HunSimpleNews: Az első autentikus magyar nyelvű szövegekből álló szövegegyszerűsítési korpusz

Prótár Noémi[1], Nemeskey Dávid Márk[1,2]

[1]Eötvös Loránd Tudományegyetem
[2]Digitális Örökség Nemzeti Laboratórium
protarnoemi@student.elte.hu
nemeskey.david@btk.elte.hu

Kivonat A tanulmány az első autentikus magyar nyelvű szövegekből álló szövegegyszerűsítési korpuszt mutatja be. A korpusz 2832 darab könnyen érthető (egyszerűsített) és standard nyelvi szövegpárból áll, melyet a PannonRTV honlapjáról gyűjtöttünk. A tanulmányban bemutatjuk a korpusz összeállításának és minőségellenőrzésének folyamatát, kitérünk a korpusz két domainje közötti kvantitatív különbségekre és hasonlóságokra, végül a korpusz alapján felmérjük az általunk elérhető, magyarul tudó nagy nyelvmodellek szövegegyszerűsítési képességét in-context tanítási környezetben. Cikkünkben kimutatjuk, hogy az egyszerűsített szövegek és a standard nyelvi szövegek között lényeges különbségek vannak, mind az általuk használt szavak varianciájában, mind pedig a mondatok hosszában. Eredményeinkből láthatóvá válik, hogy a korpusz használata one-shot és few-shot tanítási környezetben is javítja a modellek egyszerűsített kimenetét a zero-shot eredményeinkhez képest.
Kulcsszavak: szövegegyszerűsítés, korpusz, könnyen érthető kommunikáció, párhuzamos korpusz

1. Bevezetés

A szövegegyszerűsítés[1] (text simplification) általános célja az, hogy különböző, standard nyelven íródott szövegeket érthetővé tegyen nem prototipikus befogadók számára. Prototipikus befogadók alatt a cikkben a felnőtt, ép, anyanyelvi beszélőket értjük, így tehát nem prototipikus befogadók például a gyerekek, a nyelvtanulók, funkcionális analfabéták, illetve olyan testi vagy értelmi fogyatékos személyek, akiknek fogyatékossága hátráltatja őket egy adott, standard nyelvű diskurzus megértésében (demensek, afáziások, nyelvelsajátítás előtti halláskárosodásuk volt, stb.) (Maaß, 2020). Az tehát, hogy szövegeket bizonyos befogadók

[1] A gyógypedagógiai szakirodalom az egyszerűsített szövegeket általában magyarul a *könnyen érthető szöveg* vagy *könnyen érthető kommunikáció* terminussal illeti. Mivel azonban az informatikai hagyomány ezen szövegek előállítására *text simplification*-ként hivatkozik, így ez a tanulmány is konzekvensen a *szövegegyszerűsítés* terminust használja, az egyértelműség és a kurrens diskurzusba való beilleszkedés miatt.

számára egyszerűsítünk, nagyban megkönnyíti, vagy alapjában lehetővé teszi azt, hogy mindennapi diskurzusokban részt vehessenek, valamint azt is, hogy a megfelelő információk eljussanak hozzájuk, és azok önállóan feldolgozhatóvá váljanak számukra.

Kiemelendő, hogy a szövegegyszerűsítés nem ekvivalens sem a szövegösszefoglalással, sem pedig a különböző szövegek közérthetővé tételével. Ugyan az egyszerűsített szövegekben lehetnek olyan részletek, amelyeket összefoglalunk a könnyebb megérthetőség érdekében, a szövegösszefoglalásnak a célja a szövegek hosszának csökkentése a felesleges információk elhagyásával (Shardlow (2014), Alva-Manchego és mtsai (2020)). A szövegegyszerűsítés ezzel szemben a szöveg bonyolultságának redukciójára törekszik: amennyiben ez azt igényli, hogy bizonyos fogalmakat megmagyarázzon, akkor ezzel gyakran él – így tehát az egyszerűsített szövegek akár hosszabbak is lehetnek, mint a forrásszövegeik. A közérthető szövegek pedig prototipikus befogadók számára tesznek érthetővé olyan szövegeket, amelyek szakspecifikusak – a könnyen érthető, azaz jelen tanulmányban *egyszerűsített*ként jellemzett szövegek ezzel szemben standard nyelvi, hétköznapi témákat érintő szövegeket tesznek érthetővé nem prototipikus befogadók számára.

A legtöbb szöveg leegyszerűsítése manuális módszerekkel történik – ez azonban időigényes és nagy humán erőforrást kíván. Így a szövegegyszerűsítés témaköre a nyelvtechnológiai jellegű kutatásokban is előtérbe került az utóbbi években. Jelen cikk az első olyan, magyar nyelvű hírszövegpárokból álló szövegegyszerűsítési korpuszt mutatja be, amely autentikus, tehát nem fordított, hanem eredetileg is magyarul írt adatokból áll. A korpusz kutatási célra szabadon elérhető a Hugging Face Hubon[2]. Cikkünkben először áttekintjük a szövegegyszerűsítés eddigi és jelenlegi állását (2), ezután bemutatjuk a korpusz összeállítását (3), majd ismertetjük a korpuszon végzett, a két szövegdomaint összehasonlító vizsgálatok eredményeit (4), végül bemutatjuk a korpusz használatával, one- és few-shot tanítási környezetben végzett vizsgálataink eredményét (5). A cikket egy rövid összefoglalás zárja (6).

2. Kapcsolódó irodalom

2.1. A szövegegyszerűsítés megközelítései

A nyelvtechnológiai jellegű, szövegegyszerűsítéssel foglalkozó kutatások legnagyobb része teljes szövegek egyszerűsítése helyett gyakran csak mondatok egyszerűsítésével foglalkozik. Ennek főképp az az oka, hogy az elérhető, nagyobb mennyiségű adatok száma minden nyelven kevés, illetve az egyszerűsített szövegek gyakran nem standard nyelvi szövegekből jöttek létre, hanem azokkal párhuzamosan, vagy tőlük függetlenül – így a szövegek nem felelnek meg egymásnak teljesen. Emellett megemlítendő, hogy a szövegeket általában automatikusan párosítják össze – ez mondatok esetében könnyebb, hiszen mondatok automatikus párosítására, hasonlóságuk mérésére viszonylag régóta vannak elérhető, jól

[2] `https://huggingface.co/datasets/ELTE-DH/HunSimpleNews`

használható rendszerek, (hosszabb) szövegek esetében azonban ez nem minden esetben magától értetődő.

Így tehát már az első modern, neurális hálót használó kutatások is a mondatok egyszerűsítését tartották céljuknak angol nyelven (Nisioi és mtsai, 2017). Ezt több nyelvre követték hasonló korpuszok és próbálkozások: a MultiSimV2 korpuszában (Ryan és mtsai, 2023) például összesen tizenhárom nyelv szerepel (arab, baszk, brazil portugál, dán, angol, francia, német, olasz, japán, orosz, szlovén, spanyol és urdu), melyek nagyrészt mondatszintű egyszerűsítésekből állnak. Kiemelendő, hogy több olyan nyelv is van, amelyre csak mondatszintű egyszerűsítések érhetőek el: ilyen például az arab (Khallaf és mtsai, 2022), a dán (Klerke és Søgaard, 2012), a szlovén (Gorenc és Robnik-Šikonja, 2022) és az urdu (Qasmi és mtsai, 2020). A dokumentumalapú párhuzamos korpuszok esetében pedig sokszor problémát okoz a korpusz nem nyilvános mivolta, vagy az adathalmaz kis mérete: az olasz Teacher és Terence korpuszok például mindössze 18 és 32 szövegpárból állnak (Brunato és mtsai, 2015).

Magyar nyelvre eddig egy párhuzamos korpusz jött létre (Prótár és Nemeskey, 2023), ez azonban egy fordított korpusz, melynek alapját az angol Simple English Wikipediából automatikus eszközökkel összeállított anyag adta (Zhu és mtsai, 2010). Ez a korpusz, éppúgy, mint a legtöbb a nemzetközi trendekben, szintén mondatpárokból áll. A szakirodalom azonban gyakran nem vet számot azzal, hogy a szövegek egyszerűsítése nem oldódik meg a mondatok izolált egyszerűsítésével: ugyan több cikk is foglalkozik azzal, hogy milyen műveleteket kell elvégezni egy mondaton belül ahhoz, hogy azok megfelelően egyszerűek legyenek (pl.Alva-Manchego és mtsai (2017)), és habár valóban az (elemi) mondatok a megnyilatkozások feldolgozásának alapvető egységei (Tátrai, 2017), ezek a metódusok azonban nem számolnak a szövegkonstruálás mondatstruktúrán felüli alapvető elemeivel, például az információk sorrendjének a meghatározásával, azok elosztásával, a kontextualizálás folyamatával stb. Nem számol emellett azzal sem, hogy a szövegértelem nem kompozicionális jellegű, azaz nem egyenlő csupán a mondatok értelmének egymás után tett vagy összeadott halmazával (vö. Tolcsvai Nagy (2006)) – így tehát a mondategyszerűsítés nem tehető egyenlővé a szövegegyszerűsítéssel, és szövegek egyszerűsítése nem megvalósítható a szöveg mondatainak egymás utáni leegyszerűsítésével. Emellett nem hagyhatjuk figyelmen kívül azt sem, hogy az emberi szövegegyszerűsítés sem mondatszinten, a mondatok egymás után való egyszerűsítésével, hanem szövegszinten zajlik, így tehát hasznosnak tűnik az automatikus egyszerűsítési kísérleteket is a szövegszintű egyszerűsítés felé közelíteni. Jelen tanulmány a hosszabb, kohezívebb szövegek egyszerűsítésének irányába kívánja tehát megtenni az első lépéseket magyar nyelven.

