

Sejtes Györgyi^x-Bácsi János^x- Kocsor András*

^xSzegedi Tudományegyetem Juhász Gyula Tanárképző Főiskolai Kar
Gyakorló Általános és Alapfokú Művészeti Iskola

E-mail: sejtes@hung.u-szeged.hu

E-mail: bacsi@jgytf.u-szeged.hu

*MTA-SZTE Informatikai Tanszék Mesterséges Intelligencia Kutatócsoport

E-mail: kocsor@inf.u-szeged.hu

Egy nem csak fonológiai adatbázis specifikációja

1. Bevezető

A Beszédmester (továbbiakban: BeMe) olyan gépi tanuláson alapuló szoftver, amely lehetővé teszi a részképességükben sérült és az olvasási nehézségekkel küzdő gyerekek beszédjavítás-terápiáját, ill. olvasásfejlesztését. A géppel segített olvasásfejlesztés alapvető célja, hogy a gyermek a számítógépet motivációs eszközként használva minél könnyebben és gyorsabban megtanulja a fonéma-graféma, ill. graféma-fonéma megfeleltetéseket. Célunk egyfajta fonológiai tudatosság kialakítása a felhasználóban, amely valós idejű, gépi fonéma-felismerésen és fonémaazonosításon, ill. a felismerés eredményének vizualizációján alapul.

A BeMe hanganyaga korlátozottan elérhető, az adatbázis struktúrájának specifikációja nyilvános. Ennek célja, hogy az adatbázis felhasználható legyen más kutatásokban pl.: a gépi beszédfelismerés más alkalmazásaiban, és nyelvészeti (fonetikai, fonológiai, szociolingvisztikai...) kutatásokban.

Az adatbázisban megtalálható a digitálisan tárolt hanganyag. A hullámforma mellett a hozzájuk tartozó ortografikus karaktereket is rögzítettük. A teljes hanganyagot annotáltuk és szegmentáltuk.

Jelen cikkünkben bemutatjuk az adatbázis létrejöttének körülményeit, felépítését, az adatok feldolgozásának módját. Munkánkkal szeretnénk megkönnyíteni más kutatások számára is az adatbázis használatát.

1.2. A szoftverkészítés részterületei

A munka összetett volt, mert több tudomány (informatika, pedagógia, nyelvészet) szakemberei dolgoztak együtt azért, hogy a BeMe szoftver az alkalmazás során jól működjön, hatékonyan segítse a gyerekek az olvasástanítását, ill. a hallássérültek a beszédoktatását.

1.2.1. Informatikai feladatok

Az SZTE Informatikai Tanszékcsoportja olyan alapkutatási és szoftverfejlesztési munkát végzett beszédfelismerés és gépi tanulás témakörben, amelyre a BeMe informatikai háttere is épül. Céljuk olyan gépi tanuló algoritmus kifejlesztése volt, amely lehetővé teszi a hatékony klasszifikációt, a beszéd alapegységeinek megkülönböztetését.

1.2.2. Pedagógiai és nyelvészeti feladatok

A statisztikai alapon működő módszer alkalmazásához nagy méretű beszédatadabázisra volt szükség. Az SZTE JGYTF-i Kar Gyakorló Általános Iskolájának kutatócsoportja (nyelvészek, pedagógusok) megtervezte a hanganyagot, elkészítette a gyakorisági szótárat, rögzítette és feldolgozta a hanganyagot, ill. elkészítette az adatbázis specifikációját. Össze állították a hívóképanyagot, kidolgozták a számítógéppel segített olvasástanítás metodikáját.

Az Óvoda Általános Iskola Diákotthon és Gyermekotthon (Siketek Iskolája, Kaposvár) a hallássérültek beszédoktatásának anyagából készített gyakorisági szótárat, hallássérült gyermekek hanganyagát rögzítette, didaktikát írt a végleges verzióhoz.

A BeMe grafikai megjelenítésének célja a program kiszolgálása volt. A grafikus olyan közérthető jelrendszert és színszimbolikát használt, amely az alkalmazás során a gyerekek érdeklődését ébren tartja, az esztétikai élmény mellett motivációként is szolgál.

