

Gósy Mária

A beszédatbázis jelentősége a beszédkutatásban

Bevezetés

A beszéd nagyon sok szempontból tanulmányozható, hiszen maga a kutatás tárgya igen összetett folyamat. Minél több ismeretünk van a hangzó nyelv sajátosságairól, annál több olyan új kérdés merül fel, amelyeknek a megválaszolása csak kísérleti úton, objektív vizsgálatokkal lehetséges. A beszédkutatásnak nagy hagyomány van, kiváló kutatók és munkáik sora jelzi a legjelentősebb állomásokat. Az utóbbi évtizedben középpontba került a spontán beszéd elemzése, amelyhez nagy mennyiségű, megfelelően gyűjtött és rögzített beszédanyag szükséges. A technika, illetve a technológiai háttér ezt az igényt mindeddig csak igen korlátozottan tudta kiszolgálni.

Előzmények

A nyelvjárási hanganyagok gyűjtésének igénye a XIX. század elejére tehető. Már Bartók Béla és Kodály Zoltán népdalfelvételein is vannak olyan töredékek, amelyek nyelvjárási adatokat tartalmaznak. Az első világháború idején, az akkori Fonetikai Laboratórium vezetője, Balassa József irányításával rögzítettek beszédet (hordozható gramofonnal viaszhengerekre); feldolgozásuk azonban a háború alatt nem indulhatott meg, és akkor vagy később valószínűsíthetően megsemmisültek. A XX. század negyvenes éveiben Hegedűs Lajos fonetikus kezdeményezésére indult el a nyelvjárási hangfelvételek rendszeres készítése. A kutató célja az volt, hogy az ország különböző helyein, különböző nyelvjárásterületeken rögzítsen beszédet, mesemondást, ráolvasásokat stb. Az első hangfelvétel 1940 augusztusában, Drávacsehin, Baranya megyében készült. Hegedűs Lajos és munkatársai mintegy 1700 hangfelvételt rögzítettek (840 üveg- és alumíniumalapú decilith- és lakklemezre). A Hegedűs-archívum átmentése korszerű adathordozóra a XXI. század fordulóján megtörtént. Ennek a rendkívül értékes anyagnak csak igen kis része van lejegyezve. Különféle (nyelvészeti, fonetikai, sőt néprajzi indíttatású) kutatások készültek, főként az utóbbi évtizedben, az archívum hanganyagai alapján.

A XX. század nyolcvanas éveiben kezdték fejleszteni a Budapesti szociolingvisztikai interjú (BUSZI) elnevezésű korpuszt, amely spontán beszédet (is) tartalmaz (1). A korpuszban 1987 és 1989 között, 250 beszélővel, külső terepen magnetofonra felvett, egyenként 2–3 órás interjú található; ezek számítógépes lejegyzése és kódolása is megtörtént. Különböző jellegű spontán beszédanyagokat rögzítettek továbbá – eltérő módszerekkel és változó célokkal –, amelyek kiválóan használhatók részkutatásokra (például nyelvjárási beszélők anyagai, osztálytermi felvételek, narratívák, társalgások). Ezeket az anyagokat többnyire külső helyszíneken rögzítették, nem is lehetett cél a felvételek technikai állandósága, és az írásos megjelenítésük (lejegyzésük) szempontjai is nagyon változatosak voltak. Az elmúlt évtizedben különféle beszédtechnológiai célokra olvasott beszédre alapuló korpuszokat (hosszabb szövegek felolvasását, szó- és mondatlistákat) is létrehoztak.

A korszerű beszédkutatás

A számítógépes technológia következtében már nincs akadálya a nagyméretű beszédatbázisok létrehozásának és egyszerű kezelésének, ezáltal olyan kutatások megvalósításának, amelyek korábban nem

voltak lehetségesek. Napjainkban már több nyelven elérhetők különféle jellegű hangzó nyelvi adatbázisok, amelyek természetesen módszertanilag is nagy változatosságot mutatnak. A nyelvészet számos területén fokozódik az igény arra, hogy a nyelv valós használatát elemezzék és írják le. A hatalmas adatmennyiség feldolgozása bizonyos szemléletbeli változást is igényel.

