A magyar nyelv betűstatisztikája beszédfeldolgozási szempontok figyelembevételével

Zainkó Csaba

Budapesti Műszaki és Gazdaságtudományi Egyetem
Távközlési és Médiainformatikai Tanszék,
zainko@tmit.bme.hu

Kivonat: A cikkben bemutatok egy új típusú betűstatisztikát, amely a klasszikus 44 betűs magyar ábécén alapuló eljárás továbbfejlesztése és egyesíti a betű- és a hangstatisztika előnyeit. A betűstatisztika készítését olyan módon egészítem ki, hogy figyelembe veszem a beszédfeldolgozás igényeit is. A módszer megkülönböztet betű szinten olyan jelenségeket is, amelyek csak a hangstatisztika szintjén lehet kezelni. Az új módszert a Magyar Nemzeti Szövegtáron tesztelem, összehasonlítom a módszert a klasszikus betűstatisztikával és a beszédfeldolgozásban használt hangstatisztikával.

1 Bevezetés

A magyar nyelvre sok célból készítenek különböző statisztikákat, például betű- és hangstatisztikákat, hangkapcsolódások statisztikáját vagy szótag- és szóstatisztikákat. Az első átfogó hangstatisztikát Szende Tamás közölte 1976-ban [8]. Betűstatisztikákat használtak például a titkosítás tudományában, de nyelv- és beszédfeldolgozási kutatásokban is fontos szerepe van. A betűstatisztikák általában a magyar 40 vagy a kiterjesztett 44 betűs ábécéből indulnak ki. A hangstatisztika két úton készülhet. Egyik módszer, hogy valamilyen hangzó anyagot gépi és kézi módszerrel annotálunk, a másik, hogy szövegből kiindulva fonetikus átiratot készítünk.

A jelen cikkben bemutatok egy újfajta megközelítést, amely alapjaiban egy betűstatisztika, de felhasznál olyan információkat is, amelyek általában csak a fonetikus átiratban állnak rendelkezésre. Ez azt jelenti, hogy figyelembe vesszük az osztályozáskor, a betűsorozat fonetikus reprezentációját is. Ehhez a betűstatisztikához újra definiálom a betű ilyen értelmű fogalmát. Például a *pech* szó klasszikus értelemben vett ch betűkapcsolatát az *sz* betűhöz hasonlóan kezelem, egy külön két karakterből álló betűnek tekintem, és másként kezelem például a *lánchíd* ch betűkapcsolatától, ahol külön c és h betű szerepel. A betű szintű osztályozás miatt viszont megmaradnak olyan információk is, amelyek a fonetikus átírás közben elvesznek. Például rendelkezésre áll az új típusú betűstatisztikában a /j/ hangként kimondott j és ly betű, vagy az /i/ hangként kimondott i és történelmi nevek végén gyakran szereplő y betű.

A cikk második részében ismertetem a Magyar Nemzeti Szövegtár állományaiból készített különböző statisztikákat, és összehasonlítom ezek eredményeit egymással.

A dolgozatban a hangokat ferde zárójelek közötti betűvel jelölöm, vagyis azzal a betűképpel, amelyik kimondása az adott hanghoz tartozik.

A kutatást a TELEAUTÓ projekten keresztül a Jedlik program támogatta.

2 Motiváció

A magyar helyesírás szabályai [1] szerint az ábécénk 40 betűt tartalmaz, amely egy vagy több írásjegyből áll (a, á, b, c, cs stb.). A számítógépes dokumentumokban az írásjegyeket karakterek formájában tároljuk, amelyek szintén egy vagy több karakteres betűt képviselnek. A továbbiakban – az egyszerűbb szóhasználat érdekében – a karakter elnevezést használom az írásjegy értelemben is. A szabályzat az ábécébe sorolja még – az idegen szavakban gyakran előforduló – q, w, x, y betűket is. Ez a 44 betűs ábécé forma, amit a gyakorlatban széles körben használnak.

