Magyar nyelvű nagyszótáras beszédfelismerési feladatok adatelégtelenségi problémáinak csökkentése nyelvimodell-interpoláció alkalmazásával

Tarján Balázs[1], Mihajlik Péter[1,2]

[1] Budapesti Műszaki és Gazdaságtudományi Egyetem,
Távközlési és Médiainformatikai Tanszék
{tarjanb, mihajlik}@tmit.bme.hu
[2] THINKTech Kutatási Központ Nonprofit Kft.

Kivonat: A lineáris interpolációt elterjedten alkalmazzák in-domain és out-of-domain nyelvi modellek egyesítésére folyamatos, nagyszótáras gépi beszédfelismerési feladatokon. Nyelvünk gazdag morfológiája azonban szükségessé teszi, hogy morfémaalapon is megvizsgáljuk a módszer hatékonyságát, és összevessük az interpolációs és a tanítókorpuszok sima egyesítésével kapható eredményeket. Cikkünkben bemutatunk egy új megközelítést morfémaalapú nyelvi modellek interpolációjára, mellyel 3gram modellek esetén sikerült megjavítani a korpuszegyesítéses módszer eredményét. A nyelvimodell-komplexitást 4gramra növelve azonban az interpolációval nyerhető előny eltűnik, így megítélésünk szerint a morfémaalapú interpolációra vonatkozóan további vizsgálatok szükségesek. Kísérleteink során sikerült 12% alá csökkenteni a szóhibaarányt a tesztelési célokra használt hangoskönyvrészleten, mely legjobb tudomásunk szerint az eddigi legalacsonyabb eredmény magyar nyelvű, nagyszótáras feladaton.

1 Bevezetés

A nagyszótáras beszédfelismerő rendszerek pontosságát döntően befolyásolja a nyelvi modell mérete és minősége. Minél nagyobb és a felismerési feladathoz jól illeszkedő szövegkorpusz áll rendelkezésünkre a rendszer tanításához, annál precízebben írható le a szótári elemek kapcsolata az n-gram modellben. Azonban a gyakorlati tapasztalat szerint jó minőségű tanítóanyagok csak korlátozott mennyiségben hozzáférhetők, így a nyelvi modell robosztusságát gyakran a feladathoz nem vagy csak lazán kapcsolódó tanítóadat bevonásával kell növelni.

Több megoldás is létezik arra, hogy különböző szöveges tudásforrások egy közös nyelvi modellben hasznosuljanak. Szokás a rendelkezésre álló szövegeket egyszerűen összemásolni, és az így létrejött korpusszal tanítani egy n-gram modellt. Az eljárás hátránya, hogy egy nagyméretű kiegészítő korpusz könnyedén elnyomhatja a kisebb, de a feladat szempontjából releváns tanítószöveg szókapcsolati statisztikáit. Erre kínál megoldást a **nyelvimodell-interpoláció**, mellyel különböző nyelvi modellek n-gram becslései egyesíthetőek tetszőlegesen megválasztott súlyozó tényezővel. A

nyelvimodell-interpolációs technikák közül az egyik legegyszerűbb, ám igen hatékony eljárás a nyelvi modellek ún. **lineáris interpolációj**a [6]. Megvalósítása az alábbi képlet alapján történik. (**4**)

$$P(w|h) = \sum_{s \in S} \alpha_s P_s(w|h) \quad (1)$$

Ahol *w* jelöli az interpolált modell megbecsülendő szótári elemét, *h* az előtörténetet, *S* a forrásmodellek összességét, míg α_s és $P_s(w|h)$ a s-edik modellhez tartozó **interpolációs súly**t, valamint nyelvimodell-becslést. Új modell generálásakor α_s értékek változtatásával tudjuk az egyes forrásmodellek részvételi súlyát változtatni. Az interpolációban részt vevő modellek optimális arányának megállapítása általában in-domain szöveg perplexitásvizsgálatán alapul.

A lineáris interpoláció kiforrott és elterjedten használt technikának számít szóalapú nyelvi modellek esetén. Azonban a morfológiailag gazdag nyelveknél – mint amilyen a magyar – a jelentős szóalaki változatosság miatt fellépő adatelégtelenség megkérdőjelezi a szóalapú megközelítés létjogosultságát. Összehasonlító kísérletek bizonyítják, hogy magyar nyelven szóalapú helyett morfémaalapú nyelvi modelleket használva szignifikáns felismerésipontosság-növekedés érhető el [9, 11]. Felvetődik tehát a kérdés, hogy morfémákra cserélve az egyesítendő nyelvi modellek alapját, vajon a szóalapú megközelítéshez hasonló mértékben növekszik-e a felismerési pontosság, illetve ha nem, milyen módon növelhető mégis a morféma alapon interpolált nyelvi modellek teljesítőképessége.

