Az evangéliumok ómagyar és középmagyar fordításainak információsűrűsége [1]

Rácz Péter

Budapesti Műszaki és Gazdaságtudományi Egyetem
Kognitív Tudományi Tanszék

The information density of a text captures the extent to which we can use elements of the text to predict other elements. The mathematical tools that express information density are related to linguistic, structural properties of the text, such as syntactic and paradigmatic complexity. This allows us to use information density as a general linguistic measure of text. This paper compares mediaeval and early modern Hungarian translations of the first four books of the New Testament in terms of information density. The results allow an interpretation according to which the successive translations reveal the gradual emergence of Hungarian literary language.

Keywords: information theory, diachronic linguistics, Hungarian

Kulcsszavak: információelmélet, történeti nyelvészet, magyar

1. Bevezetés

1.1. A nyelv információsűrűsége

A nyelvi rendszer tükrözi a nyelvhasználatot (Zipf 1936) és a nyelvet használó közösség szerkezetét (Weinreich – Labov – Herzog 1968). A klasszikus leíró nyelvészet számára ezek a nyelvnek esetleges, elméleti jelentőséggel nem rendelkező jellemzői (Bloomfield 1933; Sapir 1912). Az újabb kutatások azonban fontos magyarázóerőt tulajdonítanak annak, hogy a természetes nyelveket hogyan alakítja a használatuk (Beckner és mtsai 2009). Ennek egyik példája az információszerkezet.

Egy nyelvi közlés információszerkezete hozzávetőlegesen azt fejezi ki, menynyire nehéz a szöveg egyes elemeiből kikövetkeztetni más elemeit (Shannon 1948). Dye és mtsai (2018) adnak egy természetes nyelvi példát: a beszélők visszatérően használnak olyan redundáns, „eposzi" jelzős szerkezeteket, amelyek nagyon erősen megjósolják az utolsó elemüket, pl. „egy jó korsó hideg (sör)" vagy „egy finom csésze (kávé)". A szerzők kvantitatív módszerekkel mutatják meg, hogy ezek a látszólagos modorosságok fontos szerepet játszanak a nyelvfeldolgozásban. Előre jelzik a tartalmas szavakat, ezzel megkönnyítve a befogadó számára

[1] A kutatás a Bolyai János Kutatási Ösztöndíj szakmai támogatásával készült.

DOI: 10.15776/NyK.2025.121.6

a mondat értelmezését. A szövegek információsűrűsége eltérhet műfajok között (egy liturgikus szöveg jellemzően jóval repetitívebb lesz, mint egy avantgárd költemény), de nyelvek között is. Dye és mtsai példája rávilágít arra, hogy ezeknek a különbségeknek az okai a nyelvhasználatból fakadnak. Ugyanígy a homéroszi szövegek eposzi jelzői valószínűleg eredetileg a szóbeli hagyomány memnotechnikai eszközei voltak (Parry 1933).

A megjósolhatóság eszerint a természetes nyelv fontos attribútuma. Az alaktani összetettség és a szavak eloszlása alapvetően érzékeny arra, hogy a nyelvi jel különböző pontjai mennyire megjósolhatóak a befogadók számára (Gibson és mtsai 2019). A nyelvi rendszerek igyekeznek egy olyan konstans szinten tartani az információsűrűséget, ahol a befogadó optimálisan tudja feldolgozni a hallottakat/látottakat. Az erre felhasznált nyelvi eszközök azonban nem determinisztikusak, hanem sztochasztikusak, ezért az információsűrűség (a nyelvi komplexitás) szövegeken belül, szövegek között, sőt nyelvek között is ingadozni fog (Piantadosi – Tily – Gibson 2011).

Ezek a nyelvi összetettségben megjelenő, szisztematikus változások összefüggéseket mutathatnak a beszélőközösség szerkezetével (Lupyan – Dale 2010; Nettle 2012). Ezeknek az összefüggéseknek a léte vitatott, két okból. Egyrészt a nyelvi összetettségnek számos mérőszáma van (Bentz és mtsai 2016), másrészt pedig a nyelvi közösségek és a nyelvek kapcsolata rendkívül összetett (Shcherbakova és mtsai 2023). Ettől függetlenül ezek az eredmények rámutatnak, hogy egy szöveg szerkezete elárulhat valamit a szöveg közvetlen és tágabb szociokulturális kontextusáról. A szakirodalom komoly figyelmet fordít arra, hogyan lehet egy szöveg összetettségét és információsűrűségét mérni, és ebből milyen következtetéseket vonhatunk le a nyelvet alkotó közlések egymáshoz való viszonyáról, a nyelvi rendszer szerkezetéről, és a nyelvi rendszer és a beszélőközösség kapcsolatáról (ld. Dębowski – Bentz 2020).

Ezeknek a kérdéseknek a jellegzetes megközelítései több nyelv szinkrón állapotát vetik össze (Atkinson – Kirby – Smith 2015; Raviv – Meyer – Lev-Ari 2019). Tanulmányomban egy alternatív megközelítésre teszek kísérletet, amennyiben a rendelkezésre álló nyelvemlékeken keresztül a magyar nyelv különböző történeti korszakait hasonlítom össze egymással. Azért, hogy a kutatást lehorgonyozzam, egymáshoz kapcsolódó szövegeket fogok vizsgálni, az Újszövetség első négy könyvének ómagyar, középmagyar és modern fordításait.

Kutatási kérdéseim a következők.

1. Vannak-e szisztematikus információszerkezeti különbségek a fordítások között?
2. Magyarázza-e ezeket a különbségeket a szövegek egymáshoz és a saját társas, kulturális környezetükhöz fűződő viszonya?

A továbbiakban először röviden áttekintem a magyar bibliafordítások történeti és kulturális hátterét, majd bemutatom az elemzésre használt korpuszt és az információsűrűség általam kiválasztott és meghatározott mérőszámait, végül pedig az összehasonlítás eredményeit és korlátait tárgyalom.

1.2. Az evangéliumok ómagyar és középmagyar fordításai

Az Újszövetség első négy könyvének, Máté, Márk, Lukács és János evangéliumának hét teljes ó-, ill. középmagyar fordítása létezik (Madas 1998; Simon – Sass 2012; Simon – Kalivoda 2020), ezeket és a tanulmányban referenciaként használt négy modern fordítást az 1. táblázat sorolja fel.

1. táblázat: A tanulmányban használt evangéliumfordítások (forrás: Simon – Kalivoda 2020). Az ómagyar és középmagyar szövegek betűhű és normalizált (Vadász – Simon 2018) változatban, a huszadik és huszonegyedik századi szövegek modern helyesírás szerint kerültek felhasználásra.

sorszám	név	keletkezés éve	szerző	feltételezett források	formátum
1	Müncheni kódex	1466	ismeretlen	vitatott	betűhű, normalizált
2	Jordánszky-kódex	1516–1519	ismeretlen	Vulgata	betűhű, normalizált
3	Pesti Gábor: Újtestamentum	1536	Pesti Gábor	Vulgata	betűhű, normalizált
4	Sylvester János: Újtestamentum	1541	Sylvester János	Vulgata, görög	betűhű, normalizált
5	Heltai Gáspár: Újtestamentum	1565	Heltai Gáspár	Vulgata, görög, Luther	betűhű, normalizált
6	Károli Gáspár Vizsolyi Bibliája	1590	Károli Gáspár	Vulgata, görög, Sylvester	betűhű, normalizált
7	Káldi György fordítása	1626	Káldi György	Vulgata	betűhű, normalizált
8	Károli Gáspár revideált fordítása	1908			modern

9	Káldi Neovulgáta	1997			modern
10	Szent István Társulati Biblia	2003			modern
11	Magyar Bibliatársulat újfordítású Bibliája	2014			modern

A jelen tanulmány nem vállalkozhat arra, hogy ezeknek a szövegeknek a keletkezését, ennek tágabb összefüggéseit, és a róluk rendelkezésre álló filológiai ismereteket kiterjedten tárgyalja. Alapgondolatunk ugyanakkor az, hogy a szövegek valamilyen módon koruk nyelvhasználatát tükrözik, ezért fontos, hogy röviden beszéljünk arról, mi a fordítások történeti közege, és hogyan viszonyulnak egymáshoz és a fordítók által felhasznált forrásszövegekhez.

