A magyar jelnyelvi korpusz létrehozásának és annotálásának kihívásai

Bartha Csilla[1], Varjasi Szabolcs[1], Holecz Margit[1]

[1] Magyar Tudományos Akadémia, Nyelvtudományi Intézet, Többnyelvűségi Kutatóközpont, 1068 Budapest, Benczúr u. 33.

Kivonat: A 2015. október 31-én zárult JelEsély Projekt keretében egy hozzávetőlegesen 1750 órányi jelnyelvi korpusz jött létre. Országos terepmunka során 147 szociolingvisztikai interjú készült 5 régióban és 9 helyszínen, 27 grammatikai teszt során pedig 54 adatközlővel készültek felvételek (interjúnként 2 adatközlővel). Ahhoz, hogy a létrejött videoalapú korpusz kereshető, kutatható és felhasználható legyen, szükség van egyrészt a korpusz annotálására, amely folyamat során különféle információkat kapcsolunk a felvételekhez, másrészt a jelnyelvi felvételek fordítására. Írásunkban a jelnyelvi korpuszépítés és annotáció egyedi kihívásait ismertetjük, melyek többsége két okra vezethető vissza, melyek összefüggenek a jelnyelvek sztenderdizációjának kérdéseivel is. Egyrészt a jelnyelveknek nincs kidolgozott és elfogadott írásrendszerük, másrészt a jelnyelvekre – a sztenderdizálatlan hangzó nyelvekhez hasonlóan –jellemző a nagyfokú változatosság. A kereshető, immár géppel is olvasható korpuszok számos további kutatási lehetőséget biztosítanak, az alapvető statisztikai vizsgálatokon túlmenően is. A szociolingvisztikai kutatások mellett lehetővé válik korpuszalapú szótár létrehozása, valamint egy valós nyelvhasználaton alapuló grammatika megalkotása is. Vizsgálhatóak továbbá diskurzusjelenségek, pragmatikai sajátosságok és a siketes jelek is. A korpusz ezen kívül oktatási célokat is szolgálhat, például tan-, és segédanyagok létrehozásával.

1. Bevezetés

A siket közösség Magyarország harmadik legnagyobb nyelvi kisebbsége, annak ellenére, hogy „a veleszületett vagy szerzett halláskárosodás folytán a siket közösségek nem etnikai alapon szerveződnek, nyelvi kisebbségek abban az értelemben is, hogy sajátjukként *bármely más (hangzó) nyelvvel egyenértékű teljes, autonóm természetese nyelvet, jelnyelvet használnak*" [1: 85 – kiemelés az eredetiben]. Munkálataink során, a Többnyelvűségi Kutatóközpontban a siketségnek a jelnyelvet forrásként kezelő, nyelvi-szociokulturális megközelítésére alapozunk, szemben a fogyatékosságparadigma deficit-alapú megközelítésével: „[…] a kulturális, antropológiai értelmezés a siketséget egy olyan embercsoport létállapotának, adottságának tekinti, amely tagjainak közös vonása, hogy a világot elsődlegesen vizuálisan érzékelik, akiket közös kultúra, hasonló tapasztalatok, viselkedési szokások jellemeznek, s legfőképpen, közös nyelvet, a jelnyelvet használják, amely elsődleges kommunikációs és megismerő szerepe mellett – más nyelvi közösségekhez hasonlóan – önazonosságuk szimbóluma

is [2: 79]". Ebből adódóan a siket közösségek tagjai tehát nemcsak siketek és nagyothallók lehetnek, de hallók is (pl. siket szülők halló gyermekei, siket gyermekek családtagjai és a közösséghez csatlakozó, annak értékeivel, nézeteivel azonosuló hallók) [vö. 4, 5].

A 2009. évi CXXV. törvény a magyar jelnyelvről és a magyar jelnyelv használatáról mérföldkő volt a siket közösség életében. Nemcsak azért, mert a magyar jelnyelvet önálló, természetes nyelvként ismeri el, hanem azért is, mert biztosítja a jogi keretet a bilingvális oktatás 2017-től való bevezetésére. A bilingvális oktatás kidolgozásához azonban szükség van a magyar jelnyelv oktatási célú sztenderdizációjára. Ez a folyamat csak a siket közösség tagjainak bevonásával valósulhat meg, a megalapozásához pedig szociolingvisztikai alapon megtervezett, korpuszalapú empirikus nyelvészeti kutatásra van szükség. Ezt a célt tűzte ki a TÁMOP 5.4.6/B-13/1-2013-0001 *A magyar jelnyelv sztenderdizációjának elméleti és gyakorlati lépései* (Jel*E*sély) elnevezésű projekt.

A következőkben a projekt során létrehozott, folyamatos fejlesztés alatt álló korpuszt mutatjuk be.

2. A korpusz bemutatása

2.1. A korpusz mint kutatási bázis

Leech már a 90-es évek elején megfogalmazta, hogy a korpusznyelvészet valójában egy módszertani bázis, így könnyen alkalmazható a nyelvészet különféle területein, például a fonetikában vagy a szociolingvisztikában [32]. Rundell pedig a következő jövőképet vázolja 1996-ban: „Mindazok számára, akik a nyelvtanulás, nyelvi leírás, illetve nyelvoktatás bármely területén dolgoznak, a korpusz használata olyan természetessé és nélkülözhetetlenné fog válni, amilyen a lexikográfusok számára jelenleg [38]."

A korpuszok osztályozása többféle módon történhet, ily módon a szakirodalomban is különféle korpusztípusokkal találkozhatunk. A *referenciakorpusz* célja, hogy átfogó információt adjon egy nyelvről, annak minden fontos változatát és a szókincs jellegzetességeit is reprezentálja, ezáltal megbízható nyelvtanok, szótárak, tezauruszok és egyéb nyelvi referenciaanyagok alapjául szolgálhatnak [40]. Az anyagok kiválasztása során meghatározásra kerülnek azok a paraméterek, amelyek alapján adott szövegek a korpusz részévé válhatnak. Ez magába foglalja a lehető legtöbb szociolingvisztikai változó figyelembe vételét, valamint az egyes szövegtípusok arányának meghatározását. A *monitorkorpuszok* lehetővé teszik, hogy a nyelv időbeli változását is nyomon követhessük, a *párhuzamos korpuszok* esetében pedig a szövegek mellett megjelennek azok különböző nyelvű fordításai is. *Összehasonlítható korpuszról* akkor beszélhetünk, ha több mint egy nyelv vagy nyelvváltozat hasonló szövegei jelennek meg, a hasonlósági kritérium azonban nincs pontosan definiálva. A korpuszok többféle jellemzővel írhatóak le, ahol minden jellemzőnek van egy „alapértelmezett"értéke. Ha bármely jellemző eltér ettől, akkor már *speciális korpusz*ról beszélünk. Az alapértel-

mezett értékek: mennyiség=nagy, minőség=autentikus, egyszerűség=egyszerű szöveg, dokumentált=igen.

