A BEA ADATBÁZIS ALKALMAZÁSFÜGGŐ LEJEGYZÉSEI

Gyarmathy Dorottya – Neuberger Tilda

Bevezetés

A beszédadatbázisok elemzése és felhasználása a nyelvészetben mintegy két évtizedes múltra tekint vissza. A korpuszok nagyságának, illetve feldolgozásának szempontjából nagy jelentőséggel bírt a számítógépek megjelenése, mely egyben a korpusznyelvészet kialakulását is jelentette (Leech 1992; Váradi 2000). Az egy adott beszédközösség nyelvhasználatát megfelelően reprezentáló adatmennyiség gyűjtése kizárólag a számítógépes technológia segítségével érhető el. Mai értelemben a számítógép segítségével létrehozott, tárolt, a szükséges magyarázó jegyzetekkel, címkézésekkel és átírásokkal ellátott, meghatározott szempontok szerint összeválogatott és egységesen kódolt beszédfelvételek gyűjteményét nevezzük adatbázisnak (Vicsi 2001).

Az utóbbi évtizedekben a korpusznyelvészet sok nyelvben indult fejlődésnek, így a nemzetközi szakirodalomban számos, a legkülönfélébb céllal létrehozott adatbázist találhatunk. A beszédkorpuszoknak különösen azokon a területeken van nagy jelentősége, ahol az adott nyelvi, nyelv-használati jelenség nem vizsgálható, avagy a kitűzött cél nem érhető el megfelelő mennyiségű, körültekintően gyűjtött adathalmaz nélkül. A fonetika, a beszédtechnológia és a pszicholingvisztika számos ilyen kutatási területtel rendelkezik. A beszéd artikulációs és akusztikai sajátosságainak vizsgálata adatbázis híján csaknem elképzelhetetlen. Ilyen célból jött létre például a laboratóriumi beszédet tartalmazó, lengyel izolált szavakból álló korpusz vagy az UPSID adatbázis (UCLA Phonological Segment Inventory Database). Ez utóbbi jelenleg 451 nyelv adatait tartalmazza. Készítői a világ nyelveiben tapasztalható bizonyos fonológiai univerzálék és egyetemes tendenciák kimutatását tűzték ki célul (Gósy 2004).

A beszédtechnológiában ugyancsak különböző tartalmú és méretű adatbázisokra épülnek a beszédszintetizáló rendszerek. 1981-ben beszédfelismerés céljából hozták létre az első azonos idejű személyi számítógépen működő, nagy, izolált szavas adatbázist, melynek segítségével a diktáló rendszert kívánták kifejleszteni. A gépi beszédfelismerés megoldásához minél változatosabb beszédmintát tartalmazó korpuszok létrehozására volt szükség (Gósy 2004). Amerikában az e célra létrehozott legjelentősebb adatbázisok a személyfüggetlen beszédfelismerők betanítására használatos TIMIT és a repülőtéri információval kapcsolatos szótárkészleten alapuló ATIS. Európa legtöbb

nyelvét átfogó adatbázisok az EUROM0, EUROM1 és a BABEL, melyeknek célja a beszédakusztikával, fonetikával, digitális jelfeldolgozással, illetve nyelvészettel foglalkozó szakemberek munkájának segítése (Vicsi 2001). Az anyanyelv-elsajátítás univerzális jelenségeinek leírására, illetve a szókincs sajátosságainak elemzésére nyújt lehetőséget a különböző anyanyelvű gyermekek beszédét rögzítő gyermeknyelvi adatbázis (CHILDES – Child Language Data Exchange System: http://childes.psy.cmu.edu/).

A spontán beszédet rögzítő korpuszok, melyek jól használhatók különböző nyelvészeti kutatásokra, különféle módokon keletkeztek. A London–Lund-korpusz 50 angol és svéd nyelvű célzottan felvett dialógust tartalmaz, Clark és Fox Tree (2002) telefonautomatának mondott szövegeket vettek fel magnetofonra. A Hutchinson–Pereira-korpusz (2001) létrehozásához egy ausztrál pizzatársaság telefonos rendeléseit rögzítették egy éven át. Ugyancsak telefonon (mobil- és vezetékes) keresztül rögzített beszédet tartalmaznak a SPEECHDAT 1, 2 és a SPEECHDAT-E adatbázisok (Vicsi 2001).

A magyar beszédtudományos kutatások hosszú időn keresztül főleg felolvasott vagy előre betanult szövegek vizsgálatán alapultak. Nagyméretű, reprezentatív írott anyag alapján készült szöveges és lexikai adatbázis a Magyar Nemzeti Szövegtár (http://corpus.nytud.hu/mnsz/). Különféle beszédadatbázisok keletkeztek a mesterséges beszéd előállításához, melyek mind az aktuális szintézis szükségleteit tükrözik. Léteznek diádos adatbázisok (pl.: Profivox szövegfelolvasó szoftver), kötött szótáras rendszerekhez készített elemtárak (pl.: hangposta, pályaudvari utastájékoztató), továbbá kevert felépítésű beszédadatbázisok (a Profivox legújabb változata, vö. Olaszy 1999). A magyar beszédtechnológiai kutatások és fejlesztések támogatására készült a vezetékes és mobiltelefonról rögzített, 500 adatközlő által felolvasott szövegeket tartalmazó MTBA magyar telefonbeszéd-adatbázis (Vicsi et al. 2002).

