In: Fóris, Ágota; Bölcskei, Andrea; Nádor, Orsolya; Sólyom, Réka (szerk.) Nyelv, kultúra, identitás. Alkalmazott nyelvészeti kutatások a 21. századi információs térben : V. Nyelvpedagógia, nyelvoktatás, nyelvelsajátítás Budapest, Magyarország : Akadémiai Kiadó (2020)

ADORJÁN MÁRIA

Károli Gáspár Református Egyetem

adorjan.maria@kre.hu

Az automatizált esszéértékelés működése, lehetőségei és korlátai

Automated essay scoring in a language teaching context: design, potential and limitations

Abstract

The emergence of large-scale standardized testing and the demand for fast and reliable feedback have given rise to automated essay scoring at international examination centers. While the first fully automated assessment applications examined texts as quantifiable entities, the latest programs are more sophisticated and versatile. Publications (mainly written by program developers) indicate high inter-rater reliability. First, empirical studies are described which explain how the programs were developed. Next, methodological issues are discussed concerning the design, applications and limitations of automated essay scoring instruments.

keywords

automated essay scoring, language assessment, assessment of writing, features of L2 writing, automated and human raters

Absztrakt

Az elmúlt évtizedekben egyre több vizsgaközpontban alkalmaznak automatizált esszéértékelő programokat a gyors és hatékony vizsgáztatás érdekében. A legújabb programok összetett módszertannal dolgoznak, és sokrétű a használatuk. A kutatások alapján úgy tűnik, képesek megközelíteni a emberi vizsgáztatók értékelését. Néhány fontos empirikus tanulmány ismertetésével bemutatom a legismertebb esszéértékelő programok működését, lehetőségeit, korlátait és használatának következményeit nyelvpedagógiai kontextusban.

kulcsszavak

automatizált esszéértékelés, idegen nyelvi értékelés, írás értékelés, idegen nyelvi írás jellemzői, automatizált és emberi értékelők

1. Bevezetés

A nagy nemzetközi vizsgaközpontok – különösképpen az USA-ban – már évtizedek óta használnak számítógépes programokat a standardizált vizsgáztatás és szintfelmérés hatékonyabbá tételére. Feleletválasztós tesztek esetében mindenki számára egyértelműek a

számítógépes programok használatának előnyei. Sok nyelvtanár azonban nincs tisztában azzal, hogy egyre több vizsgaközpont használ programokat összefüggő szövegek, például érvelő esszék szövegminőségének a vizsgálatára is, legtöbbször egy gépi — egy emberi értékelő felállásban. Az esszék értékelésének automatizálását a tömeges vizsgáztatás idő- és forrásigénye tette szükségessé, és a szövegnyelvészet, a számítógépes nyelvészet, valamint a nagy adatmennyiséget feldolgozni tudó statisztika fejlődése tette megvalósíthatóvá. Mivel már Magyarországon is lehet például TOEFL iBT[i] vizsgát tenni, fontos, hogy a nyelvtanárok megismerjék az írásbeli vizsgarész értékelésének menetét és a pontozás részleteit tanulóik hatékony felkészítése érdekében. A nyelvtanárok másik feladata a tanulók önálló tanulási és önértékelési stratégiáinak fejlesztése, amelyhez szintén elengedhetetlen az automatizált esszé értékelő programok kellő ismerete (Chapelle, Douglas 2006).

Az automatizált esszéértékelésnek jelentős angol nyelvű szakirodalma van. Ennek megismerését és nyelvpedagógiai hasznosulását azonban két dolog nehezíti. Az első probléma, hogy a legtöbb tanulmányt nem nyelvtanárok írták, hanem mérés-értékeléssel foglalkozó szakemberek, és ugyanők a célközönség is. Ebből következően, a tanulmányok az egyedi programleírásokra koncentrálnak, és nagyszámú zsargont használva, részletes statisztikai elemzéseket tartalmaznak. Ezek a tanulmányok ugyanakkor azért is nehezen vagy egyáltalán nem használhatók nyelvpedagógiai kutatások során, mert – a fent említett eltérő célközönségen kívül – olyan alapvető információkat nem tartalmaznak, amelyre egy nyelvtanárnak vagy vizsgára felkészülő nyelvtanulónak szüksége lehet. Ilyenek például a feladat és konkrét vizsgahelyzet hiányos leírása, amely gyakorló tanárok számára fontos információforrás lenne. Az írásbeli feladat megadásakor például fontos definiálni, hogy érvelő esszéről van-e szó. Szintén fontos az esszé hossza, a feladat kontextusa: a vizsgázók használhatnak-e segédeszközt, mennyi idő áll rendelkezésre, stb. E faktorok hiányos leírása miatt a tanulmányokat nehéz érdemben összehasonlítani.

