Egy angolnyelv-vizsga validálásáról és az automatizált nyelvi értékelésről

Reményi Andrea Ágnes
Pázmány Péter Katolikus Egyetem Angol-Amerikai Intézet
remenyi.andrea@btk.ppke.hu

Megjelent: Fogarasi Katalin, Ittzés Dániel, Putz Mónika, Vágási Tünde (szerk.) (2025). *Tudásmegosztás, információkezelés, alkalmazhatóság. 3.: Nyelvpedagógia.* Akadémiai Kiadó. – E tanulmány 2023-ban készült el.

Hogyan lehet a *Közös Európai Referenciakeret (KER)* szerinti nyelvtudásszinteket az angol nyelvtan és szókincs számszerűsíthető jellemzői segítségével megfogalmazni? Melyek egy bizonyos *KER*-szint jellemzői a nyelvtani és szókincskomplexitás és a nyelvhelyesség eloszlásai szempontjából? Egy magyarországi egyetem egyik angol szakos nyelvvizsgájának validálása során az írott nyelvi szövegekben sokváltozós (gépi és kézi) elemzési eszközökkel kerestük ezeket az automatizált nyelvi értékelésre is alkalmas eszközöket. A statisztikai elemzésen túl választ kerestünk arra a kérdésre is, hogy a ChatGPT vajon alkalmas-e ugyanezen szempontok *KER*-szintek szerinti értékelésére.

Kulcsszavak: nyelvi értékelés, nyelvvizsga-validálás, nyelvtani és szókincskomplexitás és -helyesség, sokváltozós statisztikai elemzés, ChatGPT

On the validation of an English language examination and automated language assessment

How can we conceptualise language proficiency levels according to the *Common European Framework of Reference for Languages (CEFR)* in quantifiable features of English grammar and vocabulary? What are the characteristics of a certain *CEFR* level, in terms of its patterns of syntactic and lexical accuracy and complexity? While validating a language examination for English majors at a Hungarian university, we are detecting the systematic patterns of those in written texts with a multivariate, automated and manual, analysis capable also of automated language assessment. Beyond the statistical analysis, I have tested ChatGPT's ability to predict *CEFR* levels.

Keywords: language assessment, language examination validation, syntactic/lexical accuracy and complexity, multivariate statistical analysis, ChatGPT

1. Bevezetés

A nyelvtudás értékelésének egyik visszatérő kérdése, hogy hogyan lehet a *Közös Európai Referenciakeret* (*KER*) szerinti tudásszinteket az angol nyelvtan és szókincs számszerűsíthető jellemzői segítségével megfogalmazni. Más szóval, melyek egy bizonyos *KER*-szint jellemzői a nyelvtani és szókincskomplexitás és -helyesség szempontjából? Tanulmányomban két egymással összefüggő témával foglalkozom: egyrészt hogyan lehet a fenti szempontok segítségével bizonyítani, hogy egy bizonyos nyelvvizsga azt méri, amit mérni szándékozik (ezt nevezzük vizsgavalidálásnak), és másrészt hogyan lehet megoldani a nyelvvizsgákon készített angol mint idegennyelvi szövegek automatizált értékelését.

Egy magyarországi egyetemen angolszakosok számára az első év végén esedékes, B2+ szintű nyelvvizsga validálásán dolgozunk, ennek részeként egy ott gyűjtött írásbeli korpusz alapján a nyelvtani és szókincskomplexitás jellegzetes mintázatait mérjük fel, illetve ezeknek a B2+ elvárásoknak való megfeleléseit keressük. Egyrészt sokváltozós kutatási módszert alkalmazunk: a nyelvtani és szókincskomplexitás mintázataiban számos változó hatását vizsgáljuk, beleértve manuálisan és automatikusan kinyert változókat is. Ezek statisztikai keresztelemzésével emeljük ki a legfontosabb változókat, arra is gondolva, hogy utóbbiak elősegíthetik majd a tanulói szövegek automatizált értékelését. A mesterséges intelligencia fejlődésének közelmúltbeli ugrásszerű fejlődését követve egy másik úton is elindultam: kipróbáltam, hogy vajon a ChatGPT képes-e angol nyelvtanulói szövegek *KER*-szintezésére.

A projekt fő kutatási kérdése az, hogy ez a vizsga valid és megbízható módon méri-e az angol nyelvtudást B2+ szinten. Ezen belül, a jelen tanulmány kutatási kérdései a következők:

1. Hogyan érdemes kiválasztani azokat a nyelvtani és szókincskomplexitási jellemzőket, amelyek az írott szövegek tekintetében hozzájárulnak a szövegek nyelvtudásszintjének bejóslásához és így a fenti nyelvvizsga validálásához?
2. Adott szövegek nyelvtudásszintjének bejóslására a vektoralapú nagy nyelvmodellek, illetve itt konkrétan a ChatGPT alkalmas alternatívája-e a statisztikai alapú elemzési módszereknek?

2. Háttér

2.1. A nyelvtani és szókincskomplexitás fogalma

A nyelvtani és szókincskomplexitás az idegennyelv-tudás egyik meghatározó jelensége (pontosabban: bármely másod- vagy idegen nyelv (L2) tudásáról van szó). Bulté és Housen (2012, 2014) modellje az L2-nyelvtudást a komplexitás, nyelvhelyesség és folyékonyság hármasaként írja le (*complexity, accuracy, fluency*). Ezek közül a komplexitást a tanuló teljesítményének „nagyságaként, kidolgozottságaként, gazdagságaként és változatosságaként" határozzák meg (Housen-Kuiken, 2009: 464). A nyelvtani komplexitás ennek a konstrukciónak a nyelvi komplexitáshoz tartozó alkomponense, részei a komplexitás mondat-, tagmondat- és frazális szintjei. A szókincskomplexitás a nyelvi komplexitás másik alkomponense, részei a kollokációs és lexémikus komplexitás. (Helyszűke miatt jelen tanulmányban csak a nyelvtani komplexitást tárgyalom tovább.)