Abban az esetben, ha a kutató a nyelvet mint a társadalmi kommunikáció eszközét, illetőleg a nyelvhasználatot szeretné vizsgálni, adatait a mindennapi kommunikációból kell vennie (Labov 1981). Erre a célra a legmegfelelőbb eszköz a spontán beszéd vizsgálata (pl. Kontra 1988). A magyar spontán beszéd vizsgálata a múlt század negyvenes éveiben indult meg Hegedűs Lajos fonetikus kezdeményezésére. Az ország különböző megyéiben rögzítettek spontán beszédet abból a célból, hogy az így készült nyelvjárási hangfelvételeket hozzáférhetővé tegyék az utókor számára. A felvételeken népszokások, babonák, mesék, ünnepi szokások, a kenyér- és süteménysütés módjai, a disznóölés leírása, élettörténetek, mondókák és énekek hallhatók (Nikléczy–Horváth 2007). A hetvenes évek elején Szende Tamás (1973) a spontán beszéd gyakorisági tényezőinek elemzéséhez négyféle korpuszt használt fel. Három felvétel különböző témájú társalgásokat rögzít, például: tudósok magánbeszélgetését a számítógépek társadalomtudományi alkalmazhatóságáról, két tanár és egy diák beszélgetését az iskolai életről, illetőleg egy a Hamlet-monológ háromféle interpretációjáról szóló négytagú társalgást. A negyedik

felvételen egy Füst Milánnal készült beszélgetés (interjú) hallható, amelyet Fónagy Iván készített.

Az 1975-ben az ELTE Mai Magyar Nyelvi Tanszékén megalakult beszélt nyelvi kutatócsoport munkájának köszönhetően nagy mennyiségű, spontánbeszéd-felvétel és azok lejegyzése áll a kutatók rendelkezésére. A Beszélt nyelvi gyűjtemény (Keszler 1983) hat kötete egyaránt tartalmaz regionális köznyelvet rögzítő néprajzi témájú interjúkat, rádió- illetve televíziós riportokat, továbbá rejtett mikrofonnal készült beszélgetéseket (a rögzítés tényét a felvétel után hozták az adatközlők tudomására). Az utóbbiak felhasználásával végezte Keszler Borbála a spontán beszéd szófaji gyakoriságát, illetve mondatgrammatikai aspektusait elemző kutatását (Keszler 1983). A beszélt nyelv jellegzetességeit, mondattani sajátosságait (a beékelődést, a kötőszavak halmozását vagy hiányát, az ismétléseket) Huszár Ágnes a médiából származó felvételeken elemezte (Huszár 1985).

A Magyar Tudományos Akadémia Nyelvtudományi Intézetének Élőnyelvi Osztályán 1987–89 között elkészült Budapesti Szociolingvisztikai Interjú (BUSZI adatbázis) 250 adatközlője a budapesti lakosság szociológiailag reprezentatív mintáját adja (Váradi 2003). A felvételek egyaránt tartalmaznak spontán és nem spontán beszédet. A BUSZI előmunkálataiként 1985-ben elkészült a gazdagréti televízió már sugárzott adásaiból válogatott felvételek több szempontú elemzése. A felvételek intonációs átiratát Varga László készítette, melynek felhasználásával a kutatók elemezték a beszéd logikai struktúráját, mondattani szerkezetét, a témaismétlő névmásokat, a spontán beszéd és az írott nyelv különbségét, továbbá a nonverbális kommunikáció, azaz a gesztusnyelv eszközeit (Kontra 1988). 1998-ban keletkezett az első nemzetközi szabvány alapján készült, felolvasásokat rögzítő adatbázis, a BABEL, melynek célja a magyar hivatalos köznyelv hanganyaggal való reprezentálása (Vicsi–Víg 1998).

A felsorolt adatbázisok az utóbbi évtizedekben látványos előrelépési lehetőségeket alapoztak meg a spontánbeszéd-vizsgálatokban, a szintetizált beszéd előállításában, valamint az automatikus, mesterséges beszédfelismerésben.

