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

Az olvashatósági formulákon túl
Olvashatóságmérés gépi tanuló algoritmusok és mélytanulási modellek bevonásával [1]

Az automatikus olvashatóságértékelés (Automated Readability Assessment, ARA) az elmúlt évtizedekben jelentős változáson ment keresztül a természetesnyelv-feldolgozás és a mesterséges intelligencia fejlődésének köszönhetően. A hagyományos olvashatósági formulák mellett egyre inkább előtérbe kerültek a gépi tanulási, majd a mélytanulási modellek. Napjainkban pedig a két megközelítést ötvöző hibrid módszerek számítanak a legígéretesebbnek. A tanulmány célja a korszerű olvashatóságértékelési eljárások szelektív-narratív áttekintése, valamint annak vizsgálata, hogy a nemzetközi kutatások eredményei miként hasznosíthatók egy magyar nyelvre optimalizált olvashatóságértékelő modell fejlesztésében. Ismerteti továbbá a modern ARA-modellek főbb vonásait, a nyelvi jellemzők és az NLP-alapú eszközök szerepét, valamint kitér a magyar nyelv sajátosságaiból adódó lehetőségekre és kihívásokra. A hazai nyelvtechnológiai fejlesztések, a magyar nyelvre kidolgozott olvashatósági mutatók és a huBERT nyelvi modell együttesen megalapozhatják egy korszerű, magyar nyelvre optimalizált, automatikus olvashatóságértékelő modell fejlesztését.

Kulcsszavak: automatikus olvashatóságértékelés, olvashatóság, gépi tanulás, mélytanulás, mérőeszköz-fejlesztés, hibrid modellek

Bevezetés

Habár több mint száz éve léteznek olvashatósági formulák, az olvashatóság mérése máig nem megoldott probléma, sőt, kihívást jelentő kutatási terület (Jian et al. 2022). Hazánkban nem áll rendelkezésre olyan komplex, több tényezőt figyelembe vevő modell, amely a magyar nyelvű szövegek olvashatóságát (nehézségét) megbízhatóan képes lenne mérni. Pedig rendkívül indokolt volna egy ilyen eszköz. Közismert, hogy a szöveg – beleértve a tankönyvek, szövegértéstesztek és kötelező olvasmányok szövegeit – olvashatósága közvetlen hatással van a megértésre, az információfeldolgozás hatékonyságára, valamint a tanulás sikerességére. A szakirodalom egyértelműen rámutat arra, hogy az optimális olvashatóságú szövegek csökkentik a kognitív terhelést, támogatják a figyelem fenntartását és a tanulást, míg a túlzottan komplex szövegek gátolhatják a megértést (Bailin–Grafstein 2001; Szabó 2024; Bogdán 2025). A téma különösen releváns az oktatás kontextusában, ahol a tananyagok és olvasmányok megfelelő nehézségi szintje kulcsfontosságú. Az általános iskolás korosztály számára készült szövegek esetében a nem megfelelően megválasztott nehézség hosszú távon is befolyásolhatja a szövegértési készségek fejlődését (Józsa–Steklács 2012; Steklács et al. 2015).

A téma aktualitása szorosan összefügg a természetesnyelv-feldolgozás és a mesterséges intelligencia gyors fejlődésével is. Kezdetben a szakértők hozzáértésük alapján kategorizálták a szövegeket, egyszerű statisztikai jellemzők, például a mondatok hossza és a

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

szavak komplexitása alapján. Ezen megközelítések azonban túlságosan szubjektívek voltak, az értékelési eredmények a kritériumaiktól és céljaiktól függően változtak (Zhang 2024; Zheng et al. 2024). A 20. század elején a kézi értékelést fokozatosan leváltották az olvashatósági formulák, teret engedve az automatikus olvashatóságértékelésnek[2]. Az olvashatósági formuláknak – habár könnyen használhatók és objektívek – számos korlátjuk ismert (Látics–Gombos 2025; Zheng et al. 2024). Jelentőségük mégsem elhanyagolható, hiszen kikövezték az utat az ARA modernebb megközelítései felé (Zheng et al. 2024). A 2000-es évek elején jelentek meg a gépi tanuláson alapuló módszerek, melyek lehetőséget teremtettek a pontosabb, árnyaltabb értékelésekre, majd ezeket követték napjaink mélytanulási modelljei (Zhang 2024).

Az elmúlt években megjelent korszerű módszerek – különösen a gépi tanulási algoritmusok és mélytanulási modellek – lehetővé teszik a szövegek komplex, többdimenziós elemzését, valamint a nyelvi mintázatok automatikus felismerését (Dárdai et al. 2015; Lukács et al. 2022). Az automatizált szövegelemző szoftverek felhasználhatók a tanulók szintjéhez illeszkedő tartalmak kiválasztására. Egy megbízható olvashatóságmérő eszköz pedig hozzájárulhat a személyre szabott tanulási környezetek kialakításához is. E technológiák alkalmazása új távlatokat nyit az olvashatóság objektív mérésében, és lehetőséget teremt olyan magyarnyelv-specifikus modell létrehozására, amely képes a szövegnehézség megbízható becslésére.

Korábbi tanulmányokban, a mérőeszköz-fejlesztés érdekében már megtörtént az olvashatósági formulák bemutatása, előnyeik és hátrányaik számbavétele (Látics–Gombos 2025), illetve a magyar nyelvű szövegek nehézségét meghatározó szövegtényezők feltárása (Látics 2025). Tanulmányom célja az automatikus olvashatóságértékelés különböző megközelítéseinek – a gépi tanulást, a mélytanulást, illetve a hibrid modelleket alkalmazó olvashatóságértékelési módszer – bemutatása[3], bízva abban, hogy hazai szinten mindez hiánypótló lesz[4].

Mindazonáltal egy magyarnyelv-specifikus modell létrehozása szempontjából hasznos (és elengedhetetlen) a főbb módszertani irányok megismerése. A végső cél egy hasonló elve(ke)n működő mérőeszköz megalkotása – jelen tanulmány feladata a módszertani lehetőségek feltérképezése.

Írásomban elsőként az olvashatóságértékelés újfajta módszereit ismertetem: a természetes nyelvfeldolgozáson alapuló gépi tanuló algoritmusokat, valamint a neurális hálókra épülő mélytanulást. Mindkét megközelítésnek megvannak a maga előnyei-hátrányai, együttes alkalmazásuk pedig egy új (hibrid) módszert eredményez. Mivel az ARA területén végzett kutatások (modellfejlesztések) többsége a hibrid megközelítés mellett döntött, és hatékonyságát igazolni is tudta, bemutatása hangsúlyos lesz. Ismertetem, milyen adatkészletet igényel, valamint bemutatom a korpuszgyűjtés lehetőségeit – hiszen a legtöbb nyelv esetében (az angolt leszámítva) nem áll rendelkezésre címkézett adatkészlet. Egy hibrid modellnek továbbá szüksége van nyelvi jellemzőkre (feature-ökre), az ezekkel kapcsolatos tudnivalókat, lehetőségeket szintén taglalom. Ezt követően olyan kutatásokról lesz szó, amelyeket a közelmúltban végeztek hibrid olvashatóságértékelő modell létrehozása céljából – ezekből képet kaphatunk a mérőeszköz-fejlesztés (főbb) lépéseiről.

[2] Automated Readability Assessment (ARA)

[3] A teljesség igénye nélkül, főként az elmúlt öt év szakirodalmából válogatva.

[4] A témával foglalkozó csekély szakirodalom egyike Lukács és munkatársaié. 2022-es tanulmányukban írnak arról, hogy az olvashatóságmérés háromféle módszerrel közelíthető meg: (1) olvashatósági formulákkal, (2) NLP-alapú, gépi tanuló algoritmusokkal, (3) neurális hálókra épülő, mélytanulási algoritmusokkal, de csak felületesen ismertetik azokat, és a hibrid megoldásokra sem térnek ki.

