ÉLŐ NYELV

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

1. B e v e z e t é s . **–** 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. (Az adatbázisok fontosságát újabban a magyar generatív nyelvészet egyes képviselői is elismerik. KIEFER FERENC [2005: 7] például a nyelvi modalitásról szóló könyvében így ír a korpuszok hasznáról: „A korpusz nem csak arra volt alkalmas, hogy autentikus példákkal igazolja korábbi elképzeléseimet, hanem újabb összefüggések megállapítását is lehetővé tette”.)

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 elérhetetlennek tűnik, főként két ok miatt. Egyrészt azért, mert a nyelvészet egyes ágai oly mértékben differenciálódtak, hogy szinte lehetetlen valamennyit kielégíteni (nehéz lenne olyan adattárat készíteni, amit 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 l. BIBER 1993., PINTÉR 2003: 74–6).

Az adatbázisok feldolgozásának esetlegessége, azaz a feldolgozás részletessége és szélessége a széles körű 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 a 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 feldolgozottságot (nem beszélhetünk tehát korpuszról akkor, ha az adattár például újságok vagy hangfelvételek gyűjteménye: ez adattár, de nem korpusz), illetve a strukturáltságot (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. sz. 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, romániai, ukrajnai és szerbiai magyar nyelvváltozatokból történt.)

2. A) A kivitelezők: az MTA Határon Túli Kutatóhálózata.– 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, a Vajdaságban pedig a kanizsai Vajdasági Magyar Nyelvi Korpusz. A nyelvi irodák létrehozásában legfontosabb szerepet a határon túli magyar nyelvváltozatok vizsgálatát érintő feladatok, illetve a határon túli magyarságot érintő különféle társadalomtudományi kutatások megszervezése játszotta (LANSTYÁK–MENYHÁRT 2001: 190–1). 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 l. 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 nyelvváltozatok szavaival – határtalanítás I.); 3. a korpuszépítéssel kapcsolatos közös kutatásokban (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. A kutatóhálózat lexikográfiai munkája a következő szótárak munkálatait segítette: Magyar értelmező kéziszótár (ÉKsz.[2]), az Osiris Helyesírás (OH.) szótárrésze. A szótárprojektek közül az EŐRY 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 gyűjtöttünk határon túli magyar nyelvi anyagot. (A kutatóhálózat közös kutatásairól bővebben l. KOLLÁTH 2005a: 16–24, 2005b: 156–64, KOLLÁTH et al. 2005., PÉNTEK 2004: 724–7, BEREGSZÁSZI–CSERNICSKÓ 2004: 127–36, CSERNICSKÓ 2004: 106–16, CSERNICSKÓ et al. 2005: 105–13, SZOTÁK 2005., LANSTYÁK 2006.)

B) 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.** D) Problémák című fejezet 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 állapota a következő (forrás: http://corpus.nytud.hu/mnsz/; 2007. november 1-jei állapot):

	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**

A Kárpát-medencei magyar nyelvi korpusz több tulajdonságával is kitűnik a többi magyar nyelvű korpusz közül. Jelenleg több mint 187 millió szót tartalmaz, regiszterei között megtalálhatók az írott és beszélt nyelvváltozatok is, illetve ez az egyetlen olyan magyar nyelvű magyarnyelvi korpusz, amely nemcsak a magyarországi, hanem a határon túli magyar nyelvváltozatokat is tartalmaz. (A Kmmnyk. egyébként a maga majdnem 200 millió szövegszavával 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 millió van morfológiailag feldolgozva. Csak érdekesség kedvéért jegyzem meg, hogy a korpusz majdnem 18 gigányi szöveget tartalmaz.)

A határon túli alkorpusz készítésének előzménye a Magyar nemzeti szövegtárig nyúlik vissza. A Kárpát-medencei magyar nyelvi korpusz megvalósítását (és így a határon 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 Szóból és az erdélyi Romániai Magyar Szóból). 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 lehetett 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 megismertetése. 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 határtalanításról szóló tanulmánya után merült fel. KOLLÁTH „A határtalanítás" című fejezetben így fogalmaz: „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, megjelennek 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. Collins Cobuild – English Grammar).