A BEA spontánbeszéd-adatbázis

A beszédkutatás új feladatai, illetőleg a spontán beszéd fonetikai elemzésének igénye szükségessé tette egy a modern korpuszépítés szabályainak megfelelő, a minőségi hangrögzítés minden kritériumát teljesítő, nagy mennyiségű spontán beszédet tartalmazó hangtár létrehozását, amely egyaránt megfelel mind a fonetikai, az alkalmazott fonetikai, illetve a pszicholingvisztikai kutatások kritériumrendszerének. Az MTA Nyelvtudományi Intézetének Fonetikai Osztályán 2008-ban kezdődött meg a BEA spontánbeszéd-adatbázis feltöltése, ami jelenleg is tart. A korpusz elsődleges célja többféle típusú spontán beszéd rögzítése, de a fonetikai célok kielégítése (összehasonlíthatóság) érdekében mondat- és szövegfelolvasásokat, illetve mondatismétléseket

is tartalmaz (Gósy 2008). A felvételek minden esetben azonos körülmények között készülnek, csendesített helyiségben. A korpusz legfőbb előnyei közé tartozik az a tény, hogy a kutatók számára időt takarít meg azáltal, hogy nem nekik kell felkeresniük az egyes adatközlőket és elkészíteni a felvételeket. Az azonos, stúdiókörülmények között való rögzítés pedig lehetővé teszi a hanganyagok fonetikai és alkalmazott fonetikai felhasználását, illetve az összehasonlításokat. Az adatbázis már elkészült része is hatalmas adathalmazt biztosít (minden érdeklődő kutató számára) a különféle célú elemzésekhez. Ezek a felvételek napjainkban is számos kutatás alapjául szolgálnak (pl. Beke 2008; Bata 2009; Bata–Gráczi 2009; Gósy 2009; Gráczi 2009; Gyarmathy–Gósy–Horváth 2009, Markó 2009, Bóna 2010; Horváth 2010; Beke–Gyarmathy 2010 stb.), és a jövőben is lehetőséget biztosítanak a beszéd fonetikai, pszicholingvisztikai, szövegtani, pragmatikai stb. szempontú elemzéseire.

A BEA eredeti lejegyzése

A különböző beszédkorpuszok rendszerint nem csupán a hangzó anyagot, de annak írott formáját is tartalmazzák. A felvételek átirata a felhasználási területtől függően lehet helyesíráson alapuló lejegyzés, fonetikai transzkripció, tartalmazhatja az intonáció és egyéb szupraszegmentumok jelölését. A BEA hangfelvételeinek lejegyzése a kezdetekben egy elsődleges írásos tükröztetés volt. A lejegyzők a Microsoft Office Word programjában .doc formátumban, helyesírásban, központozás nélkül írták le a hanganyagokat, a későbbi feldolgozás szempontjából fontosnak ítélt adatok, mint például a megakadásjelenségek, illetve a fiziológiai hangadások jelölésével. Az alapvetően helyesírás szerint történő lejegyzés nem jelölte a kiejtés és a helyesírás eltéréseit, például a *zöldség* szó esetében nem érvényesítette az összeolvadás szabályát (tehát nem *zölcség*-ként lett lejegyezve). Az átírás a megakadásjelenségeket, tehát a hibás alakot vastagon szedve jelölte, majd ha a közlés nem tartalmazta a javítást, a []-ben megadta a helyes szóalakot, például: ***berép*** *[belép] a diri.* A vastagított szedés egyaránt vonatkozott a hiba típusú és a bizonytalansági megakadásjelenségekre. Az újraindítások és a téves kezdések esetén csak a szótöredék volt vastagon szedve (pl.: ***mege-*** *megettem*; ***mege*** *[megettem] mehettem*), a téves szótalálásoknál maga a téves találat (pl.: ***zárd be*** *csukd be az ajtót*). A nyújtásokat a félkövér szedés mellett az adott hangot jelölő betű megkettőzése jelölte (pl.: *mikoor*), míg a beszélő hezitálását a kitöltött szünetként realizálódott beszédhang megtriplázásával és félkövér szedéssel (pl. ***öööö***) tükröztették. A beszéd folyamatosságát megszakító néma szüneteket a lejegyzésekben □-ek jelezték. Az eredeti lejegyzési útmutató kitért továbbá a köznyelvben használatos, de nem szótári alakban előforduló szavak (pl. *aszongya*, *asszem*), az idegen szavak, rövidítések, betűszók és mozaikszók, illetőleg a lejegyző számára értelmezhetetlen szóalakok lejegyzésére (vö. Gósy 2008).

Ezek a lejegyzések durva átiratok voltak, amelyek a kutatók munkáját voltak hivatottak megkönnyíteni, továbbá lehetővé tették a további finomabb fo-

netikai meghatározásokat, illetőleg a saját szempontú és célú átírást. Az adatbázis későbbi, automatikus, gépi beszédfelismerésben való felhasználásának igénye azonban szükségessé tette az eredeti lejegyzési elvek átdolgozását, továbbá a lejegyzés szoftveres hátterének megváltoztatását. A BEA adatbázis hanganyagainak lejegyzése 2010 októbere óta a Transcriber szoftver segítségével történik, továbbá folyamatban van az eredeti, Wordben lejegyzett anyagok Transcriberbe való átalakítása is. Tekintettel arra, hogy az eredeti lejegyzés továbbra is több szempontból felhasználható és szükséges, ezért a transcriberes lejegyzések .doc fájlos megjelenítését is tervezik. Így mindkét típusú lejegyzés a kutató rendelkezésére áll.

