A BEA beszélt nyelvi adatbázis tízéves fejlesztése és kutatási eredményei

Bevezetés

Különféle szövegekből létrehozott gyűjtemények már sok-sok évtizeddel ezelőtt is léteztek. Az írott nyelv korpuszai[1] sokfélék, közülük az első néhány, kisméretű korpusz már igyekezett kihasználni a korabeli számítógépek nyújtotta lehetőségeket az 1960-as években (Tognini Bonelli 2010). Mára már rendkívül sokféle, írott anyagokból álló adatbázis létezik, amelyeket különböző szempontok mentén osztályoznak (Lee 2010). A huszadik század ötvenes éveiben piacra került hordozható magnetofonok biztosították a beszédszövegek széles körű rögzítését, lehetővé téve az anyagok többszöri lejátszását és ezáltal a leiratok pontosabbá válását (szemben a korábban alkalmazott gyorsírással). Edinburgh-ban készült az első elektronikusan rögzített korpusz 1963 és 1965 között (Krishnamurthy 2004), amely 66 ezer szót tartalmazott formális társalgásokban (átirattal együtt). A számítógépek kis memóriakapacitása miatt a korpuszok fejlődése relatíve lassú volt. A beszélt nyelvi és annotált korpuszok létrehozása a nyolcvanas években nagyobb ütemben indult meg, eleinte főként skandináv tudósok kezdeményezésére, de angol, francia, héber, kínai, sőt fríz nyelven is jöttek létre korpuszok (Tognini Bonelli 2010). A beszédvizsgálatok iránt felélénkült érdeklődés vezetett a különféle beszédkorpuszok létrehozásához az elmúlt évtizedekben; mára már nem egy közülük „megakorpusznak" tekinthető (Lee 2010).

A nagy terjedelmű, előre meghatározott kritériumok mentén fejlesztett beszédadatbázisok létrehozását a beszédkutatás egyik nagy forradalmaként tartják számon. A beszédadatbázis rögzített beszédanyagok rendszerezett gyűjteménye, amelynek létrejötte a technológia fejlődésének és a nagy memóriakapacitású számítógépnek köszönhető (Váradi 2000). Ennek eredményeként az elmúlt évtizedekben számos olyan kutatási kérdést lehetett sok beszélő, nagy mennyiségű beszédadatának feldolgozásával vizsgálni a különböző nyelvekben, amelyekre korábban nem volt lehetőség. Egyre jobban középpontba került a spontán beszéd tanulmányozása, amely jelzi a valós nyelvhasználat megismerésének az igényét, és amely az adatbázisok alkalmazása nélkül nem vagy csak nehezen volt teljesíthető. Nagy mennyiségű anyagon váltak vizsgálhatóvá az egyéni beszédjellemzők, a beszéd stílusfüggő változatai, úgynevezett finom fonetikai jelenségek, pszicholingvisztikai, pragmatikai, szociofonetikai sajátosságok. Újabb beszédtechnológiai vonatkozások kerültek előtérbe, mint például a beszélődetektálás, amelynek során a folyamatos beszédben az akusztikai jelből gépi úton határozzák meg, hogy mikor ki beszél. A beszédkutatásban az adatbázisok használata módszertani szempontból is jelentős; biztosítja a kutató számára a beszédanyag relatív állandóságát, a célnak megfelelő azonosságokat és hasonlóságokat, a válogatás lehetőségét, így nagymértékben megkönnyíti az adott kutatáshoz szükséges adatközlők megtalálását.

A beszédadatbázisok különfélék lehetnek (Lee 2010; Gósy 2012); tartalmukat (a rögzített beszéd témakörei), a nyelvi anyagot (pl. beszédhangkapcsolatok, szósorok, mondatok, szövegek), a beszédtípust (pl. monológok, történetmesélések, párbeszédek, társalgások), a rögzítés körülményeit (laboratórium, terep, telefon) és egyéb sajátosságaikat (pl. a protokoll) meghatározza az adott kutatási cél. Terjedelmük igen nagy változatosságot mutat, a 170 000 szótól a 100 millió szóig,

[1] A tanulmány nem tér ki a korpusz és az adatbázis terminusok differenciálására (ezek olykor szinonimaként is előfordulnak). A hivatkozott szakirodalomban olvasható terminusok használatát megtartottuk.

és az adatközlők száma is nagyon különböző (rögzítettek adatbázist 40 amerikai angol beszélővel és 1395 japán adatközlővel). A korpuszok további fontos ismérve, hogy csak beszédanyagot tartalmaznak-e, avagy azok valamilyen módon lejegyezett változatai is a korpusz részei (ún. szövegfájlok), illetve hogy milyen mértékben lekérdezhetők.

A felnőttek beszédét rögzítő korpuszok mellett számos gyermeknyelvi is létezik, amelyek mérete és egyéb jellemzői – a felnőttekéihez hasonlóan – nagyon változatosak. 1984-ben hozták létre és a mai napig fejlesztik a CHILDES (Child Language Data Exchange System = Gyermeknyelvi adatcsere rendszer, vö. MacWhinney 2000) adatbázist, amely különböző anyanyelvű gyermekek közléseit tartalmazza. Azóta további adatbázisok is készültek az elmúlt évtizedekben különféle életkorú gyermek adatközlőkkel (vö. Gyarmathy–Neuberger 2015).

Előzmények

A fonetika történetének más területeihez hasonlóan, a magyar kutatók az adatbázisok létrehozásában is úttörők voltak, noha a kezdetekben ezek szórványos beszédrögzítések, illetve inkább (strukturálatlan) beszédgyűjtemények voltak. Már Bartók Béla és Kodály Zoltán népdalfelvételein is találhatók nyelvjárási adatok. 1912-ben Bíró Ányos bencés szerzetes a magyar nyelvjárások megörökítése érdekében rögzített nyelvjárási beszédet az ország több pontján (Kiss 2001). Az első világháború idején, az akkori Fonétikai Laboratórium vezetője, Balassa József irányításával történtek beszédfelvételek (úgy tudjuk, hogy ezek az Esztergom melletti fogolytábor 48 votják, baskír, tatár és orosz hadifoglyának beszédét tartalmazták). A felvételeket hordozható gramofonnal viaszhengerekre rögzítették; sajnos nem maradtak fenn, valószínűleg elvesztek (Gósy et al. 2011). Az első, ma is hozzáférhető, értékes anyag Hegedűs Lajos nevéhez fűződik, a munkálatok a 20. század negyvenes éveiben kezdődtek. A cél nyelvjárási hangfelvételek készítése volt földrajzilag különböző helyeken. Hegedűs Lajos így írt: „A hanglemezek anyagát elektro-mágneses átírókészülékkel kymográfra vihetjük és akár egész, akár egyes helyeinek időtartamát, hanglejtését, beszédszüneteit vagy ritmusát objektív vizsgálódás tárgyává tehetjük" (Hegedűs 1946: 6). Ez az adatbázis 414 beszélő anyagát tartalmazza (eredetileg üveg- és alumíniumalapú decilith- és lakklemezeket használtak). A felvételek száma mintegy 1700, összesen több mint 50 órányi beszédanyag, 840 lemezen. (Ezt a gyűjteményt az MTA Nyelvtudományi Intézet archiváltatta a 20. század kilencvenes éveinek végén – azóta Hegedűs-archívum néven áll a kutatók rendelkezésére korszerű adathordozókon az az 1200 hangfelvétel, amelyek archiválása megoldható volt, vö. Gósy et al. 2011.)