2.2. Szövegegyszerűsítés nagy nyelvi modellek segítségével

A korai automatikus szövegegyszerűsítési módszertanok főleg lexikai, azaz szószintű egyszerűsítéssel foglalkoztak (pl. Bott és mtsai (2012), Keskisärkkä (2012)) – azonban a transformer architektúra (Vaswani és mtsai, 2017) megjelenése a szövegegyszerűsítésre is hatással volt: az első ezen alapuló mondategyszerűsítő

modellt Nisioi és mtsai (2017) mutatták be. A nagy nyelvi modellek megjelenése további előrelépést jelentett a szövegegyszerűsítésben: Feng és mtsai (2023) kutatásai szerint a nagy nyelvi modellek teljesítménye meghaladja a mondategyszerűsítés kurrens modelljeinek legjobbjait, sőt, a teljesítményük az emberi egyszerűsítéshez is mérhető. Ezzel ellentétben Padovani és mtsai (2024) eredményei alapján egy megfelelően finomhangolt BERT-modell (Devlin és mtsai, 2019) jelentősen jobban teljesít az automatikus metrikák tekintetében, mint az általuk használt ChatGPT-3.5 ((Brown és mtsai, 2020) alapján). Jelen cikk szintén ehhez az emergens kutatási módszertanhoz kíván kapcsolódni, azonban nem a mondat-, hanem a szövegegyszerűsítés terén. Nem csupán a szöveghosszúság és az ezzel járó komplexitás oldaláról kínál újdonságot, hanem a magyar nyelvű szövegegyszerűsítéssel is: ugyan többnyelvű szövegegyszerűsítésre készültek már benchmarkok (Ryan és mtsai, 2023), ezek a magyar nyelvet nem tartalmazták.

3. A korpusz összeállítása

A korpusz anyaga a PannonRTV szabadkai híroldal cikkeiből származik. A kérésünkre a portál üzemeltetői valamint a cikkek írója tájékoztatott minket, hogy a hírportál könnyen érthető szövegeit egy személy írja a könnyen érthető irányelvek alapján (ld. pl. (Europe, 2017)).[3] Ezen a híroldalon kívül nincs tudomásunk olyan magyar nyelvű forrásról, amely megfelelő mennyiségben és minőségben állítana elő párhuzamos, vagy párhuzamosítható, műfaji szempontból viszonylag egységes standard nyelvi–könnyen érthető szövegpárokat magyar nyelven. Magyarul ezen kívül nagyon kevés, és nagyon különböző műfajú könnyen érthető szöveg jelent meg, amelyeknek standard nyelvi párja is van: ilyen például a Nógrádi Gergely és Bodonyi Panni által írt János vitéz-átdolgozás, mely Petőfi művének első három énekét ülteti át egyszerűen érthető és könnyen érthető formába. (Petőfi és mtsai, 2022)

A korpusz összeállításához először a Scrapy (Kouzis-Loukas, 2016) segítségével legyűjtöttük az oldalon megjelentetett cikkeket, az első egyszerűsített cikk megjelenésének időpontjáig. Mivel a híroldalon megjelenő egyszerűsített cikkek nincsenek semmilyen módon összekötve a standard nyelvi párjukkal, így az összes standard nyelvi hírszöveget le kellett menteni, hogy azokat összepárosíthassuk a könnyen érthető szövegekkel. A legyűjtések során összesen 4364 darab egyszerűsített hírszöveget, és több, mint 66 500 standard nyelvi szöveget mentettünk le. A gyűjtés két fázisban, 2023 novemberében és 2024 augusztusában zajlott, utóbbi alkalommal csak azokat a szövegeket mentettük le, amelyeket korábban nem.

A HTML-fájlok lementése után a szövegekből kinyertük a releváns adatokat, melyek a következők voltak: a cikk címe, leadje, szövege, a hozzátartozó címkék; a

[3] A korlátozott szerzői kör természetesen bizonyos mértékig befolyásolja a korpusz vizsgálhatóságát és felhasználhatóságát, azonban mivel semmilyen más, még csak közelítő méretű szöveganyag sem elérhető magyar nyelven könnyen érthető szövegekből, így ezeket használtuk fel.

cikk URL-je, közlésének dátuma, valamint a cikk kategóriája. Az utóbbi az oldalon megjelenő tizenkét magyar nyelvű kategóriát tartalmazta (politika, világ, régió, gazdaság, társadalom, természetvédelem, önkormányzatok, kultúra, kékfény, sport, színes, KÉK-hírek), amelybe a cikkek megjelenésükkor a szerkesztők vagy az újságírók által besorolásra kerülnek. Az adatokat egy JSONL-struktúrában tároltuk el.

A korábban említett standard és egyszerűsített szövegek közötti hivatkozások hiánya miatt a szövegek párosítására szükség volt, azonban ez a nagyszámú szövegmennyiség miatt nem valósulhatott meg manuálisan. A szövegek hasonlóságát Doc2Vec (Řehůřek és Sojka, 2010) és Sentence-BERT (Reimers és Gurevych, 2019) beágyazásokkal vizsgáltuk, egy olyan további kikötést téve, hogy a standard nyelvi szöveg az egyszerűsített szöveghez képest 0–3 nappal előbb kellett, hogy keletkezzen. A szövegbeágyazásokat lemmatizált és nem lemmatizált adatokon is elvégeztük, majd ezeket összevetettük, ahol pedig a két eljárás különböző eredményt adott, manuálisan döntöttünk a helyes párosításról. Végül a kétfajta beágyazási módszerrel keletkezett eredményeinket is összevetettük, és ugyanígy jártunk el.

Így összesen 3629 lehetséges párt kaptunk.

Az elvégzett párosításokat az ELTE Digitális Bölcsészet Tanszékének négy annotátora ellenőrizte manuálisan.

Az annotáció alatt három csoportba sorolhatták a lehetséges párokat:

- Az összepárosított két szöveg **valóban pár**: a két szöveg ugyanarról szól, és az egyszerűsített szöveg nem tartalmaz olyan kiegészítő információkat, amelyek nincsenek benne a standard nyelvi szövegben. Ebből összesen **2833** darab volt a korpuszban.
- Az összepárosított két szöveg **nem pár**, azaz a két szöveg nem ugyanarról szól. Ebből összesen **175** darab volt a korpuszban.[4]
- Az összepárosított két szöveg **ugyanarról szól, de az egyszerűsített szöveg tartalmaz olyan kiegészítő információkat, amelyek nincsenek benne a standard nyelvi szövegben**. Ebből összesen **621** volt a korpuszban.

Jelen cikkben a párként kategorizált cikkpárokkal dolgoztunk. A plusz információt tartalmazó cikkek esetleges javítása további munka tárgyát képezheti.

4. Korlátok és lehetőségek – tanulságok a korpusz szerkezetének vizsgálatáról

A korpuszon a few-shot tanítás előtt elvégeztünk néhány, a korpusz szerkezetét érintő vizsgálatot. A vizsgálatokkal célunk a korpusz használati lehetőségeinek és

[4] A nem párként elkönyvelt szövegek egyértelműen nem voltak párok – ez például a következő, helytelenül összepárosított, de az annotátorok által nem párnak ítélt szövegek címeiből is látszik:
Tisztújító közgyűlést tart májusban a VMSZ – 1 dinárral drágább az üzemanyag;
Szerbia legnagyobb vidámparkja várja majd a Palicson majálisozókat – Elektromos autókat fognak gyártani Szerbiában

korlátainak bemutatása volt, valamint az egyszerűsített és standard nyelvi szövegek jellemzőinek összevetése. A korpusz vizsgálata előtt megemlítendő, hogy a nyelvi anyag egy forrásból származik, így a magyar nyelvű egyszerűsített szövegek általános tulajdonságaival kapcsolatban csak nagyon óvatos megállapítások tehetők. Nem lehet figyelmen kívül hagyni azt sem, hogy mindkét korpuszkomponens a hírszövegek műfajába tartozik – ezáltal a nyelvi anyaggal kapcsolatos konklúziók ezen műfajhoz tartozó szövegekre lesznek leginkább alkalmazhatók. A fent felsorolt nehézségek ellenére a szövegkorpusszal kapcsolatos vizsgálatokat elengedhetetlennek tartjuk: mivel a magyar nyelvű szövegegyszerűsítéssel kapcsolatban kevés vizsgálat készült, valamint kevés egyszerűsített szöveganyag lelhető fel magyarul, így szükségesnek tartjuk azt, hogy a vizsgálatok reflektáljanak a már meglévő szöveganyagra, nyelvhasználatra, és ne kizárólag a priori megközelítéssel viszonyuljanak a szövegegyszerűsítés problematikájához.

