Az új magyar Braille-rövidírás kialakítása

Sass Bálint
sass.balint@nytud.mta.hu

MTA Nyelvtudományi Intézet
Nyelvtechnológiai és Alkalmazott Nyelvészeti Osztály

2013. november 11.

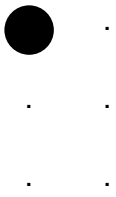
Braille-írás

- A vakok által világszerte használt tapintáson alapuló írásmód.
- 6 pont különböző elrendezéseiből $\longrightarrow$ 64 féle karakter

1	4
2	5
3	6

Braille-írás

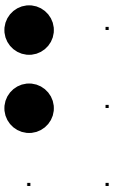
- A vakok által világszerte használt tapintáson alapuló írásmód.
- 6 pont különböző elrendezéseiből $\longrightarrow$ 64 féle karakter



a [1]

Braille-írás

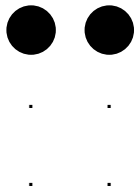
- A vakok által világszerte használt tapintáson alapuló írásmód.
- 6 pont különböző elrendezéseiből $\longrightarrow$ 64 féle karakter



b [1-2]

Braille-írás

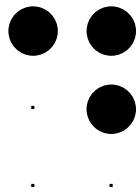
- A vakok által világszerte használt tapintáson alapuló írásmód.
- 6 pont különböző elrendezéseiből $\longrightarrow$ 64 féle karakter



c [1-4]

Braille-írás

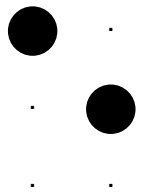
- A vakok által világszerte használt tapintáson alapuló írásmód.
- 6 pont különböző elrendezéseiből $\longrightarrow$ 64 féle karakter



d [1-2-5]

Braille-írás

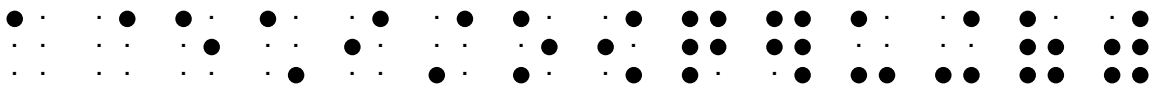
- A vakok által világszerte használt tapintáson alapuló írásmód.
- 6 pont különböző elrendezéseiből $\longrightarrow$ 64 féle karakter



e [1-5]

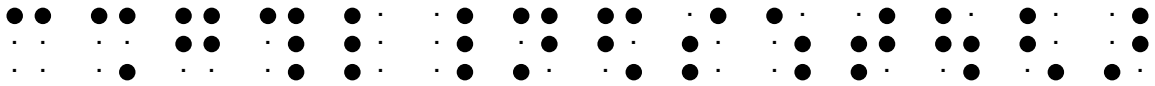
A magyar Braille-írás

- önálló jelek az *ékezetes* magánhangzóknak



a á e é i í o ó ö ő u ú ü ű

- önálló jelek a *kettős* mássalhangzóknak



c cs g gy l ly n ny s sz t ty z zs

- „tükrözés” – kivéve: *e*, *i*, *c*, *g*

- ha Braille-írást sima szövegként ábrázolunk, a kettős mássalhangzókat egy karakterrel – az első karakter nagybetűs változatával – jelöljük: C, G, L, N, S, T, Z – pl. *maGar*

Braille-*rövidírás*

A különféle nyelvekre (pl. angol, német, magyar) kidolgozott Braille-rövidírások az általános Braille-írást rövidítési, tömörítési szabályokkal egészítik ki.

Rövidírás használatával:

- a nyomtatott forma kevesebb helyet igényel;
- az írás, jegyzetelés gyorsul;
- az olvasás, felismerés gyorsul.

