LÁTICS Barbara
Pécsi Tudományegyetem
Oktatás és Társadalom
Neveléstudományi Doktori Iskola
ORCID: 0009-0002-8906-2437
lbarbi0604@gmail.com

Olvashatóságot meghatározó szövegtényezők és automatizált szövegelemző szoftverek[1]

Oktatási környezetben a rosszul olvasható tartalom akadályozhatja a megértést és a tanulást, ami hatással lehet a diákok tanulmányi eredményére és szövegértésére. A klasszikus olvashatósági formulák gyakran olyan felszíni szövegjellemzőkön alapulnak, mint például az átlagos mondat- és szóhosszúság. Nem veszik figyelembe az olvashatóság egyéb árnyalatait, ezért a rájuk való támaszkodás korlátozó és felszínes lehet (Marulli et al. 2024, Látics–Gombos 2025). A tanulmány azonosítja azokat a főbb tényezőket, amelyek hatással lehetnek egy szöveg nehézségére, olvashatóságára. A magyar nyelvű szakirodalom a tankönyvanalízis felől közelítve, a nyelvi szintek szegmentálásával – a szavak, a mondatok és a szöveg szintjén – tárja fel ezeket (Fóris 2002, Kojanitz 2004a, Domonkosi 2013). Az olvashatóság mérhetővé tétele háromféle módszerrel lehetséges (Lukács et al. 2022): az olvashatósági formulákkal szemben effektívebb megoldást jelentenek a természetesnyelv-feldolgozáson alapuló, gépi tanulási módszerek, de a leghatékonyabbnak mégis a neurális hálókra épülő, mélytanulási modellek bizonyulnak. A tanulmányban a kézzelfoghatóbb szövegtényezőkön túl néhány modernebb, automatikus szövegelemző szoftvert is bemutatok. Működési elvük ismerete hasznos lehet a magyar nyelvű szövegek géppel történő elemzésekor. Ezek képet adnak arról is, hogy az olvashatóságot meghatározó bemeneti jegyek milyen főbb kategóriákba sorolhatók. Léteznek magyar nyelvfeldolgozó eszközláncok: UDPipe, huspaCy, Magyarlánc, s kiváltképp az e-magyar, melynek 2019-es (emtsv néven ismert), új verziója már a gyakorlatban is bizonyította a hatékonyságát. Kutatásunk szempontjából – az olvashatóság mérhetővé tételéhez – pedig kifejezetten hasznos eszköz lehet.

Kulcsszavak: olvashatóság, szövegtényezők, szöveganalízis, gépi tanuló algoritmus, mélytanulás, számítógépes nyelvészet

Bevezetés

Oktatási környezetben a nehezen feldolgozható szöveg akadályozhatja a megértést és a tanulást, ami hatással lehet a diákok tanulmányi eredményére és szövegértésére. Az ismertebb olvashatósági formulák gyakran olyan felszíni szövegjellemzőkön alapulnak, mint például az átlagos mondat- és szóhosszúság. Nem veszik figyelembe ugyanakkor az olvashatóság egyéb összetevőit, valamint azok viszonyrendszerét, ezért a rájuk való támaszkodás korlátozó és felszínes lehet (Marulli et al. 2024; Látics–Gombos 2025). „A formulák önmagukban pusztán matematikai algoritmusok, melyeknek fogalmuk sincs arról, hogy a szövegjellemzők mely értékei azok, amelyek szintbeli különbséget jelentenek a feldolgozhatóságban.” (Lukács et al. 2022: 76)

[1] A tanulmány elkészítését a Magyar Tudományos Akadémia Közoktatás-fejlesztési Kutatási Programja támogatta. Olvasási Fluencia és Szövegértés Kutatócsoport, MTA–PTE, SZKF2022-12/2022

A probléma megoldása multidiszciplináris megközelítést igényel, amely a nyelvészet, a kognitív tudomány és az informatika ismereteit ötvözi (Marulli et al. 2024). Jelen tanulmány elsődleges célja, hogy körvonalazza azokat a tényezőket, amelyek befolyásolhatják egy szöveg olvashatóságát. Emellett ismerteti a természetesnyelv-feldolgozáson (NLP) és a mélytanuláson (LLM) alapuló, automatizált szövegelemző lehetőségeket is, valamint néhány kapcsolódó kutatást. Mindez egy nagyobb lélegzetvételű mű – doktori értekezés – elméleti hátterének alapozásaként történik, folytatva a korábbiakban feltártakat (Látics–Gombos 2025).

Az írás nagymértékben támaszkodik a szakirodalmi előzményekre, miközben Domonkosi Ágnes megállapítását is szem előtt tartja: „A számos felhasználható eredmény ellenére azonban nincsenek megbízható, empirikus adatok arról, hogy valójában milyen kritériumok teszik könnyebben vagy nehezebben érthetővé az egyes (tankönyv)szövegeket." (Domonkosi 2013: 28) Így a lista egyáltalán nem kötött. Már csak azért sem, mert a hazai források többsége nemzetközi szakirodalomra hivatkozik. Azt, hogy valójában mely szövegtényezők mérvadók a *magyar* nyelvű szövegek esetében, kutatásunk empirikus része fogja feltárni. Jelen írás célja a hazai és nemzetközi szakirodalmi előzmények, álláspontok bemutatása.

Az olvashatóság kedvelt kutatási téma az oktatás, az idegennyelv-tanulás és a nyelvészet területén egyaránt (Smeuninx 2020). A pedagógiai szövegelemzésnek már 20 évvel ezelőtt is kiterjedt szakirodalma volt, mind hazai, mind nemzetközi viszonylatban. Az írás első fele főként ezekre támaszkodik. Ismeretük szükséges a modernebb, automatizált szövegelemző szoftverekről szóló rész értelmezéséhez.

A témát a szöveganalízis-vizsgálatok felől közelítem meg. A szöveganalízisek során a korpuszokban – legtöbbször elektronikusan – nyelvészeti törvényszerűségeket keresnek. Középpontjukban gyakran állnak tankönyvek, hiszen a túlzottan nehéz szövegezés akadályozza az eredményes tanulást. Az analitikus vizsgálatok által „még a tankönyvek iskolákba kerülése előtt meg lehetne jósolni a tanulás eredményeit, és megakadályozni, hogy rossz minőségű tankönyvekkel kísérletezzünk a gyerekeken." (Dárdai et al. 2015) Hazai kutató (Előry 2008) is felteszi a kérdés: vajon a gyenge PISA-eredményekért mekkora mértékben felelős az értelmezendő szövegek olvashatósága és a tankönyvek minősége?

Az összehasonlító tankönyvi vizsgálatok felhívják a figyelmet azokra a tényezőkre, amelyek a használhatóság és a tanulás eredményessége szempontjából mérvadók lehetnek. Ismereteket közölnek arra vonatkozóan, hogy e feature-öknek mekkora szerepe van a tankönyvek hatásmechanizmusában (Kojanitz 2003a). A tankönyvelemzés szempontrendszerébe nemcsak a legegyszerűbb statisztikai adatok (szó- és mondathosszúság) tartoznak bele, hanem bonyolultabb nyelvi összetevők is, mint például a szavak jelentésének elvontsága, a szófajiság, a szenvedő szerkezetű szavak százalékaránya, a konkrét főnevek százalékaránya, az idegen szavak szótárában megtalálható szavak százalékaránya (Kojanitz 2004b). A releváns tényezők pedig más – nemcsak tankönyvi – szövegekre is vonatkoztathatók.

Az észt Jaan Mikk munkássága kiemelkedő a témában. A kutató meggyőződése, hogy a legtöbb országban (!) kevésbé bonyolult tankönyvekre lenne szükség. Nemcsak feltárta az olvashatóság kritériumait, hanem módszereket is kidolgozott, melyek segítenek az optimális nehézségi szint meghatározásában. Angol nyelvű könyve (Mikk 2000) többek között az alábbi témákat tekinti át:

– tankönyvek értékelése kísérletekkel;
– tankönyvek elemzése;
– olvashatósági formulák;
– a számítógépek használata a tankönyvírás és -elemzés során;

– javaslatok az átfogó szövegalkotáshoz;
– a szöveg optimális olvashatóságának kritériumai.

Hazánkban is készültek szintetizáló tanulmányok az olvashatóság tényezőit illetően. Kiemelt figyelmet igényel Domonkosi Ágnes (2013), valamint Lukács és munkatársai (2022) munkássága mellett Juhász Valéria (2024) kutatása, aki azt vizsgálta, hogy az Aranyfa gyerekkönyvsorozat szintezett kiadványai milyen jellemzőkkel támogatják az olvasáselsajátítás folyamatát. Kojanitz László (2003a; 2003b; 2004a; 2004b), Előry Vilma (2005; 2006; 2008) és B. Fejes Katalin (2002) ugyancsak mélyrehatóan foglalkoztak pedagógiai szövegek analitikus vizsgálatával. Kojanitz László – Mikkhez hasonlóan – a tankönyvek tanulhatóságra helyezi a hangsúlyt. Ahhoz, hogy a „tanulók teljesítménykülönbségeit helyesen értelmezni lehessen, szükség van a kísérletben kipróbált tankönyvek szövegjellemzőit tükröző adatsorokra." (Kojanitz 2004a: 430) Ennek érdekében feltárta és értékelte az érthetőséget befolyásoló szövegtényezőket. B. Fejes Katalin (2002) tankönyvszövegek szintaktikai jellemzőit vizsgálta, míg Előry Vilma (2008) a szöveg szintje felől közelítve kilenc releváns tényezőt azonosított az érthetőség-tanulhatóság szempontjából.