A Transcriber program ismertetése

A Transcriber program a beszéd szegmentálására, címkézésére és leírására ad lehetőséget. Segítségével a hanganyagot és az írott szöveget egyszerre láthatóvá és hallhatóvá tudjuk tenni. A szoftver ingyenesen letölthető az internetről (http://trans.sourceforge.net/en/presentation.php), felhasználóbarát grafikus felülettel rendelkezik, különféle operációs rendszerben (Windows, Unix) futtatható és többféle audiofájl-típust (.au, .wav, .snd) támogat. Folyamatos fejlesztés alatt áll, de a szabad hozzáférésnek köszönhetően a felhasználók újabb funkciókat adhatnak hozzá attól függően, hogy elsődlegesen mire kívánják használni (Barras et al. 1998).

A hanganyagok lejegyzése Transcriber programban

A Transcriber ugyan nem alkalmas fonetikai mérésekre, mint például a Praat (vö. Markó–Bóna 2006), viszont a beszédfelismeréshez történő felhasználáshoz a legmegfelelőbb eszköz. Alkalmas a néma szünetek, hezitálások, hümmögések és az egyéb nem beszéd jellegű hangadások (pl. köhögés, nevetés, egyéb zajok) címkékkel való automatikus jelölésére. Mivel a BEA adatbázis több típusú spontán beszédet (narratíva, véleménykifejtés, interpretált beszéd, társalgás), felolvasást, mondatvisszamondást tartalmaz (vö. Gósy 2008), a szövegek jellemzői műfajonként is vizsgálhatók.

A hosszú időtartamú (átlagosan 50 perces) hangfelvételek esetében a későbbi felhasználást megkönnyíti az átirat mellett a szöveg szegmensekre (időszeletekre) bontása és felcímkézése. A korábbi lejegyzési stratégia szerint egy durva átiratot kaptunk, amelyhez elengedhetetlen volt a hangzó anyag vizsgálata (vö. Neuberger 2009). Az egységes lejegyzési útmutató ellenére a különféle lejegyzők egyéni módon (különböző programok használatával, eltérő részletességgel, pontossággal) folytatták a munkát. Előfordult, hogy a speciális jelenségeket más-más módon észlelték és értelmezték, így az átiratoknak sok szubjektív megítélésű pontja volt. Az egységesítés érdekében is elengedhetetlenül fontos, hogy a továbbiakban egy meghatározott programban, az előzőhöz képest még kidolgozottabb stratégiák követésével ké-

születnek az átiratok, noha ebben az esetben is egyetlen lejegyző észlelési feldolgozásának eredménye jelenik meg a szoftverben.

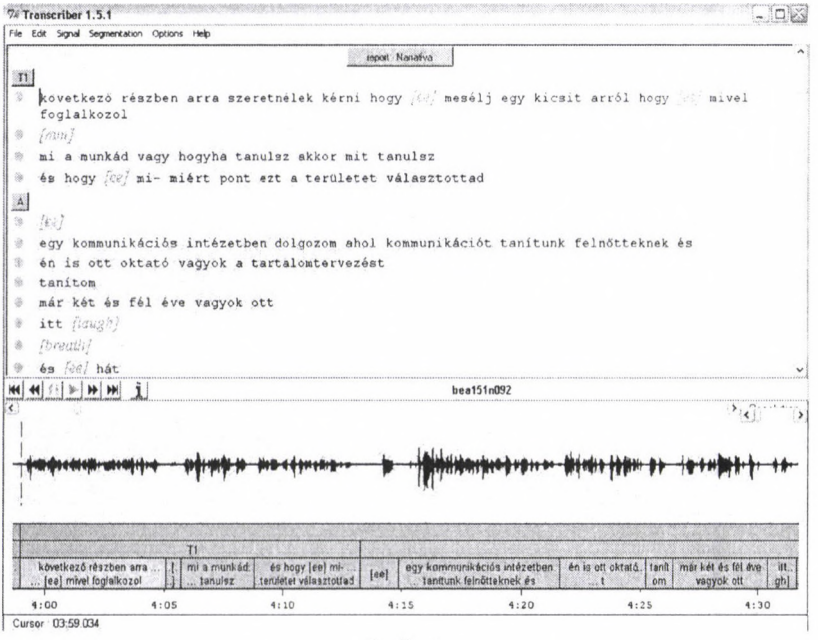
A BEA-hanganyagok transcriberes átírása bizonyos pontokon megkönnyíti a lejegyzők dolgát, más tekintetben azonban több odafigyelést igényel tőlük. Könnyebbséget jelent a korábbi, Wordben elkészített lejegyzéshez képest, hogy a hang hullámformája, valamint a szöveges rész közös felületen, egy ablakban látható és kezelhető, így nem kell váltogatni a hanglejátszó program és a Word között, valamint a hang elindítása, leállítása, újrajátszása egy billentyű segítségével végrehajtható. A fő eltérés a korábbi lejegyzési stratégiához képest az, hogy a hanganyag szegmensekre bontva kerül lejegyzésre. A Transcriber program lehetővé teszi, hogy a beszéd időszeletekre bontásával a hang és a szöveg szinkronba kerüljön, így kisebb részeket kell egyszerre feldolgozni, ez pedig megkönnyíti az ellenőrzést, a visszakeresést, valamint segíti a későbbi felhasználást is.