A magyar Braille-rövidírás: az ún. „kis” rövidírás

Magyarországon használatos sztenderd rövidírás.
Elemei:

1. nagybetűjel ⠠ törlése $\sim \mathrm{rk} = 2\%$
2. vessző utáni szóköz törlése $\sim \mathrm{rk} = 2\%$
3. határozott névelő utáni szóköz törlése, *az* rövidítése $\sim \mathrm{rk} = 2\%$
4. 44 szabály: $\sim \mathrm{rk} = 4\%$
   - 7 db szóvégi rövidítés:
     ban/ben $\longrightarrow$ b, ból/ból $\longrightarrow$ bl, hoz/hez/höz $\longrightarrow$ hz . . .
   - 16 db egyjelű szórövidítés:
     Cak $\longrightarrow$ C, de $\longrightarrow$ d, és $\longrightarrow$ é . . .
   - 21 db kétjelű szórövidítés:
     aNNi $\longrightarrow$ ai, boldog $\longrightarrow$ bg, gond $\longrightarrow$ gd . . .

$\longrightarrow$ a kis rövidírás mért rövidítési képessége: $\mathrm{rk} = \mathbf{9{,}4\%}$

Kezdeményezés + követelmények

Kezdeményezés az MVGYOSZ Braille-bizottsága részéről:
bővítsük a kis rövidírást új szabályokkal annak érdekében,
hogy a rövidítési képessége jelentősen növekedjen.

Használhatósági követelmények:

- az ismert kis rövidírást egészítse ki;
- *jó olvashatóság*: a rövidítések emlékeztessenek az eredetire;
- *jó felismerhetőség*: tapintás útján könnyen felismerhető jelek alkalmazása;
- *könnyű megtanulhatóság*: kevés, egyszerű szabály.

Kompromisszumos javaslatot készül, mely. . .

- törekszik a maximális rövidítési képesség elérésére,
- miközben megfelel a használhatósági feltételeknek is.

Az új rövidírás létrehozása

Mit rövidítsünk?

A magyar nyelv gyakorisági adatai alapján a maximális rövidítési képességgel bíró ideális rövidírás automatikusan kialakítható.

Elv: a lehető leggyakoribb betűkapcsolatokat kell a lehető legrövidebbre rövidíteni.
Pl.: *ty* $\leftrightarrow$ *et* $\sim$ 20$\times$-os különbség a gyakoriságban

$\rightarrow$ automatikus algoritmus

Mire rövidítsünk?

Szakértői döntések a gyakorisági adatok és a használhatósági megfontolások alapján. Akár az automatikus eredményt felülbírálva.

$\rightarrow$ manuális munka

Az algoritmus működése

1. Számba vesszük az összes rövidíthető nyelvi elemet:
   előállítjuk a karaktersorozatok gyakorisági listáját.
2. Kiszámoljuk az elemek rövidítési képességét a következő módon:
   $\mathrm{rk}(w, r(w)) = [l(w) - l(r(w))] \cdot \mathrm{fq}(w)$
   w az eredeti rövidítendő karaktersorozat, $r(w)$ a rövidítés,
   $l()$ a hossz (karakterszám), $\mathrm{fq}()$ a gyakoriság.
   Pl.: a *sem* és a *maGar* gyakorisága nagyjából azonos
   $\longrightarrow$ érdemesebb az utóbbit két karakterre rövidíteni.
3. Rendezzük a gyakorisági listát rövidítési képesség szerint.
4. Az *első* elemhez (maximális rövidítési képesség)
   hozzárendelünk egy megfelelő rövidítést.
5. Ez után *alkalmazzuk* a most kialakított rövidítő szabályt
   az egész listára.
6. $\longrightarrow$ 2

Működési példa

2 karakterre rövidítünk

A gyakoriségi lista eleje:

w	$\mathrm{fq}(w)$	$l(w) - l(r(w))$	$\mathrm{rk}(w)$
Ser	10901	$\times 1 =$	10901
Serint	2515	$\times 4 =$	10060
maGar	3326	$\times 3 =$	9978

Az első elemet rövidítjük: $Ser \longrightarrow Sr$

Működési példa

2 karakterre rövidítünk

A gyakorisági lista eleje:

w	$\mathrm{fq}(w)$	$l(w) - l(r(w))$	$\mathrm{rk}(w)$
Ser	10901	$\times 1 =$	10901
Serint	2515	$\times 4 =$	10060
maGar	3326	$\times 3 =$	9978

Az első elemet rövidítjük: $Ser \longrightarrow Sr$

w	$\mathrm{fq}(w)$	$l(w) - l(r(w))$	$\mathrm{rk}(w)$
maGar	3326	$\times 3 =$	9978
...			
Srint	2515	$\times 3 =$	7545
...			
Sr	10901	$\times 0 =$	0

