Az új magyar Braille-rövidírás korpuszvezérelt kialakításának lehetőségei

Sass Bálint

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

A vakok által világszerte használt, tapintáson alapuló Braille-írást Louis Braille fejlesztette ki 1837-ben. A karakterek (ún. Braille-cellák) két oszlopban elrendezett 3-3, azaz összesen hat kidomborodó pontból állnak. Az egyes pontokra a sorszámukkal hivatkozunk: a bal oszlopban helyezkedik el fentről lefelé az 1-es, 2-es, 3-as pont, a jobb oszlopban pedig szintén fentről lefelé a 4-es, az 5-ös és a 6-os. A kidomborodó és ki nem domborodó pontok különböző elrendezéseiből összesen $2^6 = 64$ különböző karakter áll elő: a *t* jele például a 2345 (⠞).

Egyrészt mivel a Braille-írás papíron (speciális dombornyomtatóval kinyomtatva) sok helyet foglal el, másrészt az írás (jegyzetelés) és az olvasás meggyorsítására kialakították a Braille-rövidírásokat – külön-külön az egyes nyelvekre (német: [1]; angol: [2]). A rövidírásban szabályok adják meg, hogy mit hogyan rövidítünk. Magyarban az 50-es években kidolgozott és ma is használatos ún. „kis" rövidírásban például a *hogy* szót *h*-val (125 ⠓), a *kell*-t *k*-val (13 ⠅) rövidítik.

A Magyar Vakok és Gyengénlátók Országos Szövetsége 60 év elteltével döntött úgy, hogy a mai nyelvhasználatot is figyelembe vevő új rövidítésekkel bővíti a szabályrendszert, azzal a céllal, hogy a rövidítési képessége a jelenlegi nagyjából 10%-ról a 20% közelébe növekedjen. A rövidírás-rendszerek kifejlesztése sok esetben nagy időigényű feladat, az egységes angol rövidírás kialakítása 1991-től kezdődően majdnem két évtizedet vett igénybe [3].

Jelen kutatásban azt vizsgáljuk, hogy hogyan lehet korpuszgyakorisági adatok alapján, a lehető legkisebb emberi beavatkozással, azaz szinte teljesen automatikusan és ezáltal záros időn belül előállítani a lehető legnagyobb rövidítési képességgel bíró új magyar rövidírást. Nyilván a lehető leggyakoribb elemeket (karaktersorozatokat) érdemes a lehető legrövidebbre rövidíteni, így nyerjük összességében a legtöbbet. A szűk keresztmetszet a rövidítésre rendelkezésre álló jelek száma: a 64 egykarakteres jelből is csak a ritkábbak alkalmasak arra, hogy rövidítésjelek legyenek.

Minden karaktersorozathoz ötféle gyakorisági értéket rendelünk: hányszor fordul elő (1) szó elején, (2) szó belsejében, (3) szó végén, (4) önálló szóként, illetve az előző négy összegeként: (5) összesen. Ez az elkülönítés két okból is kiemelten fontos. Egyrészt adott jelet rövidítésként csak abban a pozícióban érdemes szerepeltetni, ahol nem (vagy alig) fordul elő (például: pontosvessző szó elején, szó belsejében vagy önálló szóként). Másrészt pedig, azáltal, hogy nem

csak összesített gyakoriságokkal dolgozunk, megmarad annak a lehetősége, hogy egy adott rövidítésjelet különféle pozíciókban eltérő célokra használhassunk, ahogy például az `1346` (⁘) a németben a szó belsejében lévő *mm* rövidítése és az *immer* önálló szó rövidítése is. Utóbbi esetben amiatt szabadul fel egy rövidítésjel önálló szó rövidítésére, mert tekintetbe vették, hogy az *mm* betűkapcsolat a németben önálló szóként nem fordul elő. Ugyanezt az elvet követhetjük a magyarban is: a nagyon gyakori *et* hangkapcsolatra alkalmazott rövidítésjelet önálló szóként a *szerint* rövidítésére használhatjuk, mivel az *et* önálló szóként extrém ritka.

Az algoritmus vázlata a következő. Számba vesszük a rövidíthető nyelvi elemeket, azaz a gyakori karaktersorozatokat. Egy gyakorisági listában tüntetjük fel mind az öt típust a fenti ötféle gyakorisági értékükkel külön-külön, így egy elem 5-ször fog szerepelni. A listát a várható rövidítési képesség szerint rendezzük. A rövidítési képességet a következőképpen számoljuk: $\mathrm{rk}(w, r(w)) = [l(w) - l(r(w))] * fq(w)$, ahol w az eredeti rövidítendő karaktersorozat, $r(w)$ a rövidítés, $l()$ a hossz (karakterszám), $fq()$ a gyakoriság. Először abból indulunk ki, hogy 1 hosszúságú rövidítéseket tudunk képezni. A legritkábban előforduló jelek lesznek alkalmasak rövidítésnek. Ismerve ezek listáját, hozzárendeljük a lista első helyén álló rövidítendő elemhez a legritkább elemet rövidítésként.