2. Az adatbázis felépítése

Nagyméretű és áttekinthetően megszerkesztett szóadatbázis adott alapot a program háttérben működő valós idejű, gépi fonéma-felismerés kialakítására. Az olvasásfejlesztési rész szóadatbázisa 500 adatközlő 100-100 szószintű bemondását tartalmazza. A rögzített és feldolgozott szóadatbázist azért hoztuk létre, hogy a BeMe szoftver kifejlesztéséhez szükséges beszédtechnológiai kutatásokat szolgálja, ezért adatbázisunk a feladatnak megfelelően speciális.

2.1. Az adatbázis alapja (gyakorisági szótár)

A BeMe olyan beszélőfüggetlen szoftver, melynek célja, hogy a beszédképzési, ill. olvasási nehézségekkel küszködő gyerekek megfelelő fejlesztésben részesüljenek. Ezért az első lépés egy olyan szótár (Bácsi-Kerekes 2002) létrehozása volt, amelyet abból a 13 első osztályos olvasókönyv szóanyagából készítettünk, amelyet az Oktatási Minisztérium 2001/2002-ben engedélyezett. Ez a gyakorisági szótár tartalmazza a 13 első osztályos olvasókönyv teljes grafémasorát (27 293 grafémasor) csökkenő gyakoriság szerint. Szerepelnek benne olyan grafémakapcsolatok is, amelyek az olvasástanítás menetében nélkülözhetetlenek, de a magyar nyelv szókészletének nem elemei. Ebben a szótárban minden toldalékos szó önálló szótári alakként szerepel. A következő lépés az volt, hogy kiválasztottuk a tartalmasnak minősített szavakat (É. Kiss-Kiefer-Siptár 1998), így 12 226 szóalakot találtunk csökkenő gyakoriság szerint. Harmadik lépésként azt az 1953 szót különítettük el, amely legalább 10-es vagy annál nagyobb gyakorisági mutatóval fordul elő a második szótárban. Ez a szóanyag képezi a BeMe adatbázisának alapját.

2.2. Szóbemondó lapok

2.2.1. A közel 2000 szóból az olvasásfejlesztési részhez szerkesztettünk 500 szóbemondó lapot. Minden lap három részből állt.

Az első rész a 14 db magyar magánhangzót tartalmazta. (a, á, e, é, i, í, o, ó, ö, ő, u, ú, ü, ű)

A második rész 40 szóból állt. Ezek olyan tartalmas szavak voltak, amelyek betűrendben követték egymást. Az első és második rész minden adatközlőnél azonos volt.

(alma, ág, béka, cica, csiga, dió, dzé betű, dzsungel, egér, ég, fészek, golya, gyerek, hegy, iskola, ír, játék, kakas, lepke, lyuk, maci, nap, nyúl, oroszlan, óra, ördög, őz, pók, róka, süni, szőlő, tulipán, tyúk, uborka, út, üveg, úrhajós, virág, zokni, zsiráf)

A harmadik rész 60 szava minden adatközlőnél változott. A leggyakrabban használt szavakat osztottuk el előfordulási gyakoriságuk alapján, tehát egy-egy szó több lapon is szerepelhetett.

(fagy, fog, kis, vigyáz, megy, paripa, bábu, egészít, ér, név, apró, mosdik, pompás, zöld, tud, nemzet, gyakorol, menny, vég, tó, Lilla, nap, köt, lomb, sorrend, talál, kicsi, keretez, hangos, eső, kettő, Csaba, gyerek, tűz, kiadó, válasz, van, nagy, kezd, utca, fekszik, bogár, anya, illik, visz, szó, igaz, ablak, édesanya, teve, bújik, nagypapa, Dénes, szíves, réz, szabad, szőlő, jó, fül, esik).

2.2.2. A beszédjavítás-terápia szóanyagát a gyakorisági „szó-tár” leggyakoribb 2000 szava alkotta, amelyek tartalmaztak:

- magánhangzókat szókezdő pozícióban
- folyamatosan képzett mássalhangzókat szókezdő pozícióban
- magánhangzó-magánhangzó kapcsolatot az artikulációs sorban
- palatális-veláris oppozíciókat
- magánhangzókat szomszédos artikulációs pozícióban

- magánhangzókat spektrumuk hasonlósága szerint
- labiális-illabiális oppozíciókat
- azonos magánhangzókat rövid-hosszú oppozícióban
- 30 leggyakoribb difónt szókezdő pozícióban
- 30 leggyakoribb difónt szózáró pozícióban.