Az első elérhető, tizenötödik századi fordítás szerzőinek személye vita tárgya (Szabó 1989; Korompay 2015), a második (tizenhatodik századi) fordítás szerzője pedig ismeretlen (Erdős 1906). A motivációk, a társadalmi közeg hatása, és az eredeti szöveghez, szövegekhez fűződő viszony azonban azoknál a fordítóinknál sem egyértelmű, akiknek a nevét megőrizte a történeti emlékezet. Közös bennük, hogy fordításaik a reneszánsz humanizmus, a reformációt körülvevő vallási mozgalmak és a török hódítás összefüggéseiben keletkeztek, és részben ezekre az összefüggésekre reflektáltak (Erdős 1937; Madas 1998; Holland 2019).

A szövegek szerkezetét a fordítók döntésein túl a kulturális hagyományok, idegen hatások és részben akár (a nyomtatott változatokban) a rendelkezésre álló betűkészletek is befolyásolták. Ha nem is teljes mértékben, de jelentősen meghatározta a fordítók döntéseit, hogy milyen forrásszövegekkel dolgoztak, és mennyire akarták megtartani az eredeti szövegek strukturális jellemzőit (Hegedüs 2013). A Jordánszky-kódexben fennmaradt fordítás valószínűleg a Vulgatát használta forrásként, ahogyan Pesti 1536-os fordítása is. Sylvester János szövegére nagy hatása volt a humanista tradíciónak. Sylvester a latin fordítás mellett a görög eredetihez is visszatért, és nagy hatást gyakorolt a modern magyar vallásos nyelv terminológiájára (Zvara 2017). A német anyanyelvű Heltai a német Luther-Bibliát is felhasználta, Károli jelentős mértékben épített Sylvester fordítására és az eredeti szövegekre, míg Káldi főleg a *Vulgatá*t vette alapul. A fordítások, különösen a későbbi fordítások, egymásra is reflektáltak (Erdős 1906; Máthé 2004).

A szövegek jelentős része a török betörés és Buda elfoglalása után keletkezett, egy olyan történelmi helyzetben, amely a magyarul beszélő nyelvi közösség tartós széttöredezésével járt együtt (Nádor 2000). Erről az időszakról a demográfiai ismereteink is rendkívül bizonytalanok (Fügedi 1992).

Ezeket a fordításokat tehát, bár koruk nyelvállapotát valószínűleg hűen tükrözik (Hegedüs 2013), nem kezelhetjük objektív, természetes nyelvi mintaként. Latin és görög nyelvű forrásaik kiválasztásában, az azokkal létesített viszonyukban szelektívek a fordítók, miközben jellemzően tudatában vannak az őket megelőző magyar (sőt, német) nyelvű fordításoknak is. Ne feledjük, hogy egyrészt a forrásművek korabeli kezelése sem volt konzisztens (a bibliai fejezetek és versek számozása a 16. század végére lett egységes, Ruden 2023), másrészt a kor vallási mozgalmainak egy bibliafordítás legalább annyira mozgatórugója, mint terméke volt. A közös forrásszöveg, a közöttük meglévő folyamatosság és az eltérő nyelvi és társadalmi összefüggések azonban egyedi lehetőséget biztosítanak arra, hogy a fordításokat információelméleti szempontból is összevessük egymással. A módszernek nemzetközi és magyar előképei is vannak (Resnik – Olsen – Diab 1999; M. Pintér 2024).

2. Módszertan

2.1. Fordítások

A feldolgozott ómagyar, középmagyar és újkori/legújabb kori szövegek listáját az 1. táblázat tartalmazza. A fordításokból minden esetben Máté, Márk, Lukács és János evangéliumát használtam fel. A szövegeket a Párhuzamos Bibliaolvasó projekt (Simon – Kalivoda 2020) adatbázisából vettem át. Az ómagyar és középmagyar szövegeket két formában, betűhű változatban, illetve normalizált helyesírással dolgoztam fel. A normalizálás kisebb részben (mint az eredeti Károli fordítás, a Vizsolyi Biblia esetén) kézzel, nagyobb részében automatikus természetesnyelv-feldolgozó eszközökkel készült, kézi ellenőrzéssel (Vadász – Simon 2018). A szövegek a modern kanonizálás elveinek megfelelően a könyv, fejezet, vers felosztást követik. A 2. táblázat a betűhű és a normalizált változat közötti különbséget a Máté 9,6–13 passzussal illusztrálja, a Müncheni kódexben és a revideált Károli Bibliában.

2. táblázat: Máté 9,6–13 a Müncheni kódexben és a revideált Károli Bibliában. A betűhű példa a digitalizált szöveg konvencióit követi (ld. a „==” jel használatát annak jelölésére, hogy a szöveg szétválaszt egy, az eredeti változatban egybeírt alakot).

Müncheni kódex (betűhű)	Müncheni kódex (normalizált)	Károli revideált
Tu̇ aʒeꝛt ig imadkoʒiatoc Mi at’ăc ki vag meńńècbèn / Scenteltèſ-ſec te== ==nèuèd Iǫẏǫn te oꝛʒagod Legen te akaratod / mikent meńńen 7 aʒonkent fǫldǫn / Mi tèſti keńèꝛǫnc felet valo kėnèꝛèt aggad mu̇nèkǫnc ma Es bočaſ-ſad mu̇nèkǫnc mu̇ vėtètǫnkèt / mikent es mu̇ bočatonc nèkǫnc vėtèttecnᶜ / Es ne vig mūket kėſeꝛtètbè / de ʒabadoch mūkèt gonoʒtol Amen	ti azért így imádkozzatok : mi Atyánk , ki vagy mennyekben , szenteltessék te neved , jöjjön te országod , legyen te akaratod , miként mennyen és azonként földön . mi testi kenyerünk felett való kenyeret adjad minekünk ma , és bocsássad minekünk mi vétetünket , miként is mi bocsátunk nekünk vétetteknek . és ne vígy minket kísértetbe , de szabadíts minket gonosztól . ámen .	Ti azért így imádkozzatok: Mi Atyánk, ki vagy a mennyekben, szenteltessék meg a te neved; Jőjjön el a te országod; legyen meg a te akaratod, mint a mennyben, úgy a földön is. A mi mindennapi kenyerünket add meg nékünk ma. És bocsásd meg a mi vétkeinket, miképen mi is megbocsátunk azoknak, a kik ellenünk vétkeztek; És ne vígy minket kísértetbe, de szabadíts meg minket a gonosztól. Mert tiéd az ország és a hatalom és a dicsőség mind örökké. Ámen!

2.2. Az információsűrűség mutatói

A természetes nyelvek információelméleti megközelítései az információsűrűségnek több mutatóját ismerik. A magasabb szintű leírások a nyelvtan és az alaktan szabályaival operálnak, és olyan kérdéseket tesznek fel, mint hogy egy adott szónak hány alakja van, és ezeket mennyire könnyű megjósolni egymásból (Ackerman – Malouf 2013). Az általunk használt szövegek sajátossága azonban, hogy egyértelműsített alaktani annotációk (az egyes szavak szófaja, alaktana stb.) 112nt részben állnak rendelkezésre, és a nem következetes ortográfiából, illetve a történeti nyelvi alakok jelenlétéből adódóan nem triviális ilyeneket előállítani sem. A Müncheni kódexre és a Jordánszky-kódex újszövetségi részére rendelkezésre áll automatikus, kézzel ellenőrzött morfológiai elemzés, ahogyan más ómagyar szövegekre is, de a többi fordítás ilyen szempontból elemzetlen (ld. Simon – Sass 2012). A nyelvi összetettség nyelvtani támaszték nélküli megállapításának több gyakran használt eszköze létezik, mint az entrópia, a minimális leírás hossza (minimum description length), vagy a típus/token arány (Baayen 2001; Grünwald 2007; Piantadosi – Tily – Gibson 2012; Bentz és mtsai 2017).

