Látics Barbara – Gombos Péter

Olvashatósági formulák és adaptálhatóságuk magyar nyelvre

Kivonat

Miközben a világ sok részén – különösen angol nyelvterületen – komoly hagyománya van, Magyarországon nem szoktuk meghatározni egy szöveg nehézségét. Nincs pontos mérőeszközünk arra, hogy meg tudjuk határozni, hogy adott életkorú olvasó számára mennyire nehéz egy szövegkorpusz megértése. A tanulmány bemutatja, milyen utat jártak be az Egyesült Államokban, hogyan fejlesztették, alakították a különböző olvashatósági képleteket, indexeket, milyen paramétereket tekintettek fontosnak ezek kidolgozásánál. Bemutatjuk, hogy az angol nyelven eredményesen használt formulák miért nem alkalmasak a magyar nyelvű szövegek minősítésére. A végső cél egy olyan mérőeszköz (képlet) kidolgozása, amely magyar nyelven is pontosan működik. Fontos tényező a szó- és mondathosszúság, a szótagszám, valamint a szófamiliaritás. A pontos magyar mérőeszköz/képlet egyik sajátossága éppen a komplexitása lehet, vagyis hogy megfelelő arányban minél több tényezővel számoljon. Jelenleg olyan modellen dolgozunk, amely gépi tanuló algoritmusok segítségével határozza meg a döntő nyelvi sajátosságokat, így létrehozva a magyar olvashatósági képletet mint mérőeszközt.

Kulcsszavak: szövegértés, olvashatóság, olvashatósági index, Flesch–Kincaid-képlet, gépi tanuló algoritmus

1. Bevezető

A kérdés, hogy mi tesz egy szöveget könnyen olvashatóvá – valójában könnyen érthetővé –, már több mint egy évszázada foglalkoztatja a kutatókat. A 19. században kezdték felismerni az olvashatóság fontosságát: az oktatás szélesebb körökben vált hozzáférhetővé, szükség volt arra, hogy a pedagógusok közérthető nyelven tanítsanak, s ezzel együtt felmerült az igény a különböző korosztályokat megszólító szövegek osztályozására. Az 1930-as évek olvasásszociológiai kutatásai már azt is vizsgálták, hogy mit olvasnak a felnőttek. E korosztály arról panaszkodott, hogy – habár szívesen olvasnának többet – sok könyvet túl nehéznek, vagyis nehezen érthetőnek találnak. Az, hogy a befogadó felől megfogalmazódott egy igény, s hogy ennek volt piaci vonzata is, magával hozta a terület tudományos vizsgálatának fejlődését is. Az első, olvashatósággal kapcsolatos kutatások több szempontot is elemeztek a szövegek kapcsán: tartalmat, stílust, formátumot és szerkezetet. William Gray és Bernice Leary például azt találták, hogy a mondathosszúságot és a nehéz szavakat is magába foglaló *stílus* az egyik legfontosabb tényező (Readable).

Magyar Nyelvőr 149. 2025: 486–502. DOI: 10.38143/Nyr.2025.4.486

Mára nehezen megkérdőjelezhető állítás lett az, hogy a szövegeknek *olvashatónak* kell lenniük ahhoz, hogy érthetők legyenek. Az olvashatóság fogalmát ennek megfelelően nehézségi szintként definiálják, és az érthetőség szinonimájaként utalnak rá (Dale–Chall 1949; Ateşman 1997; Dubay 2007; Zamanian–Heydari 2012; Campos et al. 2014; Rahmawati–Sulistyono 2021; Öksüz–Keskin 2022; Dilenschneider–Horness 2023). Ha a szöveg szintje meghaladja az olvasó képességeit, a tartalom megértése, az információ kinyerése nehézséget fog jelenti. Az olvashatóságot a szöveg mennyiségi paraméterei segíthetik vagy akadályozhatják, ezek leggyakrabban:

– az átlagos mondathosszúság,
– a karakterekben vagy szótagokban mért átlagos szóhosszúság,
– valamint a nehéz szavak aránya (Bailin–Grafstein 2015).

Más források mérésként utalnak az olvashatóságra, „a szöveg nehézségének mérése"-ként (Olney 2022) definiálják, vagy olyan eszközként, amely a szöveg nehézségének mérőszámaként írható le, és különféle olvashatósági képletekkel számítható ki (Srisunakrua–Chumworatayee 2019).

A fogalom taglalásakor rendszerint megemlítik az úgynevezett „olvashatósági képlet"-eket (Srisunakrua–Chumworatayee 2019; Alhusban–Torki 2021; Carcamo 2023; Dilenschneider–Horness 2023; Şahin–Coşkun 2023; Zainurrahman et al. 2024), hiszen ezek – mint objektív merőeszközök – segíthetnek eldönteni, hogy egy-egy szöveg könnyen vagy nehezen érthető-e az adott korosztály számára.

Az első olvashatósági formulát Lively és Pressey alkotta meg 1923-ban. Céljuk a középiskolás tankönyvek osztályozása volt. Kvantitatív módon mérték a szövegek érthetőségét, olyan tényezők figyelembevételével, mint a mondathossz és a gyakori szavak. Habár képletükhöz skála nem kapcsolódott, munkájuk kikövezte az utat az olvashatósági formulák felé (Wang et al. 2022).

Az 1940-es évek nyelvészei már-már versengtek egymással: mindannyian a legegyszerűbb és legmegbízhatóbb mérőeszköz létrehozását tűzték ki célul. Robert Gunning (1944), Rudolf Flesch (1948) és George Spache (1953) munkássága például a későbbi kutatókra is hatással volt. Az évszázad során az olvashatósággal kapcsolatos vizsgálatok új lendületet kaptak, több száz képlet született, melyek később automatizálódtak (forrás: Readable).

