

Kísérletek orosz szövegek nehézségi szintjének automatizált meghatározására

Pisch Bendegúz

PTE BTK

OTD 2025

A lezárási dátum: 2025. január 15.

Tartalomjegyzék

Bevezetés	2
1. A szövegek nehézségi szintjének mérése	3
2. Az elemzéshez szükséges szövegek összeválogatása	6
3. Lehetőségek a szövegek komplexitásának mérésére	8
3.1. Szövegek mérése a szótagszám és a mondatok átlagos hossza alapján	10
3.2. Szövegek mérése a szavak átlagos hossza és a mondatok átlagos hossza alapján	11
3.3. Szóalakok gyakoriságát használó képletek.....	12
4. Az új képletek, és a képletek által adott eredmények bemutatása.....	14
5. A tanítási szöveggyűjtemény bővítése	18
6. További potenciális szempontok.....	20
7. A scriptek használata	22
Összegzés.....	23
Bibliográfia.....	24

Bevezetés

Egy idegen nyelv tanulásakor az olvasásnak kiemelten fontos szerep jut a nyelvtudás fejlesztésében. Az idegen nyelvet tanulók számára azonban nagy kihívást jelenthet a nyelvtudásuknak megfelelő szövegek megtalálása. Egy túl könnyű szöveg olvasása kevésbé mozdíttja elő a tanuló nyelvtudásának fejlődését, és bár talán úgy érzi, hogy jól megérti a célnyelvet, ez csak az adott szöveg lexikájának és grammatikai szerkezeteinek szintjére igaz. A túl sok nehéz szót és komplex mondatstruktúrát tartalmazó olvasmány túlságosan nagy kihívást jelenthet a szöveg megértésekor, a sok ismeretlen szó miatt előforduló gyakori szótározás pedig kedvezőtlenül hat az olvasás élményére és a tanulás hatékonyságára. Emiatt mindkét esetben csökkenhet a tanuló motivációja, és további nehézségekkel is találkozhat az idegen nyelv elsajátítása során. Napjainkban egyre több, a felhasznált szókincset és nyelvtani szerkezeteket tekintve korlátozott, adaptált könyv jelenik meg. Ezek általában egyszerűbb szerkezetű, rövid történetek, de akár ismertebb irodalmi műveknek is lehet adaptált változatuk. Ezek a könyvek a megértéshez szükséges nyelvtudás szintje alapján több kategóriába sorolhatók. E kategóriák alapján döntheti el a tanuló, hogy az aktuális nyelvtudása fejlesztéséhez milyen olvasmány lenne a legmegfelelőbb. Sajnos ezek a könyvek mind számukban, mind a megcélzott idegen nyelv tekintetében meglehetősen korlátozottak. Megírásuk, szerkesztésük egyelőre emberi munkát igényel¹, és a kevésbé népszerű nyelvek esetében kisebb igény is mutatkozik ilyen szövegek olvasására. Bár orosz nyelven elérhetők ilyen, szókincs alapján adaptált könyvek, például a Zlatouszt kiadó gondozásában, azonban a számuk jelentősen elmarad az angol nyelven elérhető hasonló kiadványokétól. Dolgozatomban ezért egy olyan automatizált módszer kidolgozásáról szeretnék írni, amely képes orosz nyelvű irodalmi szöveggyűjteményekben található szövegek előfeldolgozását követően e szövegek bizonyos tulajdonságainak mérésére és több szempontú összehasonlítására. Céлом, hogy a szövegek összevetéséhez olyan algoritmust találjak, amely több mért paraméterből egy képlet alapján képes számszerűen jelezni a vizsgált irodalmi szövegek komplexitását. Ez a módszer a későbbiekben alkalmas lehet mind az orosz nyelvet tanulók, mind az oktatók számára egyszerűbb, könnyen olvasható és értelmezhető szövegek keresésére.

¹ Úgy tűnik, hogy a szövegek adaptálását a mesterséges intelligencia is egyre inkább képessé válik megvalósítani. Hosszú távon az is elképzelhető, hogy ezt a feladatot emberi segítség nélkül el fogja tudni végezni – azonban továbbra is problémát jelent, hogy a modelleket nehéz megfelelően beállítani, valamint ellenőrizni, hogy az adott modell minden esetben igazodik-e a megadott paraméterekhez.

Remélem, hogy eredményeim a későbbiekben hozzájárulhatnak az orosz nyelv oktatásának eredményesebbé tételéhez.

1. A szövegek nehézségi szintjének mérése

A szövegek bonyolultságának kérdése a mindennapi életünk számos területén nagy jelentőséggel bír. A szélesebb embertömegekhez eljuttatandó üzenet esetén érdemes figyelembe venni, hogy a különböző csoportok eltérő társadalmi, családi, műveltségi háttérrel rendelkeznek. A különböző rétegek sikeres eléréséhez, megszólításához a csoportjukhoz leginkább illő nyelvezet és tartalom használatára érdemes törekedni. Ez napjainkban a korábbiakhoz képest még nagyobb hangsúlyt kapott – bár a digitális korszak előtt is voltak olyan területek, ahol ez a tényező fontosnak számított.

Az egyik ilyen területnek tekinthetjük az oktatást. A szélesebb társadalmi rétegeknek szánt különféle tankönyveket, oktatóanyagokat a célközönség számára érthető nyelvezettel kell ellátni. Egy általános iskolai tanuló kezébe nem adható olyan tankönyv vagy szöveg, amely a képességeihez, életkorához képest túl sok idegen szót és bonyolult mondatszerkezetet tartalmaz, ugyanis megértés hiányában az anyag elsajátítása vagy meglehetősen alacsony hatékonysággal működik, vagy egyenesen sikertelenné válik. Ha azonban az oktatóanyag vagy tankönyv egyetemi hallgatók számára készült, akkor megengedett, sőt elvárható a pontos szaknyelvhasználat.

Egy másik fontos terület a média. Ha a társadalom minden tagja számára könnyen érthető híreket, üzeneteket szeretnénk közölni, ezeket úgy érdemes összeállítani, hogy a cikk, híradás nyelvezete ne akadályozza a megértést. Ez fontos szerepet kap a politikai kommunikációban is, ahol egy üzenet széleskörű befogadásának messzemenő politikai és gazdasági következményei is lehetnek.

Az utóbbi időben a törvények is egyre gyakrabban előírják az ún. *plain language*, azaz az egyszerű, közérthető nyelvezet használatát. Ilyen módon az államok lépéseket tehetnek annak érdekében, hogy az ügyintézással (például adózás), vagy az egészségüggyel (például gyógyszerek használati utasítása) kapcsolatos információk pontosabban és egyértelműbben jussanak el a célszemélyekhez, ezzel elősegítve a különböző ügyintézési folyamatok gördülékenyebbé tételét. Ennek a célnak az elérésére törekszik a 20. században az Amerikai Egyesült Államokban indult *Egyszerű Nyelvezet Mozgalom (PLM, Plain Language Movement)*, amely javaslatokat, iránymutatásokat fogalmaz meg a különböző hivatalos szövegek közérthetőbbé tételéhez [1]. Ezek

az iránymutatások tartalmaznak úgy pragmatikai mint stiláris, lexikai és szintaktikai megoldásokat. Többek között a mozgalom hatására jött létre az Amerikai Egyesült Államokban 2010-ben hatályba lépett *Plain Writing Act of 2010* törvény [2], amelynek célja a kormányzati kommunikáció közérthetőségének javítása.

A különböző szövegek komplexitásának mérésére irányuló törekvések az Egyesült Államokban már a 20. század elején, a számítógépes korszak előtt megkezdődtek [3]. Ekkor azonban még csak olyan módszerek jöhettek számításba, amelyek nem igényeltek nagy mennyiségű szövegen végzett vizsgálatot. A nagy mennyiségű szövegen kézzel végzett mérések túlságosan nagy erőfeszítést igényeltek volna, és szembe kellett volna nézni az emberi hibák lehetőségével is. Ezek a mérések gyakran azon a feltételezésen alapultak, hogy egy szöveg egy részén (például egy bekezdésen) elvégzett mérés a szöveg többi részéhez hasonló eredményt fog adni – azonban a több stílusban vagy változatosabb módon megírt szövegekre kapott eredmény sokszor nem tükrözi az egész mű jellegzetességeit.