A nyelvi entrópia a szöveget mint az egyedi szótípusok valószínűségi eloszlását kezeli, és azt számszerűsíti, hogy ebben az eloszlásban az egyes típusok mennyire megjósolhatóak. A típusokhoz rendelt megjósolhatóságok eloszlásának entrópiája a szövegből alkotott nyelvi modell bizonytalanságát vagy információs töménységét fejezi ki. Én két mérőszámot használok, ezek a teljes szövegre számolt unigram entrópia és a feltételes bigram entrópia. Az unigram entrópia a szövegben lévő szavak eloszlásának általános bizonytalanságát/kiszámíthatatlanságát méri azon keresztül, hogy milyen valószínű egyes szavak önálló megjelenése a szövegben. Azt tükrözi, hogy átlagosan mennyi információt nyerünk, ha a szövegben egy új szó megjelenik. A feltételes bigram entrópia azt számszerűsíti, mennyire tudunk egy szót az azt megelőző szó alapján előre jelezni. Azt ragadja meg, hogy bármilyen szó milyen valószínűséggel követ egy másik konkrét szót a teljes szövegben. Egy szövegnek egy unigram entrópia és egy bigram entrópia mérőszáma van. Az első általában a szavak eloszlásáról mond valamit, a második pedig arról, hogy a szöveget milyen konzisztens és milyen gyakori szópárok alkotják. A log2 skálán számolt entrópiából a $2^{\text{entrópia}}$ képlet segítségével számolt bizonytalanság (*perplexity*) ezt az információsűrűséget egy, a normáleloszláshoz közelebbi skálán fejezi ki. Ezt a bizonytalansági mutatót hívom a továbbiakban entrópiának (Kobayashi 2014; Bentz és mtsai 2016).

A legrövidebb leírás (*minimum description length, MDL*) fogalma a szöveg információsűrűségét úgy ragadja meg, mint a szövegről készíthető legrövidebb leírás hosszát. Mivel a szövegek redundanciákat tartalmaznak, ezért tömöríthetőek, és a tömörítés mértéke jól visszaadja azt, hogy a szöveg a hosszához képest mennyi információt tartalmaz. Ez a mérőszám alapvetően a Kolmogorov-komplexitás általánosabb fogalmából vezethető le (Li – Vitányi 2019). Én a legrövidebb leírás hosszát elosztom az eredeti szöveg hosszával, ezért az általam használt mérőszám a szöveg tömöríthetőségét súlyozza a szöveg hosszúságával.

Két további mérőszámom a szöveg hossza, amelyet úgy kapok meg, hogy megszámolom a szavakat, illetve a típus/token arány, amelyet úgy, hogy megnézem, egy adott szövegben hány egyedi szó fordul elő, és ezt a számot elosztjuk a szavak számával. Szó alatt itt minden esetben karaktersort értek. Az összes mérőszámot tokenizált szövegekre állapítom meg. A tokenizált szöveg kisbetűsíti az eredeti szöveget, eltávolítja az írásjeleket, valamint a nem alfanumerikus karaktereket, és aztán a szóközök által elválasztott karaktersorokat tekinti szavaknak.

Az információsűrűségnek ezek a mérőszámai egymással is összefüggenek. Ennek megértéséhez Saussure terminológiáját hívjuk segítségül. Saussure a nyelvi összetettségnek két formáját különbözteti meg, ezek a paradigmatikus összetettség (hány alakja van egy szótőnek) és a szintagmatikus összetettség (milyen szósorok kódolnak egy nyelvtani relációt) (Saussure 1916). A paradigmatikus összetettség fogalmába tágan értve beletartoznak olyan aspektusok is, mint hogy milyen széles a szöveg által használt szókincs, és a szóalakok írásmódja mennyire

konzisztens. A szintagmatikus összetettség fogalmába beletartozik az is, hogy a szöveg milyen formulákat, ismétlődő konstrukciókat használ rendszeresen. Ez a két faktor és a szöveg hossza befolyásolja a mérőszámokat. A szöveghossz viszonya a szintagmatikus és paradigmatikus összetettséggel nem lineáris.

Ha egy szöveg paradigmatikus komplexitása alacsony, a szöveg ugyanazt az aránylag kevés szóalakot használja fel újra és újra, ezért az unigram entrópia (hányféle szóval folytathatjuk a szöveget bármelyik pontján) és a típus/token arány (sok vagy kevés szóalak található a szövegben) alacsonyak lesznek. A paradigmatikus komplexitás közvetve befolyásolja a bigram entrópiát (hiszen ez érzékeny a bigram első szavának valószínűségére, amit az unigram entrópia ad meg) és az adható legrövidebb leírás hosszát (hiszen ha egy szöveg kevés szót ismételget, könnyebb lesz tömöríteni).

A bigram entrópia és a legrövidebb leírás hossza a szöveg szintagmatikus komplexitását is megragadják. Ha a szöveg sok, ismétlődő formulát használ, akkor egyrészt könnyebb lesz megmondani, hogy mi egy tetszőleges szópár második tagja, másrészt pedig könnyebb lesz tömöríteni a szöveget.

A szövegben a típusok száma a szöveg hosszával arányosan először gyorsan, majd lassabban növekszik (ezt a szavak nyelvi eloszlásából következő viszonyt fejezi ki a Heaps-törvény, van Leijenhorst – Van der Weide 2005). Ez befolyásolja a szóeloszlásokat, az unigram entrópiát, és – közvetve – a bigram entrópiát is, amelyek érzékenyek a típusok arányára. Befolyásolja a legrövidebb leírás hosszát is: ha sok típus van egy szövegben, akkor nehezebb tömöríteni/rövidebb leírást adni róla. A legrövidebb leírás hosszát mi elosztjuk az eredeti szöveg hosszával, de mivel a szöveghossz és a típusok száma közötti viszony nem lineáris, ez a súlyozás csak részben ad számot a szöveghossz hatásáról.

Amint látható, az entrópia, a legrövidebb leírás hossza és a típus/token arány a szöveg paradigmatikus komplexitásának („szókincsgazdagságának”) és szintagmatikus komplexitásának („formulaikusságának”) különböző, egymást kiegészítő megközelítései. A szöveghossz mindegyikkel összefügg, mivel azonban fordításokkal dolgozunk, látni fogjuk, hogy a szöveghossz kevésbé szór a szövegek között, elsősorban azért, mert ezek ugyanabból a forrásból vagy forrásokból készültek.

2.3. Információsűrűség és történeti szövegek

Történeti szövegek esetén a szavak száma érzékeny lesz arra, hogy a szöveget átíró szakemberek milyen döntéseket hoztak, a típus/token arány pedig sokban fog múlni azon, hogy mennyire konzisztens a szöveg ortográfiája, hiszen egy szónak két olyan változata, amelyek csak egyetlen karakterben térnek el, már két alaknak számítanak. A hosszabb, nagyobb típus/token arányt mutató szövegek információsűrűsége nagyobb lesz. Az inkonzisztens ortográfia hatásait mérséklendő beleveszszük az összehasonlításba a helyesírás szerint normalizált szövegváltozatokat is.

A Párhuzamos Bibliaolvasó adatbázisból kinyert szövegeket fordítás, változat (betűhű, normalizált vagy modern), evangélium és fejezet szerint (de nem vers szerint) csoportosítottam, majd az így kapott szövegekre kiszámoltam az unigram és bigram entrópiát, a legrövidebb leírás arányát, a szóhosszt és a típus/token arányt. Az így kapott adattábla 11 fordítás összesen 18 változatában (betűhű, normalizált, modern) Máté könyvének 28, Márk 16, Lukács 24 és János könyvének 21 fejezetéhez rendel mérőszámokat.

A szövegfeldolgozást az R statisztikai környezetben végeztem el (R Core Team 2025). A bizonytalansági mérőszám alapjául szolgáló entrópiát az *entropy* csomag (Hausser – Strimmer 2021), az összetettségi mérőszám alapjául szolgáló legrövidebb leírást pedig a Lempel-Ziv tömörítő algoritmus (Zeeh 2003) gzip implementációjával (Gailly – Adler 1992) hoztam létre. A mérőszámokat a szövegek (kisbetűsítést és az írásjelek eltávolítását követően a szóközök segítségével) tokenizált változatára számoltam ki. A mérőszámok összefoglalóját a 3. táblázat tartalmazza. A kódot és adatokat tartalmazó adattár itt található: https://doi.org /10.5281/zenodo.17086110.

