

Automatikus szövegegyeszerűsítés a magyar nyelvre

Jelige: kaktusz

A kézirat lezárásának dátuma: 2023. 01. 11.

Tartalomjegyzék

1	Bevezetés	3
2	Elméleti keret	4
2.1	Az egyszerűsítés fogalma	4
2.2	Az intralingvális fordítás	5
2.3	Szövegegyszerűsítés a nyelvtechnológiában	6
2.4	A dolgozat szövegfogalma	8
3	Anyag és módszer	9
3.1	A kutatás célja	9
3.2	A korpusz előállítása: a használt korpusz és a fordítás.....	9
3.3	A korpusz fordításának értékelése	11
4	A korpusz használhatóságának felmérése	17
4.1	PULI GPT-3SX és HILANCO-GPTX modellek bemutatása	18
4.2	A PULI modell eredményei.....	19
4.3	A HILANCO-GPTX modell	20
4.4	A ChatGPT modell	21
4.5	A korpusz értékelése a ChatGPT segítségével	23
5	Összefoglalás, kitekintés.....	29
6	Bibliográfia	31

1 Bevezetés

Ha különböző koncepciókat, elméleteket, ötleteket vagy történeteket öntünk szóba, tudatosan ahhoz a személyhez igazítjuk mondandónk komplexitását, akivel kommunikálunk. Bizonyos esetekben azonban szükséges, hogy ezt a komplexitást jelentősen lecsökkentsük: például ha gyermekekkel, nem anyanyelvi beszélőkkel, vagy a megértésben korlátozott személyekkel kommunikálunk. Ez a mentális munka azonban sokszor nehéz, illetve már kész szövegek esetén sok időt követel a szöveg átfordítása – elvitathatatlan azonban, hogy az információkhoz mindenkinek egyforma joga van hozzájutni, rendelkezzen bármilyen olvasási vagy szövegértési készséggel: ezt az ENSZ Fogyatékossgal élő személyek jogairól szóló egyezményének 9. cikke (2006. évi New York-i egyezmény, [9 cikk. 1. bekezdés]. Kihirdette: 2007. évi XCII. törvény a Fogyatékossgal élő személyek jogairól szóló egyezmény és az ahhoz kapcsolódó Fakultatív Jegyzőkönyv a kihirdetéséről) is kimondja: „A fogyatékossgal élő személyek önálló életvitelének és az élet valamennyi területén történő teljes körű részvételének lehetővé tétele érdekében a részes államok megfelelő intézkedéseket tesznek, hogy másokkal azonos alapon biztosítsák a fogyatékossgal élő személyek számára (...) az információhoz és kommunikációhoz, beleértve az információs és kommunikációs technológiákat és rendszereket (...) való hozzáférést, mind városi, mind vidéki területeken.”). Dolgozatom egyrészt az első olyan magyar nyelvű korpusz elkészítését és használhatóságának ellenőrzését mutatja be, amelynek segítségével elkészíthető egy olyan számítógépes modell, amely képes magyar nyelvű szövegek egyszerűsítésére, ezzel megkönnyítve az ilyen kommunikációs szituációk, nehézségek áthidalását, másrészt pedig bemutat néhány lehetőséget a korpusz használatára. Dolgozatom célja nem csupán a korpusz bemutatása, hanem az automatikus szövegegyszerűsítés kommunikációs folyamatának a nyelvészeti szempontú bemutatása is.

A dolgozat felépítése a következő. Először egy rövid elméleti bevezetőben ismertetem a kulcsfogalmakat (2.). Ezután bemutatom a vizsgált anyagot, a felhasznált korpuszt, illetve a korpusz magyarra való átültetésének módját (a fordítás menetét), illetve a fordítás kiértékelését (3.). Ezt a korpusz használhatóságára vonatkozó vizsgálatokról szóló fejezet követi (4.), melyet három, már létező nyelvmodell segítségével értékeltem, majd a dolgozatot egy összefoglalás zárja (5.).

2 Elméleti keret

Az emberi kommunikáció alapvető célja a közös jelentésképzés (vö. Tolcsvai Nagy 2017: 26–28), hiszen a megnyilatkozó általában arra törekszik, hogy az összes, a diskurzusban részt vevő ember ugyanúgy, vagy a lehető leghasonlóbban értse meg a referenciális jelenetet, mint ő. Ahhoz, hogy a közös jelentésképzés sikerrel végbemenjen, szükséges, hogy a diskurzusban részt vevő felek olyan kommunikációs stratégiákat válasszanak, hogy azzal a többi résztvevő kommunikációs igényét ki tudják elégíteni, azaz adaptívan alkalmazkodjanak a diskurzus résztvevőihöz (Tátrai 2017: 920–926). Bizonyos esetekben azonban a diskurzus egy vagy több résztvevője valamilyen szempontból nem prototipikus befogadó¹ – például nem felnőtt, esetleg testi vagy szellemi fogyatékkal élő, vagy a társalgás nyelve nem az anyanyelve. Ebben az esetben kézenfekvő, hogy az a résztvevő, aki ilyen nehézségekkel nem rendelkezik, ahhoz a résztvevőhöz igazítja a kommunikációját, aki az említett okokból megértési nehézségekkel küzd. Ennek az egyik módja lehet az, hogy a mondandóját „egyszerűbben” fogalmazza meg.

Jelen dolgozat az első olyan magyar nyelvű korpuszt mutatja be, amelyben standard nyelvi és egyszerűsített szövegpárok találhatók meg. A korpusz előkészítése és validálása egy olyan számítógépes modell előkészületeihez járul hozzá, amely egyrészt le tudja képezni a fent leírt (humán) adaptációt, másrészt pedig gépi segítséget nyújthat ehhez az adaptációhoz azzal, hogy képes magyar nyelvű mondatokat egyszerűsíteni. A kutatás bemutatása előtt a dolgozat által használt fogalmakat mutatom be.

2.1 Az egyszerűsítés fogalma

Az egyszerűsítés, egyszerűsített szöveg pontos fogalma sem a magyar, sem a külföldi szakirodalomban nem tisztázott, sőt, nem is egységes: majdnem minden nyelven több különböző fogalom van aktív használatban a tudományos életben (közel) ugyanarra a nyelvi tevékenységre/produktumra. Az angol gyógypedagógiai szakirodalom az *easy language* (Maaß 2020), *plain language* (uo.), *easy-to-read text* (Inclusion Europe 2009), kifejezéseket használja a standard nyelvű szövegeknél könnyebben megérthető szövegekre; a német gyógypedagógiai szakirodalom a *leichte Sprache*, *einfache Sprache* (Hansen-Schirra–Maaß 2020) szókapcsolatokkal operál,² míg magyarul a *könnyen érthető kommunikáció (KÉK)* (Horváth 2019) és a *könnyen érthető nyelv(ezet)* (Horváth–Ladányi 2021) kifejezésekkel találkozhatunk a gyógypedagógiában. A nyelvtechnológia szakirodalma ugyanerre a folyamatra konzekvensen

¹ Prototipikus befogadó alatt az anyanyelvi, felnőtt, testileg és szellemileg egészséges befogadót értem.

² Fontosnak tartom kiemelni, hogy ezek a kifejezések nem minden esetben fedik tökéletesen egymást: németben például az *einfache Sprache* egy könnyebben érthető nyelvváltozat, mint az *einfache Sprache* (ugyanaz igaz az *easy language* – *plain language* fogalompárra is). (Hansen-Schirra – Maaß 2020: 18)

*text simplification*ként (szövegegszerűsítés) utal, a létrejött szöveget pedig *simplified text*nek (egyszerűsített szöveg) hívja. Mivel a magyar nyelvtechnológiai közösségben – tudomásom szerint – még senki nem foglalkozott a *text simplification* kérdéskörével, így ennek nincs kialakult magyar fordítása ezen a tudományterületen. Jelen dolgozat az egyszerűsítés kérdéskörét nem gyógypedagógiai, hanem nyelvészeti-nyelvtechnológiai szempontból vizsgálja. Így a külföldi NLP-szakirodalomhoz alkalmazkodva a dolgozat is a *szövegegszerűsítés*, *egyszerűsített szöveg* terminusokat fogja használni.

Az egyszerűsítés meghatározására sincs pontos definíció, hiszen a különböző csoportoknak különböző igényeik vannak az egyszerűsítés terén (Bingel et al. 2018). Ezért talán kézenfekvő jelen dolgozatban a nyelvtechnológia szakirodalma alapján definiálni a fogalmat: így tehát egyszerűsítés alatt a dolgozat azt a metódust érti, amelynek során „egy szöveg nyelvi komplexitását úgy csökkentjük, hogy az közben az eredeti információ- és jelentéstartalmát megőrzi” (Siddhartan 2014: 259).

Szövegegszerűsítéskor többféle stratégiát alkalmazhatunk (Sikka–Mago 2020: 111), sőt, sokszor az egyes stratégiákat egyidejűleg, egymással vegyítve használjuk. A szakirodalom az egyszerűsítés három fő fajtáját különbözteti meg: a konceptuális egyszerűsítést, az elaboratív módosítást és a szövegösszefoglalást. Konceptuális egyszerűsítéskor formai és tartalmi egyszerűsítés is végbemegy – ez történik például akkor, ha egy óvodáskorú gyermeknek magyarázunk bonyolult folyamatokat, fogalmat (mint mondjuk a repülőgépek működése). Elaboratív módosítás esetében a fogalmakat részletezzük, esetleg redundáns szerkezeteket alkalmazunk például vonatkozó névmások helyett. Szövegösszefoglaláskor pedig a nem kulcsfontosságú elemeket elhagyjuk a szövegből.

Fontos hangsúlyozni, hogy a szövegösszefoglalás és a szövegegszerűsítés összeérő, de egymást nem tökéletesen fedő problémakörök: egyes esetekben szövegösszefoglaláskor nem lesz alacsonyabb az összefoglalt szöveg komplexitása (gondoljunk például egy tudományos szöveg absztraktjára), illetve megtörténhet az is, hogy az egyszerűsített szöveg nem rövidebb, hanem hosszabb lesz az eredeti szövegnél (például magyarázatok hozzáfűzése esetén).

2.2 Az intralingvális fordítás

A szövegegszerűsítés folyamatához megemlítendő még az intralingvális fordítás terminusa is. Roman Jakobson háromféle fordítást különít el: intralingvális, interlingvális és interszemiotikus fordításokat (Jakobson 1959: 233). Az intralingvális fordítás vagy átfogalmazás (*rewording*) nyelvi jelek ugyanazon nyelv más nyelvi jeleivel való értelmezése (uo. ill. magyarul l. Dobos

2019: 187). Az interlingvális fordítás két különböző nyelv közötti fordítás, míg az interszemiotikus fordítás esetében „a verbális jelek interpretációja egy nem verbális jelekből álló szemiotikai rendszer segítségével valósul meg” (Dobos 2019: 187). Az intralingvális fordítás nem rendelkezik egységes definícióval (Whyatt 2017: 178): a fentebb említett jakobsoni meghatározáshoz Sanders (2006) alapján nem csak az átfogalmazást, de egy szöveg különböző közönségekhez való adaptálását is intralingvális fordításnak tekinthetjük.