Az egyik ilyen módszer az 1952-ben Robert Gunning által kifejlesztett Gunning Fog Index [4], amely az iskolai évfolyamok esetében elvárt olvasási képességekhez igazodó skálán mutatja meg a szöveg nehézségének mértékét. A Gunning Fog Index alapján egy 12. szintű szöveg a 12. osztályt elvégzett tanuló olvasási képességeinek felel meg. A módszert angol nyelvű szövegekre fejlesztették ki. Az algoritmus eredeti célja elsősorban különböző napilapok és folyóiratok érthetőségének növelése volt. A mérési folyamat több lépésből áll. Először is a mérni kívánt szövegből ki kell emelnünk egy nagyjából 100 szót tartalmazó szövegrészletet. Ezt a szövegrészletet fogjuk használni a továbbiakban a mérési folyamat során. Második lépésként a kijelölt szövegrészben található szavak számát el kell osztanunk a részletben található mondatok számával. A mondatszámnak és a mondatokban található szavak számának hányadosa adja a képlet első komponensét. A harmadik lépésben ki kell számolnunk az ún. komplex szavak arányát a szövegrészletben található többi szóhoz képest. Gunning a 3 vagy annál több szótagból álló szavakat tekinti komplex szavaknak. A komplex szavak meghatározásánál fontos leszögezni, hogy ezek nem lehetnek tulajdonnevek, összetett szavak, illetve ismert szaknyelvi kifejezések, és a szóvégi toldalékok nem számítanak szótagnak. A komplex szavak számának a szövegrész összes szavainak számával összevetett százalékos aránya fogja adni a képlet második komponensét. Negyedik lépésként a két komponenst összeadjuk, és ezt az értéket megszorozzuk 0,4-gyel. Így

kapjuk meg a szöveghez számolt Gunning Fog Index értéket, amely megmutatja a szöveg megértéséhez szükséges iskolázottsági szintet. Ezt az algoritmust az angol nyelvhez dolgozták ki, ezért alkalmazása az orosz nyelvben problémás.

Egy másik fontos olvashatósági teszt a Flesch-Kincaid olvashatósági szint. Ez a képlet két változót vesz figyelembe, amelyek közül az első a Gunning Fog Indexben is használt, a mondatokban található szavak átlagos számát mutató érték. A második azonban a szavakban található szótagok átlagos számát méri. Ezt a módszert szintén az angol nyelvhez dolgozták ki, azonban történtek kísérletek az adaptálására és alkalmazására orosz nyelvű szövegek esetében is. Irina Osborneva megkísérelte a Flesch-Kincaid olvashatósági szint automatizálását, illetve kalibrálását orosz nyelvű szövegek mérésére [5]. Osborneva a Flesch-Kincaid-féle algoritmust alkalmazva első lépésként egy általa írt program segítségével kiszámította 100 angol nyelvű klasszikus irodalmi mű Flesch-Kincaid olvashatósági szintjét. Kutatásának célja az oktatásban felhasználható szövegek mérése volt, így a kalibráláshoz használt szövegek között gyermekeknek szánt könyvek is szerepeltek. Ezután a képlet változtatása nélkül ugyanezeket a számításokat elvégezte ezeknek a klasszikus irodalmi műveknek orosz nyelvű fordításain is. A képletet ezután úgy kalibrálta az orosz nyelvű szövegekhez, hogy az eredeti képletben szereplő értékeket módosította mindaddig, amíg az angol és orosz nyelvű szövegek mérési eredményei meglehetősen közel nem kerültek egymáshoz. Ezt követően tanulmányában a további orosz nyelvű szövegek automatizált mérését a módosított értékekkel végezte. Egy 2018-ban megjelent, az orosz akadémiai szövegek mérésével foglalkozó cikk azonban rámutatott, hogy ez a módosított képlet nem minden esetben alkalmazható [6]. Mivel a képlet csak irodalmi szövegek alapján lett beállítva, nem volt képes komplex tudományos szövegek mérésére. Ezért a cikk szerzői egy új, a tudományos szövegek mérésére alkalmasabb képletet javasoltak. Bár mind Osborneva, mind Solovyev közölnek értékeket a mérési eredményeikről, ezek reprodukálása akadályokba ütközik. Mindkét tanulmány említi ugyan, hogy a szövegek előfeldolgozáson mentek keresztül, részletes leírást azonban nem adnak ezekről a folyamatokról. És bár leírják, hogy a szöveget szavakra és mondatokra darabolták, nem egyértelmű, hogy milyen algoritmust használtak, azaz a mérési eredményeknél milyen karakterlánc számít egy szónak vagy egy mondatnak.

Véleményem szerint Osborneva módosított módszere kevésbé átlátható, mint az angol nyelvű szövegekre eredetileg kifejlesztett eljárás, Solovyev módszere pedig akadémiai

szövegekkel foglalkozik. Dolgozatomban irodalmi szövegek mérésére szeretnék átlátható, jól dokumentált kísérletet tenni. Kiindulási alapként a Gunning Fog Indexet választottam annak egyszerűsége miatt, azonban ez több akadályba ütközött. Mivel az orosz nyelv gyakran használ olyan toldalékokat, amelyek két (vagy ritkábban három) szótagból állnak, a komplex szavak meghatározása csak a szavak szótövét vagy szótári alakját megbízhatóan megállapítani képes program birtokában lenne megvalósítható. Azonban, mivel az oroszban a toldalékok száma jelentősen nagyobb, mint az angol nyelv esetében, manuális szótövesítésük sokkal nehezebb. További probléma, hogy a szavak szótári alakja már önmagában is tartalmazhat toldalékokat, például igék, kicsinyítő képzővel ellátott főnevek stb. esetében. Ilyenkor ezek a tényezők befolyásolhatják a korábban használt képletet. Az ilyen jellegű automatizáláshoz tehát nincs jelenleg olyan elérhető morfológiai elemző program, amely megbízhatóan és kellő mértékben paraméterezhetően tudná elemezni a szövegekben található szavakat, és képes lenne minden esetben meghatározni az adott szó tövét. Orosz nyelvre a Yandex által kiadott MyStem morfológiai elemző program érhető el [7], azonban a program a kötőjeleket automatikusan szóközökként értelmezi, így a kötőjellel írt összetett szavakat nem tudja helyesen kezelni. Emellett a program nem képes figyelni a szó kontextusát sem, így homográfok esetén nem csak helyes eredményt adhat. Mindezen tényezők tehát egy új szempontokra épülő algoritmus kidolgozását teszik szükségessé.

2. Az elemzéshez szükséges szövegek összeválogatása

A kísérletekhez több különböző szöveggyűjteményt használtam. Ezek összeállítása manuálisan történt, több szempont alapján. A kísérletek során kiderült, hogy a különböző stílusrétegekhez tartozó szövegek esetében (például irodalmi, tudományos) a képletek eltérő eredményt adnak. Ezért szükségessé vált az eredetileg használt szöveggyűjtemény mellé még két másikat is kialakítani.

Az első, a Gunning Fog Indexen alapuló algoritmusokat alkalmazó kísérleteknél használt szöveggyűjtemény 50 nem véletlenszerűen kiválasztott szöveget tartalmazott. Ezek a szövegek úgy lettek összeválogatva, hogy vegyesen tartalmazzanak 19. századi, 20. századi és 21. századi szövegeket, különböző műfajokban, más-más célközönségnek szánva – eredetileg orosz nyelven írt vagy oroszra lefordított műveket. További feltétel volt, hogy a szövegek sokak által ismert szerzőktől származzanak. Így a gyűjteményben helyet kapott például Dosztojevszkij *Bűn és bűnhődése*, Dmitrij Gluhovszkij *Metro 2033* című könyvsorozatának első kötete és Lev Tolsztoj

Anna Karenina című műve. Ezen kívül a gyűjtemény tartalmaz fordításokat is, például Stephen King *Christine* című könyvét, Molnár Ferenc *A Pál utcai fiúk* című regényét, George Orwell két könyvét, a *Harry Potter*-sorozat első kötetének két orosz fordítását, vagy Rejtő Jenő *A tizennégy karátos autó* című regényének orosz nyelvű változatát. A szöveggyűjteménybe belefoglaltam egy számítógép által generált szöveget is. Iszaak Babel *Lovashadsereg* című könyve az előzetesen lefuttatott statisztikák alapján nehéz szövegnek számít. Mivel a mesterséges intelligencia gyors fejlődésének köszönhetően egyre könnyebbé válik a szövegek számítógépes előállítása, megkíséréltem felhasználni a mesterséges intelligenciát szövegek gépi egyszerűsítéséhez. A kísérlethez a ChatGPT nevű mesterséges intelligencia GPT-4O modelljét használtam. Mivel a modell csak kisebb szövegekkel képes dolgozni, a feldolgozáshoz a Lovashadsereg novellái közül kiválasztottam a *Trunov századparancsnokot*. Az eredményként kapott egyszerűsített szöveget az eredetivel együtt beemeltem a későbbiekben elemzendő szöveggyűjteménybe.

