Magyar kiejtési szótár az Interneten

Abari Kálmán[1], Olaszy Gábor[2],
Zainkó Csaba[3], és Kiss Géza[3]

[1] Debreceni Egyetem,
Pszichológia Intézet és Matematikai és Számítástudományi Doktori Iskola
abarik@delfin.unideb.hu
[2] Magyar Tudományos Akadémia, Nyelvtudományi Intézet
olaszy@nytud.hu
[3] Budapesti Műszaki Egyetem, Távközlési és Médiainformatikai Tanszék
{zainko, kgeza}@tmit.bme.hu

Kivonat: Internetes magyar kiejtési szótár mindezidáig nem készült magyar nyelvre. Az igény viszont világszerte nagy. Ezt a hiányt kívánjuk pótolni fejlesztésünkkel. Szótárunk megvalósításának terveit alapvetően az elektronikus lehetőségek maximális támogatására alapoztuk. Ez sok különbséget jelent egy hagyományos szótárhoz képest. Az egyik ilyen a szóállomány. Szótárunk nemcsak szótöveket tartalmaz, hanem azok ragozott, toldalékolt formáit is, mindösszesen 1,8 millió lexikai egységet. Ezért ebben a kiejtési szótárban a szótárelemeket **szóalak**nak hívjuk. A szótár minden lexikai eleme tehát egy-egy szóalak, amelynek a kiejtését nemzetközi fonetikai hangjelekkel (IPA) adjuk meg. Külön lexikai csoportot alkotnak a leggyakoribb magyar vezetéknevek, amelyeknek a kiejtését szintén megadjuk. A szótár különlegessége a hangos szótárrész, amelyből 60.000 szóalak hangban is meghallgatható. A szótár a http://fonetika.nytud.hu honlapon lesz hozzáférhető 2007-től.

1 Bevezetés

A kiejtési szótárak jól használhatók a kutatásban, az oktatásban (egyetemi oktatás, nyelvtanítás), gyakorlati alkalmazásokban és még számos területen. Az ilyen szótárak elektronikus, publikus formában való közreadása (Interneten) tovább tágítja a használat terét. Tudomásunk szerint magyar nyelvre ilyen nyilvánosan hozzáférhető nyelvtechnológiai adattár nem áll rendelkezésre az Interneten. Az igény viszont világszerte nagy. Ezt a hiányt kívánjuk pótolni.

Miért van szükség kiejtési szótárakra? Mert a nyelv írott és ejtett formája különbözik. A kettő közötti kapcsolat nyelvfüggő. Ahhoz, hogy egy nyelv szavait ki is tudjuk ejteni, tudnunk kell, hogy milyen hangsort kell megvalósítanunk az artikuláció során. Az úgynevezett fonetikusabb nyelveknél a kiejtés és az írás között szoros a kapcsolat, míg a kevésbé fonetikus nyelveknél nehezebb az írott alakból származtatni a kiejtést. A magyart közepesen fonetikus nyelvnek tarthatjuk. Ez azt jelenti, hogy az írásképnek megfelelő hangsorozat meghatározása nem túl bonyolult, bár vannak jócskán furcsa, nem várt kiejtési formák is.