A szövegek érthetősége napjainkban is kutatások tárgyát képezi, nyilvánvaló okokból. Az optimális olvashatóság könnyíti a megértést és az információfeldolgozást, kevesebb energiát igényel, s tovább fenntartja az érdeklődést (Szabó 2024). Az olvasástudás legfelsőbb szintje „az, amikor az ember »csak úgy« olvas, vagyis azért, mert számára létszükséglet, örömforrás az olva-

sás. Az ilyen olvasás egyértelműen elmerülés, belefeledkezés […]" (Balázs 2021: 432). Vajon a flow-élményhez nem járulna hozzá a *befogadó* számára ideális nehézségű szöveg?

Az olvasás-szövegértés tehát nemcsak az olvasótól függ: a szöveg nehézsége is éppoly meghatározó. A tanulásra szánt szövegek összhangban kell legyenek az adott korosztályú tanulók olvasási képességével. Az olvashatósági mutatók mentén górcső alá vehetők a tankönyvek, oktatási anyagok, kötelező olvasmányok szövegei. Vajon megfelelnek a tanulók életkori sajátosságainak és olvasási képességének (Józsa–Steklács 2012; Steklács et al. 2015)? Érdekes módon Magyarországon viszonylag régóta mérik a szövegértés (vagyis az olvasó) „szintjét", ugyanakkor nincs hagyománya a szöveg nehézsége, szintje meghatározásának. Holott ennek – a fentiek miatt – nyilvánvalóan volna jelentősége.

2. Az olvashatósági formulákról általánosságban

Az olvashatóság egyik mérőeszköze az olvashatósági formula (más néven képlet vagy index), amely lexikai és szintaktikai mutatókat kombinál. Ez a megközelítés kisszámú komponenssel hozza összefüggésbe a szöveg nehézségét – ezek lesznek a képlet alkotóelemei (Lukács et al. 2022).

Az olvashatósági indexek képesek előre jelezni (!) a szövegek nehézségi szintjét. Ezek valójában matematikai képletek, melyekbe – leggyakrabban – a szavak és a mondatok hosszúságát behelyettesítve kapjuk meg a szöveg olvashatóságára utaló számértéket. A számérték, vagyis az olvashatósági pontszám mindig egy nehézségi szintet jelöl – évfolyamszintet vagy korosztályt. A formula tehát képes iránymutatást adni, hogy a szöveg melyik korosztály/évfolyam számára ideális, vagyis könnyen olvasható (Józsa–Steklács 2012; Steklács et al. 2015). Az egyik legnépszerűbb és legelterjedtebb formula, a Flesch–Kincaid-féle is a következő elven működik[1] (Readable).

Flesch–Kincaid-képlet

$$0{,}39\left(\frac{\text{összes szó}}{\text{összes mondat}}\right) + 11{,}8\left(\frac{\text{összes szótag}}{\text{összes szó}}\right) - 15{,}59$$

1. ábra. A Flesch–Kincaid-képlet

[1] Természetesen a szavak, szótagok és mondatok *száma* kerül a képletbe.

Helyettesítsük be a képletbe a vizsgált szöveg változóit, vagyis a szavak, a mondatok és a szótagok számát! A szorzótényezők és a kivonandó állandók. Eredményül egy számértéket kapunk, mely évfolyamszintet jelent. Ez lesz a szöveg olvashatósági szintje. A táblázatról az is leolvasható, hogy az egyes Flesch–Kinchaid-pontszámok mely korosztályt jelentik. A „jó" olvashatóságot tehát mindig a célközönség határozza meg (Readable).

1. táblázat. Szövegek életkori besorolása a Flesch–Kincaid-pontszám alapján

Flesch–Kincaid-pontszám	*nehézségi szint*	*iskolai szint*	*életkor (USA)*	*példakönyv*
0–3	alap	óvoda, alsó tagozat	5–8	Hooray for fish! (Lucy Cousins)
3–6	alap	alsó tagozat	8–11	A Graffaló (Julia Donaldson)
6–9	átlagos	felső tagozat	11–14	Harry Potter
9–12	átlagos	gimnázium	14–17	Jurassic Park
12–15	haladó	egyetem	17–20	Az idő rövid története (S. Hawking)
15–18	haladó	posztgraduális	20+	szakszövegek

Az olvashatóságiindex-számítás ma már bizonyos platformokon automatikusan is működhet. A legegyszerűbb szövegszerkesztők is képesek rá – például gépelés közben jelzik a szöveg nehézségét (Benjamin 2012). Az interneten fellelhetők az olvashatóságot számító weboldalak is. A dilemmát számunkra az okozza, hogy a formulákat angol nyelvre tervezték, más nyelvek esetében azonban jellemzően hibás értékelést adnak (Józsa–Steklács 2012). Adaptálásuk éppen a változók miatt jelent kihívást. Az átlagos mondat- és szóhosszúság, vagyis a viszonyítási alap minden nyelv esetében más (Şahin–Coşkun 2023). Az agglutináló nyelvek (például a török vagy a magyar) szavai kifejezetten hosszabbak, így a sztenderd képletek biztosan helytelen eredményeket hoznak (Yildiz et at. 2019).

Bogdán (2022) a kutatásában óvodásoknak és felső tagozatosoknak szánt, magyar szövegek olvashatóságát mérte külföldi formulákkal. Az eredmények szerint még a legkisebbeknek szóló szöveg megértéséhez is felsőfokú végzettség volna szükséges…

Hasonlóan tanulságos Dóra László (2019) vizsgálata, aki négy mű (Rikitiki-tévi, Fecskék és Fruskák, Tíz kicsi néger, Egy magyar nábob) olvashatóságát számolta ki klasszikus képletek segítségével. A mesekönyv és a legnehezebbnek számító Jókai-regény között alig volt eltérés, bármelyik formulát

használták. (A kutatás szerint ugyanakkor még a mesekönyv megértéséhez is elengedhetetlen 12,4 évnyi tanulás.) A szakember konklúziói:

– a képletek egyike sem hozott elfogadható eredményeket,
– nyelvünk sajátosságaiból adódóan nem alkalmasak magyar szövegek egyértelmű besorolására (Dóra 2019).

Kétségtelen tehát, hogy a pontos eredményhez a képletek módosítására vagy új formula létrehozására van szükség.