Jelnyelvek esetében a korpuszok nemcsak dokumentálják, megőrzik az egyes jelnyelveket, de ezzel együtt autentikus szövegekhez is hozzáférést biztosítanak. (Jel)Nyelvek reprezentatív mintáit szolgáltatják, miközben grammatikák vagy szótárak alapjait is képezhetik, hosszú távon pedig nyomon követhető a nyelv változása is.

A jelnyelvészeti korpusznyelvészet kialakulása a technológiai fejlődés függvényeként is értelmezhető. A hangzó és írott nyelvi korpuszok a 20. második felétől kezdve váltak egyre elterjedtebbé (Rundell 1996-ban kifejti, hogy az angol mellett egyéb nyelveken is megindultak a korpuszmunkálatok, a 90-es években pl. már több mint 12 nyelven voltak különböző korpuszok Európában. [38:7]. Fontos azonban kiemelni, hogy már a Chomsky előtti időszakban is voltak korpuszalapú kutatások, melyet korai korpusznyelvészetnek nevezünk [34]. Ez elsősorban helyesírási konvenciók meghatározására használt gyűjtemények, illetve a nyelvelsajátítás nyomon követésére vezetett naplók formájában valósult meg.

McEnery és munkatársai a kisebbségi nyelvi tervezés problematikájával kapcsolatban emeli ki, hogy széles körű kutatásokat és szoftveres erőforrásokat nem lehet hatékonyan létrehozni korpuszos források hiányában, emellett egynyelvű és párhuzamos korpuszokból származó adatokra is szükség van [34]. Az ind nyelvekkel kapcsolatban hangsúlyozzák, hogy a magas fokon sztenderdizálatlan szövegek esetében a szöveges kódolás kulcsfontosságú kihívást jelent.

Ez a kihívás a jelnyelvek esetében még hangsúlyosabban jelenik meg, ahol problémát jelent az eltérő modalitás és írásbeliség hiánya is. A jelnyelvek esetén -- kezdetleges formában ugyan – a notációs rendszerek kialakulásával (ld. lentebb) indulhattak meg a gyűjtések. A 90-es években több fontos előrelépést érdemes kiemelni, egyrészt a tárolóeszközök közül a digitális CD, majd a nagy teljesítményű háttértárak váltották föl az élő nyelvi szövegek kazettáit; másrészt megjelentek a beszélt nyelvi korpuszok is. Közülük külön kiemelendő a Wellington-korpusz, hazai tekintetben pedig a Budapesti Szociolingvisztikai Interjú (BUSZI) [30], majd a kétezres évekből a Kárpát-korpusz [27] és a BEA adatbázisa [7]. Az írott korpuszok közül a Magyar Nemzeti Szövegtár a legfontosabb. A magyar nyelvterületen létrehozott különböző adatbázisok elérhetőek a Nyelv- és beszédtechnológiai platform honlapján [35].

Bartha rávilágít arra a visszásságra, hogy habár az emberek jelentős része a mindennapi tevékenységei során a legutóbbi időkig (a számítógép és egyéb eszközök megjelenéséig, melyek új nyelvhasználati lehetőségeket hoztak magukkal) a beszédet részesítette előnyben (az írott nyelvvel szemben), ám a hozzáférhető korpuszok fordított arányokat mutatnak [3]. Ez részben azzal magyarázható, hogy a beszélt nyelvi diskurzusminták gyűjtése és átírása lényegesen nagyobb nehézséget jelent, mint az írott nyelvieké.

A nagymennyiségű adatok tárolása, a nyelvi adatok dokumentálása és megfelelő rendszerezése a legtöbb empirikus adatokkal dolgozó kutató számára fontos kérdéssé vált, pl. a szociolingvisztikában is [vö. 28]. A nagy mennyiségű szövegek tárolása az ezredfordulóra már adott volt, azonban ahhoz, hogy az egyszerű, általában CD-n tárolt jelnyelvi archívumokból valódi korpuszok jöhessenek létre, további fejlődésre volt szükség, így a jelnyelvi korpusznyelvészet kialakulásában fáziskéséssel kell számolnunk. Jelenleg több területen hiányzik még ezen tudományág kiforrott módszertana,

amely lehetőséget teremt egyrészt a fejlődésre, másrészt az írott nyelvi korpuszok tanulságainak implementálására.

A jelnyelvi korpuszok többnyire jelenleg is fejlesztés alatt állnak [13]. Habár már 1910 és 1920 között is készült korpusznak tekinthető gyűjtemény [29], de ezt követően hosszú idő telt el, míg a modernnek tekinthető korpuszok létrehozását célzó projektek elindultak a kétezres évek elején. Ezek közül legjelentősebbek a 2006-2008 között futó holland projekt a nijmegeni Radboud Egyetem koordinálásában [15], a veszélyeztetett nyelvi státusszal rendelkező ausztrál jelnyelv (Auslan) nyelvtanát és diskurzusstratégiáit dokumentáló 2004-től 2007-ig zajló korpuszprojekt [17], a brit jelnyelv (BSL) korpuszát létrehozó három és fél éves (2008 januárja és 2011 júniusa között futó) projekt [8], valamint az a jelenleg is tartó 15 éves projekt, amely a német jelnyelv korpuszának létrehozását tűzte ki célul [16]. A 2-3 év alatt összeállított nyersanyagok feldolgozásán, közzétételén, és felhasználásán (szótárak, oktatási anyagok, grammatikai vizsgálatok stb.) folyamatosan dolgoznak.