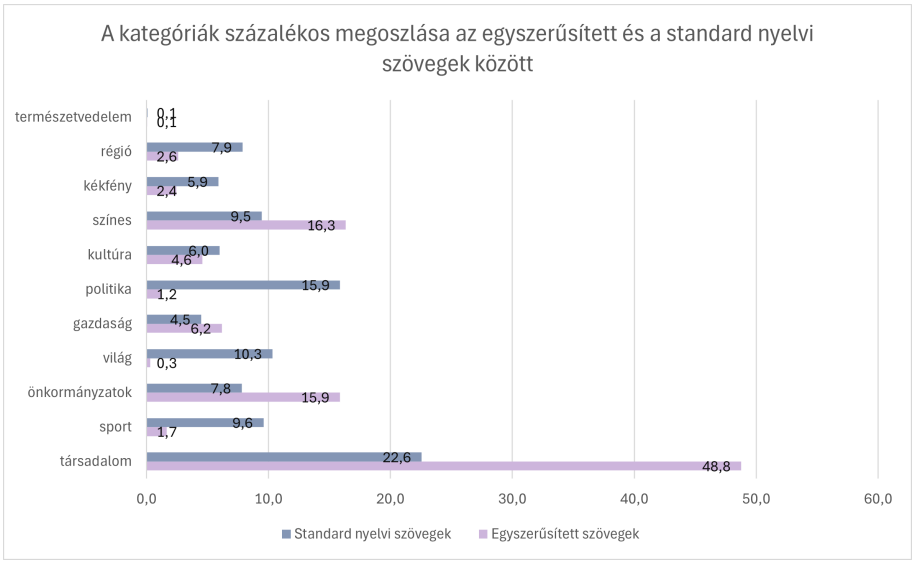
Ahogy az korábban említésre került, a cikkek a cikkírók vagy az oldal szerkesztői által kategóriákba lettek rendezve. A párosítás után az egyszerűsített cikkek „megörökölték" az adatszettben a standard nyelvi párjuk kategóriáját, [5] így vizsgálhatóvá vált az, hogy azok milyen kategóriájú cikkekből lettek előállítva, valamint az is, hogy az egyszerűsített szövegek kategóriái mennyiben képezik le az összes hírszöveg kategóriájának megoszlását, azaz hogy vannak-e preferált vagy elhanyagolt témakörök az egyszerűsített hírszövegekben.

Az 1. ábra ezt a megoszlást mutatja be, azaz az egyszerűsített szövegek és a standard nyelvi szövegek kategóriáinak százalékos megoszlását a saját korpuszukon belül. A standard nyelvi szövegek esetében az összes, az egyszerűsített cikkek keletkezésének időszakában közzétett cikket használtuk a százalékos adatok meghatározásához.

Jól látható, hogy a megoszlások erősen különböznek a két csoport esetében: ugyan mindkét csoport esetében a *társadalom* kategória volt a leggyakrabban használt, azonban az egyszerűsített szövegek esetében ebbe a kategóriába a szövegeknek közel a fele sorolódott, míg a standard nyelvi szövegek esetében csupán körülbelül a negyedét tették ki az ide kategorizált elemek. A többi szöveg esetében pedig még drasztikusabb a különbség: a *politika*, *sport* vagy a *világ* kategóriák esetében erősen alulreprezentáltak az egyszerűsített szövegek a standard nyelviekhez képest, míg az *önkormányzatok* és *színes* kategóriák az egyszerűsített szövegek korpuszában kaptak nagyobb szerepet. Ebből látható, hogy az egyszerűsített szövegek írói alapvetően más, valószínűleg a befogadóhoz közelebb álló, számára egyszerűbben beazonosítható témákat részesítenek előnyben.

A korpuszméretek és a különböző lemmák számossága alapján az 1. táblázatból jól látszik, hogy az egyszerűsített korpusz jelentősen kisebb szótárral operál: míg a nagysága körülbelül a 80%-a a standard nyelvinek, a különböző lemmák száma nagyjából feleannyi. Úgy tűnik tehát, hogy az egyszerűsített szövegek a repetitivitásra törekednek a változatosság helyett: a könnyebb érthetőség miatt a szinonimák helyett ugyanazokat a szavakat használják ugyanazon dolog

[5] Az öröklésre azért volt szükség, mert az egyszerűsített szövegeknek külön egységes kategóriájuk volt az oldalon belül; feltételeztük azonban, hogy mivel a két szöveg alaptémája megegyezik, így a kategóriabesorolásuk is megegyezne.



1. ábra: A kategóriák megoszlásának százalékos összehasonlítása az egyszerűsített szövegek között és az ugyanebben az időszakban keletkezett összes többi, standard nyelvi hírszöveg esetében.

	Standard	**Egyszerűsített**
Összes lemma	574 808	487 728
Lemmatípusok száma	35 972	18 778
Hapaxok száma	18 005	8227
Egy lemma átlagos előfordulása	15,98	25,98
Egy lemma átlagos előfordulása hapaxok nélkül	30,99	45,45

1. táblázat. A két korpuszdomain lemmatizált adatainak megoszlása.

vagy fogalom megnevezésére. A nagy mennyiségű hapax, illetve a különböző tulajdonnevek előfordulásai miatt az átlagos előfordulások száma csupán a fenti adatokkal együtt értelmezhető, önmagában nem – azonban ezek az adatok is arra mutatnak, hogy az egyszerűsített korpuszban a megnyilatkozó sokkal gyakrabban használja ugyanazokat a szavakat, mint a standard nyelvi korpuszban.

A 2. táblázatból kiolvasható, hogy a két korpusz nemcsak lemmaszinten, de mondatszinten is különbözik: a standard nyelvi hírszövegek esetében a mondatok általában több szóból állnak (tehát hosszabbak). A leadek mondatai tendenciózusan rövidebbnek tűnnek mindkét korpuszban, mint a szövegtörzs mondatai – ennek oka valószínűleg az, hogy a lead funkciója a hírszöveg összefoglalása, nem pedig az egyes aspektusok kifejtése. A táblázat adataiból leszűrhető, hogy az egyszerűsített szövegek mondatai alapvetően rövidebbek. Mivel a mondatok száma is kevesebb, így feltehető az is, hogy bizonyos információkat vagy informá-

	Szavak	Mondatok	Átlagos mondathossz
Standard nyelvi (csak szövegtest)	498 923	29 308	17,02
Egyszerűsített (csak szövegtest)	444 963	37 619	11,83
Standard nyelvi (szöveg + lead)	557 512	36 066	15,46
Egyszerűsített (szöveg + lead)	472 912	43 287	10,93

2. táblázat. A mondatok és a bennük lévő szavak aránya a két korpuszdomainben.

ciórészleteket elhagy az egyszerűsített szöveg. Az, hogy ez mikor történik meg, és a standard nyelvi szövegekhez képest milyen információk kerülnek elhagyásra vagy megmagyarázásra a szövegben, további, kvalitatív vizsgálatok tárgyát képezi.

5. A korpusz segítségével elért eredmények

A korpusz segítségével elsődlegesen négy olyan magyarul tudó utasítás-finomhangolt modellt teszteltünk, amelyek viszonylag könnyen hozzáférhetők: a zárt súlyú GPT-4o minit (Achiam és mtsai, 2024) API-n keresztül[6], a Llama 3.3 70 milliárd paraméteres, 4 bitre kvantált és a Llama 3.1 8 milliárd paraméteres változatait (Dubey és mtsai, 2024), valamint a Qwen 2.5 7 milliárd paraméteres változatát (Yang és mtsai, 2024; Qwen, 2024)) pedig saját GPU-n futtatva. A nyílt modellek elfutnak egy A100 40GB-s kártyán (a Llama 70B kettőn) 20-30 ezres kontextushosszal, a GPT-4o mini pedig megfizethető árazású. A négy fő modell mellett összehasonlításképpen kimértünk egy kis modellt, a Llama 3.2 3 milliárdos változatát, illetve a GPT-4o mini két alternatíváját, a GPT-4o-t és a Claude 3.5 Haikut (Anthropic, 2024) is. Ezek eredményei a korábbi Llama 3.1 70B-s méréseinkkel együtt a B függelékben találhatók.