Működési példa

2 karakterre rövidítünk

A gyakorisági lista eleje:

w	$\mathrm{fq}(w)$	$l(w) - l(r(w))$	$\mathrm{rk}(w)$
Ser	10901	$\times 1 =$	10901
Serint	2515	$\times 4 =$	10060
maGar	3326	$\times 3 =$	9978

Az első elemet rövidítjük: $Ser \longrightarrow Sr$

w	$\mathrm{fq}(w)$	$l(w) - l(r(w))$	$\mathrm{rk}(w)$
maGar	3326	$\times 3 =$	9978
...			
Srint	2515	$\times 3 =$	7545
...			
Sr	10901	$\times 0 =$	0

A következő rövidítendő elem a *maGar* lesz.

A *Serint* az új szabály miatt sokkal lejjebb csúszott. Ha mégis sorra kerül, tovább rövidíthetjük (így lesz).

A *Sr* kétkarakteres rövidítéssel nyilván nem rövidíthető tovább.

Rövidítések megválasztása és használhatóság 1/2

Több nagyon hatékony rövidítési ötletet
a használhatósági követelmények miatt végül elvetettünk:

- szóközt elnyelő szabályok (pedig: rk(t␣,T) $\approx$ 1,2%!);
- különböző helyzetekben különböző jelentéssel bíró rövidítések (német példa: r(immer)=r(mm)=x);
- „értelmetlen” elemek (pl.: *et*, *en*, *er*, *ele*, *ala*, *tás*, *ter*) rövidítése (pedig: rk(et,T) $\approx$ 0,8%!).

Elv: „értelmest értelmesre”

= csak értelmezhető szóelemet rövidítünk
és olvasható rövidítésjelet rendelünk hozzá

Rövidítések megválasztása és használhatóság 2/2

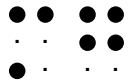
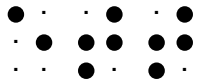
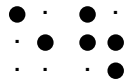
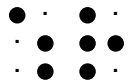
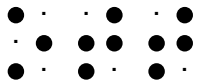
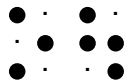
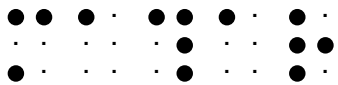
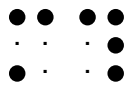
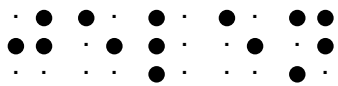
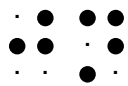
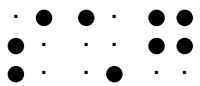
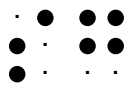
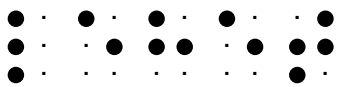
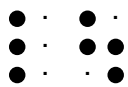
- A javaslatban csak *kétkarakteres* rövidítéseket használunk. A legtöbb esetben könnyű kiválasztani egy olyan ritka kétkarakteres rövidítést, ami megfelelően illeszkedik a rövidítendőhöz, pl.: *r*( [maGar]) =  [mG].

- A leghatékonyabb szabályok a nagyon gyakori betűkapcsolatok (pl.: ’et’, ’el’) *egy karakterre* való rövidítései lennének. Van néhány nagyon ritka jel, ami itt rövidítésként alkalmazható lenne: pl.:  ,  ,  [@],  [q],  [=],  [*],  [ty],  [w], de használatuk, főleg sok ilyen rövidítésé, nagyban rontja az olvashatóságot. Bizonyos elemek egykarakteres rövidítése még megfontolás alatt van: *el*, *tt*, *meg*, *Ser*, *ás/és*, *eG*.