2.2. A magyar jelnyelvi korpusz létrehozása és feldolgozása

2.2.1. A korpusz felépítése

A korpusz fő alkotóelemei szociolingvisztikai és grammatikai tesztek felvételei, melyek több hónapon át zajló, országos terepmunka során készültek el. 7 mintavételi pontról (Budapest, Szeged/Hódmezővásárhely, Békéscsaba, Debrecen, Kaposvár, Sopron/Győr, Vác) 16 siket terepmunkás részvételével összesen 147 szociolingvisztikai interjú készült el, melyek közül 67 budapesti és 80 vidéki. Az interjúk 345 kérdésből álltak, a felvételek három kamerával való rögzítése pedig átlagosan 3-4 órát vett igénybe.

A grammatikai tesztek során 27 terepmunka alatt összesen 54 adatközlővel készültek felvételek (interjúnként 2 adatközlővel, melyek közül 11 teszt vidéki, 16 pedig budapesti adatközlővel készült). A grammatikai tesztek (a magyar jelnyelv alapgrammatikájának megírásához szükséges elicitációs tesztsorok) felvétele 5 kamerával zajlott, és átlagosan két órásak voltak.

A nyers videofelvételek feldolgozása többlépcsős munkafolyamatban zajlott. Az anyagokat először archiváltuk és vízjeleztük, ezt követően konvertáltuk. A korpusz nyersanyaga hozzávetőlegesen 1750 órányi, ami 6,5 terabájtnyi adatot jelent.

2.2.2. A korpusz feldolgozása

Ahhoz, hogy a videoalapú korpusz kereshető, kutatható és felhasználható legyen, szükség van a korpusz annotálására, mely folyamat során különféle információkat kapcsolunk a felvételekhez (pl. a felvételek közben megjelenő kézformák, a használt jelek magyar megfelelője stb.). Az annotációs részfolyamatokban nemcsak annotátorok, de fordítók és ellenőrzők is dolgoztak.

A kutatási céloknak megfelelően más-más protokollt alkalmaztunk a szociolingvisztikai és a grammatikai korpusz annotálásakor. A külföldi jelnyelvi korpuszprojektek áttekintését követően a szociolingvisztikai anyagoknál a jelnyelvalapú fordítás volt az elsődleges, amely azt jelentette, hogy a fordítók jelről-jelre haladtak folyamatosan, és nem csupán tartalmi összefoglalót készítettek. Ezáltal biztosítható a jelelni nem

tudó kutatók számára a korpusz anyagához való hozzáférés magyar nyelven, hiszen a jelnyelveknek, köztük a magyar jelnyelvnek nincsen általánosan elfogadott és széleskörűen használt írásrendszere. Habár több kezdeményezés is született a jelek írásbeli rögzítésére, mint például a HamNoSys [18] vagy a SignWriting [39], de ezek egyrészt jól jelelők számára is sokszor nehezen olvashatóak, másrészt jelelni nem tudó kutatók számára nem hozzáférhetőek[1]. A jelnyelvi videók magyar nyelvű fordítására tehát annak ellenére is szükség volt, hogy bizonyos esetekben elfedik a jelnyelv változatosságát (vö. [26]), illetve azt sugallja, hogy a jelnyelv és a magyar hangzó nyelv elemei között lehetséges az egyértelmű megfeleltetés, de ez természetesen nem igaz [vö. 5, 42, 43]. A jelnyelvek a hangzó nyelvekhez hasonlóan természetes nyelvek [41] melyek ugyanakkor a magyartól és más hangzó nyelvektől nagymértékben eltérő struktúrával és nyelvi eszközkészlettel rendelkeznek. Mindezek ellenére szükséges a halló kutatók számára is hozzáférhetővé tenni a korpuszt. Folyamatosan készülnek a magyar fordítások, mely menetét és irányelveit a későbbiekben fogjuk ismertetni.

A számítógéppel feldolgozható jelnyelvi korpusz létrehozásakor az egyik legnagyobb kihívás a magyar jelnyelv (manuális és/vagy nonmanuális komponensekből álló) elemeinek következetes azonosítása a korpusz egészében. Ennek a kérdésnek a megoldására a nemzetközi gyakorlatban kétféle megoldást találunk [vö. 10, 11, 18]. Egyik út az ún. notációs rendszerek használata, amelyek célja, hogy olyan pontos fonológiai leírást adjanak a jelekről, hogy azok kivitelezése lemásolható legyen. Ilyen notációs rendszerek kialakítása elsősorban a jelnyelvkutatás korábbi időszakára jellemző. A legismertebb közülük a HamNoSys rendszer, amelyet a hamburgi egyetem munkatársai fejlesztettek ki. A jelnyelvi lexikográfiában és korpuszelemzésben használt egyik szoftveres megoldás az iLex rendszer, melynek központi részét képezi a HamNoSys-ben történő átírás.

A másik megoldás a jelnyelvi írásrendszer hiányának kiküszöbölésére egy következetes jelölés alkalmazása, amely minden jelformát egyedileg azonosít. A jelnyelvi jelek egyedi formai azonosítóját ID-Glosszoknak nevezzük [22, 25, 12]. Mivel több ezer jelről van szó, ezért a gyakorlatban nem alkalmazhatunk tetszőleges kódrendszert (például számokat) – ez megnehezítené a gyakorlati felhasználást, ellehetetlenítené a keresést. Fontos megemlíteni, hogy habár az ID-Glosszok elnevezése utalhat az adott jel központi jelentésére, ez a megfeleltetés nem szükségszerű, de megkönnyítheti az adott kódhoz tartozó forma felidézését. Nem utal továbbá az ID-Glossz hangzó nyelvi szófaja az adott jel szófajára, annál is inkább, mivel a szófajiság megítélése különbözik a jelnyelvekben és a hangzó nyelvekben [37]. Jelnyelvek esetében kevésbé élesek a szófaji határok, a szófaji felosztásról pedig még nem született konszenzus a nemzetközi szakirodalomban.