Mit várunk el egy elektronikus kiejtési szótártól. Azt, hogy a keresett szó begépelése után a program adja meg annak kiejtését, lehetőleg nemzetközi fonetikai hangjelekkel. Így a szótár használatához nincs szükség speciális fonetikus szimbólumkészlet megismerésére, hanem ezeknek a szabványos, jól dokumentált jeleknek az ismerete elégséges, így még az idegen nyelvű, a témában járatos látogató számára is használható. Az általunk megvalósított kiejtési szótár mind szerkezeti felépítésében, mind szolgáltatásaiban lényeges különbségeket mutat egy hagyományos szótárral szemben. A magyarra 1992-ben adtak ki kiejtési szótárt [2]. Az ilyen szótárakban a szerzőknek nem alapvető célja, hogy a magyar szóállomány kiejtési formáit (a lehető legtöbb szóra) megadja, inkább a különleges kifejezéseket, az idegen szavak kiejtését teszi közzé. A szerző [2] így összegzi szótárának célkitűzését „A Magyar kiejtési szótár régi hiányt pótol a könyvpiacon. Segítséget nyújt a legkülönfélébb hasonulások, egyes idegen szavak és rövidítések helyes kiejtésében, sőt azon szavak esetében is, amelyeket éppen hogy úgy kell kiejteni, ahogyan írva vannak, de a mindennapi beszéd ettől eltérő, helytelen alakokat alkalmaz. A 10880 szót és szókapcsolatot tartalmazó szótár hasznos segítőtársa lesz mindazoknak, akik nem csupán írni, de beszélni is helyesen szeretnének magyarul”. Egy újabb szerző legújabb kiadású ilyen szótára [6] már 40.000 elemet tartalmaz.

Az Internetes kiejtési szótár kialakításánál maximálisan támaszkodtunk a számítástechnika, az Internet adta lehetőségekre. Ebből következik, hogy a szótár szóállományának kialakítása gyökeresen más elveken nyugszik, mint amilyeneket a fenti hagyományos szótárak alkalmaztak. Esetünkben nagy, elektronikusan rögzített szövegkorpusz képezi a szóállomány kialakításának az alapját. A leendő szótár elemeit automatikus módszerrel válogattuk ki a szövegkorpuszból. Ennek következménye, hogy nemcsak szótöveket tartalmaz a szótár, hanem azok ragozott, toldalékolt formáit is, vagyis ez a szótár ténylegesen tartalmazza a magyar lexémák nagy többségét, nem téve különbséget eredetük, jelentésük között. Ezeket a szótárelemeket **szóalak**nak hívjuk. A szótár minden lexikai eleme tehát egy-egy szóalak. Az általunk használt szóalak pontos definíciója a következő: **olyan betűkből álló lexikai egység egy szövegben, amelyiket nem betű karakterek határolnak** (zömmel szóközök). A betűkarakter sorozat betűtartalma minden egyes szóalaknál legalább egy betűkarakterrel eltér a szótár más szóalakjától. Belátható, hogy elegendően nagy szövegkorpusz esetén az ilyen szóalak-állomány jól lefedi a magyar nyelv leggyakrabban használt szavainak szóállományát, tehát szótárként használható. A szótár készítésének fontos szoftver eleme a hangátíró algoritmus, amelynek segítségével a szóalakokat átírjuk hangalaki formába (ez a kiejtés formája). A hangtani szabályok a magyarra a nyelvészeti szakirodalomban jól definiáltak, bár leginkább leíró formában hozzáférhetők [1]. A jelen szótár elkészítéséhez saját hangátíró algoritmust készítettünk. Ennek fő gerincét a szakirodalomban megtalálható kiejtési szabályok alkotják, ezeket építettük be. Emellett alkalmazunk kivétel listákat, amelyek azokat a kiejtési formákat írják le szóalakokhoz kapcsolva, amelyeket a szabályok nem tudnak lekezelni. Ezek a kivétel szótárrészek szabadon bővíthetők. Az elektronikus feldolgozás (beszédtechnológia) ma már lehetővé teszi, hogy hangos szótárrészt is készítsük. Ez segíti a felhasználót a tényleges kiejtés megértésében, a hangidőtartamok érzékelésében, a szó ritmusának elsajátításában. Ezt is kihasználtuk és szótárunkban a leggyakoribb szótárelemek meg is hallgathatók.