A korábbi BEA-lejegyzés nem tartalmazott időszeletekre bontást, az átírt beszéd egy hosszú szöveges fájlként jelent meg, csupán a felvétel egyes részeinek (mondatismétlés, narratíva stb.) időintervallumát kellett megadni a pontos perc- és másodpercérték feltüntetésével. Az egyes felvételrészeket az új átírás ún. témákkal (topicokkal) jelöli. A Transcriber lehetővé teszi minden egyes szegmenshez a külön téma hozzárendelését, de a BEA-lejegyzésben elegendő a téma kezdetekor, az első megszólaló szegmenséhez hozzárendelni az adott témát. A mondatvisszamondást és a felolvasást tartalmazó részhez felhasznált mondatokat, illetve szöveget a lejegyzők kézhez kapják egy Word dokumentumban, így – ahogyan a korábbi átiratoknál is – ezeket a részeket egyszerűen bemásolhatják. Ezután jelölni kell a beszélők esetleges hibázásait, félreolvasásait (+ jellel, lásd később).

A következőkben áttekintjük a szoftver használatának, illetve a lejegyzés menetének legfontosabb tényezőit. A program angol nyelvű, így a vezérlő felület és az automatikus címkék ezen a nyelven olvashatóak. Használatát nagyban segíti, hogy felhasználóbarát grafikus felülettel rendelkezik (1. ábra). A hangfájl megnyitását követően a kezelőfelület alsó részén megjelenik a hang rezgésképe, alatta az egyszintű címkézés helye (itt láthatjuk függőleges vonalakként a szegmenshatárokat), fölötte pedig – a képernyő nagy részén – a beírt szövegek helyét találjuk (a beszélők és a témák megjelölésével).

A felvétel a TAB billentyűvel elindítható vagy leállítható az adott pozícióban, amelyet függőleges szaggatott vonal jelöl a hullámformában. Az ENTER lenyomásával húzhatunk szegmenshatárt, ekkor a szöveges részben új sor kezdődik. A lejegyzés általában úgy zajlik, hogy 1. a TAB billentyű megnyomásával elindítjuk a felvételt, és 2. meghallgatunk egy részletet, 3. a TAB újbóli megnyomásával leállítjuk a felvételt, ezt követi 4. az elhangzott szöveg leírása, és végül 5. az ENTER billentyű megnyomásával lezárjuk a szegmenst, és egyben újat is kezdünk. A szegmentálás során törekszünk arra, hogy lehetőleg maximum 5 másodperces részekre bontsuk a beszédet.

A szegmenshatárokat a beszéd közti szünetek közepére kell elhelyezni. Különleges események esetén, mint pl. zaj, nem érthető szó, megakadás, egyszerre beszélés stb. a lehető legszűkebb szegmenst kell megtalálni az adott jelenséghez, hogy azokat jól el lehessen különíteni a „hasznos", vagyis a beszédet tartalmazó részektől.



1. ábra
A Transcriber felhasználói felülete

A BEA adatbázis lejegyzése Transcriberben (átdolgozott lejegyzési útmutató)

Az új, Transcriberes lejegyzés több ponton támaszkodik a korábbi BEA-lejegyzésre, de az új lehetőségeknek és céloknak megfelelően történt néhány változtatás. A következőkben sorra vesszük a kétféle lejegyzési stratégia megegyező pontjait, illetve kitérünk a különbségekre is.

Az átiratok mindkét lejegyzési módban a magyar helyesírás szabályai szerint készülnek, a kiejtésben megvalósuló koartikulációs szabályokat nem tükrözik. Sem a Wordben, sem a Transcriber programban történt lejegyzés nem tartalmaz központozást. A régi útmutató szerint az egyetlen használható és használandó írásjel a felkiáltójel (!), amely a különböző, nem a beszédhez tartozó, de a beszélő által produkált hangok (köhögés, torokköszörülés,

nyelvcsettintés, ki- és belégzés, nevetés stb.) egységes jelölésére szolgált. Ezeket a jelenségeket az új lejegyzésben automatikus címkékkel, ún. *event*ekkel kell jelölni (1. táblázat). Az eventek szögletes zárójelek között jelennek meg a szövegben, előhívásuk billentyűkombinációkkal történik.

1. táblázat: Automatikus címkék (eventek) a Transcriberben 1.

Hangesemény	Billentyűkombináció	Event
Nevetés	< Bal_Alt-l >	[laugh]
Köhögés, torokköszörülés	< Bal_Alt-c >	[cough]
Tüsszentés, szipogás	< Bal_Alt-z >	[sneeze]
Belégzés, kilégzés	< Bal_Alt-b >	[breath]
Nyelvcsettintés, nyammogás	< Bal_Alt-p >	[lipsmack]
Külső zaj (ajtó, számítógép)	< Bal_Alt-n >	[noise]

Az új útmutató szerint a vessző (,) írásjelet kell használni az ismétlések jelölésére, például: *és, és, és*. Erre a jelölési módra azért volt szükség, mert a Transcriberben – a Word-del ellentétben – nem lehet félkövér (és semmilyen más) tipográfiát alkalmazni, de a megakadásjelenségeket valahogyan jelölni kell az írásban.

