Záró jelentés, K 60403

A kutatás három nagy részből tevődött össze:

(1) a Budapesti Szociolingvisztikai Interjú 2. változatának (BUSZI-2) lejegyzői és ellenőrei által eltérően értékelt vagy kódolt adatok újravizsgálása és a végső, kutatható változat létrehozása; valamint a különböző konverziók következtében keletkezett adatproblémák tisztázása, adattisztítás, adatkorrekció s hasonlók;
(2) adatlekérdező rendszerek kialakítása;
(3) elemzések és tanulmányok készítése és konferenciákon történő bemutatásuk, valamint rangos folyóiratokhoz publikálásra benyújtásuk.

Ad (1)
Elkészítettük az 50 BUSZI-2 interjú irányított beszélgetéseinek másodszor ellenőrzött változatát, így létrejött e részkorpusz végleges és kutatható formája (Borbély Anna és Bartha Csilla).

Elkészítettük az 50 BUSZI-2 interjú tesztadatainak másodszor ellenőrzött változatát, vagyis 50 x 677 = 33850 item és 2720 szöveges rekord kutatható formáját (Kontra Miklós és Hattyár Helga). E részmunkálat során minden olyan esetben kódoltuk a végső adatot, amikor a lejegyző és az ellenőr egymástól eltérően értékelt, illetve kódoltuk az esetlegesen hiányzó adatokat is.

Ad (2)
A projekt eredményeképpen a BUSZI-2 teljes anyaga – az 50 adatközlővel készített 50 interjú – végleges, kutatásra alkalmas formában áll rendelkezésre. Ez egyrészt az irányított beszélgetésekből álló lejegyzett beszélt nyelvi korpuszt és a hozzá tartozó lekérdező felületet jelenti, másrészt pedig az adatközlőkkel végzett nyelvi tesztfeladatokból nyert adatok gyűjteményét, mely szintén egy lekérdező felület segítségével használható.

Elkészültek a korábbi dBase és FoxPro file-ok összevetéséhez szükséges programok, és megtörtént az adattisztítás és adatkorrekció (Kontra Miklós, Váradi Tamás és Blága Szabolcs munkája). A második ellenőrzés a darabolt CD-file-ok segítségével történt. Ezáltal egy új, nem tervezett munkafázis is belépett az ellenőrzésbe: nem csak a lejegyzett szöveget ellenőriztük, hanem a CD-darabolásokat is (pl. szinkronban van-e a modul neve a darabolt file nevével, illetve benne van-e a darabolt file-ban a lejegyzett rész és fordítva.) Ebben a munkában Hattyár Helga is részt vett.

A fő cél az volt, hogy az irányított beszélgetések végleges lejegyzett formájából olyan nyelvi adatbázis készüljön, amely lehetővé teszi a szövegek gépi elemzését, a kódolt jelenségek hatékony lekérdezését. A lejegyzés eredeti formátuma és kifejtettsége nem teszi ezt lehetővé, emiatt szükséges volt az interjúk szabványos XML formátumba történő átalakítása, melynek során a lejegyzési információ egyértelmű, explicit és számítógéppel hatékonyan olvasható, feldolgozható, elemezhető és kereshető módon jelenik meg az adatbázisban.

Az eredeti kódolás és lejegyzés eredendően nem számítógépes, hanem emberi feldolgozásra készült. Sok helyen nem explicit (pl. a normalizált alakok visszaállításához feltételezi a magyar morfoszintaxis ismeretét), nem egyértelmű (azonos jelölést használ különböző jelenségek – pl. a korrekciónak és az idegen szó leírt változatának – jelölésére; vagy nem világos a jelölések horgonyzása, ill. hatóköre), vagy nem kereshető hatékonyan (pl. a hosszan ejtett hangzók körülményes kódolása ('assszony') miatt).