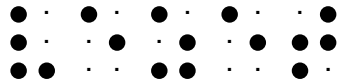
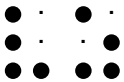
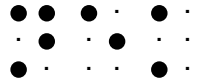
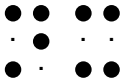
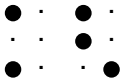
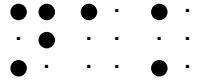
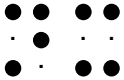
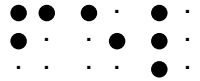
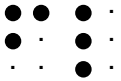
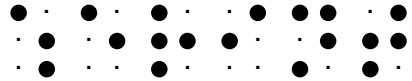
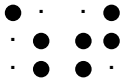
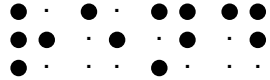
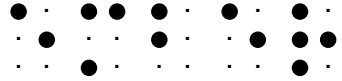
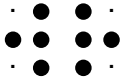
A módszer összefoglalása

- Szigorúan automatikus úton történik a bemutatott algoritmussal az épp aktuális (következő) legjobban rövidíthető elem meghatározása. Egy ajánlott, elég jó olvashatóságú rövidítést is automatikusan megad hozzá a rendszer.
- A használhatósági feltételeknek nehéz teljesen automatizált módon megfelelni, ezért szükséges az automatikusan kialakított szabályrendszer interaktív módosítása, manuális véglegesítése:
  - meghatározzuk a konkrét lehető legjobb rövidítésjelet, adott esetben megfelelő *kinézetű* jel létrehozásával;
  - szükség esetén azt is kiköthetjük, hogy az adott elemet – tudva, hogy ezzel veszítünk a rövidítési képességből – mégsem rövidítjük;
  - a gyakorisági adatok ismeretében megszoríthatjuk a rövidítés pozícióját (pl.: csak szó végén).

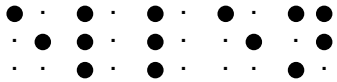
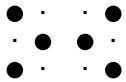
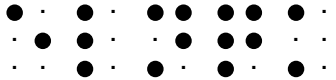
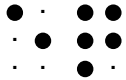
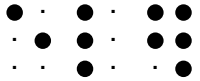
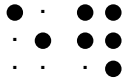
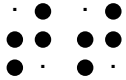
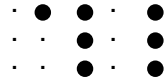
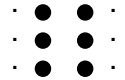
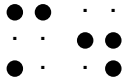
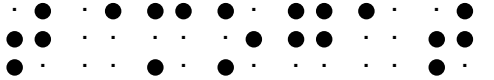
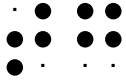
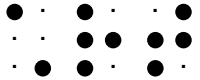
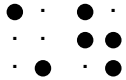
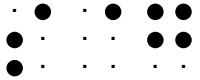
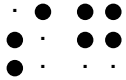
A javaslat 1/3

	rövidítendő	rövidítés	megjegyzés
1.	 [meg]	 [mg]	szó elején (98%)
2.	 [ett]	 [eT]	= ott 
3.	 [Ser]	 [Sr]	
4.	 [ott]	 [oT]	= ett 
5.	 [maGar]	 [mG]	
6.	 [jelen]	 [jn]	
7.	 [ség]	 [sg]	= ság
8.	 [lehet]	 [lT]	*lh* túl gyakori

A javaslat 2/3

	rövidítendő	rövidítés	megjegyzés
9.	 [vezet]	 [vz]	
10.	 [nek]	 [nx]	= nak  szó végén
11.	 [köz]	 [kz]	
12.	 [nak]	 [nx]	= nek  szó végén
13.	 [fel]	 [fl]	szó elején (86%)
14.	 [Serint]	 [St]	szó elején (95%)
15.	 [rend]	 [rH]	csúsztatott *d* (*rd*, *rn* nem jó)
16.	 [ember]	 [wr]	jel kinézete

A javaslat 3/3

	rövidítendő	rövidítés	megjegyzés
17.	 [ellen]	 [oó]	jel kinézete
18.	 [elnök]	 [eö]	
19.	 [elő]	 [eő]	
20.	 [tart]	 [tt]	szó elején
21.	 [áll]	 [Ll]	jel kinézete
22.	 [mond]	 [mH]	csúsztatott *d*
23.	 [támogat]	 [tg]	
24.	 [ért]	 [éT]	
25.	 [ság]	 [sg]	= ség

A javaslat kiértékelése

23 (25) szabály.

Rövidítendők listája:
áll ellen elnök elő ember ért ett/ott fel- jelen köz lehet maGar meg-mond -nak/-nek rend ság/ség Ser Serint(-) támogat tart(-) vezet

Kis rövidírás rövidítési képessége: rk $= 9,4\%$

A fenti szabályokkal kiegészített...