A Kárpát medencei magyar nyelvi korpusz határon túli anyaga még a továbbiakban is bővülni fog, s remélhetőleg nemcsak mélységében, hanem szélességében is – amennyiben az MTA Határon Túli Kutatóállomásainak segítségével sikerül legalább őrvidéki és muravidéki anyagokat is gyűjteni, illetve feldolgozni.

3. A) K e z d e t i l é p é s e k a h a t á r o n t ú l i m a g y a r k o r p u s z t e r é n. – A Határon túli magyar korpuszról szóló első hivatalos feljegyzések 2001-ben kerültek papírra. 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 munkák 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 egymás közti földrajzi távolsága miatt az érdemi munka csak nagyon nehezen indult meg. 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. 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 nemcsak nyelvészek, hanem más kutatók is. 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 nemcsak 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 feleltek.

A kutatóhálózat létrehozója és irányítója az MTA Etnikai-nemzeti Kisebbségkutató Intézete volt. A hálózat feladatai között előzőleg nemcsak 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 fogja tudni felügyelni. BARTHA CSILLA (MTA Nyelvtudományi Intézet, MTA Etnikai-nemzeti Kisebbségkutató Intézet), illetve VÁRADI TAMÁS (MTA Nyelvtudományi Intézet) 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 kutatásként" megfogalmazott feladatokat az irodahálózat munkatársai és BARTHA CSILLA, azaz minden esetben nyelvészek koordinálták. A lexikográfiai kutatások szervezője és lelke LANSTYÁK ISTVÁN (Gramma Nyelvi Iroda), a korpuszkutatások és az oktatáskutatás szervezéséért BARTHA CSILLA (MTA Nyelvtudományi Intézet) felelt – a korpuszkutatások szervezésében, valamint az irodák közötti kommunikációban PINTÉR TIBOR (Gramma Nyelvi Iroda) segítette munkáját. Az irodahálózat saját képviselőjének PÉNTEK JÁNOSt választotta.

B) K o r p u s z n y e l v é s z e t i t r é n i n g e k . – Az előzetes megbeszélések és levelezések után a Kmmnyk. Határon túli alkorpuszának 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 l. 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 tartották. Az első találkozó alkalmával a határon túli szövegek gyűjtését és kódolását végző személyek[1] 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.[2] 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, melyre 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. – A) Anyaggyűjtés. – Az irodák által feldolgozott anyag főbb szerkezeti pontjaiban követi a Magyar nemzeti szövegtárat (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, női lapok stb.), így a belső arányokat is meg kellett állapítani; másrészt mivel a határon túli magyar lapok

[1] 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.

[2] 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: a részvétel egyik évben sem volt kötelező – ellenben a budapesti találkozókkal.

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, mivel például a szépirodalmi szövegek között vannak korábbi keletkezésűek is. (A hasonló 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.) 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 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 fordítótevékenysége).

A legösszetettebb és legmunkaigényesebb részfeladatot a beszélt nyelvi alkorpusz megszerkesztése jelentette, illetve jelenti mind a mai napig. Komoly 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 a **4.** D) Problémák című fejezetben).

B) A z a n y a g g y ű j t é s m ó d j a . – 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, amik kiszűrése ugyan nem jelent problémát, csupán a letöltés folyamatának idejét növeli), a letöltött anyagok felhasználása szerzői jogi problémákat is felvet – tehát az internetes gyűjtés sem minden esetben

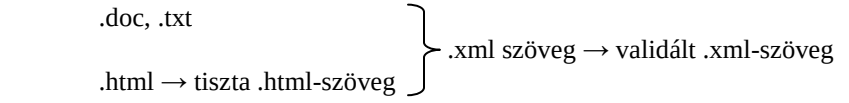
problémamentes. Ezért minden internetről letöltött szöveg felhasználására előzőleg engedélyt kell (kellene) kérni a szerzőktől, illetve a honlap működtetőjétől.

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.

C) F e l d o l g o z á s . – 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 tagolhatjuk:

1. 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 xmldtd-re van szükségünk. A megformázott és annotált szövegek további elemzését az MTA Nyelvtudományi Intézetében végezték el.

2. 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 morfológiában kezelhetőek, sőt a szövegszintaxis ismeretében általában majdnem teljes mértékben egyértelműsíthetőek (a HUMOR program működéséről és az elemzés folyamatáról l. még NOVÁK 2003., NOVÁK – M. PINTÉR 2006.). 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. – l. 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.

(Folytatjuk.)

PINTÉR TIBOR