Kísérleteink során megvizsgáljuk, milyen módszerekkel interpolálhatók hatékonyan a morfémaalapú nyelvi modellek, és összevetjük a szóalapú nyelvimodell-interpolációs eredményekkel. Emellett, hogy az interpoláció hatékonyságát általában is értékelni tudjuk, összehasonlítjuk az interpolált és az egyszerű korpuszegyesítéses modellek eredményeit is. Cikkünk további részében először a kísérletekhez használt tanító-, illetve tesztadatbázist ismertetjük, majd kitérünk az akusztikus és nyelvi modellek tanításánál alkalmazott módszerek bemutatására. A felismerési feladat részletes áttekintése után kiértékeljük a különböző interpolációs technikákat egy e célból létrehozott tesztanyagon, míg végül összefoglalását adjuk kísérleteink legfontosabb következményeinek.

2 Felismerési feladat és módszertan

A bevezetésben felvetett kérdések megválaszolásához először egy olyan felismerési feladatot kellett találnunk, mely alkalmas a különféle interpolációs módszerek vizsgálatára. Választásunk egy beszédfelismerési kísérletekhez már korábban is felhasznált [12] hangoskönyvre esett, mely Krúdy Gyula Szindbád történeteinek felvételét tartalmazza Gáspár Sándor előadásában. Fontos szempont volt, hogy olyan feladatot válasszunk, melyhez könnyen elérhető jól illeszkedő tanítószöveg, illetve hogy egy a feladattól távolabb álló, de műfajában kötődő, nagyobb méretű tanítókorpusz is gyűjthető legyen hozzá. Emellett további előnye a hangoskönyvnek, hogy a felvételeken a háttérzaj és a beszéd spontán jegyeiből adódó artikulációs pontatlanságok hatá-

sa elhanyagolható, így biztosított, hogy a felismerési pontosságok változása valóban a nyelvi modellek eltérő teljesítményéhez köthető. A rendelkezésünkre álló felvételt a [12]-ben leírtakkal megegyező módon két részre osztottuk. A nagyobbik, 186 perces részt az akusztikus modell tanításához használtuk fel, míg a kisebbik, 26 perceset a felismerő hálózatok tesztelésére.

2.1 Akusztikus modell tanítása

Akusztikusmodell-tanításhoz a hangoskönyv teszteléshez nem használt része, összesen 186 perc állt rendelkezésre. Figyelembe véve, hogy ez a több mint 3 óra egyetlen beszélőtől származik, úgy döntöttünk, hogy egy új, beszélőfüggő akusztikus modellt tanítunk. Először egy, az MRBA [13] beszédadatbázison tanított beszélőfüggetlen akusztikus modell segítségével kényszerített felismerést hajtottunk végre a tanítóanyagon, melyhez felhasználtuk az érintett Szindbád-novellák szövegét is. Ezután a kényszerített felismerés kimenete alapján háromállapotú, balról-jobbra struktúrájú, környezetfüggő rejtett Markov-modelleket tanítottunk. A létrejött akusztikus modell 1400 egyenként 7 Gauss-függvényből álló állapotot tartalmaz. A felismerési kísérletek során mindvégig ezt az akusztikus modellt használtuk.

2.2 Tanítószövegek gyűjtése és előkészítése

Mint a bevezetőben kitértünk rá, a nyelvimodell-interpolációs technikát gyakorta használják arra, hogy egy, a feladathoz jól illeszkedő kisebb és egy feladathoz csak lazán kötő nagyobb nyelvi modell előnyeit egyesítsék. Esetünkben a feladathoz jól illeszkedő modell tanításához tanítószövegként Krúdy Gyula műveinek gyűjteménye szolgált. A létrehozott korpusz 1,4 millió szót tartalmaz, forrása a Magyar Elektronikus Könyvtár [8]. Ez az általunk **jól illeszkedő (JI)** korpusznak keresztelt szöveg nem tartalmazza sem a tesztanyag, sem az akusztikusmodell-tanításhoz használt felvételek leiratát. A JI korpusz kiegészítéséhez három forrásból gyűjtöttünk, további összesen 16,6 millió szót tartalmazó tanítószöveget: Magyar Elektronikus Könyvtár, Digitális Irodalmi Akadémia [3], Elektronikus Periodika Archívum és Adatbázis [4]. Ez a tanítószöveg – melyre a továbbiakban **gyengén illeszkedő (GYI)** korpuszként fogunk hivatkozni – Krúdy Gyula kortársainak és hozzá stílusban közel álló szerzők szépirodalmi műveire épül.