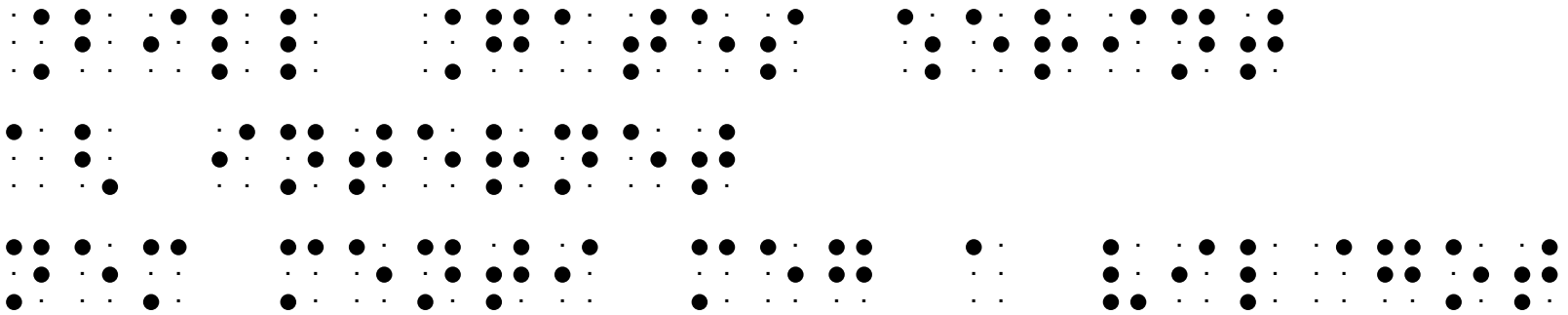
új rövidírás mért rövidítési képessége: rk $=$ **12,3%**

A rövidítési képesség növekedése: +2,9 százalékpont (+30%)

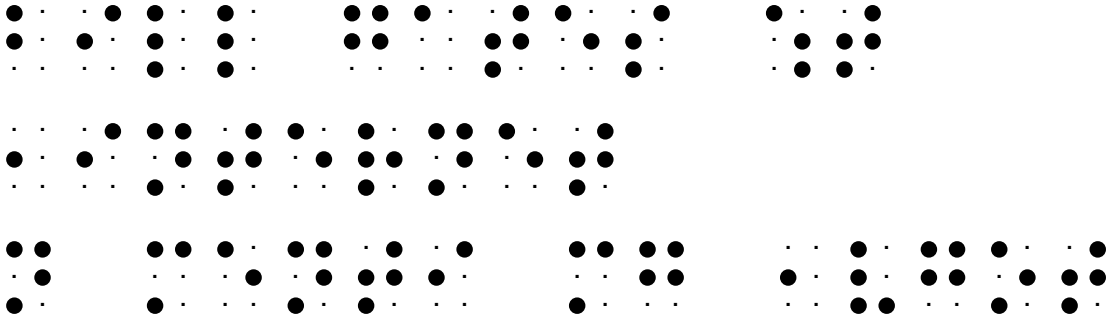
Kevés szabály, jelentős növekedés.
Szükség esetén egyszerűen tovább bővíthető a jövőben.

Példa

eredeti: *Bill Gates szerint az internet nem menti meg a világot*



rövidítve: *bill gates szt ,internet n menti mg ,vgot*



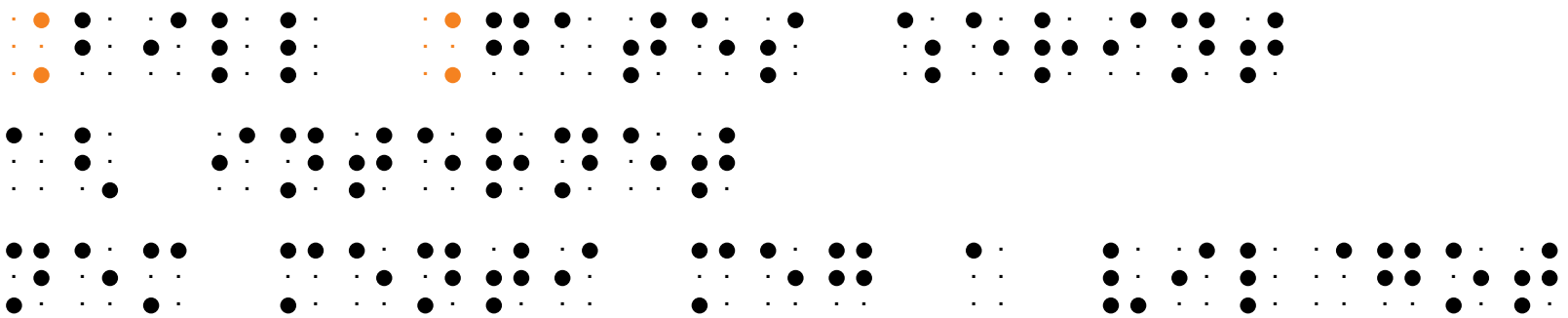
Kis rövidírás: nagybetűjel elhagyása; névelők összevonása; nem, világ