Végezetül áttekintem, milyen lehetőségeink vannak, ha magyar olvashatóságértékelő eszközt szeretnénk fejleszteni.

Az olvashatóságértékelés modernebb megközelítései

A *természetes nyelvfeldolgozás (NLP)* és a *gépi tanulási technikák* gyors fejlődésének köszönhetően a kutatók elkezdték vizsgálni, hogy miként alkalmazhatók ezen – ma már hagyományosnak tekintett – módszerek a szövegértékelés területén. A jellemzők széles körének kivonásával és azok felhasználásával próbálták megragadni a szöveg olvashatóságát, nehézségét (Zhang 2024). A természetes nyelvfeldolgozás szövegbesorolási módszereket alkalmaz a szövegolvashatóság értékelésére – különböző olvashatósági szintekhez rendelve azokat. Ez a módszer gépi tanulási modelleket képez ki a szövegjellemzők vizsgálatára és a szint előrejelzésére (Ouassil et al. 2024). A gépi tanulási megközelítések modelleket építenek, amelyek több nyelvi szintet lefedő, manuálisan létrehozott jellemző alapján előre jelzik az emberek által megadott olvashatósági pontszámokat. A modellépítéshez viszonylag[5] nagy számú jellemző szükséges. Általában jobb eredményeket hoznak, mint a hagyományos olvashatósági képletek, viszont külső forrásokat igényelnek, például címkézett adatkészleteket – melyek ritkák. Ráadásul, akárcsak a formulák esetében, a különböző korpuszok és nyelvek közötti „átvihetőség" e módszer kapcsán is problémás: a kapott jellemzőkészlet kevésbé általánosítható más típusú vagy nyelvű szövegekre (Zhang 2024; Zheng et al. 2024).

A korszerűbb módszerek általában elvetik a manuálisan létrehozott nyelvi jellemzőket és *mélytanulási modelleket (neurális hálókat*) alkalmaznak. A legújabb eredmények szerint a mély neurális hálózatok beépítése a legkorszerűbb teljesítményt érte az olvashatóság értékelésében. Ezek a megközelítések teljesen figyelmen kívül hagyják a nyelvi jellemzőket: a releváns tényezők kivonását magára a modellre bízzák. Ennek észszerűségét indokolja a tény, miszerint elegendő adat és képzés mellett a modell megtanulhat minden olyan nyelvi jellemzőt, amelyet a kutatók megpróbáltak volna előre kiszámítani. Hozzá kell tenni azonban, hogy adatszegény kontextusban, ahol az adatgyűjtés problémás, az előre kiszámított nyelvi jellemzők lesznek célravezetők (Deutsch et al. 2020).

A mély neurális hálózatok kimagasló teljesítményt mutatnak minden olyan szemantikai feladatban, ahol elegendő mennyiségű adat áll rendelkezésre (Collobert et al. 2011; Martinc et al. 2021). Azonban előfordulhat, hogy nincs elegendő mennyiségű, címkézett (nehézségi szintekbe sorolt) dokumentum a nyelvi modell képzéséhez. Hátrányként említhető továbbá, hogy a neurális megközelítések általában csökkentik az előrejelzés értelmezhetőségét – ami különösen fontos lenne oktatási kontextusban. A felhasználók (kutatók, tanárok, tananyagfejlesztők stb.) számára szükséges lehet megérteni, mi alapján olvashatóbb az egyik szöveg a másikhoz képest (Martinc et al. 2021). Fontos kérdés a felhasznált szövegek terjedelme. Amennyiben a könnyebb szövegek rövidebbek, a nehezebbek hosszabbak – s nem egységes terjedelműek –, az negatív hatással lesz a modell teljesítményére (Imperial 2021; Li et al. 2022; Zheng et al. 2024). Egyes források (Li et al. 2022; Zheng et al. 2024) szerint a mély neurális hálózatok bevonása az olvashatóságértékelésbe nem kimondottan eredményesebb, mint a hagyományos gépi tanulási módszerek.

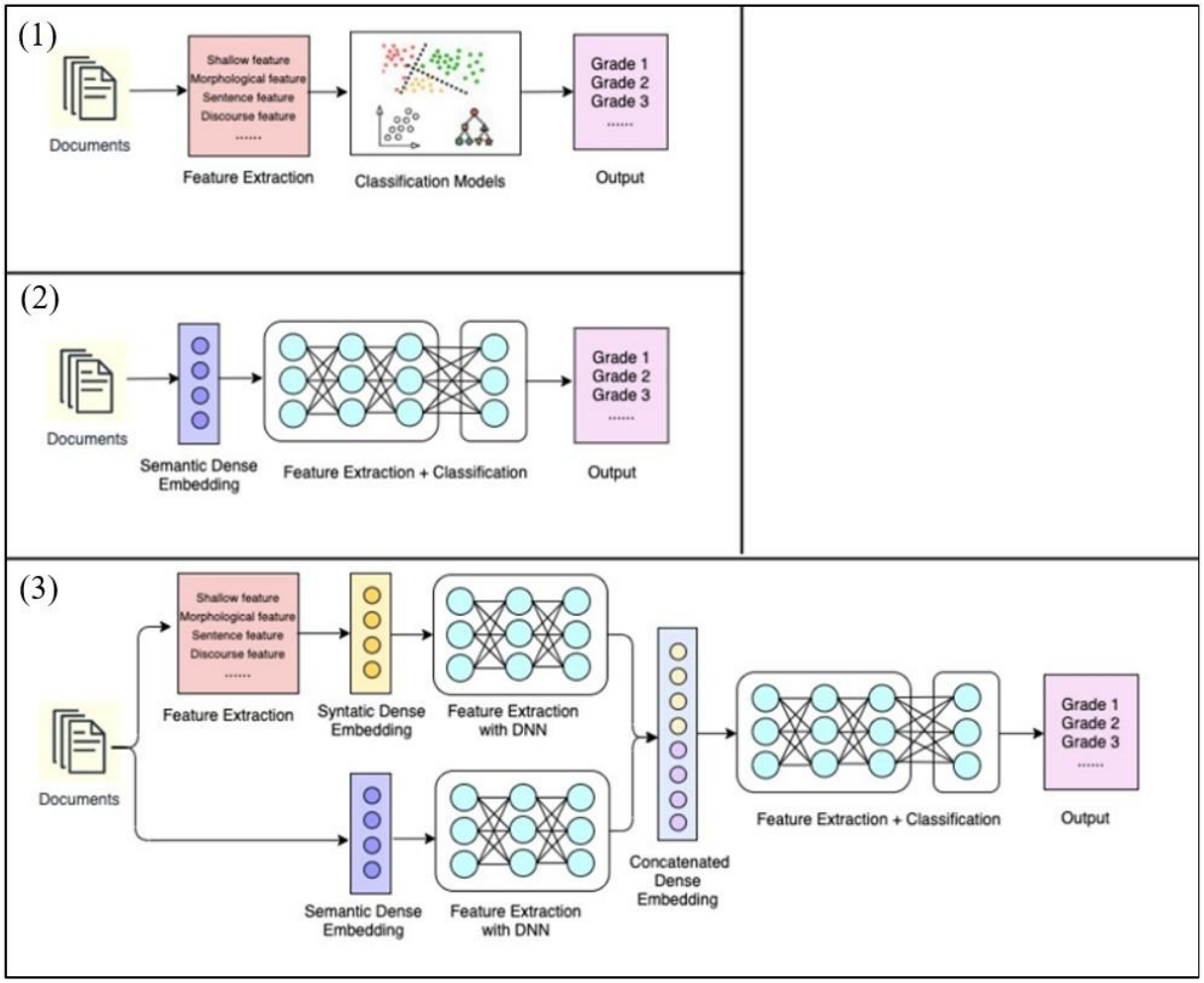
A fentiekkel ellentétben Liu et al. (2025) szerint a neurális hálózatok általában jobb teljesítményt érnek el, mint a hagyományos (NLP-alapú) osztályozók. A mélytanulás reprodukálhatóbb stratégiákhoz vezet, azaz könnyen átvihetők más területekre vagy nyelvekre (Azpiazu–Pera 2019). Azpiazu–Pera (2019) modellje, a Vec2Read kizárólag mélytanuláson alapul. Érdekessége, hogy nyolcféle nyelvre képezték ki, ilyen szempontból egyedülálló. Kiterjedt adatkészletét szintezett oktatási anyagok és közösségi forrásokból