Az utolsó módszer kidolgozásakor szükség volt bizonyos fokú kontrollra (validálásra) a képlet kalibrálásához. Solovyev (2018) a megállapításaihoz egy 14 tankönyvből álló szöveggyűjteményt használt a saját modellje beállítására. A gyűjtemény 14 darab, szociális ismeretekről szóló iskolai tankönyvet tartalmazott. Ezek a könyvek két sorozatban jelentek meg, és mindegyikük az Orosz Föderáció Oktatási Minisztériuma által ajánlott tankönyv. Mivel a sorozatokat alkotó könyvek szerzői ugyanazok voltak, feltételezhető, hogy a szövegek stílusa is hasonlóságot mutat. Az első könyvsorozat szerzőcsoportját L. N. Bogolubov, a másodikat A. F. Nikitin² vezette. A két sorozatot – Solovyevhez hasonlóan – a *bog* és a *nik* rövidítésekkel jelöltem. A könyvek mondatra vágott formában érhetőek el az interneten³, egyszerű szöveg (txt) formátumban. A megjelölésekre használt fájlnevek tartalmazzák a megcélzott iskolai évfolyam számát, így ezek segítségével egyszerűbb az egyes mérések eredményeinek összehasonlítása és a mintázatok megállapítása. A *bog* megjelölésű könyvsorozat a 6. osztálytól a 11. osztályig tartalmaz könyveket – illetve, amennyiben rendelkezésre állt, bekerült a gyűjteménybe az adott osztályhoz tartozó magasabb szintű kötet is. A *nik* megjelölésű sorozat 5. osztálytól 11. osztályig tartalmaz könyveket. Az ebben a bekezdésben említett szövegek alkotják a második szöveggyűjteményemet.

² Azokat az orosz neveket, amelyeket Solovyev a tanulmányában megemlít, az ott megjelent formában közlöm.

³ <http://kpfu.ru/portal/docs/F1554781210/shuffled.zip>

A harmadik szöveggyűjtemény iskolások számára ajánlott olvasmányként előírt, válogatott szövegeket tartalmaz. Úgy tűnik, hogy Oroszországban nincs olyan egységes lista, amely meghatározná az iskolákban előírandó kötelező olvasmányokat és a hozzájuk társított iskolai évfolyamokat. Ezért a Labirint internetes könyvesbolt által közzétett, iskolások számára ajánlott olvasmányokat tartalmazó jegyzéket használtam, amely évfolyamonként csoportosítva tartalmazott könyveket [8]. A szövegek fájlneveiben, akárcsak az előző esetben, szerepel az ajánlott évfolyam megjelölése. A lista évfolyamonként több kategóriára osztja az olvasmányokat. Az egyik fő csoportosítási szempont a szerző nemzetisége – tehát külön kategóriába kerülnek az orosz, és külön a külföldi szerzőktől származó szövegek. A szöveggyűjteményembe mindkét kategóriából kerültek szövegek: az orosz szerzők közül például Mihail Lermontovtól a *Korunk hőse*, vagy a külföldi szerzők közül Jules Verne-től a *Grant kapitány gyermekei*. A másik fő szempont a tematikus csoportosítás. Például bizonyos évfolyamok esetében külön csoportot alkotnak a mesék (például Felix Salten *Bambi* című regénye), és külön kategóriát alkotnak a háborús művek (például Arkagyij Gajdar *A dobos története* című kisregénye). A csoportba tartozó fájlok nevében a szerzőre és a címre való utaláson kívül jelzem, hogy a mű eredeti nyelve orosz (r) vagy idegen nyelv (k – külföldi), illetve azt, hogy hányadik osztályban ajánlják olvasásra. A szám melletti p-vel Solovjev nyomán azokat a műveket jelölöm, amelyek hovatarozása két osztály közé került (pl. 9p számmal kifejezve 9,5-nek felel meg). Téma szerinti csoportjelölés nincs. A továbbiakban az egyes modellek és képletek elemzésekor a szövegek megnevezésére csak a szövegek címeit fogom használni, a gyűjteményemben található szövegek pontos listája a dolgozat mellékletében kerül felsorolásra.

3. Lehetőségek a szövegek komplexitásának mérésére

Az új algoritmus kidolgozásához több lehetőséget is megvizsgáltam. A kiindulási alapot a Gunning Fog Indexben használt paraméterek szolgáltatták. A szövegek egy előfeldolgozási folyamaton mentek keresztül. Ennek első lépéseként a szövegek egy olyan szűrőn mentek át, amely a szövegben található, a MyStem program által fel nem ismert szavak mennyiségét vizsgálta. Ez a szűrő segített a hibásan digitalizált szövegek kiszűrésében. Azok a szövegek, amelyekben a legalább egyszer előforduló szavak 10%-a, vagy az összes előforduló szó 20%-a nem volt felismerhető a MyStem számára, el lettek távolítva a szöveggyűjteményből. Második lépésként, a könnyebb elemzés érdekében a szövegeket egy Perl script segítségével mondatokra választottam szét. Az általam használt szegmentálást végző script [9] minden olyan szövegrészt mondatként kezel, amelyben egy mondatzáró írásjellel végződő szó után egy szóközt

követően nagybetűvel kezdődő szó állt – amennyiben a mondatzáró írásjellel végződő szó nem szerepelt a script rövidítéseket tartalmazó listáján. Ez utóbbi kitétel például azoknál az eseteknél fontos, ahol a rövidítést egy név követi, mert ilyen esetekben a script tévesen jelzett volna mondathatárt. A script jelenleg egy ismert esetben tehet téves jelzést, azonban ennek kiküszöbölése még nem történt meg. A könyvek elején található kiadási információk, címek, alcímek általában mondatzáró írásjel nélküli egységek, ezek felismeréséhez azonban fontos a kontextus ismerete. Mivel ezeknek az adatoknak a megfogalmazása nem egységes, így nehezen (vagy egyáltalán nem) építhető rá olyan program, amely minden esetben hibátlanul működne. Jelenleg a program ezeket az adatokat egy mondatként kezeli, amely a szövegben talált legelső következő mondatzáró írásjelig tart. A script kimenetként egy olyan fájlt ad vissza, amelyben egy sorban pontosan egy mondat található. Ez megkönnyíti a későbbi elemzéseket, mert így a statisztikai számításokat végző scriptnek már nem kell szövegfeldolgozást végeznie. A harmadik lépés egy újabb szűrőfolyamat volt, amelynek eredményeként megkaptuk az elemzésben felhasználható szavak listáját. A szövegekben előfordulhatnak olyan szó értékű karakterláncok, amelyek nem lényegesek az elemzés szempontjából: ilyenek például a számok vagy a nem cirill abc betűivel írt szavak. A szólista elkészítéséhez egy script először a szóköz karakterek mentén feldarabolta a szöveget. A második lépésben az előző lépés eredményeként kapott listából eltávolította a lábjegyzetekhez és hivatkozásokhoz használt szögletes és kapcsos zárójeleket, valamint a bennük található számokat. Ezután minden szóból eltávolított bizonyos írásjeleket és egyéb jeleket, például vesszőt, kettőspontot, mondatvégi írásjeleket. A két szó közötti, szóközzel nem előzött, és nem is folytatott rövid kötőjel karaktert („-”) azonban meghagyta, mivel ez gyakran részét képezi a szavaknak, például a *umo-mo* szóban. A szólista elkészítésének negyedik lépéseként a script minden *Ё* és *ё* karaktert kis *e* karakterre cserélt, mert ezeknek a használata nem egységes a szövegekben. A következő művelet a nagybetűk kisbetűkre cserélése volt a cirill abc minden betűje esetében. Ez a későbbiekben a szóalakok összehasonlítását könnyítette meg, mert a számítógépes elemzések számára a kis- és nagybetűk különböző karaktereknek számítanak. Utolsó lépésként a script csak olyan elemeket tartott meg a feldolgozott listából, amelyek kizárólag az orosz cirill abc 32 betűjét (azaz a *ё* nélkül), és a rövid kötőjel („-”) karaktert tartalmazták.

A továbbiakban kiértékelem az algoritmuskísérletekből származó eredményeket. A kipróbált módszerek 3 nagy csoportra oszthatók. Az első algoritmus a Gunning Fog Index egy módosított változatával számolt, ahol a szövegben található szavak átlagos szótagszámát és a

szöveg mondatainak átlagos hosszát használta fel. Ennek egyik továbbmódosított változata szótagszám alapján próbálta meg súlyozni a szavakat. A második kísérlet a szótagszám helyett a szavak karakterekben, majd hangokban mért hosszúságával számolt a mondatok átlagolt hosszúsága mellett. A harmadik algoritmuscsoporthoz egy új paramétert emelt be a számításokba, amely a szövegekben előforduló szóalakok gyakoriságán alapult.