A kutatók (korábban és ma is) gyakran a saját kutatásaikhoz rögzítettek beszédanyagokat, például Szende Tamás, aki spontán beszédet vett fel az 1970-es években (1973). A Nyelvtudományi Intézet Fonetikai Laboratóriumban készített, különféle műfajú felvételek egy része alkotja az úgynevezett Szalag korpuszt, amely a hetvenes években rögzített spontán beszédanyagokat tartalmaz (Auszmann 2015). Keszler Borbála osztálytermi beszédet rögzített a nyolcvanas években (1983). A *Budapesti Szociolingvisztikai Interjú* (BUSZI) a nyolcvanas évek végén 250 beszélővel (kazettás) magnetofonra felvett, egyenként 2–3 órás interjút tartalmaz (Kontra 1988; Váradi 2003). Számítógépes lejegyzése, illetve kódolása is megtörtént. A BABEL nemzetközi szabvány alapján készült beszédadatbázis, 60 bemondóval felolvastatott anyagokat tartalmaz (Vicsi–Vig 1998). Az MTBA magyar telefonbeszéd-adatbázis vezetékes és mobiltelefonról rögzített beszédkorpusz, 500 adatközlő felolvasásaiból áll (Vicsi et al. 2002; Vicsi 2010). A *Magyar nyelvjárási hangoskönyv* (http://geolingua.elte.hu) a *Magyar nyelvjárási olvasókönyv* (Hajdú–Kázmér 1974) különféle szövegeiből ad közre válogatást az eredeti hangfelvételekkel együtt. Videofelvételeket (111 fiatal beszélő, mintegy 50 órányi anyag, felolvasások, irányított beszélgetések, kötetlen társalgások) és azok annotálásait is tartalmazza a multimodális HuComTech adatbázis (Hunyadi 2011). A fejlesztők célja az volt, hogy lehetővé tegyék az ember-ember kommunikáció azon elemeinek és szerkezeti viszonyainak a tanulmányozását, amelyek relevánsak az ember-gép kommunikációban, és technológiai szempontból is megvalósíthatók. A nyelvjárási anyagok elemzését és térképezését a BihalBocs fejlesztés (http://www.bihalbocs.hu) tette lehetővé (Vékás 1999; Vargha 2008). A magyar nyelvjárások atlaszának (MNyA.) szerves része az a mintegy 461 órányi hangfelvétel, amely az 1960 és 1964 kö-

zött lezajlott ellenőrző gyűjtések során készült (25 szlovákiai és 327 magyarországi kutatóponton); spontánbeszéd-felvételeket is tartalmaz (Balogh–Végh 1975).

A felnőttnyelvi beszédkorpuszok mellett – csakúgy, mint más nyelveken – megjelentek a magyar gyermeknyelvi beszédadatbázisok. Érdemes megjegyezni, hogy 10 év és 16 év közötti falusi gyermekektől származó beszédfelvételeket (meséket és elbeszéléseket) már a Hegedűs-archívumban is találunk az 1940-es évekből (Menyhárt 2012). A gyermeknyelvi adatbázisok is meghatározott céllal készültek, a jellemzők tekintetében meglehetősen sokfélék (gyakran egy-egy kutató saját munkájához hozott létre gyermeknyelvi korpuszt). A teljesség igénye nélkül említünk néhányat; elsőként a SPECO projekt keretein belül készült adatbázist, amelyben 5 és 10 év közötti gyermekek ismételnek, illetve olvasnak fel hangkapcsolatokat, szavakat, mondatokat (Csatári et al. 1999). A Magyar Óvodai Nyelvi Korpusz (MONYEK) 4,5–5,5 éves, budapesti gyermekek 20–30 perces felvételeit tartalmazza (Mátyus–Orosz 2014). Ezekhez átírás, valamint morfoszintaktikai annotálás is készült. 2013-ban kezdték és jelenleg is fejlesztik a GABI, Gyermeknyelvi beszédAdatBázis és Információtár elnevezésű korpuszt (Bóna et al. 2014). Ez egy széles életkori spektrumot átfogó, sok szempontú kutatásra alkalmas, nagy mennyiségű hanganyagot tartalmazó gyermekbeszéd-adatbázis. A beszélői 3–18 éves gyermekek, illetve fiatalok, a beszédfelvételek 30–40 percesek; a felvételi protokollhoz a BEA adatbázis szolgált mintául.

A BEA adatbázis fejlesztése

A jelen tanulmányunk célja a BEA (BEszélt nyelvi Adatbázis) jellemzőinek és a rajta végzett különféle kutatások néhány irányának a bemutatása. A BEA az MTA Nyelvtudományi Intézet Fonetikai Osztályának a fejlesztése, egy multifunkcionális beszédadatbázis, beszélőinek számát, a protokollt, a rögzítés körülményeit, az átiratait, illetve a nagyságát tekintve nemzetközi tekintetben is kiemelkedő. Nagy a jelentősége a tudományos és a társadalmi hasznosíthatóság tekintetében, hozzájárulás a nemzeti kulturális örökséghez, a budapesti beszélők nyelvhasználatának megőrzéséhez az utókor számára.

A BEA adatbázis fejlesztése 2007-ben, éppen tíz esztendővel ezelőtt kezdődött a célok és az alapvető jellemzők meghatározásával. Multifunkcionális beszédanyagot terveztünk létrehozni, amely a különféle nyelvészeti területek (pl. pszicholingvisztika, szociolingvisztika, pragmatika) kutatási igényeinek is képes megfelelni, egyúttal alkalmas a fonetika területein olyan beszédvizsgálatokra, amelyekhez megfelelő minőségű rögzítések szükségesek. Nem volt cél az adatközlők reprezentatív megjelenítése semmilyen tekintetben (ugyanakkor az adatbázis felvételei alapján reprezentatív minta egy adott kutatáshoz létrehozható). Kialakítottuk és érvényesítjük a *Humán vizsgálatokon alapuló nyelvészeti kutatások etikai szabályozásá*ban foglaltakat (az „1995. évi CXIX. törvény a kutatás és a közvetlen üzletszerzés célját szolgáló név- és lakcímadatok kezeléséről”, valamint az „1992. évi LXII. törvény a személyes adatok védelméről és a közérdekű adatok nyilvánosságáról” figyelembevételével). Megtörtént az anonimizálás folyamatának kialakítása: az adatközlők kódokkal szerepelnek az adatbázisban, személyük azonosíthatatlan. Folyamatosan tekintetbe vesszük a korszerű felvételi technikákat, és bizonyos mértékig érvényesülnek (noha ez nem cél) a szociológiai tényezők (pl. az adatközlők iskolázottsága). A próbafelvételeket követően 2007 őszétől indult meg az adatbázis máig tartó fejlesztése.