3. táblázat: Az információsűrűség felhasznált mérőszámai

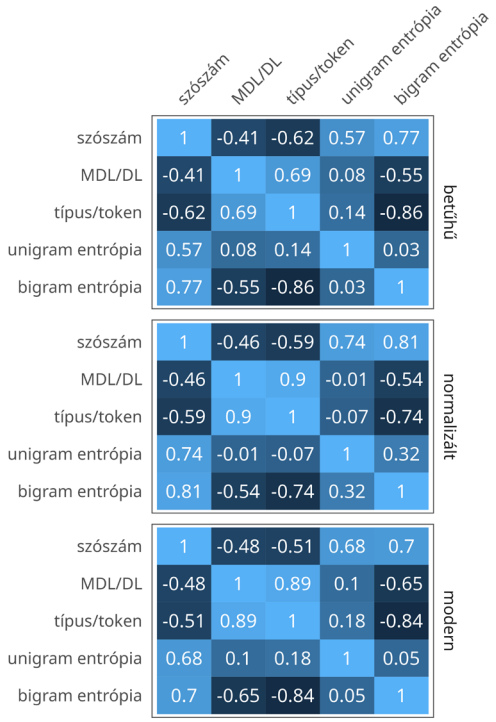
mérőszám	**meghatározás**	**rövid leírás**	**hivatkozás**
unigram entrópia	hányféle szóval lehet folytatni egy tetszőleges pontján a szöveget	$H(W) = -\Sigma(p * \log2(p))$	Hausser – Strimmer 2021
bigram feltételes entrópia	hányféle szó állhat egy tetszőleges szó után	$H(W2\|W1) = \Sigma P(w1,w2) * \log 1/P(w2\|w1)$	Hausser – Strimmer 2021
legrövidebb leírás aránya (minimum description length/description length)	a szövegre adható legrövidebb kimerítő leírás hossza, elosztva a szöveg hosszával	len(gzip(text))/len(text)	Gailly – Adler 1992
típus/token arány	a szövegben szóközökkel elválasztott egyedi karaktersorok száma elosztva a szavak számával	len(set(tokens))/ len(tokens)	Bentz és mtsai 2016
szavak száma	a szövegben szóközökkel elválasztott karaktersorok száma	len(tokens)	Bentz és mtsai 2016

Mint azt a 2.2. pontban láttuk, az információsűrűség mérőszámai erősen összefüggnek. Az öt mérőszámra ezért külön hierarchikus lineáris modelleket (Bates – Maechler – Bolker 2011) illesztettem, amelyek a mérőszámot a fordításból (ld. 1. táblázat) jósolják meg, feltételezve egy, az evangélium metszéspont alá beágyazott fejezet metszéspontot. Erre a 3.3 pontban visszatérünk.

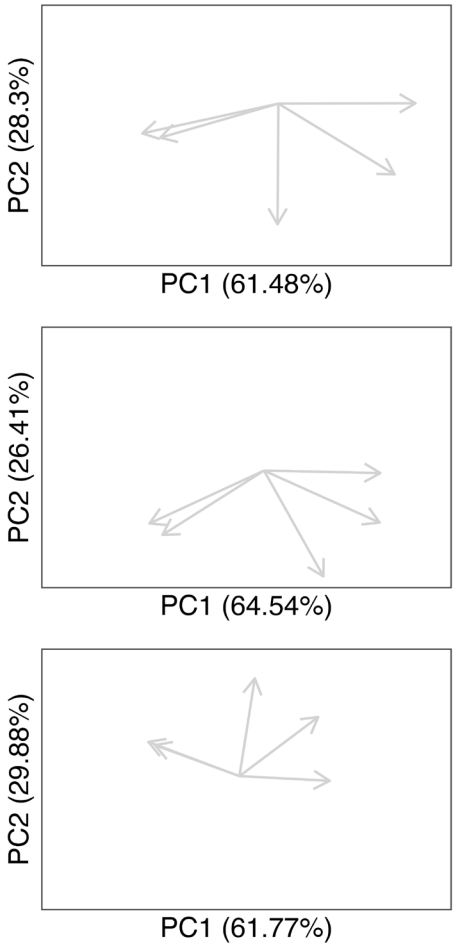
3. Eredmények

3.1. A mérőszámok összefüggései

A 2.2. pontban körüljártuk az elméleti hátterét annak, hogy a nyelvi információsűrűség mérőszámai miért függenek össze egymással. Ezeket az összefüggéseket az adatokban is megtalálhatjuk. Az 1. ábra a fejezetekre számolt mérőszámok közötti Spearman-korrelációkat mutatja. A 2. ábra a mérőszámok közötti többdimenziós viszonyt egy főkomponens elemzés segítségével, vektorosan ábrázolja.



1. ábra: A mérőszámok Spearman-korrelációs együtthatói a betűhű (fent, n = 7 fordítás * 89 fejezet), normalizált (középen, n = 7 * 89) és modern (lent, n = 4 * 89) szövegekre kapott fejezetszintű számok eloszlásaiból.

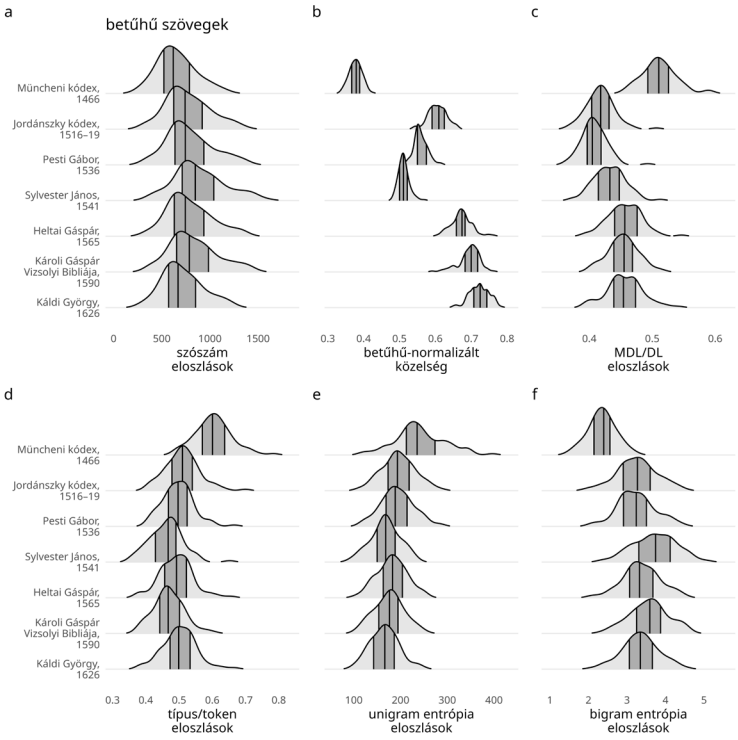


2. ábra: A mérőszámokra épített főkomponens-elemzés eredményei, az első és a második komponens viszonya betűhű (fent), normalizált (középen) és modern (lent) szövegekben.

Azt látjuk, hogy a mérőszámok mindhárom szövegtípusban erősen összefüggenek. Ezt mutatja a főkomponens-elemzés is, hiszen a számok közötti viszonyokat megragadó komponensek közül az első a három szövegtípusban a variáció 62/65/62%-át, a második a 28/27/30%-át ragadja meg. Ez azt jelenti, hogy ha az öt mérőszám által kifejezett információt öt helyett két dimenzióban ábrázoljuk, akkor ez a két dimenzió kifejezi az információ >90%-át, ami erősen arra utal, hogy

az öt szám egy jelentős részben ugyanazt méri. A három szövegtípus között nincsenek nagyon jelentős különbségek. Az első főkomponens (vízszintes tengely) elsősorban a szövegek szintagmatikus komplexitását tükrözi. Ebben a dimenzióban mozog az a két mérőszám, amelyek leginkább érzékenyek a szekvenciák szerkezetére, a legrövidebb leírás aránya (MDL/DL) és a feltételes bigram entrópia. A második főkomponens (függőleges tengely) elsősorban a paradigmatikus komplexitást tükrözi. Az erre elsősorban érzékeny mérőszám, az unigram entrópia leginkább ebben a dimenzióban mozog. A mérőszámokat nem lehet ilyen egyszerűen csoportosítani, amit szépen mutat az, hogy mindkét dimenzióban megjelenik a típus/token arány és a szavak száma. Ezekről mondtuk azt a 2.2 pontban, hogy közvetve minden mérőszámot befolyásolnak.

3.2. Betűhű szövegek



3. ábra: A betűhű szövegekre kapott mérőszámok és a betűhű–normalizált fejezetpárokra számolt Jaccard-közelség. Minden eloszlás a fejezetekre kapott számokból áll. A függőleges vonalak az eloszlások első, második és harmadik kvartilisét jelölik.