2 A szóalakok állománya

A szóalakok állományának meghatározása elektronikus formában történt, az Internetről automatikusan gyűjtött adatokból [7]. A gyűjtést újságok internetes kiadásaiból és elektronikus könyvtárak anyagaiból végeztük. Az adatgyűjtéshez azért választottuk az újságok internetes kiadásait és elektronikus könyvtárakat, mert tapasztalatunk szerint azok az átlagos webes oldalakhoz képest jóval gondosabban megszerkesztettek és átnézettek, helyesírásilag korrekt szövegeket tartalmaznak. A forrás ilyen jellegű szelektív megválasztása ellenére ezek az oldalak is tartalmaztak például idegen nyelvű részeket is, amelyek kiszűrését meg kellett oldani. A nem magyar szövegek detektálását saját fejlesztésű nyelvdetekciós szoftverrel végeztük. A nyelvdetekció elsősorban mondat szinten történt, a magyartól eltérő nyelvű mondatokat kiemeltük a szövegtárból. Azoknál a dokumentumoknál, ahol az idegen nyelvű mondatok voltak többségben, ott a teljes dokumentumot kihagytuk a szövegkorpuszból.

A 80 millió szót tartalmazó szövegkorpuszban 1,8 millió különböző szóalakot számoltunk meg, ezek alkotják a kiejtési szótár kereshető szöveges állományát. Ezek a szóalakok adott gyakorisággal fordulnak elő a nagy szövegkorpuszban. Ha ezt a gyakoriságot vesszük figyelembe, akkor előállíthatunk egy fedési diagramot, amelyik megmutatja, hogy a szóalakok a teljes szövegkorpusz hány százalékát fedik le (Fig 1. ábra).

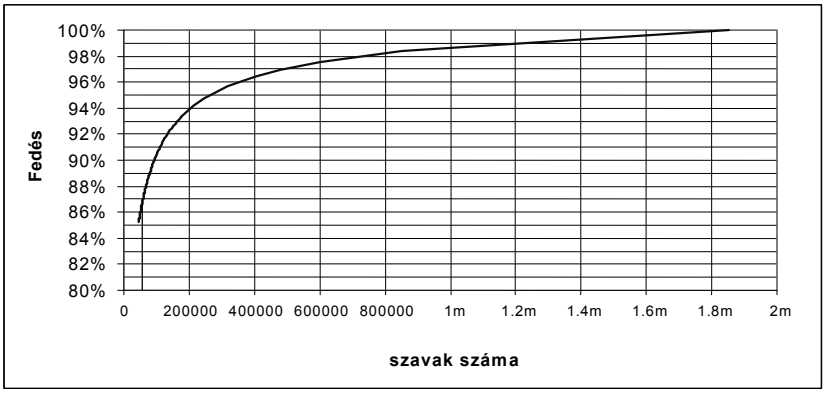

Fig. 1. Fedési diagram, maely megmutatja, hogy a szóalakok gyakoriság szerinti válogatása a teljes 80 millió szavas szövegkorpusz hány százalékát fedi le.

A szótár szóalakjaiból statisztikai gyűjtéseket végeztünk, amelynek két eredményét mutatjuk be, a szótagok szerinti eloszlást, valamint a hangok szerintit. Kiválogattuk a leggyakoribb szóalakokat a szótagszámuk szerint. Az eloszlást a Fig.2. ábra mutatja.

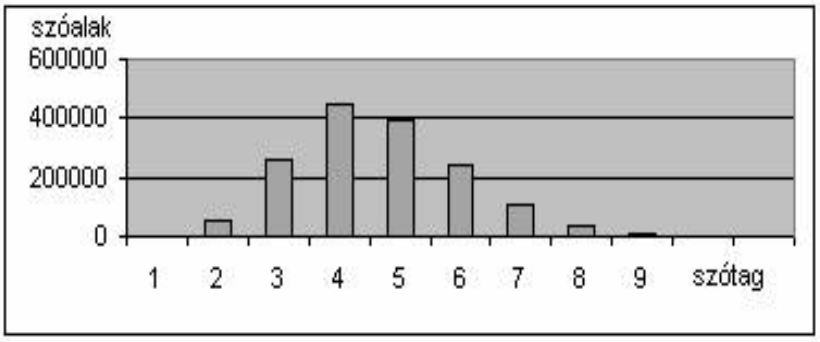

Fig. 2. A magyar elektronikus kiejtési szótár szóalakjainak eloszlása a szótagok száma szerint