Jelen írás az angol nyelvű tudományos folyóiratokban megjelent empirikus kutatásokból válogatva a nyelvpedagógus szemszögéből arra a kérdésre keresi a választ, hogy milyen módon lehetnek képesek ezek a programok egy olyan összetett, sokrétűségű nyelvi produktum megfelelő értékelésére mint egy nyelvvizsga esszé. A programok működése mellett, a másik két kutatási kérdés az automatizált esszéértékelés már feltárt lehetőségeit és korlátait vizsgálja. Mivel jelen írás keretei nem teszik lehetővé a téma átfogó bemutatását, azok számára, akik részletesebben szeretnének tájékozódni a kérdésről, magyar nyelven rendelkezésre áll többek között Molnár (2010) tanulmánya, melyben a szerző részletesen leírja a technológia alapú mérés-értékelés amerikai, európai és magyar helyzetét. Angol nyelven Dikli (2006) bevezető tanulmánya, illetve Haswell és társai (2012) annotált bibliográfiája nyújt segítséget a téma jobb megismeréséhez.

2. Empirikus kutatási irányok

A számítógéppel támogatott szövegértékelés kezdeti stádiumában a programfejlesztők azokra az elméletekre és empirikus kutatásokra támaszkodtak, amelyek egyrészt a szövegek jellemzőinek leírására, másrészt a szövegértékelés folyamatának minél jobb megismerésére irányultak. Jelen tanulmányban a szövegre irányuló kutatások közül egyetlen kutatási irányt: a szövegkohézió, azon belül is a lexikai kohézió kutatását emelem ki mint a legfontosabb tudományterületet az automatizált esszéértékelés szempontjából. A területen „alapkutatásnak” számít Halliday és Hasan kohéziós modellje (1976), valamint az azt átalakító Hoey lexikai ismétlődés modellje (1991), aki szerint a lexikai ismétlődés a legfontosabb szövegszervező erő, és egyben az egyetlen olyan kohéziós eszköz, amely többszörös kapcsolatot alakít ki különböző irányokban a diskurzuson belül. Hoey szerint a szövegben előforduló kohéziós

elemek legalább negyven százalékban lexikai kohézióra vezethetők vissza. Később ezt az arányt Teich és Fankhauser (2004) korpusznyelvészeti eszközökkel vizsgálva ötven százalékra módosította, bebizonyítva, hogy a szavak ismétlődésmintázata a szemantikai koherencia legfontosabb eleme.

A szövegértékelés folyamatának tanulmányozása volt az empirikus kutatások következő terepe. Ezen kutatások jellemző területe a gyakorlott vizsgáztatók megfigyelése lett (például hangos gondolkodtatással vagy retrospekcióval, tehát utólagos felidézéssel). Ezekben a tanulmányokban a kutatók arra keresték a választ, hogy a vizsgáztatók az esszé mely jellegzetességeit figyelik; ezeknek egyenlő fontosságot tulajdonítanak-e, vagy egyes jellemzőket súlyoznak. Sokat idézett 1979-es kísérletében Freedman előre meghatározott szempontok alapján szándékosan rontott vagy javított bizonyos esszék minőségén a tartalom, szerkesztés, mondatszerkezet vagy helyesírás változtatásával, majd így adta a szövegeket a vizsgáztatóknak értékelésre. Tanulmányában arra a megállapításra jutott, hogy a tartalom és megformáltság minősége az a két kritérium, amelyet minden másnál erőteljesebben vettek figyelembe a vizsgáztatók. Crossley és McNamara (2010) később pontosította ezt a megfigyelést, azzal, hogy a szövegkoherenciára adott ítélet 60 százalékban előre jelzi az esszé végső értékelését. Ez az eredmény tehát egy más szemszögből: a javítói magatartás alapján is újra megerősíti Hoey szövegvizsgálaton alapuló lexikai kohéziós modelljének érvényességét.