Crossley (2020: 423) szerint „a magasabb minőségű szövegek általában komplexebb nyelvtani struktúrákat használnak." Ortega (2003) szerint a nyelvtani komplexitás a nyelvi érettség jele, és a nyelvtani formák repertoárjaként és kidolgozottságuk mértékeként definiálható. Lu (2010, 2017) sokdimenziós struktúraként értelmezi, melyet a nyelvi mérés-értékelés a minősége, az L2-íráskutatás pedig a variabilitása felől közelít meg.

Ami az L2-angol írott szövegek kutatását illeti, a nyelvtani komplexitás vizsgálata négy kutatási irányba sorolható, amelyek úgy kvantifikálnak, hogy a különféle nyelvtani jelenségeket változókként emelik ki. Az egyik kutatási irány az anyanyelvi és nem-anyanyelvi szövegeket hasonlítja össze (az utóbbiakat gyakran nyelvtudás-szint szerint csoportosítva). Ai és Lu (2013) például a szövegegységek hossza (mondatok/tagmondatok/T-egységek átlagos hossza), az alárendelés és a mellérendelés mértéke (alárendelt illetve mellérendelt tagmondatok aránya) és a frazális kidolgozottság foka (komplex főnévi csoport tagmondatonként/T-egységenként) szerint számszerűsítik a nyelvtani komplexitást. Mancilla et al. (2015) hasonló kategóriákat alkalmaznak. Egy másik kutatási irány adott nyelvtanulók nyelvtudás-fejlődését vizsgálja az általuk korábban illetve később írott szövegek összehasonlításával, gyakran a fenti irányhoz hasonló változók segítségével. Polat et al. (2019) és Lei et al. (2023) azt találták, hogy ezek a változók jó indikátorai a nyelvtani komplexitás fejlődésének.

Egy harmadik kutatási irány a humán értékelők által beszintezett szövegek nyelvtani komplexitásmintázatait vizsgálja. Például Taguchi et al. (2013) frazális szintű jellemzőket (determinánsok, attributív melléknevek, főnevet követő elöljárós szerkezetek, stb.) és tagmondatszintű jellemzőket vizsgáltak (alárendelős kötőszavak, *that*-vonatkozói mellékmondatok, stb.). Crossley és McNamara (2014) szintén frazális és tagmondatszintű jellemzőket vizsgált (módosítók főnévi csoportonként, elöljárós frázisok, infinitívek, stb.). A fenti változók mindegyike jellemzőnek bizonyult a nyelvtudásszint indikátoraként.

A negyedik kutatási irány a nyelvtanulói szövegek nyelvtani komplexitásában a regiszter-, műfaj- és feladatspecifikus jellemzőket keresi. Például Biber et al. (2016) a regisztervváltozatok régóta fennálló kutatási hagyományára alapozva 23 nyelvtani jellemző eloszlását és együtt-előfordulását elemezték írott, beszélt és integrált feladatok alapján létrejött szövegekben. Azt találták, hogy például a hosszabb szavak, az elöljárós frázisok, az attributív melléknevek, a passzív igei szerkezetek és az ige+*that*-tagmondatok szignifikánsan gyakrabban fordulnak elő az írott szövegekben, valamint magasabb az együtt-előfordulásuk aránya a humán értékelők által magasabbra értékelt szövegekben.

A jelen vizsgavalidálási projekt ezekhez a kutatási eredményekhez kíván hozzájárulni. Az alább bemutatott korpusz adatait a fenti és további komplexitásváltozókkal elemezzük (közel 120 változóval), abból a célból, hogy beazonosítsuk az adott vizsga küszöbét (B2+) elérő, illetve az az

alatt maradó szövegeket. Mivel a cél az elégséges, de nem feltétlenül minimális számú változók kiválasztása, statisztikai metaanalízis segítségével, a különböző sokváltozós elemzőrendszerek eredményeinek összehasonlításával keressük a legmegbízhatóbb változókat.

2.2. Nyelvi komplexitás és nyelvhelyesség a *KER*-ben

Egyáltalán nem triviális kérdése a nyelvtudás értékelésének, hogy hogyan lehet a *KER* szerinti angolnyelvi tudásszinteket az angol nyelvtan és szókincs jellemzőinek segítségével megfogalmazni, hiszen a *KER* (*CEFR*, 2001, 2018, magyarul 2002) – bármely idegen nyelvre való alkalmazhatósága miatt – nem fogalmaz meg konkrétan az egyes nyelvekre, így az angolra vonatkoztatható deszkriptorokat. (A *KER* megjelenését megelőzően, az 1970-1990-es években történtek erőfeszítések az L2-angol nyelvtanának szintezésére, lásd például Van Ek 1975, Van Ek-Trim 1991a, Van Ek-Trim 1991b; Van Ek-Trim 2001.)