Az információsűrűség mérőszámait a betűhű szövegekben a 3. ábrán látjuk. Mivel minden fordítás minden fejezetére külön értékeket számolunk, az ábrán látható eloszlások minden esetben 28 + 16 + 24 + 21 = 89 megfigyelést tartalmaznak. Az ábra az öt, általunk használt mérőszám mellett a betűhű szöveg és a normalizált szöveg közötti, fejezetpárokra számolt Jaccard-közelséget is megmutatja. A Jaccard-közelség kifejezi az átfedést két karaktersor között, és kevésbé érzékeny a karaktersorok hosszára, mint pl. a Levenshtein-távolság. A meghatározása a betűhű szöveg és a normalizált szöveg metszete elosztva ezek uniójával. Minél kisebb az eltérés a két szöveg között, ez az érték annál nagyobb lesz (0 = nincs átfedés, 1 = teljes átfedés).

Egy magyar szövegnek a két teljesen konzisztens, de eltérő ortográfiával írt változata között is lehet nagy a távolság. A Jaccard-közelség tehát kifejezheti azt, hogy a betűhű szövegnek mekkora a konzisztenciája, és azt is, hogy mennyire tér el a normalizált változattól. A két változat között a legnagyobb távolságot mégis a korai, kézzel írt Müncheni kódex esetén találjuk, ami arra utal, hogy a konzisztenciának itt fontos szerepe van.

Általában azt látjuk, hogy minél közelebb van a fordítás betűhű szövege a normalizált változathoz, annál kisebb a szöveg információsűrűsége. Ezt magyarázhatjuk úgy, hogy a betűhű szövegek ortográfiája a fordítások időrendi sorrendjében egyre konzisztensebbé válik, és az információsűrűség csökkenése elsősorban ennek tudható be.

A szövegek hossza hasonló (3a). Az eredeti és a normalizált szöveg közötti átfedés a Müncheni kódexben található legelső fordítás esetén a legkisebb, mindössze 30% körüli, az ezt követő három, 16. század eleji fordításnál is csak 50-60% körül van, és csak Heltai, Károli és Káldi fordításánál éri el a 70%-ot (3b). A szövegek tömöríthetősége ezzel szépen együtt áll – minél messzebb van a normalizált szövegtől az eredeti szöveg, annál nehezebb tömöríteni (3c). A betűhű–normalizált távolsággal a típus/token arány is együtt áll (3d). Az unigram entrópia eloszlása gyengébben ugyan, de követi ezt a mintázatot (3e), ami arra a technikai összefüggésre utal, hogy ha a szövegben nagyon sokféle szó előfordul, akkor a szöveg információtartalma magas lesz.

Érdekes a bigram entrópia eloszlások viselkedése (3d), ami a tömöríthetőség (3c) és a típus/token arány (3d) esetén látott mintának nagyjából az inverze. Minél inkonzisztensebb a szöveg, annál több egyedi szóalakot (típust) tartalmaz, és annál nehezebb tömöríteni. A bigram entrópia azt mondja meg, hogy mennyire tudunk egy szóból a következő szóra következtetni. A típus/token arány és a bigram entrópia között negatív összefüggést várunk: az alacsony típus/token arány, a korlátozott szókincs azt jelenti, hogy ugyanaz a néhány szó nagyon hasonló helyi kontextusokban ismétlődik. Ez felduzzasztja az ismétlődő bigramok számát, ezért általában könnyű lesz az első szóból a másodikra következtetni, és a bigram entrópia nagy lesz. A magas típus/token arány, a gazdag szókincs az alacsony

gyakoriságú szavak számának növekedését jelenti, ezért a bigrammok eloszlása ellaposodik, és a bigram feltételes entrópia csökken. Ezt látjuk itt a betűhű szövegekben is.

A mutatók összefüggése tehát alapvetően azt mutatja, amit várnánk ezektől a szövegektől. A kevésbé konzisztens/többféle ortográfiát használó, és emiatt több normalizálásra szoruló szövegekben nagyobb az egyedi szóalakok aránya, és ez nagyjából megmagyarázza a többi mutató viselkedését is. Ha azt szeretnénk látni, hogy a mutatók mit árulnak el elsősorban a fordítások stilisztikai vagy nyelvi tulajdonságairól, a normalizált szövegekhez kell fordulnunk.

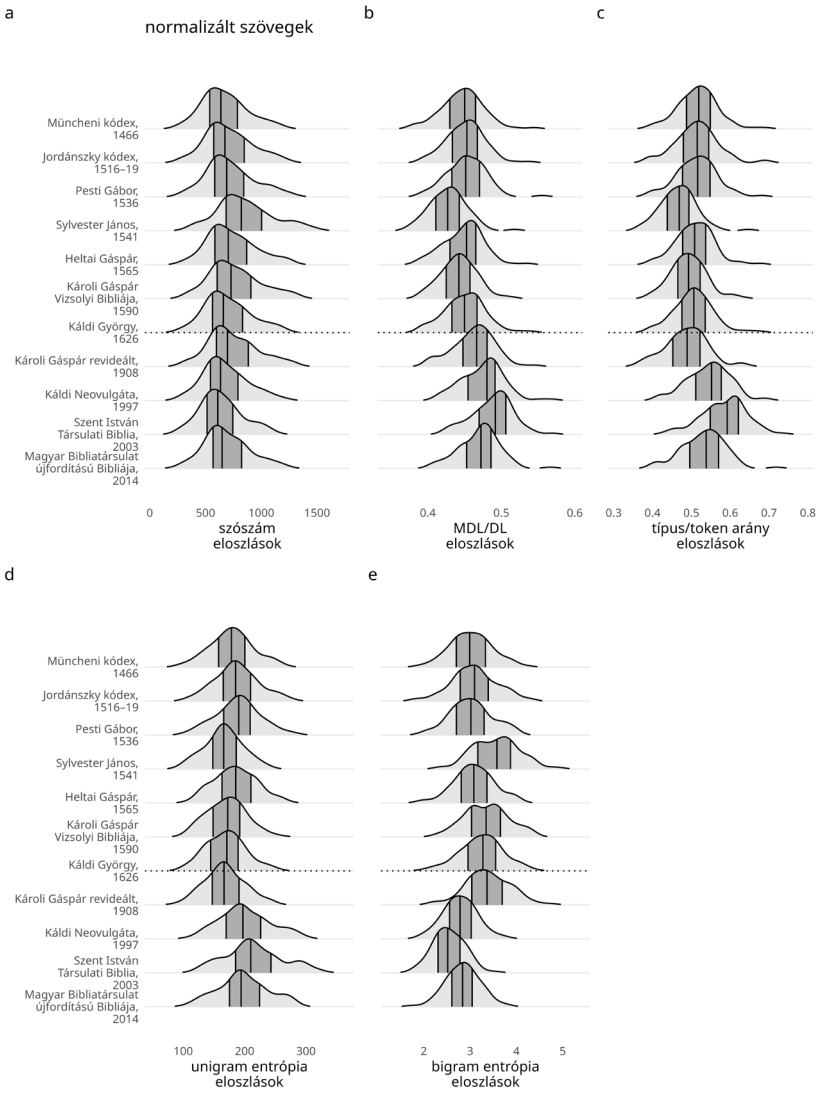
3.3. Normalizált szövegek

Az információsűrűség mérőszámait a normalizált szövegekben és a modern fordításokban a 4. ábrán látjuk. A fordítások között a modern fordításokat ide véve sincs jelentős különbség a szövegek hosszában (4a). A normalizált/modern szövegek közötti különbségek kisebbek, mint a betűhű szövegek közötti különbségek, és ez szintén az ortográfia relevanciáját mutatja. Az eloszlásokban ezen túl három fontos mintázat tűnik fel.