Az XML általánosan elfogadott szabványos, hordozható adatbázis-formátum annotált korpuszokhoz, készen kapott feldolgozó modulokkal. A BUSZI korpusz XML formátuma szintaktikailag hasonló a korábbi projektumokban (Magyar Nemzeti Szövegtár, Kárpát-medencei Magyar Nyelvi Korpusz) használatos formátumhoz, így a korábban kidolgozott konverziós eljárások, XML szoftvereszközök és az adatbázis-építő program jól adaptálhatók voltak erre a feladatra.

A korpusz megnyilatkozások sorozatának fogható fel, a megnyilatkozás pedig szavak sorozatának, melyek különféle attribútumokkal rendelkezhetnek. E struktúra mellett a beszélt nyelv időbeliségét (átfedő, együtt elhangzó beszéd) is ábrázolnunk kell: az XML formátumú korpuszban az átfedő beszéd kezdetét és végét indexelt horgonypontokkal jelöltük meg.

Más beszélt nyelvi korpuszokhoz képest a BUSZI lejegyzése igen részletes, a jelenségek olyan széles körét és típusait fedi le, melyet más adatbázisok egyáltalán nem tartalmaznak. A kódolt jelenségek a következők: 'l', 't' és 'd' kiesés, a szóbeli pozíciót is megjelölve, 'l' esetében pótlónyúlással és anélkül; magánhangzóharmónia-sértés; hiperkorrekt 'ik'; hiperkorrekt '-bAn'; '-bAn' helyett '-bA'; '-nÁk' ("nákolás"); betűejtés; nem állítmányi '-e'; hosszan ejtett 's'; '-sUk' ("suksükölés"); hiperkorrekt '-sUk'; '-szUk' ("szukszükölés"); hiperkorrekt '-szUk'; hezitációs hangzónyújtás; szünet; hezitáció ("őzés"); hiányzó elem; nemsztenderd névelő. Ezek a jelenségek a korpusz minden releváns pontján kódolva vannak, az adatbázis ezt a részletességű kódolást tartalmazza a fenti követelményeknek megfelelően. Az 'átaba' szóalak esetében például (mely az 'általában' szónak egy változata) egyetlen szón három különböző annotációt találunk: két 'l' kiesést valamint az inesszívuszi 'n' hiányát jelző annotációt.

Az XML formára történő átalakítást Váradi Tamás irányításával és Blága Szabolcs közreműködésével Oravecz Csaba végezte el. (Oravecz egyéb megbízottként történő foglalkoztatását a BUSZI adatbázis kialakítása során jelentkező feladatok nagysága, ezek speciális szaktudást igénylő jellege indokolta.)

Az XML-re konvertálást követően a meglévő gazdag annotációt automatikus nyelvtechnológiai eszközök igénybevételével jelentősen tovább bővítettük az egyes szavakra vonatkozó nyelvi információval, létrehozva ezzel egy olyan beszélt nyelvi adatbázist, mely szótövesítést, egyértelműsített morfológiai elemzést is tartalmaz (Oravecz Csaba munkája). Az ilyen adatbázisra olyan lekérdező rendszer építhető, mely valóban nyelvészetileg releváns és teljes adathalmazt tud szolgáltatni a kutatók részére, szakmai elemzés céljából számos szempont szerint.

A korábbi munkaszakaszban kialakított, validált XML változatból indultunk ki, a nyelvi elemzéshez pedig a Kárpát-medencei Magyar Nyelvi Korpusz építéséhez kifejlesztett nyelvi elemző programláncot használtuk. Az elemzés során az annotáció kiszűrése, a normalizált alakok behelyettesítése és az elemzésnek az XML annotációba való visszaépítése jelentette a feladatot. Az első lépésben az XML annotációt egy program lebontotta, és az elemzendő szóalakokat (ahol releváns, ott a normalizált formában) továbbította a szegmentáló/morfológiai elemző/egyértelműsítő programokhoz. Ezután a lánc kimenetén

előálló morfoszintaktikai elemzést egy újabb program az eredeti XML annotációba visszaépítette, melynek végeredménye lett a végleges nyelvi adatbázis.

