Záróbeszámoló: Magyar spontán beszéd adatbázis
78315 sz. OTKA

A pályázat terve

A kutatás célja egyfelől egy spontán beszéd adatbázis létrehozása volt (különféle típusú beszédanyagokkal), másfelől pedig kutatások végzése az adatbázis felhasználásával. A kutatás három évére terveztük összesen kilencven adatközlő beszédanyagának rögzítését, digitalizálását, a hanganyagok hangzáshű átírását, számítógépes rögzítését, archiválását; valamint különféle aspektusú fonetikai kutatások végzését és tanulmányok publikálását az eredményekről. A feladatainkat teljesítettük, a vállalt célokat megvalósítottuk.

A számítógépes technológia segítségével létrehozható nagyméretű beszédadatbázisokat a fonetika harmadik forradalmának is nevezik a hangszínképelemzés és a számítógépes beszédelemző programok után. Számos, különféle méretű és típusú beszédadatbázist hoztak létre a világ nyelvein; a korábbi magyar nyelvi anyagot tartalmazó korpuszok, gyűjtemények, főként olvasott beszédet tartalmaztak. A jelen kutatásban tervezett beszélt nyelvi adatbázis a fonetikai és a tágabb értelemben vett nyelvészeti elemzésekre is alkalmas beszédanyag kialakítását tűzte ki célul.

Eredmények

Az adatbázis fejlesztését többrétű tervezés, a megfelelő szakirodalom, a más nyelvű adatbázisok tanulmányozása előzte meg. Kidolgoztuk a beszédrögzítés technikai körülményeit. Megterveztük az adatközlőkkel kapcsolatos munkálatokat, az azonosíthatatlanságukat biztosító folyamatokat (etikai vonatkozások). Meghatároztuk az adatbázis beszédtípusait, a témáit, a vonatkozó nyelvi anyagot, valamint a protokoll módszertanát. Megállapítottuk a lejegyzések kritériumrendszerét és az átírási szabályokat. Megvalósítottuk az archiválást. Számos kutatást végeztünk az adatbázis felhasználásával.

1. Beszédfelvételek. A terveknek megfelelően megtörtént a beszédadatbázis állandó felvételi körülményeinek kialakítása. A helyszín minden esetben az MTA Nyelvtudományi Intézet Fonetikai Osztályának csendesített stúdiószobája volt, valamennyi adatközlő felvételét itt valósítottuk meg. Biztosítottuk a hangrögzítéshez a korszerű technikai hátteret (AT 4040 kondenzátor stúdió mikrofon, Goldwave hangfelvevő program használata, 44 kHz/16 bit mintavételezéssel). A technikai körülmények megfeleltek a nemzetközi elvárásoknak, illetve sok hasonló adatbázis felvételi körülményeit felülmúlták a tekintetben, hogy a helyiség és a technikai körülmények állandóak voltak. Megterveztük a lehetséges adatközlők felkutatását és részvételük megszervezését a felvételekhez. Ez azt is jelentette, hogy az interjúkészítőn kívül a társalgásban részt vevő harmadik személy jelenlétét is biztosítottuk. A kilencven felvétel mintegy 90%-ában ugyanaz a személy volt az interjúkészítő (ez lehetőséget nyújt különféle specifikus nyelvi, fonetikai elemzésekre is). A terveknek megfelelően elkészült a kilencven hangfelvétel.