A deszkriptorok segítségével a *KER* (2002) mind a nyelvi komplexitást, mind a nyelvhelyességet tárgyalja, ezekre rendre a *range* és a *control* kifejezéseket használja, a magyar fordításban ezek *terjedelem* illetve *alkalmazás/helyesség*. Az e projekt fókuszában lévő B2+ szintet nem tárgyalja a *KER* részletesen, de a szomszédos B2 és C1 szinteket igen. Ami a nyelvtani és szókincskomplexitást illeti, az *Általános nyelvi készségek* kategóriában a B2 szint leírása többek között ezeket emeli ki: „elegendően széleskörű nyelvtudás", „összetett mondatformák alkalmazás[a]", a C1 szintet így írja le: „széleskörű nyelvtudásából ki tudja választani a megfelelő nyelvi formát" (*KER* 2002: 132). A *Szókincs terjedelme* kategóriában a B2 szint leírására többek között ez szerepel: „jó szókincs", „variálni tudja", de vannak „lexikai hiányok", míg a C1 szintnél: „széles körű szókincs", „ritkán kell keresgélnie a kifejezéseket" (*KER* 2002: 134).

A *KER*-szintek és a 'mire képes'-deszkriptorok (*can-do*) rendkívül hasznosak a különböző nyelvtudásszintek fogalmának meghatározásában. Ugyanakkor ahhoz, hogy megtaláljuk azokat az L2-angol nyelvtani és szókincsjellemzőket, amelyek mintázatai az e projekt számára fontos szinteket (B2+ és afölött *vs.* B2+ alatt) jellemzik, a *KER* által kínált útmutatás jellegéből adódóan korlátozott. Ezeket a jellemzőket érdemes tanulói korpusz alapján, a nyelvtani változók mintázatai segítségével összegyűjteni. A következő fejezetekben, a konkrét nyelvvizsga és az összegyűjtött korpusz bemutatását a sokváltozós kvantitatív elemzéseink eddigi eredményeinek ismertetése követi.

3. Kutatási módszerek

3.1. Adatgyűjtés: A nyelvvizsga és a korpusz

A kutatás egy most épülő korpuszon alapul, amely egy magyarországi egyetem egyik nyelvvizsgáján, az ún. *Alapvizsgán* (AV) írt L2-angol tanulói szövegeket tartalmaz. Hasonlóan több magyar egyetem gyakorlatához, az AV egy angolszakosok számára az első tanévük végén szervezett, magas kockázatú, B2+ szintű nyelvtudást mérő esemény, beleértve a BA (nappali, levelezős) és az angoltanárképzésben részt vevő hallgatókat is. A neve azért „Alapvizsga", mert sikeres letétele a hallgató további tanulmányainak alapja. Évente körülbelül 150 hallgató teszi le, az őszi és tavaszi félévben.

A vizsga azért magas kockázatú, mert az egyetem szabályzata szerint a tanulmányok során összesen csak kétszer tehető le: sikertelenség esetén csak egyszer ismételhető valamelyik következő félévben. A bukási arány miatt rendszeres kritika éri a szigorúsága miatt. Ezért is kulcsfontosságú a vizsga validálása, annak bizonyítására, hogy az a szint, amelyhez képest a vizsga a jelöltek nyelvtudását méri, valóban B2+ és nem magasabb.

Az írásbeli rész egy nyelvtani és szókincstesztből (*Use of English*), egy olvasási tesztből és egy írottszöveg-alkotási részből áll – jelenleg az ez utóbbin írt szövegek állnak a kutatás fókuszában. A vizsga szóbeli részére egy másik napon kerül sor.

A 45 perces írottszöveg-alkotási részben a vizsgázók két vagy három, részletes feladatleírásban megjelölt téma közül választhatnak, melyek mindegyike három szövegműfaj (formális levél vagy

narratíva vagy recenzió/kritika) valamelyikéhez tartozik. A cél 180-200 szavas szöveg megalkotása. A vizsgázók által írt minden egyes szöveget két vak bíráló értékeli egy négy kategóriából álló értékelési skála alapján: feladatteljesítés, koherencia/kohézió, nyelvtan, szókincs; az utóbbi két kategória deszkriptorai mind a komplexitásra, mind a helyességre kitérnek.

Az Alapvizsga-korpusz (AVK) jelenleg 395, egyenként körülbelül 200 szavas formális levél- és narratíva-műfajú szöveget tartalmaz. Valamennyi szöveg gyűjtése a Brit Alkalmazott Nyelvészeti Társaság etikus kutatásra vonatkozó ajánlásai (BAAL 2021) szerint, a szövegek kutatási és publikációs célokra való felhasználása a szerzők írásbeli tájékozott beleegyezése (*informed consent*) alapján, az anonimitás védelmével történik. Mivel a szövegek a legtöbb AV-n kézírással készülnek, ezeket gondosan átírjuk, két változatban: a kézi elemzéshez készülő szó szerinti változatban megtartjuk az eredeti szöveg sajátosságait, kivéve a szerző nevét, a gépi elemzéshez készülő másik változatban a helyesírást és a központozást a brit helyesírási konvencióknak megfelelően alakítjuk át.

3.2 Elemzési módszerek

Az AVK elemzése során négy külső, sokváltozós gépi nyelvtani és szókincskomplexitást elemző program mellett néhány saját, gépi és kézi elemzésen alapuló változót alkalmazunk (lásd 1. táblázat). A külső elemzőrendszerek a Douglas Biber és munkatársai által kidolgozott Sokdimenziós Elemzőtagger (*Multidimensional Analysis Tagger*, MAT; Nini 2019, 2021), a webalapú Nem-anyanyelvi Nyelvtani Komplexitáselemző (*L2 Syntactic Complexity Analyzer*, L2SCA; Lu 2017, Ai 2022), a *KER*-alapú Szókincs-szint Elemző (*Vocabulary Level Analyzer*, CVLA 2023; Uchida-Negishi 2018) és a Lextutor (Cobb 2023); legtöbbjük ismert nyelvtani és szókincselemző alapprogramokra (*Stanford Parser, General Service List*) támaszkodik. A saját változóink fejlesztése folyamatban van, jelenleg egy automatizált (többi/K1 tokenek) és több kézi változóval dolgozunk (igeidő- és -aspektuseloszlás, névelőhasználat vs. megszámlálhatóság, vizsgáztatói pontszámok és nyelvhelyességi elemzések). A fenti sokváltozós elemzőprogramok által kapott eredményeket az összehasonlíthatóság érdekében normalizált formára hozzuk (per 100 szövegszó), majd statisztikai elemzés alá vetjük.