Új rövidírás: szerint, meg

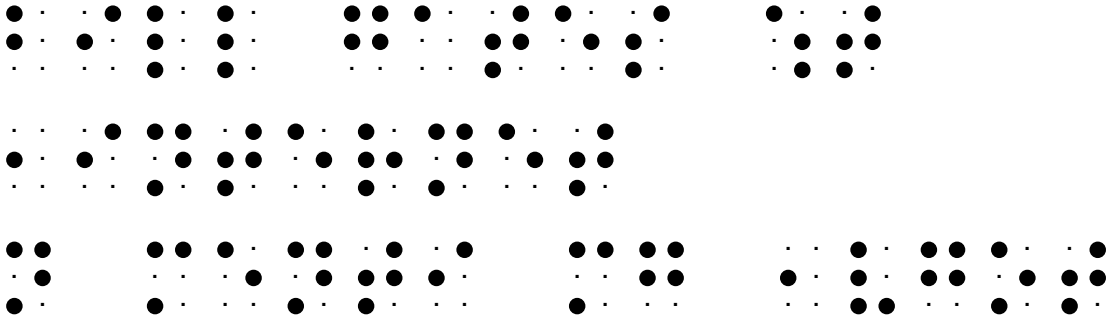
Hossz: 55 karakter

Példa

eredeti: *Bill Gates szerint az internet nem menti meg a világot*



rövidítve: *bill gates szt ,internet n menti mg ,vgot*



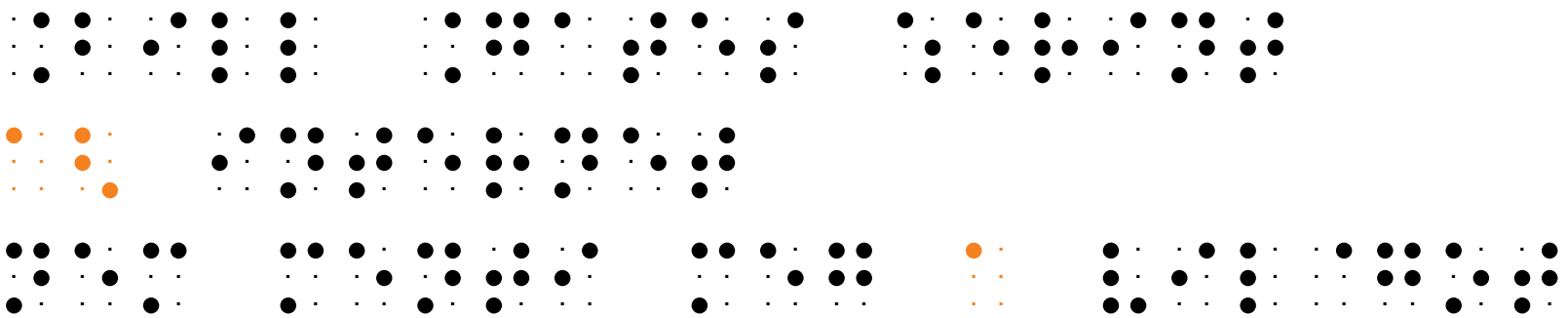
Kis rövidírás: nagybetűjel elhagyása; névelők összevonása; nem, világ

