magyar alkorpuszának elkészítéséhez a hátteret a Magyarországgal határos országokban létesített kutatóhálózat állomásai szolgáltatták: Szlovákiában a dunaszerdahelyi Gramma Nyelvi Iroda, Erdélyben a Kolozsvárott és Szepsiszentgyörgyön működő Szabó T. Attila Nyelvi Intézet, Kárpátalján a beregszászi Hodinka Antal Intézet és a Vajdaságban a kanizsai Vajdasági Magyar Nyelvi Korpusz. A nyelvi irodák létrehozásában legfontosabb szerepet a határon túli magyar nyelvváltozatot érintő feladatok, illetve a határon túli magyarságot érintő különféle társadalomtudományi kutatások megszervezése játszotta (Lanstyák–Menyhárt 2001, 190–191). A fent említett intézmények a Magyar Tudományos Akadémia Etnikai-nemzeti Kisebbségkutató Intézetének (főként igazgatójának, Szarka Lászlónak) szervezésében 2001. október 1-jétől működnek, létrehozva az MTA határon túli kutatóállomásainak hálózatát. A kutatóhálózat feladatai között kiemelkedő jelentőséggel bíró korpusznyelvészeti kutatások szakmai koordinátora a Magyar Tudományos Akadémia Nyelvtudományi Intézetének Korpusznyelvészeti Osztálya lett (mai neve: Nyelvtechnológiai Osztály), a kutatások gazdasági hátteréért pedig a Magyar Tudományos Akadémia Etnikai-nemzeti Kisebbségkutató Intézete felelt.

A *Kmmnyk* határon túli anyagokkal történő bővítése csupán egy az MTA határon túli kutatóállomásainak feladatai közül (a feladatokról bővebben lásd http://www.mtaki.hu/kutatoallomasok). Bár a kutatóhálózatot alkotó irodák saját problémákkal foglalkozó kutatási területekkel is rendelkeznek, legnagyobb eredményeiket mégis az ún. közös kutatásokban mutatják fel. Ezek a Kárpát-medencei magyarság nyelvi helyzetére irányulnak, s a következő területeket ölelik fel:

1. a Kárpát-medencei magyar nyelvű oktatás helyzete (a magyar nyelv helyzete a kisebbségi magyar régiókban);

2. a magyar nyelv állami változatait érintő lexikográfiai kutatások (a Magyarországon kiadott kodifikációs érvényű szótárak anyagának bővítése a Magyarország határain kívül használt magyar nyelvváltozat szavaival – *Határtalanítás I.*);

3. a korpuszépítéssel kapcsolatos közös kutatások (a Kárpát-medencei magyar nyelvi korpusz bővítése a Magyarország határain kívül használt magyar nyelvváltozatokkal – *Határtalanítás II.*).

A közös kutatások közül eddig legkézzelfoghatóbb eredmények a korpusznyelvészeti és a lexikográfiai kutatásokban mutatkoznak meg.[2]

2.1. A *Kárpát-medencei magyar nyelvi korpusz*

A *Kárpát-medencei magyar nyelvi korpusz* határon túli alkorpusza (így a *Szlovákiai magyar korpusz* is) a magyar nyelv legkiegyensúlyozottabb számítógépes nyelvi adatbázisának részeként jött létre. Röviden összefoglalva, a *Határon túli magyar korpusz* négy Magyarországgal határos országban megjelent vagy elhangzott szövegek számítógéppel feldolgozott, rétegzett gyűjteménye. Ez a korpusz nem kíván a határon túli magyar szövegek reprezentatív mintája lenni, hiszen a reprezentativitás kritériumait ez esetben lehetetlen lenne megfogalmazni, s ha ezek a követelmények megfogalmazódnának is, az egyes szövegtípusok állandó változását, az egyes arányok mozgását szinte lehetetlen lenne követni (vö. a 4.4. alfejezet utolsó bekezdésével).

A *Határon túli magyar korpuszban* a határon túli magyar nyelvű anyagok aránya a következőképpen lett meghatározva: szlovákiai magyar rész 4 millió, a romániai 6 millió, a kárpátaljai 3 millió, míg a vajdasági 2 millió szövegszó. Mint ahogy azt a következő táblázat mutatja, ezeket a követelményeket nem volt nehéz teljesíteni. Az igazsághoz azonban az is hozzátartozik, hogy a korpusz a határon túli anyagok összegyűjtése előtt is tartalmazott szlovákiai és romániai magyar napilapokat, amelyek a kiegészülés után a kisebbségi sajtóhoz lettek csoportosítva.

A *Kmmnyk* jelenlegi állapotát a következő táblázat alapján tekinthetjük át.

1. táblázat. A *Kmmnyk* 2006. november 1-jei állapota

	Magyarországi	Szlovákiai	Kárpátaljai	Erdélyi	Vajdasági	Összesen
Sajtó	71,0*	5,7	0,7	5,5	1,5	84,5
Szépirodalom	35,3	1,4	0,4	0,8	0,2	38,2
Tudományos	20,5	2,3	0,7	1,6	0,3	25,5
Hivatalos	19,9	0,2	0,3	0,6	0,1	20,9
Személyes	17,8	–	0,4	0,4	0,1	18,6
Összesen	164,7	9,5	2,5	8,9	2,0	187,6

Forrás: http://corpus.nytud.hu/mnsz.
Megjegyzés: * millió

A *Kárpát-medencei magyar nyelvi korpusz* több tulajdonságával is kitűnik a többi magyar nyelvű korpusz közül. Jelenleg több mint 187 millió szót tartalmaz[3], regiszterei között megtalálhatók az írott és beszélt nyelvváltozatok is, illetve ez az egyetlen olyan magyar nyelvű magyar nyelvi korpusz, amely nemcsak magyarországi, hanem határon túli magyar nyelvváltozatokat is tartalmaz.