A fentiekben vázlatosan ismertetett előzmények megalapozták, az új kutatási célok pedig szükségessé tették egy korszerű, a beszédkutatás új feladatait kielégítő, azoknak megfelelő, modern technológián alapuló, többfunkciós beszédatadbázis létrehozását. A korábbi spontán beszédanyagok gyűjtésével és rögzítésével szerzett tapasztalatokkal az MTA Nyelvtudományi Intézet Fonetikai Osztályának munkatársai – Beke András, Gósy Mária, Grácsi Tekla Etelka, Gyarmathy Dorottya, Horváth Viktória, Neuberger Tilda és Nikléczy Péter – 2007 őszén kezdték a BEszélt nyelvi Adatbázis (BEA) tervezését (2). Meghatározták a beszédtudományos célokat, a majdani feladatokat, megtervezték a korszerű felvételi (technikai) körülményeket, a beszélő személyek kiválasztásának a kritériumrendszerét, a felvételi protokollt, valamint a hangzó anyagok lejegyzési stratégiáit. Az adatbázis tartalmának (protokolljának) a megtervezésekor tekintetbe vették a nyelvészeti és a határterületi kutatási területek szempontjait, és bizonyos mértékig érvényesítik (noha ez nem cél) a szociológiai tényezőket. A munkálatokba már a kezdetektől bevontak egyetemi hallgatókat és doktoranduszokat, akik munkájukkal és tapasztalataikkal nagyban segítették és segítik folyamatosan az adatbázis fejlesztését.

A próbamunkákat követően 2008-ban kezdődtek meg a beszédfelvételek. A távlati cél 500 személy beszédének a rögzítése, amelynek során a fejlesztők tekintetbe veszik (mások mellett) a női és a férfi beszélők arányát, az életkort, valamint az iskolázottság szerinti megoszlást. A BEA teljes rögzített anyaga jelenleg közel 265 óra, az adatközlők száma 305. A legrövidebb felvétel 24 perc 27 másodperc, a leghosszabb pedig 2 óra 24 perc és 47 másodperc.

Az adatbázis hanganyagai és rögzítésük

Az adatbázis protokollja hat részből áll, a felvételi sorrendben ismertetjük őket. 1. A mondatismétlés anyaga 25 egyszerű és összetett mondat, amelyeket az interjúvezető olvas fel az adatközlőnek, akinek a meghallgatás után azonnal meg kell ismételnie őket. 2. A narratívák az adatközlő életéről, családjáról, munkájáról, hobbjáról szólnak, többé-kevésbé összefüggő monologikus szövegek. 3. A véleménykifejtés (amely nagyrészt szintén narratíva) egy az interjúkészítő által megadott, éppen aktuális témának a véleményezése (például a jogosítvány megszerzése, a házassági szerződés, a klímaváltozás, a tanárok elleni erőszak, a közlekedés, az otthonszülés, az internetes és a hagyományos könyvtár, az állatvédelmi törvény, a mobiltelefon kisgyerekeknek, az olvasási szokások, a hitelfelhalmozás). 4. A tartalomösszegzés voltaképpen irányított spontán beszéd. Az adatközlő felvételről meghallgat két szöveget, és a saját szavaival el kell mondania mindkettőnek a tartalmát. 5. A társalgásban az adatközlőn és az interjúkészítőn kívül egy további személy vesz részt. A téma változó, az élet mindennapjaihoz kapcsolódik (például szilveszter, esküvői élmények, álláskeresés, házasság vagy együttélés, állattartás a lakásban, színházi élet, a diákok jogai, nők és karrier, gyermekvállalás). 6. Végezetül a protokollban az adatközlő kétféle szöveget olvas fel. Az egyik a huszonöt, korábban ismételt mondat, a másik egy tudománynpszerűsítő cikk.

A felvételek mindig azonos helyen és technikai körülmények között készülnek, az MTA Nyelvtudományi Intézet Fonetikai Osztályának saját tervezésű, zajszigetelt szobájában. Minden egyes felvétel esetében dokumentáljuk az adatközlő életkorát, iskolai végzettségét, foglalkozását, termetét, súlyát, esetleges beszédhibáját, azt, hogy dohányzik-e, valamint a spontán beszéd témáit. A névtelenség alapvetően fontos és szigorúan betartott kritérium, az anonimizálás a felvételt követően azonnal megtörténik, így a beszédanyag nem kapcsolható össze a beszélő nevével. A beszédanyagokat csak kutatási és oktatási célokra használjuk fel.

A spontán beszédet tartalmazó adatbázisok esetében felmerülhet a természetesség kérdése. Előfordul, hogy a természetességet és a spontaneitást azonosnak vélik, ezért a megítélés félrevezető lehet. A beszéd spontán volta nem sérül akkor, ha az adatközlő egy mesterségesen létrehozott kommunikációs helyzetben beszél. A rögzített beszéd egyértelműen spontán akkor, ha a beszélőnek a feltett kérdés(ek)re azonnal kell válaszolnia, azaz megelőzően nem készülhetett fel az adott témára. A természetesség nehezebben határozható meg, és számos tényezőtől függ, például a beszélő személyiségétől. Tapasztalataink szerint a BEA adatközlői igen gyorsan alkalmazkodnak a helyzethez, többségüket nem zavarja a hangfelvétel; beszélőink az adott körülmények között természetes módon beszélnek.