2.2.3. Az ELAN szoftver alkalmazása

A projekt során áttekintett és mintául szolgáló jelnyelvi korpuszok (a holland [15], a brit [8], az ausztrál [22]) a hamburgi és a lengyel kivételével a Max Planck Institute által fejlesztett ELAN szoftvert használják, amely lehetővé teszi multimédiás anyagok

[1] Természetesen ezek mellett is számos alternatív lejegyzési módszer használatos, melyeket a gyakorlati igény hívta életre, gyakran találkozhatunk velük a jelnyelv mint idegen nyelv képzések során akár a diákok, akár az oktatók esetén.

annotálását. Alkalmas egyszerre több videó párhuzamos lejátszására, ez különösen fontos a jelnyelvi annotáció szempontjából. Maximálisan négy kamerakép egyidejű megtekintését biztosítja, valamint lehetőség van a felvételek utólagos összeszinkronizálásra abban az esetben, ha a felvételeket nem egyszerre indították. A program hátránya, hogy (az iLex-el szemben) nem kapcsolódik közvetlenül lexikai adatbázishoz, azonban – köszönhetően annak, hogy az ELAN szabad forráskódú – a készülő szótár és a szótár mögött álló lexikai adatbázis közötti kommunikációt sikerült megoldanunk.[2]

Az egyes elemzési szempontok külön szinteken, úgy nevezett tierekben jelennek meg, pl. kézforma vagy mozgás. A különböző adatokat tartalmazó tierek száma végtelen lehet.

Sem az ELAN-nak, sem az iLexnek nem volt magyar nyelvű változata a JelEsély projekt kezdetén, annak ellenére, hogy számos más nyelven elérhetőek. Az akadálymentesítés biztosításának érdekében elkészült a magyar fordítás, amely jelenleg még a mindenki által használt funkciókra tér ki, a bonyolultabb keresési és néhány egyéb, ritkán használt funkció fordítása még nem történt meg.

Az ELANban bizonyos elemzési szinteken az annotátorok egy legördülő listából kiválaszthatják az annotációs értékeket, ezeket a listákat kontrollált szótáraknak (controlled vocabularies, a továbbiakban CV) nevezzük. A CV-k nagy segítséget jelentenek a következetes annotálás elősegítésére, valamint elkerülhetőek az elütések is általuk. Használatuk azonban megköveteli, hogy az annotáció kezdete előtt meghatározzuk az adott kategória lehetséges elemeit. Az ELAN eredetileg nem teszi lehetővé a kontrollált szótárak értékeinek módosításait a munka kezdetét követően, azonban más projektek saját fejlesztésű scriptjei ezt a problémát már megoldották.

A jelnyelvi korpuszok létrehozásánál megkerülhetetlen az elemzési szempontok előzetes összeállítása. A JelEsély projekt grammatikai és szociolingvisztikai munkacsoportjaival együttműködve jött létre három sablon, melyek tartalmazzák azoknak az elemzési szinteknek a listáját, melyeket a magyar jelnyelv és (jel)nyelvhasználat vizsgálatakor előzetesen fontosnak tartottunk. A jövőbeni kutatásokhoz összesen 140 különféle elemzési szempontot határoztunk meg résztvevőnként (a szociolingvisztikai-grammatikai, célzott grammatikai és szótári annotáció során). Ezek egymással részben kompatibilisek, és van lehetőség a későbbi egyesítésre.

2.2.4. Az annotáció kihívásai

Annak ellenére, hogy a projekt során külön kezeljük a szociolingvisztikai és a grammatikai korpuszt, továbbá, hogy ezek feldolgozása más-más módon és céllal kezdődött el, hosszú távon mindkettő feldolgozásakor ugyanazokkal a kihívásokkal szembesülünk. A következő szakaszokban ezeket a kihívásokat foglaljuk össze, a jelenlegi állapotot bemutatva, függetlenül attól, hogy az eddigi munkánk során melyik részkorpusszal kapcsolatban merültek fel.

[2] Az ELAN-ban az ID-Glosszok listája a szótári adatbázisból frissíthető. Ez jelenleg csak egyirányú szinkronizációt jelent, az optimális ugyanakkor az lenne, ha az ELAN-ban megadott, új ID-Glosszok is bekerülnének a szótári adatbázisba, amely megfelelő ellenőrzési protokoll után megjelenhetne a szótári felületen is.

Az annotátorok és fordítók kiválasztásakor is fontos volt a siket közösség tagjainak lehető legnagyobb mértékű bevonása. A terepmunka és a további kutatási feladatok tervezéséhez hasonlóan itt is fontos volt, hogy az egyéni kompetenciákra és preferenciákra építve (a magas fokú magyar jelnyelvi kompetencia mellett a magyar nyelvtudás, illetve megfelelő számítógépes ismeretek voltak szükségesek) osszuk szét a feladatokat az annotátorok között. Külön nehézség volt a szociolingvisztikai annotáció során a potenciális CODA (Child of d/Deaf Adult, siket szülő halló gyermeke) munkatársak felkutatása. Később nagyothallók és a közösség által elismert tolmácsok bevonása jelentett megoldást. A szociolingvisztikai anyagok lejegyzése során próbáltunk alkalmazkodni a lejegyzők igényeihez (voltak, akik számára a fordítás azonnali gépelése volt a gyorsabb, míg mások a diktafonba fordítást preferálták). Hasonló elvek alapján kerültek kiválasztásra a grammatikai annotációt végző munkatársak is.

Kiemelten fontos volt az annotátorok oktatása annak érdekében, hogy megismerjék, és készségszinten tudják kezelni az annotációhoz használt szoftvert; valamint, hogy megértsék a feladatot, biztosítandó az annotáció következetességének megőrzését. A formális oktatás mellett folyamatosak voltak az informális megbeszélések, továbbá több feladatspecifikus leírás is készült számukra.

A legtöbb annotátor nem a Többnyelvűségi Kutatóközpontban végezte a munkáját, hanem otthonról. Jelenleg még nem épült ki nagymennyiségű videófájlok kezelésére és mozgatására alkalmas hálózat, ennek megvalósítását a későbbiekben tervezzük, mivel ennek hiányában az annotáció (főként kiadott fájlok és feladatok) dokumentálása, folyamatkövetése nagy adminisztratív terhet jelent.