A megakadásjelenségeket mindkét átírás jelöli. A korábbi útmutató a megakadások (a téves szótalálások, a téves kezdések, az újraindítások stb.) jelölését vastagon szedve írja elő, a helyes szóalakot (amennyiben a beszélő maga nem javította) pedig utána []-ben kéri megadni. Az új lejegyzés szerint különféle stratégiák használatosak az egyes megakadások jelölésére. Mint említettük, az ismétléseket az első ejtés és az ismételt ejtés közötti vessző (,) jelzi. Az újraindításokat és a töredékszavakat kötőjellel kell jelölni ott, ahol a törés történt. Mivel a valamelyik oldalán space-szel határolt kötőjel csak a töredéket jelzi, a többszörös összetételeket (például: *cipő- és ruhavásár*) nem szabad így jelölni. Ezekben az esetekben a helyesírásnak ellentmondóan az első tag utáni kötőjelet el kell hagyni (pl.: *cipő és ruhavásár*). A hibásan ejtett szavaknál (nyelvbotlásoknál) + jel kerül a szó elé. Az új lejegyzés először a helyesírás szerinti alakot (vagyis amit a beszélő szándékozott mondani), utána pedig a hibásan ejtett alakot (például: *+átöltöző=áltöltöző*) tünteti fel. A hezitálások és a különböző kommunikációs célú hangesemények (pl. hümmögések) jelölésére a Transcriberben a már említett eventek szolgálnak (2. táblázat). A korábbi lejegyzésben nem mindegyiknek volt egységes jelölése, ezért a lejegyzők szubjektív módon írták át ezeket a jelenségeket. A nyílt hezitálást például az alábbi módokon: *öö, ööm, önm, eee, eem, öhm* stb. Az új lejegyzésben az alábbi hangeseményekre dolgoztak ki eventet: nyílt és zárt hezitálás, nyílt és zárt tagadás, illetőleg igenlés.

2. táblázat: Automatikus címkék (eventek) a Transcriberben 2.

Hangesemény	Billentyűkombináció	Event
Nyílt hezitálás	< Bal_Alt-e >	[ee]
Zárt hezitálás	< Bal_Alt-m >	[mm]
Nyílt tagadás	< Bal_Alt - >	[e-e]
Zárt tagadás	< Bal_Alt / >	[m-m]
Nyílt igenlés	< Bal_Alt + >	[ehe]
Zárt igenlés	< Bal_Alt * >	[mhm]

A néma szünetek korábbi jelölése a négyzet (□) volt, a Transcriberes lejegyzés a szegmentálás következtében a néma szüneteket nem jelöli speciális jellel, csupán a szegmenshatárokat kell behúzni azokra a helyekre, ahol szünet található, így a szegmens belsejébe kerül a két szünet közti szöveges rész. Az átlagosnál hosszabb jelkimaradásokat, például a 3–5 másodperces néma szüneteket a [sil] eventtel címkézik fel a lejegyzők (bár ezekre alig akad példa a felvételeken, hiszen ha a kísérletvezető észleli, hogy az adatközlő elakadt a mondanivalóban, segítő kérdésekkel ösztönzi őt további beszélésre). A szünet a szóban jelenséget a korábbi lejegyzés az adott szóban megfelelő helyen elhelyezett tapadó □ jelölte (például: *ki□mentem*). Ha szó belsejében kitöltött szünet fordult elő, akkor folyamatosan kellett írni a szóval (pl. *kiööömentünk*). Az új útmutató töredékeknek tekinti az ilyen eseteket, és a töredékszónak megfelelően a kötőjeles jelölést írja elő, mint például: *ki- -mentem, ki- [ee] -mentünk*.

A spontán beszédben megjelenő virtuális mondatokat egyik átírás sem jelöli. A lejegyzők a mondatkezdő nagybetűket tehát nem alkalmazzák, a tulajdonnevek és a betűszók írásmódjánál azonban fontos a nagybetűk használata. Eltérés a korábbi útmutatóhoz képest, hogy a Transcriberben a képzett tulajdonneveket is nagy kezdőbetűvel írják, a helyesírási követelménnyel ellentétben (például: *Győri, Petőfis*).

A számokat mindkét útmutató szerint szóként kell lejegyezni, például: *harmincnyolc, kétezer-tizenkettő*. Az új átírásban kivételt képeznek a tulajdonnevek és a betűszók részét képező számok.