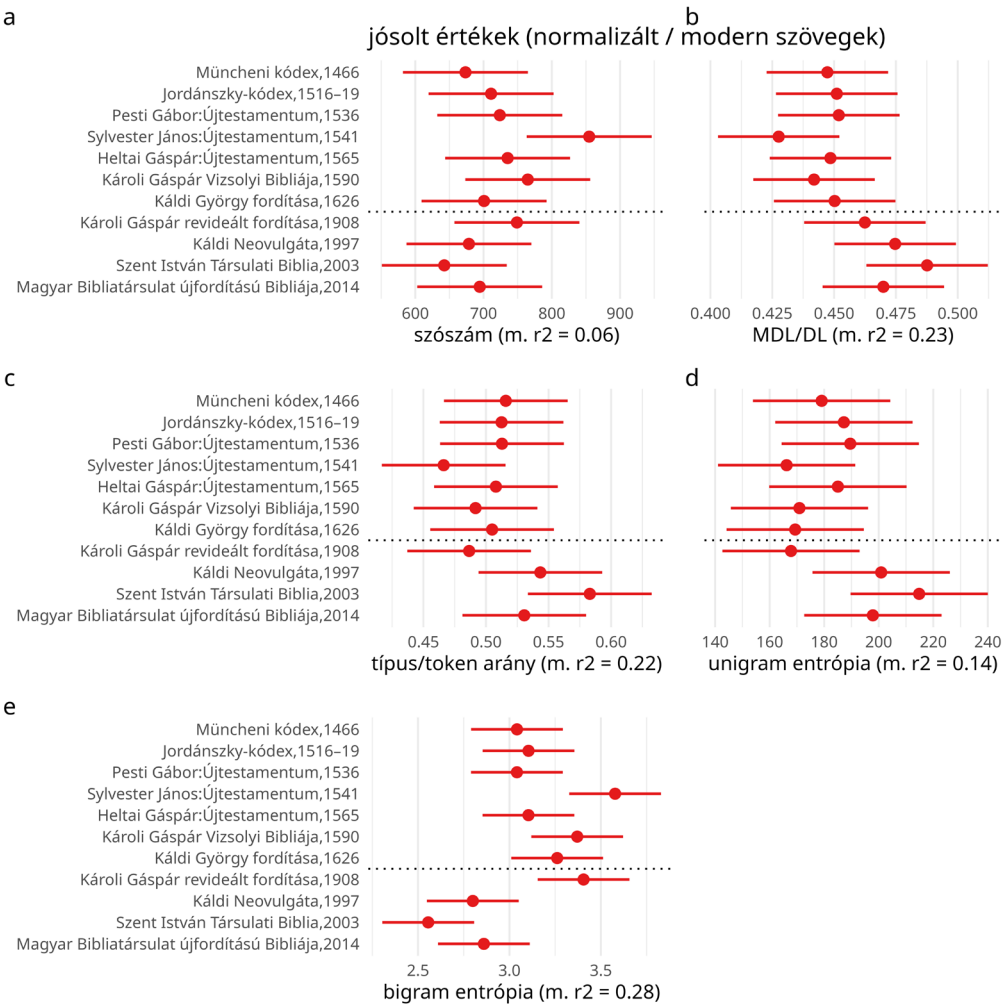
A 15. századi, 16. század eleji fordítások közül a Sylvester-féle változat kiugró információs mintázatot mutat. Jobban tömöríthető, kevesebb típust használ, ez pedig kisebb információsűrűséghez vezet, ami az információtartalmat általánosan kifejező unigram entrópiában is megjelenik. Ez a 2.3 és a 3.2 fejezetben látott módon a bigram entrópia relatív növekedését hozza magával.

Sylvester fordításától eltekintve a 15. századi, 16. század eleji fordításokhoz képest a Károli és Káldi fordítások jobban tömöríthetők, a modern, főleg a 20. századi, 21. századi fordítások viszont jóval kevésbé (4b). A szókincs méretét, az egyedi szóalakok számát megragadó típus/token arány és unigram entrópia hasonló mintázatot mutat (4c-d). Ezzel szemben azt látjuk, hogy a szövegek formulaikusságát, az ismétlődő szerkezetek gyakoriságát megragadó bigram entrópia (4e) a Károli és Káldi fordítás esetén valamivel nagyobb, a modern, és főleg a 20–21. századi fordításoknál viszont kisebb mint a korai fordításoknál.

Összefoglalva, a normalizált fordítások információs mérőszámai a 2.3. fejezet alapján várt módon függenek össze. A mérőszámok által kirajzolt információs profilok alapján egy csoportot alkot a huszita biblia, a Jordánszky-kódex, valamint Pesti és Heltai fordítása, egy második csoportot a Káldi- és a Károli-fordítások, és egy harmadik csoportot a modern változatok. A Sylvester-fordítás profilja kiugró.



4. ábra: A normalizált szövegekre kapott mérőszámok. Minden eloszlás a fejezetekre kapott számokból áll. A függőleges vonalak az eloszlások első, második és harmadik kvartilisét jelölik. A vízszintes szaggatott vonal alatti modern fordítások szövege nem normalizált.



5. ábra: A normalizált szövegekre jósolt mérőszámok, hierarchikus lineáris modellek becslései. A pontok a középértéket, a vízszintes vonalak a 99%-os konfidenciaintervallumot jelölik (a valós érték az esetek 99%-ában ebbe a tartományba esik). A vízszintes szaggatott vonal alatti modern fordítások szövege modern, nem normalizált. A 99%-os konfidenciaintervallumot azért választottam a viselkedéstudományokban megszokott 95% helyett, hogy kompenzáljuk a többszörös összehasonlítások hatását (*ceteris paribus* ugyanazokra az adatokra kapott öt 99%-os eredmény statisztikailag annyira biztos, mint egy 95%-os).

A fordítások összevetését megnehezíti az adatokban lévő hierarchia. A különböző fordításokban a különböző mérőszámok ugyanazokhoz a könyvekhez és fejezetekhez vannak hozzárendelve. Ezeknek a fejezeteknek vannak inherens információs tulajdonságaik: milyen(ek) az eredeti szöveg(ek), hogyan kapcsolódnak az előttük és utánuk álló fejezetekhez, és így tovább. Ezeket az aspektusokat részben meg tudjuk ragadni, ha egy hierarchikus modellt illesztünk a normalizált szövegekből származó adatokra, és a nyers adatok helyett a modell jóslatait vizsgáljuk meg. Ezek a jóslatok láthatóak az 5. ábrán. Mivel a mérőszámok összefüggenek egymással (ld. 1. ábra), minden mérőszámra külön modellt illesztettem, amely prediktorváltozóként a fordítások időben rendezett sorrendjét használta. Minden modell az egyes értékeket fejezetek alá, ezeket könyvek alá rendezte.

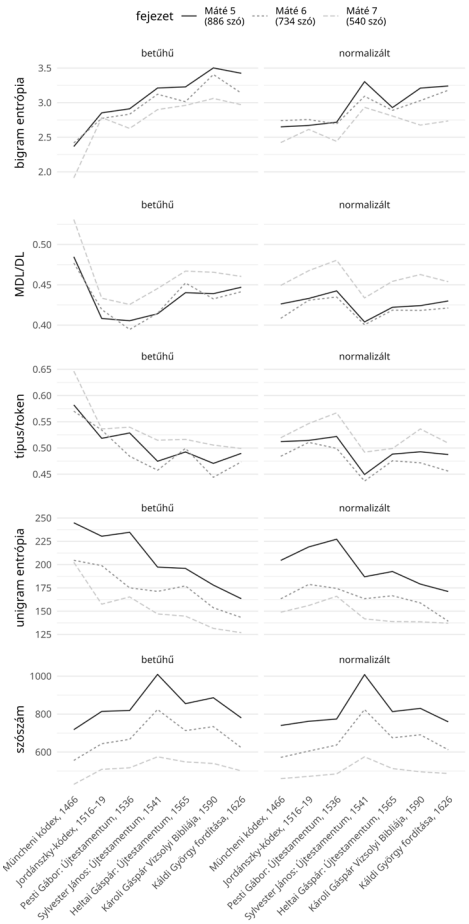
Minden modellre számolhatunk egy marginális r^2 értéket, amely megragadja, hogy a fordítások közötti különbségek a mérőszámnak mekkora varianciáját magyarázzák meg, azaz mennyire sokat tesznek hozzá a modellhez. Mivel öt különböző modellt illesztünk ugyanazokra az adatokra, az értékek inflálódni fognak, ezért nem érdemes őket készpénznek venni, de arra alkalmasak, hogy fényt vessenek a változók relatív magyarázó erejére.

A két legjobb változó a bigram entrópia, azaz az, hogy átlagosan mennyire lehet megjósolni, mi lesz egy szópár második tagja (5e), és a szöveghosszal súlyozott tömöríthetőség (5b). A típus/token arány és az unigram entrópia (5b-c) jóval gyengébb magyarázó erővel bírnak, a normalizált szövegek szavakban számolt hossza pedig gyakorlatilag nem függ össze azzal, hogy melyik fordítás mikor készült (5a).