Kognitív szempontból a fordító (átfogalmazó) a centrális ágens a kommunikatív folyamatban, hiszen ő az, aki a „forrásnyelvi” szöveget a „célnyelvi”³ szöveggé alakítja (Feist 2013: 375). Ezen folyamat közben a fordító kétszeresen is részt vesz a jelentéskonstruálásban (Whyatt 2017: 179): egyrészt értelmezi a forrásnyelvi szöveget, másrészt pedig – a diskurzus többi résztvevőjének kognitív képességeinek figyelembevételével – kiválasztja azokat a nyelvi elemeket, amelyek érthetőek lesznek a többi résztvevő számára, és közben a lehető legjobban hozzáférhetővé teszi a referenciális jelenetet (uo., a referenciális jelenet fogalmához l. Tátrai 2017: 907–911).

2.3 Szövegegyszerűsítés a nyelvtechnológiában

A szövegegyszerűsítés népszerű kutatási terület a nyelvtechnológiában. Az első, szövegegyszerűsítés céljából kifejlesztett algoritmusok a dokumentumbéli gyakoriság alapján súlyozták a szavakat, és a legtöbb magas értékű szót tartalmazó mondat(ka)t adták vissza (Sikka–Mago 2020). Ez az extraktív módszer lényegében megegyezik a kulcsszóalapú szövegösszefoglaló algoritmusokkal, és ez nem is oldja meg tökéletesen a szövegegyszerűsítés feladatát: ugyan a kevésbé fontos információk elhagyása segíthet abban, hogy az intencionált befogadó könnyebben megértse a szöveget, fontos kiemelni, hogy a szöveg szubsztanciája ezzel nem változik, a tartalma tehát nem lesz egyszerűbb.

Később absztraktív, azaz az eredeti alapján új szöveget létrehozó megközelítések szintén létrejöttek, mind a lexikai, mind pedig a szintaktikai egyszerűsítésre. A legtöbb modern automatikus szövegegyszerűsítő modell mondathosszúságú elemek egyszerűsítésére képes, így az előbbieken bemutatott humán szövegegyszerűsítési stratégiák nem, vagy csak korlátozottan

³ Az interlingvális fordítás forrás- és célnyelv terminusait a szakirodalom alapján alkalmazom az intralingvális fordításra is, jelen kontextusban a standard szöveget tekintve forrásnyelvi és az egyszerűsített szöveget célnyelvi szövegnek.

alkalmazhatóak a leírásukra.⁴ A nyelvtechnológiai szakirodalom a mondategyszerűsítésre a következő átalakítási stratégiákat (transzformációkat) alkalmazza:

- lexikai szintű átfogalmazás (lexical paraphrasing) – azaz komplex szavak, mondatrészek egyszerűbb szinonimákkal való helyettesítése;
- a mondat szintaktikai struktúrájának megváltoztatása – ez lehet akár a mondat szétválasztása két vagy több rövidebb mondatra, de akár a tagmondatok sorrendjének megcserélése is;
- kevésbé fontos elemek elhagyása a szövegből – hasonlóan a szövegösszefoglaláshoz, itt a mondatból hagyjuk el azon elemeket, amelyek a megértéshez nem feltétlenül szükségesek, azonban a mondatot bonyolultabbá teszik. (Alva-Manchego et al. 2020: 4668)

Az absztraktív modellek performanciájának javításában különlegesen nagy áttörést azonban a neurálisháló-alapú modellek jelentettek: Nisioi és munkatársai (2017) egy szekvenciaalapú, alapvetően gépi fordításra létrehozott modellt használtak fel, forrásszöveggént a standard nyelvi anyagot, célszöveggént pedig az egyszerűsített szövegeket felhasználva. Ez a megközelítés azóta is népszerű, jelen tanulmány is ezzel dolgozik. A fordítómodellek felhasználásakor forrásnyelvi szövegnek a standard nyelvű, míg célnyelvinek az egyszerűsített szövegeket tekintjük. Egy megfelelő méretű és minőségű tanítókörpusz létrehozása azonban az első és a legfontosabb feladat ahhoz, hogy egy ilyen modell megvalósulhasson: ahhoz, hogy egy modell sikeresen képes legyen a szövegegyszerűsítésre szükséges egy nagyméretű, egynyelvű, párhuzamos korpusz. Ezekben a párhuzamos korpuszokban leggyakrabban standard és egyszerűsített mondatpárok találhatók meg, melyeket legtöbbször az interneten is elérhető szövegekből nyernek ki, illetve ritkábban emberek célzottan egyszerűsítene megadott mondatokat, és ebből áll össze a korpusz (a már létező korpuszokról bővebben az *Anyag és módszer* fejezetben írok). A rendszer tehát ezen mondatpárok segítségével tanítható – általánosságban elmondható, hogy minél nagyobb és minél jobb minőségű (azaz az egyszerűsítésre minél konzekvensebb példákat adó) a korpusz, annál jobb kimeneteket tud generálni a modell is.

⁴ A mondat szintű egyszerűsítés természetesen több korlátot is támaszt a szövegegyszerűsítő modellek irányában. Erről l. ebben a fejezetben később.

2.4 A dolgozat szövegfogalma

A dolgozat a szöveg fogalmát funkcionális kognitív keretben értelmezi (vö. Tolcsvai Nagy [szerk.] 2017). Mivel a dolgozatban ismertetett modell alapvetően mondatnyi egységek egyszerűsítésére képes, így a szövegfogalmat is a mondat felől érdemes megközelíteni. A prototipikus mondat „egy jelenetet fejez ki, a beszélő, tágabban konceptualizáló által megkonstruált olyan kijelentést, amely valamilyen egzisztenciális konstrukciót mutat be: egy létezőként megkonstruált dologgal kapcsolatos időbeli eseményt, általában egy másik dologgal viszonyban.” (Tolcsvai Nagy 2017: 391). Az egyszerűsítés kontextusában azonban az elemi mondat terminusa is megemlíthető (Tolcsvai Nagy 2017: 395): az elemi mondatok a megnyilatkozások alapvető feldolgozási egységei (Tátrai 2017: 1003), amelyek egy elemi jelenetet konstruálnak meg (Imrényi 2017: 666; Tolcsvai Nagy 2017: 391, 396). Mivel az elemi mondat információtartalma korlátozott (7 ± 2 elem, nagyjából a munkamemória kapacitásának megfelelően, l. Tolcsvai Nagy 2017: 395), így olyan esetekben, amikor a befogadónak nehézséget okoz a szöveg megértése, különösen fontos lehet, hogy a megnyilatkozó a mondatait minél emészthetőbb méretű elemi mondatokra redukálja.⁵ Ezt az Inclusion Europe által kiadott, a könnyen érthető kommunikáció irányelveit bemutató kiadvány is megerősíti, amely a beszélőket arra buzdítja, hogy „Fogalmazzunk rövid mondatokban” (Inclusion Europe 2013: 11, https://www.inclusion-europe.eu/wp-content/uploads/2017/06/HU_Information_for_all.pdf, utolsó megtekintés: 2022. 11. 06.).

A folyamatot, amelyben a felhasználó betesz egy mondatot az egyszerűsítő modellbe, az utóbbi pedig visszaadja annak az egyszerűsített megfelelőjét, egy nem prototipikus kommunikációs aktusnak tekintem. A kommunikáció „a közös cselekvés, a közös megismerésben létrejövő emberi cselekvés” (Tolcsvai Nagy 2017: 26). A prototipikus kommunikáció tehát emberek között valósul meg – jelen esetben azonban nem csak humán entitások, hanem egy számítógép is „részt vesz” ebben a folyamatban. Fontos azonban megjegyezni, hogy a számítógép nem mentális ágensként van jelen a folyamatban, hiszen nem a számítógép „találja ki” a leegyszerűsítendő mondatot, illetve nem tud közvetlen, reakcióra képes kapcsolatot létesíteni sem a megnyilatkozóval (azaz az egyszerűsítendő mondat megfogalmazójával), sem a befogadóval (azaz az egyszerűsített mondatot értelmező személlyel). Ennek ellenére az intralingvális fordítás fentebb leírt folyamatában a számítógép

⁵ Fontosnak tartom kiemelni, hogy ugyan az összetett mondatokban is beazonosíthatók lehetnek elemi mondatok, azokat nem csupán elemi mondatok összekapcsolt egységének tekintem, hanem különálló séma-megvalósulás viszonyban értelmezhető konstrukciótípusként (Kugler 2022). Így tehát a sikeres szövegegyszerűsítés sem lehet csupán az összetett mondat elemi mondatokra bontásával ekvivalens.

működik (korlátozott) centrális ágensként: ő az, aki megfogalmazza a szöveget úgy, hogy az egyidejűleg egyszerűbb legyen, mint a forrásszöveg, de közben a lehető legjobban visszaadja az intencionált jelentést.

Természetesen ahogy az már a fentiekből is körvonalazódhat, egy ilyen, szövegegyszerűsítést végző modell bizonyos korlátokkal működik. Egyrészt a modell által befogadott szöveg mennyisége (egy mondat) jelentősen rövidebb, mint egy átlagos szöveg. Emellett, mivel egy hosszabb szöveg makro-, de akár mezoszintű értelmezése (vö. Tolcsvai Nagy 2006) például nem áll a modell hatáskörében, így a nagyobb szerkezeti egységek, de akár a különböző deiktikus nyelvi elemek beazonosítása is problémás lehet, mind a modell, mind a befogadó számára az egyszerűsített mondatban. Ezen felül a modell – révén, hogy nem emberi entitásról van szó – nem képes alkalmazkodni a befogadóhoz: mivel nem létesít kapcsolatot vele, így nem tudja felmérni azt sem, hogy a befogadónak milyen a mentális állapota. A kommunikáció szigorú írásbelisége szintén korlátoz: az írott szöveget nehezebb lehet megérteni a szóbeli szövegeknél.

Mindezeket figyelembe véve kijelenthető tehát, hogy egy szövegegyszerűsítő modell nem tudja átvenni az emberi egyszerűsítés szerepét, csupán könnyíteni képes rajta, illetve bizonyos helyzetekben helyettesítheti az emberi fordítót.

3 Anyag és módszer

3.1 *A kutatás célja*

Jelen dolgozat az első magyar nyelvű, egyszerűsített és standard nyelvi szövegekből álló korpusz elkészítésének és validálásának a folyamatát mutatja be. Kutatásom célja egy olyan korpusz létrehozása volt, amelyet a jövőben fel lehet használni egy automatikus szövegegyszerűsítő modell tanításához. Mivel ahhoz, hogy egy ilyen modell megfelelő egyszerűsítéseket legyen képes generálni, szükséges az, hogy a korpusz megfelelően robusztus és magas minőségű legyen, így a létrehozott korpuszt már létező nyelvmodelleken validáltam is.

3.2 *A korpusz előállítása: a használt korpusz és a fordítás*

Egynyelvű, standard és egyszerűsített szövegekből álló párhuzamos korpusz mindössze hét nyelvre érhető el: angolra, németre, spanyolra, portugálra, olaszra, dánra és japánra (Kajiwara–Komachi 2016), azonban ezek közül csak az angol nyelvű korpuszok egy része publikus. Mivel

magyarra nem érhető el megfelelő mennyiségben olyan szöveg, amelyből egy egynyelvű, párhuzamos korpusz építésére lehetőség lenne, így a saját korpusz építésével jelen kutatás nem próbálkozott meg. A szakirodalomban ugyan találunk arra is példát, hogy egy korpuszt crowdsourcinggel építenek, azaz kiválasztanak bizonyos mondatokat, és (általában fizetett és betanított) dolgozókkal, alkalmazottakkal egyszerűsítetik őket (pl. Xu et al. 2016 vagy Alva-Manchego 2020). Ezek a korpuszok általában ugyan jobb minőségűek, mint a webaratással gyűjtött anyagok, azonban jelentős emberi munkabefektetést és anyagi háttérrel követelnek, illetve általában sokkal kisebb méretűek, mint a „hagyományos” korpuszok. Mivel jelen kutatás limitált anyagi és humán erőforrásokkal dolgozott, így a crowdsourcing segítségével történő korpuszépités szintén nem volt megvalósítható megoldás. Így egy már létező, angol nyelvű korpusz lefordítása tűnt a legkézenfekvőbb megoldásnak – erre a nemzetközi szakirodalomban is van példa (pl. Megna 2021). Természetesen ez bizonyos korlátokat támaszt a vizsgálattal szemben: mivel a korpusz nem autentikus magyar adatokat tartalmaz, így ezen a korpuszon például arra vonatkozó vizsgálatokat, hogy milyen mentális folyamatok mentén, vagy milyen stratégiákat alkalmazva zajlik a humán egyszerűsítés, nem lehet végezni. Azonban – feltéve azt, hogy az eredeti, angol korpuszban megfelelőek az egyszerűsítések – egy jó minőségű fordítással létrehozhatunk egy olyan tanítókorpuszt, amely alapján egy számítógépes modell viszonylag optimális egyszerűsítéseket tud generálni.