Az idegen szavakat, betűszókat, rövidítéseket mindkét lejegyzés speciális esetnek tekinti. A korábbi lejegyzésben elsőként a kiejtett alakot írtuk le, majd szögletes zárójelben közöltük a helyesírási alakot, például: *Puccsínit [Puccinit], emtéából [MTA-ból]*. Az új átírás szerint abban az esetben, ha az elhangzott szóalak leírt alakjából a betűket a magyar nyelv logikája szerint egybeolvasva nem kapható vissza az elhangzott kiejtés, a lejegyzők a szóalakot speciális kiejtésűnek tekintik, és közvetlenül az írott alak elé illesztenek egy @ jelet. A toldalékot kiskötőjellel (-) kapcsolják a szóhoz. Például: *@Puccini-t, @MTA-t*. Ha a szó ortografikus alakjából nem lehet kikövetkeztetni a kiejtését, akkor a kiejtett alakot a szó utáni = jelet követően le kell je-

gyezni, tehát itt a helyesírási alak – kiejtés szerinti alak sorrendet kell követni. A toldalék ilyenkor a kiejtett alakhoz kapcsolódik kötőjellel. Például: *@IBM=íbéem-nél* vagy *@IBM=ájbíem-nél* (mindkét változatot használják).

Az együttbeszélés, közbevágás jelölése mindkét típusú lejegyzésben megoldott. Míg a korábbi átírásban az azonos időben hangzó szövegeket zárójelben ( ) jegyezték le egymás alatt beszélőnként, addig az új jelölésrendszerben dupla zárójelben ((|)) egymás mellé írják a beszélők szövegeit függőleges vonallal (|) jelölve a beszélőváltást.

Dupla zárójelet kell alkalmazni a nehezen vagy egyáltalán nem érthető részek leírására is, például: ((*dromek*)) vagy (( )). A korábbi átiratok ezeket csillaggal (**dromek**) vagy két csillag közötti kérdőjellel (*?*) jelölték.

Összegzés

Ahogy arra a bevezetésben is utaltunk, az elmúlt években a BEA adatbázis hangfelvételein alapuló számos fonetikai, pszicholingvisztikai és beszédtechnológiai tárgyú tanulmány született. A kutatási gyakorlat azonban azt mutatta, hogy ezekhez a korábbi durva átiratok csupán kiindulópontként szolgálhattak; minden esetben szükség volt a hanganyag további (akár hangelemző szoftverrel történő) finomabb vizsgálatára. Az új, a Transcriber programban történő lejegyzési mód – a szegmentálás, valamint a hang és szöveg szinkronba hozása miatt – az ellenőrzésben és a visszakeresésben jelentős könnyebbséget nyújt a kutatók számára. Ezen felül, a beszédtechnológiai célt figyelembe véve, a BEA számítógépes beszédfelismeréshez történő felhasználásához a Transcriber program a legmegfelelőbb eszköz. Az új lejegyzés jelölésrendszere is ezt a célt figyelembe véve lett kidolgozva. A Transcriberes lejegyzési stratégiák kialakításakor elsődleges szempont volt, hogy a lehető legnagyobb mértékben megfeleljen a korábbi lejegyzési útmutatónak, teljesíteni tudja az eredeti célokat, és a lejegyzők számára se jelentsen többlet terhet az újfajta átírási módszer elsajátítása. A program sajátosságaiból adódóan nem lehetett valamennyi korábbi jelölést átvinni az új útmutatóba, így szükség volt néhány változtatásra. Az átdolgozás során fontos szempont volt, hogy az új átírás megkönnyítse a kutatók munkáját, ugyanakkor alkalmassá tegye az adatbázist a műszaki (elsősorban a mesterséges beszédfelismerés) felhasználásra is. Az új lejegyzések azonban (hasonlóan a régiekhez) továbbra is egyetlen személy beszédészlelésén és beszédmegértésén alapulnak, így a kutatások során (például a fonetikai finomelemzéseknél) ezeket az átiratokat a kutatónak mindig ellenőriznie kell. A rendelkezésre álló kétféle lejegyzés ugyanakkor lehetőséget ad a mindenkori választásra. Az új lejegyzési módszer igyekszik megtartani a korábbi átírás egyszerűségét, ugyanakkor az egyes kutatásokban hatékonyabban használható, illetőleg több alkalmazási területet tesz lehetővé.

Irodalom

Barras, Claude – Geoffrois, Edouard – Wu, Zhibiao – Liberman, Mark 1998. Transcriber: A free tool for segmenting, labeling and transcribing speech. *First International Conference on Language Resources and Evaluation (LREC).* 1373–1376. http://xml.coverpages.org/Transcriber-LREC1998.pdf

Bata Sarolta 2009. Beszélőváltások a beszédpartnerek személyes kapcsolatának függvényében. *Beszédkutatás 2009.* 107–121.

Bata Sarolta – Gráczi Tekla Etelka 2009. Hatással van-e a beszédpartner életkora a beszélő beszédének szupraszegmentális jellegzetességeire. In Keszler Borbála – Tátrai Szilárd (szerk.): *Diskurzus a grammatikában, grammatika a diskurzusban.* Tinta Kiadó, Budapest, 74–83.

Beke András 2008. A felolvasás és a spontán beszéd alaphangszerkezeteinek vizsgálata. *Beszédkutatás 2008.* 93–108.

Beke András – Gyarmathy Dorottya 2010. Zöngétlen résmássalhangzók akusztikai szerkezete. *Beszédkutatás 2010.* 57–76.

Bóna Judit 2010. Bizonytalansági megakadások idősek és fiatalok spontán beszédében. *Beszédkutatás 2010.* 125–139.