A szakirodalomban fellelhetők olyan kutatások, amelyek a formulák más nyelvre történő adaptálásával kísérleteztek. Ezek többnyire a klasszikus formulák valamelyikét vették alapul, és a bennük szereplő állandók módosításával hoztak létre egy újat. Egyesek párhuzamos szövegeket, míg mások nyelvspecifikus gyakorisági listákat használtak fel vagy alkottak. Továbbá egyéb, modernebb megközelítések is léteznek. Mindezekről a későbbiekben fogunk szót ejteni. Mindenekelőtt azonban az olvashatósági képletek három főbb csoportját mutatjuk be. Ezek és működési elvük ismerete a magyar formula létrehozása szempontjából elengedhetetlen.

3. Az olvashatósági formulák áttekintése
(Tekfi 1987; DuBay 2007; Readable)

A korábban bemutatott Flesch–Kincaid-formula szótagszámláláson alapul, de más típusú képletek is léteznek. A meglévők – habár összetevőikben igen hasonlók – a mért változók alapján három nagy csoportba sorolhatók: (1) a szavankénti szótagok számára, (2) a szavankénti karakterek számára vagy (3) a szófamiliaritásra helyezik a hangsúlyt. Az utóbbi a nehéz szavak számát számlálja egy előre meghatározott szólistát alapul véve. A legnépszerűbb olvashatósági formulák csoportosítása (Bogdán 2024a):

2. táblázat. Angol nyelvű olvashatósági formulák

Szófamiliaritás	**Szóhosszúság**	
	Szótagszám	**Betűszám**
Dale-Chall Formula	Flesch Reading Ease Gunning	
Space Formula	Fog Index SMOG Indix	ARI
	Powers Sumner Kearl Formula	Coleman Liau Index
	Fry Readability Graph	Raygor Readability Graph
	Lensear Write Formula	Lix Readability Formula
	FORCAST Readability Formula	Rix Readability Formula

3.1. A szótagszámalapú formulák (Readable)

A legyakrabban használt formulák a mondathosszúságot és a szöveg szótagokban (!) számolt szóhosszúságát veszik alapul. A két legismertebb is így tesz, a Flesch Reading Ease (Flesch 1948) és a Flesch–Kincaid Grade Level. Flesch tézise szerint a szóismeret ugyan fontos tényező, viszont érettebb olvasóknál kevésbé releváns. Olvashatóság szempontjából a magas szótagszámot gondolta inkább mérvadónak. Az eredeti képlet egy 1 és 100 közötti pontszámot adott eredményül. (Minél magasabb ez a pontszám, annál „olvashatóbb" a szöveg.) A pontszámot évfolyammá kellett „fordítani", amihez egy konverziós táblázatra volt szükség. Az 1970-es években jött létre a továbbfejlesztett verziója, a Flesch–Kincaid Grade Level. Ez már közvetlenül a szükséges évfolyamszintet jósolja meg. A legtöbb szektorban és tudományágban a Flesch–Kincaid-képletet ajánlják.

A Gunning Fog Index (Gunning 1969) is a szöveg megértéséhez szükséges iskolai végzettséget becsüli meg, 0 és 20 közötti számot generál. A képlet szótagalapú: a három vagy több szótagú szavakat számolja. Leginkább kutatók számára ajánlják, tanulmányaik közérthetőségét javítva.

A SMOG Index (McLaughlin 1969) az egészségügy számára a legideálisabb. Gyakori probléma, hogy az orvosok nem írnak elég világosan, így a páciensek nem értik meg őket. E képletek az egészségügyi dokumentumok olvashatóságának javításához nyújthatnak segítséget. A számítás módja viszonylag egyszerű, viszont a mintavételezett szöveg terjedelme minimum 30 mondat kell, hogy legyen. A SMOG komplex szavakat számol, mégpedig azokat, melyek három vagy több szótagból állnak.

3.2. A karakterszám-alapú formulák (Readable)

Az ARI (Smith–Senter 1967) is figyelembe veszi a mondathosszt, ám – a korábban említettekkel ellentétben – karaktereket s nem szótagokat számol. Minél több betűből áll egy szó, annál nehezebbnek veszi azt. Az ARI-t eredetileg katonai használatra tervezték. A haditengerészet kézikönyvei messze meghaladták a katonák olvasási szintjét, ami régóta fennálló probléma volt. Ahogy a neve (Automated Readability Index) is mutatja, automatizált – s már létrejöttekor jóval pontosabbnak tűnt, mint „kortársai". Műszaki írások szintjének meghatározásához talán a legalkalmasabb.

A Coleman–Liau Index (Coleman–Liau 1975) változói szintén a mondatok és a betűk. A kutatópáros szerint a betűkben megadott szóhosszúság jobb előre jelzője az olvashatóságnak, mint a szótaghossz. E képletet előszeretettel használják az oktatásban.

Léteznek gráfalapú olvashatósági képletek is, a Raygor Graph (Raygor 1981) ebbe a kategóriába tartozik. Egy becslési grafikon ábrázolja a szövegek olvashatóságát. Alapfokú anyagok mérésére nem alkalmas, főként a középfokú oktatásban használt szövegekhez ajánlott. Betűszámlálást használ, a hatnál több karakterből álló szavakra fókuszálva.

A Lix (Björnsson 1968) és a Rix (Anderson 1983) ugyanazon olvashatósági képlet két változata – eredményeik korrelálnak egymással. Logikájuk azonos, azonban a Rix évfolyamszinteket ad meg, s valamivel egyszerűbb. Mindkettő hasznos segítséget nyújthat könyvtárosok, tanárok számára, hiszen főként oktatási szövegekhez ajánlják vagy könyvek kategorizálásához. Megbízhatóan működik kisgyerekeknek szóló szövegeknél és felnőtteknek szóló olvasmányoknál egyaránt.