[5] A mélytanulás esetén jóval több.

származó, egyszerűsített szövegek alkotják. A Vec2Read mélytanulási stratégiát alkalmaz, vagyis automatikusan felismeri az olvashatósággal kapcsolatos mintákat – nagy pontossággal. A mélytanulási módszerek egyre szélesebb körben használatosak az ARA területén (Zheng et al. 2024), s a 21. század új kutatási trendjeként utalnak rájuk (Jian et al. 2022).

A legújabb kutatások a *hibrid modellek* kiváló eredményeiről számolnak be. Az elmúlt években számos tanulmány (Deutsch et al. 2020; Imperial 2021; Jian et al. 2022; Li et al. 2022; Liu et al. 2025; Paraschiv et al. 2023; Ouassil et al. 2024; Qiu et al. 2021) ötvözte a gépi tanulást és a mélytanulást olvashatóságméréssel kapcsolatos feladatok végrehajtása során. Így tett Paraschiv és munkatársai (2023), akik 2–11. osztályos, orosz nyelvű tankönyveket (N = 154) osztályoztak gépi tanulási technikák és BERT-alapú modellek kombinálásával. Eredményeik arra utalnak, hogy a szövegkomplexitási indexek (olvashatósági mutatók) beépítése javította a modell teljesítményét. A hibrid modellekben a nyelvészeti jellemzők és a neurális modellek előrejelzései egyaránt fontos szerepet kapnak, ugyanis mélytanulás és gépi tanulás együttes alkalmazásával történik az értékelés.

Az automatikus olvashatóságértékelés tehát megvalósulhat (1) gépi tanulási algoritmusok bevonásával, (2) mélytanulással és (3) e kettő ötvözésével (hibrid megoldások). Működési elvüket szemlélteti az 1. ábra.



1. ábra: Automatikus olvashatóságértékelési módszerek működési elve
Forrás: Qiu et al. 2021 alapján.

(1) Az ábra első sora a hagyományos gépi tanulási megközelítést mutatja be. A bemenetet szöveges dokumentumok képezik, a szakirodalom adatkészletnek nevezi ezeket. A következő lépés a felszíni, morfológiai, mondatszintű és szövegszerkezeti jellemzők (feature-ök)

kinyerése NLP- és/vagy AI-alapú eszközök bevonásával. A program felismeri, kiszámolja, végül számérték formájában megadja őket (például átlagos mondathossz szavakban = 15). Ezt a számsort kapja meg egy olyan gépi tanuló algoritmus (osztályozó modell), amely korábban már sok-sok dokumentumon megtanulta, hogy mely jellemzők mely szinthez tartoznak. Kimenetelként így képes megadni a szöveg nehézségi szintjét.

(2) A mélytanulást alkalmazó ARA esetében nincs szükség arra, hogy a kutató előre megmondja, milyen jellemzőket kell keresni. Bemenetként itt is dokumentumokat kell megadni, amelyeket egy neurális nyelvi modell átalakít „szemantikai sűrű beágyazásokká" (Semantic Dense Embedding), ez azt jelenti, hogy a szövegek jelentését hosszú számsorokkal írja le. Ezeket a számsorokat kapja meg egy mély neurális hálózat, amely maga tanulja meg a mintázatokat, egyszerre végezve a jellemzők kinyerését és az osztályozást. A kimenetel itt is a szöveg nehézségi szintje lesz.

(3) A hibrid megoldások egyesítik az előző két megközelítést. A dokumentummal először ugyanaz történik, mint a hagyományos gépi tanulásnál: a kinyert jellemzőkből számsor készül, amelyet aztán egy neurális hálózat tovább finomít. Ezzel párhuzamosan, egy másik ágon a teljes dokumentum – egy nyelvi modell segítségével – szemantikai sűrű beágyazássá alakul. Ezt is feldolgozza egy neurális hálózat, akárcsak a másik ág esetében. A két reprezentációt összefűzik egyetlen sűrű beágyazássá, amely magába foglalja a szöveggel kapcsolatos nyelvtani információkat, felszíni jellemzőket, jelentést és szemantikai mintázatot. Ez kerül egy utolsó neurális hálózatba, amely – a jellemzők kinyerését és az osztályozást követően – meghozza a végső döntést: meghatározza a szöveg olvashatósági szintjét.

A bemutatott kutatások alapján nem állapítható meg egyértelműen, hogy a mélytanulási modellek minden esetben felülmúlják a gépi tanulási módszereket. A modellek teljesítményét jelentősen befolyásolja az adatkészlet mérete, a jellemzőkinyerés lehetőségei és a rendelkezésre álló szoftverek. Érdemes összehasonlítani a háromfajta módszert, ld. 1. táblázat.

1. táblázat: A modern ARA-módszerek összehasonlítása (saját szerkesztés)

	Gépi tanulás	Mélytanulás	Hibrid megközelítés
Bemenet	Szintezett szövegek + előre kinyert nyelvi jellemzők	Nyers szövegek	Nyers szövegek + előre kinyert nyelvi jellemzők
Jellemzők kinyerése	Manuálisan meghatározott jellemzők, kinyerés NLP-alapú eszközökkel	A mély neurális hálózat automatikusan tanulja meg	A két módszer kombinációja
Adatigény	Közepes	Nagy	Nagy, de rugalmasabb
Erősségek	Jól értelmezhető, kisebb adatkészlettel is működik	Nagy pontosság (elegendő adat esetén)	Egyesíti a két módszer előnyeit
Hátrányok	Kézi jellemzőtervezést, szoftverek bevonását igényli	Kevésbé értelmezhető, sok adatot igényel	Összetettebb felépítés
Leginkább ajánlott	Kevés címkézett adat esetén	Nagy adatkészlet esetén	Nagy adatkészlet esetén

Mérőeszköz-fejlesztőként először azt kell átgondolnunk, hogy melyik módszerhez adottak a hazai feltételek. Írásom utolsó fejezetében erre is kitérek majd. Mindenekelőtt azonban a hibrid megközelítést részletezem, tekintve, hogy számos szakirodalom emellett érvel, és esetünkben sem elvetendő.