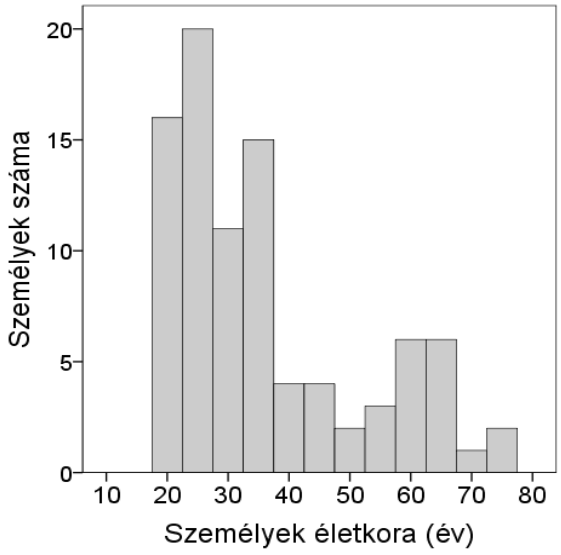
Megtörtént a beszélőkre vonatkozó adatok anonimizálása (a Nyelvtudományi Intézet etikai szabályzatának betartásával), aminek következtében az adatközlő kiléte és a rögzített hanganyag nem azonosítható. Az egyes hangfájlok kódolása tartalmazza a hanganyag sorszámát, valamint a beszélő nemére utaló rövidítést, illetve az ennek megfelelő sorszámot. Például: bea006f002 vagy bea052n037 (amelyek feloldása: az adatbázis hatos számú adatközlője, a második férfi beszélő, illetve az adatbázis ötvenkettedik adatközlője, a harminchetedik női beszélő).

Megvalósítottuk a rögzített hanganyagok többszörös archiválását (különböző adathordozókra: PC HDD, külső HDD, CD-ROM, DVD-ROM, pendrive). Megterveztük és elvégeztük a hanganyagok lekérdezhetőségét protokollegységek (lásd 4. pont) szerint. Ez azt jelenti, hogy minden adatközlő esetében a lekérdezhetőség biztosításával külön-külön rögzítettük a teljes felvétel egyes részeit (beszélőnként hat rész). Ezek külön-külön fájlként mentve is archiválva lettek. Ez arra ad lehetőséget, hogy a kutató mindig az általa kiválasztott (tematikus) beszédanyagokat kaphassa meg.

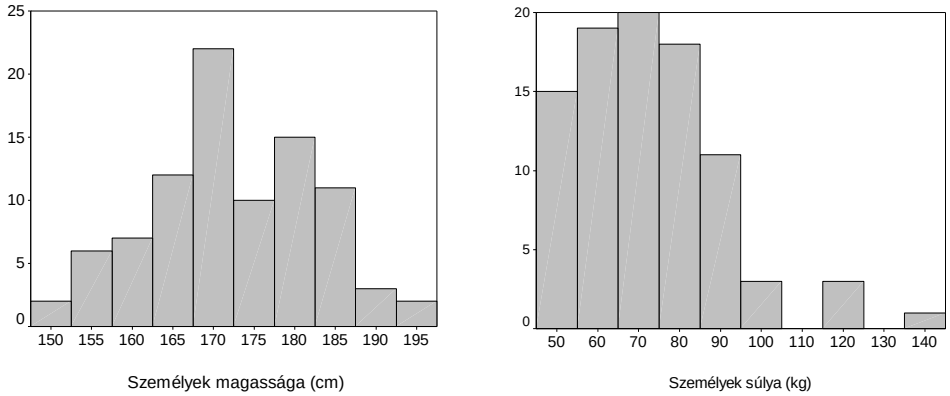
A teljes beszédanyag időtartama: 89 óra 9 perc 59 másodperc. A legrövidebb felvétel 26 perc, 10 másodperc, míg a leghosszabb 1 óra, 30 perc, 41 másodperc hosszúságú.

2. Adatközlők. A 90 adatközlő közül 50 nő, 40 férfi, életkoruk 20 és 80 év közötti (1. ábra). A legfiatalabb beszélő 20, a legidősebb 73 éves volt a felvételkor. Szociológiai szempontok érvényesítése nem volt célja a kutatásnak, ezért csupán megemlítjük, hogy a 90 adatközlő közül kettő általános iskolai végzettséggel, 39 érettségivel és 49 valamilyen felsőfokú végzettséggel rendelkezik. Az életkoron túl további adatokat is rögzítettünk az adatközlők válaszai alapján, amelyek a beszédprodukcióval kapcsolatba hozhatók, így a későbbi kutatások szempontjából releváns információkat jelenthetnek. Rögzítettük a beszélők magasságát, súlyát, valamint az, hogy dohányoznak-e (és ha igen, mióta). A 2. ábra az adatközlők termetének és súlyának a megoszlását mutatja. Kilencen évtizedek óta dohányoznak, tizenegyen csak néhány éve és hetvenen egyáltalán nem dohányoznak.

