Az új magyar Braille-rövidírás korpuszvezérelt kialakításának lehetőségei

Sass Bálint
MTA Nyelvtudományi Intézet
sass.balint@nytud.mta.hu

1. Célkitűzés

Braille-rövidírás: rövidítési szabályok rendszere.
Célja: (1) kinyomtatva kisebb hely; (2) írás (jegyzetelés) meggyorsítása; (3) olvasás meggyorsítása.
Nyelvenként külön rendszer. Magyar de facto szabvány: *„kis" rövidírás* (az 50-es évek óta)

Feladat (az MVGYOSZ kérése): „a mai nyelvhasználatot figyelembe vevő új rövidítésekkel bővíteni a szabályrendszert, hogy a rövidítési képesség a jelenlegi nagyjából 10%-ról ≈20%-ra növekedjen."

Elv: a lehető leggyakoribb elemeket (karaktersorozatokat) a lehető legrövidebbre rövidíteni.
→ korpuszgyakorisági adatok!

Kutatási kérdés: hogyan lehet korpuszgyakorisági adatok alapján, a lehető legkisebb emberi beavatkozással előállítani egy nagy rövidítési képességű, „objektíve jó" új magyar rövidírást.

Szűk keresztmetszet: a rövidítéshez rendelkezésre álló jelek száma: mindössze 64 egykarakteres jel (abból is főként a ritkák a jók)

5 gyakorisági érték pozíció szerint: (1) szó elején (beg), (2) szó belsejében (inn), (3) szó végén (end), (4) önálló szóként (sta), (5) összesen (all).

1. adott jelet rövidítésként csak abban a pozícióban jó használni, ahol ritka (pl.: írásjel szó elején)
2. adott rövidítésjelet különféle pozíciókban eltérő jelentéssel használhatunk

Pl.: német: r(mm)=x és r(immer,sta)=x ← x ritka betű + mm önálló szóként ritka
Pl.: magyar: r(et)=@ és r(Serint,sta)=@ ← @ ritka betű + et önálló szóként ritka

A magyar Braille-írás

a [1]	b [2]	c [3]	d [4]	e [5]	f [6]	g [7]	h [8]	i [9]	j [0]
k	l	m	n	o	p	ö, {	r	s	t
u	v	x	y	)	q	#	ü, [	(, ^	ű,]
é, \	z	cs	gy	sz	ny	ő, }	ty	ó	w
,	;	:	PFX?	?	!, +	=	,,	*	”
í	ú	PFX:szám	zs	.	-				
á, ‘	@ , |	ly	PFX?	PFX:nb	PFX?	PFX?			

A magyar „kis" rövidírás

- nagybetűjel törlése 2,7%
- vesszőt követő szóköz törlése 1% (3,7%)
- határozott névelők: rövidítés + követő szóköz törlése r(a)=, r(az)=X 1,3% (5%)
- 7 szóvégi rövidítés (ragok) 1,6% (6,6%)
  r(ban/ben)=b r(ból/ből)=bX r(hoz/hez/höz)=hX r(ként)=kX r(ról/ről)=rX r(tól/től)=tX r(val/vel)=v
- 16 egyjelű szórövidítés (önálló szóként, összetett szóban is) 1,8% (8,4%)
  r(Cak)=C r(de)=d r(és)=é r(hoG)=h r(is)=i r(íG)=í r(kell)=k r(leS)=l r(meL)=m r(nem)=n r(óta)=ó r(pedig)=p r(tehát)=t r(után)=u r(úG)=ú r(van)=v
- 21 kétjelű szórövidítés (bárhol!) 0,9% (9,3%)
  r(aNNi)=ai r(boldog)=bg r(eNNi)=ei r(gond)=gd r(függ)=gg r(Gors)=Gs r(keres)=ks r(mind)=md r(mint)=mt r(orSág)=og r(olvas)=os r(öSSe)=öe r(pont)=pt r(pénz)=pz r(rövid)=rd r(forr)=rr r(Sabad)=Sd r(tanáC)=tC r(teljes)=ts r(világ)=vg r(volt)=vt
- hónapnevek (hivatalosan nem része a kis rövidírásnak) 0,3% (9,6%)

Példa: Ade ha Serinted vaG mások Serint ez megalkuvás volt, akkor ezt Somorú Sívvel kéNtelen vaGok elfogadni.

Rövidítve: d ha Serinted vaG mások Serint ez megalkuvás vt,akkor ezt Somorú Sívv kéNtelen vaGok elfogadni.

2. Algoritmus

Minden létező karaktersorozat (adott hosszig) egy egységes gyakorisági listába gyűjtünk, az elemeket mind az 5 gyakorisági értékükkel külön bejegyzésként szerepeltetve (= egy elem 5-ször fog szerepelni)

Rendezés: aktuális rövidítési képesség – $\mathrm{rk}()$ – szerint:

$$\mathrm{rk}(w, r(w,t), t) = [l(w) - l(r(w,t))] * \mathrm{fq}(w,t)$$

w az eredeti rövidítendő karaktersorozat, $r(w)$ a rövidítés, t a pozíció szerinti típus ($\in$ {beg, inn, end, sta, all}), $l()$ a hossz (karakterszám), $\mathrm{fq}()$ a gyakoriság.