Új rövidírás: szerint, meg

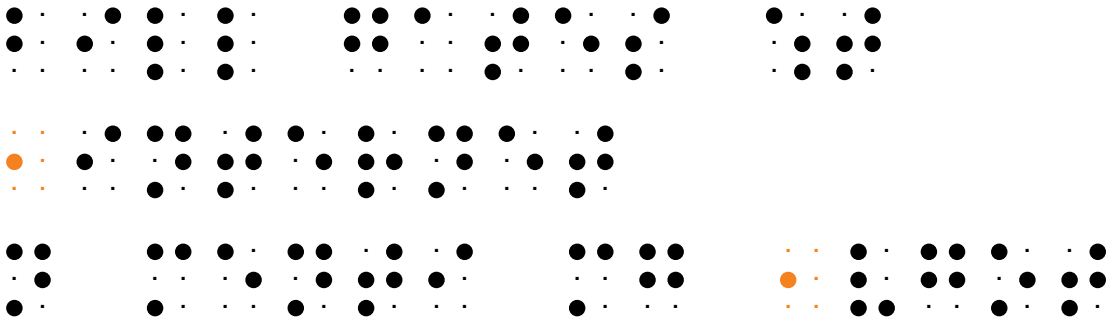
Hossz: 55 $\rightarrow$ 53 karakter: rk() $= 3,6\%$

Példa

eredeti: *Bill Gates szerint az internet nem menti meg a világot*



rövidítve: *bill gates szt ,internet n menti mg ,vgot*



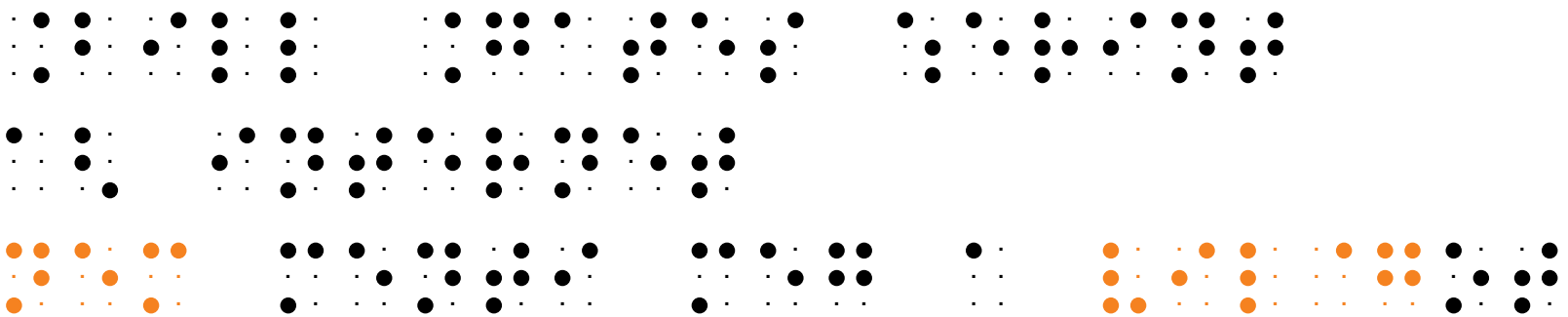
Kis rövidírás: nagybetűjel elhagyása; névelők összevonása; nem, világ

Új rövidírás: szerint, meg

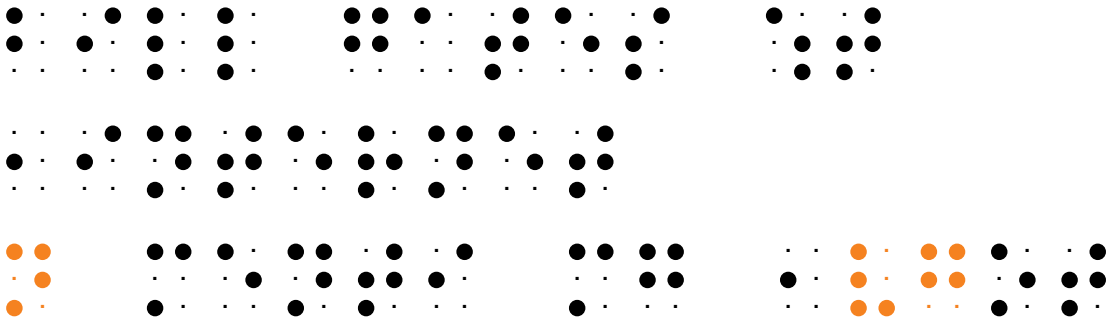
Hossz: 55 $\rightarrow$ 50 karakter: rk() $= 9,1\%$