A határon túli alkorpusz készítésének előzménye a *Magyar nemzetei szövegtár*ig nyúlik vissza. A *Kárpát-medencei magyar nyelvi korpusz* megvalósítását (és így a *Ha-*

táron túli magyar korpusz megvalósítását is) ugyanis megelőzte a *Magyar nemzeti szövegtár* projektje. Az akkor még 140 millió szavas korpusz pár millió szava származott határon túli folyóiratokból (a felvidéki *Új Szó*ból és az erdélyi *Romániai Magyar Szó*ból). Ezt természetesen akkor csupán mutatványként vagy jó szándékként lehetett értelmezni, ami a szókereséskor inkább zavaró volt, mint segítő, hiszen a nem magyarországi sajtóban külön nem lehet keresni, viszont a magyarországi adatok keresése közben a határon túli adatok zavaróan hatottak. Nyilvánvaló volt tehát, hogy szükség és igény van egy nagyobb, a kisebbségi magyar nyelvváltozatokat bemutató szövegtárra is. A határon túli magyar nyelvváltozatokat bemutató korpusz része a kutatóállomás egyik fő feladataként aposztrofált határtalanításnak, hiszen a szövegtár célja a határon túli magyar nyelvváltozatok magyarországi terjesztése. A kutatóhálózat korpuszmunkálatokért felelős munkatársai sajnos eleinte nem hangsúlyozták eléggé, hogy a *Kárpát-medencei magyar nyelvi korpusz* is része a határtalanításnak. A korpuszmunkálatok és a határtalanítás kapcsolata csupán Kolláth Anna 2005-ben írt, a határtalanításról szóló tanulmánya után merült fel (Kolláth 2005a). Kolláth *A határtalanítás* című fejezetben úgy fogalmaz, hogy: „a határtalanításnak az a célja, hogy a magyar nyelv szótárai és kézikönyvei, amelyek Trianon óta, de elsősorban 1945 után inkább csak a magyarországi magyar nyelvről szóltak, egyetemes léptékűvé, összmagyarrá váljanak" (Kolláth 2005a, 16). Abban egyetértek a tanulmány szerzőjével, hogy a határtalanítás „hordozóinak" mindenképpen a szótáraknak kell lenniük. A számítástechnika fejlődése azonban módosítja a már megszokott szótárdefiníciót, megjelentek a számítógépes „szó-tárak" legújabb fajtái, a korpuszok, amelyek esetünkben szintén a határtalanítás szerves részei – ezt azóta a kutatóhálózat tagjai is hangsúlyozzák. A korpuszok szintén egy nyelv szóanyagát dolgozzák fel, s felhasználásuk nemcsak a szókeresésben merül ki, hiszen ismertek olyan szótárak és nyelvtanok is, amelyek korpuszok alapján íródtak (pl. a John Sinclair nevével fémjelzett *Collins Cobuild – English Grammar).*

A *Kárpát-medencei magyar nyelvi korpusz* határon túli anyaga még a továbbiakban is bővülni fog: remélhetőleg nem csak mélységében, hanem szélességében is. Remélhetőleg az MTA határon túli kutatóállomásainak segítségével sikerül legalább őrvidéki és muravidéki anyagokat gyűjteni, illetve feldolgozni.

3. Kezdeti lépések a *Határon túli magyar korpusz* terén

A *Határon túli magyar korpusz*ról szóló első hivatalos feljegyzések 2001-ben készültek. A kutatóhálózat létrehozása után minden iroda kidolgozta saját tervezetét és a munka megvalósulásának ütemtervét. A munka gyakorlati részének elindításában az MTA Nyelvtudományi Intézetében működő Korpusznyelvészeti Osztály (mai nevén: Nyelvtechnológiai Osztály) által szervezett korpusznyelvészeti tréningek jelentettek felbecsülhetetlen segítséget. A tréningek és a kezdeti munkatapasztalatok után az előzetes tervek módosultak: voltak feladatok, amelyek a munka szempontjából később feleslegesnek bizonyultak (pl. a korpusznyelvészeti munkákhoz szorosan nem kapcsolódó listák készítése a szlovákiai magyar sajtóról, kapcsolatfelvétel olyan nyelvészekkel, akikkel a későbbiekben nem érintkeztünk), és voltak teendők, amelyek csak az első tréning után merültek fel (pl. a későbbi munka szempontjából

legnagyobb jelentőségű számítógépes szövegátalakítás vagy kapcsolattartás, kommunikáció a többi irodával, illetve a Nyelvtudományi Intézettel).

Három év távlatából visszanézve figyelemre méltó, hogy az irodahálózat kezdetben olyan feladatra vállalkozott, amelynek elvégzéséhez nem állt rendelkezésünkre sem tudás, sem tapasztalat. Ezek, valamint a kezdeti sikertelenségek fényében ma már elmondható, hogy ezt a projektet ilyen formában merészség volt létrehozni. Bár később az összes szükséges anyagi eszközt és szervezési segítséget megkaptuk, az irodák közti földrajzi távolság miatt az érdemi munka csak nagyon nehezen indult be. Ebben szerepe volt az irodák közti nehézkes párbeszédnek is (illetve a munka természetéből adódó tapasztalatlanságnak), pedig a kommunikáció gyorsítása végett a kutatóhálózatot alkotó nyelvi irodák számára közös levelezőlistát is létrehoztunk.[4] Az első két évben sajnos a kommunikáció nagyon esetlegesnek bizonyult (ennek okát az irodák túlterheltségében, illetve a korpuszon dolgozók elszigeteltségében látom), ám a feladatok halmozódásával és az idő sürgetésével a kommunikációs problémák mára megoldódtak.