Szóalapú tanítószöveg-előállítás

Egy beszédfelismerési alkalmazás a szöveges tanítóadatok előfeldolgozását követeli meg. A rendszer tanításához felhasznált szépirodalmi szövegek olyan elemeket is tartalmaznak, melyeket nem lehet, vagy eredeti alakjukban nem lehet beszédhangokkal leírni. Ennek megfelelően az írásjeleket eltávolítottunk a tanítószövegből, míg a számokat szöveges átiratukkal helyettesítettük. Végül minden karaktert kisbetűsre alakítottunk. Az így előállt előfeldolgozott tanítószöveget használtuk a szóalapú nyelvi modellek tanításához.

Morfémaalapú tanítószöveg-előállítás

A morfémaalapú tanítószövegek előállításához további lépések szükségesek. Először speciális szóhatárjelölő karaktereket (<w>) helyeztünk a szövegbe, melyeket külön morfémaként kezeltünk a nyelvi modellben. Szerepük abba rejlik, hogy segítségükkel vissza tudjuk állítani a morfémaalapú kimenetben a szóhatárokat. Ezután létre kellett hozni egy, a szavakat morfémák sorozatára átíró szótárat. Cikkünkben felhasznált morfémaalapú tanítószövegek az ún. **Morfessor Baseline (MB)** statisztikai szegmentáló eljárással [2] készültek. A MB egy felügyelet nélküli, nyelvfüggetlen morfémaszegmentáló eljárás, melyet kifejezetten beszédfelismerési célokra fejlesztettek ki finn kutatók. Segítségével csupán a szótár megadásával összerendelhetők a szavak morfémabontásukkal. A szóhatárjelölő szimbólummal ellátott, előfeldolgozott tanítószövegben ezután már csak a szavakat kellett morfémaszegmentálásukkal helyettesíteni.

Kétféle elv szerint hoztuk létre a tanítószövegekhez tartozó morfémakészleteket. Először a két tanítószöveghez tartozó szótáron egymástól függetlenül alkalmaztuk a MB szegmentálást. Ezt a megoldást **független szótáras (FSZ)** megközelítésnek neveztük el. Bár morfémaalapú hálózatok interpolációjával kapcsolatban nemzetközileg is kevés a tapasztalat, a független szótáras megoldás alkalmazása felvet egy problémát. Ha a statisztikai feldolgozó egymástól függetlenül szegmentálja az interpolálandó nyelvi modellek szótárát, akkor nagy valószínűséggel merőben eltérő morfémakészlet keletkezik. Ennek következtében a nyelvi modellek összefűzése során kevés közös n-gram lesz a két szótárban, ami ronthatja az interpoláció hatásfokát.

A probléma kezelésére több módszert kidolgoztunk, melyek közül egy ún. **közös szótáras (KSZ)** megközelítés vált be a legjobban. Ennek lényege, hogy a két tanítószöveg szótárát egyesítettük, majd ezen a közös szótáron futtattuk a statisztikai szegmentálást. A két tanítószövegben így minden közös szó ugyanarra a morfémasorozatra íródott át, ezzel biztosítva a lehető legtöbb közös n-gramot nyelvi modellekben. A kétféle módszert csak interpolációban részt vevő nyelvi modellek esetén alkalmaztuk. Korpuszegyesítés esetén a szótár a két részkorpusz közös szótárának adódik, így az itt alkalmazott szó-morféma átírás megegyezik a közös szótáras módszernél kapottal. A tanítószövegekkel kapcsolatos részletes statisztikákért lásd az **1. táblázatot**.