A helyesírási szabályzat rendelkezik a régi magyar és az idegen eredetű nevek írásáról is. Ezekben lehetnek olyan betűk, amelyek két vagy több karakterből is állhatnak, ennek ellenére a karaktereket különálló betűkként kell a szabályok szerint kezelni bizonyos esetekben. Például a *Czuczor* /cucor/ családnevet betűrendbe soroláskor c + z betűk szerint kell besorolni, annak ellenére, hogy mai formában c betű lenne, illetve /c/ hangnak is ejtjük. A szabályok más esetekben viszont azt mondják ki, hogy az ilyen több karakterű betűket egyetlen betűként kell kezelni. Ilyen az elválasztási rendszer, amelynek szabályai előírják, hogy ezek a betűk nem elválaszthatóak. Például: *Ri-chárd*, *Mün-chen*, *Ben-czúr*.

A beszéd gépi feldolgozása során sok esetben célszerű a hangzó anyag mellett megadni annak valamilyen hangszintű, írott reprezentációját is. Ez az írott forma származhat a fonetikus átírásból, amely az elhangozó beszédhangok leirata hangjelölési szimbólumokkal megvalósítva. Sok esetben a feldolgozandó beszéd valamilyen szöveg felolvasása során keletkezik, az írott szöveg és a beszéd kapcsolata ilyenkor jól szinkronizálható, a fonetikus gépi algoritmusokkal és utólagos manuális feldolgozási lépésekkel elkészíthető. Az ilyen átírási folyamat összetett, időigényes és egyszerűsítéseket is tartalmazhat (például hangok különböző variánsait nem kezeli).

Az új típusú betűstatisztika használható kutatási feladatokra, a magyar szöveges állományok statisztikai tulajdonságainak vizsgálatára. Például: „Milyen gyakran jelent a ch betűkapcsolat /h/ hangot?” Felhasználható beszédadatbázisok készítésekor felolvasandó szövegállományok elemzésére, válogatására. Például megbecsülhető, hogy egy adott szöveg felolvasása esetén, a felolvasott szöveg egy kiválasztott hangból elegendő számút fog-e tartalmazni.

3 Módszer

A betűstatisztika készítésekor betűként a következő tulajdonságú karaktert vagy karaktersorozatokat értem:

- a 44 betűs magyar ábécének tagja, pl.: a, á, b, c, cs, …, sz, zs …
- a régi magyar családnevekben gyakran előforduló kettős betűk, pl.: cz, ch ejtése: /cs/ …
- idegen szavakban előforduló több karakteres betűk pl.: sch, ch ejtése:/h/ …
- kiejtve eltérő hangokhoz tartozhat. például.: h betű, amelynek kétfajta ejtését különböztetjük meg /$h_{néma}$, $h_{zöngés}$/ …

Néhány betűnél a különböző változatokat a szerint különböztetjük meg, hogy milyen hang keletkezik az adott szó kimondása esetén. Fontos megjegyezni, hogy ezek az osztályozások elsősorban beszédtechnológiai szempontok figyelembevételével történnek, nyelvészeti szempontból bizonyos döntések indokolatlanoknak tűnhetnek.

Szövegnek tekintem az általános szabályok szerint leírt magyar szövegeket, amelyek tartalmazhatnak számokat, írásjeleket és egyéb karaktereket. A statisztika készítésekor a nem betű típusú karaktereket figyelmen kívül hagyom (számok, relációs jelek stb.).

A szöveg feldolgozása szabályalapú algoritmussal történik, amely felhasználja a Profivox szövegfelolvasó rendszer fonetikai átíróját és szabálygyűjteményét [7], valamint különböző egyéb szótárakat is. Ilyenek a nagyméretű, magyar elektronikus kiejtési szótár [2], a névmondó tulajdonnév kiejtési szótárai [6], Huhyphn – magyar elválasztásiminta-gyűjtemény szószedete [5].