A rögzített beszédanyagokkal kapcsolatosan időről időre felmerül az a kérdés, hogy vajon mennyire természetesek ezek a felvételek. Gyakran keveredik eközben a beszéd természetessége és a spontaneitása, ami abból is adódik, hogy a két fogalmat olykor egymás szinonimájaként használják. A spontán beszéd azt jelenti, hogy a beszélő előzetes felkészülés nélkül önti nyelvi formába a gondolatait, és ejti ki egy adott helyzetben, függetlenül attól, hogy például egy társalgásban vesz részt, avagy narratívát mond. A spontán közlés során a gondolatok kialakítása, a válogatás közöttük, a szükséges szavak, a grammatikai formák és a fonológiai, fonetikai tervezés egyidejűleg zajlik, miközben a kiejtés is folyamatban van (Levelt 1989; Gyarmathy 2015). A spontaneitás nem függ a beszélő aktuális lelkiállapotától, avagy az ismereteitől. A beszéd természetessége egészen más. Az a mód, ahogyan a beszédünket tervezzük, ahogyan a gondolatokból nyelvi szerkezetek lesznek, ahogyan beszélünk, kiejtjük a beszédhangokat, azok sorozatát, a frázisokat (stb.) változhat (változik

is) beszédhelyzetről beszédhelyzetre (de napszakról napszakra is). A természetesség skálán képzelhető el, ahol egy adott beszédprodukció számos tényező metszéspontjában jellemezhető. Meghatározó (lehet) a beszéd témája, az, hogy kik a beszédpartner(ek), a beszéd tartalma, célja, és mindez továbbá nem független a beszélő személyiségétől, aktuális lelkiállapotától, egészségétől és még számos más tényezőtől. Ezek vannak hatással a beszéd természetességére, vagyis arra, hogy egy adott beszédprodukció mennyire tükrözi a beszélőre jellemző, szokásos beszédmódot. A nem megszokott beszédhelyzet, magának a beszédrögzítésnek a ténye azonban befolyásolhatja az adatközlő szokásos beszédprodukcióját, azaz változhat az, amit a környezet természetesnek ítél. A mesterséges kommunikációs helyzet felerősítheti az önkontrollt, mivel a beszélő egy feltételezett elvárásnak próbál megfelelni (pl. a szavak megválasztása, a kiejtés egyfajta szabályozása), csakúgy, mint a kérdőívek kitöltésekor. A beszélők törekszenek egy vélt nyelvi norma megközelítésére (Szende 1973). Érdemes megemlíteni a szorongást is, amely általában hatással van a spontán beszéd folyamatosságára és a beszéd természetességére is, különösen a beszéd rögzítése esetén, de ebben a beszélők között nagy különbségek tapasztalhatók. A beszéd természetességét úgy is definiálják, hogy az megegyezik a beszélő szándékának megfelelő beszéddel (Wolfson 1976), illetve megfelel az adatközlő személyiségének (Nusbaum et al. 1995). Tapasztalataink szerint a BEA adatbázis felvételei során az adatközlők gyorsan alkalmazkodnak a helyzethez, a szorongásuk látványosan csökken, illetve elmúlik (ennek felvétel közben és utána is gyakran hangot adnak), beszédük tehát a szokásosnak, vagyis a körülményekhez képest természetesnek tekinthető.

A BEA felvételi protokollja

A beszédfelvételek meghatározott protokoll szerint készülnek, amely a rögzítendő beszédanyagok tartalmi vonatkozásait és technikai jellemzőit is előírja. Az adatbázis főként különféle típusú spontán beszédanyagokat tartalmaz, de a szélesebb kutathatóság érdekében mondatismétléseket és felolvasásokat is magában foglal. A tartalmi protokoll a következő hat részből áll: mondatismétlés, narratíva, véleménykifejtés, tartalomösszegzés, társalgás, felolvasás. Az alábbiakban az egyes részeket ismertetjük.

1. A mondatismétlés 25 egyszerű és összetett mondatot tartalmaz. A tesztmondatok összeállításánál a fejlesztők figyeltek a grammatikai szerkezet, a szórend és a koartikulációs szabályok változatosságára, például: *Az ügyfeleknek kompromisszumot kellett kötniük; Nem lehetett teljes bizonyossággal megítélni a vádlott elmeállapotát; A minap önmagát kiáltotta ki a legnagyobb énekesnek a világon; Nem kötött biztosítást, ezért kisebb vagyonba került a kórházi ellátás.* A mondatok átlagosan 8–12 szóból állnak, ezek többségükben 3–4 szótagosak. A huszonöt mondatot úgy állították össze, hogy a megismétlésük még idősebb korban se okozzon túlzott kognitív megterhelést.

2. A narratíva az adatközlő életéről, családjáról, munkájáról, hobbijáról szól, jellegét tekintve többé-kevésbé összefüggő monologikus szöveg, amely egyben feszültségoldó is, hiszen az emberek rendszerint könnyedebben beszélnek a munkájukról vagy a hobbijukról, nem tartanak attól, hogy konkrét ismereteket várnak el tőlük.

3. A véleménykifejtés (amely nagyobbrészt szintén narratíva) az interjúkészítő által megadott téma véleményezése. A felvetett témák általánosak és hétköznapiak, aktuális társadalmi, közéleti kérdésekkel kapcsolatosak, a róluk történő véleménykifejtés sem konkrét ismereteket, sem háttértudást nem kíván. Példák: a házasság, illetve együttélés; eutanázia; közlekedés a fővárosban; sztárok/celebek alkohol- és drogproblémái; ittasan okozott autóbalesetek; megúszott büntetések; a magántulajdon védelme; új adók, új szabályozások; BKV/MÁV sztrájk; motoros balesetek; az emberi szervek felajánlása. Ez a beszédfeladat bizonyos értelemben nehezebb, mint a megelőző, hiszen az adatközlőnek át kell gondolnia, mi a személyes véleménye az adott témáról, abból mennyit kíván elmondani, mit kíván elhallgatni. A beszélő ilyenkor nemegyszer felméri, hogy mi lehet az általános társadalmi vélekedés, és ő személy szerint ahhoz képest mit képvisel stb. Az interjúkészítő a témát

az adatközlő életkorának és a narratíva alapján hallott érdeklődési körének megfelelően választja ki, és rugalmasan alkalmazkodik a beszédpartner mindenkori reakcióihoz.

4. A tartalomösszegzés voltaképpen irányított spontán beszéd. Az adatközlő felvételről meghallgat egy-egy szöveget (ezeket női hanggal rögzítették, átlagos beszédtempóban), és ezt követően rögtön a saját szavaival el kell mondania a hallott tartalmat. Az egyik egy rövid tudománynépszerűsítő szöveg (174 szavas; 1 perc 37 mp tartamú), a másik egy anekdotaszerű történet (270 szavas; 2 perc 5 mp tartamú). Ennek a beszédfeladatnak a relatív nehézsége abból adódik, hogy a mindennapi kommunikációban ritkábban fordul elő, hogy egy hallott történetet vagy ismeretterjesztő szöveget a beszélőnek a saját szavaival kell elmondania. Ez a rész teljesen monologikus, a kísérletvezető nem szólal meg közben.

5. A társalgásban az adatközlőn és az interjúkészítőn kívül egy további személy vesz részt. A téma változó, az élet mindennapjaihoz kapcsolódik; ugyanazon adatközlő esetében mindig különbözik a véleménykifejtés témájától. A társalgás témáiból: karácsony, húsvét ünneplése; mobiltelefon kisgyermekeknek; új KRESZ és a biciklisek; halálbüntetés; dohányzási tilalom a szórakozóhelyeken; éjszakai élet, szórakozási lehetőségek Budapesten. Az interjúkészítő a témát ekkor is az adatközlő életkorának és érdeklődési körének figyelembevételével választja ki. A tapasztalatok alapján a társalgás a mindennapi beszédhelyzeteket legjobban modellező része a protokollnak. Itt ugyanis nemcsak az adatközlők beszélnek, hanem két másik személy is, a résztvevőknek van idejük többet gondolkodni (szemben a narratívákkal), és a helyzetből adódóan jobban elterelődik a figyelem a felvétel körülményeiről.