1. táblázat: A nyelvi modell tanító adatbázisokhoz kapcsolódó statisztikák

Tanító-korpusz	Méret [millió szó]	Szótár [ezer szó]	FSZ morféma-készlet [ezer morf.]	KSZ morféma-készlet [ezer morf.]	Szó-perplexitás [–]	OOV arány [%]
JI	1,4	152	18	36	1559	4,9
GYI	16,6	800	64	65	2905	2,6
Egyesített	18,0	840	–	66	2121	1,9

2.3 Nyelvi modellek tanítása

Mind a JI, GYI, mind az egyesített korpuszból készült nyelvi modellek módosított Kneser-Ney simítás [1] használatával készültek az SRI-LM [10] nyelvi modellező

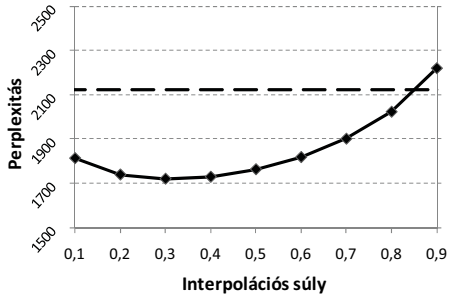
toolkit segítségével. Modellmetszést egyetlen esetben sem alkalmaztuk. Az interpolált nyelvi modellek előállításához azt az elterjedten használt technikát [7] alkalmaztuk, mely szerint egy kisebb méretű, in-domain (JI) és egy nagyobb méretű, out-of-domain (GYI) nyelvi modellt tanítottunk egymástól függetlenül, majd ezeket az SRI-LM toolkit-be épített lineáris interpolációs eljárás segítségével különböző arányban egyesítettük. A tanítókorpuszokra vonatkozó perplexitásértékek és szótáron kívüli szóarányok jól illusztrálják (**1. táblázat**), hogy bár a GYI korpusz kevésbé illeszkedik jól a tesztanyaghoz, több, a tesztanyagban előforduló szót képes modellezni, mint a JI.

3 Felismerési eredmények

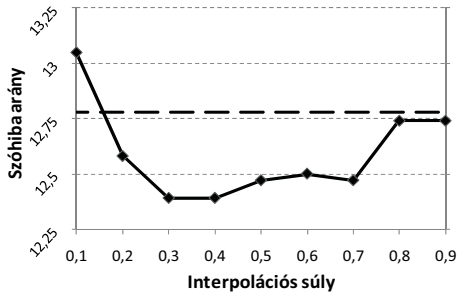
A tesztfelvétel lényegkiemeléséhez 39 dimenziós, delta és delta-delta értékkel kiegészített mel-frekvenciás kepsztrális komponenseken alapuló jellemzővektorokat hoztunk létre és ún. vak csatornakiegyenlítő eljárást is alkalmaztunk. A súlyozott véges állapotú átalakítókra (WFST) épülő felismerőhálózatok generálását és optimalizálását az Mtool keretrendszer programjaival végeztük, míg a tesztelés során alkalmazott egyutas mintaillesztéshez a VOXerver [5] nevű WFST-dekódert használtuk. A felismerő rendszerek teljesítményének értékeléséhez szóhibaarányt (WER) számoltunk. Az egyes rendszerekkel elérhető WER értékek összehasonlításához a (**2**) képletben definiált mérőszámot használtuk.

$$Relatív\ \ WER\ csökkenés = \frac{WER_{referencia} - WER_{új}}{WER_{referencia}} * 100\% \qquad (2)$$

3.1 Szóalapú 3gram eredmények



1.1 ábra. Szóperplexitás az interpolációs súly függvényében.



1.2 ábra. Szóhibaarány az interpolációs súly függvényében.

Az **1.1 ábrán** látható, hogyan alakul a tesztanyagon vizsgálva a különböző interpolációs súllyal készült szóalapú 3gram nyelvi modellek perplexitása. A súly értéke a GYI korpuszból készült modell részarányát jelöli. Megfigyelhető, hogy a kiegészítő korpusz részarányának növelése egy pontig csökkenti a perplexitást, majd a 0,3-as