Megállapításaik ismertetésére a következőkben kerül sor, hierarchikus rendezőelv alapján. Ugyanis a pedagógiai szöveganalízisek bemutatásakor célszerű – sőt, szokás – a szavak, a mondatok és a szöveg szintje szerint előrehaladni. Megértési nehézségek e fokokon jelentkezhetnek (Kojanitz 2004a; Domonkosi 2013). Fóris Ágota (2002) is említést tesz a szövegelemzés három – morfológiai, szintaktikai és szemantikai – szintjéről, s azok szegmentálásáról. Az első két kategória viszonylag egyszerűbb vizsgálatokat jelent, az ismertebb olvashatósági formulák (Dale–Chall Formula, Flesch–Kincaid Grade Level, ARI, SMOG stb.) is ezekre építenek. Nem véletlenül, hiszen a szemantikai szintű vizsgálatok már jóval bonyolultabbak, módszertanuk sem annyira kiforrott még.

A szöveganalízis-vizsgálatokat az egyre szélesebb körben elérhető, egyre modernebb szövegelemző programok nagyban megkönnyítik (Fóris 2002). Ezekről a későbbiekben teszek említést, mindenekelőtt azonban – a nyelvi szintek felől megközelítve – ismertetem az olvashatóságot meghatározó szövegtényezők lehetséges listáját, a fejezet végén pedig táblázatos formában összegzem azokat.

A szavak szintje

A szöveg által igénybe vett szókincs, jellemzően a ritka vagy hosszú szavak meghatározzák az olvashatóságot és nehezítik a szövegértést (Lukács et al. 2022).

Szófamiliaritás

A szóelőhívás hatékonyságát leginkább a szövegben szereplő szavak gyakorisága (ismertsége) befolyásolja (Lukács et al. 2022). A szavak ismertsége az alapján becsülhető meg, hogy egy szó milyen gyakorisággal fordul elő a szövegkorpuszban. Az alacsony előfordulási gyakoriságú szavakról azt feltételezzük, hogy kevésbé ismertek, és ezért nehezebbek, mivel az olvasó nem találkozik velük olyan gyakran. Minél több alacsony gyakoriságú szót használ a szöveg, annál nehezebb lesz (Leroy–Kauchak 2014). Például: „Ösztövér kútágas, hórihorgas gémmel" (Arany 1954: 5). (Egy hatodikas tanuló számára akár *mindegyik* szó ismeretlen lehet.)

A ritkán használt szavak aránya és az olvashatóság hazai kutatók szerint is összefügg egymással (Cs. Czachesz – Csirik 2002; Kojanitz 2003b; 2004a; 2004b; Nagy 2004). „Minél több az ismeretlen szó, a szövegértés annál nehézkesebbé, az olvasás fárasztóvá, abbahagyásra késztetővé válik." (Nagy 2004: 125)

Nagy József (2004) szerint az olvasás-szövegértés alapvető feltétele – már felső tagozattól – a leggyakoribb 5000 szó rutinszerű felismerése. Az 5000 leggyakoribb szó százalékos aránya így a lexikai hozzáférhetőség hasznos mutatója lehet.

A szógyakoriság mérése előre összeállított szólista alapján történik. A több ezer szavas listán az adott nyelv leggyakoribb szavai szerepelnek, gyakoriságuk alapján rangsorolva. A kérdés, hogy a listán szereplő kifejezések milyen arányban vannak jelen a szövegben (Lukács et al. 2022). A Magyar Nemzeti Szövegtár (Oravetz et al. 2014) gyakorisági listája irányadó lehet ugyan, viszont figyelembe kell venni, hogy azt felnőtt beszélőkre dolgozták ki, egy általános iskolás számára egészen más szavak lehetnek ismertek. A Czachesz–Csirik-szótár (Czachesz–Csirik 2002) gyűjtötte össze a 10–16 éves tanulók írásbeli szókincsének leggyakoribb szavait, viszont csak korlátozottan lehet rá támaszkodni egyrészt módszertana, másrészt 2002-es megjelenése miatt. Nem ismert olyan online adatbázis, amely a szófamiliaritás a tanulók felől közelítené meg, így csupán a kapcsolódó tanulmányokra támaszkodhatunk.

Kojanitz (2004a) szerint régies és/vagy ritkán használt kifejezések, tudományos szakszavak, idegen (alakú) szavak számíthatnak nehéznek. A szöveg komplikáltságának vizsgálatakor ezek, vagyis a kevésbé gyakori szavak számát és sűrűségét érdemes mérni.

A szógyakoriságot a tartalomhordozó szavak (főnevek, igék, melléknevek, számnevek) esetében külön-külön is érdemes kiszámolni (Kojanitz 2004a). Így tett Uibo (1995) is, aki iskolai tankönyvek elemzésekor figyelembe vette a szóosztályok megoszlási arányát, valamint azt is megnézte, hogy e szavak hány százaléka tartozik a leggyakoribbak közé.

Szóhosszúság

„A nyelvészetben régóta ismert tény, hogy a gyakoriság és a szóhossz egymástól nem függetlenek, így a szavak hosszúsága is jó mutatója lehet az ismerősségnek." (Lukács et al. 2022: 68)

Az MNSZ gyakorisági listáján[2] elöl szerepelnek a *vagy, mint, minden, jó, ő, mondta* szavak. A hosszabb szavak – *kijelentette, intézmények, természetes, alkalommal* – inkább (de nem kizárólagosan) a lista végén találhatók.

Az olvasásgyakorlás rövid, egy–három szótagból álló, morfológiailag egyszerűbb szerkezetű szavakkal kezdődik. A rövidebb és gyakran látott szavak elkezdenek vizuálisan is tárolódni a gyermek agyában – idővel elég lesz rájuk pillantania, hogy felismerje azokat (Juhász 2024). „A nagy gyakoriságú, rövidebb szavak (különösen a *zárt szóosztályú szavak, például a de, hogy, majd kötőszók*) dekódolása azért sikeresebb, mint a ritkább vagy hosszabb, morfológiailag komplexebb szavaké, mert a gyermek könnyebben tudja azonosítani a szót, ha az már létezik a mentális lexikonjában, mint ha nem ismeri a szót." (Juhász 2024: 235)

A tankönyvkutatás hazai képviselője (Kojanitz 2004a) is hasonló véleményen van: az ismert szavak általában rövidebbek, az elvont jelentésűek pedig általában hosszabbak. Az átlagos szóhosszúság összefügg a szöveg tartalmi komplikáltságával.

Morfológiai komplexitás

„Az átlagosan alacsony morfémakomplexitás elősegíti a szövegek jobb érthetőségét, hiszen az olvasóknak kevesebb nyelvi jel jelentését kell feldolgozniuk egy-egy szóhoz kapcsolódóan." (Juhász 2024: 247) Juhász szerint a könnyebben olvasható szavak egyik ismérve az alacsony morfológiai komplexitás: ezek jellemzően 1–3 morfémából állnak (például olvas, olvasás, olvashatóság). A 3+ morfémából álló szavak (például olvashatatlanul)

[2] http://corpus.nytud.hu/cgi-bin/mnszgyak

a szintezettség magasabb fokán jelennek meg. Ugyan kutatása nem igazolta a morfémaszám szintenkénti növekedését, de a morfológiai komplexitással érdemes lehet számolni.

Szófaj

A külföldi szakirodalom szerint a szófajok közül az új tartalmat hordozó főnév, illetve a melléknév okozza a legtöbb megértési nehézséget a mondaton belül (Elley 1969; Klein-Braley 1985). Mikk (1997) vizsgálatai rámutattak, hogy a szófamiliaritás ebben a két csoportban alakul a legkedvezőtlenebbül. A főnevek és a melléknevek szófaján belül magas gyakorisággal fordulnak elő ritka szavak – ami korrelál a nehézséggel. A kutató szerint a szófajkomplikáltsági index kiszámítása[3] hozzájárulhat az érthetőség méréséhez (Mikk 1997).

A hazai szakirodalom szerint az egyszerűbb szövegekre az igék és a határozószók, míg a bonyolultabb szövegekre a főnevek és a melléknevek túlsúlya jellemző (Juhász 2024). Ilyen értelemben egyszerűbb mondat: Menj egyenesen, aztán fordulj balra, majd jobbra, végül megérkezel.

A kutatónak azonban nem sikerült igazolnia ezt az állítást. Mint az 1. ábrán látható, a könnyebbnek számító (például 3., 4. szintű szövegek) esetében is a főnevek vannak túlsúlyban, s nem az igék.

1. ábra: A főnevek-igék arányának változása a szintek emelkedésével
forrás: Juhász 2024

A szavak elvontsága

A főnevek csoportokba sorolhatók, aszerint, hogy konkrét (közvetlenül érzékelhető) tárgyakat, folyamatokat, jelenségeket vagy elvont (közvetlenül nem érzékelhető) dolgokat jelölnek. A szöveg elvontsági szintje az elvont főnevek számának ismeretében számítható ki (Mikk 2000). Kutatások igazolták, hogy a konkrét jelentésű szavakat – képkiváltó értékük miatt – könnyebb előhívni, mint az elvontakat (Kroll–Merves 1986; Bleasdale 1987; Strain et al. 1995). Kojanitz (Kojanitz 2003b) *a* „problémamentes" és a „problémás" főne-

[3] szófajkomplikáltsági index = (főnevek száma + melléknevek száma) / (igék száma + határozószók száma)

vek csoportjait különbözteti meg. Az ő olvasatában a szöveg elvontsági szintjét ezek hányadosa adja. A konkrét főnevek megfeleltethetők a problémamenteseknek, az elvont főnevek pedig a több problémát okozóknak.