6. Az utolsó részben az adatközlő kétféle szöveget olvas fel; az egyik a huszonöt, korábban ismételt mondat, a másik egy tudomány-népszerűsítő cikk felolvasása, amely 291 szóból áll. A tapasztalatok szerint a hangos olvasás az adatközlőknek nem okoz nehézséget (módjuk van a meghangosítás előtt némán elolvasni a szövegeket).

A beszédrögzítés technikai paramétereit tekintve a legfontosabb követelmény a BEA adatbázis esetében a megfelelő jel/zaj viszonyú, széles frekvenciatartományú, magas dinamikájú, torzításmentes felvételek elkészítésének a biztosítása. Ez akusztikai szempontból megköveteli a zajszigetelt és visszhangmentesített helyiséget, elektronikus tekintetben pedig a megfelelő hangfelvevő és hangrögzítő rendszert. Mindezek miatt a korpusz beszédfelvételei az MTA Nyelvtudományi Intézet Fonetikai Osztályán található zajszigetelt szobában készülnek; a hangcsillapítás mértéke a külső környezethez képest 50 Hz-en 35 dB, 250 Hz fölött pedig ≥ 65 dB. A kritériumainknak megfelelő technikai hátteret az Audio-technika AT4040 típusú kardioidkondenzátor-mikrofonok, a Phonic MM102 típusú 2 csatornás Phantom tápos analóg keverőpult, a GoldWave szoftver, illetve a 44,1 kHz-es mintavételezéssel közvetlenül a számítógépre történő digitális rögzítés biztosítja.

A BEA adatbázis felvételeinek statisztikai adatai

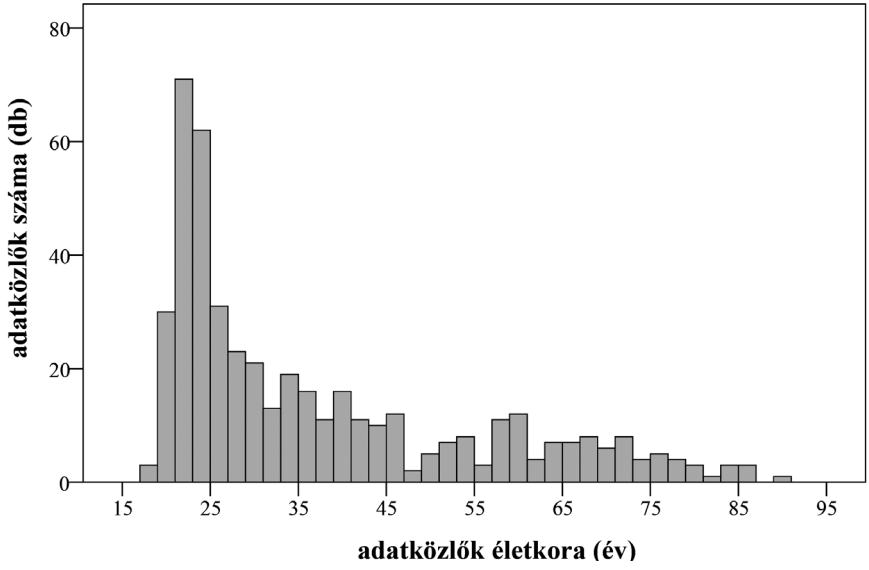
Az adatbázis felvételei anonimizáltak, az egyes beszélőket kódszámok jelölik. Minden felvétel esetében (név nélkül) rögzítik az alábbi információkat: a felvétel időtartama, az adatközlő neme, életkora, iskolai végzettsége, foglalkozása, súlya, magassága, dohányzási szokásai, a véleménykifejtés és a társalgás témája, illetve a felvételvezető és a társalkodó partner személye. Az adatközlők magasságának és súlyának a beírása a résztvevők bemondása alapján történik. Ezek az adatok lehetőséget nyújtanak az adatbázis széles körű kutathatóságára. A felvételek mintegy 85%-ában az interjúkészítő ugyanaz a fiatal kutató, ami kiváló lehetőséget nyújt például a beszédalkalmazkodás kutatására is.

A BEA adatközlői egynyelvű, köznyelvet beszélő budapestiek, jelenleg 461 beszélő, közülük 281 nő és 180 férfi, átlagéletkoruk 36,8 év (18 évestől 90 évesig). A beszélők iskolai végzettségét tekintve a felsőfokú végzettségűek vannak többségben (54,66%), illetve magas az érettségizettek

aránya is (42,08%). Az adatközlők kis százaléka (3,25%) rendelkezik mindössze általános iskolai végzettséggel. Foglalkozásukat tekintve beszélőink között találhatók általánosan ismert munkaköröket betöltők, például pedagógusok, gyógypedagógusok, orvosok, autószerelők, informatikusok, adminisztrátorok, színészek; de akadnak közöttük viszonylag ritkább foglalkozásokat űzők, például: cserépkályha-készítő, jelmeztervező, evangélikus lelkésznő, orgonaépítő, pókerjátékos, forgatókönyvíró, sajtótitkár és kaszkadőr.

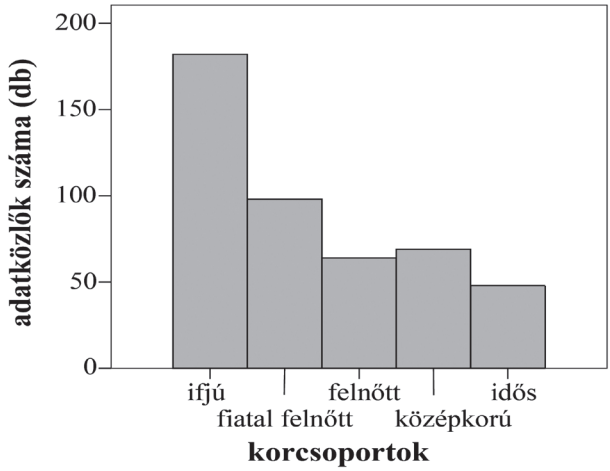
A BEA adatbázis jelenlegi 461 felvétele összesen 367 óra, 28 percnyi hanganyagot tesz ki. A legrövidebb felvétel 19 perc és 17 másodperc hosszú, míg a leghosszabb ennek a hétszerese, 2 óra 24 perc 47 másodperc. Egy felvétel átlagosan 47 perc 50 másodperc időtartamú.

Hangsúlyozzuk ismét, hogy az adatbázis fejlesztésekor nem volt cél sem az, hogy a társadalom demográfiai állapotát reprezentálja, ahogy az sem, hogy akár az életkort, a nemet, avagy az iskolai végzettséget tekintve kiegyenlített legyen. A BEA életkori eloszlását az 1. ábra szemlélteti. A legtöbb adatközlő a huszonéves korosztályba tartozik, míg a felnőttek és az idősek csoportja e tekintetben homogénebb. Ez nagyban összefügg azzal, hogy a felvételek általában hétköznapokon, munkaidőben készülnek, így azok érnek rá eljönni, akik még (vagy már) nem dolgoznak, illetve munkaidejük rugalmas.