A hibrid modellek adatkészlete és nyelvi jellemzői

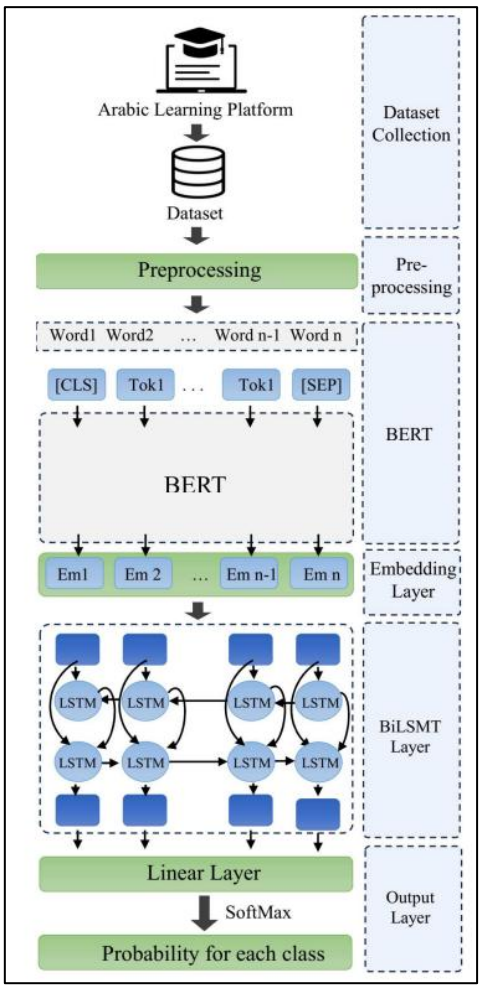
Számos friss kutatásban (Imperial 2021; Li et al. 2022; Paraschiv et al. 2023; Ouassil et al. 2024) előre betanított BERT[6] nyelvi modellt (Devlin et al. 2019) használtak az olvashatóság értékeléséhez. A BERT implicit módon képes kódolni a nyelvtani ismereteket (Imperial 2021). Felügyelet nélküli képzésre tervezték, s hatalmas, címkézetlen szövegkorpusz felhasználásával dolgozik (Ouassil et al. 2024). Közvetlenül a szöveges adatokból tanul, megragadva a bonyolult nyelvi mintákat és kapcsolatokat, lehetővé téve a finomabb, pontosabb és skálázhatóbb értékelést (Zhang 2024).

A hibrid modell a BERT által kivont nyelvi mintákat manuálisan meghatározott jellemzőkkel kombinálja. A jellemzőket szintezett, angol nyelvű szövegeket tartalmazó adatkészletekből (Wall Street Journal, WeeBit, Newsela, OneStopEnglish stb.) vonják ki, NLP-alapú szoftverekkel. Amennyiben nem áll rendelkezésre kész adatkészlet, ahogy a magyar nyelv esetében sem, a jellemzőkivonáshoz szükséges, címkézett korpuszok összeállítása nehéz, mivel idő és szakértelem szükséges hozzá (Deutsch et al. 2020).

Mindez az arab nyelv esetében is problémásnak bizonyult, ezért Ouassil és munkatársai (2024) szöveggyűjtésbe kezdtek. Olyan szövegeket gyűjtöttek, melyeket szerzőik hat nehézségi szintbe soroltak. A szinteket három – kezdő, középhaladó és haladó – olvashatósági kategóriába egyesítették. A korpuszgyűjtést követte az adatok előfeldolgozása és finomítása, annak érdekében, hogy az elemzésre alkalmas legyen. Az azt követő fázisban mélytanulási algoritmusokat (a BERT-et) alkalmazva fejlesztették ki a saját modelljüket a 2. ábrán látható módon.

[6] Bidirectional Encoder Representations from Transformer (BERT)

2. ábra: Olvashatóságmérés hibrid modell bevonásával – folyamatábra
Forrás: Ouassil et al. 2024.



Az angolhoz képest a kínai szövegek olvashatóságmérése ugyancsak korlátozott. Zheng és munkatársai (2024) kínai tankönyvek felhasználásával hozták létre a saját korpuszukat.

Li és munkatársai (2022) 200 darab kínai szöveget gyűjtöttek össze a tanterv részét nem képező olvasmányokból, ezeket három általános iskolai tanár sorolta három nehézségi szintbe. A gépi és az emberi értékelést összevetve az eredmények jelentős korrelációt mutattak.

Több forrás (Deutsch et al. 2020; Zhang 2024) is utalást tesz arra, hogy a manuálisan meghatározott nyelvi jellemzők nem elhanyagolhatók. Figyelembevételük átfogóbb képet nyújt az olvashatóságról, és növeli a modell teljesítményét. Amennyiben feltételezzük, hogy a nyelvi jellemzők és a mély neurális modellek kombinálása hatékonyságnövelő, szükséges feltárni a főbb olvashatósági mutatókat. Az ARA-megközelítések arra tesznek kísérletet, hogy megtalálják és értékeljék azokat a jellemzőket, amelyek jól korrelálnak az emberek által észlelt olvashatósággal (Martinc et al. 2021). Ezek a technikák – melyek általában felügyelt gépi tanuláson alapulnak – több száz (akár több ezer) jellemzőt is figyelembe vesznek, amelyek több szempontból írják le a szöveget. A jellemzők a nyelv függvényében változnak, ezért a kategóriák összehasonlítása nehézkes. Mindez megakadályozza a kutatókat abban, hogy összehasonlításokat végezzenek, valamint kialakítsanak egy konszenzust azzal kapcsolatban, hogy mely mutatók a legelőnyösebbek az olvashatóság szempontjából (De Clercq–Hoste 2016). Viszont a főbb kategóriák ismerete hozzájárulhat

egy magyarnyelv-specifikus lista összeállításához, illetve kész listák (Látics 2025) bővítéséhez, árnyalásához.

Martinc és munkatársai (2021) szerint az olvashatósági mutatók a következő csoportokba sorolhatók: hagyományos, diskurzus-kohéziós, lexikális-szemantikai és szintaktikai jellemzők. A hagyományos jellemzők közé tartoznak azok, amelyen a klasszikus olvashatósági formulák alapulnak (például átlagos szó- és mondathosszúság). A mondathosszúság jól korrelál a szöveg olvashatóságával, és javíthatja a modell teljesítményét. A diskurzus jellemzői a kohézió és a koherencia (Martinc et al. 2021).[7] Végül a lexikális-szemantikai jellemzők a szókincs nehézségi fokát, míg a szintaktikai jellemzők a szöveg grammatikai komplexitását mérik (Martinc et al. 2021). További kutatások a téma kapcsán:

– Zheng és munkatársai (2024) maguk gyűjtötték össze a kínai nyelv szempontjából releváns tényezőket – a szókincshez, a szófajhoz, a diskurzushoz és a kohézióhoz kapcsolódóan.

– Li és munkatársai (2022) szintén a kínai nyelvre szabottan állítottak össze saját feature-listát, amely karakter-, szó- és mondatjellemzőket tartalmaz.

– Liu és munkatársai (2025) jellemzői öt nyelvi kategóriába sorolhatók: a „jelentésszerkezetek bonyolultságára" utaló, kohézióval és koherenciával kapcsolatos, nyelvtani szerkezet bonyolultságát mérő, szavak és kifejezések nehézségét tükröző, illetve hagyományos (képletalapú) tényezők.

– Deutsch és munkatársai (2020) feature-listája a hagyományos (képletalapú), szintaktikai, szófajjal kapcsolatos és pszicholingvisztikai tényezőket foglalja magába.

– Több cikk (Deutsch et al. 2020; Imperial 2021) is utal Vajjala és Meurers (2016) munkájára, amely a jellemzőkivonás „receptjét" tartalmazza. Deutsch és munkatársai (2020) szerint Vajjala és Meurers állították össze a talán legszélesebb körű jellemzőket. Jellemzőkészletük négy csoportból – lexikai, szintaktikai, morfológiai és pszicholingvisztikai tényezőkből – tevődik össze.

– Zhang (2024) szintén felsorolja a munkája alapját képező nyelvi tényezőket, köztük „felületes" (képletalapú), szófaji, lexikai, szintaktikai és diskurzusjellemzőket.