A dolgozat a PWKP korpuszt (Zhu et al. 2010) használja fel. Ebben a korpuszban több mint 65.000 angol nyelvű és egyszerűsített angol nyelvű Wikipedia-cikket (*English Wikipedia*, *Simple English Wikipedia* – a továbbiakban *EW*, *SEW*) párosítottak össze, majd ezekből automatikusan kinyerték az egymásnak megfelelő mondatpárokat, ezzel több mint 108.000 mondatpárt összegyűjtve. Fontos megjegyezni, hogy egy komplex mondathoz több egyszerű is tartozhat: mivel a kutatók számoltak azzal, hogy lesznek olyan mondatok, amelyek az EW korpuszban egy, míg a SEW korpuszban több mondatból állnak, így először egyesével párosították az egyszerűsített és a komplex mondatokat, majd egybevonták azokat a párokat, ahol ugyanaz a komplex mondat jött ki párnak több egyszerűsített mondathoz. Mivel ez a korpusz megfelelően robosztus, illetve a nemzetközi szakirodalomban is gyakran használt, így jelen vizsgálathoz is megfelelő alapnak tűnt.

A PWKP azonban hátrányokkal is bír: ahogy arra Xu és munkatársai (2015) is rámutatnak, a korpusz egy része (17%) nem jól lett párosítva (az egyszerűsített és a standard mondatok nem párok), egy része (33%) esetében pedig az egyszerűbb mondatok valójában nem egyszerűbbek a standard nyelvi párjuknál. Mivel azonban a PWKP a legnagyobb publikusan

elérhető korpusz, így a dolgozat emellett kötelezi el magát. Fontos kiemelni azonban, hogy a korpuszból szűrésre került előzetesen 3386 darab mondatpár. Ezek mindegyike a következőképp épült fel: a standard nyelvi mondat egy településről állapítja meg, hogy hol található (például: 105.107. sor mondata: *Thienu is a commune in the Aisne department in Picardie in northern France.*). Az „egyszerűsített” változat pedig ugyanezt adja vissza, azonban a város neve helyett csak az „It is...” formula található meg (például az előbbi mondat „egyszerűsítése”: *It is found in the region Picardie in the Aisne department in the north of France.*). Egyértelmű, hogy az ilyen mondatok esetében hiányzik egy kontextualizáló mondat (jelen esetben például egy olyan mondat, mint *Thienu is a commune.*), amely által értelmezhető lenne az *it* mondatkezdő deiktikus elem. Ezen mondatpárokból azonban ez nem volt megtalálható, így megalapozottnak tűnt kiszűrni őket, hiszen nem optimális egyszerűsítést végeztek.

A korpusz mérete miatt a manuális fordítás nem tűnt megvalósítható lehetőségnek, így a korpusz egy gépi fordítómodell segítségével került lefordításra. Az optimális fordítás megtalálásához a dolgozat két fordítómodell eredményeit hasonlítja össze: egyrészt az Opus-MT angol-magyar fordítóját (Tiedemann–Thottingal 2020), másrészt pedig a DeepL fordítóját (<https://www.deepl.com/>).

3.3 A korpusz fordításának értékelése

A lefordított korpuszok értékelése automatikus, gépi módszerekkel és humán annotátorok segítségével is megtörtént. A gépi értékeléshez a BLEU-pontrendszert (Papineni et al. 2002) használtam. A pontrendszert kifejezetten a gépi fordítás értékelésére fejlesztették ki: a pontszám kiszámításához az algoritmus megszámlolja, hogy a gépi fordítás és ugyanazon szöveg humán, gold standardként kezelt fordítása között hány megegyező n-gram van, pozíciótól függetlenül. Mivel a korpuszhoz gold standard (azaz humán) fordítás nem létezik, így a két lefordított korpusz egymással került összehasonlításra, egyszer az egyik, másszor a másik korpuszt használva gold standardként, a pontszámot az NLTK BLEU-kalkulátorával számítva (Bird et al. 2009), 7-es simítófüggvénnyel. Az eredmények 0,027 és 1,74 (a DeepL által fordított korpuszt használva gold standardként) 0,051 és 1,68 (az Opus-ML által fordított korpuszt használva gold standardként) BLEU-pont között voltak.⁶ Mivel azonban humán fordítás nem állt rendelkezésemre, így csupán a BLEU-pontok alapján való szűrés nem tűnt elég

⁶ A BLEU-pontok alapvetően 0 és 1 közötti skálán helyezkednek el, így az 1 fölötti értékek az algoritmus nem tökéletes implementációjára utalnak. Mivel azonban a mondatok pontszámait csak skálázásra használtam, így a dolgozat jelen verziójában ezeket a pontszámokat használtam.

megalapozottnak. Ezért a korpusz egy részét humán elemzésnek is alávettem, arra keresve a választ, hogy a BLEU-pontok és a humán értékelés között van-e korreláció, illetve hogy vannak-e olyan korpuszrészletek, amelyek vagy a fordítás, vagy az eredeti korpusz minősége miatt a vizsgálatból kizárhatók.

A lefordított korpusz értékelését egy független annotátor és én végeztük el. A vizsgálathoz a korpuszt először egytized BLEU-pontonként felosztottam, majd minden percentilisből 5-5 mondatot sorsoltam ki egy randomizáló függvénnyel. Ezeket végül egy Excel-táblába rendeztem, az angol, az Opus-ML és a DeepL fordításokat egymás mellé helyezve, és minden standard mondathoz hozzárendeltem az egyszerűsített változatát, illetve annak fordítását is. Amennyiben egy mondat teljesen rossz volt már az eredeti korpuszban is (például egy értelmetlen karaktersorozat volt), a mondatot eltávolítottam, és új mondatot sorsoltam. Ilyen összesen 5 mondat esetében történt, négyszer 0,1 maximum BLEU-percentilissel, egyszer pedig 0,2 maximum BLEU-percentilissel. A vizsgálatba összesen hetven mondatpár került bele, az alábbi elrendezésben:

Angol		Jelentésmegtartás	Grammatikai	Referencia folytonosság		Jelentésmegtartás	Grammatikai	Referencia folytonosság
The present parliament building, the Alþingishús, was built in 1881, of hewn Icelandic stone.	A jelenlegi parlamenti épülete, az Alþingishús 1881-ben épült, faragott izlandi kőből.	4	5		A jelenlegi parlamenti épülete, az Alþingishús 1881-ben épült, faragott izlandi kőből.	5	5	5
The parliament building where the Althing meets now is called the Alþingishús.	A parlamenti épületét, ahol az Althing most ülészik, Alþingishúsnak hívják.	4	5		A parlamenti épületét, ahol jelenleg az Althing ülészik, Alþingishúsnak hívják.	5	5	
It was built with Icelandic stone in 1881.	Izlandi kőből épült 1881-ben.	5	5		Izlandi kőből épült 1881-ben.	5	5	

1. ábra. Az értékelőlap egy részlete (1 db mondatpár).

A pontozási feltételek a következők voltak:

- Jelentésmegtartás: Ebben a szekcióban a fordítás minőségét kellett értékelni, azaz a magyar és az angol mondatot összevetni, és az alapján pontozni, hogy a magyar mondat mennyire jelenti ugyanazt, mint az eredeti. Ha egy szót nem sikerült lefordítani a fordítónak, vagy az eredeti mondat kontextusa alapján rosszul fordított valamit, ebben a szekcióban került levonásra pont. Fontos kiemelni, hogy a nevek és címek lefordítását nem vártuk el a fordítótól, azonban ha az egyik fordító lefordította, a másik pedig nem, akkor az kapott magasabb pontot, amelyik magyarra átültette a címet. Ugyanígy, ha egy névben különleges karakterek szerepeltek (mint az 1. ábrán), és egyik fordító sem fordította le a nevet, akkor az a verzió kapott magasabb pontszámot, amely az eredeti verzióhoz közelebb állt. A jelentésmegtartás szempont alatt minden mondat külön-külön kapott pontszámot.

- Grammatikalitás: Ezen szempont azt vizsgálta, hogy mennyire „helyes”, „magyaros”, gördülékeny és értelmes a lefordított mondat. Itt tehát az eredeti, angol mondattal már nem kellett összevetni a mondatot, hanem magában kellett azt értékelni. A grammatikalitás szempont alatt szintén minden mondat külön-külön kapott pontszámot.
- Referenciafolytonosság: Ebben a szempontban a fordítást az alapján kellett értékelni, hogy a standard és az egyszerűsített mondat(ok)ban ugyanaz a szó szerepel-e ugyanarra az entitásra. Tehát ha az angolban a következő két mondatot látnánk: „The dog is big. The dog is brown.”, és a magyar fordítás a következő lenne „A kutya nagy. Az eb barna.”, akkor a referenciafolytonosság nem valósulna jól meg, mivel amíg az angolban ugyanaz a szó (*dog*) szerepel mindkét mondatban, addig a magyar fordítás két külön szót alkalmaz rá. Fontos megemlíteni, hogy a referenciafolytonosság akkor is lehet jó, ha a fordító nem pontosan fordít dolgokat, vagy szavakat „talál ki”. Például a 30683-as sortól kezdődő mondatpár a következő az eredeti korpuszban:

August 22 — Battle of the Standard: David I of Scotland is defeated by the English.

August 22 - Battle of the Standard between David I of Scotland and the English.

Ez az OpenMT fordításában:

augusztus 22. – A normannok csatája: I. Dávid skót király vereséget szenved az angoloktól.

augusztus 22. – A normandiai csata I. Dávid skót király és az angolok között.

És a DeepL fordításában:

Augusztus 22. - A zászlós csata: A skót I. Dávid vereséget szenved az angoloktól.

augusztus 22. - A skót I. Dávid és az angolok között lezajlott zászlós csata.

Ugyan a *Battle of the Standard* csatának sem a *zászlós csata*, sem a *normandiai csata*, sem pedig a *normannok csatája* nem felel meg tökéletesen (a magyarban ezt a csatát az angol nevével szoktuk illetni), az egyszerűsítés szempontjából tulajdonképpen mindegy, hogy az információk, nevek, stb. faktuálisan helyesek-e: sokkal fontosabb az, hogy ahol az eredeti korpuszban nem volt szószinten változás, ott a fordításban se legyen (hiszen ez – tévesen – arra taníthatja a modellt,

hogy ott is végrehajtsa valamilyen változtatást a mondaton, ahol valójában nem kell).