A *Kmmnyk* határon túli korpusza egységes formátumú és szerkezetű szövegcsoportot alkot. Ennek feltétele azonban nem csak a közös munka volt, hanem a jó szervezés is. A munka természete úgy kívánta, hogy a kutatóhálózat korpusznyelvészeti teendőit több személy koordinálja. Az egyes irodák munkájához szükséges technológiai követelmények biztosítását, a budapesti szakmai összejövetelek szervezését, illetve a hálózat koordinálását Bartha Csilla végezte. Mivel Bartha nem számítógépes nyelvész, a szakmai feladatok ellenőrzéséért Oravecz Csaba, illetve Váradi Tamás felelt.

A kutatóhálózat létrehozója és irányítója az MTA Etnikai-nemzeti Kisebbségkutató Intézete volt. A hálózat feladatai között előzőleg nem csak nyelvészeti, hanem egyéb társadalomtudományi kutatások végrehajtása és szervezése is helyet kapott. Az a kezdetektől fogva nyilvánvaló volt, hogy a korpusznyelvészeti tevékenységet egy társadalomtudományi kutatásokkal foglalkozó intézet (MTA ENKI) nem tudja felügyelni. Bartha Csilla (MTA Nyelvtudományi Intézete, MTA Etnikai-nemzeti Kisebbségkutató Intézete), illetve Váradi Tamás (MTA Nyelvtudományi Intézete) személyében azonban ez a probléma megoldódott, hiszen így ezt a projektet szakmailag nyelvészek irányították.

A gazdasági és szakmai felügyelet megoszlása 2005 tavaszáig működött ilyen formában, ekkor a kutatóhálózat irányítása átkerült az MTA Nyelvtudományi Intézetéhez (azaz az összes kutatás irányítását a Nyelvtudományi Intézet végzi). Az Etnikai-nemzeti Kisebbségkutató Intézettől ez érthető lépés volt, hiszen a kutatóhálózat közös feladatai nyelvészeti témájúak (noha a kutatóhálózat természetéből adódóan ezek is minden esetben rendelkeznek „kisebbségi" vonatkozással, s az irodák egyéni kutatásai között is vannak kisebbségeket érintő – nem csak nyelvészeti – kérdések). Az új helyzet nem érződött a kutatásokon, hiszen azok ugyanolyan intenzitással folytak minden régióban. Ez annak is köszönhető, hogy a „közös kutatás"-ként megfogalmazott feladatokat az irodahálózat munkatársai és Bartha Csilla, azaz minden esetben nyelvészek koordinálták.[5]

3.1. Korpusznyelvészeti tréningek

Az előzetes megbeszélések és levelezések után a *Kmmnyk* határon túli korpuszának készítői az első elméleti és gyakorlati információkat 2003. január 30–31-én

kapták meg, de mint később a gyakorlatból kiderült, a folyamatos, eredményes munka végzéséhez ez az egyszeri alkalom nem volt elegendő, további folyamatos egyeztetésekre, szakmai összejövetelekre volt szükség. Mivel a kutatóhálózat korpusznyelvészeti teendőket ellátó munkatársai egyik esetben sem rendelkeztek számítógépes nyelvészeti vagy korpusznyelvészeti képzettséggel – számítógépes előismerete is csak néhányuknak volt –, ezért szükség volt az előkódolást végző személyek betanítására (a kódolásról bővebben lásd Pintér 2003, 79–80). Mivel a szövegtár szerkesztése javarészt mechanikus folyamatok elvégzése, ezért a számítógépes előképzettség itt nem volt feltétel. Ezt bizonyítja az is, hogy több irodában azok, akik kezdetben a korpusszal foglalkoztak, még nyelvészeti ismeretekkel sem rendelkeztek. A nyelvészeti beállítottság, a nyelvészeti alapismeretek hiánya természetesen nem jelenthetett problémát, hiszen a nyelvészeti tudást igénylő munkát a nyelvi irodák nyelvészei is elvégezhették.

A tréningeket (a második 2004. június 21–22-én volt) az MTA Nyelvtudományi Intézetének Nyelvtechnológiai Osztályát vezető Váradi Tamás és az osztály egyik munkatársa, Oravecz Csaba tartotta. Az első találkozó alkalmával a határon túli szövegek gyűjtését és kódolását végző személyek[6] megismerkedtek a kódoláshoz szükséges elméleti és gyakorlati információkkal, így a második találkozó során már megvitathatták a kódolás folyamán felmerült gyakorlati problémákat is. Mivel ezek az összejövetelek Budapesten zajlottak, kisebb-nagyobb számban mindig minden kutatóállomás képviseltette magát.[7] Bár mind a négy iroda azonos feladatot végez, a második megbeszélésen irodánként mégis más-más problémák merültek fel. A megbeszélések csak részben hozták meg a tőlük várt eredményeket, mivel az utolsó közös megbeszélés után sem gyorsult az anyagfeldolgozás, és a problémákkal küszködő irodák egy év elteltével is ugyanazon hibák kiküszöbölésével foglalkoztak.