Az adatközlőkre, illetve a beszédanyaguk felvételére vonatkozó adatokat Excel-fájlokban összesítettük, ez a későbbi kutatási felhasználást nagymértékben megkönnyítette, az adatbázist áttekinthetővé tette. Az Excel-táblák a következő adatokat tartalmazzák: a beszélő kódját (amely tartalmazza a nemét), de külön oszlopban is feltüntettük, hogy női vagy férfi beszélő-e, az adatközlő életkorát, magasságát, súlyát, dohányzási szokását, valamint a spontán beszéd felvételeinek témáit (lásd 3. pont).



1. ábra
Az adatbázisban rögzített 90 beszélő életkora



2. ábra
Az adatbázisban rögzített 90 beszélő magassága és súlya

3. A spontán beszéd témakörei. Összeállítottuk azokat a témaköröket (életkori bontásban is), amelyek felhasználhatók a spontán beszéd rögzítésekor. Olyan témákat kellett választanunk, amelyek különösebb ismereteket nem igényelnek az adatközlőtől, alkalmasak hosszabb kifejtésre vagy megbeszélésre, és lehetőség szerint igazodnak az adatközlő életkorához (esetenként ahhoz is, hogy női vagy férfi beszélő). Meg kellett határozni azokat a témaköröket, amelyek elsődlegesen a véleménykifejtés, illetve azokat, amelyek főként a társalgás témájaként megfelelőek.

Néhány példa a témák közül a véleménykifejtéshez: jogosítvány megszerzése, házassági szerződés, klímaváltozás, tanárok elleni erőszak, budapesti közlekedés, otthonszülés, internetes és hagyományos könyvtár, állatvédelmi törvény, mobiltelefon kisgyerekeknek; illetve a társalgáshoz: álláskeresés, dohányzás, házasság vagy együttélés, színházi élet, gázkrízis Európában, válság hatása a kultúrára, metróépítés, diákok jogai, nők és karrier, gyermekvállalás, kerékpározás mint közlekedési forma.

4. A felvételi protokoll. Az adatbázis felvételeinek protokollja 6 részből áll. Ezek a következő címkékkel jellemezhetők: narratíva, véleménykifejtés, tartalomösszegzés, mondatismétlés, társalgás, felolvasás.

a) A narratívák az adatközlő életéről, családjáról, munkájáról, hobbijáról szólnak, többé-kevésbé összefüggő monologikus szövegek. Cél, hogy az adatközlő minél hosszabban beszéljen anélkül, hogy az interjúkészítőnek meg kelljen szólalnia. Az interjúkészítő csak akkor kérdez, ha a beszélő jelzi, hogy befejezte.

b) A véleménykifejtés egy az interjúkészítő által megadott téma véleményezése (a példákat lásd a 3. pontban). Ez a spontán beszéd is tartalmaz narratívaszerű részeket, de jellemzők a dialógusok, továbbá a nézetkülönbségek vagy a nézetazonosságok verbális megjelenései.

c) A tartalomösszegzés voltaképpen irányított spontán beszéd. Előkísérletek sorozatában választottuk ki azt a két szöveget, amelyek alkalmasnak bizonyultak ennek a verbális feladatnak a teljesítésére. Az egyik szöveg egy tudománynépszerűsítő cikk (174 szóból áll), a másik egy történelmi anekdota (270 szóból áll). Rögzítésük átlagos női beszélővel történt, a felolvasás instrukciója az volt, hogy értő-értető olvasás legyen (nem előadói teljesítmény). Az adatközlőnek meghallgatja az adott szöveget, majd ezt követően a saját szavaival összegzi a történetet (monológ formájában).

d) A mondatismétlés anyaga 25 egyszerű és összetett mondat (pl. *A farsangi bálban mindenkinek szép jelmeze volt*. vagy *Nem kötött biztosítást, ezért kisebb vagyonba került a kórházi ellátás*.). Az előkísérletek eredményei alapján alakult ki, hogy melyek azok a mondatok, amelyeknek a tartalma és a morfológiai, szintaktikai szerkesztettsége alkalmas az

egyszeri meghallgatás utáni visszamondásra. A mondatokat az interjúvezető olvassa fel az adatközlőnek (átlagos tempóban és hangerővel), akinek egyszeri meghallgatás után azonnal meg kell azt ismételnie.