Juhász (2024) meghatározónak véli nemcsak a főnevek, hanem – tágabban – a szavak konkrétságát, azaz képkiváltó értékét, vizualizálhatóságát. Az általa vizsgált, különböző szintezésű gyerekkönyv-részletek igazolták, hogy a könnyebb szövegekben a konkrét szavak (például tátongott, úszáspróba, valakivel tart) aránya magasabb.

Idegen szavak

Az idegen szavak mennyisége és előfordulási sűrűsége fontos adat lehet a szöveg érthetőségének vizsgálatakor, amit Mikk (1980) vizsgálata is igazolt. A folyamatosan beépülő jövevényszavak miatt nem könnyű megítélni, aktuálisan mi számít idegen szónak (Lukács et al. 2022). A szakirodalom az idegen szavak szótárát javasolja mint eszközt: „Ha egy szó benne van a szótárban, akkor azt mindenképp figyelembe veszik a mérésnél, s nem mérlegelik, hogy egy közismert idegen szóról, vagy egy tudományterület speciális szakszaváról van-e szó." (Kojanitz 2004a: 436). Sok idegen szót tartalmazó mondatra példa:

„A magas Si-érték a problémamegoldásban a sztereotípiára való hajlamot, az originalitáshiányt, a bizalmatlanságot, rigiditást […] és az autoritással szembeni szubmisszív magatartást jelenti […]." (Bagdy–Safir szerk. 2004: 81)

Szakszavak

A tanulók kevésbé tartják érdekesnek és érthetőnek azokat a szövegeket, amelyek arányaiban sok szakszót tartalmaznak (Mikk 2000). A lexikai anyag kapcsán a szakszavak típusára és előfordulási sűrűségére érdemes fókuszálni, emellett célszerű megvizsgálni, hogy a nem szakszavak mennyire állnak összhangban az adott korosztály szókincsével (Előry 2008) A szakszavak azonosítása – az idegen szavakkal ellentétben – komplikált feladat (Kojanitz 2004a).

Ismétlődés

A szavak, illetve azok allomorfjainak (például játszott, játszották, játszik) ismétlődése – a silabizálás helyett – a rápillantásszerű felismerést teszik lehetővé. Ha pedig egy szó ismerős, valamennyi formájában, az kedvező hatással van az olvashatóságra ld. szófamiliaritás (Juhász 2024).

A mondatok szintje

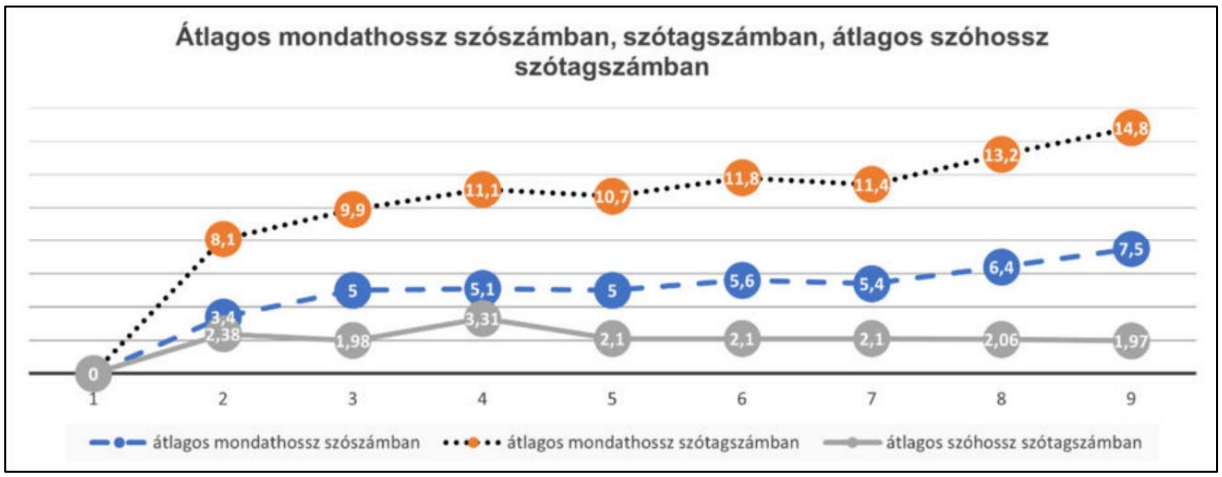
A szöveg mondattani jellemzői közül az alábbiak hozhatók összefüggésbe a szöveg olvashatóságával.

Mondathosszúság

Nemzetközi kutatások igazolták a szignifikáns összefüggést az átlagos mondathosszúság és a tanulók szövegértése között. A tankönyvelemzések azt mutatják, hogy a szerzők is tudatában vannak ennek jelentőségével (Kojanitz 2004b).

Habár az n-ben mért mondathosszúság önmagában keveset mond (Előry 2005), a szöveg bonyolultságáról mégis árulkodni fog (Mikk 2000). Egy hosszabb mondatban nagyobb eséllyel fordulnak elő bonyolult szerkezetek, valamint az összetartozó jelentésegységek sem feltétlenül egymás mellé kerülnek. A megértést az is akadályozza, hogy a hosszú mondatok tárolása a rövid távú memóriában jóval nehezebb.

Juhász (2024) azt vizsgálta, hogy az Aranyfa gyerekkönyvsorozat szintezett kiadványai milyen jellemzőkkel támogatják az olvasáselsajátítás folyamatát. Egyik feltételezése szerint „a mondatok a kezdeti szinteken átlagosan 4-5 szóból állnak, a magasabb szinteken nő a mondatok szószáma. Ezzel arányosan a mondatok szótagszáma is növekszik" (Juhász 2024: 241). Hipotézise igazolódott, melyet a 2. ábra szemléltet. A mondatok szószáma a szintezettséggel együtt emelkedik. Míg az 1. szinten az átlagos mondathosszúság 3,4 szó, a 9. szinten már 7,5. Hasonlóképp a szótagszám esetében, mely 8,1-ről 14,8-ra emelkedik. (A szavak átlagos karakterszáma, valamint szótagszáma a szintek növekedésével nem mutat változást.)



2. ábra: Az átlagos mondathossz meghatározása szószámban, szótagszámban, az átlagos szóhossz meghatározása szótagszámban szintenként
Forrás: Juhász 2024

A mondathosszúságot szavakban, szótagokban vagy morfémaszámban is lehet mérni. Besznyák (2023) a szószámban mért mondathosszúság mellett érvel. Kojanitz (2003b; 2004b) szerint viszont a legcélravezetőbb a – szóközöket és írásjeleket is magába foglaló – betűhelyek számlálása. A kutató – magyar szakiskolai tankönyvek esetében – a 100 betűhelynél hosszabb mondatokat tekinti nehéznek. 200 betűhelyből álló mondat:

„A gyarmatosítók Amerikából nemesfémet és gyarmatárut (gyapot, cukor, dohány) szállítottak Európába, míg az öreg kontinensről a gyarmatosítóknak szükséges iparcikkek (textíliák, fegyverek, szerszámok) hajóztak visszafelé." (Száray 2021: 10)

Az általa vizsgált, általános és szakiskolai tankönyvek többségére 80-100 betűhely/mondat volt jellemző.

Természetesen az adott életkor számára nehézséget jelentő mondathossz megállapítása elengedhetetlen (Lukács et al. 2022). Az n-hosszúságú mondatok előfordulási gyakorisága és az adott korosztály szövegértési teljesítménye között korreláció kell legyen.

Hangsúlyozandó az is, hogy a mondathossz – önmagában – nem ad megbízható képet az olvashatóságról. A tényezők mindig egymáshoz való viszonyukban (!) határozzák meg a mondatok érthetőségét, feldolgozhatóságát.

Olvashatósági pontszám

Domonkosi (2013) az olvashatósági formulákat a szintaktikai szinthez sorolja, tekintve, hogy mindegyik formulában megjelenik a mondattényező. Nem véletlenül, hiszen a mondathossz és az érthetőség között szignifikáns összefüggés van. Az olvashatósági pontszámok önmagukban kevésnek bizonyulnak, viszont részletes elemzéshez jó kiindulópontot

jelenthetnek (Besznyák 2023) – például képet adhatnak a vizsgált szövegek egymáshoz viszonyított nehézségéről.

A következő szövegek egy 4. osztályos olvasókönyvből (Kóródi 2022) származnak. A RIX-formula – mely a hosszú, vagyis minimum hatkarakteres szavak/mondatok számával kalkulál – 12,5-ös olvashatósági pontszámot jósolt (12.-es évfolyamszint):

„Boldogan s nagy egyetértésben teltek a napok, a kővirág büszkén feszített az ablakban, s arra gondolt, hogy a kinti virágok bizonyára irigykednek, ha netalántán fölpillantanak Bab Berci ablakába. De erről soha nem bizonyosodott meg, lehet, hogy a kinti virágok észrevették őt, de az is lehet, hogy nem. Élték mindennapi életüket, örültek a napfénynek, esőnek, olykor még a szélnek is, mert a szélben kedvükre táncolhattak, hajladozhattak." (Kóródi 2022: 51).