Példa

eredeti: *Bill Gates szerint az internet nem menti meg a világot*



rövidítve: *bill gates szt ,internet n menti mg ,vgot*



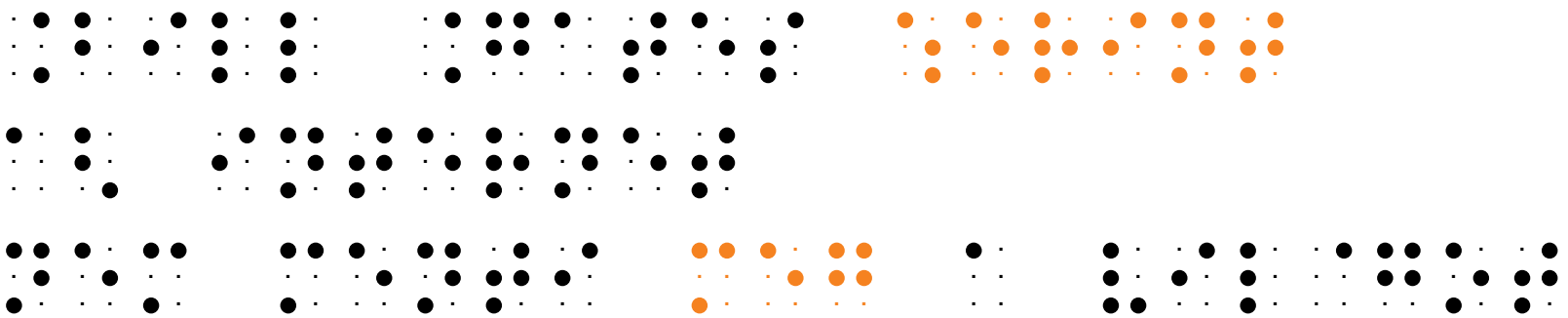
Kis rövidírás: nagybetűjel elhagyása; névelők összevonása; nem, világ

Új rövidírás: szerint, meg

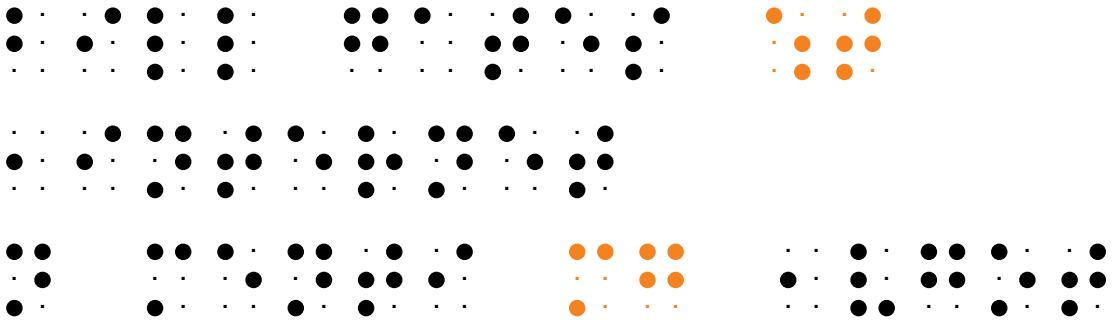
Hossz: 55 $\rightarrow$ 45 karakter: rk() $= 18,2\%$

Példa

eredeti: *Bill Gates szerint az internet nem menti meg a világot*



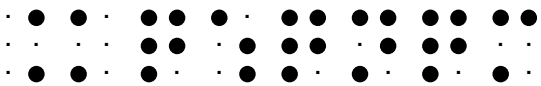
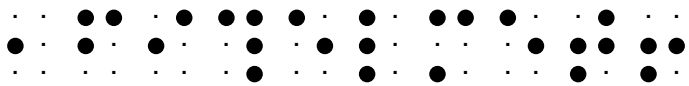
rövidítve: *bill gates szt ,internet n menti mg ,vgot*



Kis rövidírás: nagybetűjel elhagyása; névelők összevonása; nem, világ

Új rövidírás: szerint, meg

Hossz: 55 $\rightarrow$ 40 karakter: rk() $=$ **27,3%**

 = Ö