Mivel a referenciafolytonosság csak a standard mondat és az egyszerűsített mondat egyszerre történő vizsgálata alapján megállapítható, így ezen szempont esetében nem egy-egy mondat, hanem egy mondatpár kapott egy pontszámot.

Az annotátorok ezen szempontok alapján egytől ötig terjedő Likert-skálán pontozták a mondatokat. A két korpusz átlagai a következők:

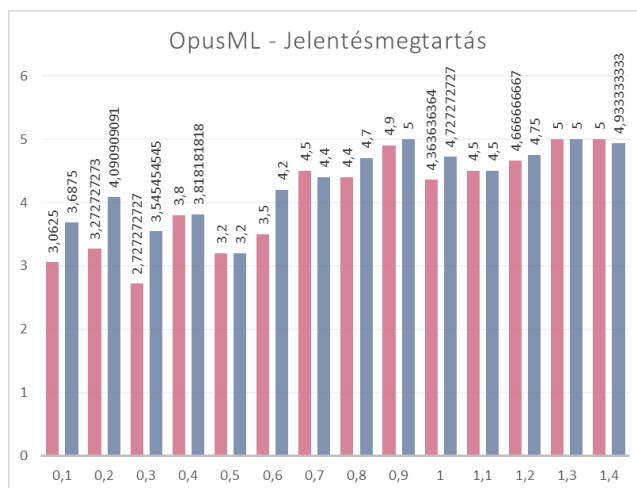
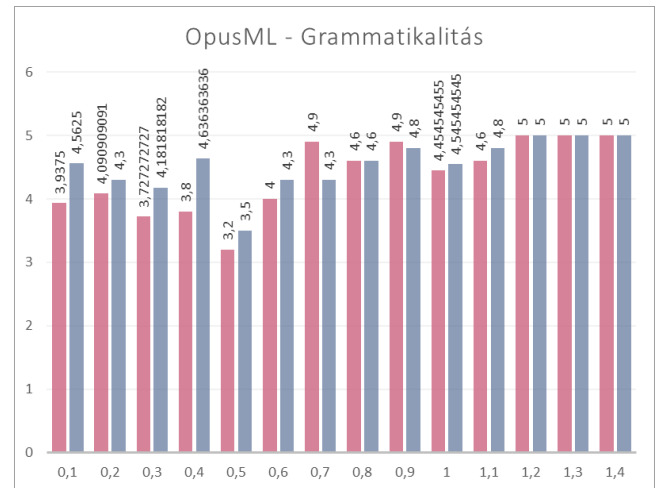
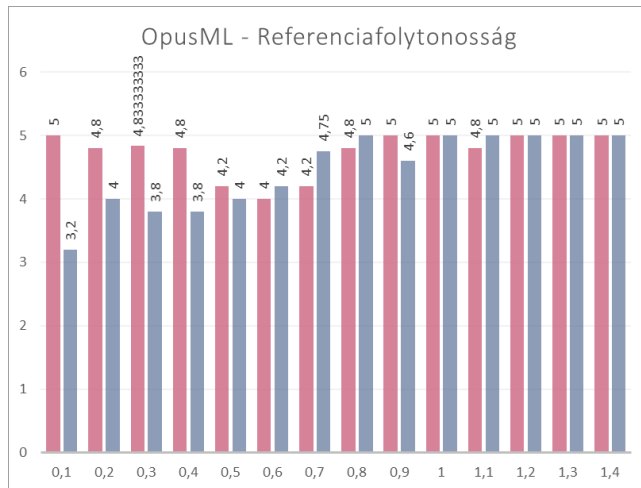
	Jelentésmegtartás	Grammatikalitás	Referenciafolytonosság
OpusML			
Első annotátor	4,038961039	4,37012987	4,753424658
Második annotátor	4,318471338	4,557692308	4,449275362
DeepL			
Első annotátor	4,320261438	4,688311688	4,915492958
Második annotátor	4,713375796	4,859872611	4,565217391

1. táblázat. Az annotátorok értékeléseinek átlaga.

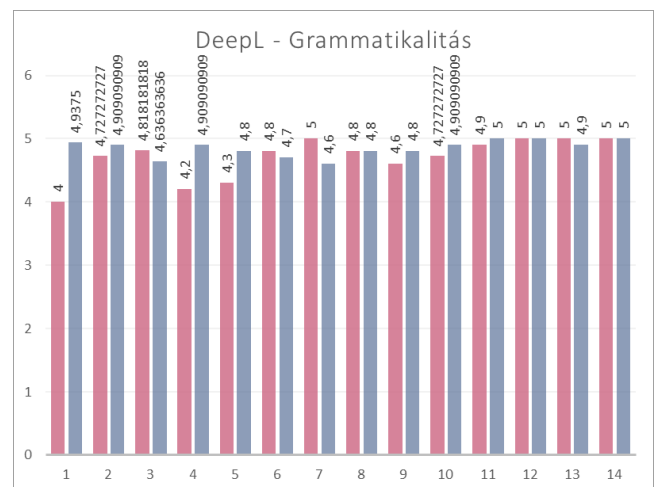
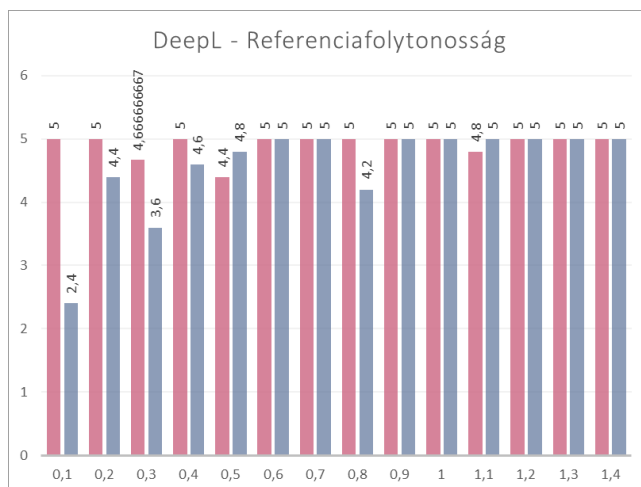
Jól látható, hogy mindkét annotátor esetében a DeepL által fordított korpusz jobban teljesített, minden vizsgált szempont tekintetében. Fontos kiemelni azonban, hogy az OpusML fordítása sem teljesített rosszul: ezen fordítás esetében is minden szempont 4,0 fölötti átlagos értékkel rendelkezett.

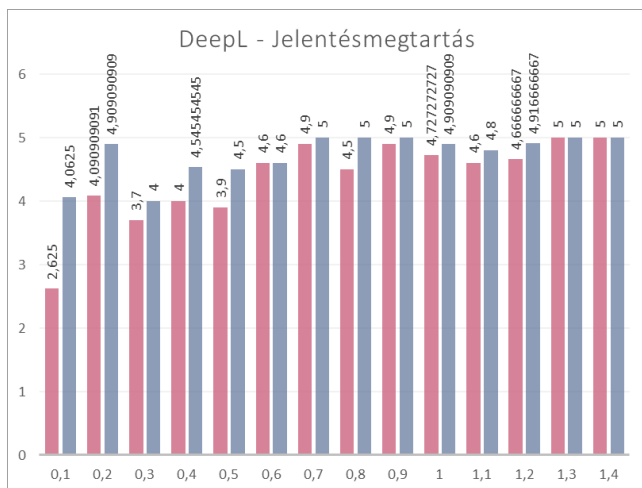
Az eredmények átláthatóságának növelése érdekében a korpuszokat percentilisenként is kiértékeltem. Az annotátorok pontjainak összesítése a következő diagramokon látható (a vízszintes tengelyen a BLEU-percentilis felső határa látható):⁷

⁷ 1,4-es maximum BLEU-percentilis fölött mindkét korpuszban csak egy-egy mondat volt, ezek értéke volt a korábbiakban maximumként meghatározott érték. 1,4 és 1,5; 1,5 és 1,6; illetve 1,6 fölött más mondat nem volt, így ezek a percentilisek a vizsgálatba sem kerültek be.



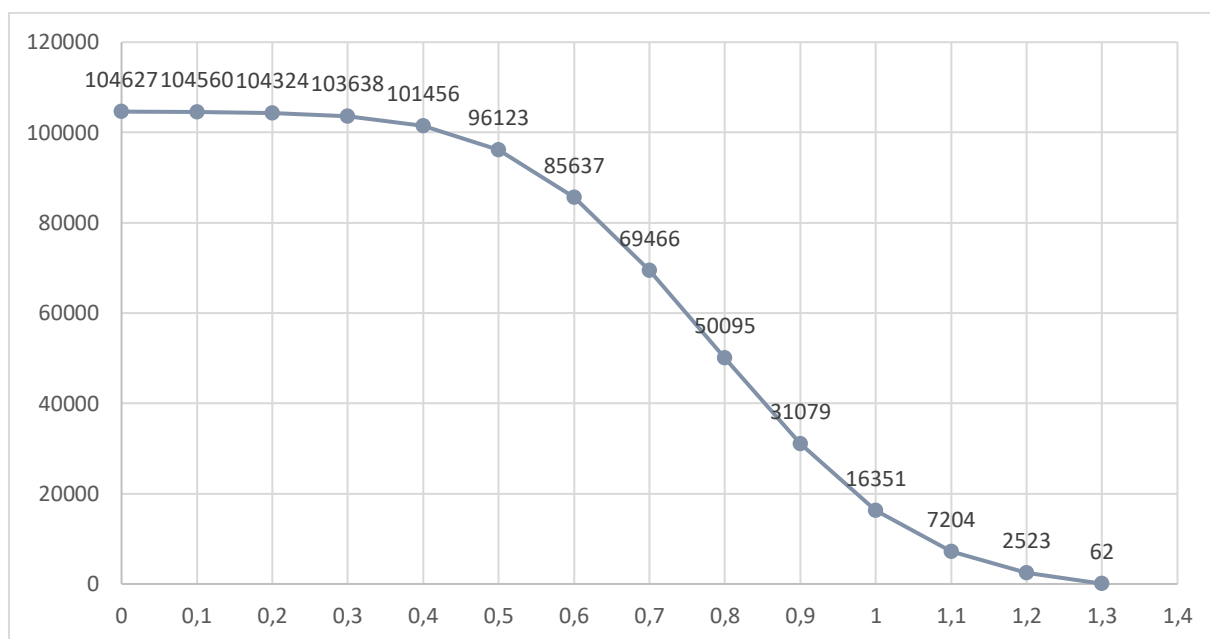
2., 3. és 4. ábra. Az OpusML által fordított korpusz eredményei jelentésmegtartás, grammatikalitás és referenciafolytonosság tekintetében, az annotátorok pontozása alapján.





5., 6. és u. ábra. Az DeepL által fordított korpusz eredményei jelentésmegtartás, grammatikalitás és referenciafolytonosság tekintetében, az annotátorok pontozása alapján.

A diagramokat megtekintve láthatjuk, hogy az eredmények részben korrelálnak a BLEU-pontokkal: úgy tűnik, hogy mindkét korpusz esetében 0,5-0,6 BLEU-pont felett már viszonylag jó (4,5 fölötti) átlagot hoz mindkét korpusz, minden vizsgálati szempontot tekintve. Mivel az eredmények egy, a korpusz méretéhez képest alapvetően kicsi minta tulajdonságait tükrözik vissza, ezért sem a fordítás, sem az eredeti, angol korpusz minőségéről nem lehet messzemenő következtetéseket levonni. A tendenciák alapján azonban érdemesnek tűnik elhagyni ez alatt a BLEU-pont alatt teljesítő mondatokat, a korpusz minőségének javítása céljából. Valószínűsíthető továbbá az is, hogy a BLEU-pontok növekedésével a korpusz minősége is pozitív irányba mozdul el – fontos azonban szem előtt tartani, hogy a korpusz mérete drasztikusan csökken az elért BLEU-pontok növelésével:



8. ábra. A korpusz méretének változása különböző BLEU-minimumok alapján.