Szintén 4. osztályos szöveg, RIX-formulával mérve 5,5-ös nehézségű:

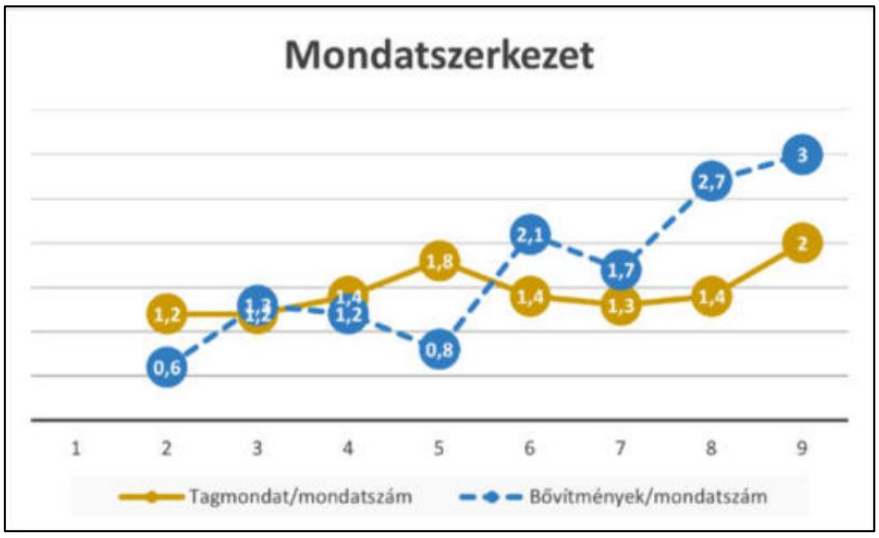
„A sziklák a köveknél jóval nagyobb méretűek. Az anyaguk viszont nagyon hasonló vagy éppenséggel ugyanaz lehet, mint a kőé. A sziklák olyan magasba nyúló, különböző alakú csupasz kőzetek, amelyek gyakran valamilyen hegy részét képezik. Ilyen gyönyörű szikla a Balaton-felvidéken található Hegyestű." (Kóródi 2022: 53).

Mondatszerkezet

Az olvashatóbb szövegek jellemzője, hogy egyszerű, kevés bővítménnyel rendelkező mondatokat tartalmaznak. Az összetett mondatok nehezítik a megértést (Juhász 2024), például:

„Megtetszett a királyfinak a királykisasszony, mert akármilyen rongyos volt, akármilyen piszkos volt a ruhája, szép volt, kellemetes volt az arca." (Kóródi 2022: 8).

Juhász (2024) kilencféle nehézségi szintű gyerekkönyv 10-10 mondatát vizsgálva kimutatta, hogy a *tagmondatokon belüli bővítmények száma* – az 5. szintű könyv kivételével – a nehézségi szinttel együtt nő (ld. 3. ábra).



3. ábra: A tagmondatok
és a tagmondatokban lévő bővítmények számának szintenkénti jellemzői
Forrás: Juhász 2024

Előry (2005, 2006) is utal a mondatok szerkesztettségére (egyszerű és összetett mondatok aránya) mint befolyásoló tényezőre. A mondatok szintmélysége szintén meghatározó, vagyis hogy a többszörösen összetett mondatok mellé- vagy alárendelő kapcsolatban vannak-e egymással. Többszörösen összetett, mellérendelő mondat:

„Vagy a vonaton felejtettem, vagy már előbb elvesztettem az ernyőmet, de a bőröndben is lehet." (Legeza 2020: 95).

Többszörösen összetett, alárendelő mondat:

„Olyan szép volt az este, így elindultunk sétálni, mert nincs szebb a balatoni naplementénél." (Legeza 2020: 95).

A tagmondatok tagoltsága kapcsán a szabad mondatrészek aránya lesz mérvadó. Említi még a mondatrészek zsúfoltságát (grammatikai funkciójú elemek száma) és telítettségét (szavak száma mondatrészenként).

Deme László (1971) elemzési rendszere mondatszerkezeti jellemzőkre fókuszál. Habár egy fél évszázados elméletről beszélünk, „mit sem vesztett időszerűségéből és használhatóságából. A mutatók kifejezetten alkalmasak a viszonyításra és az összehasonlításra" (Zimányi 2025: 51). Ezek a mutatók a következők:

– *szerkesztettségi mutató* (a mondategységek és a mondategészek hányadosa): azt jelzi, hogy egy mondategész átlagosan hány tagmondatból áll – a kevés tagmondatból álló mondatok olvashatóbbak;

– *telítettségi mutató* (a mondategészek és a mondategységek szóelőfordulásainak átlaga): a rövid mondatok, melyek kevés szóból állnak, előnyösebbek;

– *bonyolultsági mutató*: jelzi a mondategészeken belül a mondategységek kapcsolásainak számát, illetve annak átlagát – minél több mondategységből áll a mondat, annál több a kapcsolás, vagyis annál bonyolultabb.

Szórend

A szórend kapcsán érdemes megvizsgálni, hogy az mennyire emeli ki a fontos/új közlést (Előry 2005, 2006).

A nyomatékos mondat – mivel egy és csakis egy körülményt helyez előtérbe – segít az új információ kiemelésében. A mondat nyomatékossá válik, ha a fontos mondatrészt kiemeljük a többi közül, s nagyobb nyomatékkal ejtjük, azaz fókuszáljuk:

„Délelőtt anya főz a konyhában." (Kövérné Nagyházi 2003: 56).

Vagyis anya főz, és nem más. Nagyházi (2011) explicit módon írja le a magyar szórendi törvényszerűségeket, tanulhatóvá és taníthatóvá téve azokat a magyart mint idegen nyelvet tanulók és tanítók számára. Értekezésében hangsúlyozza: habár a magyar nyelvben sokféle szórendi variáció megengedett, léteznek szórendet befolyásoló szabályok.

„[…] A mondat szórendje, összetevőinek egymásutánisága befolyásol(hat)ja a megértést, a szövegértelmezést." (Gombos–Nagyházi 2023: 502) Bírósági határozatok szórendi jellemzőit vizsgálva a kutatópáros azt találta, hogy az ige (állítmány) helye rendszerint a mondat végén van. Ezáltal túl sok elemet fog tartalmazni a mondat állítmány előtti része – így az megterheltté (az olvasó számára pedig megterhelővé) válik. Például:

„Az alperes kártérítési felelősségének elbírálásához szükséges bizonyítást a bíróságok *lefolytatták*." (Gombos–Nagyházi 2023: 506)[4]

Ugyan a magyar írott nyelvre jellemző a balra bővítő szerkezet – halmozott mondatrészek az ige előtt –, mégis kerülendő, mert nehézséget okoz (Juhász–Radics 2024). Olvashatóság szempontjából a legkedvezőbb, ha az állítmány a mondat közepére kerül.

[4] Javaslat a mondat egyensúlyának visszaállítására: „A bíróságok *lefolytatták* az alperes kártérítési felelősségének elbírálásához szükséges bizonyítást." (Gombos–Nagyházi 2023: 506)

A mondatközi kapcsolatok és a szöveg szintje

Kohézió és koherencia

Míg a koherencia a szöveg általános logikai felépítésére koncentrál (a legfelsőbb szintre), addig a kohézió a szövegrészek és a mondatok összetartását jelenti. Az egymásból következő mondatok könnyebben olvashatók. A kohéziót explicit nyelvi elemek (például kötőszók vagy anaforák) jelölik, melyek segítik – de nem biztosítják – a szöveg koherenciáját. Az emberek gyorsabban olvasnak egy logikusan felépített (koherens) szöveget, és jobban is emlékeznek rá, mint egy rosszul strukturáltra. A koherencia szerepet játszik az érthetőség/nehézség megítélésében (Meyer 2003; Lukács et al. 2022). A kohézió szintén fontos előrejelzője a szöveg olvashatóságának (Todirascu et al. 2013).

Előry (2008) elsősorban a kohézióval és a koherenciával kapcsolatos szempontokat emeli ki mint releváns szövegtényezőket. (Kilenc elemből álló listája tartalmaz olyan tényezőket is, melyek korábban, a szavak és a mondatok szintjénél már említettem.)

A szöveg szerkezetére, szerkesztettségére vonatkozó szempontjai a következők:

– *A mondategészek kapcsolódása egymáshoz* (például ismétlések által): a szomszédos mondatok közötti tartalmi, logikai, grammatikai kapcsolat;

– *a mondategész kapcsolata a szövegegésszel*: minden egyes mondat kapcsolódik a szöveg egészéhez, mindegyik szükséges a megértéshez, nincsenek felesleges, oda nem illő részek;

– *a szövegegységek kapcsolata egymással*: logikai, nyelvi kapcsolat a bekezdések között;

– *szövegtagolás és gondolati egységek*: a szöveg tagolása (bekezdések, mondattömbök, mondathatárok) megfelel a szöveg gondolati rendszerének.

Az olvashatóságot meghatározó szövegtényezők lehetséges listáját az *1. táblázat* összegzi.