3.1. Szövegek mérése a szótagszám és a mondatok átlagos hossza alapján

A szótagszámok alapján dolgozó algoritmus a Gunning Fog Indexhez hasonlóan működött. Az eredeti és az általam használt képlet eltérését az adta, hogy az eredeti képletben szereplő komplex szó fogalma helyett a script az orosz szövegekben található szavak szótagszámának átlagát vizsgálta. Erre azért volt szükség, mert megbízható morfológiai elemzést végző eljárás nem állt rendelkezésre, így a komplex szavak meghatározása nem lett volna megoldható. A számítások elvégzése előtt szükség volt a bemeneti szavak szűrésére. Ehhez a fentebb már bemutatott eljárást alkalmaztam, amelynek eredményeképpen egy megtisztított szólistát kaptam. A lista valamennyi eleme olyan szó volt, amely csak a cirill abc betűit tartalmazta a *ë* kivételével, illetve a rövid kötőjel „-” karakter megtartásával. A képlet első komponense tehát ezen szavak szótagszámának átlaga volt. A szótagok számát a script a szóban előforduló magánhangzók számával adta meg. A magánhangzókat nem tartalmazó szavakat – ezek jellemzően előjárók (például *в, с, к*) – a script 0 szótagból álló szóként kezelte⁴. A script a képlet második komponenseként kiszámolta a szöveg mondataiban található szavak számának átlagát is – ezt az értéket a többi algoritmus is megtartotta. Bizonyos szélsőséges esetek miatt azonban korlátozásokat kellett bevezetni a mondatokkal való számolásban is, voltak ugyanis olyan szövegek, amelyek az írói szándék szerint összesen egy mondatból álltak, terjedelemben azonban több száz oldalon keresztül tartottak. Ezért az általam kidolgozott algoritmus valamennyi elemzett szövegben figyelmen kívül hagyja az 5 leghosszabb mondatot. Már a korai eredményekben is megmutatkozott, hogy a szavak átlagos szótagszámát illetően a szövegek között nincs nagy különbség, és az eredmények nem hozták a várt értékeket sem. A mondatokban található szavak átlagos száma azonban már pontosabb képet adott a szövegek nehézségi szintjéről. A szöveggyűjteményen futtatott mérés eredményeinél látszik, hogy az *Anna*

⁴ Tisztában vagyok a megoldás vitathatóságával, de a szóba jövő másik két megoldás, azaz a szóközök közé írt, magánhangzókat nem tartalmazó szavak teljes kihagyása, illetve ezeknek – attól függetlenül, hogy nem tartalmaznak magánhangzót – egyszótagúként való meghatározása szintén kétséget ébreszt. Érdemes megfigyelni, hogy az orosz *школе* szövegrész kiszámításánál az első, az általam használt megoldás szerint ez esetben egy szó átlagosan egy szótagból áll, a másik szerint kettőből, a harmadik szerint pedig 1,5-ből.

Kareninát sokkal hosszabb mondatok jellemzik, mint például a *Metro 2033*-at, vagy a *Harry Potter* fordításokat. A megvizsgált eredmények alapján az is egyértelművé vált, hogy a szavak átlagos szótagszáma a szöveg komplexitásának megállapításánál nem fog fontos szerepet játszani.

Következő lépésként a szótagalapú mérést alakítottam át úgy, hogy a szavak a bennük található szótagok számának megfelelően számítottak bele a képlet eredményébe. Ehhez súlyozott pontszámokat használtam, ahol a több szótagból álló szavak több pontot kaptak, ezzel növelve a számítás végső eredményét. A script fokozatosan növelte a szótagokhoz társított pontok számát. Így nem kaptak pontot az 1 és 2 szótagos szavak, 1 pontot kaptak a 3 szótagos szavak, 2 pontot a 4 szótagosak, 3 pontot az 5 szótagosak, 4 pontot a 6 szótagosak; a 7 vagy annál több szótagos szavak pedig 5 pontot kaptak. Az eredményből kiderült, hogy jelentős különbségek így sem tapasztalhatók az egyes szövegek között, azonban rámutattak egy fontos problémára a tömeges szövegelemzéssel kapcsolatban. A digitalizált szövegek minősége eltérő lehet, és előfordulhatnak bennük a digitalizálásból adódó hibák. Az egyik ilyen hiba, amikor két szó közé a digitalizáláskor nem tesznek szóközt, vagy a szövegfelismerést végző program egy szónak értelmezi a szkennelt képen látható, egyébként külön írt szavakat. A mérések során kiderült, hogy az ilyen szövegek rontják az eredményt, és a legmagasabb értékeket a mérések során ezek a szövegek adták. Ebből következően, mivel a tömeges szövegek elemzésekor nincs mód ezeket a problémákat kiszűrni, ennek az algoritmusnak a használatát nem tartottam célravezetőnek.

3.2. Szövegek mérése a szavak átlagos hossza és a mondatok átlagos hossza alapján

A következő megközelítésként a szavak betűinek és hangjainak átlagos számát használó képletet alkalmaztam. A szavak ennél a scriptnél is a fentebb ismertetett szöveg tisztító eljárás kimeneti szólistájából származtak. Ebben a képletben először a szavakban található betűk átlagos számát vettem első komponensként. Kiderült azonban, hogy mivel az orosz nyelvben gyakran fordulnak elő olyan karakterek, amelyek csak a szó hosszát növelik, a kiejtésbe viszont nem számítanak bele, ez a megoldás sem vezetett eredményre. Ráadásul a rövid szavak, például az előljárók, a névmások az eredményt erősen befolyásolták. A tesztek alkalmával nem sikerült egyértelmű eredményre jutni ezzel a módszerrel sem. Így következő lépésként a képletbe olyan módosítást vezettem be, amelyben inkább fonetikai szempontok érvényesültek, tehát a script a karakterek számolása helyett a szavakban előforduló hangok átlagos számát vizsgálta. Ehhez több szabályt is be kellett építeni. Egyrészt az orosz nyelvben használt bizonyos magánhangzók (*е, я, ю*) a pozíciójuktól függően egy vagy két hangot is jelölhetnek, másrészt ki kellett szűrni azokat a

karaktereket, amelyek nem jelölnek hangot (ѣ, ъ). Az eredmény azonban nagyrészt változatlan maradt. Bár az értékek között kisebb különbségek megfigyelhetők, azonban használható mintázatok nem sikerült felfedezni a vizsgált szövegekben.

3.3. Szóalakok gyakoriságát használó képletek

Mivel a fentebb bemutatott paraméterek mérése nem hozott megbízható eredményt, más módszert kellett keresni. Solovyev (2018) tanulmányában megkísérelte akadémiai szövegek nehézségének mérését az elterjedt képletektől eltérő paraméterek alapján [6]. Az egyik ilyen paraméter volt a szavak előfordulásának gyakorisága, amelyet a Dale-Chall olvashatósági index is felhasznál. A Dale-Chall olvashatósági indexet 1948-ban fejlesztette ki Edgar Dale és Jeanne S. Chall [6]. Az általuk használt képlet egyik komponense a Flesch-Kincaid olvashatósági tesztben, és később a Gunning Fog Indexben is felhasznált érték, a mondatokban található szavak átlagos száma. A második paraméterként azonban egy új megközelítési módot, az ún. *ismerős szavak* (*familiar words*) fogalmát vezette be. Ennél a módszernél egy korábban összeállított, 3000 szót tartalmazó listát használtak. A listán szereplő szavakat *ismerős szóként* határozták meg. A szöveg komplexitásának mérése során a képlet második komponenseként a szövegben található nem ismerős szavak százalékát használták fel. Ez a megközelítés egy szókinszalapú mérést tett lehetővé. Ezért egy további próbálkozást jelenthet az adott szövegekben megtalálható szókinsz mérése. Mivel az orosz nyelv morfológiai rendszere sokkal összetettebb az angolénál, szükség lett a 3000 szót tartalmazó lista kibővítésére. Kellett egy megbízható, szabadon letölthető, nagy szöveggyűjteményből kinyert szólista, amely előfordulási gyakoriság szerint tartalmazza az orosz szövegekben felbukkanó szóalakokat. Feltételezhető, hogy a gyakrabban előforduló szóalakok az adott nyelv szókinszének fontosabb elemei. Ezért elvégezhető egy olyan számítás, amely megmutatja, hogy az adott szövegben előforduló szavak hány százalékát teszik ki egy olyan csoportnak, amelyet szókinsz szempontjából fontosnak (ismerősnek) határozunk meg. Lényeges azonban, hogy mivel nincs olyan megbízható eljárás, amely a szavakat a szótári alakjukra tudná visszavezetni, ezért ezen a listán ugyanannak a lexémának az elemei különböző alakokban többször is szerepelhetnek (pl. főnevek esetén több esetben, egyes és többes számban). A fontosnak tartott szóalakok meghatározásához kísérletezésre van szükség.

A méréshez kellett az orosz nyelv leggyakoribb szóalakjait tartalmazó szólisták. Ezek a listák [10] szintén átestek a fent már leírt szövegtisztítási eljáráson. Egy-egy lista az orosz nyelv meghatározott számú leggyakoribb szóalakjait tartalmazza a fentebb említett szöveggyűjteményen

mérve. A saját mérésemhez a következő, megadott számú leggyakoribb szóalakot tartalmazó listákat használtam: 1000, 1500, 2000, 3000, 5000, 10000, 20000, 30000, 40000, 50000, 100000. A mérések alapján úgy tűnt, hogy az 5000 leggyakoribb szóalakot tartalmazó lista már elég változatos eredményt ad – ahogy az is kiderült, hogy a lista hosszúságának növelése negatívan hatott az elért eredményekre.