Clark, Herbert H. – Fox Tree, Jean E. 2002. Using *uh* and *um* in spontaneous speaking. *Cognition* 84. 73–111.

Gósy Mária 2004. *Fonetika, a beszéd tudománya.* Osiris Kiadó, Budapest.

Gósy Mária 2008. Magyar spontánbeszéd-adatbázis – BEA. *Beszédkutatás 2008.* 194–207.

Gósy Mária 2009. Szóejtés és szóészlelés: változatosság és adaptálódás. *Beszédkutatás 2009.* 46–76.

Gráczi Tekla Etelka: Temporális jellemzők a beszédpartnerek ismeretségének függvényében. *Beszédkutatás 2009.* 121–134.

Gyarmathy Dorottya – Gósy Mária – Horváth Viktória 2009. A rejtett és a felszíni önmonitorozás temporális jellemzői. In Keszler Borbála – Tátrai Szilárd (szerk.): *Diskurzus a grammatikában – grammatika a diskurzusban.* Tinta Kiadó, Budapest, 46–55.

Horváth Viktória 2010. Funkció és kivitelezés a hezitációs jelenségekben. In Navracsics Judit (szerk.): *Nyelv, beszéd, írás. Pszicholingvisztikai tanulmányok I.* Tinta Kiadó, Budapest, 65–74.

Hutchison, Ben – Pereira, Cécile 2001. *Um, one large pizza.* A preliminary study of disfluency modelling for improving ASR. In Lickley, Robert – Shriberg, Lisa (eds.): *Disfluency in spontaneous speech. Proceedings of* the *ISCA Workshop.* University of Edinburgh, Edinburgh, Edinburgh, 77–81.

Huszár Ágnes 1985. A rádió és a televízió beszélt nyelvének mondattana. In Grétsy László (szerk.): *Nyelvészet és tömegkommunikáció.* Tömegkommunikációs Kutatóközpont. Budapest, 73–117.

Keszler Borbála 1983. Kötetlen beszélgetések mondat- és szövegtani vizsgálata. In Rácz Endre – Szathmári István (szerk.): *Tanulmányok a mai magyar nyelv szövegtana köréből.* Akadémiai Kiadó, Budapest, 164–202.

Kontra Miklós 1988. Bevezető. In Kontra Miklós (szerk.): *Beszélt nyelvi tanulmányok.* MTA Nyelvtudományi Intézet, Budapest, 1–4.

Labov, William 1981. Can dialectology deal with spontaneous speech? In Warkentyne, Henry J. (ed.): *Papers from the Fourth International Conference on Methods*

in Dialectology. Department of Linguistics, University of Victoria, British Columbia, Canada, 7–28.
Leech, Geoffrey 1992. Corpora and theories of linguistic performance. In Svartvik, Jan (ed.): *Directions in corpus linguistics*. Mouton de Gruyter, Berlin, 105–122.
Markó Alexandra 2009. Stigmatizált hanglejtésforma a spontán beszédben. *Beszédkutatás 2009*. 88–107.
Markó Alexandra – Bóna Judit 2006. A spontán beszéd lejegyzésének néhány módszertani kérdése. *Beszédkutatás 2006*. 124–133.
Neuberger Tilda 2009. A spontán beszéd lejegyzése – a BEA adatbázis tapasztalatai alapján. *Beszédkutatás 2009*. 182–195.
Nikléczy Péter – Horváth Viktória 2007. Nyelvjárási hangarchívum az interneten. *Beszédkutatás 2007*. 173–178.
Olaszy Gábor 1999. Beszédadatbázisok készítése gépi beszéd-előállításhoz. *Beszédkutatás 1999*. 68–89.
Szende Tamás 1973. *Spontán beszédanyag gyakorisági mutatói*. Nyelvtudományi Értekezések 81. Akadémiai Kiadó, Budapest.
Váradi Tamás 2000. Modern nyelvi technológiák a magyar nyelvért. In Kiefer Ferenc – Gósy Mária (szerk.): *Helyzetkép a magyar nyelvtudományról*. MTA Nyelvtudományi Intézet, Budapest, 146–156.
Váradi Tamás 2003. A Budapesti Szociolingvisztikai Interjú. In Kiefer Ferenc – Siptár Péter (szerk.): *A magyar nyelv kézikönyve*. Akadémiai Kiadó, Budapest, 339–359.
Vicsi Klára – Víg Attila 1998 Az első magyar nyelvű beszédadatbázis. *Beszédkutatás 1998*. 163–178.
Vicsi Klára 2001. Beszédadatbázisok a gépi beszédfelismerés segítésére. *Híradástechnika* 2001/1. 5–13.
Vicsi Klára – Tóth László – Kocsor András – Gordos Géza – Csirik József 2002. MTBA – magyar nyelvű telefonbeszéd-adatbázis. *Híradástechnika* 8. 35–39.

A szerzők köszönetet mondanak Fegyó Tibornak, Mihajlik Péternek, Nyári Beátának és Balogh Andrásnak értékes tanácsaikért és a lejegyzési útmutató átdolgozásában nyújtott segítségükért.