4 A korpusz használhatóságának felmérése

A létrehozott korpusz használhatóságának mérésére három, már létező, általános, large-scale nyelvmodellt használtam fel. Kettő az MTA Nyelvtechnológiai Kutatócsoportja és a Pécsi Tudományegyetem Alkalmazott Adattudományi és Mesterséges Intelligencia csoportja által létrehozott PULI GPT-3SX (<https://juniper.nytud.hu/demo/puli>) és HILANCO GPTX (<https://juniper.nytud.hu/demo/gptx>). Ezek közül a HILANCO GPTX magyar és angol nyelven képes kommunikációra és szövegelőállításra, míg a PULI egynyelvű modell. A harmadik modell pedig az OpenAI által fejlesztett ChatGPT (<https://chat.openai.com/>, v. Dec. 15.), amely több nyelven, köztük magyarul is képes szöveget előállítani, azt értelmezni, illetve a felhasználóval kommunikálni. Fontos megemlíteni, hogy a ChatGPT-t főleg diskurzusok megvalósítására fejlesztették ki, míg a PULI GPT-3SX és a HILANCO GPTX inkább a felhasználó által megadott példaszerű szöveg alapján képes egy hasonló outputot létrehozni. Emellett a két magyar modell sokkal kevesebb paraméterrel dolgozik: a PULI GPT-3SX és a HILANCO GPTX 7 milliárd paraméterrel (vö. <https://juniper.nytud.hu/demo/nlp> utolsó megtekintés: 2023. 01. 09.), míg a ChatGPT paramétereinek száma nem elérhető hivatalosan, a modell „önbevallása” szerint 175 milliárd paraméterrel rendelkezik (<https://www.sciencefocus.com/future-technology/gpt-3/> utolsó megtekintés: 2023. 01. 09.). Ez az óriási különbség performanciabeli különbségeket is okozhat a modellek között.

A három modell mindegyike GPT-3 nyelvmodell (Brown et al. 2020) – ezen modellek jelen pillanatban a legmodernebb nyelvmodelleknek számítanak. A GPT-3-alapú modellek képesek az emberi szövegprodukciónak nagyon hasonló szövegeket előállítani, valamint az ún. few-shot learningre, melynek lényege (nyelvmodellek esetén), hogy a modell néhány példán keresztül képes absztrahálni a szövegek közös jellemzőit, és ezek alapján hasonló outputot tud előállítani. Jelen dolgozat is ilyen, few-shot learningre támaszkodó vizsgálatokat végez.

Mivel a három modell nem tökéletesen egyformán működik, így szükség volt egy olyan vizsgálati módszerre, amellyel egyrészt kiértékelhetővé válik az összeállított korpusz, illetve a modellek teljesítménye is összemérhető. Ezért a korpuszból létrehoztam egy néhány mondatpárból álló példát (promptot), amely alapján a modellnek egy – szintén a korpuszból származó – mondatot kellett egyszerűsítve visszaadnia.

A következőkben bemutatom a modellek eredményeit, valamint a használt promptokat, először a PULI GPT-3SX és a HILANCO-GPTX esetében, majd a ChatGPT-re vonatkozóan.

4.1 PULI GPT-3SX és HILANCO-GPTX modellek bemutatása

Először az MTA-PTE együttműködésében megvalósult két modellel végeztem méréseket. Mivel a PULI GPT-3SX és a HILANCO-GPTX működése nagyon hasonló, így egyszerre mutatom be őket: a két modell a felhasználó inputja alapján generálja a kiadott eredményt, így célszerűnek tűnt minél hosszabb prompttal lefolytatni a méréseket. Mindkét modell maximálisan 1024 tokenet képes befogadni, a tokenizáló viszont nem elérhető nyilvánosan, így úgy döntöttem, hogy 10 mondatpárral fogom lefolytatni a vizsgálatot: ezzel valószínűleg a lehetőségekhez képest maximalizáltam, de nem léptem túl a maximális tokenszámot.

A prompt összeállításához először az 1,25-ös BLEU-pont fölötti mondatpárokat szűrtem a korpuszban: így összesen 113 mondatpárt kaptam. Ebből kézzel kiválogattam az általam legjobbra értékelt húsz mondatpárt a következő szempontok alapján:

- a mondatok értelmesek és magyarosak,
- az egyszerűsített és az eredeti mondat logikailag nem mondanak egymásnak ellent,
- az egyszerűsített mondatban nincs több információ, mint az eredeti mondatban,
- az egyszerűsített mondat nem az eredeti másolata egy az egyben.

Így a kapott húsz mondat első tíz mondatpárját használtam példának, a tizenegyedik mondat pedig az egyszerűsítendő mondat lett. Fontos megemlíteni, hogy a korpusz 1,25 fölötti BLEU-ponttal rendelkező részében (és valószínűsíthetően az egész korpuszban) felülreprezentáltak azon mondatpárok, amelyben az egyszerűsített verzió egynél több mondatból áll – így a prompt mondatai is ilyen struktúrájú egyszerűsített mondatokból állnak. A használt prompt megtalálható a dolgozat mellékletei között. Az egyszerűsítendő mondat mindig minden esetben a *Görögországban és Cipruson ez az egyik legnépszerűbb saláta, különösen a nyári hónapokban, mivel könnyű, frissítő és könnyen elkészíthető.* volt.

Mind a PULI GPT-3SX-et, mind pedig a HILANCO GPTX-et három paraméter mentén lehet állítani: ezek a temperature, a top-k és a top-p. A temperature segítségével azt tudjuk állítani, hogy a modell milyen mennyiségű véletlenszerűséggel dolgozzon: minél magasabb ez a szám, annál nagyobb a véletlenszerűség mértéke is. A magasabbra állított temperature segítségével a modell kreatívabb válaszokat tud visszaadni, azonban ha túl magasra állítjuk ezt a paramétert, akkor a modell akár a témától eltérő vagy értelmetlen outputokat is generálhat. Az alacsonyabb temperature ezzel szemben jobb választás akkor, ha a modellel olyan feladatokat végeztetünk, amelyek esetében van „helyes” kimenet (például kérdésekre való

válaszadás) – a túl alacsony temperature esetében azonban a modell elkezdheti ismételni az általa generált válaszokat. (vö. <https://docs.cohere.ai/docs/temperature>, utolsó megtekintés: 2023. 01. 05.)

A *top-k* paraméterrel azt tudjuk állítani, hogy a kimenet generálásakor szavanként a legjobb tokenek közül mennyit vegyen figyelembe a modell. Ha a modell nem csak a legvalószínűbb tokent veszi figyelembe, akkor nagyobb az esély arra, hogy egy természetesebbnek ható kimenetet kapjunk. Ha azonban ez a szám túl magas, akkor a kisebb valószínűségű tokenek is belekerülhetnek az outputba, ami azt eredményezheti, hogy egy rossz vagy értelmetlen kimenetet kapunk.

A *top-p* paraméter hasonlóan működik: ezen paraméter azokat a tokeneket veszi figyelembe, amelyek egy bizonyos valószínűségi értéket elérnek. Például ha a *top-p*-t 0,15-re állítjuk, akkor a modell csak azokat a tokeneket veszi figyelembe az output generálásakor, amelyek 15%-os vagy annál magasabb valószínűséggel állnak a következő helyen. Amennyiben mind a *top-k*, mind pedig a *top-p* paraméter be van kapcsolva, akkor a modell először a *k*-t, majd pedig a *p*-t veszi figyelembe. (vö. <https://docs.cohere.ai/docs/controlling-generation-with-top-k-top-p> utolsó megtekintés: 2023. 01. 05.)

Mindkét modell esetében összesen 80 mérést végeztem úgy, hogy a paramétereket az összes lehetséges kombinációban kipróbáltam, az alábbi beállításokkal:

- Temperature: 0,1 és 1,0 között tizedesenként növekedve
- Top-k: 5, 10, 20, 40-es értékek
- Top-p: 0,5 és 1-es érték.

4.2 A PULI modell eredményei

A modell a 80 kísérletből 63 esetben semmit nem változtatott az egyszerűsítendő mondaton, hanem a kapott bemenetet adta vissza egyszerűsített mondatként is. Tíz esetben nem megfelelő egyszerűsítést kaptam vissza: esetenként a szemantikai mező megegyezett, azonban nem az eredeti mondat egyszerűsítése volt. Ezt az (1)-es példa szemlélteti:

(1) A saláta alapja az uborka, a hagyma és a paradicsom. (0,6 temperature, 1 top-p, 5 top-k)

Esetenként pedig egy végtelenített szósort adott vissza a modell:

(2) A görög saláta, a tzatziki, a görög saláta, a tzatziki, a görög saláta, a tzatziki... (1 temperature, 0,5 top-p, 40 top-k).

Három esetben a modell az eredeti mondat csonkolt verzióját adta vissza:

(3) Görögországban és Cipruson ez az egyik legnépszerűbb saláta. (0,7 temperature, 1 top-p, 40 top-k)

Összesen két esetben (l. 4. és 5. példa) sikerült megfelelő egyszerűsítést létrehozni, azonban ezen esetekben mindkétszer nagy információvesztés történt:

(4) A görög és a ciprusi konyhában ez a saláta nagyon népszerű. (0,8 temperature, 1 top-p, 5 top-k)

(5) Ez a saláta Görögországban és Cipruson népszerű. (0,9 temperature, 1 top-p, 5 top-k)

Látható az is, hogy ugyan a prompt minden mondatpárja esetében az információelhagyás helyett a több mondatra osztás valósult meg, itt a modell az információelhagyást preferálta. Mivel a megfelelő egyszerűsítéseknél is jelentős információvesztést produkált a modell, így a dolgozatban ezzel a modellel további vizsgálatokat nem végeztem.

4.3 A HILANCO-GPTX modell

A HILANCO-GPTX modell esetében ritkább volt az, hogy teljesen ugyanazt a mondatot adta vissza a modell: ez összesen kilencszer fordult elő. Azonban az nagyon gyakori volt, hogy a modell visszaadta az eredeti mondatot, és utána pedig az eredeti mondat csonkolt verzióját tette mögé:

(6) Görögországban és Cipruson ez az egyik legnépszerűbb saláta, különösen a nyári hónapokban, mivel könnyű, frissítő és könnyen elkészíthető.

Görögországban és Cipruson ez az egyik legnépszerűbb saláta, különösen a nyári hónapokban. – 0,1 temperature, 1 top-p, 5 top-k)

Ezt a megoldást összesen ötvenháromszor alkalmazta a modell. Megemlítendő, hogy a második mondatot minden esetben új sorba tördelte a modell, éppúgy, mint ahogy a prompt példamondatai is formázva voltak.

Ezen kívül 13 esetben nem megfelelő egyszerűsítést végzett el a modell:

(7) Görögországban és Cipruson.

Ez a saláta könnyű, frissítő és könnyen elkészíthető.

Különösen a nyári hónapokban, mivel könnyen elkészíthető. – 0,8 temperature, 1 top-p, 5 top-k).

Végül pedig négy darab megfelelő egyszerűsítést kaptam. Ezek a következők:

(8) Görögországban és Cipruson ez a saláta a legnépszerűbb saláta.