A jelnyelvi videók hangzó nyelvre való fordítása során több elméleti és módszertani problémával szembesültünk, melyek közül néhányat már érintettünk. Annak ellenére, hogy a fordítói protokoll készítésekor törekedtünk a feladat pontos leírására, a jelnyelvi fordítás – hasonlóan a hangzó nyelvihez – nem törekedhet arra, hogy egyszerre adja vissza a jelnyelvekre jellemző sajátos mondatszerkezetet és jelentésalkotási stratégiát; valamint a mondat jelentésének megértéséhez szükséges magyar nyelvtani rendszert követő fordítást. Ez az elméleti probléma a gyakorlatban azt jelentette – annak ellenére, hogy CODA (siket szülő halló gyermekeként felnőtt, esetünkben mindkét nyelven magas kompetenciájú személy), vagy a közösség által elfogadott tolmács végezte a fordítási munkákat –, hogy több munkatárs nem vállalta a feladatot, vagy első elvállalás után nem folytatták a munkát. Ez elsősorban azzal magyarázható, hogy a jelnyelvi sajátosságokat visszaadó, jelről jelre haladó magyar fordítást kértünk a fordítóktól, nem pusztán tartalmi fordítást. Ez pedig olyan feladat, amellyel a legritkább esetben találkoznak mindennapos nyelvi környezetükben a tolmácsok és a CODÁ-k is. Az annotátorok és a fordítók egyéni kompetenciáihoz nagymértékben kellett alkalmazkodni a fordítás során, bizonyos esetekben még akkor is, ha ez módszertani problémákat is felvetett. A kutatás során a hosszú távú cél, hogy az annotációhoz használt szoftver felületén megjelenve a fordítások időben összekapcsolódjanak a releváns beszédeseménnyel (jelelési eseménnyel). Fontos volt továbbá szem előtt tartani, hogy a projekt szűk időkerete megkövetelte a gyors munkavégzést. Emiatt döntöttünk később úgy, hogy a számítógépet nem jól kezelő annotátorok diktafonba fordítsák a jelnyelvi videók anyagát, ami pedig később kerüljön begépelésre. Ez ugyan nem alkalmas a videókkal való azonnali összekapcsolásra, ugyanakkor nagymértékben meggyorsította a munkát. A projekt szellemiségével összhangban a gépelők között

látássérült munkatársak bevonására is sor került, emellett a számítógépet készségszinten használó, és nagy sebességgel gépelő munkatársak ELAN oktatása is folyamatban van. Kidolgozásra került továbbá az az eljárási mód, ahogyan a különböző szövegformátumú (de a videókkal nem összekapcsolt) fordítások ELAN-ba importálhatóak, ahol már a felvételekkel összekapcsolva, időkódokkal jelennek meg. Mivel szövegszerkesztőkben nem jeleníthető meg párhuzamosan a jelnyelvi változat és a fordítás, ezért a fordítások ellenőrzése problémát jelentett. Az ELAN-ba való későbbi importálás során a fordítások újraellenőrzésére és megfelelő szegmentálására sort kell keríteni.

Az általánosan elfogadott jelnyelvi írásrendszer hiánya mellett számos további problémával szembesültünk, amely a jelnyelvek sajátosságaiból adódnak. Ilyen alapvető kérdéskör a jel kezdetének és a jel végének a meghatározása, amely a videóanyagok tokenizálása során jelentkezett. Annak ellenére, hogy nincs egységes álláspont a nemzetközi szakirodalomban ezzel kapcsolatban sem, szükséges volt meghatározni, hogy az annotátorok milyen kritériumok alapján járjanak el a szegmentáció során. A későbbiekben tervezzük ennek a felülvizsgálatát, ellenőrzését is. A jelszegmentáció alapvető kérdése, hogy a jelelést folyamatos jelfolyamnak (ahol egy jelhez nemcsak az ún. tiszta fázis, hanem az átvezető mozgások is hozzátartoznak), vagy *jel→átmeneti mozgás→jel* folyamnak tekintjük. Számos oka van annak, hogy végül az első lehetőség mellett döntöttünk. A legfontosabb, hogy ne egy előre meghatározott konstrukcióval közelítsünk az egyik legfontosabb jelnyelvi elem felé, ne egy adott elméleti elgondolás mentén tekintsünk egy jelenséget jelnek, míg egy másik jelenséget átmeneti mozgásnak, hanem valóban alulról-felfelé építkezve, az adatokból elindulva határozzuk meg a jel fogalmát.

Ezek alapján „tág" szegmentumokat hoztunk létre, tehát a jel akkor kezdődik, amikor a kéz vagy kezek irányváltást kezdenek, miután az előző jel kivitelezéséhez szükséges összes mozgást befejezték ÉS/VAGY amikor a kéz vagy kezek elkezdik megváltoztatni a kézformát, ha az nem része az előző jel artikulációjának. A jelnek vége van: (1) Még mielőtt a kéz vagy kezek elkezdenének irányt változtatni, miután befejezték az aktuális jel kivitelezésének összes releváns mozgását ÉS/VAGY (2) még mielőtt a kéz vagy kezek elkezdenék megváltoztatni a kézformát, ha az nem része az előző jel artikulációjának. Továbbá (3) amikor a kéz vagy kezek elkezdenének visszatérni a pihenési pozícióba (pl. keresztbe tett karok, kezek a csípőn, vagy karfán, vagy a test mellett.). A kéz vagy kezek kivitelezési helyen való megállítása és pihentetése (a kézforma megtartásával) a jel részét képezi. A szakasz addig tart, amíg a „pihenés" véget ér, és a kéz vissza nem tér a nyugalmi helyzetbe vagy el nem mozdul egy következő jel kivitelezése felé. A félbehagyott jeleket, és minden kezekkel kapcsolatos jelenséget szegmentálni kell (ez alól kivétel a nyelvileg nem értelmezhető cselekvés). Hezitálásokat, szókereséseket és egyéb (feltehetően) megakadás-jelenségeket is szegmentálni kellett.

További alapvető problémát jelent a magyar jelnyelv kézforma-állományának a kérdése. A magyar jelnyelv szublexikális szintjeinek leírására korábban született monográfia természetesen foglalkozik a kézformák kérdéskörével is: [42], [43], de a probléma tisztázását célzó további vizsgálatok még folyamatban vannak. A magyar jelnyelvben használt, fonémának tekinthető kézformák meghatározása nélkül nem lehetséges a jelenségek következetes jelölése, ráadásul ennek a kérdésnek nagy jelen-

tősége van a sztenderdizációs folyamat egészét és a hallásállapottól független módon értelmezett jelnyelv-tanulói közösséget nézve is.