Mivel a cél egy konkrét értéket adó képlet kidolgozása volt, ezért a szövegeket vizsgáló scriptbe korábbi paramétereket is bevezettem. Az új script 6 értéket mért a paraméterként kapott szövegen. Az első paraméter a már fent ismertetett szavakban található átlagos szótagszám volt. A script a számításokat itt is a szövegből generált tisztított szólista alapján végezte. A második paraméter a szintén a tisztított szólistán elvégzett szavakban található átlagos karakterszámot tartalmazta. A harmadik és negyedik paraméter a mondatokban található szavak, illetve karakterek átlagos számát mutatta meg. Az 5. és 6. paraméterek az 5000 és 10000 leggyakoribb szóalakot tartalmazó listákon nem szereplő szóalakok százalékat adták meg – itt szintén a tisztított szólista volt az összehasonlítás alapja. Történt több kísérlet is a paraméterek egyszerű képletbe foglalására, ezek azonban nem adtak megfelelő eredményt. A 6 paraméter együtt túl sok, túl komplex képletet hozna létre, valamint a kipróbált képletek eredményei nem voltak könnyen hozzárendelhetők egy skálához. Ezért szükség volt a paraméterek korrelációinak matematikai vizsgálatára és tesztelésére.

A paraméterek korrelációjához és a megfelelő képlet megtalálásához egy lineáris regressziót végző modellt használtam. A modell tanítását külön-külön elvégeztem mind a Solovjev-cikkben tárgyalt akadémiai szöveggyűjteményen, mind pedig az iskolai kötelező olvasmányokat tartalmazó gyűjteményen. A modell tanításához itt is szükség volt a szövegek előfeldolgozására. A szövegeket először mondatokra vágtam a fent már ismertetett script segítségével. Ezután az egyes fájlokban található mondatokat véletlenszerűen újrendeztem. Erre azért volt szükség, mert ellenkező esetben a szövegekben olyan mondatblokkok alakulhattak volna ki, amelyek befolyásolták volna a mérési eredményeket. Ezután minden fájlt 11 egyenlő (azaz kb. ugyanannyi mondatból álló) részre választottam szét. Mivel a mondatok véletlenszerű sorrendben szerepeltek, így nagyobb a valószínűsége, hogy az egyes szövegrészleteken végzett mérések nagyjából hasonló eredményt adnak egy szövegen belül. Ellenkező esetben egy párbeszédekben gazdag szövegrész az adott szövegrész fájljára kiugróan alacsony mondathosszúsági értéket adhatott volna. A 11 részre feldarabolt fájlokat ezután úgy csoportosítottam, hogy valamennyi fájl

első 10 részlete a tanításra szánt adathalmazba került, minden szöveg utolsó része pedig az ellenőrzési halmazba. Így a modell tanításakor a tanított adatokhoz hasonló ellenőrzési halmazt kaptam. A tanítási halmaz bemenetei a szövegen mért paraméterek voltak, kimenete pedig a Flesch-Kincaid olvashatósági szinthez és a Gunning Fog Indexhez hasonló iskolai évfolyamot meghatározó szám. A modellek hatékonyságát az átlagos négyzetes hiba kiértékelésével kaptam meg. Ez az érték megmutatja, hogy mekkora az eltérés az ellenőrzésre szánt adathalmazban az előre jelzett és a valós értékek között. Minél kisebb ez a szám, a modell annál jobban teljesít az értékek előrejelzésében. Az eredményből azonban az látszott, hogy valóban erős korrelációt leginkább a Flesch-Kincaid olvashatósági szintben és a Dale-Chall indexben már felhasznált paraméterek adnak. A többi kipróbált képlet magasabb hibaértékeket adott a lineáris regresszió kiértékelésekor. Ezért végül a Flesch-Kincaid olvashatósági szint és a Dale-Chall index képletének módosítását végeztem el. Bár egy hasonló módosítást már Solovyev is javasolt [6], azonban a tanulmányában módosított képlet csak akadémiai szövegekre vonatkozott. További probléma volt, hogy Solovyev nem fejtette ki pontosan, hogy milyen módszerrel végezte a mondatra vágást, illetve hogy milyen szabályok határoznak meg egy szót. A most kidolgozott képletek célja irodalmi szövegek értékelése. A kísérletek során kiderült, hogy valószínűleg nem lehet egy olyan képletet létrehozni, amely egyaránt alkalmas lenne tudományos és irodalmi szövegek kiértékelésére. Ezért két olyan képletet javaslok, amelyek inkább az irodalmi szövegek nehézségének meghatározására alkalmasak. Az új képletek visszaellenőrizhetők és reprodukálhatók, mert a megírt scriptek szabadon hozzáférhetők [11]⁵.

4. Az új képletek, és a képletek által adott eredmények bemutatása

A képletek adaptálását a Flesch-Kincaid indexszel kezdtem. Az eredeti, 1975-ben az angol nyelvhez készített képlet a következő:

$$FKG = 0.39 * ASL + 11.8 * ASW - 15.59$$

ahol ASL a mondatokban található szavak átlagos számát, ASW pedig a szavakban található szótagok átlagos számát jelenti. Ez a képlet azonban nem veszi figyelembe az orosz nyelv sajátosságait. Az általa elvégzett mérések alapján Solovyev két különbségre mutat rá [6]. Az első, hogy az angol szavak átlagos szótagszáma 2,97, míg az orosz szavak esetében ez az érték 3,29. A

⁵ A dolgozatomhoz felhasznált és megírt scriptek elérhetők az interneten és szabadon felhasználhatók.

másik különbség, hogy az angol mondatok szóban mért átlagos hossza 1,5-szöröse az orosz mondatok szavakban mért átlagos hosszának. Ezért Solovjev a képletet a következőképp módosította:

$$FKG_S = 0.36 * ASL + 5.76 * ASW - 11.97$$

Ezzel a képlettel kiértékelve az általa megadott szövegeket az átlagos négyzetes hiba kerekítve 1,02. Solovjev ennek a képletnek a leírásakor azonban nem határozta meg, hogy hogyan történt a modell betanítása és a hatékonyság ellenőrzése. Ezért a számításokat újra elvégeztem, és a tanítási és ellenőrzési folyamatokhoz a fent leírt módszert használtam, ahol minden szöveg egyenlő mértékben járult hozzá a két adathalmazhoz. Eredményként egy másik képletet kaptam:

$$FKG_NEW_ACADEMIC = 0.4042 * ASL + 3.7542 * ASW - 8.3518$$

Ennek a képletnek az átlagos négyzetes hibaértéke az előző 1,02 helyett csak 0,2936. Mivel ez a modell tudományos jellegű szövegeken lett tanítva, a célom pedig az irodalmi szövegek vizsgálata, el kellett végezni egy második tanítási folyamatot is.

A második modell tanítását a 32 szöveget tartalmazó, osztály alapján megjelölt szövegeken végeztem. A szövegeket most is a fent leírt módszer segítségével daraboltam egyenlő részekre, majd a szövegek 10/11 része bekerült a tanítási, 1/11 része pedig az ellenőrzési adathalmazba. A modell lefuttatását követően az alábbi képletet kaptam:

$$FKG_NEW_FICTION = 0.2269 * ASL + 3.3173 * ASW - 2.4248$$

A modellhez kapott átlagos négyzetes hiba 4,7586, ami sokkal pontatlanabb a tudományos szövegek modelljéhez képest. A modellt kiértékeltem egy olyan korpuszon, amely tartalmazta mindhárom korábbi korpusz szövegeit. Az alábbiakban szeretnék bemutatni néhány példát az eredmények közül.

A modell a legalacsonyabb eredményt, 6,3055-öt *A kis hercegre* adta. A lista elején leginkább mesék, elbeszélések szerepelnek (például *Bambi* – 6,4584, *Micimackó* – 6,5364, *A dobos története* – 6,6038). A legmagasabb eredmény a *bog* tankönyvsorozat 11. osztályos tanulóknak írt könyvéhez tartozik: 11,1669. A listában az utolsó elemek nagyrészt tankönyvek, előtte szép-irodalmi művek találhatóak (például *bog* tankönyvsorozat 8. osztály – 10,401, *Arszenyev élete* –

9,8631, a *Holt lelkek* 1. kötete – 9,0672). Lényeges különbség mutatkozik a *Trunov századparancsnok* eredeti, és a ChatGPT által egyszerűsített változata között: az eredeti szövegre kapott érték 8,5119, az egyszerűsített szövegre pedig 6,8486.