Ezek szerint a magyarban a leggyakoribb szóalakok a 3, 4, 5 és 6 szótagú szavak. Megjegyezzük, hogy ez az eloszlás a szóalakokra vonatkozik, nem pedig a magyar szövegekben előforduló szavak általános eloszlására. Ez utóbbiról részletes adatokat [5]-ben találhatunk. Az eredményt a Fig. 3. ábra mutatja.

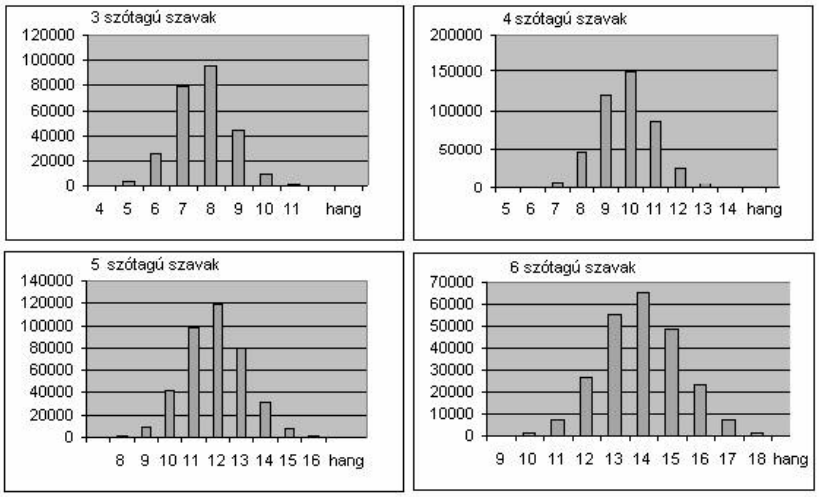

Fig. 3 A 3, 4, 5 és 6 szótagú magyar szóalakok hangszám szerinti eloszlása a kiejtési szótár szóalak állományában

Mindezek az adatok azt mutatják, hogy a kiejtési szótárunk jellemzően 7-15 hangot tartalmazó szóalakokat tartalmaz.

2.1 A hangjelölések és a hangátírás

A kiejtési szótár akkor használható jól, ha a hangjelölésekre nemzetközileg elfogadott jelrendszert alkalmaznak a fejlesztők. Mi is ezt az elvet követtük. Így bármely anyanyelvű felhasználó azonosítani tudja a kiválasztott magyar szó kiejtési formáját. A fonetikában használatos hangszimbólumokkal (IPA jelek) adjuk meg a szóalakok

hangalaki formáit. Az átíráshoz 9 magánhangzó és 25 mássalhangzó jelét használjuk, nem számítva a fonológiailag hosszú hangokat, amelyek hosszúságát kettősponttal jelöljük. A hangok szimbólumait a szótárban szögletes zárójelek közé tesszük.

1. Táblázat: A szótárban használt IPA jelek a magyar beszédhangok jelölésére

Betű	IPA jel	Betű	IPA jel	Betű	IPA jel	Betű	IPA jel
á	aː	b	b	n	n	zs	ʒ
a	ɔ	p	p	ny	ɲ	s	ʃ
o	o	d	d	j	j	cs	t͡ʃ
u	u	t	t	h	h	l	l
ü	y	g	g	v	v	r	r
i	i	k	k	f	f	dz	d͡z
é	eː	gy	ɟ	z	z	dzs	d͡ʒ
ö	ø	ty	c	sz	s		
e	ɛ	m	m	c	t͡s		