1. táblázat: Az olvashatóságot meghatározó szövegtényezők lehetséges listája a szakirodalom alapján (saját szerkesztés)

Nyelvi szintek	tényező neve	mérése
a szavak szintje	szófamiliaritás	a kevésbé gyakori szavak száma és aránya (szófajonként is)
		5000 leggyakoribb szó százalékos aránya (szófajonként is)
		zárt szóosztályú szavak (kötőszók, segédigék, névelők, névutók, perpozíciók, névmások) aránya
	morfológiai komplexitás	1–3 morfémát tartalmazó szavak aránya
		3-nál több morfémát tartalmazó szavak aránya
	szóhosszúság	átlagos szóhosszúság karakter-, szótag- vagy morfémaszámban mérve
	szófaj	igék és határozószók aránya
		főnevek és melléknevek aránya
	elvontság	konkrét és elvont főnevek aránya
		konkrét (vizualizálható) szavak aránya
	idegen szavak	idegen szavak szótárában szereplő szavak aránya
	szakszavak	a szakszavak típusa és előfordulási sűrűsége
		a nem szakszavak mennyire állnak összehangban az adott korosztály szókincsével
	ismétlődés	a szavak, illetve azok allomorfjainak ismétlődése
a mondatok szintje	mondathosszúság	átlagos mondathosszúság betűhelyekben mérve
		átlagos mondathosszúság szószámban mérve
		átlagos mondathosszúság szótagszámban mérve
		átlagos mondathosszúság morfémaszámban mérve

	olvashatósági pontszám	olvashatóságiindex-számítás ismertebb formulákkal
	szórend és mondatszerkezet	egyszerű és összetett mondatok aránya
		tagmondatokon belüli bővítmények száma
		mellérendelő/alárendelő mondatok
		a szabad mondatrészek aránya
		grammatikai funkciójú elemek száma
		szavak száma mondatrészenként
		a szórend mennyire emeli ki a fontos/új közlést
		az állítmány helye a mondatban
		mondategységek/mondategészek (szerkesztettségi mutató)
		a mondategészek és a mondategységek szóelőfordulásainak átlaga (telítettségi mutató)
		a mondategészeken belül a mondategységek kapcsolásainak száma, illetve annak átlaga (bonyolultsági mutató)
a mondatközi kapcsolatok és a szöveg szintje	kohézió és koherencia	a mondategészek kapcsolódása egymáshoz
		a mondategész kapcsolata a szövegegésszel
		a szövegegységek kapcsolata egymással
		szövegtagolás és gondolati egységek

Olvashatóságot mérő eszközök: számítógépes elemző szoftverek

A fenti tényezők befolyásolhatják a fogalmak koherens reprezentációjának kialakítását, és a megértést. Tanulmányozásuk manuális úton túlságosan körülményes, viszont nagy segítséget jelenthet egy elméleti alapokra épülő automatizált eszköz.

E fejezet a rendelkezésre álló szoftverek összehasonlító értékelését tartalmazza. A szövegtényezők feltárása annak érdekében történt a korábbi fejezetekben, hogy a magyar nyelvű szövegek esetében is konkretizálódjanak az olvashatóságot meghatározó szövegtényezők. Ezek manuális úton történő elemzése több szempontból sem előnyös, ezért érdemes áttekinteni a rendelkezésre álló, automatizált szövegelemző eszközöket. E szoftverek bevonása elengedhetetlen ahhoz, hogy a kutatás elérje a végső célját: mérhetővé váljék a magyar nyelvű szövegek olvashatósága.

Az olvashatóság mérése háromféle módszerrel történhet: (1) olvashatósági formulákkal, (2) természetesnyelv-feldolgozáson (natural language processing, NLP) alapuló, gépi tanulási módszerekkel[5], (3) neurális hálókra épülő mélytanulási (deep learning) modellekkel[6]. Az olvashatósági formulák előnye az átláthatóság (viszont pontatlanok), a mélytanulási modelleké pedig a pontosság (viszont a bementi jegyek csak a neurális háló számára láthatóak) (Lukács et al. 2022).

Az AI automatizált algoritmusai képesek nagy mennyiségű, strukturálatlan szövegből kinyerni az indirekt információkat. A manuális szövegelemzéssel szemben effektívebb munkát végeznek: a korábbi találatokból tanulva képesek „önfejlődésre", teljesítményük fokozására. A bennük rejlő lehetőségek akár az olvashatóságmérés területén is kiaknázhatók (Látics–Gombos 2025).

[5] „Nyelvi jegyeken tanított, komplex algoritmikus modelleket használnak arra, hogy megbecsüljék egy szöveg olvashatóságát." (Lukács et al. 2022: 75) Képesek számszerűsíteni például egy adott mondat szintaktikai mélységét vagy a lexikai sűrűséget (Dárdai et al. 2015).

[6] Ezek követlenül a referenciaszöveget dolgozzák fel, a bemeneti jegyek meghatározását gép végzi (Lukács et al. 2022).

Mező Péter Dániel (2023) a gépi tanulást alkalmazó algoritmusokat két csoportba sorolja: (1) felügyelt (supervised), (2) nem felügyelt (unsupervised) tanulást alkalmazó algoritmusok.

A felügyelt tanuláshoz ember szükséges: az „algoritmus ember által megadott és összekapcsolt bemeneti és kimeneti adatok bázisán állít fel olyan modellt, ami alapján a jövőben a korábban nem ismert bemeneti adatokhoz hozzá tudjon társítani valamilyen kimeneti adatot" (Mező 2023: 69). Ehhez kezdetben „labelled", vagyis címkézett, egymáshoz kapcsolt ki- és bemeneti adatok szükségesek (Mező 2023). Az NLP-alapú és gépi tanulást alkalmazó modellek – amiket Lukács és munkatársai a (2)-es csoportba sorol – ekképp működnek.

A nem felügyelt tanulás esetében nincs szükség ember által megadott címkézett adatokra. A számítógép a betanulási-betanítási fázis nélkül is képes felállítani egy modellt a szöveg alapján (Mező 2023).

A gépi tanulás és a mélytanulás – beleértve a neurális hálók és az NLP-k – működését mélyrehatóbban ismertetik Chris Minnick (2025) könyvének, valamint Hoss Alexandra (2022) doktori értekezésének kapcsolódó fejezetei, Russel–Norvig (2023) könyve pedig a téma teljes körű feltárására tesz kísérletet.

A számítógépes nyelvészet szakterülete egyre népszerűbb, melyet az immár több mint 20 éves Magyar Számítógépes Nyelvészeti Konferencia megléte is igazol[7]. Vitathatatlan, hogy az AI képes szövegelemző feladatok ellátására. De vajon milyen eszközök állnak rendelkezésre ehhez? „A szöveganalízis-vizsgálatok megkönnyítésére egyre több és egyre modernebb szövegelemző programot készítenek a számítógépes nyelvi kutatással foglalkozók." (Fóris 2002: 31) – a következőkben ezeket ismertetem. Valamint bemutatok néhány olyan kutatást, amelyekben a szövegek nehézségét NLP-alapú vagy mélytanulási modellekkel mérték.

Természetesnyelv-feldolgozáson alapuló, gépi tanulási módszerek[8]

A gépi tanulás a mesterséges intelligencia egyik ága, amely olyan algoritmusok és statisztikai modellek fejlesztésével foglalkozik, amelyek lehetővé teszik a számítógépek számára, hogy tanuljanak a tapasztalatokból és fejlesszék magukat anélkül, hogy kifejezetten programoznák őket (Setiawan et al. 2023).

A Coh-Metrix (Graesser et al. 2011) egy fejlett számítógépes nyelvészeti eszköz. Átfogó képest nyújt a szövegek kohéziójáról és koherenciájáról. Hatékonysága abban rejlik, hogy hűen képes reprezentálni az olvasás során aktivizált kognitív műveleteket (Szabó 2019). A Coh-Metrix TERA eredményei a következő hat területre terjednek ki: szintaktikai egyszerűség, szókonkrétság, referenciális kohézió[9], mélykohézió[10] és idegen nyelvi olvashatóság.

Szabó (2019) kutatásának célja az volt, hogy felmérje az angol nyelvi érettségi szövegek nehézségét, szignifikáns különbséget keresve az eltérő vizsgaidőszakokban használt textusok jellemzői között. A vizsgálat 33 vizsgaszövegre terjedt ki, a jellemzők kivonása a

[7] A konferenciára többek között a következő témakörökben várnak tanulmányokat: morfológiai és szintaktikai elemzés, korpusznyelvészet, szövegbányászat, gépi fordítás, nyelvi intelligenciára épülő alkalmazások (https://rgai.inf.u-szeged.hu/mszny2025).

[8] A teljesség igénye nélkül.

[9] Megmutatja, hogy mekkora az átfedés az egyes mondatokban megjelenő gondolatok, valamint szavak között. Az alacsony referenciális kohézió negatív hatással van az olvashatóságra (Szabó 2019).

[10] „Ez a komponens a szövegben megjelenő események, fogalmak és információk összekapcsolódását elemzi, és elsősorban a kötőszókon alapul." (Szabó 2019: 111)

Coh-Metrix TERA program segítségével történt. A kutató a szövegek egységes és elemezhető formátummá alakítása után elemzést hajtott végre a program segítségével.[11] A kutatás igazolta, hogy a vizsgákban szereplő szövegcsoportok nem különböznek egymástól szignifikáns módon – vagyis azonos nehézségűek (ami kedvező).

A TextEvaluator[12] egy olyan eszköz, amely szinte bármilyen írott szöveg elemzésére alkalmas, és mélyreható információkat nyújt a szöveg olvashatóságáról és összetettségéről. Ezeket az információkat egy holisztikus pontszámmal összegzi. A program első körben a szöveg formai jellemzőit határozza meg: szavak és mondatok száma, átlagos mondathosszúság szavakban, bekezdések száma, átlagos bekezdéshossz szavakban, idézőjelben szereplő szavak bekezdésenként, Flesch–Kincaid Grade Level-pontszám. Ezt követően számszerűsíti a szöveg általános komplexitását az alábbi paraméterekkel:

– mondatszerkezet: mondathossz és -összetettség;

– a szókincs nehézsége, komplexitása: akadémiai szókincs, ismeretlen szavak, absztrakt szavak;

– lexikai kohézió, interaktív stílus, érvelő beszédmód[13];

– a narrativitás mértéke: idézőjelek között található szavak száma, múlt idejű igék stb.