A szabályok automatikus megalkotásakor számos szempont figyelembevételével döntünk. Az egyes rövidítésjelek hatékony felhasználására vonatkozó fenti megfontolások szerint járunk el. Kezeljük azt az esetet, mikor a rövidítésjel literálisan jelenik meg (azaz például az `125` (⁖) nem rövidítésként, hanem valóban *h* betűként értendő – ilyenkor egy külön erre szolgáló jellel prefixáljuk az adott jelet, és ez levonódik a rövidítési képességéből). Egy előzetesen kidolgozott lista alapján bizonyos toldalékok (pl.: *-ság/-ség*) hangrend szerint eltérő formáit összevonjuk, egy jellel rövidítjük. Számításba vesszük a korábbi rövidítésekkel való átfedések hatását, ugyanis egy szabály megalkotása érinti az általa rövidített elem részleteire vonatkozó vagy az elemet részként tartalmazó potenciális szabályokat, az általuk elérhető rövidítési képesség változhat. Ezért minden szabály megalkotása után újrarendezzük a listát a frissített rövidítési képességek szerint, majd vesszük a lista elejére kerülő – legnagyobb rövidítési képességgel bíró – elemet, és visszatérünk az algoritmus elejére. Ha nem találunk megfelelő n hosszúságú rövidítésjelet az aktuális rövidítendő elemhez, akkor a továbbiakban $n + 1$ hosszúságú többkarakteres vagy speciális prefixjellel ellátott rövidítésjelet keresünk hozzá. Ennek megfelelően a fenti képlet szerint csökken az elemhez rendelt rövidítési képesség, és ez befolyásolja a rendezett listán elfoglalt helyét is.

A legnagyobb lehetséges rövidítésre vonatkozó fent tárgyalt megfontolások mellett ugyanilyen fontos szempont az új rövidírás jó olvashatósága (tapintás útján jó felismerhetőség) és könnyű megtanulhatósága (kevés, egyszerű szabály). A nagy rövidítési képesség és kényelmes használhatóság egymás ellen ható követelmények, itt egy körültekintően kidolgozott kompromisszumra valamint közvetlen vakok általi tesztelésre van szükség annak érdekében, hogy a potenciális

felhasználók elfogadják és szívesen használják az új rövidírást. A Bánó-féle ún. ,,nagy" rövidírás éppen bonyolultsága miatt nem terjedt el korábban.

Tapasztalat szerint a jól olvasható rövidítés pozíciótól függetlenül mindig azonos jelentésű, a szó kezdő és záró betűjéből, illetve a szót alkotó jellegzetes mássalhangzóból áll. Érdemes külön kezelni az egykarakteres és a többkarakteres rövidítéseket ebből a szempontból. Az egykarakteres rövidítésjelek kiemelten értékesek, mivel nagyon rövidek és nagyon kevés van belőlük. Esetükben nyilván nem követelhető meg a szó kezdő és záró betűjére vonatkozó fenti feltétel, főként, hogy a legalkalmasabb rövidítésjelek éppen az írásjelek. Érdemes megengedni, hogy az egykarakteres rövidítésjelek esetében csak a rövidítési képesség számítson, azaz korlátozás nélkül bárminek a rövidítésére felhasználhassuk őket, sőt még azt is, hogy különböző pozíciókban különféle jelentéssel bírhassanak. Lehetőség szerint törekedni kell a könnyű megtanulhatóságra, ahogy ezt a fent idézett német (*mm*) és magyar (*et*) példánál láttuk. A többkarakteres rövidítésjeleknél a fenti követelmény könnyebben teljesíthető, esetleg automatikus úton is.

Említettük, hogy a lista elején aktuálisan található leggyakoribb rövidítendő elemhez mindig az épp rendelkezésre álló legritkább elemet rendeljük hozzá rövidítésként. A fentiek alapján ez egykarakteres rövidítés esetén valóban szigorúan gyakorisági alapon történik a lehető legnagyobb rövidítés elérése érdekében. Az olvashatósági szempontok akkor kerülnek előtérbe, mikor a rövidítendő elem ritka típusa helyett keresünk másik ideillő, könnyen megjegyezhető rövidíthető elemet; illetve a többkarakteres rövidítéseknél, mikor számos azonos gyakoriságú (ritka) rövidítésjel közül választhatunk.

A fenti módszerrel előállított rendszer rövidítési képessége elérheti a kívánt 18-20%-ot, ami megfelel az új magyar Braille-rövidírással szemben támasztott követelményeknek.

Hivatkozások

1. Freud, E.: Leitfaden der deutschen Blindenkurzschrift: Teil 2. Verlag der Deutschen Blindenstudienanstalt, Marburg (1973)
2. Christine Simpson, ed.: The Rules of Unified English Braille. Version I edn. Round Table on Information Access for People with Print Disabilities Inc., Australia (2010)
3. Bogart, D.: Unifying the English Braille Code. Journal of Visual Impairment & Blindness **103**(10) (2009) 581–583