A jelnyelv fonológiai[3] komponensei, tehát a kézkonfiguráció (kézforma és kézformaváltás, orientáció, érintkezés testrésszel vagy másik kézzel, egy- vagy kétkezes) mellett a mozgás, a kivitelezési hely, a nonmanuális elemek, valamint orális elemek (szájkép) vesznek részt a jelnyelvi produkcióban [42]. A nyelvleírásnak csakúgy, mint a korpuszépítésnek alapvető feladata meghatározni a fenti kategóriák lehetséges értékeit (például a lehetséges mozgástípusokat). A külföldi jelnyelvi korpuszmunkálatok és grammatikai leírások, valamint egyéb nem nyelvészeti, de releváns kutatások alapján (pl. emócióelemzés és gesztuskutatás) meghatározott elemek, illetve a hazai siket közösség képviselőinek meglátása alapján dolgoztuk ki ezeknek a kategóriáknak a rendszerét, melyek az annotáció jelen szakaszában tesztfázisban vannak.

A jelnyelvi korpuszok létrehozásának és annotálásának számos hasonlóan új területe van, amelyekre jellemző, hogy több esetben empirikusan nem igazolt állítások, csoportosítások és hipotézisek várnak tesztelésre. Annak érdekében, hogy az annotálást végző munkatársak egy következetes segédlethez hozzáférjenek, létrehoztunk egy ún. annotációs vitaanyagot, amely tartalmazza egyrészt a munkafolyamat protokollját, másrészt eligazítást ad a jelnyelvi annotáció néhány kérdésében (a lexikális és fél-lexikális jelek és a nonmanuális komponensek, ismétlések és az artikuláció annotálása, stb.) Másik célja, hogy az annotációt tervező munkatársak közös referenciaanyagot hozzanak létre, amelyben az egyes nyelvi elemek annotációját megvitathatják. Ahogy a neve is sugallja, ez a dokumentum nem tekinthető véglegesnek. Az annotációs vitaanyag több hasonló külföldi anyag mintájára készült el [26, 9, 14, 43], elsősorban Trevor Johnston korábban hivatkozott anyagán alapul, amelyet évről-évre frissítve elérhetővé tesz, és az ausztrál jelnyelvi korpusz annotációja során használják.

A korpuszannotáció ciklikus volta lehetővé és szükségessé is teszi a projekt indulásakor meglévő tudásunk újraértelmezését. Az új ismeretek, új kihívások lehetővé teszik az annotációhoz kidolgozott rendszer folyamatos fejlesztését, fejlődését.

2.3 A korpusz felhasználási lehetőségei

A korpusz széleskörű felhasználási lehetőségeit röviden már érintettük a 2.1. fejezetben. A kereshető, immár géppel is olvasható korpuszok számos további kutatási lehetőséget biztosítanak, az alapvető statisztikai vizsgálatokon túlmenően is. A szociolingvisztikai kutatások (pl. területi és társadalmi változatosság) mellett lehetővé válik korpuszalapú szótár létrehozása, melynek során kiemelkedően fontos irányelv a jelnyelv-központúság; valamint egy valós nyelvhasználaton alapuló grammatika megalkotása is. Vizsgálhatóak továbbá diskurzusjelenségek, pragmatikai sajátosságok és a siketes jelek is. A korpusz ezen kívül oktatási célokat is szolgálhat, például tan-, és segédanyagok létrehozásával.

[3] Stokoe a fonológia, fonéma és allofón mintájára bevezeti a kerológia, keréma, alloker fogalmakat [41], de ezt a megkülönböztetést később ő maga sem látja szükségesnek. Egyrészt a közös fogalomrendszer rávilágít a hangzó nyelvek és jelnyelvek közös vonásaira, valamint ezek a fogalmak ugyanolyan adekvátak és megfelelőek a jelnyelvek leírásakor is, mint hangzó nyelvek esetében [6].

A jól annotált, számítógéppel feldolgozható korpusz a szótárkészítés alapja lehet, a felvételekből származtatott jeladatbázis a nemzetközi kutatási normáknak megfelelve jelnyelv-központú szótár létrehozását teszi lehetővé. A korpusz szociolingvisztikai vizsgálatokra, területi- és társadalmi, illetve rejtett változó mentén való vizsgálatokra is alkalmas, amennyiben mind a jelnyelvi szöveg, mint az interjú metaadatai rendelkezésre állnak a vizsgálathoz. Az előállt adatbázis korpuszalapú, valódi nyelvhasználatból származtatott grammatika készítését teszi lehetővé, hiszen alapot jelenthet a jövőben az összes nyelvi szint vizsgálatára a fonológiától a pragmatikáig.

A korpusz alapvető funkciója a magyar jelnyelv archiválása, mivel egyedülálló értékként bír a kortárs magyar jelnyelvhasználatot tekintve, a jövőben pedig történeti anyagként szolgál, így a későbbiekben lehetséges lesz a magyar jelnyelv különböző szintjein történő változások vizsgálata. Az ELAN-ban annotált korpuszban futtathatóak az egyes jelformákra való önálló keresések, mivel az annotáció során létrehozott címkét összekapcsolja a videó megfelelő szegmensével.

A korpusz kiváló tanítási anyagként is felhasználható, pl. a siketek oktatásakor, számtalan formában (segéd-, és példaanyagok, siket kultúra tantárgy, nyelvtan, stb.). A hallók jelnyelvoktatásának fejlesztésében is kulcsszerepe van a jelnyelvhez való hozzáférés kérdésének, a jelnyelvet hallóként, idegen nyelvként tanulók – a kezdő szinttől a tolmácsszintig, az egyszeri érdeklődőtől a siket gyermekek halló szüleiig – nagy hasznát vehetik a korpusz anyagának.

A korpusz kiindulási alapja lehet az automatikus jelfelismerő rendszereknek, illetve a számítógépes jelnyelvi modellezésnek, mivel természetes változatossággal rendelkező nyelvi anyagról van szó.