A modelleket a SARI- (Xu és mtsai, 2016) és a BLEU-pont (Papineni és mtsai, 2002) alapján vetettük össze. A SARI pontszámot a Huggingface *evaluate* könyvtárával számoltattuk ki (von Werra és mtsai, 2020), míg a BLEU-pontszámokat a *sacreBLEU* könyvtár segítségével (Post, 2018). Mind a BLEU-, mind pedig a SARI-pontszám általánosan használt az egyszerűsített szövegek megfelelőségének mérésére: az előbbi n-gram-egyezéseket mér, míg az utóbbi a standard nyelvi szöveget és a referenciaszöveget veti össze a generált szöveggel, annak fényében, hogy a generált szöveg mennyire jól:

- ad hozzá a standard nyelvi szöveghez releváns információkat,
- törli ki a standard nyelvi szövegből a nem fontos elemeket,
- hagyja meg azt a szöveget érintetlenül, amit érintetlenül kell hagynia.

Mindkét pontszám esetében 0 és 100 közötti értéket vehet fel az egyszerűsített szöveg minősítése, ahol a 100 jelenti azt, hogy a referenciaszöveg és a generált szöveg teljes egészében megegyezik. Mindkét pontszám esetében a szövegek címét,

[6] 2024 november 20-án futtattuk.

leadjét és törzsszövegét egybevetve, egy egységként kezeltük őket. A mondatokra való lebontást és egyesével való értékelést nem tartottuk sem megvalósíthatónak, sem célravezetőnek, hiszen a szövegek értelmezése – ahogy amellett korábban is érveltünk – nem egyenlő az egyes mondatok értelmezésével. Természetesen ezen lebontás elkerülése miatt az eredményeink csak korlátozottan összevethetőek a többi, az automatikus szövegegyszerűsítést mondatszinten vizsgáló cikk eredményeivel.

Minden modell esetében ugyanazt a promptot használtuk, amely a zero-shot és a one- vagy few-shot megközelítések esetében a példák megadásában, valamint a szeparátorkarakterek felsorolásában különbözött. A promptban az egyszerűsítést adtuk meg a modelleknek feladatként, valamint a 4. részben bemutatott eredmények alapján összefoglaltuk az egyszerűsített szövegek néhány jellemzőjét: a túl hosszú mondatok lerövidítésére kértük, a bonyolult fogalmak elmagyarázására, valamint arra, hogy ugyanazt a fogalmat végig ugyanazzal a szóval illesse a szövegben, és ne használjon szinonimákat. A promptok megtalálhatók a függelékben (A).

A különböző vizsgálati esetekhez a cikkeket az írók által megadott kategóriánként szegmentáltuk, valamint a standard nyelvi és az egyszerűsített szövegek között karakteralapú Levenshtein-távolságot számoltunk a Levenshtein csomaggal (Bachmann, 2024). A távolság alapján ezután három egyforma számosságú csoportot hoztunk létre: az alacsony távolságúakba a távolságmetrika szerinti alsó harmadba, a közepesekbe a középsőbe, míg a felsőbe a felső harmadba tartozó szövegek kerültek. Minél nagyobb távolság volt a két szöveg között, feltételezésünk szerint annál több változtatást (információ-hozzáadást vagy -elhagyást) kellett eszközölni a két szövegverzió között.

A vizsgálati esetek a következők voltak:

1. zero-shot környezet (0-shot);
2. one-shot környezet, ahol a példa cikk-kategóriája nem egyezett meg a teszteset kategóriájával (1-shot);
3. one-shot környezet, ahol a példa cikk-kategóriája megegyezett a teszteset kategóriájával (1-shot + kat.);
4. three-shot környezet, ahol a példa cikk-kategóriája nem egyezett meg a teszteset kategóriájával, és a Levenshtein-távolság alapján randomizálva kerültek elemek be a példák közé (3-shot);
5. three-shot környezet, ahol a példa cikk-kategóriája nem egyezett meg a teszteset kategóriájával, és a Levenshtein-távolság alapján mindhárom csoportból egy-egy elem került be a példák közé (3-shot + Levenshtein);
6. three-shot környezet, ahol a példa cikk-kategóriája megegyezett a teszteset kategóriájával, és a Levenshtein-távolság alapján mindhárom csoportból egy-egy elem került be a példák közé (3-shot + kat. + Levenshtein);
7. three-shot környezet, ahol a példa cikk-kategóriája megegyezett a teszteset kategóriájával, és a Levenshtein-távolság alapján randomizálva kerültek elemek be a példák közé (3-shot + kat);

Minden modellt összesen ötször futtattunk le minden adatsorra, majd az eredményeket átlagoltuk. Ez alól kivétel a Claude és a GPT-4o; itt pénzügyi

okokból az elsőt csak egyszer futtattuk végig minden vizsgálati esetre, a másodikat pedig csak arra a két esetre, amin a GPT-4o mini legjobb és legrosszabb eredményeit érte el.

Modell / vizsgálati eset	BLEU	SARI	Perplexitás
GPT-4o mini			
1. (0-shot)	10,48 ± 6,52	42,49 ± 5,69	15,80 ± 5,76
2. (1-shot)	12,33 ± 6,96	43,99 ± 5,71	15,14 ± 5,39
3. (1-shot + kat.)	12,43 ± 6,88	43,74 ± 5,23	15,14 ± 5,31
4. (3-shot)	12,39 ± 6,75	43,89 ± 5,27	15,31 ± 5,44
5. (3-shot + Levenshtein)	**13,38 ± 7,40**	**44,58 ± 5,58**	14,85 ± 5,27
6. (3-shot + kat. + Levenshtein)	13,28 ± 7,17	44,36 ± 5,58	**14,58 ± 4,58**
7. (3-shot + kat)	12,48 ± 7,11	44,05 ± 5,78	15,06 ± 5,26
Llama 3.3 70B, kvantált			
1. (0-shot)	6,42 ± 4,19	38,59 ± 6,71	14,79 ± 5,36
2. (1-shot)	8,51 ± 5,22	40,62 ± 6,19	13,50 ± 4,68
3. (1-shot + kat.)	8,57 ± 5,06	40,88 ± 6,58	13,40 ± 5,01
4. (3-shot)	8,72 ± 5,21	41,02 ± 6,57	13,48 ± 5,03
5. (3-shot + Levenshtein)	**10,65 ± 5,98**	42,41 ± 5,91	**11,72 ± 4,06**
6. (3-shot + kat. + Levenshtein)	10,60 ± 6,05	**42,43 ± 6,19**	11,66 ± 3,91
7. (3-shot + kat)	8,45 ± 5,52	40,61 ± 6,88	13,57 ± 4,98
Llama 3.1 8B			
1. (0-shot)	6,32 ± 4,26	38,17 ± 6,22	25,11 ± 13,84
2. (1-shot)	6,65 ± 4,81	38,58 ± 6,29	23,57 ± 12,41
3. (1-shot + kat.)	6,51 ± 4,58	38,52 ± 6,08	23,39 ± 11,08
4. (3-shot)	6,58 ± 4,89	38,53 ± 6,17	23,86 ± 12,42
5. (3-shot + Levenshtein)	**7,65 ± 5,63**	39,27 ± 5,95	**19,97 ± 9,26**
6. (3-shot + kat. + Levenshtein)	7,52 ± 4,88	**39,49 ± 5,61**	20,32 ± 9,47
7. (3-shot + kat)	6,94 ± 5,11	38,93 ± 6,18	24,58 ± 12,06
Qwen2.5 7B			
1. (0-shot)	5,64 ± 4,62	37,53 ± 6,28	65,54 ± 45,84
2. (1-shot)	8,37 ± 6,20	39,63 ± 5,50	45,36 ± 29,16
3. (1-shot + kat.)	8,55 ± 6,29	39,69 ± 5,89	44,45 ± 29,07
4. (3-shot)	8,41 ± 6,08	39,93 ± 5,63	47,64 ± 30,27
5. (3-shot + Levenshtein)	9,95 ± 6,84	40,34 ± 5,65	38,32 ± 25,25
6. (3-shot + kat. + Levenshtein)	**10,17 ± 6,61**	**40,65 ± 5,44**	**37,05 ± 22,17**
7. (3-shot + kat)	8,53 ± 5,70	39,75 ± 5,54	44,33 ± 29,79

3. táblázat. A BLEU- és SARI-pontok, valamint a perplexitás megoszlása modellenként. A legjobb értékek minden oszlopban, minden modell esetében félkövérrel vannak jelölve. A jelölések követik a vizsgálati esetek leírásánál feltüntetettet: a *+ kat.* jelölés a kategóriaegyeztetést jelenti, a *+ Levenshtein* pedig a három Levenshtein-csoport figyelembe vételét veszik. A jelek hiánya az adott tulajdonság randomizált vagy figyelembe nem vett voltára utal.