A beszédjavítás-terápia rész a hallássérültek beszédnevelésében előforduló 204 leggyakoribb szóból és 28 szintagmából állt, amely egy szűkített gyakorisági szótár szóanyaga; tartalmazott állathangokat, egy beszédérthetőségi vizsgálat szóanyagát, a szurdologopédiai automatizálás bázisszavait, magánhangzókat mássalhangzóval kitartott ejtésben és magánhangzósorokat folyamatos ejtésben a képzési sor szerint. A szóanyagot véletlenszerű keveréssel 40 sorba osztottuk. A sorok 4-5 szót, szókapcsolatot tartalmaztak.

2.3. Az anyaggyűjtés módszere (Kiss 2001)

Anyaggyűjtésünk, mint a regionális vizsgálatok többsége, a korpusznyelvészethez tartozik. Célja a nyelvi adatok segítségével fonémák, ill. szavak olyan pontosságú gépi felismertetése, hogy az a BeMe programban alkalmazható legyen a beszédterápiában és az olvasástanításban.

Az anyaggyűjtés professzionális, mert a terepmunkában (a felvételek készítésénél) szakemberek: nyelvészek és fejlesztő pedagógusok vettek részt.

A gyűjtés aktív módon történt, így csak azokat a nyelvi adatokat vettük fel, amelyekre szükségünk volt. Minden gyereknek a saját szólapján található 100 szót kellett beolvasni a számítógép mikrofonjába.

2.4. Az adatközlők kiválasztása

Adatközlőink 6-7 éves kisiskolások az ország 14 gyakorlóiskolájából. Az 500 gyerek (250 lány, 250 fiú) kiválasztásánál fő szempontunk az volt, hogy minél közelebb legyenek az olvasástanulás kezdeti szakaszához, mivel a BeMe majdani felhasználói ebbe a korosztályba tartoznak.

2.5. A hangfelvételek helyszínének kiválasztása

Úgy jelöltük ki a felvételek helyszínét, hogy az minél jobban lefedje a magyar nyelvjárási régiókat (Kiss 2001). Tettük ezt azért, hogy a BeMe szoftvert az ország bármely területén élő gyerek eredményesen tudja használni.

Munkánk során megtapasztaltuk az aktív módszer azon hátrányát, hogy a felvételek során az interjúhelyzet megnövelte a gyerekek nyelvi önkontrollját, s a formális beszédhelyzet, a magyar köznyelv kívánalmait szerint olvasták a szavakat.

Meg kell említenünk azt a nyelvi sztereotípiát is, amely csak erős megszorítással igaz ugyan, hogy Magyarországon a falusiak beszélnek nyelvjárásban, a városiak pedig köznyelven. A gyakorlóiskolákat úgy választottuk ki, hogy a az ország minél több régióját lefedje a vizsgálat. Készítettünk felvételeket lakótelepi, ill. belvárosi iskolában is.

A regionális nyelvhasználat és az életkor azon összefüggését is szem előtt kellett tartanunk, hogy az idősebbek jobban őrzik a nyelvjárási vonásokat, mint a hat év körüli kisiskolások.

A felvételek helyszíne

Város	Az iskola neve	Nyelvjárási régió	Régió belüli csoport
Szeged	Tabán Ált. Isk.	Dél-alföldi r.	Szeged környéki cs.
Szeged	SZTE JGYTFK Gyak. Ált. Isk.	Dél-alföldi r.	Szeged környéki cs.