3.3. A szógyakoriság-alapú formulák

A Dale–Chall-féle formula (Dale–Chall 1948) alapja egy gyakorisági szólista. Ez – mára már – 3000 olyan szót tartalmaz, amelyet a negyedik osztályosok 80%-a ismer. A szólistában nem szereplő szavakat határozzák meg „nehéz”-nek. A Dale–Chall-képlet a szövegben szereplő nehéz szavakat számolja. Minél több ilyen szót használ a szöveg, annál magasabb lesz az olvashatósági szint. (A gyakorisági szólistát Thorndike [1921] munkássága ihlette, aki az angol nyelv 10.000 leggyakrabban használt szavát gyűjtötte össze). Dale és Chall formulája még ma is a legnépszerűbbek egyike.

A Spache Formula[2] (Spache 1953) hasonló elven működik, viszont leghatékonyabban negyedik osztályos szintig képes mérni. Gyakorisági listája nagyjából 1000 szóból áll. 1978-ban jött létre a Spache Formula átdolgozott változata. A képlet általános olvashatóságmérésre nem alkalmas, viszont szülők és pedagógusok számára értékes eszköz lehet. A gyerekeknek szóló írásokról gyorsan és pontosan képes megállapítani, hogy megfelelnek-e az adott korosztály számára.

4. Kritikák

A klasszikus formulák még mindig igen népszerűek, bár megjelenésük óta számos kritika érte, éri őket. Napjainkban is több száz képletet használnak, ezek viszont egymástól gyakran eltérő s nem mindig pontos eredményt mutatnak. Ezek kavalkádjában pedig nehéz eligazodni (Lukács et al. 2022).

[2] Az átdolgozott képlet: (0,121 × a mondatok átlagos szószáma) + (0,082 × a nehéz szavak százalékos aránya) + 0,659

Korábban tisztáztuk, hogy az olvashatósági formulák matematikai képletek: a szöveg mérhető egységeit (szavak, szótagok és mondatok) számlálják, és előfordulásuk gyakorisága alapján számszerűsítik a szöveg érthetőségét. Ebből következik, hogy könnyebbnek számít az a szöveg, amely (átlagosan) rövidebb mondatokból és szavakból áll, s kevesebb nehéz szót tartalmaz. „A legtöbb kritika a számokban kifejezett mérhetőség irrelevanciája miatt veti el a képletek alkalmazását" (Dóra 2019: 23).

Davison és Kantor (1982) ugyanakkor arra figyelmeztet, hogy ez a technika az érthetőség durva becslését eredményezheti. A képletek még a legösszefüggéstelenebb szöveget is olvashatónak ítélik, amennyiben mondatai és szavai elég rövidek/ismertek. Az összetettebb szerkezeti kapcsolatokat – például a koherenciát – figyelmen kívül hagyják. Ahogy Bogdán (2024a) is utal rá: a képletek a szöveg komplexitását és nehézségét nem mindig tükrözik, s emberi tényezőkkel sem számolnak. Ebből kifolyólag nem mérésre, hanem előrejelzésre használhatók. Kérdés, hogy mennyire felel meg az érvényességnek az a mérőeszköz, amely nem képes mérni azt, amit mérni szeretnénk.

Veszély is rejlik abban, hogy a képletek mindössze két-három féle változót tartalmaznak. A kiadók és a szerzők aligha tudnak ellenállni a kísértésnek, hogy olvashatóbb (eladhatóbb) szövegeket állítsanak elő. Azonban a „formulákra írás" többet árt, mint használ (Schriver 2000). A jó olvashatósági pontszám nem feltétlenül jár együtt az érthetőséggel. „Nincs biztosíték arra, hogy pusztán a rövidítés révén könnyebben érthető mondatokat kapunk" (Domonkosi 2013: 31). Önmagában nem elég rövidebb mondatokat, ismertebb szavakat előállítani: számos egyéb nyelvi és szövegbeli tényezőt is figyelembe kell(ene) venni (Schriver 2000).

A „formulákra írás" a hazai gyakorlatban is megjelenik, noha nincsenek is formuláink... Domonkosi Ágnes (2013) a magyar tankönyvszöveg érthetőségét vizsgálta a hatékonyság kritériumait keresve. Említést tesz egy, a tankönyvvé nyilvánítást szabályozó kormányrendeletről (URL1), melyben konkretizálják a hosszú mondatok arányát és a szakszavak gyakoriságát is mint kritériumokat (3. táblázat).

A rendeletben empirikus kutatásokra azonban nem hivatkoznak – ezek jogosan hiányolhatók (Kojanitz 2008). „A mondathosszúság ilyen korlátozása mellett a szövegezés komplikáltságának tekintetében nagymértékben eltérő szövegek is elfogadhatónak minősülhetnek" (Domonkosi 2013: 29), ami a tankönyvek esetében kimondottan kockázatos.

3. táblázat. A hosszú mondatok kívánatos aránya évfolyamonként

Vizsgálni kell, hogy a tankönyv nyelvezete megfelel-e a tanulók fejlettségi szintjének. Vizsgálni kell, hogy a megnevezett évfolyamok esetében a tényleges tankönyvi ismeretanyagot bemutató szövegekben (idézetek, lábjegyzetek, képletek nélkül) a 150 betűhelynél (karakterek száma szóközökkel) hosszabb mondatok %-os aránya ne haladja meg:	
1–2. évfolyamokon	5%-ot
3–4. évfolyamokon	10%-ot
5–6.. évfolyamokon	20%-ot
7–8. évfolyamokon	25%-ot
9–10. évfolyamokon	25%-ot
9–12. évfolyamokon	35%-ot

Előírás az idegen nyelvi tankönyvek és a szakmai tankönyvek kivételével, hogy a megnevezett évfolyamok esetében az egymástól eltérő speciális szakszavak (tudományos szakszavak, szakkifejezések, amelyek a mindennapi nyelvben nem használatosak) átlagos előfordulási sűrűsége mondatonként ne haladja meg:	
1–4 évfolyamokon	a 0,3 értéket
5–8. évfolyamokon	a 0,5 értéket
9–12. évfolyamon	a 0,8 értéket

Az olvashatósági formulák a szavak és a mondatok szintjére korlátozódnak. Sem a szöveg érdekességéről, sem az olvasót jellemző faktorokról[3] nem adnak közvetlen tájékoztatást. Kutatás igazolta, hogy további változók bevezetése statisztikailag nem indokolt: az eredményeken szignifikánsan nem változtatnának (Klare 1984). Azonban a 21. század nyelvészei más véleményen vannak.