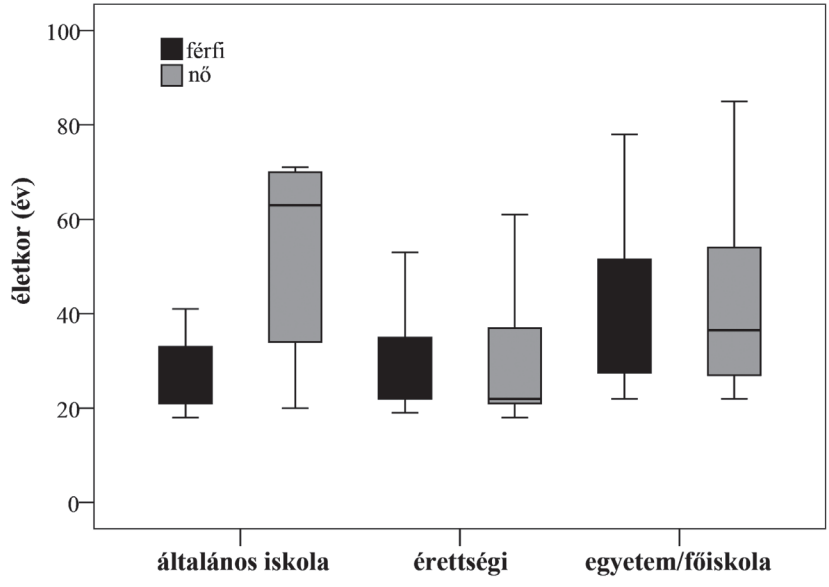
1. ábra. *Az adatközlők életkori eloszlása*

Az életkori adatok átláthatóbb szemléltetéséhez az adatközlőket (önkényesen) öt korcsoportra osztottuk: 1. ifjú (25 éves korig), 2. fiatal felnőtt (26–35 éves korig), 3. felnőtt (36–45 éves korig), 4. középkorú (46–65 éves korig) és 5. idős (66 éves kor felett). A 2. ábrán jól látszik, hogy az ifjúkorú adatközlők száma a legtöbb, a többi korcsoport beszélőszáma többé-kevésbé hasonló. A 25 év alattiak aránya a korpuszban 39,48%, a fiatal felnőtteké 21,26%, a felnőtteké 13,88%, a középkorúaké 14,97%, míg az időseké 10,41%.



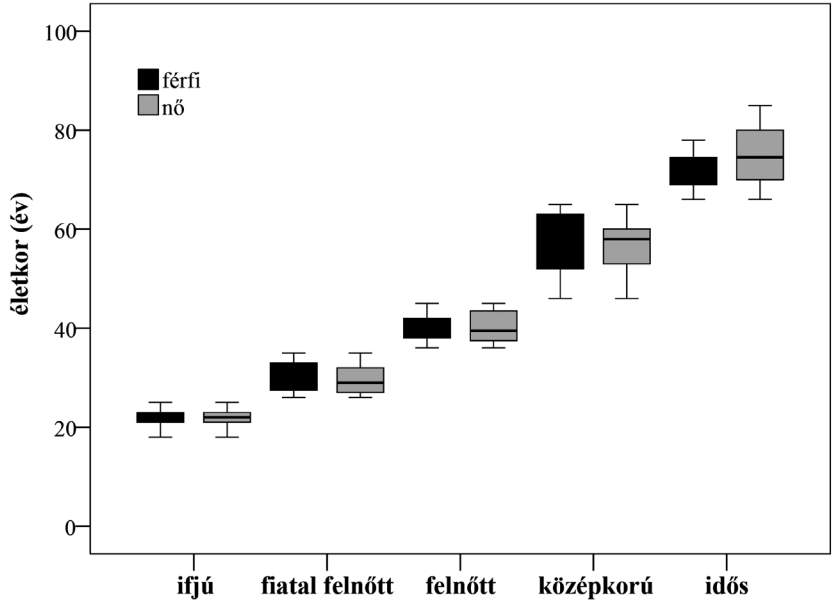
2. ábra. *Az adatközlők korcsoportok szerinti eloszlása*

Az adatközlők életkorát a nem és az iskolai végzettség függvényében vizsgálva megállapítható, hogy az érettségizettek és a felsőfokú végzettséggel rendelkezők között nincs jelentős eltérés a két nem között (3. ábra). Az előbbi csoportban a férfiak átlagéletkora 29,7 év (szórás: 19–73 év), a nőké 31,9 év (szórás: 18–85 év), míg az utóbbiban a férfiaké 40,2 év (szórás: 22–90 év), a nőké 41,8 év (szórás: 22–85 év). Az általános iskolát végzettek csoportjában azonban a két nem számottevő különbséget mutat; a férfiak átlagéletkora itt 29,1 év (szórás: 18–58 év), a nőké 51,6 év (szórás: 20–71 év). Az adatok természetesen nem reprezentatívak.



3. ábra. *Az életkor, a nem és az iskolai végzettség összefüggése*

Az adatbázisban a nemek aránya nem kiegyenlített; a nők aránya 61%, a férfiaké 39%. Az átlagéletkort tekintve nincs jelentős különbség köztük; női adatközlőink átlagosan 37,5 évesek, míg a férfiak 35,8 évesek. Ez a kiegyenlítettség a korcsoportok szerinti bontásban is megfigyelhető (4. ábra). Az egyes csoportokban eltérés csak a nemek arányában jelentkezik. Az ifjúkorúak, a középkorúak és az idősek csoportjában kétszer annyi a női beszélő, mint a férfi.

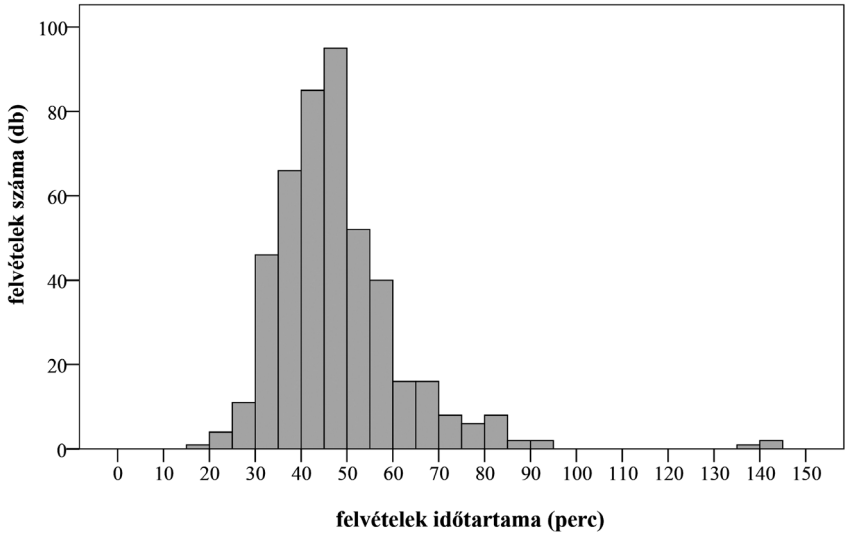


4. ábra. *Az adatközlők életkora a nemek és a korcsoportok függvényében*

Az adatközlők súlya átlagosan 70,8 kg (42–160 kg), magasságuk 172 cm (132–202 cm) volt. Ugyanez nemek szerinti bontásban a következőképp alakul. A férfi beszélők átlagos testmagassága csaknem 180 cm (165–202 cm), súlyuk 81,8 kg (51–160 kg); míg a nők átlagosan 166,9 cm magasak (148–197 cm) és átlagosan 63,8 kg-osak (42–113 kg) voltak. A 461 adatközlő közül csupán 83-an (18%) dohányoznak rendszeresen, 23-an (5%) csak alkalmanként, illetve már leszoktak, és 355-en (77%) egyáltalán nem dohányoznak. A férfiak és a nők aránya mindhárom csoportban hasonlóan alakul (mintegy 40% férfi, 60% nő).