Különösen a nyári hónapokban, mivel könnyű, frissítő és könnyen elkészíthető. (0,4 temperature, 1 top-p, 5 top-k)

(9) Görögországban és Cipruson ez a saláta népszerű.

Könnyű, frissítő és könnyen elkészíthető. (0,6 temperature, 1 top-p, 40 top-k)

(10) Görögországban és Cipruson ezt a salátát szeretik a legjobban.

Görögországban és Cipruson is ezt a salátát szeretik a legjobban.

A saláta könnyen elkészíthető. (0,7 temperature, 1 top-p, 10 top-k)

(11) Görögországban és Cipruson ez a saláta egyik legnépszerűbb.

könnyű, frissítő és könnyen elkészíthető (0,8 temperature, 1 top-p, 20 top-k)

Jól látható, hogy ez a modell gyakrabban vétett „gépelési” hibákat, mint a HILANCO-GPTX: például a (11)-es egyszerűsítés esetében nem kezdte nagybetűvel a második mondatot. Ez a modell azonban képes volt olyan mondatokat is létrehozni, amelyek szerkezete nem egyezett meg az eredeti mondat szerkezetével: erre a (10)-es mondat ad jó példát. Mivel a (10)-es és (11)-es példa mondataiban azonban több helyesírási hiba vagy ismétlés volt, így ezen beállításokkal nem tartottam érdemesnek további vizsgálatokat folytatni. A (8)-as és a (9)-es példa beállításával megpróbáltam hasonló méréseket lefolytatni mint a ChatGPT esetében (erről bővebben l. 5.5-ös és 5.6-os fejezetekben), azonban a modell által visszaadott eredmények egyértelműen rosszabb minőségűek voltak, mint a ChatGPT által visszaadott eredmények, így ezt a modellt is kizártam a további vizsgálatból.

4.4 A ChatGPT modell

A ChatGPT paraméterein a felhasználó nem tud állítani, így ott sokkal kevesebb előzetes vizsgálat lefolytatására volt szükség, mint a másik két modell esetében. Az azonban már az első mérésnél kiderült, hogy a modell nem teljesít olyan jól akkor, ha túl sok inputot kap: 10, illetve 8 mondatpár beadása esetében a modell egyszerűen elkezd új mondatpárokat generálni, valamint a kapott mondatokat adja vissza. 8 mondatpár esetében például a használt példamondatra (*Görögországban és Cipruson ez az egyik legnépszerűbb saláta, különösen a*

nyári hónapokban, mivel könnyű, frissítő és könnyen elkészíthető.) a következő outputot adta vissza:

(12) Eredeti:

A földgolyóközi űrrepülés legnagyobb és legnagyobb sebességet elérő űreszközének a NASA Apollo 11 holdraszálló űrhajója volt, amely 1969. július 20-án érkezett a Holdra. Egyszerűsített:

A NASA Apollo 11 holdraszálló űrhajója a legnagyobb és legnagyobb sebességet elérő űreszköz a földgolyóközi űrrepülések során.

1969. július 20-án érkezett a Holdra.

Jól látható, hogy ahelyett, hogy a használt példamondat egyszerűsített változatát adta volna vissza, a modell egy új eredeti-egyszerűsített mondatpárt generált. A mondatpár követi a prompt mondatainak struktúráját: egy darab eredeti mondathoz több, rövidebb egyszerűsített mondat tartozik.

Túl kevés input esetében viszont fennáll a veszély, hogy a modell nem a legmegfelelőbb egyszerűsítést adja vissza. Így a modellt végül 6 mondatpárral promptoltam – ez még az a szám volt, amellyel a modell jól tudott teljesíteni, és nem generált felesleges mondatpárokat. Emellett előzetes vizsgálataim alapján úgy tűnt, hogy a modell jobban teljesít akkor, ha konkrét utasítást kap arra vonatkozóan, hogy mit csináljon a mondattal, így a promptba belekerült az „Add meg az egyszerűsített mondatot ehhez a mondathoz:” utasítás is. A hat mondatpárral és az utasítással lefuttatva kapott eredmény a következő:

(13) Görögországban és Cipruson népszerű saláta, főleg nyáron, mert könnyű, frissítő és könnyen elkészíthető.

A mondat grammatikus, nincsenek benne felesleges ismétlések. Látható, hogy ebben az esetben a ChatGPT sem preferálja a mondat kettéosztását. Az eredeti mondathoz képest a mondat szerkezete nem változott meg, azonban a *nyári hónapokban* szókapcsolat a *nyáron* szóra, illetve a *mivel* kötőszó a *mert* kötőszóra cserélődött.

Ugyan az előzetes vizsgálatok során mindhárom vizsgált modell esetében előfordultak olyan outputok, amelyek nem teljesítették az egyszerűsítés feltételeit, a ChatGPT volt az egyetlen olyan modell, amelyik minden esetben grammatikus kimenetet generált. Így úgy döntöttem, hogy a korpusz értékelését is ezzel a modellel fogom megvalósítani.

4.5 A korpusz értékelése a ChatGPT segítségével

A modell és a korpusz teljesítményét az MNSZ (v. 2.0.5 Oravecz–Váradi–Sass 2014) segítségével értékeltem. Az értékeléshez először korpusz minden alkorpuszából (sajtó, szépirodalom, tudományos, hivatalos, beszélt nyelvi, fórumhozzászólások és közösségi oldalokról gyűjtött szövegek) 5-5 mondatot gyűjtöttem. Mivel az MNSZ esetében nincs arra lehetőség, hogy alkorpuszon belül, keresőkifejezés nélkül, randomizálva mondatokat kérjünk le, így mindenképp szükség volt arra, hogy egy olyan keresőkifejezést találjak, amely a lehető legtöbb mondatot magában foglalja, hogy ebből véletlenszerű mintát vehessek. Sajnos az olyan keresőkifejezések, amelyek például mondatvégi írásjelre keresnek, hibát okoznak a rendszerben, ugyanis túl sok adat felel meg a keresőkifejezésnek, így ilyen lekérdezést nem tudtam végezni. Így az a megoldás tűnt a legműködőképesebbnek, hogy mondateleji névelőre keresek: így a keresőkifejezésem az „A|Az|Egy” lett. Természetesen ez a keresőkifejezés minden olyan mondatot kizárt a vizsgálatból, amelyek nem névelővel kezdődnek, azonban az összes alkorpuszból együttvéve még így is majdnem húszmillió megfelelő találatot sikerült kinyerni. Ezután a kapott lekérdezésből alkorpuszonként egy tízes randomizált listát kértem le. Ebből a listából kiszűrtem azokat a mondatokat, amelyek nem voltak teljes mondatok (például: `<p> A KÖVETKEZŐKBEN ÁLLAPODTAK MEG: </p> [doc#2122]`), majd a megmaradt listából az első öt mondatot vettem be a vizsgálatba. A prompt ugyanaz a hat mondatpár és az utasítás volt, mint az előzőekben, az egyszerűsítendő mondat pedig az MNSZ-ből kinyert mondat volt. A kapott mondatokat először kézzel elemeztem, majd kérdőíves vizsgálat segítségével. A kapott mondatok megtalálhatók a dolgozat mellékletei között.

A manuális vizsgálat során négy egyszerűsítési stratégia alapján osztályoztam az egyszerűsített mondatokat az eredeti párjukkal összevetve. Az osztályozási szempontokat Alva és Manchego gyűjtése alapján (Alva-Manchego et al. 2020: 4668, valamint l. 2.3. fejezet) állítottam fel, azonban külön jelöltem azokat a mondatokat, amelyek esetében az egyszerűsítés több mondatból állt. A szempontokat mondatszinten binárisan osztályoztam, azaz csak azt mértem fel, hogy megjelenik-e az adott egyszerűsített mondatban a leírt jelenség. A jelölt egyszerűsítési stratégiák a következők voltak:

- szócsere, szóelhagyás: jelölésre került, amennyiben az eredeti mondathoz képest egy vagy több szót kicserélt vagy elhagyott a modell az egyszerűsített mondat generálásakor;

- struktúraváltozás a mondatban: jelölésre került, amennyiben az eredeti mondathoz képest megváltozott a mondat belső struktúrája: például a tagmondatok száma vagy az azok közötti viszony;
- több mondat létrehozása: jelölésre került, amennyiben a generált egyszerűsítés egynél több mondatból állt;
- információelhagyás: jelölésre került, amennyiben az eredeti mondat valamilyen információja nem volt megtalálható az egyszerűsített mondatban.

A 35 mondatpárban összesen hatvankétszer jelöltem azt, hogy a fenti négy szempontból legalább egy darab, legalább egyszer jelen van a mondatban – azaz mondatonként átlagosan nagyjából 2 módosítás ment végbe. A módosítások számának lebontásáról az alábbi táblázat ad áttekintést:

	Szócsere, szóelhagyás	Struktúra- változás a mondatban	Több mondat létrehozása	Információ- elhagyás
Mondatok száma, ahol az egyszerűsítési stratégia előfordul	33	15	2	14

2. táblázat. Az egyszerűsítési stratégiák előfordulása a generált mondatokban.

Látható, hogy a leggyakoribb a szócsere volt, míg a legritkább az, egyszerűsített output egynél több mondatból állt. Az utóbbi azért is izgalmas jelenség, mert a prompt példamondataiban minden esetben több mondatra tördelt egyszerűsítés szerepelt. Megfigyelhető továbbá, hogy a szócsere vagy -elhagyás nem járt minden esetben információvesztéssel (fordítva azonban ez mindig igaz volt: amennyiben információt hagytunk el, az egyértelműen mindig szóelhagyással is járt).

A kérdőíves vizsgálat során arra kerestem a választ, hogy az egyszerűsített mondatok valóban elég jó minőségűek-e. Mivel a ChatGPT képes prompt nélküli egyszerűsítésre is, így kézenfekvőnek tűnt megvizsgálni azt is, hogy a promptolt vagy a prompt nélküli egyszerűsítést tartják-e jobbnak az adatközlők, valamint azt is, hogy a promptolt mondatok grammatikusak-e, megőrzik-e az eredeti mondat információtartalmát, és valóban egyszerűbbek-e az eredeti mondatnál. Mindezt egy kérdőívben vizsgálni ennyi szempontot nagyon megterhelő lett volna az adatközlők számára, így a két szempont mentén két kérdőívet hoztam létre. Mindkét kérdőívet öt-öt adatközlő töltötte ki, minden esetben olyanok, akik magyar alapszakot és/vagy magyar nyelv és irodalom mesterszakot és/vagy alkalmazott nyelvészet mesterszakot és/vagy irodalom- és kultúratudomány mesterszakot és/vagy magyar nyelv és irodalom tanárképzést

végeztek vagy jelenleg ezen képzések egyikében vesznek részt. Mindenképp hangsúlyozandó, hogy a kérdőív nem tud és nem is kíván reprezentatív lenni: a kérdőív célja az volt, hogy a saját értékelésen kívül más, a témában nem jártas emberek nézőpontját is érvényesíteni tudjam, kvalitatív szempontokat alkalmazva, valamint az, hogy a szakirodalomban általános humán értékelés perspektíváját is beemeljem a vizsgálatba (pl. Brunato et al. 2016, Alva-Manchego et al. 2020). A vizsgálat kis csoportban való lefolytatása egyrészt szintúgy az NLP szakirodalomban megszokott keretekhez alkalmazkodik (pl. Brunato et al. 2016 hét értékelővel, míg Vu et al. 2018 három értékelővel dolgozik), másrészt pedig az adatok átláthatóságát biztosítja: mivel a kérdőív nem a fő vizsgálati tényezőt biztosítja a dolgozatban, csupán a korpusz validálására szolgál, így a kiscsoportos vizsgálatot megfelelőnek tartottam.