1. táblázat. Az alkalmazott sokváltozós elemzőprogramok és változóik

Nyelvtani komplexitás	**Szókincskomplexitás**
Biber-tagger/MAT (Nini 2019, 2021):	
- egyszerű múltidejű igealakok - perfektív aspektusú igealakok - nominalizációk - *there is/there are*-szerkezetek - *be* főigeként - ágens nélküli / *by* passzívok - különféle vonatkozói mellékmondatok - módbeli segédigés szerkezetek, stb.	- kötőszavak (*moreover*) - lefokozók (*almost, nearly*) - privát jelentésű igék (*consider, realise*) - publikus jelentésű igék (*announce, explain*) - helyhatározók, időhatározók - attributív, predikatív melléknevek, stb. (összesen 67 változó)
L2SCA (Lu 2017, Ai 2022): - átlagos tagmondat/mondat/T-egységhossz - alárendelés-arány (pl. tagmondat/mondat) - mellérendelés-arány - frazális alárendelés-arány, stb. (kb. 20 változó)	**Lextutor-alapú változók (Cobb 2023):** - szócsalád/típus/token-eloszlások - K1-K2-K3 és a fölötti gyakorisági eloszlások - típus-token arány - szókincs-sűrűség (tartalmasszó/token), stb. (kb. 20 változó)
CVLA (2023, Uchida-Negishi 2018):	
- ige/mondat - ARI (ez egy olvashatósági mérőszám)	- átlagos szókincsnehézség - B per A (a két gyakorisági változó aránya) (összesen 4 változó + 1 összesítő mutató)
Saját változóink – kézi: nyelvhelyességi változók; igeidő/aspektuseloszlás, névelőhasználat vs. megszámlálhatóság; vizsgáztatói pontszámok, stb.	**Saját változóink – gépi**: többi/K1 tokenek (Lextutor alapján)

4. Adatelemzés

4.1. Rövid összefoglaló a statisztikai adatelemzésről

Projektünkben jelenleg a korreláció és a statisztikai faktorelemzés látszik legalkalmasabbnak az adatmátrixban lévő belső összefüggések és látens változók feltárására, az adott elemzőprogramokon belül és azok között is – utóbbiakat statisztikai keresztelemzésnek nevezem. Az adatkódolás és a számítások folyamatban vannak; az alábbi eredmények részben Radnay (2017), Adamova (2022), Reményi-Velner (2022), Velner (2022) és Reményi (2024) számításai alapján készültek. A statisztikai keresztelemzés első eredményeit részletesebben máshol ismertetem (Reményi 2024) – abban a tanulmányban azt a kérdést is tárgyalom, hogy hogyan lehetséges a változóink alapján a tanulói szövegek *KER*-szintjének bejóslása.

A projektben több mint 110 változóval dolgozunk. Megpróbáljuk megtalálni azt a korlátozott, bár nem feltétlenül minimális számú változót és együtt-előfordulásaikat, amelyekkel előrejelezhetjük a B2+ szintű, illetve az az alatti szintű nyelvtani komplexitás specifikus mintázatait. Itt egyetlen példát mutatok be ennek alátámasztására.

A 2. táblázat korrelációs mátrixán az látható, hogy a 2022 áprilisi alkorpusz adatai alapján a CVLA-változók, egyes Lextutor-változók és az egyik saját változónk közül melyek korrelálnak szignifikánsan egymással. Függőlegesen a CVLA-változók közül a „CVLAnum" az összefoglaló mutató a *KER*-alkategóriává való átalakítás előtt, az „ARI" az olvashatósági változó, az „Ige/mondat" a nyelvtani komplexitásváltozó, az „Átlagnehézség" és a „BperA" a szókincsváltozók. Vízszintesen a Lextutor-alapú szókincsváltozók a szócsaládok/típusok/tokenek, a token-szinten leggyakoribb 1000 szó (K1), a második leggyakoribb 1000 szó (K2), az ezek feletti összes token (K3up), majd a szókincs-sűrűség. A jobb szélső oszlop saját változónk, amely a K1 tokenek arányát számítja az összes többihez képest (többi/K1 token). Az árnyékolt korrelációs együtthatók $p < 0,01$ és $p < 0,05$ szinten szignifikánsak.

A táblázatban a CVLA-változók közül az „Ige/mondat" változó az egyetlen, amely nem látszik korrelálni sem a Lextutorból kinyert, sem a saját változóinkkal. Ez alátámasztja módszerünk helyességét, mivel az „Ige/mondat" egy nyelvtani komplexitás-változó, a többi változó viszont a szókincskomplexitással kapcsolatos. A többi CVLA-változót mind megerősíti néhány más változó: például az "Átlagnehézség" és a "többi/K1 token" erős pozitív korrelációt mutat (r=0,747), azaz minél magasabb a szó nehézségének átlagos értéke a CVLA-ben, annál magasabb a K1 feletti tokenek aránya a szövegekben, és fordítva. Hasonlóképpen, a "BperA" és a "többi/K1 token" között erős pozitív korreláció van (r=0,663), az "Átlagnehézség" és a "K2 token" között pedig közepes erősségű pozitív korreláció (r=0,543). A "K1 token" és a "CVLAnum", az "Átlagnehézség" és a "BperA" közötti közepes erősségű negatív korreláció azt jelenti, hogy minél magasabb a nagy gyakoriságú K1 szavak aránya egy szövegben, annál alacsonyabb lesz a többi változó értéke, és fordítva. Ami a lexikográfiában hagyományosan használt szócsalád/típus/token-változókat illeti, úgy tűnik, hogy ezeket a CVLA-változók nem támasztják alá.