Ezeket az eredményeket összekapcsolhatjuk az információtartalomnak és a nyelvi struktúrának a 2. pontban tárgyalt összefüggéseivel. A modern fordítások nehezebben tömöríthetőek, aminek fő oka az lehet, hogy több szóalakot használnak, nagyobb szókinccsel operálnak, mint az ómagyar és középmagyar fordítások. Paradigmatikus komplexitásuk tehát magasabb. Ugyanakkor a bigram entrópiájuk jóval kisebb. Ezt részben magyarázhatja a szókincs mérete és a szekvenciák megjósolhatósága közötti negatív összefüggés (ld 1. ábra, 3.2 pont). A másik magyarázat az lehet, hogy a modern szövegek valamilyen értelemben formulaikusabbak, a szintagmatikus komplexitásuk alacsonyabb. Erre utal az, hogy a bigram entrópia (5e), amely elsősorban a szintagmatikus komplexitásra érzékeny, jobban megkülönbözteti egymástól a fordításokat, mint a típus/token arány vagy az unigram entrópia (5c-d), amelyek elsősorban a paradigmatikus komplexitásra érzékenyek. Fontos persze megismételni, hogy ezek a számok nem közvetlenül mérik a nyelvi komplexitást, és egymással is összefüggenek.

Bármilyen tendenciát vélünk felfedezni az adatokban, azt szem előtt kell tartanunk, hogy (a korai fordításokat nézve) hét adatpontról van szó, ezért a jelen megközelítés alapján nem érdemes és nem is lehet komoly következtetéseket

levonni nemhogy a magyar nyelv szerkezetéről, de még a bibliafordítások változásáról sem. Ettől függetlenül érdekes lehet megnézni egy konkrét fejezet változásait. A 6. ábra a Máté-evangélium 5–7. fejezetét (a Hegyi Beszéd) veszi szemügyre a hét korai fordításban, betűhű és normalizált változatban, az információsűrűség különböző mérőszámait mutatva.



6. ábra: A bizonytalanság és az összetettség változása a Máté-evangélium 5–7. fejezetének hét ó-, ill. középmagyar fordításában. Az abszcissza a fordítások időrendi sorrendje, az ordináta a mérőszám mértéke. A tíz panel az öt mérőszám (fentről lefelé) betűhű (balra) és normalizált (jobbra) számolt értékeit mutatja, a vonalak a fejezeteket. Jól látható, hogy a példában a mérőszámok általában érzékenyek a fejezetek hosszára, amelyet itt a vizsolyi bibliából vett szószámmal illusztráltam.

4. Az eredmények értékelése

A fentiekben az Újszövetség első négy könyvének hét ómagyar és középmagyar fordítását vetettem össze egymással és négy modern fordítással, öt, az információszerkezetet, az információsűrűséget mérő paraméter alapján. Az elemzésnek vannak egyértelmű korlátai. A középkori magyar kódexanyag egy jelentős része elveszett (Korompay 2006; Simon – Sass 2012), és több, a forrásokban említett, de nem hozzáférhető magyar bibliafordításról tudunk (Madas 1998). A feldolgozott adathalmaz tehát ilyen szempontból sem reprezentatív. Az automatikus szövegnormalizáláson javít a kézi ellenőrzés, de így sem teljesen pontos, és a pontossága a régebbi, inkonzisztensebb ortográfiájú szövegeknél romlik (Vadász – Simon 2018), tehát a normalizált szöveg tartalmazza a betűhű eredeti valamekkora kontaminációját. Az információsűrűség mérőszámainak kiválasztása a szakirodalom által megalapozott, de végső soron önkényes. Az adatokra illesztett lineáris modellek hajlamosak lehetnek a megfigyelt hatások mértékét túlbecsülni, különösen akkor, ha ezek között nagyobb autokorreláció van, ami egy könyv fejezeteinél valószínűsíthető (Gelman – Hill 2006). Az eredmények ugyanakkor informatívak és konzisztensek. Lehetőséget adnak arra, hogy visszatérjünk a Bevezetésben megfogalmazott kutatási kérdésekhez.

4.1. Vannak-e szisztematikus információszerkezeti különbségek a fordítások között?

A különféle fordítások között jól látható különbségeket találunk. Az információszerkezet különböző mérőszámai mind azt mutatják, hogy más profiljuk van az ómagyar szövegeknek, a középmagyar szövegeknek, illetve a modern fordításoknak. A látható különbségeket, különösen a három huszadik század végi, huszonegyedik századi fordítás esetén, hierarchikus regressziós elemzésünk is megerősíti.

4.2. Magyarázzák-e ezek a különbségek a szövegek egymáshoz és a saját társas, kulturális környezetükhöz fűződő viszonyát?

Ha elfogadjuk azt, hogy az információsűrűség változásai valamilyen általános tendenciára utalnak, erre adhatunk ortográfiai, nyelvi és stilisztikai magyarázatot.

Az ortográfiai magyarázat lényege az, hogy a helyesírásnak a magyar nyelvű könyvnyomtatás kezdetével egy időben kezdődő konszolidációja a betűhű szövegekben a lokális entrópia folyamatos csökkenéséhez vezetett (mivel elkezdték az emberek a szavakat ugyanúgy leírni), és ez a hatás áttételesen a normalizált szövegekben is megjelenik. A bigram entrópia növekedése ebben az olvasatban szinte kizárólag az unigram entrópia csökkenésének az eredménye.

A nyelvi magyarázat az, hogy az egymás után következő ómagyar és középmagyar fordításokban a szövegek paradigmatikus komplexitása csökken, ami ahhoz vezet, hogy ugyanannak a tőnek kevesebb alakja jelenik meg, ez pedig csökkenti az információs bizonytalanságot. Ezzel párhuzamosan a szintagmatikus komplexitás valamilyen értelemben növekszik, tehát a nyelvtani szerkezetek több szóval ragadják meg ugyanazokat a jelentéseket, ami az információs összetettség növekedésével jár. A modern kiadások nem illeszkednek jól ebbe a mintába.

A stilisztikai magyarázat ezt a két szempontot fogja össze, amikor azt mondja, hogy az ortográfiai és nyelvi mintázatok a fordítók tudatos döntéseinek eredményeként jelentek meg a szövegekben. A jelen szöveg nem tud teret adni annak, hogy ómagyar és középmagyar bibliafordításaink keletkezésének körülményeit, a fordítók szándékait és döntéseit kimerítően tárgyalja (teljesség igénye nélkül vö. Erdős 1906; 1937; Madas 1998; Máthé 2004; Hegedüs 2013; Korompay 2016; P. Kocsis 2022). Fontos azonban megjegyeznünk, hogy a rendelkezésre álló tudás alapján a fordítások közötti különbségeket aligha redukálhatjuk a fordítók közötti inkonzisztenciára. Ehelyett két szempontot érdemes figyelembe venni. Egyrészt főleg az ómagyar kori szövegpusztulás igen jelentős, miközben éppen az ómagyar korból a középmagyar korba érve változik meg a magyar nyelv közvetítői szerepe, feladatai, ideértve a szakrális használat térnyerését is (Korompay 2006; 2023). Másrészt a középmagyar kortól kezdődően az igeidők egyszerűsödése és ingadozása figyelhető meg, amivel a magyar morfológia jelentős változásokon megy keresztül (Mohay 2018). A magyar nyelvhasználat kontextusaiban lezajló változások, a formális, szakrális közegekben megjelenő kiterjedt magyar nyelvű használat (ezeknek az első fennmaradt bibliafordítások sokkal inkább cselekvő részesei, mint példái vagy tünetei), valamint a morfológiai rendszer változása is magyarázatot adhatnak a szövegek információs profiljában látható változásokra. Ez további kutatásoknak adhat teret.

Az a felvetés, hogy a fordítások információszerkezetében megfigyelhető változások a nyelv és a nyelvhasználat változásait tükrözik, kapcsolódik a nyelvi komplexitás és a társas közeg viszonyát tárgyaló tágabb szakirodalomhoz (Lupyan – Dale 2010; Nettle 2012; Shcherbakova és mtsai 2023). Eredményeim arra utalnak, hogy egy szöveg keletkezésének a szociokulturális kontextusa befolyásolhatja a szöveg szerkezetét. Az előbbitől az utóbbiig vezető oksági láncot valószínűleg nem lehet leegyszerűsíteni egy olyan állításra, hogy az eltérő méretű beszélőközösségek tendencia szerint eltérő információstruktúrájú szövegeket állítanak elő. Az evangéliumok magyar fordításainak példája inkább azt mutatja, hogy a szövegek születését jelentős mértékben befolyásolják azoknak szociális és kulturális körülményei, és ezek itt elsősorban a szerzők (fordítók) rendelkezésére álló szimbolikus eszköztár, a támaszként, referenciapontként szolgáló kulturális hagyomány kereteit jelentik.