A modell hibaértékén látszott, hogy a szövegek besorolása nem túl hatékony, ezért más megoldásban kell gondolkodni. A Dale-Chall index, hasonlóan a Flesch-Kincaid szinthez, a mondatokban található szavak átlagos számával dolgozik, de az átlagos szótagszám helyett a már korábban is kipróbált szavak gyakoriságát mérő módszert használja. Ezért a lineáris regressziót végző modellt átalakítottam úgy, hogy ezt a két paramétert vegye figyelembe. Az angol nyelvhez javasolt Dale-Chall képlet a leggyakoribb 3000 szót veszi figyelembe. Az eredeti, 1948-ban megalkotott képlet a következő:

$$DC = 0.1429 * DWP + 0.4756 * ASL - 4.7279$$

ahol ASL továbbra is a mondatokban található szavak átlagos számát jelöli, DWP pedig az ún. nehéz szavak százalékát, azaz azoknak a szavaknak az összes szóhoz képest mért százalékban kifejezett arányát, amelyek nem szerepelnek a leggyakoribb szavak listáján. Dale és Chall eredetileg a leggyakoribb 3000 szót használta a képletben. Azonban már a korábban az orosz szövegeken elvégzett kísérleteim alapján is látszott, hogy az orosz nyelv komplexebb morfológiai rendszere miatt célszerűbb egy nagyobb listát használni. A modell végül az 5000 leggyakoribb szóalakot tartalmazó listát vette alapul. Mivel továbbra sem állt módomban elvégezni sem szótövesítést, sem morfológiai elemzést, így a listán nem szavak, hanem szóalakok szerepelnek, tehát az ugyanazon lexéma paradigmájába tartozó szó több alakban is előfordul. Az előzőkhez hasonlóan most is az akadémiai gyűjtemény kiértékelésével kezdtem. Eredményként az alábbi képletet kaptam:

$$DC_NEW_ACADEMIC = 0.1429 * DWP + 0.4756 * ASL - 4.7279$$

A modell kiértékelésekor kapott hibaérték 0,2901, amely nagyon kis mértékben alacsonyabb a FKG_NEW_ACADEMIC képlet hibaértékénél ugyanezen a szövegeken. A következő lépésként ugyanezt a tanítási folyamatot végeztem el a 32 irodalmi szöveget tartalmazó korpuszon is. A szógyakorisági paramétert megtartottam, tehát az előző modellben is használt 5000 leggyakoribb szóalakot tartalmazó listával dolgozott a modell. A következő eredmény született:

$$DC_NEW_FICTION = 0.1027 * DWP + 0.2753 * ASL + 0.9304$$

A modell hibaértéke 4,5332, amely szintén csak kis mértékben jobb, mint a korábbi hibaérték. Néhány példa következik a mérés eredményei közül. A legalacsonyabb értéket ismét *A kis herceg* kapta, az érték 5,68. A DC_NEW_FICTION érték alapján növekvő sorrendbe rendezett eredménylista elején azonban történtek változások: például az FKG_NEW_FICTION szerint rendezett lista 9. helyéről a 2. helyre került egy Sarah Morgan Harlequin-sorozatba tartozó regénye (*Dare she date the dreamy*) – 6,1524, a korábbi képlet alapján 6,6939-es értékelést kapott), a *Bambi* pedig a korábbi lista 3. helyéről a 32. helyre került (korábbi érték: 6,4584, új érték: 7,3158). A legmagasabb értékek között is megfigyelhető egy átrendeződés: például a legmagasabb értéket a *bog* tankönyvsorozat 11. osztályosoknak szánt könyve helyett az *Arszenyev élete* kapta (10,63), mögötte a lista végén továbbra is tankönyvek vannak, de közöttük több az irodalmi szöveg: például az *Egy város története* – 9,6442, a *Trunov századparancsnok* eredeti szövege – 9,4591. Ennek és a ChatGPT által egyszerűsített változata közötti különbség itt is megmutatkozik: az eredeti szöveg 9,4591-es értéket kapott, a ChatGPT által egyszerűsített változat 6,8294-et.

Bár az erre a képletre kapott hibaérték árnyalatnyival alacsonyabb volt a FKG_NEW_FICTION hibaértékénél, továbbra is pontatlan maradt. A lineáris regressziót végző modellbe ezért bevezettem egy harmadik paramétert. A korábbi DWP paraméter DWP2-re módosult, az új paraméter neve pedig DWP1. Ez a paraméter a leggyakoribb 1000 szóalakot tartalmazó listán nem szereplő szavak százalékát mérte. A képlet a tanítási folyamat után a következőképpen módosult:

$$DC_NEW_FICTION2 = -0.5921 * DWP1 + 0.6362 * DWP2 + 0.1827 * ASL + 11.6611$$

A modell hibaértéke 3,6572, ami már jelentősen jobbnak mutatkozik az előző irodalmi szövegeken tanított modelleknél. A listában nagy átrendeződés figyelhető meg: első helyre a Szorokin *Regénye* (Поман) kerül – 4,0452, utána Lukjanyenko *Éjszakai őrsége* – 4,9871, *A kis herceg* csak a harmadik – 5,346. A lista végén kisebb az átrendeződés: az utolsó előtti 3. helyen az *Egy város története* áll – 9,8203, az utolsó előtti helyen az *Arszenyev élete* – 10,63, utolsó helyen pedig *Az elvarázsolt zarándok* áll – 10,1514. Természetesen e modell hibahatára is magasnak számít, azonban az összes kipróbált modell közül ez teljesített a legjobban.

5. A tanítási szöveggyűjtemény bővítése

A DC_NEW_FICTION2 képlet, bár hibaértéke alapján jobban teljesített az előző képleteknél, tovább pontosítható. Ehhez azonban szükség van egy olyan, nagyobb méretű szöveggyűjteményre, amely iskolai évfolyamok szerint megbízhatóan felcímkézett ajánlott olvasmányokat tartalmaz. Bár úgy tűnik, Oroszországban nincs egységes előírás az iskolák számára, azonban a Föderációs Pedagógiai Mérési Intézet publikált egy, az általános irodalomtanítást célzó föderációs oktatóprogramhoz összeállított ajánlott olvasmánylistákat.⁶ Ez a korábban használt olvasmánylistáknál sokkal bővebb anyagot szolgáltat a modell pontosításához. Fontos megjegyezni azonban, hogy Oroszország területén sok régió található, és a lakosság etnikai szempontból nagyon változatos, ezért az iskolai oktatásban valószínűleg fontos szerepet játszanak a regionális-nemzetiségi sajátosságok is. Ez is magyarázhatja egy egységesen előírt olvasmánylista hiányát. Ebből az is következik, hogy ha más állítana össze hasonló olvasmánylistát, akkor nagyon valószínű, hogy más művek és más sorrendben szerepelnének rajta. Ezen a listán is látható, hogy a készítői nagy mozgásteret hagytak a listát használó tanároknak. Egy szöveggyűjtemény összeállításához, amennyiben ezeket az ajánlásokat fel akarjuk használni, elengedhetetlen meghozni néhány döntést.

A szöveggyűjtemény összeállítása a következő elvek mentén történt. A gyűjtemény nem tartalmazza az első 4 osztály számára ajánlott olvasmányokat, mert ezek gyakran olyan rövid szövegek voltak, amelyeken nem lehetett megbízható méréseket futtatni. A szövegek közül, mint korábban is, kiszűrtem minden olyan szöveget, amelyben a MyStem morfológiai elemzőprogram szerint a legalább egyszer előforduló szavak 10%-a, vagy az összes előforduló szó 20%-a nem volt a MyStem által felismerhető. A szövegek közül, illetve a gyűjteményekből eltávolítottam a verseket, mert esetükben a mondatokat vizsgáló paraméter nem mérhető. Ezen felül a drámákat is kivettem a gyűjteményből. Ahol egy olvasmányt több osztályhoz is besoroltak, ott automatikusan mindig az alacsonyabb évfolyam lett az érvényes besorolás. Amikor a lista egy szerző esetében nem konkrét műveket, csupán példákat ad meg, ott a példák lettek beválogatva a szöveggyűjteménybe. Ahol három művet ír elő az ajánlás, de csak két példa van megadva, ott a

⁶ https://doc.fipi.ru/metodicheskaya-kopilka/univers-kodifikatory-oko/nachalnoye-obshcheye-obrazovaniye/literatunoe_chtenie_1-4_un_kodifikator.pdf; https://doc.fipi.ru/metodicheskaya-kopilka/univers-kodifikatory-oko/osnovnoye-obshcheye-obrazovaniye/literatura_5-9_un_kodifikator.pdf; https://doc.fipi.ru/metodicheskaya-kopilka/univers-kodifikatory-oko/sredneye-obshcheye-obrazovaniye/literatura10-11_un_kodifikator.pdf

szöveggyűjteménybe is csak két példa került be. Amennyiben a lista egy tétele csak szerzőt határoz meg, úgy az adott szerzőtől véletlenszerűen került egy mű beválogatásra. Olyan esetben, ahol az író az orosz nyelven kívül más nyelven is alkotott (mint például Vladimir Nabokov esetében), ott az eredetileg is orosz nyelven írt mű lett beválogatva. Olyan könyvek esetében pedig, amelyek nem orosz, hanem külföldi szerzőtől származtak, és több fordításuk is elérhető, csak egy fordítás került a szöveggyűjteménybe. Ahol a lista csak bizonyos fejezeteket jelölt meg egy könyvből, ott az egész könyv bekerült a szöveggyűjteménybe. Így a korábbi 32 műből álló listához képest egy 207 műből álló tanítási gyűjteménnyel folyt a képletek meghatározása.