Győr	A. Cs. J. Tank. Főisk. Gyak. Ált. Isk.	Közép-dunántúli- kisalföldi r.	Észak-dunántúli cs.
Pécs	Pécsi Tudegy. I.sz. Gyak. Ált. Isk.	Dél-dunántúli r.	Észak-baranyai cs.
Szekszárd	IPF Gyak. Ált. Isk.	Dél-alföldi r.	Baja környéki cs.
Kaposvár	K. E. Cs.V.M. Ped. Főisk. Gyak. Ált. Isk.és Gimn.	Dél-dunántúli r.	Közép-somogyi cs.
Baja	Eötvös J. Főiskola Gyak. Ált. Isk.	Dél-alföldi r.	Baja környéki cs.
Eger	Esterházy K. Főisk. 2.sz. Gyak. Ált. Isk.és G.	Palóc r.	Déli tömb Eger vidéki cs.
Nyíregyháza	Ny. F. Bt. és Műv. K. A. Cs. J. Gyak. Ált. Isk.	Északkeleti r.	Szabolcs-szatmári cs.
Sárospatak	M. Egy. Com. Tan. F. K. Á. J. Gyak. Ált. Isk.	Palóc r.	Keleti tömb Keleti palóc cs.
Debrecen	Kölcsey F. Ref.Tanítók. Főisk. Gyakorlóisk.	Tisza-Körös vidéki r.	Hajdú-bihari cs.
Békéscsaba	Békéscsaba Belvárosi Ált. Isk.	Dél-alföldi r.	Szeged környéki cs.
Budapest	ELTE Tan.-és Óv. F. K. Gyak. Ált. Isk.	*Budapest	-----
Kecskemét	K. Tanítók. Főisk. Kar Petőfi S. Gyak. Ált. Isk.	Dél-alföldi r.	Kiskunsági cs.

* Budapestet külön régióként kezeljük a település nagysága, lakóinak száma és összetétele miatt.

2.6.A felvételek időtartama

A felvételeket négy terepmunkás készítette féléven át heti egy alkalommal. Az anyaggyűjtés egy-egy adatközlőnél 2-10 percet vett igénybe. Az időkülönbség egyéni sajátosságokból adódott (eltérő olvasási tempó, ill. készségszint).

3. Az adatok feldolgozása

3.1. A hanganyag annotálása

Az annotálás azt jelenti (Vicsi *et al.*, 2002), hogy minden hangfájl mellé egy címkefájl került, ami a hangfájl paramétereivel és tartalmával kapcsolatos változókat tartalmazza. Az általánosan elfogadott szempontokon túl specifikus mutatókat is felhasználtunk. A hanganyag annotálásának eredményeként elkészült 700 gyerek adatrögzítő lapja, amely a hangfájl paramétereivel és tartalmával kapcsolatos változókat tartalmazza.

Az adatlap szerkesztésénél ügyeltünk arra, hogy a gyerekek személyiségi jogai ne sérüljenek, ugyanakkor az adatlapok felhasználhatóak legyenek szociolingvisztikai kutatásokra.

A BeMe hanganyagának annotálásánál informatikai, nyelvi és társadalmi változókat vettünk figyelembe. Ezeket a SpeechDat-E (Pollak *et al.*, 2000) és a The Budapest Sociolinguistic Interview (Kontra-Váradi 1997) alapján dolgoztuk ki.

A SpeechDat (továbbiakban SD) személyfüggetlen beszédfelismerők betanítására alkalmas adatbázisok létrehozásának szabályrendszere, melyet az Európai Közösség által létrehozott szakértői bizottság állított össze (Pollak *et al.*, 2000).

A Budapesti Szociolingvisztikai Interjú (Kontra-Váradi 1997) (továbbiakban BUSZI) az MTA Nyelvtudományi Intézetében készült magyar nyelvi szociolingvisztikai kutatás angol nyelvű leírással. Az általánosan elfogadott szempontokon túl specifikus mutatókat is felhasználtunk.