A 3., a 4., az 5., a 6. valamint a 7. táblázatból jól látható, hogy minden modell esetében a 0-shot módszertan hozta a legalacsonyabb átlag- és mediánértéket is: ebből következtethetünk arra, hogy az összeállított korpusz használata few- és one-shot környezetben is segít az egyszerűsítésben. Továbbá 0-shot esetében több modell néha, a Claude viszont mindig hozzáfűzött egy bevezető mondatot a kimenethez (pl. „*Íme az egyszerűsített szöveg:*"). Az in-context tanulás ezt is korrigálta.

Úgy tűnik, hogy 3-shot esetben minden modellnél azok a példák voltak a legsikeresebbek, ahol a Levenshtein-hasonlóság szerinti csoportosítást figyelembe vettük – azaz a modellek jobban teljesítettek akkor, ha a standard nyelvi és az egyszerűsített szöveg közötti átalakításokra variábilis példákat láttak.

A modellek teljesítménye nagyobb szórást mutat: a legjobban a GPT-4o-mini teljesít a SARI-pontszámban, még a nála sokkal drágább GPT-4o-t is megelőzi kb. 2 ponttal (utóbbi $41,29 \pm 5,89$ SARI pontot ér el 0-shoton, és $42,70 \pm 5,85$ pontot 3-shoton különböző Levenshtein-távolságokkal). A 3.3-as Llama 70B verziója kvantálva a nagyobb OpenAI modellhez hasonló teljesítményt nyújt. A sort a kisebb modellek: a Llama 3.1 8B és a Qwen2.5 7B zárják további 2-3 pont hátránnyal.

A BLEU-pontszámok ezzel szemben lényegesen alacsonyabbak minden modell esetében. Ennek oka valószínűleg a magyar nyelv ragozó mivolta, illetve a feladat maga: a szövegegyszerűsítés megköveteli mind a mondatszerkezet, mind a szavak egyszerűbbre cserélését, ami megtöri a hosszabb n-gramokat. Ez utóbbi hatást erősíti, hogy minden cikkhez csak egy referencia van; alternatív egyszerűsítésekkel a pontszám valószínűleg magasabb lenne.

Megemlítendő, hogy a BLEU és a SARI ugyan gyakran korrelálnak, azonban nem mindig – egymástól erősen eltérő, magasabb és alacsonyabb eredmények több nyelv több korpuszán végzett méréseken is megfigyelhetők (Ryan és mtsai, 2023).

A 3. táblázatban feltüntettük a válaszok a PULI-GPT-3SX modell szerinti perplexitását is (Yang és mtsai, 2023). Ez a metrika a fenti kettővel szemben elsősorban nem az egyszerűsítés sikerességével, hanem a válasz nyelvi helyességével és „magyarságával" áll korrelációban. A referenciaválaszok átlagos perplexitása 11,09; ehhez legközelebb a magyar nyelvet hivatalosan nem támogató Llama 3.3 70B került, legtávolabb pedig a Qwen. Ez utóbbi nem meglepő annak fényében, hogy a nyelvtudása erősen hullámzó: nagyjából helyes mondatok váltakoznak teljesen értelmetlenekkel[7]. Emellett a szövegben gyakran találtunk kínai (olykor angol) szöveget, akár egy magyar mondat közepén is, melyekben leggyakrabban a szöveg egyszerűsítésére vonatkozó utasítások voltak találhatók.[8] Ebből kifolyólag magyar nyelvmodellezésre nem ajánlható. A Qwenhez hasonló, vagy még rosszabb eredményt ért el a Llama 3.2 3B, amely egy egészen meglepő, 600

[7] *Egy fiatal Szabadkán hősége a napon megszomba volt.* vagy
A házak falára terepolták a táblákat, hogy könnyebbé tessék a fájdalom utasításának olvasását.

[8] *Az üléseken is meg*酌取材料，简写文本并用简单词汇表达：— *Holnap a válságstáb ülése lesz.*

fölötti perplexitással rendelkező szöveget is generált, mely pár értelmes magyar szó után legnagyobbrészt értelmetlen betűsorokból állt.[9]

Egy adott modellhez tartozó perplexitásokat összehasonlítva kitűnik, hogy a prompt szerepe nagyon fontos nem csak az egyszerűsítés, hanem a szövegminőség miatt is. Minden modell a legrosszabb perplexitást a 0-shot esetben érte el. A kategória egyeztetése érdemben nem befolyásolta a perplexitást, viszont a Levenshtein-távolság-alapú válogatás akár 15–20 százalékkal is javított rajta. A modellek közül a magyar nyelvet hivatalosan is támogató GPT-4o mini perplexitását befolyásolta a tesztkörnyezet a legkevésbé, azonban itt is közel egy pont különbséget lehet elérni egy jobb prompttal.

A perplexitás azonban fenntartásokkal kezelendő: szúrópróbaszerűen összehasonlítva a GPT-4o mini és a kvantált 70B-s Llama által generált szövegeket, világosan látszik, hogy az utóbbi vét nyelvtani hibákat, amik a GPT-re nem jellemzőek.

Összességében elmondható, hogy a modellek között az előre sejthető sorrend alakult ki: a GPT-4o mini mind nyelvi, mind szövegegyszerűsítési szempontból a legjobb az általunk tesztelt modellek közül. A Llama 3.3 70B jó prompttal megközelíti, de a SARI tekintetében kis mértékben elmarad tőle. A sort a Llama 3.1 8B és a Qwen 2.5 7B zárja, utóbbi nehezen értékelhető nyelvi teljesítménnyel.

6. Összefoglalás

A cikk az első olyan magyar nyelvű szövegegyszerűsítési korpuszt mutatta be, amely autentikus, tehát eredetileg is magyarul írt adatokból álló szövegpárokat tartalmaz. A korpusz a PannonRTV híroldal szöveganyagából áll, melyek automatikus módszerekkel lettek összepárosítva. A korpusz kutatási célre szabadon hozzáférhető. A korpuszon először a korpusz két domainje közötti vizsgálatot végeztük el: itt az eredményekből egyértelművé vált, hogy az egyszerűsített szövegek átlagosan lényegesen rövidebb mondatokat használnak. Kimutattuk emellett, hogy az egyszerűsített szövegek ebben a korpuszban egy sokkal kisebb, majdnem feleakkora szótárral operálnak a szövegek megkonstruálásakor. Végül a korpuszt négy nagy nyelvi modell segítségével zero- és few-shot tesztkörnyezetekben próbáltuk ki, melynek során megállapítottuk, hogy a korpusz használata elősegíti a megfelelő egyszerűsített szövegek automatikus létrehozását. Ezek alapján érdemesnek tűnik további vizsgálatokat folytatni a magyar nyelvű szövegegyszerűsítéssel kapcsolatban, akár modellek finomhangolásával, akár az adatszett további bővítésével kapcsolatban.

[9] *Kannor mel on deix filtrborh vál Bone Bern dadFSka datas izsak Proapers vie toplyír Kft be menyel EstatyAll. eszet cleverno regalk Bslsn ln dansro ban Neros ours mamm st gef vistaVdre Communications Arg Kanas fant wasm Fcoris*

Hivatkozások

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., ...: GPT-4 technical report (2024), https://arxiv.org/abs/2303.08774

Alva-Manchego, F., Bingel, J., Paetzold, G., Scarton, C., Specia, L.: Learning how to simplify from explicit labeling of complex-simplified text pairs. In: Kondrak, G., Watanabe, T. (szerk.) Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers). pp. 295–305. Asian Federation of Natural Language Processing, Taipei, Taiwan (Nov 2017), https://aclanthology.org/I17-1030

Alva-Manchego, F., Scarton, C., Specia, L.: Data-driven sentence simplification: Survey and benchmark. Computational Linguistics 46(1), 135–187 (2020), https://aclanthology.org/2020.cl-1.4

Anthropic: The Claude 3 model family: Opus, Sonnet, Haiku (2024), https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf

Bachmann, M.: Levenshtein. https://rapidfuzz.github.io/Levenshtein/ (2024), python package

Bott, S., Rello, L., Drndarevic, B., Saggion, H.: Can spanish be simpler? lexsis: Lexical simplification for spanish. In: International Conference on Computational Linguistics (2012), https://api.semanticscholar.org/CorpusID:7939405

Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D.: Language models are few-shot learners (2020), https://arxiv.org/abs/2005.14165

Brunato, D., Dell'Orletta, F., Venturi, G., Montemagni, S.: Design and annotation of the first Italian corpus for text simplification. In: Meyers, A., Rehbein, I., Zinsmeister, H. (szerk.) Proceedings of the 9th Linguistic Annotation Workshop. pp. 31–41. Association for Computational Linguistics, Denver, Colorado, USA (Jun 2015), https://aclanthology.org/W15-1604

Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Burstein, J., Doran, C., Solorio, T. (szerk.) Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota (Jun 2019), https://aclanthology.org/N19-1423

Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., ...: The llama 3 herd of models (2024), https://arxiv.org/abs/2407.21783

Europe, I.: Információ mindenkinek - a könnyen érthető kommunikáció európai alapelvei (2017), https://www.inclusion-europe.eu/wp-content/uploads/2017/06/HU_Information_for_all.pdf

Feng, Y., Qiang, J., Li, Y., Yuan, Y., Zhu, Y.: Sentence simplification via large language models (2023), https://arxiv.org/abs/2302.11957

Gorenc, S., Robnik-Šikonja, M.: Slovene text simplification dataset SloTS (2022), http://hdl.handle.net/11356/1682, slovenian language resource repository CLARIN.SI

Keskisärkkä, R.: Automatic Text Simplification via Synonym Replacement. Szakdolgozat, Linköping University (2012), https://api.semanticscholar.org/CorpusID:16139324

Khallaf, N., Sharoff, S., Soliman, R.: Towards Arabic sentence simplification via classification and generative approaches. In: Bouamor, H., Al-Khalifa, H., Darwish, K., Rambow, O., Bougares, F., Abdelali, A., Tomeh, N., Khalifa, S., Zaghouani, W. (szerk.) Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP). pp. 43–52. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Hybrid) (Dec 2022), https://aclanthology.org/2022.wanlp-1.5

Klerke, S., Søgaard, A.: DSim, a Danish parallel corpus for text simplification. In: Calzolari, N., Choukri, K., Declerck, T., Doğan, M.U., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., Piperidis, S. (szerk.) Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12). pp. 4015–4018. European Language Resources Association (ELRA), Istanbul, Turkey (May 2012), http://www.lrec-conf.org/proceedings/lrec2012/pdf/270_Paper.pdf

Kouzis-Loukas, D.: Learning Scrapy. Packt Publishing Ltd (2016)

Maaß, C.: Easy Language - Plain Language - Easy Language Plus: Balancing Comprehensibility and Acceptability. Frank & Timme GmbH (January 2020)

Nisioi, S., Štajner, S., Ponzetto, S.P., Dinu, L.P.: Exploring neural text simplification models. In: Barzilay, R., Kan, M.Y. (szerk.) Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). pp. 85–91. Association for Computational Linguistics, Vancouver, Canada (Jul 2017), https://aclanthology.org/P17-2014

Padovani, F., Marchesi, C., Pasqua, E., Galletti, M., Nardi, D.: Automatic text simplification: A comparative study in italian for children with language disorders. In: Proceedings of the 13th Workshop on Natural Language Processing for Computer Assisted Language Learning (NLP4CALL 2024). vol. 211, pp. 176–186. Rennes, France (2024)

Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. p. 311–318. ACL '02, Association for Computational Linguistics, USA (2002), https://doi.org/10.3115/1073083.1073135

Petőfi, S., Nógrádi, G., Bodonyi, P.: János vitéz 1-3. ének. Egyszerűen érthetően / könnyen érthetően. Petőfi Irodalmi Múzeum és Kortárs Irodalmi Központ, Budapest (2022)

Post, M.: A call for clarity in reporting BLEU scores. In: Proceedings of the Third Conference on Machine Translation: Research Papers. pp. 186–

191. Association for Computational Linguistics, Belgium, Brussels (Oct 2018), https://www.aclweb.org/anthology/W18-6319

Prótár, N., Nemeskey, D.M.: huPWKP: A Hungarian text simplification corpus. In: Mitkov, R., Angelova, G. (szerk.) Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing. pp. 898–907. INCOMA Ltd., Shoumen, Bulgaria, Varna, Bulgaria (Sep 2023), https://aclanthology.org/2023.ranlp-1.97

Qasmi, N.H., Zia, H.B., Athar, A., Raza, A.A.: SimplifyUR: Unsupervised lexical text simplification for Urdu. In: Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S. (szerk.) Proceedings of the Twelfth Language Resources and Evaluation Conference. pp. 3484–3489. European Language Resources Association, Marseille, France (May 2020), https://aclanthology.org/2020.lrec-1.428

Qwen, T.: Qwen2.5: A party of foundation models (September 2024), https://qwenlm.github.io/blog/qwen2.5/

Řehůřek, R., Sojka, P.: Software Framework for Topic Modelling with Large Corpora. In: Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks. pp. 45–50. ELRA, Valletta, Malta (May 2010), http://is.muni.cz/publication/884893/en

Reimers, N., Gurevych, I.: Sentence-bert: Sentence embeddings using siamese bert-networks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics (11 2019), http://arxiv.org/abs/1908.10084

Ryan, M., Naous, T., Xu, W.: Revisiting non-English text simplification: A unified multilingual benchmark. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 4898–4927. Association for Computational Linguistics, Toronto, Canada (Jul 2023), https://aclanthology.org/2023.acl-long.269

Shardlow, M.: A survey of automated text simplification. International Journal of Advanced Computer Science and Applications 4 (01 2014)

Tolcsvai Nagy, G.: Szövegtan. In: Kiefer, F. (szerk.) Magyar nyelv, pp. 149–174. Akadémiai Kiadó, Budapest (2006)

Tátrai, S.: Pragmatika. In: Tolcsvai Nagy, G. (szerk.) Nyelvtan, chap. Pragmatika, pp. 899–1058. Osiris Kiadó, Budapest, Hungary (2017)

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. CoRR abs/1706.03762 (2017), http://arxiv.org/abs/1706.03762

von Werra, L., Belkada, Y., Tunstall, L., Beeching, E., Thrush, T., Lambert, N., Huang, S., Rasul, K., Gallouédec, Q.: Trl: Transformer reinforcement learning. https://github.com/huggingface/trl (2020)

Xu, W., Napoles, C., Pavlick, E., Chen, Q., Callison-Burch, C.: Optimizing statistical machine translation for text simplification. Transactions of the Association for Computational Linguistics 4, 401–415 (2016), https://aclanthology.org/Q16-1029

Yang, A., Yang, B., Hui, B., Zheng, B., Yu, B., Zhou, C., ...: Qwen2 technical report. arXiv preprint arXiv:2407.10671 (2024)

Yang, Z.G., Dodé, R., Ferenczi, G., Héja, E., Jelencsik-Mátyus, K., Kőrös, Á., Laki, L.J., Ligeti-Nagy, N., Vadász, N., Váradi, T.: Jönnek a nagyok! BERT-large, GPT-2 és GPT-3 nyelvmodellek magyar nyelvre. In: XIX. Magyar Számítógépes Nyelvészeti Konferencia. pp. 247–262. Szegedi Tudományegyetem, Informatikai Intézet, Szeged (2023)

Zhu, Z., Bernhard, D., Gurevych, I.: A monolingual tree-based translation model for sentence simplification. In: Huang, C.R., Jurafsky, D. (szerk.) Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010). pp. 1353–1361. Coling 2010 Organizing Committee, Beijing, China (Aug 2010), https://aclanthology.org/C10-1152

A. Promptok

Zero-shot prompt

You are an assistant, specializing in text simplification. Your task is to simplify news for mentally disabled people.

RULES FOR SIMPLIFICATION

- ***Only generate text in Hungarian.*** Do not generate text in any other language.

- Simplify the standard-language text in the ***TASK*** section. Only use information that is found in this text. Do not make up any information or facts that is not in this text.

- Explain the concepts that you deem too hard.

- If you think that a sentence is too long, shorten it.

- Use simple, well-known words. Use the same word for the same concept throughout the simplified text. Do ***not*** use synonyms for the same concept to make the text less repetitive.

- If you think that some information in the standard-language text is not necessary for the reader to understand the text, you may leave it out of the simplified text.

TASK task

Few-shot prompt

You are an assistant, specializing in text simplification. Your task is to simplify news for mentally disabled people.

RULES FOR SIMPLIFICATION

- ***Only generate text in Hungarian.*** Do not generate text in any other language.

- Simplify the standard-language text in the ***TASK*** section. Only use information that is found in this text. Do not make up any information or facts that is not in this text.

- Explain the concepts that you deem too hard.

- If you think that a sentence is too long, shorten it.

- Use simple, well-known words. Use the same word for the same concept throughout the simplified text. Do ***not*** use synonyms for the same concept to make the text less repetitive.

- If you think that some information in the standard-language text is not necessary for the reader to understand the text, you may leave it out of the simplified text.

- Use the examples from the ***EXAMPLES FOR SIMPLIFICATION*** section to gain information about what a good simplification is. Make sure to generate text that is similar to these examples in grammar and simplicity.