Az átalakítás folyamán számos rejtett transkripciós következetlenség került felszínre. Az ilyen típusú hibák megszokottak és elkerülhetetlenek az ilyen nagyméretű, automatikus gépi ellenőrzés nélkül manuálisan annotált korpuszoknál. Ezeket a következetlenségeket típustól függően automatikusan vagy manuálisan lehetett kijavítani. Az 'l' kiesésre irányuló kutatás kapcsán további adattisztítási munkákat is végeztünk. Gépileg támogatott célzott eljárással egyenként végigvizsgáltuk, ellenőriztük és szükség esetén javítottuk az 'l' kieséseket (5737 darabot) a teljes BUSZI-2 korpuszon.

Az imént ismertetett korpuszépítő munka, a kézi annotáció géppel olvasható formára alakítása, melyre a szakirodalom kifejezetten munkaigényes feladatként hivatkozik, esetünkben is nagy erőfeszítéseket igényelt.

Fontos hozzáadott érték tehát, hogy a végső korpuszban nemcsak az eredeti lejegyzésben is szereplő, már ismertetett jelenségek találhatók meg, hanem minden egyes szó annotálva van a következő nyelvi információkkal is: szótő; egyértelműsített morfológiai elemzés; regularizált szótő CV váza, magánhangzók BNF (hátulképzett/neutrális/elölképzett) alakban; elhangzott szóalak fonetikai reprezentációja; illetve a szó regularizált alakja.

Az irányított beszélgetések korpusza 268 ezer szóból áll: 173 ezer szó származik az adatközlőktől, 95 ezer szó a terepmunkásoktól. Egy interjú átlagos hossza 3470 szó, a legrövidebb interjú 1900, a leghosszabb 15000 szavas. A korpusz 31 ezer megszólalást tartalmaz. A korpusz számos jelenség tekintetében statisztikai vizsgálatok elvégzéséhez is elegendő adatot szolgáltat: nagyjából 25 ezer különféle szón belüli kódolt jelenséget (pl.: 'l' kiesés, 't' kiesés, '-bAn' helyett '-bA' stb.); 25 ezer szünetet és 10 ezer hezitációt tartalmaz.

Az előállt szabványos XML formátumú korpuszhoz a szabadon hozzáférhető Emdros korpuszkezelő rendszer használatával készült el a felhasználói felület a nyelvi elemzéseket végző kutatók részére (Sass Bálint munkája). Ez tetszőleges böngésző használatával lehetővé teszi a jelenségek több szempontú menüvezérelt lekérdezését a teljes adatbázisra, illetve ennek a felhasználó által megadott feltételek szerinti meghatározott részhalmazára vonatkozóan. A menürendszer kényelmes hozzáférést biztosít a nagy kifejező erejű Emdros lekérdezőnyelv számos funkciójához, anélkül hogy ezt a formális nyelvet részleteiben ismerni kellene, de lehetőség van a lekérdezőnyelv közvetlen használatára is. A lekérdező az adatok érzékenysége miatt korlátozottan érhető el.

A felületen beállíthatjuk a különféle keresett jelenségeket. Jelenségek sorozatát is megadhatjuk, illetve – a gazdag annotáció adta lehetőségeket kihasználva – a BUSZI korpuszban kódolt jelenségek mellett adott tulajdonságú szavakra, szóalakokra is rákereshetünk (például: a 'fontos' szó összes alakja, a 't'-végű szavak, magánhangzóval kezdődő szavak stb.). Az adatbázisban meglévő fonetikai reprezentáció azt teszi lehetővé, hogy bizonyos hang(kapcsolat)okat tartalmazó szavakat keressünk. A lekérdező felületen elérhető ún. regularizált alak funkciójával egy konkrét szóalak összes megjelenési formáját egyszerre, kényelmesen kereshetjük meg. Ha a 'például' szóalakot adjuk meg regularizált alakként, akkor az összes megjelenési formát ('például', 'pédáu', 'pédául', 'példáu', 'pélá' stb.) valamint a zárójeles, kisbetűs, nagybetűs, megszakított alakokat is megkapjuk. A felületen kiegészítő funkcióként megtalálható a megszólalásszám/beszélőváltásszám, szószám, írásjelszám, illetve az átlagos megszólaláshossz. Ezek a számszerű jellemzők a diskurzuskutatásban kaphatnak szerepet.