Példa egy adatlapra

SES:	s0001	* (SD) Session number - a felvétel sorszáma
ENV:	Iskola	(SD)Environment- környezet
SEX:	Lány	(SD és BUSZI) Speaker sex - az adatközlő neme
AGE:	7 éves	(SD és BUSZI) Speaker age - az adatközlő kora
ACC:	Dél-alföldi r. Szeged környéki cs.	(SD) Speaker accent - az adatközlő kiejtése
REP:	Szeged	(SD és BUSZI) Recording place - a felvétel helye
RED:	b01-b06: 2002. 06. 12. b07-b012: 2002. 09. 27.	(SD és BUSZI) Recording date a felvétel dátuma
RET:	b01-b06: 12:06-12:10 b07-b12: 12:03-12:04	(SD és BUSZI) Recording time - a felvétel időpontja
LBR:	b02: alma, ág, béka, cica, csiga, dió, dzé betű, dzsungel b03: egér, ég, fészek, gólya, gyerek, hegy, iskola, ír b04: játék, kakas, lepke, lyuk, maci, nap, nyúl, oroszlán b05: óra, ördög, őz, pók, róka, süni, szőlő, tulipán b06: tyúk, uborka, út, üveg, úrhajós, virág, zokni, zsiráf	(SD és BUSZI) Orthographic text prompt- az adatközlőtől várt szóanyag ortografikus formában
LBO:		(SD) Orthographic labelling - az adatközlő által bementett szóanyag ortografikus formában
CMT:		(SD) Comment row - megjegyzés: Sampa-átirat
SNL:	Alacsony	(SD) Subjective noise level - szubjektív zajszint
EDU:	Óvoda	(BUSZI) Education - végzettség
BRP:	Szeged	(BUSZI) Birthplace - születési hely
MOT:	Magyar	(BUSZI) Mother tongue - anyanyelv
TIA:	-	(BUSZI) Time abroad - külföldön töltött idő
KFL:	-	(BUSZI) Knowledge of foreign language - idegennyelv-tudás
MIC:	Hama Dynamic microphone IMP. 600	(BeMe)Microphone type - a mikrofon típusa
SPD:	-	(BeMe) Speech defect - beszédhiba
POC:	apja: gépészmérnök anyja: fogorvos	(BeMe) Parents' occupation – a szülők foglalkozása
OCC:	-	(BUSZI) Occupation - foglalkozás
HLT:	éphallású	(SD) Healthy - a beszélő egészségi állapota

* Az adatlap ezt az oszlopot nem tartalmazza, csak magyarázatként szolgál.

3.2.A táblázatfájlok felosztása és tartalma

A változókat a következő fájlokba csoportosítottuk. A rövidítések mellett a lehetséges válaszokat tüntettük fel.

3.2.1. A beszélőre vonatkozó információs fájl

Fájl: <BeMe>\TABLE\SPEAKER.TBL)

SEX: lány/fiú

AGE: 5/ 6/ 7

ACC: Dél-alföldi r. (Szeged környéki cs.) / Közép-dunántúli-kisalföldi r.

(Észak-dunántúli cs.) / Dél-dunántúli r. (Észak-baranyai cs.) / Dél-alföldi r.(Baja környéki cs.) / Dél-dunántúli r. (Közép-somogyi cs.) / Palóc r. (Déli tömb Eger vidéki cs.) / Északkeleti r. (Szabolcs-szatmári cs.) / Palóc r. (Keleti tömb Keleti palóc cs.) / Tisza-Körös vidéki r. (Hajdú-bihari cs.) / Budapest / Dél-alföldi r (Kiskunsági cs.)

EDU: óvoda

TIA: az évek száma

KFL: a nyelv neve

SPD: fejlődési artikulációs zavar

POC: a szülők foglalkozásának neve

BRP: a település neve

OCC: a foglalkozás neve

HLT: ép hallású, nagyothalló

3.2.2. A felvétel körülményeire vonatkozó információs fájlok

Fájl: <BeMe>\TABLE\REC_COND.TBL

ENV: iskola

MIC: Hama Dynamic microphone IMP. 600 / Shure HW505 / Typhoon multimedia / Sennheiser / Gaia SL-8720 / Circe SL-8715

SNL: magas / közepes / alacsony

3.2.3. A felvételek sorrendjére vonatkozó fájlok

Fájl: <BeMe>\TABLE\SESSION.TBL

SES: 0001- 0500

Minden SES egy adatközlő bemondásait tartalmazza 12 fájlba elosztva.
(500 fő x 12 = 6000 fájl)

A felépítés:

<BeMe>\<ses0001-ses0500>\<b01-b12>

A tartalma:

b01: a 14 db magyar magánhangzó (2.2.)