Az algoritmus a Profivox szövegfelolvasó fonetikai átírója szabálygyűjteményének egy részhalmazát használja. Ezek nagy részét a kettő vagy több karakterből álló betűk meghatározására vonatkozó szabályok teszik ki. A további szabályok azokra a betűkre vonatkoznak, amelyek kiejtésekor nem a betűhöz tartozó, nyelvileg szabályos fonéma realizáció (hang) keletkezik, hanem annak valamelyik speciális variánsa. Ilyen például a szóvégi zöngétlen /j/ hang (*lépj, hívj*). A szabályrendszer a magyar elektronikus kiejtési szótár információival van kiegészítve. Ez a szótár 1,5 millió magyar szóalak helyesírását és kiejtését adja meg párhuzamosított formában.

A régi magyar családnevekben előforduló betűk kezelését a névmondó szótár segítségével dolgozza fel az algoritmus. A régi betűvariációk nagy száma miatt, csak a gyakran előforduló személynevekben található eseteket kezeli a rendszer.

A magyarra jellemző a szóösszetétel. Az összetett szavak határán előfordulnak olyan karakterkombinációk, amelyek megegyeznek több karakteres betűkkel, de valójában nem azok. Ilyen például a *malacsült*, amelyben c + s betű található és nem cs. Az ilyen félreértelmezések elkerülésére az algoritmus a Huhyphn elvalasztásiminta-gyűjteményben található szószedetet használja. Az szószedet tartalmazza a szavak elválasztási lehetőségeit. Az algoritmus kihasználja ennek a szószedetnek azon tulajdonságát, hogy a helyesírási szabályok nem engedik meg a több karakteres betűkön belüli elválasztást, így az elválasztási helyeken korlátozza ezek hibás észlelését. Például a ma-lac-sült szó elválasztásából látható, hogy cs betű nem szerepelhet ebben a szóban. Az algoritmus figyelembe veszik a két karakterből álló hosszú változatát is pl.: tty, ssz, nny. Ezek két betűként szerepelnek a statisztikában zzs -> zs + zs.

A Profivox szabályokat tartalmaz a magánhangzók rövidülésére is, amely szintén használható a betűstatisztika finomítására. A magánhangzó rövidülése nagy változatosságot mutat a különböző beszélőknél, ezért ezek az információk nem adnak pontos eredményt, de tájékoztató adatnak megfelelnek. A további felhasználásuk esetében ezt figyelembe kell venni.

3.1 Speciális betűk, hangok

A ch betű az eredetétől függően /h/, /cs/ vagy /k/ hangot is jelölhet. Ennek megfelelően háromféle jelölést alkalmazunk: ch_h, ch_cs, ch_k (ezeknél a jelöléseknél az aláhúzás utáni betű jelöli a kiejtési formát)

A h betű hangalakja is többféle lehet. Néma /h/ valósul meg például a *cseh* /cse/ szóban. Csak ragozott formánál ejtjük a /h/ hangot (*csehül* /csehül/) Jelölése: $h_{néma}$

A h betű másik értelmezési formája a zöngés /h/ hang. Jelölése: $h_{zöngés}$

A j betű egyes esetekben zöngétlen /j/ hangként jelenik. Jelölése: $j_{zöngétlen}$

Az sch német eredetű betű is gyakran előfordul, a 3 karakteres hossza miatt fontos a külön kezelése. Jelölése: sch

Az y betű többnyire régi nevekben és idegen eredetű szavakban fordul elő általában /i/ vagy /j/ hangként valósul meg ejtéskor. Jelölésük: y_i, y_j

Az olyan rövid magán hangzókat is megkülönböztetjük, amely az átlagos hanghossznál rövidebbek. Jelölése: $a_{röv}$, $á_{röv}$ …

A régi írásmódú betűk jelölése: cz_c, ts_cs, eö_ö, tz_c, ck_k

3.2 Nyelvi anyag

A statisztikai elemzésekhez a Magyar Nemzeti Szövegtár (MNSZ) [4] anyagát használtam fel. A szövegtár 187,6 millió szövegszót tartalmaz, 5 nagyobb témát dolgoz fel. Tartalmaz sajtószövegeket, szépirodalmi műveket, tudományos, hivatalos és személyes szövegeket. A vizsgálatokhoz a teljes szövegtárat felhasználtam. A karakter és betűstatisztikához a vizsgált leghosszabb betűsorozat a szó volt, a beszédet reprezentáló fonetikai hangszimbólumok statisztikájához mondatokat használtam, ugyanis figyelembe vettem a szavak határán történő hangváltozási jelenségeket is.