A korpusznyelvészeti tréningek eredményeiről, illetve a kutatóhálózat korpusznyelvészeti tevékenységéről honlap is készült, erre a kódoláshoz, illetve a munka közben felmerült problémák megoldásához szükséges információk Oravecz Csaba révén folyamatosan felkerültek (http://corpus.nytud.hu/mnszworkshop/index.html).

4. A *Kárpát-medencei magyar nyelvi korpusz* készítésének részei

4.1. Anyaggyűjtés

Az irodák által feldolgozott anyag főbb szerkezeti pontjaiban követi a *Magyar nemzeti szövegtár*at (így tudják együttesen alkotni a *Kmmnyk*-t). A gyakorlati megvalósulásban ez azt jelenti, hogy az *MNSZ* magyarországi anyagához hasonlóan a *Határon túli korpusz* is kötelezően öt alkorpuszból áll: tudományos próza, publicisztika, szépirodalom, hivatalos nyelv, személyes közlések. Az anyaggyűjtést minden irodában gondos szervezőmunka előzte meg, hiszen a felgyűjtött anyagoknak már egy kész struktúrába kellett beilleszkedniük.

A sajtónyelvi alkorpusz összeállítása kiemelten fontos előkészületet kívánt, egyrészt mivel a sajtónyelvi szövegek maguk is többfélék (napilapok, ifjúsági lapok, nőknek szóló lapok stb.), így a belső arányokat is meg kellett állapítani, másrészt mivel a határon túli magyar lapok magyarországi lapokból, illetve hírügynökségektől is vesznek át cikkeket, s ezeket előzőleg ki kellett válogatni, hiszen nem magyarországi anyagok feldolgozását tűztük ki célul.

A *Kárpát-medencei magyar nyelvi korpusz* a magyar nyelv jelenlegi állapotát kívánja rögzíteni. Ez a gyakorlatban azt jelenti, hogy a korpusz nem tartalmazhat rendszerváltás előtt keletkezett szövegeket. Ezt a követelményt nem minden alkorpusz esetében tudtuk betartani,[8] mivel például a szépirodalmi szövegek között vannak korábbi keletkezésűek is. Ez azonban nem okoz értelmezési és szerkezeti gondot (már csak azért sem, mivel a szépirodalmi stílus „szabadsága" kortalan, illetve kevésbé változó, mint mondjuk a beszélt nyelvi).

A tudományos prózát tartalmazó alkorpusz összeállításának, gyűjtésének fő problémája, hogy a határon túli magyar tudományos élet bizonyos szinten gyakran többségi nyelven folyik: például a szlovákiai magyar tudományos elitet alkotó réteg szlovák nyelvű munkahelyeken dolgozik, illetve – általában – szlovák nyelven publikál. Ezért a szigorúan tudományos ismérvek szerint írott szövegekből jóval kevesebb van, mint Magyarországon, illetve ezért arányában több a tudományos ismeretterjesztő próza, mint a magyarországi mintában.

A határon túli magyar hivatali nyelvet (nyelvhasználatot) bemutató alkorpusz egyik alappillére a kutatóhálózat nyelvtervezési tevékenysége volt (például a Gramma Nyelvi Iroda nyelvtervezési és fordító tevékenysége).

A legösszetettebb és legmunkaigényesebb részfeladatot a beszélt nyelvi alkorpusz megszerkesztése jelentette, illetve jelenti mind a mai napig. Alapvető probléma a beszélt nyelvi szövegek lejegyzése. Az egyes hangtani jelenségek lejegyzésénél nemcsak a hanganyag lehető legárnyaltabb visszaadását kell figyelembe venni, hanem a számítógép diktálta lehetőségeket, a minél könnyebb számítógépes keresés feltételeit is állandóan szem előtt kell tartani. Így a lejegyzés nem lehet olyan részletekbe menő, mint egy fonetikai vagy részletes nyelvjárási lejegyzés, ám a hangzó nyelv legfőbb sajátosságait mindenképpen írásban is meg kell próbálni visszaadni. A beszélt nyelvi szövegek lejegyzési útmutatójának véglegesítése csak hosszadalmas és időigényes egyeztetések után fejeződött be, mivel a Gramma Nyelvi Irodában készült részletes útmutatót fonetikus és számítógépes nyelvész is véleményezte. A lejegyzés egységesítése fontos, hiszen csak úgy készülhetnek összehasonlítható átiratok, ha a szövegek egységes kódolási minta alapján készülnek el. Éppen ezért minden irodának lehetősége volt közös minta összeállítására, azonban sajnos nem minden iroda élt ezzel a lehetőséggel, és nem tett javaslatot az útmutató kialakítására. A lejegyzési útmutató így a Gramma Nyelvi Irodában, a Lanstyák István által szerkesztett javaslat alapján készült el Kassai Ilona egységesítésével (bővebben lásd a 4.4. alfejezetben).