b02-b06: minden adatközlőnél azonos bemondások (2.2.), fájlanként 8 szó (8 x 5 = 40 szó)

b07-b12: az adatközlőnként változó bemondások, fájlanként 10 szó (6 x 10 = 60 szó)

RED: év, hó, nap

RET: óra, perc

3.2.4. A lexikonfájl tartalma

Fájl: <BeMe>\TABLE\LEXIKON.TBL

LBR: az adatközlőtől várt szóanyag ortografikus formában

LBO: az adatközlő által bemondott szóanyag ortografikus formában

CMT: az adatközlőtől várt szóanyag SAMPA átírata

3.2.5. A bemondások száma

Az 500 kisiskolás bemond 7000 magánhangzót és 50 000 szót, ami összességében 57 000 bemondást jelent.

3.3. A szegmentálás

Az adatbázis tartalmazza a korlátozottan elérhető szegmentált hanganyagot is. A szegmentálás a beszéd folyamat lineáris tagolása, vagyis a hangtest hangegységekre bontása. A hangtestben előforduló diszkrét metszetek kijelölésével jutunk el a beszédhanghoz, a hangtest legkisebb szegmentális szerkezeti egységéhez (Kiss, 2001). A beszéd időfüggvényében bejelöljük a fizikailag megfigyelhető beszédhangokat és azok határait. A szegmentálás célja, hogy az adatbázis gépi feldolgozásához megadjuk a beszédjel és a fonetikai átírat közti időbeli kapcsolatot, azt, hogy melyik szimbólum a beszédjel mely időintervallumának felel meg. A szegmentálás egységei a beszédhangok, ezekből absztraháljuk a fonémákat (Vicsi *et al.*, 2002).

Jelöltük az adatközlő által elkövetett hibákat és a felvétel során hallható, nem környezeti zajokat is. A munkához az MTA-SZTE Mesterséges Intelligencia Kutatócsoportja által kifejlesztett HotSA programot használtuk (HotSA 2001).

Az adatbázis tartalmazza a beszédhangok fonetikai átírását is, ez a beszéd részlet hangjait jelölő speciális fonetikai kódok sorozata. Az elhangzott szövegek lejegyzésére a nyelvészetben általában az IPA szimbólumrendszert használják. Ez a szimbólumrendszer azonban nem illeszkedik a számítógép billentyűzetéhez, ezért a nemzetközi SAMPA kódolással dolgoztunk (Wels *et al.*, 1992). A SAMPA alapja az IPA jelölésrendszer, de az előbbi jobban alkalmazkodik a számítógéppel kezelhető karakterkészlethez; a segítségével könnyebb megoldani a gépelést és a továbbítást.

3.3.1. A munka menete

A rögzített hanganyag tisztítása (nagyobb szünetek törlése, ismétlődő részek eltávolítása, a szó bemondó lapon nem szereplő szavak törlése) után került sor a szegmentálási műveletre.

Először meghallgattuk a felvételt az esetleges zajok, hibák, a fonetikus átírat hibáinak a kiszűrésére. Majd erősebb nagyítás mellett jelöltük a beszédhanghatárokat. A következő meghallgatás során pontosítottuk, véglegesítettük a határokat, javítottuk az esetleges ejtési eltéréseket.

A szegmentálási munkát egyszerűsítette az automata szegmentáló. Ez egy olyan algoritmus, amely automatikusan jelölte a hanghatárokat, ezzel könnyítette és gyorsította a kézi szegmentáló munkáját.

3.3.2. A szegmentálás szempontjai

A beszédhangok a lineáris hangszerkezetben hangkapcsolatokat képeznek. A hangok gyakran nem különíthetők el élesen egymástól, rövidebbek-hosszabbak köztük az átmenetek,

ezért a munka során azt tapasztaltuk, hogy a beszédhanghatárt nem lehet mindig egyértelműen bejelölni. Gyakran a spektrogram és a hullámforma összehasonlítása, de még a meghallgatás sem adott objektív fogódzót a határ pontos bejelölésére.

Ezekben az esetekben az alábbi elvekhez (Vicsi *et al.*, 2002) tartottuk magunkat.