e) A társalgás három személy beszélgetése, egyikük az adatközlő, a másikuk az interjúkészítő, a harmadik személy pedig a Fonetikai Osztály fiatal kutatóinak egyike. A témák változók (lásd a 3. pontban). A társalgásokban mindhárom beszélő azonos kommunikációs eséllyel vesz részt, ami egy kvázi-természetes helyzetet eredményez.

f) Kétféle felolvasás szerepel a protokollban a későbbi összevethetőség érdekében. Az egyik felolvasáskor az adatközlőnek a 25, korábban elismételt mondatot kell meghangosítania, majd pedig a már hallott tudománynépszerűsítő cikket felolvasnia.

5. A beszédanyag átiratai. A hanganyagok átírása nagy jelentőségű a beszédadatbázisokban. A terveknek megfelelően kialakítottuk az elsődleges átírás kritériumait, megterveztük a lejegyzési útmutatót, megtörténtek a próbaátiratok. Ezek tapasztalatai alapján korrigáltuk a kritériumrendszert, és módosítottuk az útmutatót. Ez az elsődleges átírás helyesírás alapú, központozás nélkül, amely számos, a további kutatások számára hasznos hangzási tükröztetést tartalmaz. Az elsődleges átírás során a lejegyzők a Microsoft Office Word programot használták (.doc formátum). A lejegyzési útmutató szabályozza a köznyelvben használatos, de nem szótári alakban előforduló szavak, az idegen szavak, a rövidítések, a betűszók és mozaikszók, illetőleg a lejegyző számára értelmezhetetlen szóalakok lejegyzésének módját. Az átiratok készítése és ellenőrzésük folyamatosan zajlott.

A beszédanyag hangzáshű leírása idő- és energiaigényes feladat (még jó felvételi körülmények között rögzített beszéd esetén is). Egyetlen percnyi beszéd átírása 15–20 percet vesz igénybe az időadatok mérése és a másodlagos (külső) ellenőrzés nélkül. A leírás időtartama függ az adatközlő és a többi beszélő kiejtésétől, a megakadásjelenségek és egyéb nonverbális jelzések mennyiségétől és típusaiktól (ezek lejegyzése tipográfiailag van kódolva a szövegekben), az egyszerre beszélések arányától stb. További, felhasználóspecifikus átírásokat is alkalmaztunk, amelyek mind a fonetikai, mind a beszédtechnológiai alkalmazásokban hasznosak. Kialakítottunk egy a Praat (ingyenesen használható) szoftver fájlformájában (TextGrid) archivált, többszintű annotációt a saját kutatásokhoz, valamint egy szakaszszintű annotálást a Transcriber (ingyenesen használható) program felhasználásával. Az átírásnak ez a két formája lehetővé teszi, hogy a beszédszövegek mellett a hangzás akusztikai lenyomata egyidejűleg megjeleníthető legyen. Mindkettő felhasználóbarát grafikus felülettel rendelkezik, és többféle platformon (Windows, Unix, Mac) alkalmazható. Az átiratok ellenőrzése nagymértékben csökkenti az elsődleges lejegyző relatív szubjektivitását.

6. A természetesség kérdése. A spontán beszédet tartalmazó adatbázisok esetében gyakorta felmerül a természetesség kérdése, vagyis az, hogy a beszélő mennyire viselkedik a közléseit tekintve természetesen egy ilyen nem szokásos beszédhelyzetben (ún. „megfigyelői paradoxon” problémája). A megfigyelői paradoxon áthidalásának egyik lehetséges módja az, ha az adatközlőtől többféle típusú beszédet rögzítünk. Ez a kritérium a jelen adatbázis esetében teljesült. Növeli a helyzet természetességét, ha olyan körülményeket tudunk teremteni, amelyben a beszélő szorongása oldódik, és a mikrofon okozta fokozott önellenőrzés megszűnik. Némely esetben az adatközlőink egy része a felvétel előtt bevallotta, hogy izgul, azonban szinte kivétel nélkül hamar feloldódtak. Ennek nyelvi igazolása az (is), hogy néha olyan szavakat és nyelvi fordulatokat használtak, amelyek nem illettek egy formális verbális kommunikáció kereteibe. Adatközlőink a körülményekhez képest természetes módon beszéltek.