Az egységes felvételi protokoll ellenére a hanganyagok időtartama változatos képet mutat (5. ábra). Előfordulnak ugyanis szűkszavú és vannak bőbeszédű adatközlők, de bármilyen egyéb külső vagy belső körülmény (pl. az adatközlő és a felvételvezető ismeretségi foka, az adott téma, az adatközlő hangulata, fiziológiai, egészségügyi állapota, időbeosztása stb.) is befolyásolhatja a felvétel hosszát. Az esetek döntő többségében 30 és 60 perc között valósul meg egy-egy felvétel; leggyakrabban – mint írtuk – mintegy háromnegyed órát vesz igénybe. Ha korcsoportok szerint elemezzük a felvételek hosszát, megállapítható, hogy a középkorúak beszéltek a leghosszabban (átlag: 49,9 perc), őket követték a fiatal felnőttek átlagosan 47 perccel, majd az idősek 46,9 perccel. Az ifjúkorúak átlagosan 44,8 percet beszéltek, míg a legrövidebb felvételek (43,5 perc) a felnőttektől származnak. Az elemzésbe a nemeket is bevonva megállapítható, hogy a férfiak sokkal homogénebb csoportot alkotnak, náluk nincsenek akkora különbségek a beszédtartamban az egyes korcsoportok között, mint a nőknél. A középkorú és a fiatal felnőtt férfiak felvételei a leghosszabbak, 48,4 perc, illetve 48,2 perc; őket követik az ifjak 47,3 perccel, majd az idősek 45,7 perccel, míg a legrövidebb a férfiaknál is a felnőttek beszédanyaga (44,4 perc). A nőknél a leghosszabb felvételek a középkorú

(51,4 perc) és az idős (48,1 perc) hölgyektől származnak, majd sorrendben a fiatal felnőttek következnek 45,9 perccel, míg az ifjúkorú (42,3 perc) és a felnőtt (42 perc) nők felvételei csaknem azonos időtartamúak.



5. ábra. *A felvételek időtartamának eloszlása*

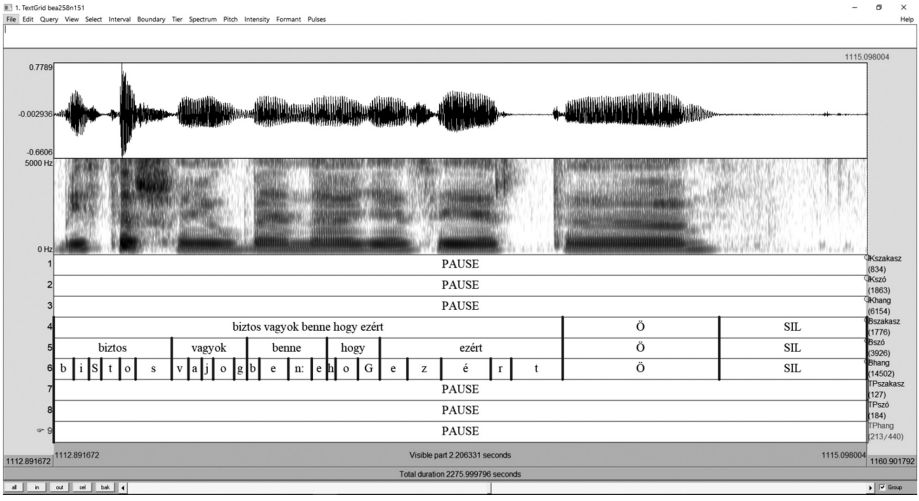
A BEA adatbázis lejegyzései

A BEA hanganyagainak lejegyzése az elmúlt 10 év során többféle módon zajlott. A kezdetekben ez egy elsődleges írásos tükröztetést takart; az átiratok a Microsoft Office Word programjában .doc formátumban, helyesírásban, központozás nélkül készültek, a későbbi feldolgozás szempontjából fontosnak ítélt adatok, mint például a megakadásjelenségek, illetve a fiziológiai hangadások jelölésével (az átiratok száma 167). A helyesíráson alapuló lejegyzés nem jelölte a kiejtés és a helyesírás eltéréseit, tehát nem érvényesítette a hasonulási, összeolvadási szabályokat (a *szabadság* szó nem *szabaccság*ként lett lejegyezve).0 A megakadásjelenségek jelölése a vastagon szedett alak volt, és ha nem hangzott el javítás, a helyes szó []-ben lett megadva, például: *kell még **tenyér** meg **kej** [kenyér meg tej]*. A lejegyzésben jelölve voltak a néma és a kitöltött szünetek, nyújtások, a spontán beszédben gyakran használt, nem szótári alakban előforduló szavak (pl. *asszem, nem tom*), az idegen szavak, rövidítések, betűszók, mozaikszók, illetőleg a lejegyző számára értelmezhetetlen közlésrészletek. Az elsődleges lejegyzések a kutatók munkáját megkönnyíteni hivatott durva átiratok voltak.

A későbbiekben a kutatási igények, illetve az automatikus, gépi beszédfelismerésben történő felhasználás szükségessé tették a lejegyzési elvek újragondolását és a szoftveres háttér megváltoztatását. 2010 októberétől a Word-dokumentumban történő átírást felváltotta a Transcriber szoftverrel történő lejegyzés, és ezzel egyidejűleg megkezdődött a már elkészült 167 darab, Wordben lejegyzett anyag Transcriberbe történő átkonvertálása. A Transcriber programban 179 lejegyzés készült el, amelyeknek nagy előnye, hogy az írott szöveg és a hanganyag egyszerre látható és hallható; a program lehetővé teszi a beszéd szegmentálását, címkézését és leírását (Barras et al. 1998). (A BEA Transcriberes lejegyzéséről bővebben lásd Gyarmathy–Neuberger 2011.)

Noha az imént röviden bemutatott lejegyzési mód a beszédfelismeréssel kapcsolatos kutatásokban kiválóan használható, más vizsgálatok számára a felhasználhatósága korlátozott. Nem alkalmas fonetikai mérésekre, elemzésekre. Ahhoz, hogy az adatbázis eredeti céljainak megfelelő-

en minél szélesebb körben felhasználhatóvá válhasson, a Fonetikai Osztály munkatársai 2014-ben kidolgoztak egy újabb, a Praat szoftverben (Boersma–Weenink 2013) történő lejegyzési rendszert. Ennek lényege, hogy a lejegyzés során a hangfájl mellé egy szöveges fájl is készül, amely egyrészt az elhangzott közléseket tartalmazza, másrészt a beszédanyagot kisebb szegmentumokra bontja időbélyegek alkalmazásával. Az annotálás a nemzetközileg használt Praat programban a leiratok kettős ellenőrzésével, illetve statisztikai kontrolljával történik. A Praat komplex akusztikai jelfeldolgozó, amely lehetőséget nyújt az annotálásra is. A BEA beszédanyagának lejegyzése ekkor három szinten történik: beszédszakaszok, szavak és beszédhangok szintjén (vö. Gyarmathy et al. 2014). A társalgási részben ez összesen kilenc címkesort jelent (a társalgásokban három beszélő vesz részt). Ennek a lejegyzési rendszernek a használatával 110 átirat készült el, további 29 pedig folyamatban van (6. ábra).