A szöveg-hang átalakítást egy nagy elemszámú szótár állományára csak automatikus feldolgozással lehet hatékonyan elvégezni. Hangátírási algoritmust kellett készíteni. Ennek fő elemeit a magyar szakirodalomban található kiejtési szabályok képezik. A megvalósított algoritmus alapfilozófiája a következő: meghatározunk szabályokat és alkalmazzuk azokat, majd felsoroljuk a kivételeket a szabályok alól. A kérdés minden esetben az volt, hogy mit nevezzünk szabálynak. Azt az elvet követtük, hogy egy hangátírási formánál felmértük annak hatókörét. A többségi előfordulást tekintettük szabálynak, és a kivétellistákba helyeztük el és oldottuk fel a kevesebb előfordulást. Így tudtuk helyes átírással lefedni a kiejtéssel kapcsolatos igen gazdag formációk közel teljes állományát.

A hangátírási algoritmusunk három szinten végzi a hangátírást: betű-hang szabályok, posztlexikális módosulások, kivételek kezelése (kivétel listák a szabályok mellett, illetve a kivétel szótár, amely önálló eleme a rendszernek).

Az alapot képező hangátírási szabályrendszer végzi általánosságban a betűképek hangszimbólumokká való átalakítását. Az eredmény sok esetben a végleges hangátírási forma (*ablak* = [ɔ b l ɔ k], azonban sok esetben nem. Ha nem, akkor a következő további feldolgozási formákat alkalmazzuk. Külön kivételként kezeljük a hangkivetés néhány esetét (*mondta* = [m o n t ɔ], *küldte* = [k y l t ɛ]ét (*mondta* = [m o n t ɔ], *küldte* = [k y l t ɛ]. A hangidőtartamok tekintetében két további alsóbb szinten módosulhat a kapott hangsorozat: hosszan írjuk, röviden mondjuk (*vá**ll**alat* = [v a: **l** ɔ l ɔ t], *ko**mm**unikál* = [k o **m** u n i k a: l]), *me**nny**ország* = [m ɛ ɲ o r s a: g], jo**bb**ra = [j o **b** r ɔ]. Ennek ellenkezője is előfordul, amikor röviden írjuk, hosszan ejtjük a szó valamelyik hangját ***USB*** = [**u:** e ʃ b e:], *NAT**O*** = [n a: t **o:**].

Posztlexikális szabályok

A magyar hangátírás legproblematikusabb része a hasonulások korrekt kezelése. Esetünkben azokkal foglalkoztunk, amelyeket a helyesírásunk nem jelöl. Ezek a szabályok a mássalhangzókat érintik. Ilyen beépített szabályok a részleges hasonulásból a zöngésedés, illetve zöngétlenedés, a képzés helye szerinti hasonulásból az [n]

hasonulása [m ɲ] hangokká (szí**n**pad = [s i **m** p ɔ d], po**n**ty = [p o ɲ c]). A teljes hasonulások, illetve az összeolvadás tekintetében hangsúlyozottan alkalmaztuk a kivétellistákat (például több esetben a kiejtés közeledik az írásképhez, főleg összetett szavak határán (*telj**e**s* = [t ɛ **j:** ɛ ʃ], de *feljavít* = [f ɛ l j ɔ v i: t]), kétséges = [k e: tʃ: e: g ɛ ʃ], de kétsávú = [k e: t ʃ a: v u]). A szótár végleges kialakítása során közel 50.000 szóalak kiejtése került kézi ellenőrzés formájában meghatározásra.

Kivétel szótár

A hangátírási szabályok harmadik fő modulja a kivétel szótár. Ide kerülnek elhelyezésre az idegen szavak, nevek, rövidítések és minden olyan kiejtési forma, amelyik az előző két modullal nem lefedhető (city = [s i t i], plasa = [p l a: z ɔ], Peugeot = [p ø ʒ o:], MTA = [ɛ m t e: ɔ:]). Esetünkben a magyar családnevek kiejtési meghatározásai is ebben a kivétel szótárban vannak (Kossuth = [k o ʃ u: t], Bernáthffy = [b ɛ r n a: t f i], Előrsy = [ő r ʃ i]), Unghváry = [u n g v a: r i]). A kivétel szótár közel 12.000 elemet tartalmaz.