Összességében az eredmények hasznos meglátásokkal szolgálnak a magyar nyelvi standardizáció leírásához, és alátámasztják az információs mutatók használatának relevanciáját a nyelvtörténeti és komparatív kutatásokban.

Irodalom

Ackerman, Farrell – Malouf, Robert (2013), Morphological organization: the low conditional entropy conjecture. Language 89/3: 429–464.

Atkinson, Mark – Kirby, Simon – Smith, Kenny (2015), Speaker input variability does not explain why larger populations have simpler languages. PLOS ONE 10/6: e0129463.

Bates, Douglas – Maechler, Martin – Bolker, Ben (2011), lme4: Linear mixed-effects models using S4 classes.

Beckner, Clay – Blythe, Richard – Bybee, Joan – Christiansen, Morten H. – Croft, William – Ellis, Nick C. – Holland, John – Ke, Jinyun – Larsen-Freeman, Diane – Schoenemann, Tom (2009), Language is a complex adaptive system: position paper. Language Learning 59: 1–26.

Bentz, Christian – Ruzsics, Tatyana – Koplenig, Alexander – Samardzic, Tanja (2016), A comparison between morphological complexity measures: typological data vs. language corpora. In: Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC). 142–153.

Bloomfield, Leonard (1933), Language. Henry Holt, New York.

Dębowski, Łukasz – Bentz, Christian (2020), Information theory and language. Entropy 22/4: 435.

Dye, Melody – Milin, Petar – Futrell, Richard – Ramscar, Michael (2018), Alternative solutions to a language design problem: the role of adjectives and gender marking in efficient communication. Topics in Cognitive Science 10/1: 209–224.

Erdős József (1906), Az Újszövetségi kánon fordításairól. Franklin-Társulat nyomdája, Budapest.

Erdős Károly (1937), Az Újszövetség magyar fordításai a reformáció óta. Újszövetségi Füzetek 1/1: 7.

Fügedi Erik (1992), A középkori Magyarország történeti demográfiája. In: Dányi Dezső (sorozatszerk.), A Központi Statisztikai Hivatal Népességtudományi Kutatóintézetének történeti demográfiai füzetei 10. Központi Statisztikai Hivatal, Budapest.

Gailly, Jean-loup – Adler, Mark (1992–2024), gzip 1.13. Free Software Foundation. https://www.gnu.org/software/gzip/

Gelman, Andrew – Hill, Jennifer (2006), Data analysis using regression and multilevel/hierarchical models. Cambridge University Press, Cambridge.

Gibson, Edward – Futrell, Richard – Piantadosi, Steven P. – Dautriche, Isabelle – Mahowald, Kyle – Bergen, Leon – Levy, Roger (2019), How efficiency shapes human language. Trends in Cognitive Sciences 23/5: 389–407.

Hausser, Jean – Strimmer, Korbinian (2021), entropy: estimation of entropy, mutual information and related quantities. R package version 1.3.1. https://CRAN.R-project.org/package=entropy

Hegedüs Béla (2013), János evangéliumának eleje hat bibliafordításunkban. In: Benő Attila – Fazekas Emese – Kádár Edit (szerk.), „…hogy legyen a víznek lefolyása…”. Köszöntő kötet Szilágyi N. Sándor tiszteletére. Erdélyi Múzeum-Egyesület, Kolozsvár. 181–186.

Holland, Tom (2019), Dominion: the making of the Western mind. Hachette UK, London.

Kobayashi, Hayato (2014), Perplexity on reduced corpora. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 797–806.

P. Kocsis Réka (2022), A névelők alaki kérdései az ún. huszita biblia kódexeiben. In: Forgács Tamás – Németh Miklós – Sinkovics Balázs (szerk.), A nyelvtörténeti kutatások újabb eredményei IX. SZTE Magyar Nyelvészeti Tanszék, Szeged. 173–188.

Korompay Klára (2006), Árpád-kori szövegek: mit képvisel az, amit ránk maradt? In: Mártonfi Attila – Papp Kornélia – Slíz Mariann (szerk.), 101 írás Pusztai Ferenc tiszteletére. Argumentum Kiadó, Budapest. 116–122.

Korompay Klára (2015), Gondolatok egy régi-új vitához: az ún. Huszita Biblia eredetének kérdésköre. In: Bárth M. János – Bodó Csanád – Kocsis Zsuzsanna (szerk.), A nyelv dimenziói. Tanulmányok Juhász Dezső tiszteletére. ELTE BTK Magyar Nyelvtörténeti, Szociolingvisztikai, Dialektológiai Tanszék, Budapest. 79–88.

Korompay Klára (2023), Szóbeliség és írásbeliség viszonya a korai ómagyar korban. Magyar Nyelv 119/2: 153–170. https://doi.org/10.18349/MagyarNyelv.2023.2.153

Lupyan, Gary – Dale, Rick (2010), Language structure is partly determined by social structure. PLOS ONE 5/1: e8559.

Madas Edit (1998), Középkori bibliafordításainkról. Iskolakultúra 8/1: 48–54.

Máthé Dénes (2004), Heltai Gáspár nyelvhasználati sajátosságairól. Keresztény Magvető 110/4: 428.

Mohay, Zsuzsanna (2018), Múltidő-használat a középmagyar korban. Doktori disszertáció. Eötvös Loránd Tudományegyetem, Budapest.

Li, Ming – Vitányi, Paul M. B. (2019), An Introduction to Kolmogorov Complexity and Its Applications. 4th edition. Springer, Cham. https://doi.org/10.1007/978-3-030-11298-1

Moscoso del Prado, Fermin (2013), The missing baselines in arguments for the optimal efficiency of languages. In: Proceedings of the Annual Meeting of the Cognitive Science Society 35. 1032–1037.

Nádor Orsolya (2000), A magyar nyelv státusának változásai a honfoglalástól a XIX. század közepéig. Hungarológiai Évkönyv 1: 54–71.

Nettle, Daniel (2012), Social scale and structural complexity in human languages. Philosophical Transactions of the Royal Society B 367/(1597): 1829–1836.

Piantadosi, Steven T. – Tily, Harry – Gibson, Edward (2011), Word lengths are optimized for efficient communication. Proceedings of the National Academy of Sciences 108/9: 3526–3529.

Parry, Milman (1933), Whole formulaic verses in Greek and Southslavic heroic song. Transactions and Proceedings of the American Philological Association 64: 179–197.

M. Pintér Tibor (2024), Magyar nyelvű bibliafordítások statisztikai elemzése. Alkalmazott Nyelvtudomány, Különszám 2024/1: 22–36. http://dx.doi.org/10.18460/ANY.K.2024.1.002

R Core Team (2025), R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna.
Raviv, Limor – Meyer, Antje – Lev-Ari, Shiri (2019), Larger communities create more systematic languages. Proceedings of the Royal Society B 286/1907: 20191262.
Resnik, Philip – Olsen, Mari Broman – Diab, Mona (1999), The Bible as a parallel corpus: annotating the „Book of 2000 Tongues". Computers and the Humanities 33: 129–153.
Ruden, Sarah (2023), The Gospels: a new translation. Modern Library, New York.
Sapir, Edward (1912), Language and environment. American Anthropologist 14/2: 226–242.
de Saussure, Ferdinand (1916), Cours de linguistique générale. Payot, Lausanne – Paris.
Shannon, Claude Elwood (1948), A mathematical theory of communication. Bell System Technical Journal 27/3: 379–423.
Shcherbakova, Olena – Michaelis, Susanne Maria – Haynie, Hannah J. – Passmore, Sam – Gast, Volker – Gray, Russell D. – Greenhill, Simon J. – Blasi, Damián E. – Skirgård, Hedvig (2023), Societies of strangers do not speak less complex languages. Science Advances 9/33: eadf7704.
Simon Eszter – Kalivoda Ágnes (2020), A párhuzamos bibliakorpusz és Bibliaolvasó fejlesztése. Általános Nyelvészeti Tanulmányok 32: 429–438.
Simon Eszter – Sass Bálint (2012), Nyelvtechnológia és kulturális örökség, avagy korpuszépítés ómagyar kódexekből. Általános Nyelvészeti Tanulmányok 24: 243–264.