A fenti források mindegyike fontos szerepet tulajdonít a lexikai-szintaktikai tényezőknek. Ez összhangban áll a hazai szakirodalom (Látics 2025) megállapításaival, amely szintén a szó- és a mondatszintű tényezők szerepét emeli ki. Érdemes megnézni, hogy milyen lexikai és szintaktikai szintű feature-öket említ a nemzetközi szakirodalom. Liu és munkatársainál (2025) a szókincs és a határozószók változatossága, továbbá a szófajok egymáshoz viszonyított arányai (melyek a mondatszerkezet összetettségét tükrözik) jelennek meg. Li és munkatársai (2022) mérvadó lexikai tényezőként említik a lexikai diverzitást, az átlagos szóhosszúságot, a szókincs nehézségét és a különböző szófajok arányát. A mondat szerkezetét leíró (szintaktikai) mutatói a mondathosszhoz[8] és a mondatszerkezet komplexitásához[9] kapcsolódnak. Zhang (2024) lexikai jellemzői a szavak változatosságával, sűrűségével[10] és gyakoriságával állnak kapcsolatban. Mondatszintű jellemzői a mondatok bonyolultságát[11] és a szövegben előforduló nyelvtani szerkezeteket (pl. összetett

[7] Todirascu és munkatársai (2016) 65 darab diskurzusjelölőt elemezve viszont arra a következtetésre jutottak, hogy ezek kevésbé járulnak hozzá az osztályozó prediktív erejéhez.

[8] Mondatok száma, átlagos mondathosszúság karakterekben és szavakban mérve, maximális karakterszám/mondat, maximális szószám/mondat.

[9] Tagmondatok száma, átlagos és maximális karakterszám tagmondatonként, átlagos és maximális szószám tagmondatonként.

[10] Tartalmas szavak (pl. főnevek, igék, melléknevek) aránya a szövegben.

[11] Szavak száma mondatonként.

mondatokat) mérik. Természetesen a szókinccsel, mondatszerkezettel, szórenddel stb. kapcsolatos jellemzők nyelvspecifikusak: a célnyelv sajátos nyelvi tulajdonságaihoz kell igazítani azokat (Zhang 2024).

Az ARA területén végzett korábbi munkák (Dell'Orletta et al. 2011; Gonzalez-Dios et al. 2014) rávilágítottak arra, hogy nem minden jellemző egyformán fontos az olvashatóság kapcsán: szelektálni szükséges azokat. Liu és munkatársai (2025) – miután listázták a számításba vehető tényezőket – statisztikai módszerekkel (ANOVA) határozták meg a legjelentősebbeket. Listájukat végül tíz nyelvi tényezőre redukálták, melyek leginkább korreláltak az adatkészlet olvashatóságával. Vajjala és Meurers (2016) szintén megvizsgálta, hogy mely jellemzők járulnak hozzá a legnagyobb mértékben az olvashatósághoz, akár önállóan, akár jellemzőcsoportok részeként – feature selectiont alkalmaztak. Példaként említhető Qiu és munkatársai (2021) kutatása is, akik miután kinyerték a szövegekből a nyelvi jellemzőket, kiszámították a köztük lévő korrelációt, feltételezve, hogy azok nem függetlenek egymástól, hanem együtt változnak. Például minél több a kötőszó, annál bonyolultabb a mondatszerkezet. Ezt követően csak a pozitív és erős (0,7 fölötti) kapcsolatokat tartották meg.

Hibrid olvashatóságértékelő modellek – módszertani irányok

A következőkben ismertetek néhány olyan kutatást a közelmúltból, melyek hibrid olvashatóságértékelő modell létrehozását tűzték ki célul. Ezen kutatások a számítógépes nyelvészet területéhez tartoznak, így értelmezésükhöz informatikai ismeretek szükségesek – mégis, egy magyarnyelv-specifikus modell megalkotása szempontjából hasznos (és elengedhetetlen) a főbb módszertani lépések megismerése.

Deutsch és munkatársai (2020) kutatásának központi kérdése: érdemes-e kombinálni a modern mélytanulási módszereket (például a BERT-et) a klasszikus olvashatósági mutatókkal? A rendelkezésre álló két legnagyobb, angol nyelvű korpuszon (WeeBit[12] és Newsela[13]) végezték el az értékelést. A módszerük első lépéseként kétfajta információt nyertek ki ugyanabból a szövegből. Egyrészt a hagyományos olvashatósági mutatókat – amelyekről korábbi kutatások már kimutatták, hogy kapcsolatban állnak az olvashatósággal –, másrészt a szöveget feldolgoztatták mélytanulási modellekkel is. A neurális háló (BERT) önállóan tanulta meg, milyen jellemzők utalnak az olvashatóságra. Ezt a véleményt nem fogadták el automatikusan, hanem új jellemzőként hozzáadták a többi nyelvi jellemzőhöz, és összefűzték azokat (feature fusion). Egy egyszerű gépi tanulási modell hozta meg a végső döntést, hogy az adott szöveg melyik nehézségi kategóriába tartozik. Eredményeik igazolták, hogy elegendő adat rendelkezésre állása esetén a nyelvi jellemzők alig vagy egyáltalán nem nyújtanak előnyt a független mélytanulási modellekhez képest. Kis adathalmaz esetén viszont érdemes megfontolni a nyelvi jellemzők beépítését, ugyanis javíthatják a mélytanulási modellek teljesítményét.

[12] 7–16 éves tanulóknak szóló dokumentumokat tartalmaz, Vajjala és Meurers (2012) összeállítása. A teljes mintát lecsökkentve Deutsch és munkatársai osztályonként 625 darab szöveget vontak be a kutatásba.

[13] A Newsela korpusz (Xia et al. 2016) 1911 hírcikkből áll. Mindegyik szöveghez négyféle nehézségi szint tartozik. A WeeBithez hasonlóan a Newsela is 7–16 éveseknek címkézett.

Li és munkatársai (2022) egy BERT-alapú olvashatóságértékelő modellt javasolnak. Három angol nyelvű – WeeBit, OneStopEnglish[14], Cambridge[15] – és egy tankönyvi szövegeket tartalmazó adatbázis – ChineseLR[16] – bevonásával dolgoztak. A hatékonyság növelése érdekében többféle információt kombináltak: a BERT által kivont nyelvi jellemzőket vették alapul, kiegészítve a hagyományos olvashatósági mutatókkal. Modelljük nemcsak a szöveg nehézségét, hanem annak témáját is figyelembe vette. Ennek során a modell megtanulta, hogy mely témák vagy szóhasználatok jellemzőbbek a könnyebb, illetve a nehezebb szövegekre. Ezt követően – az ismétlődő feature-ök kiszűrése érdekében – egy Feature Projection (vetítés) nevű eljárást alkalmaztak, amely kiválasztja azokat a nyelvi jellemzőket, amelyek valóban új információt hordoznak a BERT számára. A létrehozott modelljük három adatkészleten a legkorszerűbb eredményeket érte el, sőt az egyik angol adatkészleten 99%-os előrejelzést adott a szöveg olvashatóságáról.

Imperial (2021) számára az egyik legfontosabb megoldandó probléma az volt, hogy ARA-modellje egy alacsony erőforrású nyelv (filippínó) esetében is hatékony legyen. A kutató a BERT által tanult tudást kapcsolta össze a nyelvi jellemzőkkel, majd egy gépi tanulási algoritmus döntött a szövegek nehézségi szintjéről. Angol és filippínó adatkészletet (OneStopEnglish, Common Core Exemplars[17], Adarna House[18]) egyaránt használt a modellépítéshez. Az angol nyelvű korpuszból nagyjából 150 nyelvi (lexikai, szintaktikai, pszicholingvisztikai stb.) jellemzőt vont ki, a filippínó korpuszból 54-et. Elsőként a hagyományos nyelvi jellemzőket gyűjtötte össze, ezután a teljes szöveget átadta a BERT-nek, amely összefoglalta a szövegek jelentését és nyelvi szerkezetét (embedding). A kétfajta információ összefűzésével kapott egy közös jellemzőhalmazt, amit továbbított egy gépi tanulási algoritmusnak. Ez az algoritmus döntötte el, hogy a szöveg melyik olvashatósági szintbe sorolható. Hibrid modellje mind az angol, mind a filippínó adatkészleteken jobban teljesített, mint a kizárólag nyelvi tényezőket bevonó modellek (2,63%-os teljesítménynövekedés).