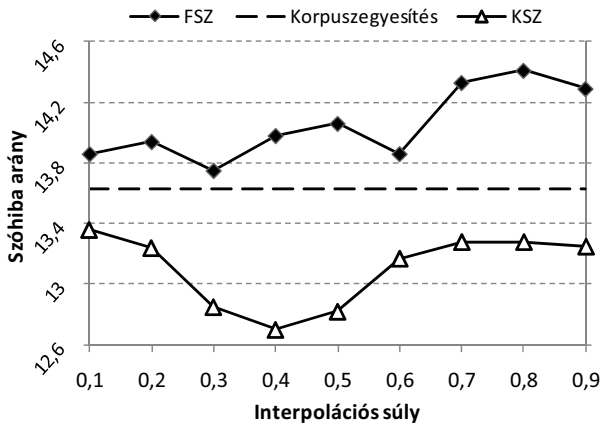
értéktől kezdve az újra növekedni kezd. Hasonló tendencia figyelhető meg a **1.2 ábrán**, mely a szóhibaarányokat ábrázolja a súly függvényében. Mindkét grafikonon szaggatott vonal jelöli a korpuszegyesítéses módszerrel elérhető perplexitást, illetve szóhibaarányt. Az a tény, hogy a folytonos vonal nagy része a szaggatott vonal alatt halad, szemléletesen mutatja, hogy szóalapú modellek esetén az interpoláció hatékonyabb, mint a korpuszok egyszerű egyesítése. Az elérhető legnagyobb pontosság esetén az interpolációval kapható relatív WER-csökkenés 3%-ot tesz ki.

3.2 Morfémaalapú 3gram eredmények

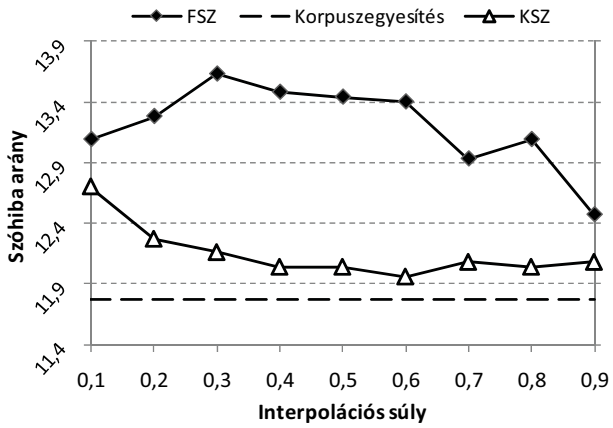
A morfémaalapú nyelvi modellek előállításához két különböző szegmentálási módszert is alkalmaztunk. Az első ún. független szótáras (**FSZ**) esetén nem készítjük fel a nyelvi modelleket az interpolációra, így azok morfémakészlete egymástól független optimalizálás eredménye (**2. ábra**). Ezt a megközelítést alkalmazva láthatóan egyetlen interpolációs súly esetén sem tudjuk javítani a korpuszegyesítéssel kapható szóhibaarányt. Ezzel szemben, ha a morfémaszegmentálás az általunk bevezetett közös szótáras (**KSZ**) módszerrel történik, akkor szóalapú eredményekhez hasonlóan csökkenteni lehet interpolációval a szóhibaarányt. A korpuszegyesítéses módszerhez képest mérhető maximális relatív WER-csökkenés (7%) felülmúlja a szó alapon kaphatót.

3.3 Morfémaalapú 4gram eredmények

Korábbi kutatásaink során többször tapasztaltuk, hogy morfémaalapú nyelvi modellezéskor 3-ról 4gram-ra növelve a nyelvi modell komplexitását szignifikánsan növekedett a felismerési pontosság [9, 11]. Ezért fontosnak láttuk morfémaalapon a 4gram modellek vizsgálatát is. Némi meglepetésre 4gram nyelvi modellek interpolációjakor nem sikerült javítani a korpuszegyesítéssel kapható felismerési eredményen. Azonban a közös szótáras (**KSZ**) megoldás itt is felülmúlja a független szótárasat (**FSZ**) felismerési pontosság tekintetében. (**3. ábra**)



2. ábra. Morfémaalapú 3gram szóhibaarányok az interpolációs súly függvényében.



3. ábra. Morfémaalapú 4gram szóhibaarányok az interpolációs súly függvényében.

4 Összefoglalás

Cikkünkben a nyelvi modellek lineáris interpolációjának alkalmazási lehetőségeit vizsgáltuk elsősorban morfémaalapú beszédfelismerő rendszerek esetén. Felismerési feladatként egy képzett beszélőtől származó hangoskönyvrészletet használtunk, melyhez egy kisebb in-domain és egy nagyobb out-of-domain tanítószöveget gyűjtöttünk. Az idealisztikus körülményeknek hála, sikerült 12% alá szorítani rendszerünk szóhibaarányát, mely legjobb tudomásunk szerint az eddig publikált legalacsonyabb érték nagyszótáras, folyamatos magyar nyelvű beszédfelismerési feladaton.