A találatokat megjeleníthetjük különféle szempontok szerint rendezett konkordancia formájában, készíthetünk gyakorisági listát, valamint összefoglaló táblázatot, mely kvóták (=foglalkozási csoportok) szerint, modulok szerint és összesítve mutatja a kérdezett jelenség

előfordulási számát. A BUSZI korpuszban minden egyes megszólalásról tudjuk, hogy mely interjú, melyik moduljában van, illetve hogy az adatközlő vagy a terepmunkás szájából hangzott-e el. A lekérdezéseket e három szempont szerint szűkíthetjük alkorpuszokra. Természetesen lehetőség van teljes beszélgetések megjelenítésére is.

A 268 ezer szavas korpuszt kezelő rendszer válaszideje általában néhány másodperc. A használt eszköznek köszönhetően a válaszidő lényegében nem függ a korpusz méretétől, csak a konkrét válasz méretétől. Következésképpen a ritkább jelenségek eredményei azonnal megjelennek, az 5737 darab 'l' kiesés teljes összesítő táblázatának elkészítése pedig 20–30 másodpercet vesz igénybe. A lekérdezőt a kutatóink többek között az alábbi témák kutatása során használták: (bVn) változó, 'l'-kiesés, valamint a beszélt nyelvben lévő főnévi csoportok jellegzetességei.

A vállaltakon túli plusz feladatként megkezdődtek a korpusz bővítésére irányuló munkálatok. Ezek keretében elkészült a BUSZI-3 és BUSZI-4[1] tesztadatainak rögzítésére és ellenőrzésére szolgáló rendszer, a lejegyzők igényeinek megfelelő rugalmas navigációval, a végleges BUSZI-3 és BUSZI-4 itemlistával (Hattyár Helga, Mátyus Kinga és Sass Bálint munkája). Az adatok táblázatos formátumban exportálhatók az adatbázisból. A rendszer intenzív használatával 2009-ben 142 új interjú tesztadatainak első rögzítése történt meg.

Ad (3)

A két adatbázist felhasználva elemzéseket készítettünk, melyeket bemutattunk a Budapesti Szociolingvisztikai Interjú, I. Szimpóziumon (2008. december 9-én) és II. Szimpóziumon (2009. október 20-án). A II. Szimpóziumon lehetővé tettük számos határon túli magyar nyelvész és nyelvész doktorandusz részvételét is.

Egy szakdolgozat készült a BUSZI-2 korpusz adataiból (Szeredi Dániel, ELTE, 2008, témavezető: Bartha Csilla).

Két magyarországi PhD hallgatót jelentősen bevontunk a munkálatokba, amit publikációik is jeleznek (Mátyus Kinga, Vargha Fruzsina Sára).

Előadásokat tartottunk több nemzetközi konferencián is, és sokat profitáltunk annak a „workshop”-nak következtében, amelyen Jack Chambers professzor (University of Toronto) részletesen kommentálta s kritizálta kutatóink BUSZI-elemzéseit 2009. decemberében.

A pályázat futamideje alatt – a 2010. február eleji állapot szerint – keletkezett hét már megjelent tanulmány, s további 4 tanulmányt közlésre elfogadtak.

[1] A BUSZI-2 interjúkat 50 fős kvótamintával készítettük, a BUSZI-3 és BUSZI-4 interjúkat 200 fős rétegzett reprezentatív mintával.