

e-Magyar beszédarchívum

(fejezet a közös cikkbe)

Kornai András

1. Céljaink. A beszédarchívum létrehozásával három fő célunk volt. Az első és legfontosabb a magyar beszédtechnológiára annak kezdetei óta jellemző zárt kutatási és publikációs modell felváltása egy szabad, nyílt forrású (Free and Open Source Software, FOSS) modellel. Második célunk a hagyományos, gondosan felcímkezett és mind artikulációsan mind aukusztikailag tiszta adatokon alapuló felügyelt tanulási módszerek felváltása gyengén felügyelt illetve felügyeletlen (weakly supervised, unsupervised) módszerekkel. Harmadik, az első kettőtől nem mindig könnyen elválasztható célunk pedig egy a digitális bölcsészeti munkát, elsősorban a szociológiát, történelemtudományt, folklorisztikát, és néprajzot beszédtechnológiai oldalról támogató platform alapjainak megteremtése.

2. A kiinduló állapot Míg a magyar nyelvtechnológiát évtizedek óta egymással hol versenyben, hol együttműködésben álló műhelyek viszik előre, a beszédtechnológiát egészen a legutóbbi időkig alapjában egy monokulturális helyzet, a BME TMIT dominanciája jellemezte. Miután ez a műhely nem törekedett (és ma sem törekszik) sem az adatok sem a szoftver nyílttá tételére, a beszédtechnológia egészének helyzetét is ez a zárt ajtó mögötti “titkos tudomány” állapot jellemzi annál is inkább, mert a kisebb résztvevők (pl. ALL, AITIA) üzleti alapon álló vállalkozások, míg a FOSS fejlesztések hajtóerejét a világon mindenütt elsősorban az egyetemi, akadémiai világ adja.

Az Infra2 pályázat a nyelvtechnológiában, különösen a szószintű eszközök (morfológiai elemzés és generálás) de kisebb részben már a frázis- és mondat-szintű eszközök területén teljessé tette a nyílt forrású adatok és eszközök bevezetését (ref ugyane cikk/kötet), ennek minden a fejlődést katalizáló előnyével együtt. Ez csak úgy volt lehetséges, hogy az évtizedek során komoly FOSS eszközök halmozódtak fel, melyek közül a teljesség igénye nélkül kiemeljük a Hun* és a Magyarlanc eszközláncokat, a monolingvális Webkorpuszt és a Hunglish párhuzamus korpuszt. Mostani projektünk elkezdése előtt a magyar beszédtechnológia szabadon letölthető adatokat nem tett

közzé (az egyedi mérlegelésen alapuló hozzáfér-engedélyezést nem sorolhatjuk a FOSS paradigmába) sem a világszerte közismert beszédtechnológiai eszközök magyar honosításai nem voltak elérhetőek, annak ellenére, hogy a létező szoftverek, különösen a beszédfelismerés terén, elsősorban ilyeneken alapultak (ennek pontos mértéke természetesen csak a szoftverek nyilvánosságra kerülésével lesz megállapítható).

3. A projekt eredményei Elmondhatjuk, hogy az Infra2 FOSS beszédarchívum megjelenésével a helyzet gyökeresen megváltozott. Az adatok szintjén elérhetővé vált sokezer órányi jogtiszta adásmínőségű (broadcast quality) és sokszáz órányi ennél rosszabb (communication quality) anyag. Ezeknél sokkal jobb minőséget képvisel a BEA spontánbeszéd-adatbázis (Gósy et al 2012), de kisebb, és nem teljesen FOSS. Hangsúlyoznánk, hogy a korszerű beszédtechnológiában a jobb akusztikai minőség nem követelmény, sőt, immár több évtizedes tapasztalat, hogy a legjobban azok a beszédfelismerő rendszerek teljesítenek, melyeket reális, az alkalmazásban valóban fellépő akusztikai körülményeket tükröző adatokon tanítottak be.

Ugyanilyen változást hozott a projekt a követő szoftverek terén is. Sok tucatnyi alternatíva telepítésével és összemérésével választottuk ki a legjobbakat. A hangformátumok konverziójára a `sox` és `ffmpeg` eszközöket, a beszédaktivitás detektálására és naplózás (diarization) céljára a `shout` programot, végül statisztikai nyelveldezésre az `srilm` eszközt. Ez utóbbihoz olyan modelleket tettünk elérhetővé, melyek perplexitása 56, tudunkkal az összes publikált (de le azért nem tölthető) modell perplexitását lényegesen megjavítva. Eredeti vállalkozásunkkal ellentétben *nem készült el*, de terveink között változatlanul szerepel az automatikus nyelv-azonosítást lehetővé tevő szoftver.

4. Automatikusan várható eredmények Az archívum bővülése több irányból is várható anélkül, hogy ez újabb anyagi vagy emberi erőforrásokat igényelne. Támogatásukról biztosítottak a NAVA, az OGYK, az OSZK, az MTA TK, Kisebbségkutató, és Szociológiai intézetek és más intézmények is, sőt az intézmények egy részétől már kaptunk is anyagokat.

Különösen fontos a hazai és környező országokbeli társadalomtudósok támogatása. A teljesség igénye nélkül: Havas Gábor, Lengyel Gabriella, Németh Szilvia, Zolnay János, Virág Tünde; a kolozsvári kisebbségkutató (Fosztó László, Kiss Tamás, Vitos Katalin, Lőrincz József), a marosvásárhelyi Sapientia (Gagy József), a kolozsvári Kriza Társaság (Szabó Tőhötöm), a Babes-Bolyai Egyetem (Tánczos Vilmos, Pozsony Ferenc), a kolozsvári, marosvásárhelyi rádiók anyagai (Maksay Ágnes, Tibád Zoltán). Külön említést igényel Molnár Gusztáv hatalmas interjúanyaga (magyarul, románul)

a XX század olyan jelentős személyeivel mint Balogh Edgár vagy Szabó T. Attila. Sajnos ezen anyagok nagy része ma még kazettán van, de ezek átjátszását folyamatosan végezzük.

Különösebb plusz befektetés nélkül, csupán a meglevő folyamatok folytatásával az archívum még éveken át bővülni fog.

5. A továbblépés főbb irányai Számítunk a közösség, *primus inter pares* a BME TMIT támogatására abban, hogy a beszédarchívum még jobban használható legyen. Az első és legfontosabb lépés ebben egy *adatkezelési* modell (data curation model) kialakítása kell legyen.

5.1 Kik adják az adatokat? A [google kérdőív](#) kitöltésével bárki, aki szeretné adatait nyilvánosan hozzáférhetővé tenni.

5.2 Kik őrzik az adatokat? Ennek infrastrukturális hátterét legalább 10 évre megadta az Infra2 finanszírozású hardver-fejlesztés, a szervezeti hátteret biztosítja az MTA Nyelvtudományi Intézet és az MTA SZTAKI közti megállapodás. Természetesen teljes idejű, vagy akár részidejű digitális könyvtáros felvétele a folyamatot nagyban gyorsítaná, erre azonban az Infra2 nem adott módot.

5.3 Milyen metaadatokat tároljunk, és milyen sémában? A rendszer rugalmas, itt elsősorban az érdekelt felhasználók véleményét várjuk.

A második kérdés a további szoftverek fejlesztése. Itt a ProsoTool (ref) által követett irányt (független [github](#) szoftver-repozitórium belinkelése) tartjuk a leghatékonyabbnak, és nagy örömmel várjuk a többi FOSS szoftver megjelenését.