A képlet megállapításához használt modell tanítása most is a 3.3 alfejezetben ismertetettek szerint történt. A tanítási folyamat lefuttatása után látszott, hogy a modell hibaértéke magas: mindhárom képlethez használt modell hibaértéke 5 felett volt. Ez szükségessé tette egy évfolyameloszlás-korrekción és szűrés elvégzését. A korrekció során a kiugróan magas számban megtalálható évfolyam-besorolású könyvek közül véletlenszerűen eltávolítottam bizonyos műveket a gyűjteményből, így egy kiegyenlítettebb adathalmazt kaptam. Ezután elvégeztem a kiugró eredmények szűrését is. Az első futtatásból származó eredményeken látszott, hogy voltak olyan esetek, ahol az olvasmány évfolyam-besorolása és a modell által adott besorolás különbsége nagyobb volt 3-nál. Ezeket a szövegeket eltávolítottam a gyűjteményből. Ezután elvégeztem egy második tanítási folyamatot is, ekkor már csak a gyűjteményben maradt 161 szövegen. A modell itt már a korábbi DC_NEW_FICTION2 képlet hibaértékénél jelentősen alacsonyabb hibaértéket adott, a DC_NEW_FICTION2 képletben használt paraméterek alapján kialakított új képlet modelljének hibaértéke 2,3795. Kísérletképpen a szöveggyűjteménybe beemeltem a korábban a Solovjev (2018) által elérhetővé tett 14 tankönyvet, mivel ezek az Oktatási Minisztérium által évfolyamok szerint címkézett könyvek. A könyvek bevitelét követően a 175 fájlra elvégzett optimalizálás eredményeként a DC_NEW_FICTION2 képlet paramétereivel dolgozó modell hibaértéke az előzőnél alacsonyabb lett, 2,3370. A végleges képlet:

$$DC_NEW_FICTION3 = 0.0093 * DWP1 + 0.0569 * DWP2 + 0.2629 * ASL + 2.0093$$

A képlet kiszámítása után elvégeztem a korábban is lefuttatott 50 besorolás nélküli könyv, és a 14 évfolyam szerint besorolt tankönyv elemzését. Ebben az esetben a korábban tanításra használt 32 könyvet nem foglaltam bele az elemzésbe, mivel ezek a tanítási adathalmazba nem kerültek bele. Így a következő elemzés tárgyát képező eredménylista összesen 64 művet

tartalmazott. A dolgozat mellékletében közzéteszem az 50+14 könyv eredményei mellett a teljes tanítási adathalmaz elemzését, illetve külön azoknak a könyveknek az elemzését, amelyek időközben a tanításra szánt szöveggyűjteményből eltávolításra kerültek (beleértve az első 4 évfolyamnak szánt műveket is).

Az eredményeket növekvő sorrendbe rendezve látható, hogy az első helyen *A kis herceg* szerepel, 5,9879-s besorolással. A lista elején szintén a korábban is alacsony besorolást kapott művek szerepelnek: az *Éjszakai őrség* 6,2478, a *Dare she date the dreamy* 6,2982, a *Micimackó* 6,5679, a *nik* tankönyvsorozatának 5. évfolyamos tanulónak szánt kötete pedig 6,9606-os értékelést kapott. A lista végén az *Arszenyev élete* szerepel a legmagasabb besorolással, 10,2904, ezeket előzik meg a *bog* és *nik* tankönyvsorozat 10. és 11. évfolyamos tanulónak szánt kötetei, például a *nik* sorozat 11. évfolyamos kötete 9,3965, a *Holt lelkek* pedig 9,0972. A mesterséges intelligencia által egyszerűsített, és az eredeti *Trunov századparancsnok* novellát összehasonlítva a *Trunov századparancsnok* ChatGPT-vel egyszerűsített változata a lista 11. helyén áll 6,8523-as besorolással, míg az eredeti szöveg a lista 56. helyén áll 8,9814-es értékkel.

Bár a képlet a hibaértékek alapján pontosabb, mint a DC_NEW_FICTION2 képlet, láthatjuk, hogy még mindig vannak esetek, ahol nem ad kellően pontos meghatározást. Mivel azonban a tesztek alapján a DC_NEW_FICTION3 képlet jobban teljesít, érdemesebb ezt a képletet használni a szövegek olvashatóságának mérésére. Fontos ugyanakkor megjegyezni, hogy mivel az irodalmi szövegek meglehetősen sokszínűek, egy bizonyos hibahatáron túl már nem lehet a képletet tovább optimalizálni, és az irodalmi művek természetéből adódóan mindig lesznek olyan esetek, amikor a képlet nem ad pontos eredményt.

6. További potenciális szempontok

A paraméterekkel folytatott kísérletezésből levonható következtetések alapján úgy tűnik, hogy a komplexebb paraméterek bevezetése nem jelenti az algoritmus teljesítményének javulását. Ezért tűnt célszerűnek a legkönnyebben mérhető szövegtulajdonságok vizsgálata. Vannak azonban olyan potenciális lehetőségek, amelyek figyelembevételével pontosabb előrejelzést adhat egy szöveg komplexitásáról, vagy amelyek egy az eddig vizsgált tulajdonságokon kívüli egyéb nehézségre hívják fel a figyelmet. Włodzimierz Gruszczyński és csapata 2015-ben a lengyel nyelvre alkalmazva mutatott be néhány ilyen lehetőséget [12]. Ők többek között egy olyan módszert is használtak a szövegek nehézségének mérésére, ahol egy szövegből minden ötödik szót

eltávolítottak, majd egy kérdőíves kutatás során felmérték, hogy a kutatásban résztvevő személyek milyen pontossággal tudták a szöveget rekonstruálni. Véleményem szerint azonban a kérdőíves kutatás költségesebb, és mivel hatékonysága a kutatásban résztvevő személyek egyéni tulajdonságaitól is függ, nem túl megbízható. Ezért érdemes lehet megvizsgálni egy hasonló kutatás elvégzésének lehetőségét automatizált módszerekkel. A nagy méretű korpuszokból generált n-grammok listái lehetőséget adnak arra, hogy bizonyos szavak együttállásának gyakoriságát vizsgáljuk. Ha feltételezzük, hogy a gyakoribb n-grammok könnyebben érthető szerkezetek, akkor ilyen módon egy megfelelő program, amely egy hiányos szöveget kapna bemenetül, meg tudná állapítani, hogy a szöveg helyes kiegészítéséhez használt n-gramm mennyire gyakori, és ezen keresztül az adott szöveg mennyire érthető.

Gruszczyńskiék felvetettek néhány további lehetőséget is. Tanulmányukban például nem csak az adott szövegben szereplő szavak paramétereit vizsgálták, hanem, amennyiben lehetséges volt, az adott szó szótári alakra visszavezetett alakjának paramétereit is. A szerzők megjegyzik, hogy ez a vizsgálat csak akkor végezhető el eredményesen, ha egy megbízható program képes a szavakat a szótári alakjukra visszavezetni. Ez azonban sokszor problémába ütközhet, kiváltképp akkor, ha az ezt a feladatot végző rendszer nem képes a kontextus elemzésére, mert például homográfok esetén nem tud egyértelmű döntést hozni. Ezért ez a megoldás valószínűleg nem túl megbízható. Egy másik, szintén szavakon mérhető szempont a szófajok gyakorisága lehetne. Ehhez szintén szükség van olyan megbízható eszközre, amely képes a kontextus és a szóalak alapján meghatározni az adott szó szófaját. Így lehetséges lenne azonosítani olyan szófajokat, amelyek csak bizonyos szövegtípusokban gyakoriak, vagy amelyek bizonyos komplexitású szövegre jellemzőbbek. Amennyiben azonban nem áll rendelkezésre a szófajok megbízható azonosításához megfelelő program, addig ez a mérés sem végezhető el eredményesen.

Láthatjuk tehát, hogy érdemes foglalkozni a most vizsgálnál komplexebb szempontokkal is, azonban ezek sokszor speciális feltételeket igényelnek. Amíg ezek a feltételek nem tudnak kellő mértékben megvalósulni, addig ezekkel a szempontokkal nem lehet hatékony méréseket végezni.

7. A scriptek használata

A számításokat egy a Windows Subsystem for Linux alrendszer alatt futó Debian GNU/Linux 12 operációs rendszeren futtattam. A scriptek futtatásához Python 3.11 és Perl 5.36 környezeteket használtam. Az optimalizációt végző scriptek használatához szükség van a Numpy, a Pandas és a Scikit-learn Python könyvtárak telepítésére. A szövegfeldolgozást és a számítások elvégzését bash scriptek fogják össze. A feldolgozni kívánt szövegeknek az *input* elnevezésű könyvtárban kell lenniük, .txt kiterjesztéssel, egyszerű szöveg formátumban, UTF8 karakterkódolással, Unix-stílusú sortörésekkel. A scripteket egy az *input* könyvtár mellett található *scripts* könyvtárnak kell tartalmaznia.