3.3 A módszer korlátai

A statisztika egy gépileg gyűjtött és ellenőrzött szövegen alapul, amely tartalmaz hibákat. A összeállított szövegek mérete miatt manuális ellenőrzés nem jöhet szóba.

A felhasznált kivételszótárak szintén részben gépi módszereken alapulnak, egy része manuálisan ellenőrzött, de ennek ellenére tartalmazhatnak hibákat vagy hiányosak is lehetnek.

A kiejtéshez kapcsolódó szabályok a magyar nyelvi normát képviselik.

A vizsgált betűk meghatározása önkényes, a beszédfeldolgozás egyes szempontjait tartotta szem előtt, más felhasználás esetén a vizsgált betűk kiválasztása korlátozást jelenthet. Például a régi írásmódú betűk vizsgálata nem teljes körű, amely beszéd szempontjából megengedhető, de névelemzés esetén már további finomítás szükséges.

4 Eredmények

A vizsgált szöveg statisztikáját 3 formában készítettem el és az 1. táblázatban foglaltam össze. Az első két oszlop a karakterstatisztika, a második kettő a betűstatisztika, az utolsó két oszlop a hangstatisztika. A táblázatban szereplő számértékek megadják, hogy átlagosan 1000 elemből hány adott elem fordul elő. A hangstatisztika esetén a könnyebb összehasonlíthatóság miatt a betűképpel jelöltem a hangokat. A hangstatisztika teljesen gép módszerrel készült, manuális ellenőrzés nem történt a fonetikus átiratokon. A hangok esetében nincsenek megkülönböztetve a speciális esetek, variánsok. Az üres mezők azt jelentik, hogy az adott típusú statisztikában olyan elem nem szerepelt.

A betűstatisztika esetén az 1. táblázat csak a 44 betűs ábécében szereplő betűket tartalmazza, a speciális betűket a 2. táblázat tartalmazza. A második táblázatban a számértékek 1 millió elemre vonatkoznak.

A különböző statisztikák elkészítése nagyságrendileg eltérő erőforrást igényelt. A karakterstatisztika másodpercek alatt elkészült, a betűstatisztika több tíz perc, míg a hangstatisztika elkészítése 3-4 órát vett igénybe. A karakterstatisztika használata tehát akkor előnyös, ha gyors működés elengedhetetlen.

A karakterstatisztika hátránya, hogy csak 36 karakterre tartalmaz információkat, azokat is erősen torzítva. Az 1. táblázat s karakteréhez és betűjéhez tartozó gyakoriságokat összevetve látható, hogy a s karakter jóval gyakrabban fordul elő, mint az s betű, amelyet a kettős betűk szétdarabolása okozott. A karakterstatisztika tehát nyelv- és beszédfeldolgozási szempontokból egyáltalán nem vagy alig használható. Ennek ellenére az egyszerű programozhatósága miatt sok helyen használják betűstatisztika helyett.

Az itt szereplő hangstatisztika egyszerűsítéseket tartalmaz, csak 38 beszédhang szerepel benne. Ennek ellenére a karakterstatisztikához képest, jobban tükrözi a nyelv tulajdonságait, mert az egyszerűsítések fonetikailag megengedhető helyeken történtek. A betűstatisztikával összehasonlítva az elemek hasonló gyakorisággal szerepelnek, néhány esetben van csak eltérés, például az sz betű és az /sz/ hang között. Gósy [3] spontán beszédre készített hangstatisztikát, amelyben a magánhangzó-mássalhangzó arány 43% és 57% volt. Itt ez az arány 42% és 58% volt. A leggyakoribb hangokat összehasonlítva szintén hasonló számokat kaptunk, például a leggyakoribb /e/ hang gyakorisága Gósy statisztikájában 11.4%., itt 10.7%.