- A szegmentumhatárokat igyekeztünk a null átmenetekhez igazítani.
- Zöngés hangok esetében ez a pozitív null átmenetet jelenti. Igyekeztünk nagyon precízen jelölni a határt. Zöngétlen hangoknál 1 ms pontossággal jelöltük a hang kezdetét.
- A magánhangzó kezdetét a zöngé indulásánál jelöltük (zöngétlen hang után).
- A zárhangok, affrikáták kezdetét a megelőző hang utolsó periódusa előtt jelöltük.
- A magánhangzó-magánhangzó vagy magánhangzó-rezonáns mássalhangzó kapcsolatokban a határt az átmeneti rész 50%-ánál jelöltük be. Ilyenkor némileg pontatlan a bejelölés, mert a hangok kettéválasztása bizonytalan (Vicsi, Víg, 1998).

Többször előfordult, hogy egy-egy hangot többszöri visszahallgatás után sem lehetett azonosítani. Ennek jelzésére a [cut] kódot használtuk.

3.3.3. A szegmentált hanganyag szerkezete

A szegmentált anyag a következőképp épül fel:

Pl.: ses0001 (az első adatközlő bemondásai)

s0001b01.dat (az első adatközlő első bemondásának szegmentált anyaga)

s0001b01.huo (az első adatközlő első bemondásának SAMPA átírata)

s0001b01.wav (az első adatközlő első bemondásának hanganyaga)

4. Eredmények

A BeMe hanganyaga korlátozottan elérhető, az adatbázis struktúrájának specifikációja nyilvános. A program fejlesztése során létrejötték gyakorisági szótárak, nagyméretű annotált és szegmentált adatbázis. A rögzített és feldolgozott szóadatbázist azért hoztuk létre, hogy a BeMe szoftver kifejlesztéséhez szükséges beszédtechnológiai kutatásokat kiszolgálja.

Úgy gondoljuk azonban, hogy a korpusz nagysága és a feldolgozás szempontjai alapján más nyelvészeti (szociolingvisztikai, fonetikai, fonológiai...) és beszédtechnológiai kutatásokra is alkalmas az adatbázis, az annotált és szegmentált hanganyag.

A Beszédmester szoftver 2004 februárjától ingyenesen hozzáférhető az interneten (www.inf.u-szeged.hu/beszedmester). A program mellékleteként automatikusan használható az olvasástanítás, ill. a beszédjavítás-terápia metodikája.

Irodalom

Bácsi, J., Kerekes J.: Az első osztályos olvasókönyvek szóanyagából készült Gyakorisági szótár „Van szó”, Módszertani Közlemények, Szeged, 2002

É. Kiss, K., Kiefer, F., Siptár, P.: Új magyar nyelvtan, Osiris Kiadó, Budapest, 1998, pp.20

HotSA: <http://oasis.inf.u-szeged.hu>, SZT-IS-10/1/23/2001

Kiss, J.: Magyar dialektológia, Osiris Kiadó, Budapest, 2001

Kontra, M., Váradi, T.: The Budapest Sociolinguistic Interview: Version 3. Linguistics Institute, Hungarian Academy of Science, Budapest, 1997

Pollak, P., Cernocky, J., Boudy, J., Choukri, K., Heuvel, H., Vicsi, K., Virag, A., Siemund, R., Majewski, W., Sadowski, J., Staroniewicz, P., Tropf, H., Kochanina, J., Ostrouchov, A., Rusko, M., Trnka, M.: SpeechDat(E) –Eastern European Telephone Speech Databases.

Proceeding LREC' 2000 Satellite workshop XLDB – Very large Telephone Speech Databases, 29 May 2000, Athens UCL, 2000

Vicsi K., Víg A.: Az első magyar nyelvű beszédatadátbázis. Budapest: Beszédkutatás '98, MTA Nyelvtudományi Intézete, 1998

Vicsi, K., Tóth, L., Kocsor, A., Gordos, G., Csirik, J.: MTBA-Magyar nyelvű telefonbeszédatadátbázis, Híradástechnika, LVII. 2002/8, Budapest, pp.35-43, 2002

Wels, J. at all: Standard Computer-Compatible Transcription. Esprit Project 2589 (SAM), Doc.no. SAM-UCL-037. London: Phonetics and Linguistics Dep., 1992