7. Kutatások az adatbázison. Az elmúlt három év alatt különféle kutatásokat végeztünk a rögzített beszédanyagok felhasználásával. A zárójelentés megírásáig a pályázatban résztvevők összesen 31 tanulmányt publikáltak, 24-öt magyar és 7-et angol nyelven (utóbbiak közül öt ERIH A, B kategóriás folyóiratban jelent meg). További öt tanulmányunkat fogadták el publikálásra, közülük egy angol nyelvű, ezek megjelenése a 2012-es évben várható. Újabb kutatásainkat, amelyeket az adatbázis felhasználásával végzünk, egy önálló kötetben szándékozunk megjelentetni, várhatóan ebben a naptári évben. Szóbeli előadásra fogadták el a Clinical Linguistics and Phonetics nemzetközi kongresszusra (Cork, 2012) a spontán beszéd hezitálási jelenségei alapján benyújtott előadásunkat.

A kutatott témák nagy vonalakban az alábbiak voltak: a magyar magánhangzók megvalósulásainak akusztikai-fonetikai vizsgálata, a zöngésségi hasonulás nyelvfüggő fonológiai szabályának érvényesülése olvasott és spontán beszéd összevetésével, a beszédalkalmazkodás kutatása a részt vevő partnerek ismeretségének és a feltételezett összeszokásnak a függvényében, egyetlen szó előfordulásainak akusztikai-fonetikai sajátosságai a spontán beszédben, a bizonytalanságot jelző és a hiba típusú megakadásjelenségek vizsgálata, a spontán beszédben előforduló magánhangzók időtartamai, a zöngétlen résmássalhangzók akusztikai szerkezete, a temporális variabilitás, a beszédhangok, hangkapcsolatok és szavak gyakorisági mutatóinak vizsgálata narratívákban, a lexikális hozzáférés sajátosságai, töltelékszók funkcionális vizsgálata.

A jelen adatbázis a nyelvészet számos területén kínál tanulmányozásra alkalmas anyagot; lehetőség nyílik a lexikális hozzáférés folyamatainak elemzésére, a feltételezett szinkrón nyelvi változások igazolására, a szintaxis, a szemantika, illetve a prozódia összefüggéseinek, a nyelvi normativitás kérdéseinek az elemzésére (stb.). A beszédadatbázis jól felhasználható beszédtechnológiai kutatásokban. A tipikus beszélők anyagai támpontul szolgálhatnak az atipikus beszéd tanulmányozásához is.

A pályázat két résztvevője (Horváth Viktória és Gyarmathy Dorottya) sikeresen megvédte PhD értekezését, amelyek anyagául az adatbázis szolgált. A pályázat egy további résztvevője (Gráczi Tekla Etelka) benyújtotta PhD értekezését, amelynek egy fejezetében az elemzések az adatbázis anyagán történtek.

8. Az adatbázis hozzáférhetősége. A fentiek tükrében fontosnak tartottuk, hogy az adatbázis csaknem a kezdetektől hozzáférhető legyen – a megfelelő etikai szabályok szigorú betartásával és betartatásával. Eddig mintegy 30 tanulmányt írtak a projektben nem részt vevő külső kutatók és doktoranduszok az adatbázis anyagainak felhasználásával. Minden ilyen publikált tanulmány módszertani része tartalmazza az utalást az adatbázisra.

Kitekintés

A magyar spontán beszéd adatbázis közvetlen tudományos értékén és hasznán kívül számos gyakorlati alkalmazáshoz is közvetve vagy közvetlenül nélkülözhetetlen. Úgy ítéljük meg, hogy társadalmi szempontból is meghatározó jelentőségű, nemzetközi tekintetben is a legértékesebbek egyike. Az adatbázis beszédanyagának fontos hozadéka, hogy az ezredforduló (fővárosi) magyar beszédének rögzítésével kulturálisan is maradandó értéket képvisel.