Az interpolációval és a tanítókorpuszok sima egyesítésével kapható eredményeket folyamatosan összevetettük, hogy képet kapjunk az interpolációval járó előnyökről. Hagyományos szóalapú interpolációval 3%-os WER-javulást tudtunk regisztrálni. Ez a javulás 7%-osra nőtt 3gram morfémaalapú felismerővel, ám csak abban az esetben, ha az általunk bevezetett új, a morfémaszegmentálást a tanítókorpuszok közös szótárán végző módszerrel hajtottuk végre. Ha a szótárakon függetlenül végeztük a morfémabontást, az interpoláció hatástalan eljárásnak bizonyult. Növelve a morfémaalapú nyelvi modell komplexitását 3-ról 4gramra, eltűnt az interpolációval kapható előny, és a korpuszegyesítéses módszerrel nagyobb felismerési pontosságot értünk el.

Jelen kísérletünkben nagyobb komplexitású morfémaalapú nyelvi modell esetén a lineáris interpoláció nem növelte a pontosságot a standard eljáráshoz képest. Annak eldöntéséhez, hogy ez a megfigyelés általános érvényű, vagy csupán a felismerési feladat sajátosságaiból következik, további vizsgálatok szükségesek. Éppen ezért a későbbiekben vizsgálni szeretnénk a lineáris interpolációt olyan feladatokon, melyekhez a mostaninál nagyobb tesztanyag érhető el, így kiküszöbölve a mérési hibát. Illetve ki szeretnénk próbálni a közös szótáras morfémainterpolációt olyan esetekre is, amikor a jelenleginél sokkal kevesebb adat áll rendelkezésre in-domain nyelvi modell tanításához.

Köszönetnyilvánítás

Ezúton szeretnénk köszönetet mondani az AITIA International Zrt.-nek és a THINKTech Kutatási Központ Nonprofit Kft.-nek a rendelkezésünkre bocsátott eszközökért. Kutatásunkat részben a KMOP-1.1.3-08/A-2009-0006-os és TAMOP-4.2.2-08/1/KMR-2008-0007-es projekt támogatta.

Bibliográfia

1. Chen, S. F., Goodman, J.: An Empirical Study of Smooting Techniques for Language Modeling. Technical Report TR-10-98. Computer Science Group, Harvard University (1998)
2. Creutz, M., Lagus, K.: Unsupervised Morpheme Segmentation and Morphology Induction from Text Corpora Using Morfessor 1.0. In: Comp. and Inf. Sci., report A81. HUT (2005)
3. Digitális Irodalmi Akadémia. http://www.irodalmiakademia.hu
4. Elektronikus Periodika Archívum és Adatbázis. http://epa.oszk.hu
5. Fegyó, T., Mihajlik, P., Szarvas, M., Tatai, P., Tatai, G.: VOXenter - Intelligent voice enabled call center for Hungarian. In: EUROSPEECH (2003) 1905–1908
6. Jelinek, F., Mercer, R.L.: Interpolated estimation of Markov source parameters from sparse data. In: Proc.Workshop on Pattern Recognition in Practice (1980)
7. Liu, F. et al.: IBM Switchboard progress and evaluation site report. In: LVCSR Workshop, Gaithersburg, MD. National Institute of Standards and Technology (1995)
8. Magyar Elektronikus Könyvtár. http://www.mek.oszk.hu
9. Mihajlik, P., Tüske, Z., Tarján, B., Németh, B., Fegyó, T.: Improved Recognition of Spontaneous Hungarian Speech – Morphological and Acoustic Modeling Techniques for a Less Resourced Task. In: IEEE Transactions on Speech and Audio Processing, Vol. 18 No. 6, (2010) 1588–1600
10. Stolcke, A.: SRILM – an extensible language modeling toolkit. In: Proc. Intl. Conf. on Spoken Language Processing. Denver (2002) 901–904
11. Tarján, B., Mihajlik, P.: On Morph Based LVCSR Improvements. In: Proc. of the 2nd Int. Workshop on Spoken Language Technologies for Under-resourced Languages (2010) 10–15
12. Tóth L.: Beszédfelismerési kísérletek hangoskönyvekkel. In: VI. Magyar Számítógépes Nyelvészeti Konferencia. Szeged (2009) 206–216
13. Vicsi K., Kocsor A., Teleki Cs., Tóth L.: Beszédadatbázis irodai számítógép-felhasználói környezetben. In: II. Magyar Számítógépes Nyelvészeti Konferencia. Szeged (2004) 348–359