6. ábra. *A BEA adatbázis háromszintű lejegyzése (részlet társalgásból)*

Kutatások az adatbázis alapján

Számos kutatás történt az adatbázison az elmúlt tíz évben, a publikációk hazai és nemzetközi folyóiratokban, szerkesztett kötetekben, monográfiákban láttak napvilágot (pl. a *Beszédkutatás* c. folyóirat számaiban, a *Beszéd, adatbázis, kutatások* [2012], *A spontán beszéd sajátosságai az időskorban* [2013], a *Hezitációs jelenségek a magyar beszédben* [2014], a *Diszharmóniás jelenségek a beszédben* [2015], *A spontán beszéd prozódiai szerkezete* [2015], a *Morfémák időzítési mintázatai a beszédben* [2017], avagy a *Megakadásjelenségek a magyar spontán beszédben* [2017] című kötetekben). A vizsgálatok a beszéd számos területein folytak és folynak most is a beszédtervezési mechanizmus megismerésétől az akusztikai-fonetikai elemzéseken át pszicholingvisztikai, pragmatikai, szociolingvisztikai és beszédtechnológiai kérdések tanulmányozásáig.

A spontán beszéd vizsgálatának eredményei hozzájárulnak az elméleti nyelvészeti területeken folyó kutatásokhoz is. A gyakorisági mutatók a hangzó nyelv rendszerszerű leírásának fontos aspektusát képviselik. A nagy mennyiségű rögzített beszédanyag tényeinek feldolgozása lehetőséget nyújt a nyelvi norma kérdésének új szempontú elemzésére. A fonetikai kutatások számos, korábban nem vizsgált jelenséget és folyamatot jellemeztek és írtak le méréssorozatokon alapuló elemzések eredményeként a BEA alapján. Nagy adatmennyiségen elemezték a magánhangzók koartikulációs mezőit, a különböző képzésmódú mássalhangzók hangszerkezetének módosulását, zöngétlen zárhangok egymásra hatását, beszédhangok időviszonyait. Foglalkoztak a szavak fonetikai és nyelv-

használati jellemzőivel. Bemutatták a magánhangzók frázisvégi nyúlásának nyelvspecifikus mintázatát, az összetett képzésű zármássalhangzók temporális és spektrális sajátosságait, a különböző mássalhangzó-kapcsolatok egymásra hatásának fonetikai ismérveit, valamint a spontán beszédre jellemző redukciós, időzítési és glottalizációs folyamatokat (a tipikus és atipikus fonáció előfordulásait). Igazolták a magánhangzók stabilitását a beszédben, a zármássalhangzók és a közelítőhangok fonetikai realizációit, a hiátus beszélőspecifikus ejtési módozatait a spontán közlésekben, valamint a beszédtempó hatásait a kiejtésre.

Több kutatást folytattak a fonológia és a fonetika érintkezési területein, mint például a fonológiailag rövid és hosszú magánhangzók időtartamának megvalósulása a spontán beszédben, a /h/ fonéma ejtési realizációinak bemutatása, illetve az obstruensek zöngésségi oppozíciójában a fonetikai helyzetnek és a hangkörnyezetnek a zöngésségre gyakorolt hatása. Számos aspektusban elemezték a spontán beszéd egyes prozódiai jellemzőit (pl. a beszéddallam határjelző funkcióját, prozódiai és szintaktikai összefüggéseket), valamint pragmatikai sajátosságait. A társalgáselemzés számos új eredményt hozott, beleértve az univerzális és nyelvspecifikus sajátosságok, illetve a beszélőalkalmazkodás egyes jellemzőinek a leírását. A beszédtervezési folyamatok különféle diszharmóniás jelenségeinek felszíni következményeit értelmezték a tervezés és a kivitelezés összefüggéseiben. Számos kutatás irányult az önellenőrzési folyamatok vizsgálatára a spontán közlésekben. A beszéd alapján kialakított beszélői profil fejlesztésében azonosították az egyéni beszédjellemzők és a fizikai felépítés összefüggéseit. Kutatások folytak a spontán beszédre jellemző nonverbális jelenségek (nevetés, krákogás, hümmögés, nyammogás) funkcionális és fonetikai szempontú megismerésére.

A BEA adatbázis munkálatai 2007-ben kezdődtek. Az elmúlt tíz év alatt jelentős eredmények születtek, mind a fejlesztés, mind az adatbázison végzett sokféle kutatás tekintetében.

SZAKIRODALOM

Auszmann Anita 2015. A spontán beszéd időviszonyai 40 évvel ezelőtti és mai beszélőknél. In: Gósy Mária (szerk.): *Diszharmóniás jelenségek a beszédben*. MTA Nyelvtudományi Intézet, Budapest, 219–34.

Balogh Lajos – Végh József 1975. A magyar nyelvjárások atlaszához kapcsolódó hangfelvételek In: Deme László – Imre Samu (szerk.): *A magyar nyelvjárások atlaszának elméleti, módszertani kérdései*. Akadémiai Kiadó, Budapest, 257–62.

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)*, 1373–6.

Boersma, Paul – Weenink, David 2013. *Praat: Doing phonetics by computer* [Komputer program]. http://www.praat.org Letöltés: 2013. szeptember 15.

Bóna Judit 2013. *A spontán beszéd sajátosságai az időskorban*. (Beszéd – Kutatás – Alkalmazás 2.) ELTE–Eötvös Kiadó, Budapest.

Bóna Judit – Imre Angéla – Markó Alexandra – Váradi Viola – Gósy Mária 2014. GABI – Gyermeknyelvi Beszédadatbázis és Információtár. *Beszédkutatás* 2014: 246–51.

Csatári, Ferenc – Bakcsi, Zsolt – Vicsi, Klára 1999. A Hungarian child database for speech processing applications. *ESCA Proceedings*, BME, Budapest, 2231–4.

Gósy Mária 2012. Multifunkcionális beszélt nyelvi adatbázis – BEA. Prószéky Gábor – Váradi Tamás szerk. 2012. *Általános Nyelvészeti Tanulmányok* XXIV. Nyelvtechnológiai kutatások. Akadémiai Kiadó, Budapest, 329–49.

Gósy Mária – Horváth Viktória – Nikléczy Péter 2011. A Hegedűs-archívum mint korszerű adatbázis. In: Bárth M. János – Vargha Fruzsina Sára (szerk.): *Hangok – helyek*. ELTE Magyar Nyelvtudományi és Finnugor Intézet, Budapest, 85–103.

Gósy Mária szerk. 2012. *Beszéd, adatbázis, kutatások*. Akadémiai Kiadó, Budapest.

Gósy Mária szerk. 2015. *Diszharmóniás jelenségek a beszédben*. MTA Nyelvtudományi Intézet, Budapest.

Gósy Mária – Krepsz Valéria 2017. *Morfémák időzítési mintázatai a beszédben*. MTA Nyelvtudományi Intézet, Budapest.

Gyarmathy Dorottya 2015. Diszharmóniás jelenségek, megakadások a beszédben. In: Gósy Mária (szerk.): *Diszharmóniás jelenségek a beszédben*. MTA Nyelvtudományi Intézet, Budapest, 9–49.

Gyarmathy Dorottya 2017. *Megakadásjelenségek a magyar spontán beszédben* MTA Nyelvtudományi Intézet, Budapest.