Kojanitz (2004) mérvadónak gondolja a mondatok közötti kapcsolatokat és a szöveg szintjét. A szövegérthetőséget a szöveg szerkezete, szerkesztettsége is befolyásolhatja – hangsúlyozza Előry Vilma (2008). Szerinte a szöveg tagolódása, a szövegegységek kapcsolódása, a mondatok kapcsolódása és zsúfoltsága, valamint a lexikai anyag egyaránt olvashatósági tényezők lehetnek. Hasonló állásponton van Fulcher (1997) is. Ő bizonyos tipográfiai jegyeket (betűméret, betűtípus[4], színek, illusztrációk, szövegszerkezet) is fontosnak gondol. Bailin és Grafstein (2001) pedig úgy vélik, olyan tényezőket is figyelembe kellene venni, mint a stílus, a koherencia, az olvasó szó-

[3] Érdeklődés, kulturális és szociális változók (iskolai végzettség, előzetes tudás, motiváció), pszichológiai faktorok (pl. az olvasó aktuális állapota) (Dóra 2019).

[4] Egy 2023-as kutatásból kiderült, hogy a betű típusa nem releváns sem a megértés, sem az olvasottak felidézése szempontjából (Gombos–Látics 2023).

kincse és háttérismeretei. Mindezek kölcsönhatásban vannak egymással, és együtt teszik a szöveget könnyebben vagy nehezebben érthetővé. Akárhogy is: két változó önmagában nem elegendő az olvashatóság méréséhez.

5. Adaptálhatóság és megoldási javaslatok

„Az angol nyelvű szövegekre készült olvashatósági formulák módosítás nélkül nem adaptálhatók a magyar nyelvre" (Szabó 2024). Ehelyett más alternatívá(ka)t kell keresni. Egy, a témával kapcsolatos szakdolgozatban a hallgató több járható utat is megfogalmaz: (1) a szógyakoriság-alapú képletekből kellene kiindulni, a gyakoriságszótárakat felhasználva, (2) szóhosszúság-alapú képletek adaptálása (melyhez az átlagos szóhosszúság meghatározása elengedhetetlen), (3) egy teljesen új mérőeszköz kifejlesztése (Szabó 2024).[5] A továbbiakban Szabó (2024) javaslatait taglaljuk.

(1) A szófamiliaritás egy fontos tényező az érthetőség szempontjából, erre a szakirodalomban is találhatók utalások (Dale–Chall 1948; Cs. Czachesz–Csirik 2002; Leroy–Kauchak 2014; Bogdán 2024). Ez esetben szükség lenne egy magyar nyelvű szófamiliaritási listára. A magyar nyelv leggyakoribb szavait az MTA szövegtára tartalmazza. Ez egy olyan reprezentatív, magyar nyelvi korpusz, amely a hazaiak mellett a határon túli magyar nyelvváltozatokat is felöleli. Gyakorisági listája ötféle alkorpuszt foglal magába: sajtó, irodalmi, tudományos, hivatalos, személyes (Oravetz et al. 2014). A 10–16 éves tanulók leggyakoribb szavait pedig Cs. Czachesz–Csirik (2002) gyakorisági szótára tartalmazza.

(2) Az olvashatóság paralelszöveg-alapú megközelítései igencsak elterjedtek az utóbbi években (Buck et al. 2014). Mivel rendelkezésre állnak magyar–angol párhuzamos szövegek,[6] a korpuszalapú megközelítés sem elvetendő. A módszertant illetően hasznos iránymutatást adhatnak a témával kapcsolatos kutatások (Antunes–Lopes 2019; Bendová 2021; Sibeko 2023; Bogdán 2024b; Sibeko 2024).

Az (1) és (2) esetben is klasszikus formulán alapulna az új képlet, melynek azonban lennének hiányosságai, ahogy az a 4. fejezetből már kiderült.

---

[5] Szabó kiemeli, hogy mindhárom esetben szótövesíteni szükséges a szavakat, a magyar nyelv agglutináló jellegéből adódóan (Szabó 2024). Lukács et al. (2022) javaslata egy morfológiai elemzőprogram használata. „A magyar sajátosságok miatt elengedhetetlen […]. A magyarban a szavak sokféle toldalékolt alakban fordulnak elő. Emiatt ahhoz, hogy szótövekre kereshessünk, az elemzőnek előbb fel kell bontania a morfológiailag komplex alakokat. Ilyen, specifikusan a magyar nyelvre kifejlesztett program az e-magyar" (Lukács et al. 2022: 82). Az e-magyar (Váradi et al. 2018) a magyar nyelv digitális elemzését teszi lehetővé.

[6] ld. https://opus.nlpl.eu/

Amennyiben a cél az olvashatóság pontos mérése – s nem csupán hozzávetőleges előrejelzése –, több változó bevonása szükséges, valamint érdemes modernebb módszerhez folyamodni.

A (3)-as javaslat tűnik a leghatékonyabb megoldásnak, mégpedig az úgynevezett természetesnyelv-modellek és gépi tanuló algoritmusok bevonásával.

A természetesnyelv-modellek (például TAASSC) a klasszikus formulákat háttérbe szorították, főként mert sokkal komplexebb képet adnak (Crossley et al. 2019). Esetükben a szókincs és a struktúra nyelvi jegyei képezik a bemenetet (strukturált adatgyűjtés). A szavak hosszúságát, gyakoriságát, eloszlását, a mondathosszt, az anaforák és referenseik közötti távolságot, valamint a szavak közötti szemantikai viszonyokat is figyelembe veszik – tehát jóval több bemeneti jeggyel dolgoznak. Miben különböznek még a formuláktól? Nem alapműveletekkel számolnak, hanem egy – valószínűségi függvényt használó – gépi tanuló algoritmus jósolja meg az olvashatósági indexet, mégpedig a természetes nyelvi jegyekből tanulva (Lukács et al. 2022).