2. táblázat. Spearman-korrelációk a CVLA, a Lextutor és az egyik saját változó között a 2022 áprilisi alkorpuszban (N=98; ** $p < 0,01$, * $p < 0,05$)

CVLA-változók		Lextutor-változók								Saját
		Szó-család	Típus	Token	K1 token	K2 token	K3up token	Típus/ token	Szókincs-sűrűség	többi/K1 token
	CVLAnum	0,034	0,066	-0,084	-0,263**	0,471**	0,307**	0,233*	0,167	0,535**
	ARI	0,069	0,072	0,011	-0,093	0,404**	0,058	0,038	0,151	0,279**
	Ige/mondat	-0,011	-0,008	0,039	0,048	0,102	-0,115	-0,130	-0,106	-0,069
	Átlagnehézség	0,002	0,004	-0,174	-0,396**	0,543**	0,444**	0,371**	0,249*	0,747**
	BperA	0,019	0,029	-0,167	-0,365**	0,407**	0,492**	0,328**	0,229*	0,663**

A statisztikai keresztelemzésben az itt bemutatott korrelációelemzés mellett faktorelemzéssel vizsgáljuk azt, hogy az egyes elemzőprogramok mely változóit érdemes figyelembe venni a tanulói korpusz nyelvi komplexitásának elemzésére. A fenti példából az látszik, hogy módszerünk alkalmas nagy számú komplexitásváltozó elemzésére, és a keresztelemzés segítségével azon változók kinyerésére, amelyek leginkább jellemzőek a B2+ nyelvtudás-szint körüli komplexitásmintázatokra.

4.2. Kísérlet a transzformer-alapú adatelemzésre

A nyelvtanulói szövegek automatizált elemzésére két út kínálkozik. A statisztikai elemzés azon az elven alapul, hogy a kutató előre kijelöli azokat a jellemzőket, amikre azután az elemzés épül; ezeknek a változóknak a relevanciáját azután legfeljebb a látens változós elemzési eszközök segítségével finomíthatjuk. A másik út a neurális hálós nagy nyelvmodellek illetve gépi tanulás-alapú rendszerek; ezeknél a program a gépi tanulás során maga alakítja ki a vektorszámításon alapuló rejtett rétegek és súlyok rendszerét, ami azután az outputot létrehozza, ebbe a rendszerbe a kutató nem lát bele.

Ezen belül a transzformerek olyan mélytanulás-alapú neurális hálózatos architektúrát jelentenek, amelyet arra terveztek, hogy szekvenciális adatokból legyen képes kontextust és jelentést tanulni. A 2022 végén megjelent ChatGPT (2023) ilyen transzformer, amit kifejezetten csevegésre optimalizáltak (*chatbot*). Megtapasztalhattuk, hogy meglepően jól teljesít akár angol, akár magyar nyelvű beszélgetésben (bár az angol nyelvű bemeneti korpusza sokszorosa a magyar nyelvűének, és ennek megfelelően az előbbi nyelven gyakran jobb eredményt ér el). Ráadásul nemcsak beszélgetésben, hanem írott szövegek javításában-fejlesztésében is jól teljesít: képes javítani egy beadott szöveg nyelvtani komplexitását, szókincskomplexitását, bekezdésszerkezetét, szövegkohézióját, nyelvhelyességét és stílusmegfelelését. Sőt, képes egy megjelölt nyelvi szintnek megfelelően egyszerűbb vagy komplexebb szöveget is létrehozni.

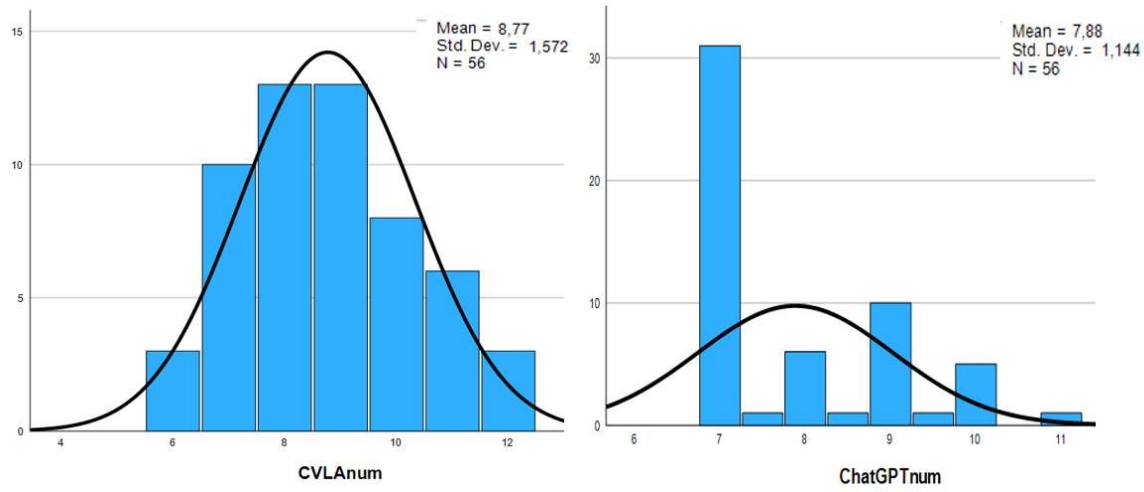
Ennek alapján felmerül a kérdés, hogy vajon nemcsak a szöveggenerálásban képes-e különféle nyelvi szintű szöveget létrehozni vagy átalakítani, hanem vajon az ellenkező irány is megoldható-e a számára: meg tudja-e állapítani beadott szövegekről, hogy azok milyen szintűek? Könnyen ellenőrizhető, hogy a *KER*-szintek fogalmával tisztában van.