Az adatbázis írásos megjelenítése

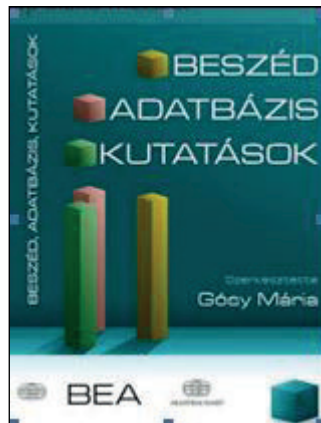
A BEA anyagainak lejegyzése többféleképpen történik (egyszerű átírás helyesírásban, különféle annotálások szoftverek használatával). Ez lehetőséget ad arra, hogy a kutató a számára legmegfelelőbbet választhassa ki. A lejegyzés nagyon időigényes munka. Egypercnyi elhangzó szöveg lejegyzéséhez – a szöveg típusától, az adatközlő beszéd-sajátosságaitól, a lejegyző gyakorlottságától, valamint a leírás szempontjaitól (részletességétől) függően – átlagosan tíz percre van szükség. 2013-ban a Fonetikai Osztály OTKA-pályázatot nyert el, amelynek keretében folytatódik az adatbázis fejlesztése, egyszerűbbé és közvetlenebbé válik a használata a kutatók számára. Célkitűzésünk az is, hogy 120 felvétel beszédhangszintű lejegyzése történjen meg egy fonetikai szoftver segítségével, amely nagymértékben pontosítja majd a további elemzéseket, és lehetővé teszi a részletező munkákat.

Kutatások az adatbázis felhasználásával

A BEA főbb jellemzőinek bemutatása szemlélteti, hogy az adatbázis a nyelvészet számos területén kínál tanulmányozásra alkalmas anyagot. Csupán felsorolásszerűen néhány ezek közül:

- a beszédhangok képzésének fonetikai jellemzői,
- a magánhangzók és mássalhangzók egymásra hatása,
- a hangszalagműködés kommunikációs funkciói,
- a szövejtés sajátosságai,
- prosódiai tényezők: beszéddallam, hangsúlyozás, hangszínezet stb.,
- a beszédtempó alakulása,
- a beszédfolyamat tagolása,
- a társalgások több szempontú elemzése,
- pragmatikai megközelítések,
- a beszédalkalmazkodás tanulmányozása,
- fiatalok, középkorúak és idősek beszéde,
- különböző beszédstílusok hatása a kiejtésre (például olvasás, narratíva, társalgás),
- a megakadásjelenségek, nyelvbotlások sajátosságai,
- a beszédhangok, szavak gyakorisága,
- a beszélők egyéni kiejtési jellemzői és még sorolhatnánk.

2012-ben jelent meg az Akadémiai Kiadó gondozásában a *Beszéd, adatbázis, kutatások* című kötet, amelynek minden egyes tanulmánya az adatbázison alapszik; a könyv végén pedig a korábban megjelent tanulmányok felsorolása olvasható (Gósy 2012).



1. ábra
Beszéd, adatbázis, kutatások című kötet

A nagy mennyiségű rögzített anyag kiindulást és elemzési adatokat nyújthat a nyelvi norma kutatásához, továbbá számos beszédtechnológiai alkalmazáshoz. A tipikus beszélők közléseinek elemzési eredményei felhasználhatók az atipikus, valamint a kóros beszéddel, illetve nyelvhasználattal történő összevetésre is.

A beszédatadbázis jelentősége

A BEA adatbázis a nyelvészet számos területén, a határterületi tudományokban és az alkalmazott kutatásokban is használható, mindez önmagában óriási jelentőséggel bír. Nagyságát és jellemzőit tekintve nemzetközileg is jelentős (több megkeresésünk van külföldi kutatóhelyektől). A jól megtervezett és kivitelezett, írásban is megjelenített és lekérdezhető adatbázisok kiváltják az időigényes felvételek készítésének a munkáját, hatalmas adathalmazt biztosítanak, és a nyelv valós használatát tükrözik. A BEA hanganyagai az ezredforduló (fővárosi) magyar beszédének rögzítésével maradandó értéket és nemzeti kincset képviselnek.

Irodalom

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

(1) Budapesti Szociolingvisztikai interjú. <http://www.nytud.hu/buszi/> (2013. október 5.)

(2) BEszélt nyelvi Adatbázis. <http://www.nytud.hu/adatb/bea/> (2013. október 5.)

Gósy, Mária: The importance of a speech database in speech research

Az írás szerzőjéről

Gósy Mária

Eötvös Loránd Tudományegyetem, Budapest, Magyarország; MTA, Budapest, Magyarország

gosity.maria@nytud.mta.hu

<https://orcid.org/0000-0003-4336-3007>