A TextEvaluator a szöveg nehézségét a megadott célszinthez viszonyítva adja meg – alatta vagy fölötte helyezkedik-e el. Hasznos eszköz oktatók, tartalomfejlesztők, kutatók és olvasók számára egyaránt (Napolitano et al. 2015).

Az indonéz származású Hamdan Hidayat (2023) a 10. osztályos angol nyelvkönyvekben szereplő szövegek nehézségi szintjét elemezte, és meghatározta, hogy azok megfelelnek-e a tanulók szintjének. A kutató a TextEvaluator szoftvert használta a szövegkomplexitás értékelésére. Az elemzett 13 szöveg közül csupán kettő felelt meg a 10. osztályos szintnek. Mint kiderült, a legtöbbjük az elvárható szint alatt van: nem jelentenek kihívást a tanulók számára, ami negatívan befolyásolhatja a tanulók motivációját. Az eredmények rávilágítottak: az olvasmányok értékelése szükséges – még a tantervbe való beépítésük előtt.

A TextInspector egy sokoldalú elemző szoftver: meghatározza az alapvető statisztikai adatokat (átlagos mondathossz, kettőnél több szótagú szavak száma stb.), elemzi a szöveget (lexikai diverzitását, felcímkézi a szavak a szófajuk, gyakoriságuk alapján stb.) és feltárja annak logikai összefüggéseit (Besznyák 2023).[14]

Egy 2025-ös tanulmányban a kutató e szoftver segítségével elemezte a népszerűbb angolnyelv-tankönyvek szövegeit. Arra a kérdésre kereste a választ, hogy szükséges-e módosítani a nehézségi szintjükön a diszlexiával küzdő tanulók megtámogatása érdekében. A TextInspector-elemzés eredményei azt mutatták, hogy a vizsgált tankönyvekben szereplő szövegek egy része nehezebb volt, mint azt a kiadók hirdették. Emellett példákat is hozott arra, hogy milyen hatékony módszerekkel lehet módosítani az olvasási anyagokat a diszlexiás diákok megsegítésére és a tantárgy eredményeinek javítására (Montgomery 2025).

A fentebb említett szoftverek egyike sem képes magyar nyelvű szöveg feldolgozására, viszont léteznek ilyenek. Magyarnyelv-specifikus gépi elemző a Magyarlánc (Zsibrita et al. 2013). Ez képes felismerni és számszerűsíteni a statisztikai, morfológiai, szintaktikai, szemantikai és pragmatikai tényezőket – ahogy az egy kutatásból (Vincze et al. 2021.) is kiderült.

[11] A Coh-Metrix TERA online is elérhető: http://cohmetrix.com

[12] Online elérhető a következő linken: https://textevaluator.ets.org/textevaluator/

[13] A szöveg komplexitásának egyes jellemzői másképp működnek az információs és az irodalmi szövegek esetében. Az irodalmi szövegek nagyobb kihívást jelentenek, mint az ismeretközlő szövegek (Napolitano 2015).

[14] Sajnálatos tény, hogy ez a program is kizárólag angol nyelvű szövegeknél működik.

A Magyarláncnál hatékonyabb az e-magyar.hu (Váradi et al. 2018), vagyis a Digitális Nyelvfeldolgozó Rendszer, mely lehetővé teszi a magyar nyelv digitális elemzését. Képes felfedni és egyértelművé tenni a szöveg jellemzőit. A program a következő modulokat foglalja magába:

– Szövegtagoló (tokenizáló): képes megállapítani a mondat- és szóhatárokat;

– morfológiai elemző: a szöveg minden egyes szóalakjához hozzárendeli az összes lehetséges morfológiai, morfoszintaktikai elemzését, szövegkörnyezettől függetlenül (megállapítja a szótőt, a szófaji főkategóriát, elemzi a toldalékokat, megadja a szóalak morfémákra bontásának lehetőségeit, így a szóösszetételi határokat is);

– szótövesítő: toldalékok azonosításán túl összetételi tagokra bontja a szót, képzőket azonosít;

– egyértelműsítő: meghatározza a korábban tokenekre bontott mondat minden tokenjének szótövét és szófaját, majd ezt címkével jelöli is;

– függőségi mondatelemző: feltárja a mondatok szerkezeti egységei (szavak, többszavas kifejezések) közötti függőségi viszonyokat;

– összetevős mondatelemző: azt tárja fel, hogy a szavak egymással kombinálódva milyen kifejezéseket alkotnak, illetve hogyan állnak össze egy mondattá;

– főnévi csoport (frázis) felismerő

– tulajdonnév-felismerő: azonosítja a folyó szövegben található tulajdonneveket, és besorolja őket az előre meghatározott névkategóriák valamelyikébe.

Az e-magyar nyílt forráskóddal rendelkezik: szabadon elérhető webszolgáltatás[15], illetve beépülő modul formájában is. A folyamatos fejlesztésnek köszönhetően 2019-ben egy újabb verziója (emtsv) vált hozzáférhetővé[16]. „A fejlesztések során kifejezetten szem előtt tartottuk a bölcsészet- és társadalomtudományok művelőit is, hogy számukra is praktikus alkalmazásokat hozzunk létre." (Simon et al. 2020: 40) Az új modell teljesítmény, ki- és belépési lehetőség, valamint új modulok integrálhatósága szempontjából jól teljesített a hasonló funkciókkal bíró magyar nyelvfeldolgozó eszközláncokkal – UDPipe, huspaCy, Magyarlánc – szemben (Simon et al. 2020), ahogy az a 2. táblázatból kiolvasható.

2. táblázat: A rendszerek összehasonlító táblázata a vizsgált paraméterek mentén (O = ha a rendszer az adott paraméter megvizsgálásából pozitívan jön ki, X = ha nem)

	teljesítmény	gyorsaság	nyelvfüggetlenség	ki- és belépési lehetőség	integrálhatóság	használhatóság
emtsv	O	X	X	O	O	O
Magyarlánc	O	X	X	X	X	O
UDPipe	X	O	O	O	X	O
huspaCy	X	O	X	O	O	O

Forrás: Simon et al. 2020

Használhatóságát igazolja, hogy számos friss kutatásban megjelenik mint eszköz; legyen szó egészségügyi álhírekről (Szécsényi et al. 2024), sajtószövegekről (Simon 2024) vagy költészetről (Horváth 2020; Horváth et al. 2022).

Meg kell jegyezni, hogy sem az e-magyar, sem a Magyarlánc nem képes felismerni vagy értékelni a szövegkohéziót és a koherenciát. Ezek az eszközök elsősorban morfológiai-

[15] https://e-magyar.hu/hu/parser

[16] „A teljes elemzőlánc elérhető LGPL 3.0 licenc alatt a rendszer GitHub repozitóriumában" (Simon et al. 2020: 40) a következő linken: https://github.com/nytud/emtsv

szintaktikai szinten működnek (igen hatékonyan), nem pedig diszkurzuselemző rendszerként.

Neurális hálókra épülő mélytanulási modellek

A *BERT* az első olyan finomhangoláson alapuló reprezentációs modell, amely a legkorszerűbb teljesítményt éri el a mondat- és tokenszintű feladatok széles skáláján (Devlin et al. 2019).[17] A BERT-típusú modellek előnye, hogy nincs szükség a jellemzők kinyerésére: mélytanuláson alapulnak, így a jellemzőkinyerés automatikus. Esetükben előfeldolgozásra (stemming, n-grammozás) sincs szükség (Çel˙ikten–Bulut 2021). A szakirodalomnak gyakran „nagy nyelvi modelleknek" (LLM) nevezik őket.

Çel˙ikten és Bulut (2021) török nyelvű, orvosi szövegek osztályozására használták a BERT-et. A több mint 1000 dokumentumot – az alapján, hogy milyen betegséggel kapcsolatosak – tízféle kategóriába tudták besorolni.

Setiawan és munkatársai (2023) hallgatói visszajelzéseket osztályoztak, elemeztek. Vizsgálatuk célja a minőségfejlesztés volt: meghatározni azokat a területeket, egyetemi szolgáltatásokat, amelyek fejlesztése indokolt. A szövegjellemzők kivonásához a BERT-et használták. (A kinyert jellemzőket ezután gépi tanulási modellek bemeneteként használták, a visszajelzések többcímkés osztályozása érdekében.)

Naismith és munkatársai (2023) azt vizsgálták, hogy a GPT-4 mennyire hatékonyan képes észlelni a koherenciát oktatási kontextusban. A korpuszt angolnyelvvizsga-szövegek adták. Emberi értékeléseket használtak referenciaértékként. (Tanulmányuk végén értékelési szempontsor is található.) Megállapították, hogy a GPT-4 képes az írásbeli szövegek koherenciáját jó pontossággal – az emberi értékelésekkel összhangban – osztályozni, a hagyományos NLP-mérőszámokat pedig messze felülmúlja.

Mizumoto és Eguchi (2023) a TOEFL11 adatbázisban[18] található – több mint 12 ezer darab – angol nyelvű esszét GPT-3 segítségével értékelték. A kutatás során összesen 45 féle nyelvi jellemzőt vettek figyelembe. Ezeket öt kategóriába sorolták: lexikai diverzitás, lexikai kifinomultság, szintaktikai komplexitás, szintaktikai jellemzők, ige-tárgy szerkezetek és szövegkohézió. A mérőszámokat automatikusan számították ki a következő NLP-eszközökkel TAALED, TAALES, TAASSC és TAACO[19]. Ezekkel párhuzamosan figyelembe vették a GPT-pontszámokat is. Az eredményeket összehasonlították a TOEFL-referenciaértékekkel (könnyű, közepes vagy nehéz), s mint kiderült, összhangban voltak azokkal. Az eredmények arra utalnak, hogy a GPT-3 pontos és megbízható, így értékes támogatást nyújthat az emberi értékelésekhez.