Chen és Meurers (2016) szerint a modern olvashatóságértékelő rendszerek széles körben használtak, és igen hatékonynak bizonyulnak. Példaként említik a Lexile-t, az ATOS-t és a CohMetrixet.

Egy másik lehetséges mérőeszköz a neurális hálókra épülő, mély tanulási (deep learning) modellek. Ezek közvetlenül a referenciaszöveget dolgozzák fel. Az interneten és a digitális szövegtárakban óriási mennyiségű magyar nyelvű szöveg áll rendelkezésre. A neurális hálók még a természetesnyelv-modelleknél is pontosabban képesek megjósolni a szöveg érthetőségét. A neurális hálók „önjárók", a nyelvészek és pedagógusok csupán a jól olvasható szöveg kritériumait határozzák meg. Nagy előnyük, hogy használatuk a laikusok számára sem okoz nehézséget (Lukács et al. 2022). „Az automatizált, neurális hálókra épülő szövegfeldolgozó rendszerek a jövőben a magyar nyelvben is a hagyományos empirikus módszerek komoly konkurenseivé válhatnak" (Lukács et al. 2022: 83).

6. Összegzés

Az olvasás, a nyelvi tudatosság és a szövegértés szoros kapcsolatban állnak egymással (Balázs 2006), s bár ezt régóta tudjuk, a közelmúltig nem vizsgáltuk e kérdéskör bizonyos területét. Miközben Magyarországon – nehezen megmagyarázható módon – nincs hagyománya, szokása egy-egy szöveg „nehézsége" (olvashatósága, érthetősége) pontos megállapításának, aligha kell indokolni ennek lehetséges hasznosságát, használhatóságát. Elég csupán azt elképzelni, hogy egy általános iskolás tankönyv szövegeit megnézve a szerkesztőknek

megmutatná egy mérőeszköz, hogy van-e olyan részlet a kiadványban, amelynek a megértése túl nehéz az adott korosztály számára.[7]

Hogy lehetne megalkotni egy ilyen mérőeszközt/formulát/képletet? Tanulmányunkban megmutattuk a lehetséges utakat. Bár külföldön – elsősorban angol nyelvterületen – sok évtizede használnak, fejlesztenek ilyen formulákat, sőt folyamatosan javítják, aktualizálják is azokat, ezek nem működnek magyar nyelven, elsősorban nyelvünk agglutináló sajátosságai és az eltérő nyelvhasználati szokások miatt.

Megmutattuk azt is, hogy milyen úton indulhatnak el azok, akik ilyen formula meghatározására vállalkoznak. A szó- és mondathosszúság, a szótagszámok mellett fontos lehet például a szófamiliaritás figyelembevétele – de a pontos magyar mérőeszköz/képlet egyik fontos sajátossága éppen a komplexitása lehet, vagyis hogy minél több tényezővel számoljon. S ráadásul úgy, hogy ezek súlyozása is megtörténjen, nyelvünk sajátosságainak megfelelően.

A fentiekre figyelemmel mi jelenleg egy olyan modellen dolgozunk, amely gépi tanuló algoritmusok segítségével határozza meg a döntő nyelvi sajátosságokat. Miközben úgy látjuk, hogy ez a módszer lehet a leghatékonyabb, munkánk valódi értékét természetesen majd a tesztelések mutatják meg.

Szakirodalom

Alhusban, Hamdallah A., Torki, Saad 2021. An Original Computerized and Web-based Method for Assessing Textbook Readability via Lexical Coverage. *TESOL International Journal* 7–30.

Anderson, Jonathan 1983. Lix and Rix. Variations on a Little-Known Readability Index. *Journal of Reading* 490–6.

Antunes, Hélder – Lopes, Carla Teixeira 2019. Analyzing the adequacy of readability indicators to a non-English language. In: Crestani, Fabio et al. (ed.): *Experimental IR Meets Multilinguality, Multimodality, and Interaction*. 10th International Conference of the CLEF Association. Lugano. 149–55. https://doi.org/10.1007/978-3-030-28577-7_10

Ateşman, Elder 1997. Türkçede Okunabilirliğin Ölçülmesi. *Dil Dergisi* 71–4.

Bailin, Alan – Grafstein, Ann 2001. The linguistic assumptions underlying readability formulae. A critique. *Language & Communication* 285–301. https://doi.org/10.1016/s0271-5309(01)00005-2

Bailin, Alan – Grafstein, Ann 2015. *Readability. Text and context.* Palgrave Macmillan. London.

Balázs Géza 2006. Szenvedéllyel olvasni... *Édes Anyanyelvünk* 97.

[7] Érdekes: miközben több mint 30 éve van hazánkban országos kompetenciamérés, a szövegek „szintezése", nehézségének meghatározása e tesztek készítőinél sem merült fel. Nem ritka például, hogy pontosan ugyanazt a szöveget adták 6. és 8. osztályban, különbséget a kapcsolódó kérdések, feladatok jelentettek.