3 A hangos szótár

A mai beszédtechnológiai eszközök lehetővé teszik, hogy egy kiejtési szótárban nagy számú hangzó példát is beépítsünk. Esetünkben 60.000 szóalak hangos formáját készítettük el (ennek nem az a formája, hogy felolvastuk a szavakat). Ezek mindegyike meghallgatható. Azért döntöttünk a hangos szótárrész elkészítése mellett, mert ezzel közelebbi támpontot tudunk nyújtani a nemzetközi közösségből kikerülő felhasználóknak a magyar hangsorépítés időszerkezeti viszonyairól is. A kiejtési hangjelekben ugyanis nincsenek információk a hangok időtartamairól, csak a fonológiai rövid-hosszú szembenállásról (a magyarban fontos a kiejtésben jó hosszan realizálni a hosszú hangokat). Ezért a tényleges hangzó forma meghallgatásával az alapvető magyar kiejtés időszerkezeti képét is megismerheti, érzékelheti a felhasználó.

A szóalakok kiválogatását a szótár teljes állományából végeztük, mégpedig a Fig. 2. ábrán megadott szóhosszúsági eloszlásnak megfelelően. Így minden hosszúságból arányos számú szó került a meghallgatható listába. A szavakat a Profivox magyar beszédszintetizátorral olvastattuk fel [4], így generáltunk 60.000 wav fájlt teljesen automatikusan. A szintetizátorban a hangidőtartamok kialakításánál a magyar kiejtési időmodell szabályai működtek [3]. A szóalakok kiejtési ritmusa megfelel a magyar köznyelvi kiejtés kritériumainak.

4 A kiejtési szótár szerkezeti felépítése

A kiejtési szótár az Interneten bárki számára hozzáférhető, használatához egy böngészőprogram szükséges. A szótárból a keresőgépekhez hasonló, könnyen kezelhető felületen keresztül kapjuk meg a kívánt szóalakok listáját. Az ilyen – webes környezetben megszokott – keresések során alapvető követelmény, hogy a keresőkérdésre illeszkedő találatok gyorsan jelenjenek meg a böngészőben.

A keresési idő csökkentése érdekében a kiejtési szótár 1,8 millió szóalakjának betűképét és hangsorát relációs adatbázisban (MySql) tároljuk. A 60.000 meghallgatható szóalak mindegyikéhez pedig egy-egy külön tárolt hangállomány (WAV) is kapcsolódik. A szótár helyigénye igen nagy, közel 3 GB, melynek nagy részét (kb. 95 %-át) a szintetizált szavak alkotják.

5 A szótár szolgáltatásai

Keresés a szótárban. A keresett szót magyar betűkarakterekkel kell megadni. Betűkapcsolatot is megadhatunk, ilyenkor minden olyan szót megkapunk, amelyikben az adott betűkapcsolat szerepel. A * karakter egyfajta joker szerepet tölt be a keresés megadásában. Például az „úszóedző*” megadására a szótár minden olyan szót megmutat, amelyik a * előtti karaktersorozattal íródik (*úszóedzővel, -nek, -ről, -mnek, -iket* stb.). A kikeresett szó magyar helyesírású betűképe mellett megjelenik a szó hangjainak sorozata IPA szimbólumokkal megadva. Ez a kiejtési forma. Amennyiben a szó mellett megjelenik egy hangszóró ikon az azt jelenti, hogy a szó meghallgatható.