Felsőoktatásban a nyelvészeti terepmunka és a korpusznyelvészet órákon különösen, de antropológiai és minden egyéb, terepfelvételeket használó tudományterületnek kiváló példát szolgáltat a Jel*E*sély projekt jelnyelvi korpusza az adatkezelésre, adatfeldolgozásra. A korpusz nyelvi adatait az ELAN szoftverben megtekintve lehetővé válik, hogy egy időben 4 kamerakép vizsgálatával a legaprólékosabban megfigyeléseket tegyük a magyar jelnyelv jelenségeivel kapcsolatban. Kiváló konvertálási adottságok jellemzik a korpuszt (példamondatok exportálhatóak a jelnyelv tanításhoz, valamint a nagy sebességű videó feliratozás is lehetségessé vált). A korpuszból emellett számos alapvető statisztika kinyerhető.

3. Résztvevők

A projekt során számos terület szakemberei (szociolingvisták, elméleti nyelvészek, pszichológusok, szociológus, jogász stb.) dolgoztak együtt. A terepmunkák során kizárólag siket terepmunkásokkal dolgoztunk, az annotálás/fordítás/lejegyzés folyamataiban pedig siket, nagyothalló, CODA és halló munkatársak dolgoztak együtt, összesen 35-en. A részfeladatok összehangolásához precíz és részletes dokumentálásra volt szükség, valamint a jelnyelvi tolmácsokkal való állandó együttműködésre.

4. További tervek és feladatok

A korpuszhoz kötődő munkálatok során a jövőben is további számos kihívással kell szembenéznünk. Ilyen például a széleskörű annotálás és az ID-Glossz adatbázis kidolgozása, valamint a nemzetközi jelnyelvi korpuszokkal való átjárhatóság megteremtése.

A következő fontos lépés az ID-Glossz adatbázis kidolgozása, amely nemcsak biztosítja a könnyű keresést, de a későbbiekben a szótárépítéshez is nélkülözhetetlen. Ennek a folyamatnak a szótárkészítési vonatkozása a lemmatizáció, amelynek a jelnyelvekre alkalmazható nemzetközi standardjai még nem adottak. Álláspontunk szerint a lemmatizáció elveinek kidolgozása is csak a siket közösség bevonásával lehetséges.

Ahogy már említettük, fontos további feladat a szociolingvisztikai korpuszanyagok fordításainak ELAN-ba való átemelése és azoknak a videók megfelelő szegmenseivel való összekapcsolása.

A módszertanában nemzetközileg is úttörőnek számító *JelEsély* projekt a magyar jelnyelv átfogó, korpuszalapú grammatikai leírásával, korpuszával és szótárával e kutatások nélkülözhetetlen kiindulását jelentik a minőségi kétnyelvű oktatás elméleti, módszertani és gyakorlati feltételrendszere meghatározásának és az új oktatási program kimunkálásának.

Köszönetnyilvánítás

A tanulmányban leírtak nem valósulhattak volna meg a Jelesély Projekt (Támop 5.4.6/B-13/1-2013-0001) támogatása nélkül. Köszönetet mondunk a Jel*E*sély projekt megvalósítóinak, valamennyi siket és halló munkatársnak, különösen a technológiai előkészítésében és archiválásában résztvevő Tarr Zoltánnak és Gál Ferencnek, valamint a kontrollált szótárak elemeinek kidolgozásában nyújtott támogatásáért Szabó Mária Helgának.

Hivatkozások

1. Bartha, Cs: A kétnyelvűség alapkérdései. Nemzeti Tankönyvkiadó, Budapest (1999)
2. Bartha, Cs., Hattyár, H.: Szegregáció, diszkrimináció vagy társadalmi integráció? – A magyarországi siketek nyelvi jogai. In: Kontra, M., Hattyár, H. (eds.): Magyarok és nyelvtörvények. Teleki László Alapítvány, Budapest (2002) 73–123
3. Bartha, Cs.: A Kárpát-medencei kisebbségi magyar nyelvi korpusz. Korpuszépítési és kutatási lehetőségek. Kézirat. MTA Nyelvtudományi Intézet, Budapest (2002)
4. Bartha, Cs.: Siket közösség, kétnyelvűség és a siket gyermekek kétnyelvű oktatásának lehetőségei. In: Ladányi, M., Dér, Cs., Hattyár, H. (eds.): „…még onnét is eljutni túlra…". Nyelvészeti és irodalmi tanulmányok Horváth Katalin tiszteletére. Tinta Könyvkiadó, Budapest (2004) 313–332
5. Bartha, Cs., Hattyár, H., Szabó, M. H.: A magyarországi siketek közössége és a magyarországi jelnyelv. In: Kiefer, F. (ed.): Magyar Nyelv. Akadémiai Kiadó, Budapest (2006) 852–906