Balázs Géza 2021. Így olvasunk. Esszé a sokféle olvasásról. *Vasi Szemle* 431–6.
Bendová, Klára 2021. Using a parallel corpus to adapt the Flesch Reading Ease formula to Czech. *Journal of Linguistics/Jazykovedný casopis* 477–87. https://doi.org/10.2478/jazcas-2021-0044
Benjamin George, Rebekah 2012. Reconstructing Readability. Recent Developments and Recommendations in the Analysis of Text Difficulty. *Educional Psychology Review* 63–88. https://doi.org/10.1007/s10648-011-9181-8
Björnsson, Carl-Hugo 1968. *Lesbarkeit durch LIX.* Pedagogiskt Centrum. Stockholm.
Bogdán Patrik 2022. A gyermekirodalmi szövegek olvashatósági vizsgálata a szöveg grammatikai komplexitásán keresztül. In: Bóna Judit – Murányi Sarolta (szerk.): *A nyelvfejlődés folyamata. Kora gyermekkortól kamaszkorig*. Cser Kiadó. Budapest. 149–65.
Bogdán Patrik 2024a. Angol nyelvű olvashatósági formulák magyar nyelvi adaptálásának lehetséges irányai. *Iskolakultúra* 69–77. https://doi.org/10.14232/iskkult.2024.1.69
Bogdán Patrik 2024b. A szövegértés sikerességének előrejelzésére irányuló olvashatósági formula adaptációjának és validálásának folyamata. In: *Találkozások az anyanyelvi nevelésben 5. Régi témák, új nézőpontok a nyelvészetben és a nyelvi nevelésben.* Pécsi Tudományegyetem Bölcsészettudományi Kar (PTE BTK) Nyelvtudományi Tanszék. Pécs. 4–4.
Buck, Christian – Heafield, Kenneth – van Ooyen, Bas 2014. N-gram counts and language models from the common crawl. In: Calzolari, Nicoletta et al. (ed.), *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14).* European Language Resources Association. Reykjavik. 3579–84.
Campos Saavedra, Daniella – Contreras Carmona, Paula – Riffo Ocares, Bernardo – Véliz, Mónica – Reyes Reyes, Alejandro 2014. Complejidad textual, lecturabilidad y rendimiento lector en una prueba de comprensión en escolares y adolescentes. *Universitas Psychologica*. 15–36. https://doi.org/10.11144/javeriana.upsy13-3.ctlr
Cárcamo, Benjamin 2023. The Issue of Readability in the Chilean EFL Textbook. *HOW* 135–155. https://doi.org/10.19183/how.30.2.739
Chen, Xiaobin – Meurers, Detmar 2016. Characterizing Text Difficulty with Word Frequencies. In: *Conference: Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*. San Diego, CA. 84–94. https://doi.org/10.18653/v1/w16-0509
Coleman, Meri – Liau, T. L. 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology.* 283–4. https://doi.org/10.1037/h0076540
Crossley, Scott A. – Skalicky, Stephen – Dascalu, Mihai 2019. Moving beyond classic readability formulas. New methods and new models. *Journal of Research in Reading* 541–61. https://doi.org/10.1111/1467-9817.12283
Cs. Czachesz Erzsébet – Csirik János 2002. *10–16 éves tanulók írásbeli szókincsének gyakorisági szótára.* BIP. Budapest.

Dale, Edgar – Chall, Jeanne S. 1948. A Formula for Predicting Readability. *Educational Research Bulletin* 37–54.

Dale, Edgar – Chall, Jeanne S. 1949. The concept of readability. *Elementary English* 19–26.

Davison, Alice – Kantor, Robert 1982. On the failure of readability formulas to define readable texts. A case study from adaptations. *Reading Research Quarterly* 187–209. https://doi.org/10.2307/747483

Dilenschneider, Robert – Horness, Paul 2023. Profiling Word Frequency and Readability of Online Learner Dictionary Definitions. *rEFLections* 247–267. https://doi.org/10.61508/refl.v30i2.267211

Domonkosi Ágnes 2013. A tankönyvszöveg érthetőségének vizsgálati szempontjai. In: *Az Eszterházy Károly Főiskola tudományos közleményei (Új sorozat 40. köt.)*. Eszterházy Károly Katolikus Egyetem. Eger. 27–35.

Dóra László 2019. Olvashatósági tesztek: elmélet és gyakorlat. *Képzés és Gyakorlat* 21–34. https://doi.org/10.17165/tp.2019.2.2

DuBay, William H. (ed.) 2007. *The Classic Readability Studies*. Impact Information. Costa Mesa.

Előry Vilma 2008. Milyen a jó tankönyvszöveg? In: Medve Anna – Szépe Görgy (szerk.): *Anyanyelvi nevelési tanulmányok III.* Iskolakultúra. Pécs. 7–16.

Flesch, Rudolf 1948. A new readability yardstick. *Journal of Applied Psychology* 221–33. https://doi.org/10.1037/h0057532

Fulcher, Glenn 1997. Text difficulty and accessibility: reading formulae and expert judgement. *System* 497–513. https://doi.org/10.1016/s0346-251x(97)00048-1

Gombos Péter – Látics Barbara 2023. A betűtípus és a szövegértés összefüggésének vizsgálata felső tagozatos tanulók körében. *Anyanyelv-pedagógia* 1–17.

Gunning, Robert 1969. The Fog Index After Twenty Years. *Journal of Business Communication* 3–13. https://doi.org/10.1177/002194366900600202

Józsa Krisztián – Steklács János 2012. Az olvasás tanításának tartalmi és tantervi szempontjai. In: Csapó Benő – Csépe Valéria (szerk.): *Tartalmi keretek az olvasás diagnosztikus értékeléséhez.* Nemzeti Tankönyvkiadó. Budapest. 137–88.

Klare, George R. 1984. Readability. In: Pearson, P. David (ed.): *Handbook of Reading Research. Longman*. Lawrence Erlbaum Accociates. Mahwah, NJ.

Kojanitz László 2004. A pedagógiai szövegek analitikus vizsgálata – a szavak szintje. *Magyar Pedagógia* 429–42.

Kojanitz László 2008. Tankönyvértékelés, tankönyvanalízis, tankönyvkutatás. In: Simon Mária (szerk.): *Tankönyvdialógusok*. Oktatáskutató és Fejlesztő Intézet. Budapest. 23–32.

Leroy, Gondy – Kauchak, David 2014. The effect of word familiarity on actual and perceived text difficulty. *Journal of the American Medical Informatics Association* 169–72. https://doi.org/10.1136/amiajnl-2013-002172

Lukács Ágnes – Rácz Péter – Kas Bence 2022. Tankönyvi szövegek nyelvi feldolgozhatóságának mutatói és vizsgálati módszerei. *Magyar Pedagógia* 65–88. https://doi.org/10.14232/mped.2022.2.65

McLaughlin, G. Harry 1969. SMOG Grading – a New Readability Formula. *Journal of Reading* 639–46.

Olney, Andrew M. 2022. Assessing Readability by Filling Cloze Items with Transformers. In: Rodrigo, Maria Mercedes et al. (ed.): *Artificial Intelligence in Education.* 307–18. https://doi.org/10.1007/978-3-031-11644-5_25

Oravetz Csaba – Váradi Tamás – Sass Bence 2014. The Hungarian Gigaword Corpus. In: Calzolari, Ncoletta et al. (ed.): *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14).* Reykjavik. 1719–23.