Zhang (2024) célja egy olyan modell létrehozása, amely képes automatikusan és pontosan meghatározni a kínainyelv-tanulóknak szóló szövegek olvasási nehézségét. A kutató többlépcsős módszertant alkalmazott, melynek első lépéseként különböző nehézségű, kínai szöveget (N = 18 430) gyűjtött össze nyelvkönyvekből. A szövegeket hét különböző nehézségi szintbe sorolta, majd mondatokra és szavakra bontotta azokat a könnyebb elemezhetőség érdekében. Ennek a kutatásnak is az alapját a BERT adta: először kínai szövegkorpuszon betanította, hogy megértse a nyelv alapvető sajátosságait, ezt követően a nehézségi szintekkel ellátott adathalmazon finomhangolta, hogy rájöjjön a nehézségi szintek közötti különbségekre. Annak érdekében, hogy a modell még pontosabb legyen, további nyelvészeti jellemzőkkel egészítette ki. A BERT modell által tanultakat és a nyelvészeti jellemzőket egy neurális hálózatban egyesítette. A végleges modellt 950 új, korábban nem látott szövegen próbálta ki – igen hatékony (92,9%-os) teljesítményt nyújtott.

Bár a szakirodalom többsége a hibrid modellek mellett érvel, fontos megjegyezni, hogy fejlesztésük lényegesen nagyobb módszertani felkészültséget és több erőforrást igényel.

[14] Vajjala és Lučić (2018) korpusza 189 szövegből áll, amelyek mindegyike három változatban található meg.

[15] CEFR-szerinti besorolású, angolnyelv-tanulókra szabott szövegadatbázis, melyet Xia és munkatársai (2019) nyelvvizsgaszövegekből állítottak össze. Li és munkatársai (2022) osztályonként 60 darab szöveget használtak fel belőle.

[16] Li és munkatársai (2022) gyűjtése. Egy kínai adatbázis, amelyet középiskolai és általános iskolai tankönyvekből gyűjtöttek össze, majd soroltak öt nehézségi szintbe.

[17] 168 darab, 2–12. osztályos angolnyelv-tanulóknak szóló szöveget tartalmaz, három nehézségi szintbe sorolva.

[18] Egy Fülöp-szigeteki (filippínó) gyermekkönyvkiadó 265 mesekönyvének gyűjteménye, 1–3. osztályosok számára.

Egy magyar modell fejlesztésének koncepciója

A fentiek ismeretében kifejtem, hogy milyen lehetőségeink vannak, ha magyar olvashatóságértékelő eszközt szeretnénk fejleszteni. Ismert-e címkézett szövegeket tartalmazó adatbázis, hogyan alakul a magyar feature-lista, rendelkezésre állnak-e NLP-alapú eszközök a jellemzőkivonáshoz, bevonható-e a mélytanulás, és létezik-e magyarnyelv-specifikus BERT? A kérdések sora hosszú, megválaszolásuk azonban annál fontosabb.

Először az egyszerűbb módszernek számító gépi tanulás felől érdemes elindulni. A korábbiakból kiderült, hogy szükség lesz viszonylag sok címkézett szövegre, valamint manuálisan meghatározott, majd NLP-alapú eszközök segítségével számszerűsített feature-ökre.

Alacsony erőforrású nyelv lévén nem áll rendelkezésre olyan adatbázis, amely magyar nyelvű, szintezett szövegeket tartalmazna. Amennyiben továbbra is kitartunk amellett, hogy az ARA-t gépi tanulási és/vagy mélytanuláson alapuló módszerekkel szeretnénk megközelíteni, szövegeket kell gyűjteni. Ez esetben követhetnénk akár Zheng és munkatársai (2024) példáját is, akik tankönyvek felhasználásával állították össze korpuszukat, viszont mérőeszközünkkel a tankönyvek olvashatóságát is meg szeretnénk majd vizsgálni, így ez a módszer nem lenne célravezető. Dönthetünk úgy is, ahogy Li és munkatársai (2022), akik tanterven kívüli olvasmányokat gyűjtöttek, majd tanárokkal (szakértőkkel) végeztették el a szintezést. A korpusz sokszínűségét azonban jobban garantálja, ha maguk a szakértők végzik a gyűjtést és az osztályozást is, s a szövegeket már eleve szintezve továbbítják a kutatóknak – emellett döntöttek például Ouassil és munkatársai (2024). Az adatkészlet nagysága attól függ, melyik módszert alkalmazzuk. A gépi tanulási algoritmusnak jóval kevesebb szövegre van szüksége, mint a mély neurális hálóknak, ahogy az a korábbiakból kiderült.

A feature-öket illetően Deutsch és munkatársai (2020) véleményét érdemes figyelembe venni, miszerint adatszegény kontextusban az előre kiszámított nyelvi jellemzők lesznek célravezetők. Immár ismertek azok a mutatók, melyek az olvashatósággal kapcsolatban állnak (Látics 2025). A különböző kutatások eltérő feature-készleteket alkalmaznak, ugyanakkor jól látható, hogy bizonyos jellemzők – különösen szó- és mondatszinten – szinte minden modellben megjelennek. Ezek univerzálisabb szerepet tölthetnek be az olvashatóság becslésében, míg más jellemzők inkább nyelvspecifikusnak tekinthetők. A magyar olvashatósági mutatók árnyalásához, bővítéséhez tehát ötletet adhatnak a nemzetközi kutatások.

Azt is tudjuk, hogy léteznek magyarnyelv-specifikus, NLP-alapú szoftverek, melyek lehetővé teszik a szöveganalízis-vizsgálatokat. Ilyen például az e-magyar, a Magyarlánc és a HuSpaCy (Látics 2025). Sőt, több olyan hazai kutatás is ismert, amely ezek bevonásával dolgozott fel nagyméretű szövegkorpuszokat.

Horváth Péter (2020) célja az volt, hogy feltárja a József Attila-versek hangzásbeli jellemzőit NLP-alapú eszközök bevonásával. A kutató saját verselemző programot fejlesztett, ezt kombinálta az MTA Nyelvtudományi Intézete *e-magyar* nevű, nyelvfeldolgozó rendszerével. Első lépésként az e-magyar automatikusan elvégezte a versek nyelvi elemzését. Ezt követően a verselemző program – a nyelvi annotációkra építve – automatikusan meghatározta többek között a versszakok rímképletét, a versek ritmusát, a rímpárokat és az alliterációkat. Az NLP-eszközök lehetővé tették – ami esetünkben is hasznos lehet – a szövegek szavakra bontását (tokenizálás), a szótári alakok felismerését (lemmatizálás), valamint a nyelvtani kategóriák meghatározását (szófaji címkézés). Ennek eredményeképp ismertté vált a szókészlet mérete és a leggyakoribb lemmák, szófajonkénti bontásban is.

Szécsényi és munkatársai (2024) egészségügyi álhírek automatikus felismerésére használták fel az NLP-alapú eszközöket (konkrétan az e-magyart). A szerzők hipotézise szerint az álhírek nemcsak tartalmukban, hanem nyelvhasználatukban is eltérnek a hiteles hírektől, például jóval több direkt felszólítást tartalmaznak. A tanulmány ezek vizsgálatára összpontosít. Szécsényiék maguk hozták létre ál- és valós híreket is tartalmazó korpuszukat. Minden szöveget az e-magyarral dolgoz(tat)tak fel, ami automatikusan elkészítette a szövegek nyelvi elemzését, például megjelölte a felszólító módú igéket. A gépi elemzést kézi annotáció egészítette ki. A kvantitatív elemzés igazolta a kutatók hipotézisét, miszerint az álhírekben lényegesen több felszólító igealak fordul elő. Ez a kutatás is igazolta, hogy az e-magyar automatikus nyelvi elemzése jelentősen megkönnyíti a korpusz feldolgozását.