- The examples are separated by '##########'.

EXAMPLES FOR SIMPLIFICATION examples

TASK task

B. További modellek

A lenti táblázat tartalmazza néhány további modell eredményét. A számokat összevetve a 3 táblázattal, a legfontosabb észrevétel, hogy a GPT-4o mini jelentősen felülmúlja mind a GPT-4o-t, mind az Claude 3.5 Haikut, annak ellenére, hogy ezen a feladaton előbbi 15×, utóbbi több mint 6× drágább. A másik fontos tanulság, hogy egy alverzió különbség is sokat tud jelenteni: a Llama 3.3 70B kvantált változatának SARI pontszáma majdnem egy teljes ponttal, perplexitása 1,5-tel javult a 3.1-es verzióhoz képest.

Modell / vizsgálati eset	BLEU	SARI	Perplexitás
Claude 3.5 Haiku			
1. (0-shot)	6,28 ± 3,87	38,90 ± 6,14	15,11 ± 4,67
2. (1-shot)	8,20 ± 5,08	41,04 ± 6,29	14,10 ± 4,30
3. (1-shot + kat.)	7,98 ± 5,12	40,83 ± 6,62	14,02 ± 4,17
4. (3-shot)	7,96 ± 5,14	40,74 ± 6,22	14,11 ± 4,17
5. (3-shot + Levenshtein)	10,58 ± 6,02	42,81 ± 5,81	**12,63 ± 3,69**
6. (3-shot + kat. + Levenshtein)	**10,86 ± 6,15**	**42,87 ± 5,96**	12,72 ± 3,73
7. (3-shot + kat.)	8,15 ± 5,07	40,95 ± 6,35	13,89 ± 4,09
GPT-4o			
1. (0-shot)	8,82 ± 5,12	41,29 ± 5,89	15,93 ± 5,72
5. (3-shot + Levenshtein)	10,24 ± 5,75	42,70 ± 5,85	14,98 ± 5,25
LLama 3.1 70B, kvantált			
1. (0-shot)	6,36 ± 4,85	38,28 ± 6,85	18,45 ± 7,74
2. (1-shot)	8,13 ± 5,37	40,21 ± 6,23	15,40 ± 6,39
3. (1-shot + kat.)	8,82 ± 5,44	40,76 ± 6,37	15,34 ± 6,27
4. (3-shot)	8,31 ± 5,44	40,11 ± 6,24	15,59 ± 6,57
5. (3-shot + Levenshtein)	10,55 ± 6,96	41,50 ± 5,88	**13,10 ± 4,52**
6. (3-shot + kat. + Levenshtein)	**10,76 ± 6,60**	**41,80 ± 6,52**	13,21 ± 5,44
7. (3-shot, + kat.)	8,60 ± 5,73	40,54 ± 6,40	15,36 ± 5,56
Llama 3.2 3B			
1. (0-shot)	5,83 ± 4,34	37,12 ± 5,70	59,32 ± 52,30
2. (1-shot)	6,73 ± 5,25	38,21 ± 5,92	38,86 ± 38,57
3. (1-shot + kat.)	6,41 ± 4,88	37,62 ± 5,84	41,57 ± 41,65
4. (3-shot)	**7,13 ± 5,67**	38,24 ± 6,39	43,28 ± 40,85
5. (3-shot + Levenshtein)	7,02 ± 5,67	**38,49 ± 6,24**	**40,47 ± 44,63**
6. (3-shot + kat. + Levenshtein)	6,97 ± 5,56	38,47 ± 6,10	40,84 ± 45,33
7. (3-shot + kat.)	6,45 ± 4,88	37,90 ± 6,38	39,04 ± 33,18

4. táblázat. A BLEU- és SARI-pontok, valamint a perplexitás megoszlása modellenként. A legjobb értékek minden oszlopban, minden modell esetében félkövérrel vannak jelölve. A jelölések követik a vizsgálati esetek leírásánál feltüntetettet: a *+ kat.* jelölés a kategóriaegyeztetést jelenti, a *+ Levenshtein* pedig a három Levenshtein-csoport figyelembe vételét veszik. A jelek hiánya az adott tulajdonság randomizált vagy figyelembe nem vett voltára utal.

C. Részletes SARI-, BLEU- és Perplexitás-adatok

A jelölések követik a vizsgálati esetek felsorolásánál feltüntetettet.

Modell	Vizsgálati eset	Mean ± Stdev	Min	Max
Claude 3.5 Haiku	1. (0-shot)	6,28 ± 3,87	0,52	27,68
	2. (1-shot)	8,20 ± 5,08	0,65	25,44
	3. (1-shot + kat.)	7,98 ± 5,12	0,56	34,55
	4. (3-shot)	7,96 ± 5,14	0,62	27,89
	5. (3-shot + Levenshtein)	10,58 ± 6,02	0,54	29,02
	6. (3-shot + kat. + Levenshtein)	10,86 ± 6,15	0,84	29,68
	7. (3-shot + kat.)	8,15 ± 5,07	0,49	28,10
GPT-4o	1. (0-shot)	8,82 ± 5,12	1,61	35,62
	5. (3-shot + Levenshtein)	10,24 ± 5,75	0,74	27,52
GPT-4o mini	1. (0-shot)	10,48 ± 6,52	0,90	35,25
	2. (1-shot)	12,33 ± 6,96	1,13	34,81
	3. (1-shot + kat.)	12,43 ± 6,88	0,89	37,32
	4. (3-shot)	12,39 ± 6,75	0,98	36,28
	6. (3-shot + kat. + Levenshtein)	13,28 ± 7,17	1,23	41,17
	5. (3-shot + Levenshtein)	13,38 ± 7,40	1,67	35,54
	7. (3-shot + kat.)	12,48 ± 7,11	1,14	34,41
Llama 3.3 70B, kvantált	1. (0-shot)	6,42 ± 4,19	0,65	23,46
	2. (1-shot)	8,51 ± 5,22	0,22	29,83
	3. (1-shot + kat.)	8,57 ± 5,06	0,64	27,29
	4. (3-shot)	8,72 ± 5,21	0,43	34,70
	5. (3-shot + Levenshtein)	10,65 ± 5,98	0,35	33,65
	6. (3-shot + kat. + Levenshtein)	10,60 ± 6,05	0,86	30,84
	7. (3-shot + kat.)	8,45 ± 5,52	0,55	30,50
LLama 3.1 70B, kvantált	1. (0-shot)	6,36 ± 4,85	0,06	31,69
	2. (1-shot)	8,13 ± 5,37	0,58	29,24
	3. (1-shot + kat.)	8,82 ± 5,44	0,09	27,42
	4. (3-shot)	8,31 ± 5,44	0,63	27,52
	5. (3-shot + Levenshtein)	10,55 ± 6,96	0,33	36,96
	6. (3-shot + kat. + Levenshtein)	10,76 ± 6,60	0,01	36,17
	7. (3-shot + kat.)	8,60 ± 5,73	0,10	30,53
Llama 3.1 8B	1. (0-shot)	6,32 ± 4,26	0,68	22,12
	2. (1-shot)	6,65 ± 4,81	0,05	24,29
	3. (1-shot + kat.)	6,51 ± 4,58	0,17	26,41
	4. (3-shot)	6,58 ± 4,89	0,26	27,25
	5. (3-shot + Levenshtein)	7,65 ± 5,63	0,07	28,68
	6. (3-shot + kat. + Levenshtein)	7,52 ± 4,88	0,49	26,63
	7. (3-shot + kat.)	6,94 ± 5,11	0,26	27,89
Llama 3.2 3B	1. (0-shot)	5,83 ± 4,34	0,13	25,30
	2. (1-shot)	6,73 ± 5,25	0,35	35,72
	3. (1-shot + kat.)	6,41 ± 4,88	0,52	28,76
	4. (3-shot)	7,13 ± 5,67	0,07	36,07
	5. (3-shot + Levenshtein)	7,02 ± 5,67	0,07	34,50
	6. (3-shot + kat. + Levenshtein)	6,97 ± 5,56	0,01	29,48
	7. (3-shot + kat.)	6,45 ± 4,88	0,23	27,17
Qwen2.5 7B	1. (0-shot)	5,64 ± 4,62	0,01	25,48
	2. (1-shot)	8,37 ± 6,20	0,00	32,42
	3. (1-shot + kat.)	8,55 ± 6,29	0,00	32,07
	4. (3-shot)	8,41 ± 6,08	0,87	33,52
	5. (3-shot + Levenshtein)	9,95 ± 6,84	0,80	44,54
	6. (3-shot + kat. + Levenshtein)	10,17 ± 6,61	0,94	32,46
	7. (3-shot + kat.)	8,53 ± 5,70	0,11	27,12