1. táblázat: karakter-, betű- és hangstatisztika

Karakter	1000-ből	Betű	1000-ből	Hang	1000-ből
a	89.37	a	92.85	/a/	90.21
á	35.95	á	37.58	/á/	37.99
b	19.66	b	20.56	/b/	18.28
c	7.64	c	3.97	/c/	6.10
		cs	3.91	/cs/	3.85
d	19.74	d	20.42	/d/	19.49
		dz	0.03		
		dzs	0.02		
e	98.70	e	101.31	/e/	106.59
é	33.46	é	35.02	/é/	35.69
f	9.18	f	9.59	/f/	9.04
g	33.80	g	22.69	/g/	19.82
		gy	12.70	/gy/	11.45
h	15.32	h	13.07	/h/	17.56
i	44.06	i	46.39	/i/	47.28
í	5.82	í	5.60	/í/	5.51
j	11.19	j	11.98	/j/	14.27
k	49.22	k	51.46	/k/	53.63
l	62.27	l	60.78	/l/	58.46
		ly	3.77		
m	35.00	m	36.56	/m/	36.61
n	58.12	n	53.78	/n/	54.37
		ny	7.02	/ny/	8.21
o	40.93	o	40.21	/o/	42.26
ó	10.03	ó	10.49	/ó/	9.95
ö	10.90	ö	11.39	/ö/	11.75
ő	8.94	ő	9.35	/ő/	9.68
p	11.14	p	11.65	/p/	12.42
q	0.04	q	0.04		
r	42.47	r	44.41	/r/	44.02
s	60.35	s	39.08	/s/	35.89
		sz	19.27	/sz/	24.55
t	79.42	t	82.72	/t/	81.58
		ty	0.27	/ty/	4.09
u	10.18	u	10.73	/u/	11.19
ú	3.01	ú	3.06	/ú/	2.69
ü	5.51	ü	5.85	/ü/	5.54
ű	1.86	ű	1.86	/ű/	1.74
v	19.89	v	20.80	/v/	21.52
w	0.28	w	0.29		
x	0.36	x	0.38		
y	22.71	y	0.21		
z	43.48	z	26.48	/z/	24.51
		zs	0.73	/zs/	2.21

2. táblázat: betűstatisztika speciális betűkre

hang	1000000-ból
ch_cs	28.12
ch_h	129.04
ch_k	2.33
ck_k	4.96
cz_c	25.24
eö_ö	7.16
$h_{néma}$	60.27
$h_{zöngés}$	2470.19
$a_{röv}$	627.99
$á_{röv}$	111.70
$e_{röv}$	2010.52
$é_{röv}$	15.02
$i_{röv}$	1000.29
$o_{röv}$	2667.03
$ö_{röv}$	36.57
$u_{röv}$	144.33
sch	85.09
ts_cs	34.25
tz_c	11.84
y_i	65.27
y_j	111.27
$j_{zöngétlen}$	3.59

A betűstatisztika 1. táblázatban szereplő részén mind a 44 betű statisztikáját megtalálhatjuk. Ez a 44 betű az összes betűstatisztikában szereplő betű 99%-at adja, a speciális betűk csak 1%-ot tesztnek ki a vizsgált szövegekben. A ábécé betűi közül a dz, dzs, q szerepel nagyon ritkán, a 1 millió szóban átlagosan 20-40 db található meg. Leggyakrabban a vártnak megfelelően az e betű szerepelt.

A 2. táblázatban szereplő rövid magánhangzók közül a rövid á 1 millió szóból átlagosan 111-szer szerepel. Ez a kis érték több okra vezethető vissza. A rövid /á/ hang a *fájl*, *bájt* szavakban található, amit gyakoribbnak volt várható. Egyik ok, hogy a szöveg jelentős részben tartalmaz irodalmi alkotásokat, amelyekben ez a szó nem szerepel. A másik ok, hogy a szöveggyűjteményben nagy számban helyesírásilag hibásan szerepelnek a *bájt* és *fáj*l szavak, az angol *file* és *byte* formában.