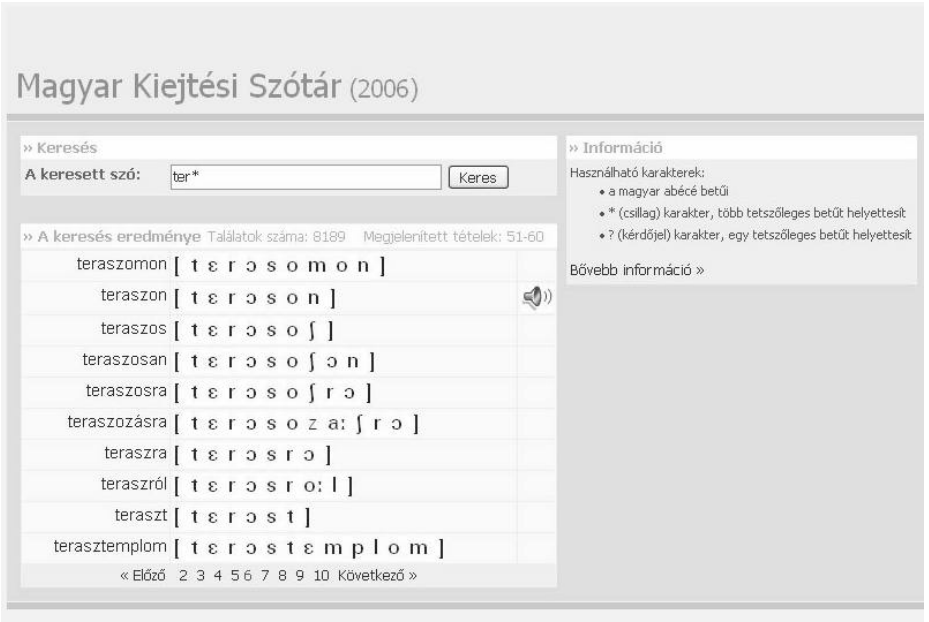

Fig. 4. Példa a kiejtési szótár találati listájára, a *ter** betűkapcsolatot tartalmazó szavakkal

A * karakter mellet használhatjuk a ? (kérdőjel) helyettesítő karaktert is, amely pontosan egy tetszőleges karaktert helyettesít. Ha a beviteli mezőbe a *v?r* karaktersorozatot visszük be, akkor a találati listában a *var, vár, ver, vér* szavakat kapjuk.

Amennyiben a keresőkérdésünk nem elég pontos – vagyis túl sok szó illeszkedik a kért betű- és helyettesítőkarakterek mintájára – egy lapozható találati listát kapunk, melyben az *Előző* vagy *Következő* linkeken kattintva kapjuk a szomszédos 10 találа-

tot. A Fig. 4. ábrán látható, hogy a *ter** keresőkérdésre 8189 db szóalak illeszkedik, amelyből éppen az 51-től 60-ig terjedő tételeket jelenítettük meg az ábrán. Az éppen listázott elemek közül a *teraszon* szólakot meg is hallgathatjuk.

Bibliográfia

1. A magyar nyelv könyve. Főszerkesztő: A. Jászó Anna.. Trezor Kiadó. 1991.
2. Fekete László: Magyar Kiejtési szótár. Gondolat Könyvkiadó. 1992.
3. Olaszy Gábor: Hangidőtartamok és időszerkezeti elemek a magyar beszédben. Nyelvtudományi Értekezések 155. Akadémiai Kiadó. 2006
4. Olaszy Gábor: Profivox- a legkorszerűbb hazai beszédszintetizátor. Beszédkutatás-2000. Szerk.: Gósy Mária. MTA Nyelvtudományi Intézet. Budapest. 2000. 167-179
5. Szende Tamás: Spontán beszédanyag gyakorisági mutatói. Nyelvtudományi Értekezések 81. Akadémiai Kiadó, 1973.
6. Tóthfalusi István: Kiejtési szótár. Tinta Kiadó. 2006. november 5.
7. Zainkó, Cs., Németh G.: Statistical Text Processing for Automatic Synthesis of Speech, Proc. of ECMCS2001 (EURASIP Conference on Digital Signal Processing for Multimedia Communications and Services), 2001. 644-647