Vincze Veronika és munkatársai (2021) szerzőazonosításra használták az NLP-eszközöket. Hipotézisük szerint minden emberre, így a fenyegető, zsaroló vagy rágalmazó üzenetek szerzőjére is sajátos nyelvhasználat jellemző, amely megjelenik a szövegeiben is. E nyelvi sajátosság automatikus felismerése támogathatja a kriminalisztikai szakemberek munkáját. A mi kutatásunk szempontjából azért lehet érdekes Vinczéék munkája, mert nyelvi jellemzőket elemeztek (ehhez a *Magyarlánc*ot használták). A vizsgált lexikai jellemzők között említik a szókincsgazdagságot, a szó- és szófajgyakoriságot, a tartalmas szavak arányát (és egyebeket is, melyek esetünkben kevésbé mérvadók). Az automatikusan kinyert szintaktikai jellemzők közé tartoznak: alá- és mellérendelő szerkezetek aránya, tagmondatok száma, egyszerű és összetett mondatok aránya, az átlagos mondathossz stb. A kutatás igazolta, hogy e jellemzők hatékonyan hozzájárulnak a szerzők azonosításához. Az is kiderült a cikkből – ami igencsak biztató –, hogy az olvashatóságot meghatározó szövegtényezők (Látics 2025) számszerűsíthetők.

A mérőeszköz-fejlesztést megelőzően kell határozni a célpopulációt és az olvashatósági skálát, továbbá alapos átgondolást igényel a magyar korpusz műfaji és tematikai összetétele, a szakértői szintezés kritériumai, valamint az értékelők közötti megbízhatóság garantálása.

A fentiek alapján úgy tűnik, a gépi tanulás egy járható út. Ennél még hatékonyabb lenne mélytanulással ötvözni azt. A nemzetközi szakirodalom mélytanuláson (is) alapuló modelljei a BERT bevonásával dolgoztak, ami viszont nyelvspecifikus. Amennyiben mi is hasonlóképp tennénk, szükségünk lesz egy magyar nyelvre optimalizált BERT modellre. Nemeskey Dávid Márk 2020-ban fejlesztette ki a huBERT-et, melyet kifejezetten magyar szövegeken tanítottak be, a Webcorpus 2.0 és a magyar Wikipédia felhasználásával (Nemeskey 2020). A huBERT megfelelő módszertani alapot jelenthet munkánk során[19].

Záró gondolatok

Az automatikus olvashatóságértékelés az elmúlt két évtizedben jelentős fejlődésen ment keresztül. A kutatók felismerték a hagyományos olvashatósági formulák korlátait, s fokozatosan a természetesnyelv-feldolgozáson alapuló, gépi tanulási algoritmusok, majd a mélytanulási modellek felé fordultak. Ugyanakkor a legfrissebb nemzetközi eredmények arra utalnak, hogy a két megközelítés egymást akár ki is egészítheti – igen hatékonyan.

A szakirodalom áttekintése arra is rámutatott, hogy az ARA-modellek pontosságát nagyban meghatározza az adatkészlet minősége, a vizsgált nyelv sajátosságaihoz igazított nyelvi jellemzők, valamint azok az NLP-alapú eszközök, amelyek lehetővé teszik e jellemzők automatikus kinyerését.

Nincs könnyű helyzetben az a kutató, aki mérhetővé szeretné tenni a magyar nyelvű szövegek olvashatóságát. Nyelvünk agglutináló jellege, gazdag morfológiája és viszonylag

[19] A huBERT nyilvánosan elérhető: https://hlt.bme.hu/en/resources/hubert

korlátozott NLP-erőforrásai miatt a nemzetközi modellek csupán iránymutatást adhatnak, magyar példa pedig egyáltalán nincs is. Ugyanakkor már ismertek azok a (főbb) szövegtényezők, amelyek az olvashatósággal kapcsolatba hozhatók. Az is egyértelművé vált, hogy a lexikai-szintaktikai tényezők szinte valamennyi modellben meghatározó szerepet kapnak. Szintén biztató, hogy rendelkezésre állnak magyar nyelvre fejlesztett, nyelvtechnológiai eszközök, amelyek alkalmasak a jellemzők automatikus kinyerésére. A huBERT pedig arra teremt lehetőséget, hogy a magyar nyelv esetében is bevonhatók legyenek a legkorszerűbbnek számító mélytanulási megoldások.

Amennyiben a cél egy hibrid – vagy legalábbis gépi tanuláson alapuló – olvashatóságértékelő modell fejlesztése, az első feladat egy megfelelően kialakított (szakértők által szintezett) szövegkorpusz létrehozása, valamint a magyar nyelv sajátosságait tükröző feature-készlet véglegesítése. Jelen tanulmány nem egy szisztematikus szakirodalmi áttekintés, s nem is egy konkrét olvashatóságértékelő modell bemutatását tűzte ki célul. Célja azoknak az elméleti és módszertani irányoknak a szelektív-narratív áttekintése, amelyekre egy későbbi, magyar fejlesztés épülhet. Az ehhez szükséges módszertani feltételek pedig egyre inkább körvonalazódnak. Már nem az a kérdés, hogy lehet-e magyar modellt készíteni, hanem az, hogy milyen módszertannal érdemes megvalósítani.

Irodalom

Azpiazu, I. M., & Pera, M. S. (2019): Multiattentive Recurrent Neural Network Architecture for Multilingual Readability Assessment. Transactions of the Association for Computational Linguistics, 7, 421–436. https://doi.org/10.1162/tacl_a_00278

Bailin, A., & Grafstein, A. (2001): The linguistic assumptions underlying readability formulae. A critique. Language & Communication, 21(3), 285–301. https://doi.org/10.1016/S0271-5309(01)00005-2

Bogdán, P. (2025): Az olvashatóság mint a szövegértés sikerességét befolyásoló tényező. In Zs. Prohászka & A. Huszics (szerk.), Találkozások az anyanyelvi nevelésben 5.: Régi témák, új nézőpontok a nyelvészetben és a nyelvi nevelésben (pp. 19–25). PTE BTK Nyelvtudományi Tanszék, Lingua Franca Csoport. https://doi.org/10.15170/talalkozasok5-btk-2025-02

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., & Kuksa, P. (2011): Natural Language Processing (almost) From Scratch. Journal of Machine Learning Research, 12, 2493–2537. https://doi.org/10.48550/arXiv.1103.0398

Dárdai, Á., Dévényi, A., Márhoffer, N., & Molnár-Kovács, Zs. (2015): Tankönyvkutatás, tankönyvfejlesztés külföldön II. A nemzetközi tankönyvelméleti kutatások elméleti és módszertani kérdései. Történelemtanítás: online történelemdidaktikai folyóirat, 6(1–2).

De Clercq, O., & Hoste, V. (2016): All mixed up? Finding the optimal feature set for general readability prediction and its application to English and Dutch. Computational Linguistics, 42(3), 457–490. https://doi.org/10.1162/COLI_a_00255

Dell'Orletta, F., Montemagni, S., & Venturi, G. (2011): Read-it. Assessing readability of Italian texts with a view to text simplification. In Workshop on Speech and Language Processing for Assistive Technologies. Edinburgh, Scotland, UK.