Öksüz, Halil Ibrahim – Keskin, Hasan Kagan 2022. A Study on the Impact of Readability on Comprehensibility. *International Journal of Progressive Education* 322–55.

Rahmawati, Laili Etika – Sulistyono, Yunus 2021. Assessment and Evaluation on Text Readability in Reading Test Instrument Development for BIPA-1 to BIPA-3. *Asian Journal of University Education (AJUE)* 51–57. https://doi.org/10.24191/ajue.v17i3.14522

Raygor, Alton L. 1981. Validity Studies of the Minnesota Reading Assessment. A New Reading Test for Post-Secondary Programs. *Proceedings of the Annual Conference of the Western College Reading Association* 63–71. https://doi.org/10.1080/24699365.1981.11669764

readable.com

Şahin, Tuğrul Gökmen – Coşkun, Mehmet 2023. Levels of Readability of the Texts in the 7th Grade Turkish Textbook. *International Journal of Psychology and Educational Studies* 1024–38. https://doi.org/ 10.52380/ijpes.2023.10.4.1297

Schriver, Karen 2000. Readability formulas in the new millennium. What's the use? *ACM Journal of Computer Documentation* 105–6. https://doi.org/10.1145/344599.344638

Sibeko, Johannes – van Zaanen, Menno 2024. Adapting Nine Traditional Text Readability Measures into Sesotho. In: Rooweither, Mabuya et al. (ed.): *Proceedings of the Fifth Workshop on Resources for African Indigenous Languages*. Torino. 66–76.

Sibeko, Johannes 2023. Developing Resources for Measuring Text Readability in Sesotho. In: *11th CLARIN Annual Conference.* Belgium. 120–33. https://doi.org/10.3384/ecp198012

Smith, E. A. – Senter, R. J. 1967. *Automated Readability Index.* Wright-Patterson Air Force Base, Ohio.

Spache, George 1953. A New Readability Formula for Primary-Grade Reading Materials. *The Elementary School Journal* 410–3. https://doi.org/10.1086/458513

Srisunakrua, Thanaporn – Chumworatayee, Tipamas 2019. Readability of reading passages in English textbooks and the Thai national education English test. A comparative study. *Arab World English Journal (AWEJ)* 257–69. https://doi.org/10.24093/awej/vol10no2.20

Steklács János – Molnár Gyöngyvér – Csapó Benő 2015. Az olvasás-szövegértés online diagnosztikus mérések tartalmi kereteinek elméleti háttere. In: Steklács János et al. (szerk.): *Az olvasás-szövegértés online diagnosztikus értékelésének tartalmi keretei.* Oktatáskutató és Fejlesztő Intézet (OFI). Budapest. 15–31.

Szabó Csilla 2024. *Olvashatósági index alkalmazása angol nyelvű szövegeknél. A magyar nyelvi adaptáció kérdésköre.* Magyar Agrár- és Élettudományi Egyetem. Kaposvár.

Tekfi, Chaffai 1987. Readability formulas. An overview. *Journal of Documentation* 261–73. https://doi.org/ 10.1108/eb026811

Thorndike, Edward L. 1921. *The Teacher's Word Book.* Teachers College, Columbia University. New York City.

Wang, Shan – Liu, Xiaojun – Zhou, Jie 2022. Readability is decreasing in language and linguistics. *Scientometrics* 4697–729. https://doi.org/10.1007/s11192-022-04427-1

Yildiz, Munise – Kozanhan, Betul – Tutar, Mahmut Sami 2019. Assessment of readability level of informed consent forms used in intensive care units. *Medicine Science* 277–81. https://doi.org/ 10.5455/medscience.2018.07.8933

Yusuf, Fazri Nur – Sukyadi, Didi – Zainurrahman, Z. 2024. Text readability. Its impact on reading comprehension and reading time. *Journal of Education and Learning (EduLearn* 1422–32. https://doi.org/ 10.11591/edulearn.v18i4.21724

Zamanian, Mostafa – Heydari, Pooneh 2012. Readability of texts. State of the art. *Theory & Practice in Language Studies* 43–53. https://doi.org/10.4304/tpls.2.1.43-53

URL1. njt.hu (2004): *23/2004. (VIII. 27.) OM-rendelet a tankönyvvé nyilvánítás, a tankönyvtámogatás, valamint az iskolai tankönyvellátás rendjéről.* https://njt.hu/jogszabaly/2004-23-20-45 (Letöltés: 2025. 04. 25.)

Látics Barbara
PhD-hallgató
PTE BTK Oktatás és Társadalom
Neveléstudományi Doktori Iskola
E-mail: lbarbi0604@gmail.com
https://orcid.org/0009-0002-8906-2437

Gombos Péter
egyetemi docens
MATE Kaposvári Campus
Neveléstudományi Intézet
E-mail: Gombos.Peter@uni-mate.hu
https://orcid.org/0000-0002-8172-0557

Látics, Barbara – Gombos, Péter

Readability Formulas and their Adaptability to Hungarian

Abstract

While there is a long tradition of assessing the difficulty of texts in many parts of the world, particularly in English-speaking countries, this is not customary in Hungary. We do not have any precise tools for determining how difficult a text corpus is for readers of a given age. This study presents the approach taken in the United States, how various readability formulas and indices were developed and refined, and what parameters were considered important in their development. We show why formulas that have been used successfully in English are not suitable for classifying Hungarian-language texts. The ultimate goal is to develop a measurement tool (formula) that works accurately in Hungarian. Important factors include word and sentence length, number of syllables, and word familiarity. One of the characteristics of an accurate Hungarian measurement tool/formula may be its complexity, i.e. that it takes into account as many factors as possible in the right proportions. We are currently working on a model that uses machine learning algorithms to determine the decisive linguistic characteristics, thus creating a Hungarian readability formula as a measurement tool.

Keywords: text comprehension, readability, readability index, Flesch–Kincaid formula, machine learning algorithm