Az olvashatóságmérés összetett modellezési probléma, hiszen a szövegnehézség több száz változóval definiálható. Az osztályozási feladat megoldásához egy olyan modellre van szükség, ami pontosan és viszonylag gyors módszerekkel képes kiválasztani a valóban releváns jellemzőket. Viharos és munkatársai (2021) hibrid jellemzőválasztási algoritmusa (az ún. AHFS) felülmúlta a legkorszerűbb jellemzőválasztási módszereket, mind a megbízhatóság, mind a feladat végrehajtásához szükséges idő szempontjából.

[17] A kód és az előre betanított modellek a https://github.com/google-research/bert címen elérhetők.

[18] Az adatbázisban található esszéket nem angol anyanyelvű, de angolul tanulók írták. A szövegek különböző nehézségűek: könnyű, közepes és nehéz szintűek. A korpusz célja, hogy adatokat szolgáltasson nyelvészeti és oktatási kutatásokhoz.

[19] Az összes eszköz ingyenesen elérhető itt: https://www.linguisticanalysistools.org/

Összegzés

A tanulmány azonosítja azokat a főbb tényezőket, amelyek hatással lehetnek egy szöveg nehézségére, olvashatóságára. A magyar nyelvű szakirodalom a tankönyvanalízis felől közelítve, a nyelvi szintek szegmentálásával – a szavak, a mondatok és a szöveg szintjén – tárja fel ezeket. A lista korántsem kötött: további elemekkel szabadon bővíthető. A szakirodalom szerint relevánsnak számító szövegtényezőket az 1. táblázat összegezte.

Mint kiderült, az olvashatóság mérhetővé tétele háromféle módszerrel lehetséges: az olvashatósági formulákkal szemben effektívebb megoldást jelentenek a természetesnyelv-feldolgozáson alapuló, gépi tanulási módszerek, de a leghatékonyabbnak mégis a neurális hálókra épülő mélytanulási modellek bizonyulnak. Az NLP-alapú modellek jóval több bemeneti jeggyel dolgoznak, beleértve, ám túlmutatva azokon a szövegtényezőkön, amelyeket írásom első felében taglalok. A szövegelemzés terén nyújtott hatékonyságukat és a bennük rejlő lehetőségeket kutatások igazolják. Velük szemben a mélytanulási (LLM) modellek bemeneti jegyei csak a neurális háló számára láthatók, viszont pontosság terén ezek a legkimagaslóbbak.

A tanulmányban a kézzelfoghatóbb szövegtényezőkön túl néhány modernebb, automatikus szövegelemző szoftvert is bemutattam. Működési elvük ismerete hasznos lehet a magyar nyelvű szövegek géppel történő elemzésekor. Képet adnak arról is, hogy az olvashatóságot meghatározó bemeneti jegyek milyen főbb kategóriákba sorolhatók. Sajnálatos tény, hogy az említett szoftverek – akárcsak az ismertebb olvashatósági formulák – kizárólag angol nyelvű szövegekhez készültek. Léteznek azonban magyar nyelvfeldolgozó eszközláncok: UDPipe, huspaCy, Magyarlánc, s kiváltképp az e-magyar, melynek 2019-es (emtsv néven ismert) új verziója már a gyakorlatban is bizonyította a hatékonyságát. A végső cél elérése érdekében – mérhetővé váljék a magyar nyelvű szövegek olvashatósága – elengedhetetlen a nyelvfeldolgozó eszközláncok ismerete és bevonása.

Irodalom

Arany János (1954): *Toldi.* Budapest: Ifjúsági Könyvkiadó.

Bagdy Emőke – Safir Erika (szerk.) (2004): *Klinikai pszichológiai esettanulmányok*. Budapest: Animula.

Besznyák Rita (2023): *Gyakorlóbeszédek graduálása a tolmácsképzésben korpusznyelvészeti módszerek bevonásával*. [PhD-értekezés]. Budapest: Eötvös Loránd Tudományegyetem Bölcsészettudományi Kar. DOI: https://doi.org/10.15476/elte.2023.272

Bleasdale, F. A. (1987): Concreteness dependent associative priming. Separate lexical organization for concrete and abstract words. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 13, 582–594. DOI: https://doi.org/10.1037//0278-7393.13.4.582

Çel˙ikten, A. – Bulut, H. (2021): BERT Modeli ile Türkçe Medikal Metin Sınıflandırma Turkish Medical Text Classification Using BERT. *29th Signal Processing and Communications Applications Conference (SIU).* Istanbul.

Cs. Czachesz Erzsébet – Csirik János (2002): *10–16 éves tanulók írásbeli szókincsének gyakorisági szótára*. Budapest: BIP.

Dárdai Ágnes – Dévényi Anna – Márhoffer Nikolett – Molnár-Kovács Zsófia (2015): Tankönyvkutatás, tankönyvfejlesztés külföldön II. *Történelemtanítás: online történelemdidaktikai folyóirat*, 6(1-2), [oldalszám nélkül]

Deme László (1971): *Mondatszerkezeti sajátságok gyakorisági vizsgálata*. Budapest: Akadémiai Kiadó.

Devlin, J. – Chang, M-W. – Lee, K. – Toutanova, K. (2019): BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Minneapolis, Minnesota. pp. 4171–4186.
Domonkosi Ágnes (2013): A tankönyvszöveg érthetőségének vizsgálati szempontjai. *Az Eszterházy Károly Főiskola tudományos közleményei,* 40, 27–35.
Elley, W. B. (1969): The assessment of readability by noun frequency counts. *Reading Research Quarterly,* 4, 411–427. DOI: https://doi.org/10.2307/747147
Előry Vilma (2005): A tankönyvszöveg megértése. *Iskolakultúra*, 11, 59–62.
Előry Vilma (2006): A jó tankönyv nyelvi követelményeinek rendszerezése. *Könyv és Nevelés*, 8(2), 28–33.
Előry Vilma (2008): Milyen a jó tankönyvszöveg? In: Medve Anna – Szépe Görgy (szerk.): *Anyanyelvi nevelési tanulmányok III..* Budapest: Iskolakultúra, pp. 7–16.
Fejes Katalin, B. (2002): *A tankönyvszöveg szintaktikai jellemzői*. Szeged: Juhász Gyula Felsőoktatási Kiadó.
Fóris Ágota (2002): *Szótár és oktatás*. Pécs: Iskolakultúra-könyvek 14.
Gombos Péter – Nagyházi Bernadette (2023): Bírósági ítéletek szövegeinek nyelvi sajátosságai. *Magyar Nyelvőr*, 147(4), 493–513. DOI: https://doi.org/10.38143/nyr.2023.4.493
Graesser, A. C. – McNamara, D. S. – Kulikowich, J. M. (2011): Coh-Metrix. Providing Multilevel Analyses of Text Characteristics. *Educational Researcher*. 40(5), 223–234. DOI: https://doi.org/10.3102/0013189x11413260
Hidayat, H. (2023): A Using Text Evaluator to Analyze Reading Texts in Indonesian Grade X English Course Book. *Native: Journal of English Teaching and Learning*, 1(1), [oldalszám nélkül]
Horváth Péter – Kundráth Péter – Indig Balázs – Fellegi Zsófia – Szlávich Eszter – Bajzát Tímea Borbála – Sárközi-Lindner Zsófia – Vida Bence – Karabulut Aslihan – Timári Mária – Palkó Gábor (2024): ELTE Verskorpusz – a magyar kanonikus költészet gépileg annotált adatbázisa. *XVIII. Magyar Számítógépes Nyelvészeti Konferencia*. Szeged. pp. 375–388.
Horváth Péter (2020): A vershangzás jellemzőinek automatikus feltárása József Attila verseiben. *Digitális Bölcsészet*, 3, 3–27. DOI: https://doi.org/10.31400/dh-hun.2020.3.422
Hoss Alexandra (2022): *Nyelvtechnológiai eszközök és eljárások alkalmazhatósága autizmusban.* [PhD-értekezés]. Pécs: PTE BTK Nyelvtudományi Doktori Iskola.
Juhász Valéria – Radics Márta (2024): Beszédfeldolgozást vizsgáló tesztek iskoláskorig. *Módszertani Közlemények*, 64(2), 3–30. DOI: https://doi.org/10.14232/modszertani.2024.2.3-30
Juhász Valéria (2024): Az olvasástanulást támogató szintezett könyvek típusai. Az Aranyfa-sorozat szintezési jellemzőinek olvasásfejlesztési szempontú vizsgálata. In: Karlovitz János Tibor (szerk.): *Pedagógiai gondolkodásunk a konfliktusokkal terhelt világban.* pp. 232–252.
Klein-Braley, Ch. (1985): A cloze-up on the C-test. A study in the construct validation of authentic tests. *Language Teaching*, 2(1), 76–104. DOI: https://doi.org/10.1177/026553228500200108
Kojanitz László (2003a): Szakiskolai tankönyvek összehasonlító vizsgálata I. *Új Pedagógia Szemle*, 53(9), 14–24.
Kojanitz László (2003b): Szakiskolai tankönyvek összehasonlító vizsgálata II. *Új Pedagógia Szemle*, [lapszám nélkül], [oldalszám nélkül]
Kojanitz László (2004a): A pedagógiai szövegek analitikus vizsgálata. A szavak szintje. *Magyar Pedagógia*, 104(4), 29–442.