Deutsch, T., Jasbi, M., & Shieber, S. (2020): Linguistic Features for Readability Assessment. In Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications. Seattle, WA, USA. https://doi.org/10.48550/arXiv.2006.00377

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019): BERT. Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis, Minnesota. https://doi.org/10.48550/arXiv.1810.04805

Gonzalez-Dios, I., Aranzabe, M. J., de Ilarraza, A. D., & Salaberri, H. (2014): Simple or complex? Assessing the readability of Basque texts. In Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics. Dublin, Ireland.

Horváth, P. (2020): A vershangzás jellemzőinek automatikus feltárása József Attila verseiben. Digitális Bölcsészet, 3(1), 3–27. https://doi.org/10.31400/dh-hun.2020.3.422

Imperial, J. M. (2021): BERT Embeddings for Automatic Readability Assessment. In Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021). Shoumen, Bulgaria.

Jian, L., Xiang, H., & Le, G. (2022): English Text Readability Measurement Based on Convolutional Neural Network. A Hybrid Network Model. Computational Intelligence and Neuroscience, 2022(4), 1–9. https://doi.org/10.1155/2022/6984586

Józsa, K., & Steklács, J. (2012): Az olvasás tanításának tartalmi és tantervi szempontjai. In B. Csapó & V. Csépe (szerk.), Tartalmi keretek az olvasás diagnosztikus értékeléséhez (pp. 137–188). Nemzeti Tankönyvkiadó.

Látics, B., & Gombos, P. (2025): Olvashatósági formulák és adaptálhatóságuk magyar nyelvre. Magyar Nyelvőr, 149(4), 486–502. https://doi.org/10.38143/Nyr.2025.4.486

Látics, B. (2025): Olvashatóságot meghatározó szövegtényezők és automatizált szövegelemző szoftverek. Anyanyelvi Kultúraközvetítés, 8(2), 4–23. https://doi.org/10.33569/akk.7197

Li, W., Ziyang, W., & Wu, Y. (2022): A Unified Neural Network Model for Readability Assessment with Feature Projection and Length-Balanced Loss. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Abu Dhabi. https://doi.org/10.18653/v1/2022.emnlp-main.504

Liu, F., Jin, T., & Lee, J. S. Y. (2025): Automatic Readability Assessment for Sentences: Neural, Hybrid and Large Language Models. Language Resources and Evaluation, 59(3), 2265–2296. https://doi.org/10.1007/s10579-024-09800-5

Lukács, Á., Rácz, P., & Kas, B. (2022): Tankönyvi szövegek nyelvi feldolgozhatóságának mutatói és vizsgálati módszerei. Magyar Pedagógia, 122(2), 65–88. https://doi.org/10.14232/mped.2022.2.65

Martinc, M., Pollak, S., & Robnik-Šikonja, M. (2021): Supervised and Unsupervised Neural Approaches to Text Readability. Computational Linguistics, 47(1), 141–179. https://doi.org/10.48550/arXiv.1907.11779

Nemeskey, D. M. (2020): Natural Language Processing Methods for Language Modeling [PhD-értekezés]. Eötvös Loránd Tudományegyetem Informatikai Kar. https://doi.org/10.15476/elte.2020.066

Ouassil, M. A., Jebbari, M., Rachidi, R., Errami, M., Cherradi, B., & Raihani, A. (2024): Enhancing Arabic Text Readability Assessment. A Combined BERT and BiLSTM Approach. In 2024 International Conference on Circuit, Systems and Communication (ICCSC). Fez, Morocco. https://doi.org/10.1109/ICCSC62074.2024.10616953

Paraschiv, A., Dascalu, M., & Solnyshkina, M. (2023): Classification of Russian textbooks by grade level and topic using ReaderBench. Research Result. Theoretical and Applied Linguistics, 9(1), 50–63. https://doi.org/10.18413/2313-8912-2023-9-1-0-4

Qiu, X., Chen, Y., Chen, H., Nie, J.-Y., Shen, Y., & Lu, D. (2021): Learning Syntactic Dense Embedding with Correlation Graph for Automatic Readability Assessment. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. (Online). https://doi.org/10.18653/v1/2021.acl-long.235

Steklács, J., Molnár, G., & Csapó, B. (2015): Az olvasás-szövegértés online diagnosztikus mérések tartalmi kereteinek elméleti háttere. In B. Csapó, J. Steklács & G. Molnár (szerk.), Az olvasás-szövegértés online diagnosztikus értékelésének tartalmi keretei (pp. 15–31). Oktatáskutató és Fejlesztő Intézet (OFI).

Szabó, Cs. (2024): Olvashatósági index alkalmazása angol nyelvű szövegeknél. A magyar nyelvi adaptáció kérdésköre [Szakdolgozat]. Magyar Agrár- és Élettudományi Egyetem.

Szécsényi, T., Nagy, C. K., & Németh, T. E. (2024): Felszólításannotálás a MedCollect egészségügyi álhírkorpuszban. In XX. Magyar Számítógépes Nyelvészeti Konferencia. Szeged.

Todirascu, A., Francois, T., Bernhard, D., Gala, N., & Ligozat, A.-L. (2016): Are cohesive features relevant for text readability evaluation? In Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics. Osaka, Japan.

Vajjala, S., & Lučić, I. (2018): OneStopEnglish corpus: A new corpus for automatic readability assessment and text simplification. In Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications. New Orleans, Louisiana. https://doi.org/10.18653/v1/w18-0535

Vajjala, S., & Meurers, D. (2012): On Improving the Accuracy of Readability Classification using Insights from Second Language Acquisition. In Proceedings of the Seventh Workshop on Building Educational Applications Using NLP. Montreal, Canada.

Vajjala, S., & Meurers, D. (2016): Readability-based Sentence Ranking for Evaluating Text Simplification. Computation and Language. https://doi.org/10.48550/arXiv.1603.06009

Vincze, V., Kicsi, A., Főző, E., & Vidács, L. (2021): A gépi elemzők kriminalisztikai szempontú felhasználásának lehetőségei. In XVII. Magyar Számítógépes Nyelvészeti Konferencia. Szeged.

Xia, M., Kochmar, E., & Briscoe, T. (2016): Text Readability Assessment for Second Language Learners. In Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications. San Diego, CA. https://doi.org/10.18653/v1/w16-0502

Zhang, C. (2024): Readability Grading Based on Multidimensional Linguistics Features for International Chinese Language Education. IEEE Access, 12, 88608–88619. https://doi.org/10.1109/ACCESS.2024.3418844

Zheng, Y., Chen, Y., & Zhang, S. (2024): Hierarchical Ranking Neural Network for Long Document Readability Assessment. Available at SSRN: https://ssrn.com/abstract=4878472 or http://dx.doi.org/10.2139/ssrn.4878472

SUMMARY

Beyond Readability Formulas – Readability Assessment Involving Machine Learning Algorithms and Deep Learning Models

Automated Readability Assessment (ARA) has undergone significant changes in recent decades due to the development of natural language processing and artificial intelligence. In addition to traditional readability formulas, machine learning and then deep learning models have increasingly come to the fore. Nowadays, hybrid methods combining the two

approaches are considered the most promising. The aim of the study is to provide a selective-narrative overview of modern readability assessment procedures and to examine how the results of international research can be utilized in the development of a readability assessment model optimized for the Hungarian language. It also describes the main features of modern ARA models, the role of linguistic features and NLP-based tools, and discusses the opportunities and challenges arising from the specificities of the Hungarian language. Domestic language technology developments, readability indicators developed for the Hungarian language, and the huBERT language model can jointly lay the foundation for the development of a modern, automatic readability assessment model optimized for the Hungarian language.

Keywords: automatic readability assessment, readability, machine learning, deep learning, measurement tool development, hybrid models

A műre a Creative Commons 4.0 standard licenc alábbi típusa vonatkozik: ***CC-BY-NC-ND-4.0***