Gyarmathy Dorottya – Neuberger Tilda 2011. A BEA adatbázis alkalmazásfüggő lejegyzései *Beszédkutatás* 19: 109–20.
Gyarmathy Dorottya – Neuberger Tilda – Gráczi Tekla Etelka 2014. Lejegyzési útmutató a BEA Spontánbeszéd-adatbázis háromszintű annotálásához. *Alkalmazott Nyelvtudomány* XIV/1–2. 35–45.
Gyarmathy Dorottya – Neuberger Tilda 2015. Egy hiánypótló adatbázis: a Tini BEA. *Beszédkutatás* 23: 209–22.
Hajdú Mihály – Kázmér Miklós 1974. *Magyar nyelvjárási olvasókönyv*. Tankönyvkiadó, Budapest.
Hegedűs Lajos 1946. *Népi beszélgetések az Ormánságból*. Szabadság Pécsi Nyomda és Könyvkiadó Kft., Pécs.
Horváth Viktória 2014. *Hezitációs jelenségek a magyar beszédben*. ELTE–Eötvös Kiadó, Budapest.
Hunyadi László 2011. A multimodális ember-gép kommunikáció technológiái – elméleti modellezés és alkalmazás a beszédfeldolgozásban. In: Németh T. Enikő (szerk.): *Ember-gép kapcsolat. A multimodális ember-gép kommunikáció modellezésének alapjai*. Tinta Könyvkiadó, Budapest, 15–41.
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.
Kiss Jenő szerk. 2001. *Magyar dialektológia*. Osiris Kiadó, Budapest.
Kontra Miklós 1988. Bevezető. In: Kontra Miklós (szerk.): *Beszélt nyelvi tanulmányok*. Linguistica, Series A, Studia et Dissertationes 1. MTA Nyelvtudományi Intézet, Budapest, 1–4.
Krishnamurthy, Ramesh szerk. 2004. English collocation studies. Continuum, London.
Lee, David Y. W. 2010. What corpora are available? *The Routledge handbook of corpus linguistics*. Routledge. www.routledgehandbooks.com/doi/10.4324/9780203856949.ch9. Letöltés: 2017. szeptember 18.
Levelt, Willem J. M. 1989. *Speaking: From intention to articulation*. MIT Press, Cambridge, MA.
MacWhinney, Brian 2000. *The CHILDES project: Tools for analyzing talk*. Lawrence Erlbaum Associates, Mahwah, New Jersey.
Markó Alexandra 2015. *A spontán beszéd prozódiai szerkezete*. Nyelvtudományi Értekezések 166. Akadémiai Kiadó, Budapest.
Mátyus Kinga – Orosz György 2014. MONYEK – Morfológiailag egyértelműsített óvodai nyelvi korpusz. *Beszédkutatás* 2014: 237–45.
Menyhárt Krisztina 2012. A beszéd temporális jellemzői 60 évvel ezelőtti gyermek beszélőknél. *Beszédkutatás* 2012: 246–59.
Nusbaum, Howard C. – Francis, Alexander L. – Henly, Anne S. 1995. Measuring the naturalness of synthetic speech. *International Journal of Speech Technology* 1: 7–19.
Szende Tamás 1973. *Spontán beszédanyag gyakorisági mutatói*. Nyelvtudományi Értekezések 81. Akadémiai Kiadó, Budapest.
Tognini Bonelli, Elena 2010. Theoretical overview of the evolution of corpus linguistics. In: *The Routledge handbook of corpus linguistics*. Routledge. www.routledgehandbooks. com/ doi/10.4324/9780203856949.ch2. Letöltés: 2017. szeptember 18.
Vargha Fruzsina 2008. Nyelvjárási és helynévtörténeti anyagok számítógépes feldolgozása. In: Alabán, František (szerk.): *Kontextus – Filológia – Kultúra II*. Banská Bystrica, 75–82.
Váradi Tamás 2000. Modern nyelvi technológiák a magyar nyelvért. In: Kiefer Ferenc – Gósy Mária (szerk.): *Helyzetkép a magyar nyelvtudományról*. MTA Nyelvtudományi Intézet, Budapest, 146–56.
Váradi Tamás 2003. A Budapesti Szociolingvisztikai Interjú. In: Kiefer Ferenc – Siptár Péter (szerk.): *A magyar nyelv kézikönyve*. Akadémiai Kiadó, Budapest, 339–59.
Vékás Domokos 1999. Informatikai lehetőségek a dialektológiában, különös tekintettel a fonetikai szempontokra. Kézirat. ELTE Fonetikai Tanszék, Budapest.
Vicsi Klára – Vig Attila 1998. Az első magyar nyelvű beszédadatbázis. *Beszédkutatás '98*: 163–78.
Vicsi Klára – Tóth László – Kocsor András – Gordos Géza – Csirik József 2002. MTBA – magyar nyelvű telefonbeszéd-adatbázis. *Híradástechnika* 8: 35–9.
Vicsi Klára 2010. Adatbázisok a beszédtechnológia szolgálatában. In: Németh Géza – Olaszy Gábor (szerk.): *A magyar beszéd – beszédkutatás, beszédtechnológia, beszédinformációs rendszerek*. Akadémiai Kiadó, Budapest, 262–32.
Wolfson, Nessa 1976. *Speech events and natural speech: Some implications for sociolinguistic methodology*. Cambridge University Press, Cambridge.

Gósy Mária – Gyarmathy Dorottya
MTA Nyelvtudományi Intézet

SUMMARY

Gósy, Mária – Gyarmathy, Dorottya

Large, multifunctional Hungarian speech database (BEA)

The ten-year development of the multifunctional BEA spontaneous speech database is noteworthy both nationally and internationally. This database contributes to the enlargement of the national cultural heritage: the preservation of the speech production and speaking styles of Hungarian-speaking adults of Budapest. By now, the database consists of a 367-hour recorded speech material produced by 461 adult speakers of ages between 18 and 90, recorded under the very same conditions. There are various descriptions of the speech materials including annotations of 139 subjects within the Praat software. There are numerous studies based on the speech materials of this database. Investigations concerned the acoustic-phonetic characteristics of the speech sounds, co-articulation processes, the phonology-phonetics interface, phrase-final lengthening, hiatus, speech melody, as well as various disfluency patterns and self-repair strategies in spontaneous utterances. Researchers focused also on the interrelations between speech planning and pronunciation that made it possible to describe spontaneous speech processes. The objective analysis of the speech production variability provides new ways to investigate individual characteristics of speech and language use in general.

Keywords: spontaneous speech, database, uniform protocol, text transcripts, database-based research

Idegen eredetű, jelölt mássalhangzó-kapcsolatra végződő tövek tárgyesete a magyar nyelvben

1. Bevezetés, témamegjelölés

A magyar anyanyelvi beszélők jelentős hányada[1] az idegen eredetű, erősen jelölt mássalhangzó-kapcsolatra végződő főnevek tárgyesetét kötőhangzó betoldása nélkül teszi tárgyesetbe (pl. *Tom Hankst, fájlt*). Ez a jelenség azért is meglepő, mert a kevésbé jelölt mássalhangzó-kapcsolatra végződő tövek tárgyesetében megjelenik a kötőhangzó (pl. *tapsot, bokszot*), éppen ezért elvárható volna a kötőhangzó megjelenése a jelöltebb mássalhangzó-kapcsolatok és a tárgyrag között.

Dolgozatomban tisztázom a jelöltség és jelöletlenség fogalmát, kitérek a szótagszerkezet felépítésére, a szonoritási sorba rendezésre, majd a vizsgált jelenséget optimalitáselméleti keretben kísérlem meg megválaszolni.

2. Jelöltség, szótagszerkezet, szonoritás

2.1. Jelöltség

Mindenekelőtt tisztázzuk a bevezetésben használt jelöltség fogalmát. A jelöltség-jelöletlenség terminuspár arra vonatkozik, hogy egy nyelvi egység megformálási módja mennyiben felel meg vagy éppen tér el bizonyos univerzális, tipológiai és/vagy rendszerszintű szabályoktól vagy elvektől (vö.

[1] A Google kereső találatainak száma alapján, lásd később.