Kojanitz László (2004b): Lehet-e statisztikai eszközökkel mérni a tankönyvek minőségét?. *Iskolakultúra*, 14(9), 38–56.
Kóródi Bence (2022): *Olvasókönyv 4. Hétszín sorozat.* Budapest: Oktatási Hivatal.
Kövérné Nagyházi Bernadette (2003): A magyar nyelv szórendjének egy lehetséges tanítási modellje – kezdő szinten. *Hungarológiai évkönyv*, 4(1), 52–65.
Kroll, J. – Merves, S. (1986): Lexical access for concrete and abstract words. *Journal of Experimental Psychology: Learning, Memory and Cognition,* 12(1), 92–107. DOI: https://doi.org/10.1037//0278-7393.12.1.92
Látics Barbara – Gombos Péter (2025): Olvashatósági formulák és adaptálhatóságuk magyar nyelvre. *Magyar Nyelvőr*, 149(4), 486–502. DOI: https://doi.org/10.38143/nyr.2025.4.486
Legeza Márton (2020): Magyar nyelv 9. tankönyv. Budapest: Oktatási Hivatal.
Leroy, G. – Kauchak, D. (2014): The effect of word familiarity on actual and perceived text difficulty. *Journal of the American Medical Informatics Association*, 21(1), 169–172. DOI: 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*, 122(2), 65–88. DOI: https://doi.org/10.14232/mped.2022.2.65
Marulli, F. – Campanilea, L. – de Biasea, M. S. – Marronea, S. – Verdea, L. – Bifulco, M. (2024): Understanding Readability of Large Language Models Output. An Empirical Analysis. *Procedia Computer Science*, 246, 5273–5282. DOI: https://doi.org/10.1016/j.procs.2024.09.636
Meyer, B. J. F. (2003): Text Coherence and Readability. *Top Lang Disorders*, 23(3), 204–224. DOI: https://doi.org/10.1097/00011363-200307000-00007
Mező Péter Dániel (2023): Szöveganalízis és mesterséges intelligencia. Bevezetés a gépi tanulás és a mintakeresés által nyújtott lehetőségekbe. *Oipo: interdiszciplináris e-folyóirat,* 5(2), 67–72. DOI: https://doi.org/10.35405/oxipo.2023.2.67
Mikk, J. (1980): *Comprehension of text.* Tallin: Valgus.
Mikk, J. (1997): Parts of speech in predicting reading comprehension. *Journal of Quantitative Linguistics*, 4(1–3). 156–163. DOI: https://doi.org/10.1080/09296179708590091
Mikk, J. (2000): *Textbook. Research and writing*. Frankfurt am Main: Peter Lang.
Minnick, C. (2025): *Programozás MI-vel.* Budapest: Panem Könyvek.
Mizumoto, A. – Eguchi, M. (2023): Exploring the potential of using an AI language model for automated essay scoring. *Research Methods in Applied Linguistics*, 2(2), 1–13. DOI: https://doi.org/10.2139/ssrn.4373111
Montgomery, T. A. (2025): Analyzing and Modifying Reading Activities for Japanese EFL Students with Dyslexic Tendencies. *The Centre for the Study of English Language Teaching, 13, 1–12.*
Nagy József (2004): A szóolvasó készség fejlődésének kritérumorienált diagnosztikus feltérképezése. *Magyar Pedagógia*, 104(2), 123–142.
Nagyházi Bernadette (2011): Az egyszerű mondat szórendjének egy lehetséges tanítási modellje a magyar mint idegen nyelv oktatásában. [PhD-értekezés]. Pécs: Pécsi Tudományegyetem.
Naismith, B. – Mulcaire, P. – Burstein, J. (2023): Automated evaluation of written discourse coherence using GPT-4. *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023).* Toronto, Canada. 394–403. DOI: https://doi.org/10.18653/v1/2023.bea-1.32
Napolitano, D. – Sheehan, K. M. – Mundkowsky, R. (2015): Online Readability and Text Complexity Analysis with TextEvaluator. *Proceedings of the 2015 Conference of the*

North American Chapter of the Association for Computational Linguistics: Demonstrations. Princeton. DOI: https://doi.org/10.3115/v1/n15-3020_

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.

Russel, S. J. – Norvig, P. (2023): *Mesterséges intelligencia. Modern megközelítésben. I–II. kötet.* Budapest: Taramix.

Setiawan, H. – Fatichah, H. – Saikhu, A. (2023): Multilabel Classification of Student Feedback Data Using BERT and Machine Learning Methods. *14th International Conference on Information & Communication Technology and System (ICTS).* Surabaya. DOI: https://doi.org/10.1109/icts58770.2023.10330849

Simon Eszter – Indig Balázs – Kalivoda Ágnes – Mittelholcz Iván – Sass Bálint – Vadász Noémi (2020): Újabb fejlemények az e-magyar háza táján. *XVI. Magyar Számítógépes Nyelvészeti Konferencia.* Szeged. pp. 29–42.

Simon Gábor (2024): A megszemélyesítés regiszterspecifikus mintázatai magyar nyelvű online sajtószövegekben. *Jelentés és Nyelvhasználat*, 11(1), 101–121. DOI: https://doi.org/10.14232/jeny.2024.1.4

Smeuninx, N. – de Clerck, B. – Aerts, W. (2020): Measuring the Readability of Sustainability Reports. A Corpus-Based Analysis Through Standard Formulae and NLP. *International Journal of Business Communication*, 57(1), 52–85. DOI: https://doi.org/10.1177/2329488416675456

Strain, E. – Patterson, K. – Seidenberg, M. S. (1995): Semantic effects in single-word naming. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 21(5), 1140–1154. DOI: https://doi.org/10.1037//0278-7393.21.5.1140

Szabó Gábor (2019): A szövegnehézség vizsgálata az emelt szintű angol érettségi olvasáskomponensében. *Modern Nyelvoktatás*, 25(3-4), 102–119.

Száray Miklós (2021): *Történelem 10. a középiskolák számára.* Budapest: Oktatási Hivatal.

Szécsényi Tibor – Nagy C. Katalin – Németh T. Enikő (2024): Felszólításannotálás a MedCollect egészségügyi álhírkorpuszban. *XX. Magyar Számítógépes Nyelvészeti Konferencia.* Szeged. pp. 159–170.

Todirascu, A. – François, T. – Gala, N. – Fairon, C. – Ligozat, A-L. – Bernhard, D. (2013): Coherence and cohesion for the assessment of text readability. *Proceedings of 10th International Workshop on Natural Language Processing and Cognitive Science.* Marseille, France. pp. 11–19.

Uibo, H. (1995): Computer readability analysis of Estonian texts. In: Kraav, I. I. – Mikk, J. Vassiltchenko, L. (ed.): *Family and textbooks. Proceedings of the Department of Education*, pp. 96–115.

Váradi, T. – Simon, E. – Sass, B. – Mittelholcz, I. – Novák, A. – Indig, B. (2018): E-magyar – A digital language processing system. In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).* Japan, Miyazaki.

Viharos Zsolt János – Kis Krisztián Balázs – Fodor Ádám – Büki Máté István (2021): Adaptive, Hybrid Feature Selection (AHFS). *Pattern Recognition*, 116(107932), 1–13. DOI: https://doi.org/10.1016/j.patcog.2021.107932_

Vincze Veronika – Kicsi András – Főző Eszter – Vidács László (2021): A gépi elemzők kriminalisztikai szempontú felhasználásának lehetőségei. In: *XVII. Magyar Számítógépes Nyelvészeti Konferencia.* Szeged.

Zimányi Árpád (2025): Tankönyvi szövegek mondatszerkezeti sajátságainak vizsgálata. In: Takács Judit (szerk.): *Varietas incognita. Válogatás Zimányi Árpád tanulmányaiból*

70. születésnapja alkalmából. Eger: Eszterházy Károly Katolikus Egyetem Líceum Kiadó. pp. 49–56.
Zsibrita János – Vincze Veronika – Farkas Richárd (2013): Magyarlanc. A Toolkit for Morphological and Dependency Parsing of Hungarian. In: *Proceedings of RANLP 2013*. Hissar, Bulgaria. pp. 763–771.

SUMMARY

Factors of Readability in Text and Automated Text-Analysis Software

Hard to read content can hinder understanding and learning in educational environments, which can affect students' academic performance and reading comprehension. Traditional readability formulas are often based in surface-level characteristics of the text, such as average sentence and word length. They disregard further nuances in readability, which is why depending on them can be stifling and shallow (Marulli et al. 2024, Látics–Gombos 2025). This study identifies the major factors, which can determine the difficulty and readability of a text. Hungarian academic literature approaches the subject through student book-analysis, segmenting levels of language – morphology, semantics, syntax (Fóris 2002, Kojanitz 2004a, Domonkosi 2013). To make readability measurable, three methods can be employed (Lukács et al. 2022): Natural Language Processing based machine learning, which is more effective than readability formulas, but more potent still are neural web based deep learning models. Beyond more obvious aspects of text, in this study I present several modern text-analysis software as well. Understanding their operating principles can prove useful in the analysis of Hungarian texts through machine-based processes. They also provide a fuller picture of what categories the different input signals determining readability can be sorted into. There also exist Hungarian language processing tools, such as UDPipe, huspaCy, Magyarlánc, and especially e-magyar, whose latest, 2019 version (also known as emtsv) has already proven useful in practice. From the point of view of this study – to make readability a measurable metric – it can also prove to be an effective tool.

Keywords: readability, text factors, text-analysis, deep learning, computational linguistics