A szó végén szereplő zöngétlen /j/ hang kis számban szerepel a szövegekben. Ennek oka az lehet, hogy a felszólító módú igék írott szövegben kevésbé gyakoriak, inkább a beszélt nyelvben találhatóak meg.

Az y betű /j/ hangként való realizációja gyakoribb, mint az /i/ hangként való megjelenése. Ez abból adódik, hogy idegen nevek többször szerepelnek (például Toyota) mint a történelmi nevek (például Desseffy).

A ch betű leggyakrabban /h/ hangként jelenik meg, majd /cs/ hang a második leggyakoribb formája, /k/ hangként ritkán ejtjük.

A 3. táblázatban a betűstatisztika található gyakorisági sorrendben.

3. táblázat: Betűstatisztika gyakorisági sorrendben

Betű	db/1000	Betű	db/1000	Betű	db/1000	Betű	db/1000
e	101.31	d	20.42	ú	3.06	y_i	0.065
a	92.85	sz	19.27	$o_{röv}$	2.67	$h_{néma}$	0.060
t	82.72	h	13.07	$h_{zöngés}$	2.47	q	0.040
l	60.78	gy	12.70	$e_{röv}$	2.01	$ö_{röv}$	0.037
n	53.78	j	11.98	ű	1.86	ts_cs	0.034
k	51.46	p	11.65	$i_{röv}$	1.00	ch_cs	0.028
i	46.39	ö	11.39	zs	0.73	dz	0.026
r	44.41	u	10.73	$a_{röv}$	0.63	cz_c	0.025
o	40.21	ó	10.49	x	0.38	dzs	0.017
s	39.08	f	9.59	w	0.29	$é_{röv}$	0.015
á	37.58	ő	9.35	ty	0.27	tz_c	0.012
m	36.56	ny	7.02	y	0.21	eö_ö	0.007
é	35.02	ü	5.85	$u_{röv}$	0.14	ck_k	0.005
z	26.48	í	5.60	ch_h	0.13	$j_{zöngétlen}$	0.004
g	22.69	c	3.97	$á_{röv}$	0.11	ch_k	0.002
v	20.80	cs	3.91	y_j	0.11		
b	20.56	ly	3.77	sch	0.09		

5 Összegzés

Az új típusú betűstatisztika alkalmas arra, hogy szövegekről, korpuszokról olyan statisztikai információkhoz jussunk egy lépésben, amelyhez csak a klasszikus betűstatisztika és a hangstatisztika (fonémastatisztika) együttes elemzésével juthatunk. Megadtam egy lehetséges betűosztályozást, amellyel egy kibővített statisztikát lehet készíteni magyar nyelvre. A cikkben továbbá összehasonlító elemzést adtam karakterstatisztikára, az általam módosított értelmű betűstatisztikára és az ugyanazon szövegkorpuszból készített hangstatisztikára. A statisztikák a Magyar Nemzeti Szövegtár alapján készültek.

Hivatkozások

1. A magyar helyesírás szabályai. MTA Budapest: Akadémiai. Kiadó (1985)
2. Abari K., Olaszy G., Kiss G., Zainkó Cs.: Magyar kiejtési szótár az Interneten. In: Alexin Z., Csendes D. (szerk.) MSZNY (2006) 223-230
3. Gósy M.: Fonetika, a beszéd tudománya. Osiris. Budapest (2004)
4. Magyar Nemzeti Szövegtár. MTA – Nyelvtudományi Intézet http://corpus.nytud.hu/mnsz/
5. Nagy B.: Huhyphn: magyar elválasztásiminta-gyűjtemény. http://www.tipogral.hu/ (2008)
6. Németh G., Zainkó Cs., Kiss G., Fék M., Olaszy G., Gordos G.: Language Processing for Name and Address Reading in Hungarian In: IEEE NLP-KE Beijing, Kína (2003) 238-243
7. Olaszy G., Németh G., Olaszi P., Kiss G., Zainkó Cs., Gordos G.: Profivox - a Hungarian TTS System for Telecommunications Applications In: IJST 3-4: 201-215 (2000)
8. Szende T.: A beszédfolyamat alaptényezői. Akadémiai Kiadó (1976)