6. Battison, R.: Analysing Signs. In: Valli, C., Lucas, C.: (eds.) Linguistics of American Sign Language: An Introduction. Gallaudet University Press, Washington (2000) 199–218
7. Bea – Magyar Spontán Beszéd Adatbázis http://www.nytud.hu/adatb/bea/index.html (é.n.)
8. Cormier, K., Fenlon, J., Rentelis, R., Schembri, A.: British Sign Language Corpus Project: A corpus of digital video data of British Sign Language 2008–2011. University College London, London (2011)
9. Cormier, K., Fenlon, J., Gulamani, S., Smith, S.: BSL Corpus Annotation Conventions (2015) http://www.bslcorpusproject.org/wp-content/uploads/BSLCorpus_AnnotationConventions_v2_-Feb2015.pdf
10. Crasborn, O., Sloetjes, H., Auer, E., Wittenburg, P.: Combining video and numeric data in the analysis of sign languages within the ELAN annotation software In: Vettori, C. (ed): LREC 2006, II. Workshop proceedings. Representation and processing of sign languages. ELRA, Paris (2006) 82–87
11. Crasborn, O., Sloetjes, H.: Enhanced ELAN Functionality for sign langauge corpora In. Crasborn, O., Hanke, T., Efthimiou, E., Thoutenhoofd, E. D., Zwitzserlood, I.: Proceedings of LREC 2008, Sixth International Conference on Languge Resources and Evaluation (2008) 39–43
12. Crasborn, O., de Meijer, A.: From corpus to lexicon: the creation of ID-glosses for the Corpus NGT In: Crasborn, Efthimiou, Fontinea, Hanke, Kristoffersen, Mesch (eds.): Proceedings of the 5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon (2012) 13–17
13. Crasborn, O.: „Sign Language Corpora." Sign Language Corpora Wiki. Online: http://sign.let.ru.nl/groups/slcwikigroup/wiki/7f8aa/sign_language_corpora.html (2013) (2014. 03. 08)
14. Crasborn, O., Bank, R., Zwitserlood, I., van der Kooij, E., de Meijer, A., Sáfár, A.: Annotation Conventions for the Corpus NGT. (2015) http://www.bslcorpusproject.org/wp-content/uploads/CorpusNGT_AnnotationConventions_v3_Feb2015.pdf
15. Crasborn, O., Zwitserlood, I., Ros, J.: Corpus NGT. An Open Access Digital Corpus of Movies with Annotations of Sign Language of the Netherlands. Centre for Language Studies, Radboud University Nijmegen. [Available at: http://www.ru.nl/corpusngt] (én) (2015.12.03)
16. DGS-Korpus. Online: http://www.sign-lang.uni-hamburg.de/dgs-korpus/index.php/dgs-korpus.html (é.n.) (2014. 03. 09)
17. ELP, Endangered Languages Project: Corpus of grammar and discourse strategies of deaf native users of Auslan (Australian Sign Language). http://www.hrelp.org/grants/projects/index.php?lang=9 (é.n.)
18. Hanke, T.: HamNoSys – Representing Sign Language Data in Language Resources and Language Processing Contexts In: Streiter, O., Vettori, C. (eds): LREC 2004, Workshop proceedings. Representation and processing of sign languages. ELRA, Paris (2004) 1–6
19. Hattyár, H.:. A siketoktatás elméleti és gyakorlati kérdései. Educatio 9. (2000) 776–790
20. Hattyár, H.: Jelnyelvek – Természetes emberi nyelvek eltérő modalitással. In: Ladányi, M., Dér, Cs., Hattyár, H. (eds.): „…még onnét is eljutni túlra…". Nyelvészeti és irodalmi tanulmányok Horváth Katalin tiszteletére. Tinta Könyvkiadó, Budapest (2004) 342–346
21. Hattyár, H.: A magyarországi siketek nyelvelsajátításának és nyelvhasználatának szociolingvisztikai vizsgálata. Doktori Disszertáció ELTE BTK, Budapest (2008)
22. Johnston, T.: The lexical database of Auslan (Australian Sign Language). Sign Language & Linguistics (2001) 145–169
23. Johnston, T., Schembri, A.: Australian Sign Language (Auslan): An introduction to sign language linguistics. Cambridge University Press, Cambridge (2007)

24. Johnston, T.: The Auslan Archive and Corpus. In D. Nathan (ed.): The Endangered Languages Archive—http://elar.soas.ac.uk/languages. Hans Rausing Endangered Languages Documentation Project, School of Oriental and African Studies, University of London, London (2008)
25. Johnston, T.: From archive to corpus: transcription and annotation in the creation of signed language corpora. In: Roxas, R. (ed.): 22nd Pacific Asia Conference on Language, Information, and Computation. De La Salle University, Cebu, Philippines (2008) 16–29
26. Johnston, T.: Auslan Corpus Annotation Guidelines. Centre for Language Sciences, Department of Linguistics, Macquarie University, Sydney, Australia (2014)
27. Kárpád-medencei Magyar Nyelvi Korpusz: http://corpus.nytud.hu/mnszworkshop/index.html (2006)
28. Kendall, T. On the History and Future of Sociolinguistic Data. In: Language and Linguistics Compass (2008) 332–351
29. Konrad, R.: Sign Language Corpora Survey http://www.sign-lang.uni-hamburg.de/dgs-korpus/files/inhalt_pdf/SL-Corpora-Survey_update_2012.pdf (2012)
30. Kontra, M.: A Budapesti Szociolingvisztikai Interjú. MTA Nyelvtudományi Intézet, Élőnyelvi Kutatócsoport. Kézirat. Budapest http://buszi.nytud.hu/ (1987)
31. Lancz, E., Barbeco, S.: A magyar jelnyelv szótára. Siketek és Nagyothallók Országos Szövetsége, Budapest (1999)
32. Leech, G.: „Corpora and theories of linguistic performance." In: Svartvik, J. (ed.): Directions in Corpus Linguistics: Proceedings of Nobel Symposium 82, Stockholm, 4-8 August 1991. Berlin: Mouton de Gruyter. (1992) 105–122
33. McEnery, T., Wilson, A.: Corpus Linguistics. Lancaster University, Lancaster (2001)
34. McEnery, T., Sebba, M., Burnard, L.: Minority Language Engineering (MILLE) – Summary Report (é.n.)
35. Nyelv- és beszédtechnológiai platform (sz.n.) http://www.hlt-platform.hu/online-adatbazisok.html
36. Oravecz, Cs., Váradi, T., Sass, B.: The Hungarian Gigaword Corpus. In: Proceedings of LREC 2014. http://clara.nytud.hu/mnsz2-dev/ (2014)
37. Pfau, R., Steinbach, M., Woll, B. (eds.), Sign language. An international handbook (HSK - Handbooks of linguistics and communication science). Mouton De Gruyter, Berlin (2012)
38. Rundell, M.: The corpus of the future, and the future of the corpus. Talk at 'New Trends in Reference Science' (1996)
39. SignWriting History. SignWriting® Site.: www.signwriting.org/library/history/history.html (é.n.) (2014.3.10.)
40. Sinclair, J.: EAGLES. Preliminary recommendations on Corpus Typology. (1996) http://www.ilc.cnr.it/EAGLES/corpustyp/corpustyp.html
41. Stokoe, W. Sign Language Structure: An Outline of Visual Communication Systems of the American Deaf. Studies in Linguistics: Occasional Paper No. 8. University of Buffalo. Buffalo, NY (1960)
42. Szabó, M. H.: A magyar jelnyelv szublexikális szintjének leírása. Akadémiai Kiadó, Budapest (2007)
43. Szabó, M. H., Mongyi, P.: A jelnyelv nyelvészeti megközelítései. Magyar Jelnyelvi Programiroda, Budapest (2005)
44. Wallin, L., Mesch, J., Nilsson., A-L.: Transcription guide lines for Swedish Sign Language discourse. https://www.diva-portal.org/smash/get/diva2:389066/FULLTEXT01.pdf (2010)