Próbavizsgálatot végeztem: az AVK-ból nyert kis random mintán (N=56) kipróbáltam, hogy (1) mennyire egyezik a *KER*-szint bejóslási eredménye ugyanazon szövegek statisztikai alapú bejóslásával, a CVLA-eredményekkel összevetve, illetve (2) a ChatGPT-eredmények eloszlása megbízhatónak tűnik-e. A beszélgetések angol nyelven folytak. A kiinduló bemeneti prompt mindig azonos volt: „Milyen *KER* szintű ez az XYZ szöveg? [bemásolt szöveg]". Amennyiben a válaszban nem szerepelt a *KER*-alszintek megnevezése is (például „magasabb B2"), akkor szintén azonos prompttal kérdeztem tovább: „Ez az XYZ szöveg magasabb vagy alacsonyabb [szint] szintű szöveg a *KER* szerint?" A ChatGPT válasza néha rövidebb volt, például: „Ez az XYZ szöveg alacsonyabb B2-szintűnek tekinthető." Máskor részletesebb:

> A nyelvhasználat és a kifejezett gondolatok komplexitása alapján ez a XYZ szöveg a KER szerinti alsó B1 szintűnek tűnik. A szövegíró képes megfogalmazni a panaszát és javítási javaslatokat tenni, de ejt néhány olyan nyelvtani és szókincshibát, amik azt jelzik, hogy nyelvtudása még fejlődésben van. Ehhez hozzájárul még az is, hogy a szövegszerveződés és -struktúra még fejlesztendő, hogy világosabb és koherensebb legyen.

Az eredmények azonban azt jelezték, hogy a ChatGPT nem alkalmas a *KER*-szint megbízható bejóslására. Egyrészt összezavarható. Amikor ugyanis a statisztikai eredményektől eltérő szintet jelzett, mindig rákérdeztem: „Biztos, hogy a […] szöveg nem magasabb szintű, mint felső [szint]? Esetleg nem lehet [szint]?" Ilyenkor több esetben szabadkozott, visszavonta a jelzett szintet vagy módosította, például: „Elnézést kérek a fenti válaszomban ejtett hibáért. Újraolvasva a szöveget, azt mondanám, hogy a ZYX szöveg *KER*-szintje talán C1, a szövegkomplexitást tekintve." Más esetekben kitartott az előző szintmegállapítása mellett. (Helyszűke miatt itt nincs mód bemutatni azon

erőfeszítéseimet, amelyekkel több lépésben egyre mélyebb metareflexióra, a nyelvtani elemzés mélyítésére igyekeztem rávenni a programot.)



1. ábra. A CVLA- és ChatGPT-szintbejóslások illeszkedése a normális eloszlásra (N=56; 6: A2+, 7: B1-, 8: B1+, 9: B2-, 10: B2+, 11: C1, 12: C2 szint)

Másrészt a próbavizsgálat szerint a ChatGPT bejóslásában nem tűnik megbízhatónak, sem az eloszlása, sem a statisztikai elemzéssel bejósolt szintekkel való korrelációja alapján. Az 1. ábrán illeszkedésvizsgálat nélkül, szabad szemmel is jól látható, hogy a CVLA-elemzéssel nyert szintbejóslások eloszlása jól illeszkedik a normális eloszlás görbéjére, ahogy az várható is az ilyen jellegű eredményeknél. Ezzel szemben a ChatGPT-ből kinyert eredmények nem követik a normális eloszlást. A kétféle eredmény összevetése azt mutatja, hogy közöttük az együttjárás az alacsony-közepes tartományba esik (korreláció: $r=0,409$, $p<0,01$).

Összefoglalva, a kismintás próbavizsgálat azt mutatja, hogy a ChatGPT nem alkalmas az L2-angol szövegek nyelvtudásszintjének megbízható bejóslására. Egy lehetséges más, szövegelemzésre optimalizált transzformerprogram azonban esetleg alkalmas, sőt, kívánatos lehet a nyelvtanulói szövegek vizsgálatára is.

5. Konklúzió

Az angol nyelvvizsgák képzett vizsgáztatói világszerte évente több százezer vizsgázó írásbeli szövegeit értékelik. Értékelő munkájuk az időkorlátok miatt impresszionisztikus, és a megbízhatóság biztosítása érdekében három forrásra támaszkodik: a vizsgáztatók képzettségére és tapasztalatára, az adott vizsgán alkalmazott értékelőskálák deszkriptoraira, valamint a vizsgáztatók közötti együttműködésre. Bár a nyelvtani és szókincskomplexitás munkájuknak csak két értékelendő komponense, statisztikai alapú kutatásunk – és benne e két komplexitáskategória jellemzőinek számszerűsítése – úgy is felfogható, mint a vizsgáztatók látens elemzői erőfeszítésének feltárása és modellezése.