4.2. Az anyaggyűjtés módja

Az anyaggyűjtés legegyszerűbb és legköltségkímélőbb módszere nagy mennyiségű anyagok gyűjtésekor az internetről történő letöltés. Az internet legnagyobb előnye, hogy a rajta lévő anyagok mindenki számára szabadon hozzáférhetők, letölthetők, illetve hogy a kész anyag (ez esetben szöveg) gyorsan és könnyen hozzáférhető. Sajnálatos módon azonban az anyaggyűjtésnek ez a módja sem tökéletes, mert amellett, hogy az internet a korpusz számára sok felesleges adatot tartalmaz (pl. képek, videók, mozgó reklámok, azaz nem szöveges részek, amelyek kiszűrése ugyan nem jelent problémát, csupán a letöltés folyamatának idejét növeli), a letöltött anyagok

felhasználása szerzői jogi problémákat is felvet – tehát látható, hogy az internetes gyűjtés sem minden esetben problémamentes. Ezért minden internetről letöltött szöveg felhasználására előzőleg engedélyt kell (kellene) kérni a szerzőktől, illetve a honlap működtetőjétől.

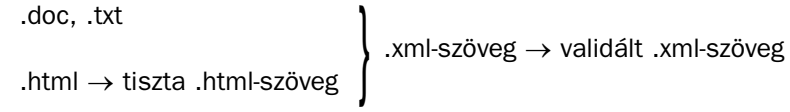
Bár az anyaggyűjtés szempontjából az internet óriási előnyökkel jár, minden alkorpuszhoz mégsem nyújtott anyagot (leginkább a sajtónyelvi és a hivatali nyelvi alkorpusz gyűjtésében volt segítségünkre). Mivel az irodák munkatársai saját régiójukban közismert emberek, ezért gyakran magánszemélyektől, illetve személyes ismeretség alapján kiadóktól és szerkesztőségektől is kaptunk szövegeket. Az anyaggyűjtés, azaz a helyi ismertség és ismeretség kiaknázásának, értékesítésének szempontjából pozitív lépésnek bizonyult a kutatóhálózat korpusznyelvészeti megbízása.

4.3. Feldolgozás

A gyűjtés utáni szövegfeldolgozás, azaz munkánk érdemi része nem jelentett különösen nehéz feladatot, mivel az csupán már meglévő szövegek XML-formátumúvá történő átalakításában merült ki. Megfelelő programok hiányában a feladat nehézsége főleg a folyamat hosszúságában rejlett, ám ez a folyamat (akár egyszerű Word-alkalmazásokkal is) jól automatizálható – így ideje jelentősen csökkenthető. A határon túli anyagok esetében a feldolgozás két elkülöníthető folyamatból áll. Az első folyamat, azaz a szövegek átalakítása az egyes irodákban, míg a feldolgozás második, és egyben bonyolultabb folyamata pedig az MTA Nyelvtudományi Intézetében történt (értelemszerűen a magyarországi anyagok esetében mindkét részfolyamat Magyarországon történik).

Az alapformátumtól (alapszövegtől) a célformátumig tartó számítógépes és számítógépes nyelvészeti folyamatokat a következőképpen modellálhatjuk:

1. ábra. Az MTA határon túli irodáiban végzett folyamat



Ahogy az ábrából is látszik, a folyamat nem túl bonyolult mindössze egy bonyolultabb szövegszerkesztő programra és egy előre meghatározott XML DTD-re van szükségünk. A megformázott és annotált szövegek további elemzését az MTA Nyelvtudományi Intézetében végezték el.

A Nyelvtudományi Intézetben végzett folyamat során minden adott szóalak morfoszintaktikai jegyei kódok formájában (ún. msd, azaz morpho-syntactic description kódok) az adott szóalak mellé kerülnek. Ezt a kódolást a MorphoLogic Kft.-ben kifejlesztett Humor (High-Speed Unification Morphology) morfológiai elemzőprogram végzi: a program lényege, hogy szótár és nyelvtan segítségével felismeri (elemzi vagy adott esetben generálja) az adott szóalakokat. Mivel a program nem rendelkezik szemantikai ismeretekkel, így általában egy-egy szónak több elemzését is létrehozza (pl. *ultramarinkék=ultramarin [FN]+kék[FN]~ultra[FN]+mar[FN]+i[_IKEP]+nk[PSt1] +ék[FAM]+[NOM]*). Ezek a szóalak-homonimák többségében azonban még a morfo-

lógiában kezelhetők, sőt a szövegszintaxis ismeretében általában majdnem teljes mértékben egyértelműsíthetők (a Humor-program működéséről és az elemzés folyamatáról lásd még Novák 2003; Novák–M. Pintér megj. alatt). A már egyszerűsített szöveget az XML-dokumentumoknak megfelelő szerkezet szerint fejléccel látják el, amely tartalmazza a szöveg keletkezésére és megjelenésére vonatkozó információkat (pl. a szöveg keletkezésének ideje, helye, a szöveg szerzője, a kiadó neve, stb. – lásd http://www.tei-c.org/P4X/HD.html). A szövegek feldolgozásának második részét röviden a következőképpen foglalhatjuk össze:

validált .xml-szöveg → szövegrészek szegmentálása → (szóalak-homonimák) egyszerűsítése → annotált (kódolt) részkorpusz → TEI header (fejléc) → belső referenciamutatók → végső validálás → *Kárpát-medencei magyar nyelvi korpusz*

4.4. Problémák

Az előző fejezetben felvázolt alapkódolás az egyes régiókban eltérő gyorsasággal, eltérő módszerekkel, illetve eltérő számítógépes programokkal valósult meg (a végeredményt ellenőrző program 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 „nyelvészetlevelezőlista" 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 kimenetelének kivétel nélkül azonosaknak kell lenniük: ennek oka a szövegekben történő keresés. 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 akadályozná 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 egy fonetikus szakember, Kassai Ilona egységesítette volna. Sajnos félreértések miatt a lejegyzési útmutató összeállításának ez a terve 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 lásd Lanstyák 2004, 181–185) – 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: így 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 lejegyzik, 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:

– a lejegyzendő hangfelvételek nem lehetnek az 1990-es éveknél korábbiak;

– 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 for-

mális regisztereket is kell tartalmaznia (pl. ritualizált szövegek, élettörténetek); a nyelvjárási hanganyag az egész anyag 40-50%-át teheti ki;

– 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 felvétel 70-80%-át kell alkotniuk;

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

– az adatközlők kiválasztásának szempontjait hierarchizálni kell;

– korcsoportok: gyerekek és idős adatközlők is kellenek; a gyerekek képviselhetik az informális, egynyelvű, az idősek a nyelvjárási beszélőket;

– az egyes digitalizált hangfájlokhoz és a hozzájuk tartozó lejegyzett szöveghez fejlécet is csatolni kell, amit célszerű lenne külön fájlban tárolni; ennek 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 jelenlevő 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, igyekeztünk 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. Mivel a hivatali írásbeliség leggyakrabban formanyomtatványok formájában van jelen, ezek pedig leggyakrabban a magyarországi nyomtatványok formahű átvételei. Ez esetben pedig nem beszélhetünk szlovákiai magyar vagy romániai magyar hivatali nyelvről, hiszen ezek általában magyarországi mintát követnek, vesznek át. 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 bővíthetők, de csupán azután, hogy az irodák begyakorolták a lejegyzési útmutatót. Így a 10% elméletileg elérhető (vagy inkább csak elképzelhető), ám mivel a többi alkorpusz is gyarapszik, ennek esélye egyre kevesebb (a hivatalos nyelvi szövegek esetében inkább elképzelhetetlen).

5. Wordject

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 korpu-

szok 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, még nem piacképes változata tartalmazza az *Értelmező kéziszótár* második kiadását és az Osiris Kiadó *Helyesírás*át).

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: pl. szóközök, mondathatár ellenőrzése vagy trágár kifejezések megjelölése. 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 – hiszen 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, pl. *Huszt*[9], *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 tenné 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 normavitással rendelkező nyelvváltozatok (szövegek) elemzésére tervezték (nem pedig nyelvjárási vagy regionális köznyelvi szövegekre). Ebből az következik, hogy a felgyűjtött szavaknak túl kell mutatniuk a regionalitáson (legideálisabb esetben olyan szó kell, hogy legyen, amelyet az egész magyar beszélőközösségben azonosan használnak) és – legalább az állami változatok szintjén – normatívnak kell lenniük. Ezeknek a követelményeknek leginkább a tulajdonnevek, illetve a közvetett kölcsönszavak (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 jónak a program, amelyet vagy tartalmaz a szótár vagy össze tudja rakni már meglévő elemekből. Terveinkben a következő típusú szavak gyűjtését kívánjuk megvalósítani:

a) földrajzi nevek;
b) személynevek – családnevek;
c) személynevek – utónevek;
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 egyértelműsíthető lesz, valamint a morfológiai kódok alapján a szavak az elemzőbe is beépíthetők lesznek. Annak illusztrációjaként, hogy hogyan néz ki 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 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); a szófajon belül milyen altípusba tartozik (pse, azaz helynév); melyik állami változat eleme (nyv:őv, azaz őrvidéki nyelvváltozat); szótő-e vagy toldalék (rp, azaz jobbra bővülő, tehát szótő); illetve főnevek esetében az egyes szám harmadik személyű alakját is (a példában nincs semmi, azaz Sopronkeresztúrja a kívánt alak); sopron+kereszt+úr@i[MN|pse];nyv:őv;rp: Ess_Ul; – a melléknevek esetében többletként jelölni kell a melléknév essivusi alakját (ESS_Ul, azaz sopronkeresztúriul).

A munka első fázisában a helyneveket és az egyéb földrajzi neveket (folyónevek, tájnevek stb.) gyűjtjük össze, s a gyűjtés, illetve kódolás tapasztalataiból kiindulva folytatjuk majd a személynevekkel és a köznevekkel. A köznevekre vonatkozóan már vannak tapasztalataink, amelyet a ht-lista (azaz „a határon túli vonatkozású magyar szókészleti elemek listája”) összeállításával szereztünk és szerzünk folyamatosan. 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.

A Word-szótár határon túli anyagának elkészítése jelenleg nincs szigorú határidőhöz kötve. A határidők bizonytalanságának egyik oka az alkalmazás megvalósításában rejlik: még nincs tisztázva, milyen formában kapcsolódjon a határon túli lexikon a központi szótárhoz: el kell dönteni, hogy külön modulként vagy a központi szótár szerves részeként valósuljon-e meg. A határidőt befolyásoló másik tényező a kutatóhálózat túlterheltsége; mivel az amúgy is sok munkát igénylő közös kutatások mellett minden kutatóállomás és kutatóhely a saját régiójában egyéb (pl. oktatói vagy szervezői) tevékenységet is ellát, ezért a virtuális hálózatot alkotó személyek túlterheltek (ebben annak is szerepe van, hogy a megmaradásért folyó küzdelemben folyamatosan pályázni kell, illetve a pénzszerzésnek egyéb módjait is ki kell használni).

6. Összefoglalás

Háromévi 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 magában foglalja, lehetővé téve ezzel akár egy összehasonlító kutatást 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, feltételezhető, hogy a kutatóállomások továbbra is folytatják az anyagok gyűjtését, mivel mind a négy kutatóállomás a saját régiójában elindította regionális korpuszának építését, illetve pályázik 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. Ennek eldöntésében valószínűleg vízválasztó szerepe lesz a Wordject-projectnek, hiszen kiderül, hogy a hálózat önerőből véghez tud-e vinni egy ilyen méretű kutatást és fejlesztést.)