5. táblázat. BLEU eredmények

Modell	Vizsgálati eset	Mean ± Stdev	Min	Max
Claude 3.5 Haiku	1. (0-shot)	38,90 ± 6,14	24,16	58,61
	2. (1-shot)	41,04 ± 6,29	23,52	63,53
	3. (1-shot + kat.)	40,83 ± 6,62	21,91	62,41
	4. (3-shot)	40,74 ± 6,22	26,25	62,41
	5. (3-shot + Levenshtein)	42,81 ± 5,81	26,08	60,06
	6. (3-shot + kat. + Levenshtein)	42,87 ± 5,96	26,23	63,70
	7. (3-shot + kat.)	40,95 ± 6,35	23,97	60,89
GPT-4o	1. (0-shot)	41,29 ± 5,89	22,67	57,12
	5. (3-shot + Levenshtein)	42,70 ± 5,85	24,08	59,16
GPT-4o mini	1. (0-shot)	42,49 ± 5,69	25,08	64,70
	2. (1-shot)	43,99 ± 5,71	28,92	63,74
	3. (1-shot + kat.)	43,74 ± 5,23	25,24	59,97
	4. (3-shot)	43,89 ± 5,27	25,52	61,21
	5. (3-shot + Levenshtein)	44,58 ± 5,58	30,00	65,02
	6. (3-shot + kat. + Levenshtein)	44,36 ± 5,58	30,04	70,66
	7. (3-shot + kat.)	44,05 ± 5,78	24,54	63,85
Llama 3.3 70B, kvantált	1. (0-shot)	38,59 ± 6,71	18,26	60,95
	2. (1-shot)	40,62 ± 6,19	19,99	60,78
	3. (1-shot + kat.)	40,88 ± 6,58	16,11	62,63
	4. (3-shot)	41,02 ± 6,57	18,48	57,80
	5. (3-shot + Levenshtein)	42,41 ± 5,91	23,42	58,73
	6. (3-shot + kat. + Levenshtein)	42,43 ± 6,19	25,00	68,95
	7. (3-shot + kat.)	40,61 ± 6,88	19,84	58,53
LLama 3.1 70B, kvantált	1. (0-shot)	38,28 ± 6,85	19,41	59,78
	2. (1-shot)	40,21 ± 6,23	23,24	60,14
	3. (1-shot + kat.)	40,76 ± 6,37	21,65	59,36
	4. (3-shot)	40,11 ± 6,24	19,20	61,36
	5. (3-shot + Levenshtein)	41,50 ± 5,88	24,90	59,54
	6. (3-shot + kat. + Levenshtein)	41,80 ± 6,52	22,34	61,70
	7. (3-shot, + kat.)	40,54 ± 6,40	21,00	58,29
Llama 3.1 8B	1. (0-shot)	38,17 ± 6,22	15,08	58,73
	2. (1-shot)	38,58 ± 6,29	17,01	63,09
	3. (1-shot + kat.)	38,52 ± 6,08	17,92	55,43
	4. (3-shot)	38,53 ± 6,17	15,96	60,53
	5. (3-shot + Levenshtein)	39,27 ± 5,95	25,12	59,03
	6. (3-shot + kat. + Levenshtein)	39,49 ± 5,61	25,42	52,00
	7. (3-shot + kat.)	38,93 ± 6,18	20,84	56,70
Llama 3.2 3B	1. (0-shot)	37,12 ± 5,70	19,60	52,43
	2. (1-shot)	38,21 ± 5,92	21,73	58,40
	3. (1-shot + kat.)	37,62 ± 5,84	20,71	53,04
	4. (3-shot)	38,24 ± 6,39	19,89	57,00
	5. (3-shot + Levenshtein)	38,49 ± 6,24	18,38	52,61
	6. (3-shot + kat. + Levenshtein)	38,47 ± 6,10	22,42	59,56
	7. (3-shot + kat.)	37,90 ± 6,38	17,94	54,83
Qwen2.5 7B	1. (0-shot)	37,53 ± 6,28	17,87	56,65
	2. (1-shot)	39,63 ± 5,50	23,39	55,35
	3. (1-shot + kat.)	39,69 ± 5,89	17,17	57,12
	4. (3-shot)	39,93 ± 5,63	22,94	57,18
	5. (3-shot + Levenshtein)	40,34 ± 5,65	23,61	57,14
	6. (3-shot + kat. + Levenshtein)	40,65 ± 5,44	26,24	63,97
	7. (3-shot + kat.)	39,75 ± 5,54	17,89	55,87

6. táblázat. SARI eredmények

	Átlag	Minimum	Maximum	
Claude 3.5 Haiku	1. (0-shot)	15,11 ± 4,67	7,45	33,78
	2. (1-shot)	14,10 ± 4,30	7,49	31,66
	3. (1-shot + kat.)	14,02 ± 4,17	7,40	29,37
	4. (3-shot)	14,11 ± 4,17	6,88	27,03
	5. (3-shot + Levenshtein)	12,63 ± 3,69	6,02	23,13
	6. (3-shot + kat. + Levenshtein)	12,72 ± 3,73	5,88	25,40
	7. (3-shot + kat.)	13,89 ± 4,09	6,02	26,73
GPT-4o	1. (0-shot)	15,93 ± 5,72	7,52	47,48
	5. (3-shot + Levenshtein)	14,98 ± 5,25	5,87	57,49
GPT-4o mini	1. (0-shot)	15,80 ± 5,76	6,64	47,00
	2. (1-shot)	15,14 ± 5,39	6,91	34,35
	3. (1-shot + kat.)	15,14 ± 5,31	5,95	64,81
	4. (3-shot)	15,31 ± 5,44	7,02	39,08
	5. (3-shot + Levenshtein)	14,85 ± 5,27	7,02	42,52
	6. (3-shot + kat. + Levenshtein)	14,58 ± 4,58	7,07	34,15
	7. (3-shot + kat.)	15,06 ± 5,26	7,22	37,36
Llama 3.3 70B, kvantált	1. (0-shot)	14,79 ± 5,36	6,60	44,09
	2. (1-shot)	13,50 ± 4,68	6,07	37,93
	3. (1-shot + kat.)	13,40 ± 5,01	5,81	33,86
	4. (3-shot)	13,48 ± 5,03	6,13	37,76
	5. (3-shot + Levenshtein)	11,72 ± 4,06	5,25	38,82
	6. (3-shot + kat. + Levenshtein)	11,66 ± 3,91	5,04	26,39
	7. (3-shot + kat.)	13,57 ± 4,98	5,90	35,79
LLama 3.1 70B, kvantált	1. (0-shot)	18,45 ± 7,74	7,10	50,28
	2. (1-shot)	15,40 ± 6,39	6,00	67,82
	3. (1-shot + kat.)	15,34 ± 6,27	5,97	42,57
	4. (3-shot + Levenshtein)	13,10 ± 4,52	4,54	27,85
	5. (3-shot)	15,59 ± 6,57	6,57	41,24
	6. (3-shot + kat. + Levenshtein)	13,21 ± 5,44	5,54	38,29
	7. (3-shot + kat.)	15,36 ± 5,56	7,19	44,61
Llama 3.1 8B	1. (0-shot)	25,11 ± 13,84	9,13	163,04
	2. (1-shot)	23,57 ± 12,41	5,69	165,13
	3. (1-shot + kat.)	23,39 ± 11,08	8,77	73,61
	4. (3-shot)	23,86 ± 12,42	6,07	118,37
	5. (3-shot + Levenshtein)	19,97 ± 9,26	6,13	88,03
	6. (3-shot + kat. + Levenshtein)	20,32 ± 9,47	5,67	59,32
	7. (3-shot + kat.)	24,58 ± 12,06	8,99	102,85
Llama 3.2 3B	1. (0-shot)	59,32 ± 52,30	5,00	577,32
	2. (1-shot)	38,86 ± 38,57	3,60	299,56
	3. (1-shot + kat.)	41,57 ± 41,65	4,76	351,74
	4. (3-shot)	43,28 ± 40,85	9,96	367,61
	5. (3-shot + Levenshtein)	40,47 ± 44,63	3,83	600,01
	6. (3-shot + kat. + Levenshtein)	40,84 ± 45,33	1,46	734,61
	7. (3-shot + kat.)	39,04 ± 33,18	9,48	160,63
Qwen2.5 7B	1. (0-shot)	65,54 ± 45,84	6,93	236,78
	2. (1-shot)	45,36 ± 29,16	10,32	514,96
	3. (1-shot + kat.)	44,45 ± 29,07	10,19	163,75
	4. (3-shot)	47,64 ± 30,27	8,57	154,69
	5. (3-shot + Levenshtein)	38,32 ± 25,25	1,23	161,58
	6. (3-shot + kat. + Levenshtein)	37,05 ± 22,17	2,66	133,32
	7. (3-shot + kat.)	44,33 ± 29,79	8,78	137,05

7. táblázat. A perplexitás eredményei