Egyes vizsgáztatói értékelési mozzanatok statisztikai/változóalapú gépi elemzéssel való megoldhatósága tehát részben azért lenne hasznos, mert megkönnyítené a vizsgáztatók munkáját, részben azért, mert átláthatósága miatt megbízhatóbbá tenné azt. A nyelvvizsgáztatási gyakorlatban már léteznek erre megoldások (például az e-rater a TOEFL-vizsga esetében, lásd Ramineni et al. 2018), de használatuk korlátozott. A transzformer-alapú elemzések is elindultak már, egyelőre ellentmondó eredményekkel (Mizumoto-Eguchi, 2023) – ezeknek az elemzéseknek az a hátulütője, mint minden neurálisháló-alapú munkának: nem transzparensek, mert a programban a vektorszámítások által konstruált rejtett rétegek és súlyok rendszerébe a humán elemző nem kap betekintést.

A fenti kutatási kérdésekre az a válasz adható, hogy (1.) a tanulói szövegekből gépi és kézi elemzéssel kinyert nyelvtani és szókincskomplexitási eredmények statisztikai keresztelemzése lehetővé teszi a legfontosabb változók kinyerését, a szövegek nyelvtudásszintjének bejóslását és a továbbiakban a fenti nyelvvizsga validálását. És másrészt (2.) úgy tűnik, hogy a nyelvtudásszintek bejóslására a ChatGPT nem alkalmas alternatívája a statisztikai alapú elemzési módszereknek.

Végül hadd mutassak rá arra, hogy a vizsgavalidációs projekt eredetileg csak lokálisan releváns és praktikus célt szolgált volna, ami a résztvevőkön kívül keveseket érdekelhet. A projektről menet közben derült ki, hogy általánosabb, globálisan releváns célok szolgálatába is állítható. A nyelvtudás értékelésének automatizálhatóságával jelenleg számos nemzetközi kutatás foglalkozik, egy nyüzsgő, innovatív kutatási térbe csöppentünk így bele. A kutatások eddig – a miénkhez hasonlóan – statisztikai alapúak voltak. Várható azonban, hogy az automatizált nyelvi értékelésben megjelennek a neurális hálós, nagy nyelvmodell-alapú elemzések is.

Irodalom

Adamova, K. 2022. *A Multidimensional Analysis of L2 Learners' Writings at the B2+ Level.* Kézirat. Budapest: Pázmány Péter Katolikus Egyetem.

Ai, H., Lu, X. 2013. A corpus-based comparison of syntactic complexity in NNS and NS university students' writing. In: Díaz-Negrillo, A., Ballier, N., Thompson, P. (eds.) *Automatic Treatment and Analysis of Learner Corpus Data*. (Studies in Corpus Linguistics 59.) Amsterdam: Benjamins. 249-64. https://doi.org/10.1075/scl.59.15ai

Biber, D., Gray, B., Staples, S. 2016. Predicting Patterns of Grammatical Complexity Across Language Exam Task Types and Proficiency Levels. *Applied Linguistics* Vol. 37. Is. 5. 639–668. https://doi.org/10.1093/applin/amu059

Bulté, B., Housen, A. 2012. Defining and operationalising L2 complexity. In: Housen, A.,Kuiken, F., Vedder, I. (eds.) *Dimensions of L2 Performance and Proficiency. Complexity, Accuracy and Fluency in SLA.* (Language Learning & Language Teaching 32.) Amsterdam: Benjamins. 21–46. https://doi.org/10.1075/lllt.32.02bul

Bulté, B., Housen, A. 2014. Conceptualizing and measuring short-term changes in L2 writing complexity. *Journal of Second Language Writing* Vol. 26. 42–65. https://doi.org/10.1016/j.jslw.2014.09.005

CEFR 2001. *Common European Framework of Reference for Languages. Learning, teaching, assessment*. Strasbourg–Cambridge: Council of Europe – Cambridge University Press. https://rm.coe.int/1680459f97 (Hozzáférés: 2024. 10. 23.)

CEFR 2018. *Common European Framework of Reference for Languages: Learning, teaching, assessment: Companion volume with new descriptors.* Strasbourg: Council of Europe. https://rm.coe.int/cefr-companion-volume-with-new-descriptors-2018/1680787989 (Hozzáférés: 2024. 10. 23.)

Crossley, S. 2020. Linguistic features in writing quality and development. An overview. *Journal of Writing Research* Vol. 11. No. 3. 415–443. https://doi.org/10.17239/jowr-2020.11.03.01

Crossley, S., McNamara, D. 2014. Does writing development equal writing quality? A computational investigation of syntactic complexity in L2 learners. *Journal of Second Language Writing* Vol. 26. 66–79. https://doi.org/10.1016/j.jslw.2014.09.006

Housen, A., Kuiken, F. 2009. Complexity, Accuracy, and Fluency in Second Language Acquisition. *Applied Linguistics* Vol. 30. Is. 4. 461–473. https://doi.org/10.1093/applin/amp048

KER 2002. *Közös Európai Referenciakeret: Nyelvtanulás, nyelvtanítás, értékelés.* Budapest: Pedagógustovábbképzési és Módszertani Információs Központ. https://nyak.oh.gov.hu/nyat/doc/ker_2002.asp (Hozzáférés: 2024. 10. 23.)