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, s a hivatalos szférában használata – nyelvtörvények által – korlátozott, nem valószínű, hogy a *Határon túli magyar korpusz*ban 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. október 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 elterjed, s minél többen kihasználják majd az általa nyújtott kutatási és oktatási lehetőségeket.

Jegyzetek

1. Az adatbázisok fontosságát újabban a magyar generatív nyelvészet egyes képviselői is elismerik. Kiefer Ferenc a nyelvi modalitásról szóló könyvében így ír a korpuszok használáról: „A korpusz nem csak arra volt alkalmas, hogy autentikus példákkal igazolja korábbi elképzeléseimet, hanem újabb összefüggések megállapítását is lehetővé tette" (Kiefer 2005, 7).
2. A kutatóhálózat lexikográfiai munkája a következő szótárak munkálatait segítette: *Magyar értelmező kéziszótár* második kiadása, az Osiris-féle *Helyesírás* szótárrésze. A folyamatban levő szótárprojektek közül az Előry Vilma szerkesztette *Képes diákszótár* második kiadásába, a Tolcsvai Nagy Gábor szerkesztette *Idegen szavak szótárá*ba, illetve a MorphoLogic Kft. által gyártott MS Word helyesírás-ellenőrző és nyelvhelyesség-ellenőrző program szótárrészébe gyűjtünk határon túli magyar nyelvi anyagot. (A kutatóhálózat közös kutatásairól bővebben lásd Kolláth 2005a, 16–24; Kolláth 2005b, 156–164; Kolláth et al. 2005; Péntek 2004, 724–727; Beregszászi–Csernicskó 2004, 127–136; Csernicskó 2004, 106–116; Csernicskó et al. 2005, 105–113; Szoták 2005).
3. A *Kmmnyk* a maga majdnem 200 millió szövegszavával azonban korántsem a legnagyobb magyar korpusz. Ez a cím minden kétséget kizáróan a *Szószablya-projektum* keretében létrehozott *Webkorpuszt* illeti meg, amely 1,48 milliárd szót tartalmaz, amelyből 589 van morfológialag feldolgozva. Csak érdekesség kedvéért jegyzem meg, hogy a korpusz majdnem 18 gigányi szöveget tartalmaz!
4. Erre az ún. nyelvészet-levelezőlistára (vagy ahogy Kolláth Anna elnevezte, „nyelvésznetre") minden iroda feliratkozott, illetve a listára mindenki felkerülhetett, aki valamilyen formában érintve volt vagy van a kutatóhálózat munkájában (tehát nem csak nyelvészek, hanem egyéb kutatók is).
5. A lexikográfiai kutatások szervezője és lelke Lanstyák István (Gramma Nyelvi Iroda), a korpuszkutatások és az oktatáskutatás szervezéséért Bartha Csilla (MTA Nyelvtudományi Intézet) felel – a korpuszkutatások szervezésében, valamint az irodák közötti kommunikációban Pintér Tibor (Gramma Nyelvi Iroda) segíti munkáját. Az irodahálózat saját képviselőjének Péntek Jánost választotta.
6. A *Kárpát-medencei magyar nyelvi korpusz* határon túli anyagának előkódolását végzők: Szlovákia (Gramma Nyelvi Iroda): Pintér Tibor, Mészáros Tímea, illetve Simon Szabolcs; Erdély (Szabó T. Attila Nyelvi Intézet): Becze Orsolya, Sárosi Mardírosz Krisztína Mária; Kárpátalja (Hodinka Antal Intézet): Molnár D. István, Márku Anita, Hires Kornélia; Vajdaság (Vajdasági Magyar Korpusz): Varga Tünde, Darabán Piroska, Fodor Attila.
7. A korpusznyelvészeti összejövetelek sajátos formái voltak a Szabó T. Attila Nyelvi Intézet által Illyefalván szervezett találkozók, ahol a kutatóhálózat tagjai egy héten keresztül részletesen megbeszélhették az egyes kutatásokat (nemcsak a korpusznyelvészeti teendőket, hanem a lexikográfiai, oktatásügyi, illetve szervezési kérdéseket is). Sajnos az illyefalvi találkozók nem váltották be a hozzájuk fűzött kezdeti reményeket, mivel a három alkalom közül a 2004-ben őrvidéki, muravidéki és horvátországi kutatóhelyekkel kiegészült kutatóhálózat egyiken sem tudott teljes létszámban részt venni. Így az első két találkozó után harmadik alkalommal a kutatóhálózatból már csak a szervezők voltak jelen. Ennek

oka valószínűleg a találkozó „fakultatív" jellegéből adódott, mivel a részvétel egyik évben sem volt kötelező (ellenben a budapesti találkozókkal).

8. Ezt a követelményt a *Kárpát-medencei magyar nyelvi korpusz* elődje, a *Magyar nemzeti szövegtár* sem tartotta be, amit a gyűjtés és feldolgozás körülményessége miatt nem is lehet a szerkesztőknek felróni.
9. Köszönet Kolláth Annának az észrevételért és a példáért.

Felhasznált 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.* Ungvár, PoliPrint, 127–136. p.

Biber, Douglas 1993. Representativeness in corpus design. *Literary and Linguistic Computing,* 8. évf. 4. sz. 243–257. p.