Rövidítéshossz kezdetben: $n = 1$
Potenciális rövidítésjelek listája := legritkább jelek sorra

Hozzárendeljük a lista első helyén álló rövidítendő elemhez az alkalmas legritkább elemet rövidítésként.

– rövidítésjelek hatékony felhasználása: az 1. + 2. megfontolás szerint
– rövidítésjel literálisan: külön prefix-szel. Pl.: ⠓ = h (és nem hogy) – ez levonódik a rövidítési képességből a $\mathrm{fq}(\mathrm{h}, \mathrm{sta})$-nak megfelelően

Kiszámoljuk az új szabálynak a további potenciális szabályok rövidítési képességét befolyásoló hatását (még kidolgozandó).

Ha nem találunk alkalmas n hosszúságú rövidítésjelet az aktuális rövidítendő elemhez, akkor a továbbiakban $n + 1$ hosszúságú keresünk hozzá.

Az új helyzetnek megfelelően frissítjük a $\mathrm{rk}()$ értékeket, újrarendezzük a listát, és vesszük ismét a lista elején álló elemet.

3. Használhatóság

Egyaránt fontos: rövidség + kényelmes használhatóság:

– jó **olvashatóság** (tapintás útján jó felismerhetőség)
– könnyű **megtanulhatóság** (kevés, egyszerű szabály)

„Jó" rövidítés: mindig azonos jelentésű, a szó kezdő és záró betűjéből, illetve a szót alkotó jellegzetes mássalhangzóból áll.

- **egykarakteres** rövidítésjelek: értékesek, $rk()$ sok, kevés van, írásjelek.
  érdemes megengedni: korlátozás nélkül bárminek a rövidítésére felhasználhassuk őket + különböző pozíciókban különféle jelentéssel bírhassanak
  alkalmas rövidítésjel: szigorúan gyakorisági alapon, a lehető legnagyobb rövidítés elérése érdekében. (Könnyű megtanulhatóság itt: vö: mm és et.)
- **többkarakteres** rövidítésjelek: a fenti követelmény könnyebben teljesíthető, esetleg automatikusan is.
  alkalmas rövidítésjel: az olvashatósági szempontok figyelembe vételével: ti. itt általában számos azonos gyakoriságú (ritka) rövidítésjel közül választhatunk.

Valamint: mikor a rövidítendő elem ritka típusa helyett keresünk másik könnyen megjegyezhető rövidíthető elemet adott rövidítésjelhez.

Első rendszer

Kézzel készült a lefektetett elvek alapján.
A kis rövidírás után alkalmazandó.

r(t+␣)=#	r(k+␣)=q	r(␣+m)=;	r(el)==
r(n+␣)=@	r(␣+a)=y	r(s+␣)=w	r(␣+k)=x
r(et)=!	r(en)=?	r(te)=T	r(le)=(
r(er)=)	r(al)=Z	r(at)=:	r(an)=ű

16 szabály: 7,9% rövidítés
Kis rövidírással összesen: 9,6% + 7,9% = 17,5%

Példamondat rövidítve:
d ha S)inTd vaG;ásoqS)in#ez;egZkuváwvt,akkor ez#Somorú SívvxéNt=e@vaGoq=fogadni.

Megközelíthető a 20%, olvashatóság rossz.

Második rendszer

Automatikusan generálva.
A kis rövidírás után alkalmazandó.

r(el)=#		
r(et|inn/end)=@	r(me|beg)=@	r(Serint|sta)=@
r(en)=q	r(te)=y	r(le)=w

7 (5?) szabály: 3,3% rövidítés
Kis rövidírással összesen: 9,6% + 3,3% = 12,9%

Példamondat rövidítve:
d ha Serinyd vaG mások @ ez @galkuvás vt,akkor ezt Somorú Sívv kéNt#q vaGok #fogadni.

Megfelelő. További/végleges szabályok kialakítása szükséges a terveknek megfelelően.

4. Tervek

Kidolgozni a hangrend szerint összevonandó, egy jellel rövidítendő toldalékok (pl.: -ság/-ség) listáját.

Eljárás, ami lehetővé teszi az épp létrehozott új szabálynak a további potenciális szabályok rövidítési képességét befolyásoló hatásának kiszámítását.

Használhatósági követelmények automatikus kezelése.

Megfelelő kompromisszum kialakítása a minél nagyobb rövidítési képesség és a használhatóság szempontjai között. Annak érdekében, hogy a potenciális felhasználók valóban szívesen használják majd, szükség van a kialakított rendszer közvetlen vakok általi tesztelésre.

IX. Magyar Számítógépes Nyelvészeti Konferencia,
2013. január 7-8., Szeged