A promptolt és a prompt nélküli mondatok vizsgálata során a beleegyező nyilatkozat elfogadása és az általános adatok felvétele (nem, életkor, valamint annak bejelölése, hogy a fenti szakok közül melyike[ke]t végzi vagy végezte az adatközlő) után a kérdőív kitöltője az alábbi instrukciót olvashatta: „A kérdőív következő részében a modell kimeneteinek értékelésekor két mondat közül kell választanod az alapján, hogy te melyiket tartod az eredeti mondat jobb egyszerűsítésének. Fontos, hogy a feladatnak nincs jó megoldása, bármelyik mondatot választhatod.”. A feladat sémája minden mondat esetében azonos volt: az adatközlők először idézőjelek között látták az eredeti mondatot, ezt pedig a „Melyik mondat jobb egyszerűsítése a fenti mondatnak?” kérdés követte. Végül a két lehetséges kimenetet jelölhették meg az adatközlők. A vizsgálatban a leírt feladatot 34 mondat esetében végezték el az adatközlők.⁸ A kérdőívet összesen öt fő töltötte ki, két férfi és három nő, 24 és 28 éves kor közöttiek. Az adatközlők magyar alapszakot, magyar nyelv és irodalomtanár-képzést és magyar nyelv és irodalom mesterszakot végeztek vagy végeznek.

A vizsgálat során szándékosan nem adtam semmiféle iránymutatást vagy definíciót az egyszerűsítéssel kapcsolatban – a vizsgálat célja az volt, hogy a kitöltők az EGYSZERŰ SZÖVEG általuk ismert sémáját aktiválják.

A kérdésre adott válaszok megoszlásának összesítése az alábbi táblázatban látható:

⁸ Egy, a szépirodalmi alkorpuszból vételezett mondat esetében a két kimenet teljesen megegyezett, így ezt a mondatot kihagytam a vizsgálatból.

A promptolt outputot favorizálók száma	A nem promptolt outputot favorizálók száma	Előfordulás darabszáma
0	5	2
1	4	7
2	3	9
3	2	5
4	1	7
5	0	4

3. táblázat. A promptolt – nem promptolt outputot favorizálók aránya.

Látható, hogy ugyan a nem promptolt outputot favorizáló esetek aránya kicsivel nagyobb (18-16), azonban a határon lévő esetek (2 fő – 3 fő és 3 fő – 2 fő megoszlással) igen gyakoriak – az összes esetnek több mint negyven százalékát ezek teszik ki. A fennmaradó esetekben, ahol az adatközlők válaszaiból egybehangzóbb preferencia rajzolódik ki, a promptolt mondatokkal létrejött outputot részesítik előnyben a válaszadók.

Az alkorpuszonként lebontott arányok a következő táblázatban láthatók:

Alkorpusz	Promptolt verziót favorizálók száma	Nem promptolt verziót favorizálók száma
Beszélt nyelvi	2	3
Hivatalos	1	4
Sajtó	3	2
Személyes – fórum	4	1
Személyes – közösségi	2	3
Szépirodalom	2	2
Tudományos	1	4

4. táblázat. Az adott egyszerűsítést preferálók aránya alkorpuszonként lebontva.

A válaszadók a promptolt verziót a sajtószövegek és a fórumszövegek esetében favorizálták, a szépirodalmi szövegek esetében pedig egyenlő arányban preferálták a promptolt és a nem promptolt verziót. A korábbiakban említett határesetek itt is gyakoriak: mindössze a hivatalos, a fórum és a tudományos alkorpuszban látható teljesen egyértelmű preferencia. Meglepő, hogy a tudományos alkorpusz esetében a nem promptolt verzió teljesít jobban, hiszen a PWKP korpusza a Wikipédiát használja forrásként, amely valószínűleg a tudományos alkorpuszban található szövegek témájával (és esetlegesen megfogalmazásával) mutat hasonlóságot.

A második kérdőív esetében az adatközlőknek ugyanúgy ki kellett tölteniük egy beleegyező nyilatkozatot, illetve a fent említett adataikat itt is meg kellett adniuk. Ezután a következő instrukciót olvashatták: „A következő szakaszban azt kell megjelölnöd, hogy mennyire értesz egyet standard és egyszerűsített magyar nyelvű mondatokra vonatkozó állításokkal.”. A feladatrészeknél először az eredeti, majd az egyszerűsített mondatot látták az

adatközlők (minden esetben fel volt tüntetve, hogy melyik az egyszerűsített és melyik az eredeti mondat). A feladatban három állítás esetében kellett egytől ötig terjedő Likert-skálán pontozniuk azt, hogy mennyire értenek egyet az állítással. A skála esetében az 1 az „egyáltalán nem értek egyet”, míg az 5 a „teljes mértékben egyetértek” válasznak felelt meg. Az állítások a következők voltak:

1. Az egyszerűsített mondat megfelelően kifejezi az eredeti jelentését, esetleg kihagyja a legkevésbé fontos információt.
2. Az egyszerűsített mondat magyaros, nincsenek benne nyelvtani hibák.
3. Az egyszerűsített mondatot könnyebb megérteni, mint az eredeti mondatot. (az állítások Manchego et al. 2020: 4674 alapján készültek)

A vizsgálatban leírt feladatot 35 mondatpár esetében végezték el az adatközlők. A mondatok megegyeztek az előző kérdőív mondataival. A kérdőívet összesen öt fő töltötte ki, egy férfi és négy nő,⁹ 23 és 28 éves kor közöttiek. Az adatközlők magyar alapszakot, magyar nyelv és irodalomtanár-képzést és magyar nyelv és irodalom mesterszakot, irodalom- és kultúratudomány mesterszakot és alkalmazott nyelvészet mesterszakot végeztek vagy végeznek.

A különböző állításokra kapott pontok átlagát és mediánját a következő táblázat szemlélteti:

	Átlag	Medián
1.: Információtartalom megőrzése	4,08	5
2.: Grammatikalitás, „magyarosság”	4,4	5
3.: Egyszerűség mértéke	3,417142857	4

5. táblázat. Az állításokra kapott értékelések átlaga és mediánja.

Látható, hogy az információtartalom megőrzése és a grammatikalitás területén kiemelkedően teljesít a modell: mindkét modell 4 feletti pontszámot ért el átlagosan, a medián pedig mindkét szempont esetében 5. Az egyszerűség mértékének növelését tekintve az átlagos pontszám 3,4, a medián azonban itt is magasabb ennél. Az átlagok alacsonyabb mivoltára azon mondatpárok adnak magyarázatot, amelyek helytelenek/értelmetlenek, vagy amelyek esetében az eredeti mondatot rosszul értelmezte a modell. Ilyen volt például az *Egy nagy, modern irodaházban takarít, egyszer már mesélt arról a házról, csupa üveg meg vasbeton az egész, de Albert nem is*

⁹ Az adatközlők eltértek az előző kérdőív kitöltőitől.

tudja, hol van az a ház, sőt azt sem tudja, létezik -e egyáltalán, elképzelhetetlen, hová megy és honnan érkezik meg az asszony naponta. (doc#706) mondat egyszerűsítése, melyet a modell így valósított meg: *Albert takarít egy üveg és vasbeton irodaházban, de nem tudja, hol van az. Ő nem tudja, létezik-e egyáltalán, és nem tudja, hova megy és honnan érkezik meg az a ház.* Jól látható, hogy a modellnek nem sikerült jól értelmeznie az eredeti mondat referenciáit: a takarítás esetében az asszony helyett Albertet, míg a jövés-menés folyamatánál a házat jelölte meg cselekvő ágensként.

A különböző alkorpuszok pontjainak átlagát a következő táblázat mutatja be:

Alkorpusz	Információtartalom megőrzése	Grammatikalitás, magyarosság	Egyszerűség mértéke
Beszélt nyelvi	4,52	4,64	3,56
Hivatalos	3,961538462	3,961538462	3,269230769
Sajtó	3,615384615	4	3,576923077
Személyes – fórum	4,384615385	4,769230769	3,769230769
Személyes – közösségi	3,846153846	4,153846154	2,961538462
Szépirodalom	3,92	4,32	3,28
Tudományos	3,68	4,28	2,96

6. táblázat. Az alkorpuszonkénti átlagpontok megoszlása, zölddel jelölve adott oszlop legmagasabb értékét, narancssárgával a legrosszabbat.

A táblázat adatait megfigyelve láthatjuk, hogy az információtartalom megőrzésének szempontjából a beszélt nyelvi, míg a grammatikalitás/magyarosság és az egyszerűség mértékének növekedése szempontjából a sajtó alkorpusz rendelkezik a legjobb eredményekkel. Az előző kérdőív válaszait megvizsgálva látható volt, hogy a fórumhozzászólások alkorpuszának esetében a válaszadók legnagyobbbrészt a promptolt, míg a hivatalos és tudományos szövegekből álló alkorpuszok esetében a válaszadók leggyakrabban a nem promptolt verziót részesítették előnyben (l. 6. táblázat). A pontozás tekintetében a pozitív póluson nem látszik korreláció: az információtartalom megőrzése és az egyszerűség mértékének szempontját tekintve a fórum alkorpusz közepesen teljesít, grammatikalitás szempontjából viszont egyenesen a legrosszabbul. A negatív póluson azonban talán megfigyelhető gyenge összefüggés: az információtartalom megőrzésének szempontjából a hivatalos szövegek teljesítenek a legrosszabbul (míg a tudományos szövegek a második legrosszabbul), az egyszerűség mértékének növekedését tekintve pedig a tudományos szövegek érték el a legrosszabb eredményt.

Összefoglalásképp tehát elmondható, hogy a korpusz eredményesen felhasználható már létező nyelvmodellek promptolására. A korpusz különösen jól teljesít sajtószövegek és fórumhozzászólások egyszerűsítése esetében – így tehát plauzibilisnek tűnik az, hogy az korpusz segítségével létrehozható egy mindennapi kommunikációt megkönnyítő egyszerűsítő modell.

5 Összefoglalás, kitekintés

Jelen dolgozat az első, magyar nyelvre létrehozott standard és egyszerűsített szövegpárokból álló párhuzamos korpuszt mutatta be. A dolgozat először megkísérelte a szövegegyszerűsítés elméleti megalapozását: ehhez először az egyszerűsítés fogalmát határozta meg, illetve a *szövegegyszerűsítés* fogalma mellett kötelezte el magát a nemzetközi nyelvtechnológiai szakirodalom alapján. Ezután bemutatta az intralingvális fordítás és a szövegegyszerűsítés kapcsolatát. Majd rövid áttekintést adott a szövegegyszerűsítés nyelvtechnológiai vonatkozásáról, végül pedig a dolgozat szövegfogalmát határozta meg.

Ezután a dolgozat bemutatta a korpusz összeállításának folyamatát: mivel magyar nyelvre kevés olyan egyszerűsített szöveg érhető el, amelynek standard nyelvi párja is van, így egy már létező, angol nyelvű korpusz gépi fordítása tűnt a legmegfelelőbb megoldásnak. A korpuszt két fordítómodell is átültette magyarra. A fordítások egyrészt a BLEU-pontszám alapján lettek automatikusan értékelve, másrészt pedig egy, általam és egy, a kutatástól független annotátor által egy randomizált mintán elvégzett értékelés által. Az értékelés alapján egyrészt láthatóvá vált, hogy a DeepL által fordított korpuszt érdemes használni a vizsgálathoz, másrészt pedig az, hogy a 0,5-ös BLEU-pont alatt teljesítő mondatok elhagyása valószínűleg javítja a korpusz minőségét.