Csernicskó István 2004. A magyar nemzeti nyelvstratégiáról, mulasztásainkról, feladatainkról és vágyainkról. In Beregszászi Anikó–Csernicskó István (szerk.) *Tanulmányok a kárpátaljai magyar nyelvhasználatról.* Ungvár, PoliPrint–Kárpátaljai Magyar Tanárképző Főiskola, 106–116. p.

Csernicskó István–Papp György–Péntek János–Szabómihály Gizella 2005. A szomszédos országok magyarnyelvi kutatóállomásairól. *Magyar Nyelv*, 101. évf. 1. sz. 105–113. p.

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

Emlékeztető a nyelvi irodák műhelytalálkozójáról 2004. Kézirat. Illyefalva (2004. július 12–17.).

Kiefer Ferenc 2005. *Lehetőség és szükségszerűség. Tanulmányok a nyelvi modalitás köréből.* Budapest, Tinta Könyvkiadó.

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.* Pozsony, Kalligram Könyvkiadó, 15–31. p.

Kolláth Anna 2005b. Fejezetek a kisebbségi magyar nyelvhasználat összehasonlító vizsgálatából. *Magyar Tudomány,* 49. évf. 2. sz. 156–164. p.

Kolláth Anna–Szoták Szilvia–Žagar-Szentesi Orsolya 2005. Kiegészítés „A szomszédos országok magyarnyelvi kutatóállomásai" című beszámolóhoz. *Magyar Nyelv*, 101. évf. 3. sz. 371–377. p.

Lanstyák István 2004. Élőnyelvi szövegek fonematikai elvű átírása. In Beregszászi Anikó–Csernicskó István: *„... itt mennyit ér a szó? Írások a kárpátaljai magyar nyelvhasználatról"*. Ungvár, PoliPrint, 181–185. p.

Lanstyák István 2005. Határtalanítás (a Magyar értelmező kéziszótár 2. kiadása után, 3. kiadása előtt). In Mártonfi Attila–Papp Kornélia–Slíz Mariann (szerk.): *101 írás Pusztai Ferenc tiszteletére.* Budapest, Argumentum, 179–286. p.

Lanstyák István–Menyhárt József 2001. A Gramma Nyelvi Iroda (avagy: Lesz-e álomból valóság?). *Fórum Társadalomtudományi Szemle,* 3. évf. 3. sz. 189–196. p.

Novák Attila 2003. Milyen a jó humor? In Alexin Zoltán–Csendes Dóra (szerk): *Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2003)*. Szeged, Szegedi Tudományegyetem, 138–145. p.

Novák Attila–M. Pintér Tibor (megj. alatt). Milyen a még jobb Humor? In Alexin Zoltán–Csendes Dóra (szerk.): *IV. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2006)*. Szeged, Szegedi Tudományegyetem, 60–69. p.

Pintér Tibor 2003. Amit a modern nemzeti korpuszokról tudni kell. *Fórum Társadalomtudományi Szemle,* 4. évf. 3. sz. 71–85. p.

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,* 48. évf. 7. sz. 724–727. p.

Rajsli Ilona 2004. Útmutató a korpuszba építendő élőnyelvi szövegek lejegyzéséhez. In Papp György (szerk.): *Mi ilyen nyelvben élünk. Nyelvszociológiai és korpuszvizsgálati tanulmányok.* Szabadka, Magyarságkutató Tudományos Társaság, 65–79. p.

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.* Debrecen–Bécs, Debreceni Egyetem Néprajzi Tanszéke–Collegium Hungaricum.

TIBOR M. PINTÉR

„BORDERLESS" HUNGARIAN LANGUAGE – THE FIRST STRUCTURISED HUNGARIAN LANGUAGE CORPUS COMPRISING OF CROSS-BORDER HUNGARIAN LANGUAGE VARIATIONS IS READY

After a three-year work The Hungarian Language Corpus of the Carpathian Basin is ready. Despite the fact that the material is just a fragment of the Hungarian version, it is a substantial progress in the field of Hungarian language corpuses, since with this work such corpus was created in the Linguistic Institution that includes even the variations of the Hungarian language outside the borders of Hungary, thus enabling perhaps a comparing reserach.

Although, with the creation of the Kmmnyk the works have not been finished. Two question have not been answered. The transcription and annotation of living languages have not been finished, work harmonisation and/or unification/annotation of records is not done. It is questionable if the work on the material Kmmnyk containing over the border Hungarian language variations will continue in the future. Whatever would the competition's future be, it can be presumed that the research station will continue in collecting materials, since all the four research stations launched building in their own region its regional corpuses, and/or competes for the preparation of Wordject-project. Although, if under the supervision of MTA Linguistic Institution no other joint project is realised, then it is more probable that the material that is gradually collected in the research station will have different forms.

The Hungarian Over the Border Language Corpus comparing to the inicial conceptions has been changed. The change related to the sub-corpus, official language and personal communication.

Although the collection of official text has been gradual, but since the Hungarian language in minority position is only secondary, and its usage in officialy - due to the acts on languages - is limited, it is less probable that in the Over the Border Hungarian Language the requested proportions will be ever reached (because scientific, literary and publicistic sub-corpus is growing in a big extent, therefore the absolute numbers are gradually growing, and thus becoming unreachable).

The first public repserentation of the corpus was on 22nd November 2005 within the series of performations of the Hungarian Scientific Day. Personally, I can only hope that it will be known and used by many people for research and educational purposes.