Három egymásra épülő feladatot látnak el ezek a scriptek.

Az első feladat az ismert komplexitású szövegek feldolgozása, azok mindegyikéből a korábbi pontokban már ismertetett ASL, ASW, DWP stb. paraméterek kinyerése és táblázatba rendezése. Ezt a folyamatot a *start_statistics.sh* nevű script vezérli.

Első lépésben a *clean.sh* elnevezésű script kitisztítja a környezetet, majd az *s1.sh* script kerül meghívásra, amely elvégzi az *input* könyvtárban található .txt kiterjesztésű fájlok mondatra vágását, és az eredményt az *out* könyvtárba menti, .s kiterjesztéssel. Az eredményül kapott fájlokban egy sorban pontosan egy mondat található, így a későbbiek során elegendő újsor karakterek *\n* mentén szétválasztani a szöveget, ha mondatokkal dolgozunk.

Az *s2.sh* script a .s kiterjesztésű állományokat bemenetként fogadva a mondatokból egy kitisztított szólistát készít a *sub.pl* perl script segítségével, az eredményt pedig az *out* könyvtárba menti .words kiterjesztéssel.

A .words kiterjesztésű fájlok soronként pontosan egy szót tartalmaznak, így a szavakkal dolgozó számítások esetében könnyen szavakra lehet bontani a szöveget.

Ezután a *s3.sh* script hívódik meg, amely a bemenetként kapott szövegfájl mondatait véletlenszerű sorrendben adja vissza. Ez szükséges ahhoz, hogy a fájlok későbbi feldarabolása után az egyes részletek egymáshoz közelebbi mérési eredményeket adjanak.

A következő lépésben az *s4.sh* script hívódik meg, amely a véletlenszerű sorrendben álló mondatokat tartalmazó fájlok darabolását indítja el. Ezután a darabolt fájlok a *scripts* könyvtár melletti *parts* könyvtárban lesznek elérhetők.

Az előkészítési folyamatot a *s5.sh* script zárja, amely a korábbi *s2.sh* scripthez hasonlóan kitisztított szólistákat készít, azonban most a feldarabolt fájlokból. Így a statisztikai script a részekhez társított kitisztított szólistákat tudja használni a statisztikák elkészítéséhez.

Ennek a feladatnak a végén a *combined_statsv2.py* script hívódik meg, mely paraméterül kapja a mondatokra illetve szavakra vágott *.s* és *.words* kiterjesztésű állományokat, illetve a leggyakoribb szóalakot tartalmazó listákat (*1000.txt*, *2000.txt*, *5000.txt stb.*). Ezekből az információkból a *combined_statsv2.py* kiszámítja az adott szövegre jellemző paramétereket és az *out_statistics.txt* nevű táblázatba menti őket.

A második feladat a modell optimalizálása. Ezt a folyamatot a *start_optimizer.sh* nevű script vezérli. A script első lépésként a feldarabolt fájlokból készült statisztikát tartalmazó fájl alapján meghatározza az optimalizálást végző script számára szükséges *training.txt* és *validation.txt* állományokat. A *training.txt* fájl tartalmazza a feldarabolt fájlok közül a *part1*-től a *part10*-ig megjelölt fájlok statisztikai adatait, a *validation.txt* pedig a *part11*-gyel jelölt fájlok statisztikai adatait. Ezen felül a scriptnek szüksége van paraméterek egy listájára, amelyek az adott lineáris modellt alkotják. A script a *scikit-learn* nevű gépi tanulási Python csomag segítségével meghatározza az adott paraméter listához tartozó optimális modellt, illetve megméri ennek átlagos négyzetes hibáját a validációs halmazon.

Végül a harmadik feladat az ismeretlen komplexitású szövegek komplexitásának meghatározása. Ezt a folyamatot a *start_complexity.sh* nevű script vezérli. Először meghatározzuk az adott szövegekhez tartozó paramétereket ugyanúgy, ahogy azt már *start_statistics.sh* esetén ismertettük, majd a paraméterek és az optimális modellek (*FKG_NEW_FICTION*, *DC_NEW_FICTION*, *DC_NEW_FICTION2*) ismeretében meghatározzuk az egyes szövegek komplexitását, és ezeket sorba állítjuk komplexitás szerint. Az így kapott output állományok a következők lesznek: *out_complexity_dc.txt*, *out_complexity_dc_ext.txt*, *out_complexity_fkg.txt*.

Összegzés

A fenti kísérletek alapján látható, hogy a szövegek nehézségének meghatározása nem egyszerű feladat. Számos úton el lehet indulni, és számos paramétert figyelembe lehet venni a képletek megalkotásakor. A megfelelő megoldás megtalálásához érdemes az előzőleg már elkészített módszerekhez nyúlni, és azokat továbbfejleszteni. Fontos megjegyezni azonban, hogy

mivel a szövegek rendkívül sokfélék lehetnek, olyan képletet, amely minden szövegre egységesen alkalmazható, nem lehetséges megadni, mert a tanítási adatok nagyban befolyásolják a képlet hatékonyságát. Problémát jelent a képletek elkészítésekor az is, hogy túl kevés olyan szöveg áll rendelkezésre, amely pontosan hozzárendelhető egy bizonyos iskolai évfolyamhoz. Mivel nincs egységes elérhető lista, ezért csak kevésbé hivatalos forrásokhoz nyúlhatunk ilyen listákért, és ezek is csak korlátozott számú könyvet tartalmaznak. Emellett gyakran a listákon egy szöveg többször, több évfolyamnál is szerepel, ami tovább nehezíti a pontos besorolást. Dolgozatomban egy olyan képletet és eszköztárat javasolok, amely kézi előválogatást követően képes lehet egy megközelítőleges összehasonlítást adni, és meghatározni, hogy nyelvtanulók számára milyen szövegek lehetnek könnyebben olvashatók. A vizsgált irodalmi szövegek esetében a legpontosabb módszernek a Dale-Chall index egy módosított változata tűnik. Azonban az ezt a képletet adó modell hibaértékei is viszonylag magasak a szűkebb stílusstartományban mozgó szövegek kiértékeléséhez képest. Egy pontosabb módszer kidolgozásához érdemes lehet megvizsgálni, hogy milyen hatékonyságot lehetne elérni komplexebb paraméterek figyelembevételével, azonban ennek megvalósítása több időt és komplex eszközöket igényel. A módszer ilyen irányú továbbfejlesztése egy további kutatás tárgya lehet.

Bibliográfia

- [1] ÜVEGES I.: (2022). Közérthetőség mint osztályozási probléma (?) – gépi tanulási kísérlet kézzel címkézett korpuszon. In G. BEREND, G. GOSZTOLYA, V. VINCZE (Eds.), *XVIII. Magyar Számítógépes Nyelvészeti Konferencia 2022*. Szegedi Tudományegyetem, 619–633 p.
- [2] Office of the Federal Register, National Archives and Records Administration. Public Law 111 – 274 – Plain Writing Act of 2010. [Government]. U.S. Government Printing Office.
<https://www.govinfo.gov/app/details/PLAW-111publ274>
- [3] BOGDÁN P.: (2024). Angol nyelvű olvashatósági formulák magyar nyelvi adaptálásának lehetséges irányai. *Iskolakultúra*, 34 (1), 69–77 p. <https://doi.org/10.14232/ISKKULT.2024.1.69>
- [4] GUNNING, R.: *The technique of clear writing*. McGraw-Hill, 1968.

- [5] ОБОРНЕВА, И. В.: Автоматизация оценки качества восприятия текста. *Вестник Московского городского педагогического университета*. Серия: Информатика и информатизация образования, 2005, 5, 86–91 p.
- [6] SOLOVYEV, V.; IVANOV, V.; SOLNYSHKINA, M.: Assessment of Reading Difficulty Levels in Russian Academic Texts: Approaches and Metrics. 1 Jan. 2018 *Journal of Intelligent & Fuzzy Systems* 34 (2018): 3049–3058 p.
- [7] MyStem. Программа MyStem производит морфологический анализ текста на русском языке. URL: <https://yandex.ru/dev/mystem/>
- [8] Списки на лето: <https://www.labirint.ru/top/summer-book/>
- [9] Orosz nyelvű szövegek mondatokra szegmentálása (Perl script): https://www.wolosz.hu/russian_segmentation.html
- [10] Szavakból álló n-gramok gyakorisága orosz szövegekben: https://www.wolosz.hu/russian_n-grams.html
- [11] <https://github.com/bpisch/russian-text-complexity>
- [12] GRUSZCZYŃSKI, W. et al. (2015). W poszukiwaniu metody automatycznego mierzenia zrozumiałości tekstów informacyjnych [Információs szövegek érthetősége automatizált mérési módszerének nyomában]. *Poradnik Językowy*, 2015 (2), 9–22 p.