Végül pedig a dolgozat bemutatta a korpusz hasznosíthatóságának ellenőrzését is: a korpuszból vételezett mintát (promptot) három, már létező nyelvmodell segítségével vizsgáltam. A kapott eredmények alapján a legjobban teljesítő nyelvmodell segítségével további vizsgálatokat végeztem, melynek eredményeit manuálisan, valamint egy kérdőíves vizsgálat segítségével értékeltem. A manuális vizsgálat a promptolt modell egyszerűsítési stratégiáit vizsgálta: a modell a szócsere/szóelhagyás stratégiáját preferálta, míg a legritkábban használt egyszerűsítési módszer a több mondatra bontás volt.

A kérdőíves vizsgálat eredményei egyrészt azt vizsgálták, hogy az adatközlők a promptolt vagy a prompt nélkül létrehozott egyszerűsítést preferálják-e. A preferencia kismértékben a prompt nélküli egyszerűsítések felé mutat. Megjegyzendő azonban, hogy azon

esetekben, amikor a preferencia egyértelműen az egyik verzióra esik (azaz ötből öt vagy ötből négy adatközlő is ugyanazt a mondatot jelölte meg jobb egyszerűsítésnek), a promptolt inputtal generált egyszerűsítések jobban teljesítenek, mint a másik csoport. A kérdőíves vizsgálat másik aspektusa a prompt segítségével előállított egyszerűsítés minőségének vizsgálata volt. A promptolt mondatok az információtartalom megőrzése és a grammatikalitás tekintetében 4 fölötti átlagot ér el, míg az egyszerűség mértékének szempontjából 3,41-es átlagot. Úgy tűnik, hogy a legjobb értékeket a beszélt nyelvi adatok és a fórumhozzászólások esetében éri el a promptolt modell.

A korpusz eredményei alapján úgy tűnik, hogy a korpuszt tanítókorpuszként használva a későbbiekben létrehozható egy olyan modell, amely sikeresen képes a szövegegyszerűsítésre.

6 Bibliográfia

2006. évi New York-i egyezmény. Kihirdette: 2007. évi XCII. törvény a Fogytékossággal élő személyek jogairól szóló egyezmény és az ahhoz kapcsolódó Fakultatív Jegyzőkönyv a kihirdetéséről.

Alva-Manchego, Fernando – Martin, Louis – Bordes, Antoine – Scarton, Carolina – Sagot, Benoît – Specia, Lucia 2020. ASSET: A Dataset for Tuning and Evaluation of Sentence Simplification Models with Multiple Rewriting Transformations. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Vancouver. 4668–4679.

Bingel, Joachim – Paetzold, Gustavo H. – Søgaard, Anders 2018. Lexi: A tool for adaptive, personalized text simplification. In: *Proceedings of the 27th International Conference on Computational Linguistics (2018)*. Association for Computational Linguistics, Santa Fe. 245–258.

Bird, Steven – Edward Loper – Ewan Klein 2009. *Natural Language Processing with Python*. O'Reilly Media Inc., Sebastopol, CA.

Brown, Tom B. – Mann, Benjamin – Ryder, Nick – Subbiah, Melanie – Kaplan, Jared – Dhariwal, Prafulla – Neelakantan, Arvind – Shyam, Pranav – Sastry, Girish – Askell, Amanda – Agarwal, Sandhini – Herbert-Voss, Ariel – Krueger, Gretchen – Henighan, Tom – Child, Rewon – Ramesh, Aditya – Ziegler, Daniel M. – Wu, Jeffrey – Winter, Clemens – Hesse, Christopher – Chen, Mark – Sigler, Eric – Litwin, Mateusz – Gray, Scott – Chess, Benjamin – Clark, Jack – Berner, Christopher – McCandlish, Sam – Radford, Alec – Sutskever, Ilya – Amodei, Dario 2020. *Language Models are Few-Shot Learners*. (<https://arxiv.org/abs/2005.14165>, utolsó megtekintés: 2023. 01. 09.)

Brunato, Dominique – Cimino, Andrea – Dell’Orletta, Felice – Venturi, Giulia 2016. PaCCSS-IT: A Parallel Corpus of Complex-Simple Sentences for Automatic Text Simplification. In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Austin. 351–361.

ChatGPT. <https://chat.openai.com/>. Utolsó megtekintés: 2023. 01. 10.

Dobos Csilla 2019. Intralingvális fordítás és szaknyelvi terminológia. In: Szoták, Szilvia (szerk.) *Diszciplínák találkozása: nyelvi közvetítés a XXI. században*. Országos Fordító és Fordításhitelesítő Iroda Zrt. (OFFI), Budapest. 186–196.

AI Docs. <https://docs.cohere.ai/docs/> Utolsó megtekintés: 2023. 01. 05.

Feist, Michele I. 2013. Experimental lexical semantics at the crossroads between languages. In: Ana Rojo – Iraide Ibarretxe-Antuñano (eds.) *Cognitive Linguistics and Translation*. Mouton de Gruyter, Berlin. 375–394.

Hansen-Schirra, Silvia – Maaß, Christiane 2020. Easy Language, Plain Language, Easy Language Plus: Perspectives on Comprehensibility and Stigmatisation. In: Silvia Hansen-Schirra–Christiane Maaß (eds.) *Easy Language Research: Text and User Perspective*. Frank & Timme, Berlin. 17–38.

HILANCO GPTX. <https://juniper.nytud.hu/demo/gptx>. Utolsó megtekintés: 2023. 01. 10.

Horváth Péter László – Ladányi Lili (szerk.) 2021. *Válogatás a könnyen érthető nyelv nemzetközi szakirodalmából*. Szegedi Tudományegyetem Juhász Gyula Pedagógusképző kar, Szeged.

Horváth Péter László 2019. A könnyen érthető kommunikáció meghatározása. *Carissimi X/2*. 8–10.

Imrényi András 2017. Az elemi mondat viszonyhálózata. In: Tolcsvai Nagy Gábor (szerk.) *Nyelvtan*. Budapest: Osiris Kiadó. 664–760.

Inclusion Europe 2009. *Information For All: European Standards for making information easy to read and understand (Easy-to-read ed.)* Inclusion Europe, Brussels. (online: https://www.inclusion-europe.eu/wp-content/uploads/2017/06/EN_Information_for_all.pdf, utolsó megtekintés: 2022. 11. 26.)

Inclusion Europe 2013. *Információ mindenkinek. A könnyen érthető kommunikáció európai alapelvei*. ÉFOÉSZ – Inclusion Europe, Budapest–Brüsszel. https://www.inclusion-europe.eu/wp-content/uploads/2017/06/HU_Information_for_all.pdf, utolsó megtekintés: 2022. 11. 06.)

Jakobson, Roman 1959. On Linguistic Aspects of Translation. In Reuben Brower (ed.) *On Translation*. Harvard University Press, Cambridge, MA és London, England. 232–239.

Kajiwarra, Tomoyuki – Komachi, Mamoru 2016. Building a Monolingual Parallel Corpus for Text Simplification Using Sentence Similarity Based on Alignment between Word Embeddings. In: *Proceedings of COLING 2016, the 26th International Conference on*

- Computational Linguistics: Technical Papers*, COLING 2016 Organizing Committee, Osaka. 1147–1158.
- Kugler Nóra 2022. Milyen kérdésekhez és válaszlehetőségekhez vezet az összetett mondat leírásában a kontextus módszeres figyelembevétele?. In: Tátrai Szilárd – Tolcsvai Nagy Gábor (szerk.) *A magyar mondat kontextuális környezete*. ELTE Eötvös Kiadó, Budapest.
- Maaß, Christiane 2020. *Easy Language – Plain Language – Easy Language Plus. Balancing Comprehensibility and Acceptability*. Frank & Timme, Berlin.
- Megna, Angelo Luigi – Schicchi, Daniele – Lo Bosco, Giosué – Pilato, Giovanni 2021. A Controllable Text Simplification System for the Italian Language. In: *2021 IEEE 15th International Conference on Semantic Computing (ICSC)*. 191–194.
- Nisioi, Sergiu – Stajner, Sanja – Ponzetto, Simone Paolo – Dinu, Liviu P. 2017. Exploring neural text simplification models. In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Vancouver. 85–91.
- Oravecz Csaba, Váradi Tamás, Sass Bálint: The Hungarian Gigaword Corpus. In: *Proceedings of LREC 2014*, 2014.
- Papineni, Kishore – Roukos, Salim – Ward, Todd – Zhu, Wei-Jing 2002. BLEU: a method for automatic evaluation of machine translation. In: *ACL-2002: 40th Annual meeting of the Association for Computational Linguistics*. 311–318.
- PULI GPT-3SX. <https://juniper.nytud.hu/demo/puli>. Utolsó megtekintés: 2023. 01. 10.
- Sanders, Julie 2006. *Adaptation and Appropriation*. Routledge, London.
- Siddharthan, Advaith 2014. A survey of research on text simplification. *ITL - International Journal of Applied Linguistics* 165, 2 (Dec. 2014), 259–298.
- Sikka, Punardeep – Mago, Vijay K. 2020. *A Survey on Text Simplification*. ArXiv abs/2008.08612. Kizárólag online megjelenés, elérhetőség itt: <https://arxiv.org/pdf/2008.08612.pdf> (utolsó megtekintés: 2022. 11. 26.)
- Tátrai Szilárd 2017. Pragmatika. In: Tolcsvai Nagy Gábor (szerk.) *Nyelvtan*. Osiris Kiadó, Budapest. 897–1058.

- Tiedemann, Jörg – Thottingal, Santosh 2020. OPUS-MT – Building open translation services for the World. In: *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*. European Association for Machine Translation, Lisszabon. 479–480.
- Tolcsvai Nagy Gábor (szerk.) 2017. *Nyelvtan*. Osiris Kiadó, Budapest.
- Tolcsvai Nagy Gábor 2006. *Szövegtan*. In: Kiefer Ferenc (szerk.) *Magyar nyelv*. Akadémiai Kiadó, Budapest. 149–174.
- Tolcsvai Nagy Gábor 2017. Bevezetés. In: Tolcsvai Nagy Gábor (szerk.) *Nyelvtan*. Osiris Kiadó, Budapest. 23–71.
- Vu, Tu – Hu, Baotian – Munkhdalai, Tsendsuren – Yu, Hong 2018. Sentence Simplification with Memory-Augmented Neural Networks. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. Association for Computational Linguistics, New Orleans. 79–85.
- Whyatt, Boguslawa 2017. Intralingual Translation. In: John W. Schwieter – Aline Ferreira (eds.) *The Handbook of Translation and Cognition*. John Wiley & Sons, New Jersey. 176–192.
- Xu, Wei – Napoles, Courtney – Callison-Burch, Chris 2015. Problems in Current Text Simplification Research: New Data Can Help. *Transactions of the Association for Computational Linguistics*. 2015/3. 283–297.
- Xu, Wei – Napoles, Courtney – Pavlick, Ellie – Chen, Quanze – Callison-Burch, Chris 2016. Optimizing Statistical Machine Translation for Text Simplification. *Transactions of the Association for Computational Linguistics* 2016/4. 401–415.
- Zhu, Zhemín – Bernhard, Delphine – Gurevych, Irina 2010. A Monolingual Tree-based Translation Model for Sentence Simplification. In: *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*. Coling 2010 Organizing Committee, Peking. 1353–1361.