Lei, L., Wen, J., Yang, X. 2023. A large-scale longitudinal study of syntactic complexity development in EFL writing. A mixed-effects model approach. *Journal of Second Language Writing* Vol. 59. Art. 100962. https://doi.org/10.1016/j.jslw.2022.100962

Lu, X. 2010. Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics* Vol. 15. Is. 4. 474–496. https://doi.org/10.1075/ijcl.15.4.02lu

Lu, X. 2017. Automated measurement of syntactic complexity in corpus-based L2 writing research and implications for writing assessment. *Language Testing* Vol. 34. Is. 4. 493–511. https://doi.org/10.1177/0265532217710675

Mancilla, R. L., Polat, N., Akcay, A. O. 2015. An Investigation of Native and Nonnative English Speakers' Levels of Written Syntactic Complexity in Asynchronous Online Discussions. *Applied Linguistics* Vol. 38. Is. 1. 1–24. https://doi.org/10.1093/applin/amv012

Mizumoto, A., Eguchi, M. 2023. Exploring the potential of using an AI language model for automated essay scoring. *Research Methods in Applied Linguistics.* Vol. 2. Is. 2. Art. 100050. https://doi.org/10.1016/j.rmal.2023.100050

Nini, A. 2019. The Multi-Dimensional Analysis Tagger. In: Berber Sardinha, T., Veirano Pinto, M. (eds.) *Multi-Dimensional Analysis. Research Methods and Current Issues.* London: Bloomsbury. 67–94.

Ortega, L. 2003. Syntactic Complexity Measures and their Relationship to L2 Proficiency. A Research Synthesis of College-level L2 Writing. *Applied Linguistics* Vol. 24. Is. 4. 492–518. https://doi.org/10.1093/applin/24.4.492

Polat, N., Mahalingappa, L. J., Mancilla, R. 2019. Longitudinal Growth Trajectories of Written Syntactic Complexity. The Case of Turkish Learners in an Intensive English Program. *Applied Linguistics* Vol. 41. Is. 5. 688–711. https://doi.org/10.1093/APPLIN/AMZ034

Radnay Zs. 2017. *Corpus-Based Validation of the Basic Language Examination at Pázmány Péter Catholic University. An Analysis of the Written Production.* Szakdolgozat. (PPKE BTK). Kézirat. Budapest.

Ramineni, C., Trapani, C. S., Williamson, D. M., Davey, T., Bridgeman, B. 2012. Evaluation of the *e-rater*® Scoring Engine for the *TOEFL*® Independent and Integrated Prompts. *Research Report ETS RR-12-06.* Princeton: Educational Testing Service. https://files.eric.ed.gov/fulltext/EJ1109838.pdf (Hozzáférés: 2024. 10. 23.)

Reményi A. Á. 2024 Evaluating Written L2-English Learner Texts through Manual and Automated Quantitative Analyses. In: Dósa A., Magnuczné Godó Á., Schäffer A., Nagano, R. L. (eds.) *Space, Identity and Discourse in Anglophone Studies. Crossing Boundaries* Newcastle: Cambridge Scholars Publishing.

Reményi A. Á., Velner P. 2022. *Validating an EFL Examination Through the Manual and Automated Quantitative Analysis of Candidates' Written Texts.* Előadás az ESSE2022 konferencián. Mainz. https://esse2022.uni-mainz.de/ (Hozzáférés: 2024. 10.23.)

Taguchi, N., Crawford, W., Wetzel, D. Z. 2013. What Linguistic Features Are Indicative of Writing Quality? A Case of Argumentative Essays in a College Composition Program. *TESOL Quarterly* Vol. 47. Is. 2. 420–430. https://doi.org/10.1002/tesq.91

Uchida, S., Negishi, M. 2018. Assigning CEFR-J levels to English texts based on textual features. In: Tono, Y., Isahaara, H. (eds.) *Proceedings of the 4th Asia Pacific Corpus Linguistics Conference (APCLC 2018).* Takamatsu: APCLA. 463–467.

Van Ek, J. A. 1975. *The Threshold Level.* Strasbourg: Council of Europe.

Van Ek, J. A., Trim, J. L. M. 1991a. *Threshold 1990*. Strasbourg–Cambridge: Council of Europe – Cambridge University Press. https://ealta.eu/documents/resources/Threshold-Level_CUP.pdf (Hozzáférés: 2024. 10. 23.)

Van Ek, J. A., Trim, J. L. M. 1991b. *Waystage 1990*. Strasbourg– Cambridge: Council of Europe – Cambridge University Press. https://ealta.eu/documents/resources/Waystage_CUP.pdf (Hozzáférés: 2024. 10. 24.)

Van Ek, J. A., Trim, J. L. M. 2001. *Vantage*. Strasbourg–Cambridge: Council of Europe – Cambridge University Press.

Velner P. 2022. *CEFR-Based Assessment with Computational Methods. Investigating Accuracy and Complexity in English Learner Texts*. Szakdolgozat. (PPKE BTK). Kézirat. Budapest.

Források

Ai, H. 2022. *Web-Based L2 Syntactic Complexity Analyzer*. Webalapú/letölthető program. https://aihaiyang.com/software/l2sca/

BAAL 2021. *Recommendations on Good Practice in Applied Linguistics.* British Association for Applied Linguistics. https://www.baal.org.uk/wp-content/uploads/2021/03/BAAL-Good-Practice-Guidelines-2021.pdf

ChatGPT 2023. február 13-i és augusztus 3-i verzió. Webalapú program. https://chat.openai.com/chat

Cobb, T. 2023. *LexTutor VocabProfilers*. Webalapú program. https://www.lextutor.ca/vp/

CVLA 2023. *CVLA: CEFR-Based Vocabulary Level Analyzer (ver. 2.0)*. Webalapú program. https://cvla.langedu.jp/index.html

Nini, A. 2021. *The Multidimensional Analysis Tagger.* Webalapú/letölthető program. https://sites.google.com/site/multidimensionaltagger