A szakirodalom egyetért abban, hogy a meglehetősen bonyolult folyamat során több mint száz különféle szövegjellemzőt figyelnek a vizsgáztatók. A TOEFL esszéket értékelők döntéseit vizsgálva Cumming, Kantor és Powers (2001) analitikus keretbe rendezte a javító tanárok magatartását. Ez alapján kétféle stratégia különböztethető meg: az interpretáló stratégiák és az értékelő stratégiák. Mindkét stratégiacsoport három-három részre bontható tovább az alapján, hogy mire irányul az értékelők figyelme: a szöveg retorikai és tartalmi jellemzőire; a szöveg nyelvi megformáltságára, illetve saját értékelésük folyamatára. Ezen megfigyelések legtöbbje beépült az automatizált esszé értékelő programokba. Az emberi és automatizált értékelés folyamatát részletesen, ábrákkal Deane tanulmánya elemzi (2013).

A szövegen belüli mondatok eltérő funkciójának megfelelő vizsgálatához ugyanakkor nemcsak a szöveg egészét kell vizsgálni: elengedhetetlen, hogy a program meg tudja különböztetni a diskurzus egységeit is. Ennek elősegítésére az egyik automatizált esszéértékelő program kialakításakor szükség volt arra, hogy létrehozzanak egy esszé korpuszt, melyben több száz szöveget manuálisan annotáltak (bevezető szöveg, tézismondat, érvek, példák, konklúzió, irreleváns mondatok stb. megjelölésével), majd ez képessé tette a programot az esszé diskurzuselemeinek automatizált megkülönböztetésére (Burstein, Marcu, Knight 2003).

Az empirikus kutatások másik vonulata a már meglévő esszéértékelő programokat használva von le következtetéseket a szövegek minőségéről. Két különböző területet emelek ki a tanulmányok köréből, amelyek a Coh-Metrix[ii] elnevezésű programot használják szövegelemzésre. Az első terület az írások eltérő szövegminőségre utaló lexikai jellemzőit vizsgálja. A program ki tudta mutatni a nyelvtanulók lexikai fejlődését, valamint az anyanyelvi, illetve nem anyanyelvi írók szövegeinek lexikai eltéréseit. Az aktív lexikon fejlődését egy éves longitudinális vizsgálatában Crossley, Salsbury, McNamara (2009) objektív módon tudta bemutatni. A program két fontos írásjellegzetességben talált jelentős változást. A tanulás előrehaladtával a lexikai sokszínűség (lexical diversity) megnövekedett, és a tanulók változatosabban használtak nem- és fajtafogalmakat kifejező főneveket (hiponima – hiperonima relációkat).

A második fő kutatási irány a szövegek szintaktikai összetettségét vizsgálja. Crossley és McNamara (2014) tanulmánya a Coh-Metrix programmal a nem anyanyelvi írók mondatszerkezeteinek variációit, valamint annak fejlődését figyelte egy év során. Bebizonyosodott, hogy a szintaktikai összetettség is a szövegminőség fontos eleme. A kutatók megjegyzik azonban, hogy a szintaktikai eltéréseket az általános témájú szövegekben mérték. Tudományos angol (English for Academic Purposes) kontextusban, a regiszter más szintaktikai jellemzői miatt, új mérésekre van szükség. Egy másik kutatás szintén a mondatszerkezetek szintjén a tömörítés feladat során a szöveg minőségét vizsgálta gyakorlott és egyetemet kezdő hallgatók esetében (Brandon, Crossley, McNamara 2012). A gyakorlott írók nemcsak bonyolultabb mondatszerkezeteket használtak, hanem kevésbé gyakori szavakat is, és ez a két jellemző együtt járult hozzá a szövegek pozitív megítéléséhez.

3. Az automatizált esszéértékelő programok funkciója és működése

Az automatizált esszéértékelő programok gyakorlati haszna elsősorban a nagy mennyiségű szöveg gyors feldolgozásában rejlik. A másik fontos tényező, hogy a program nincs kitéve olyan biológiai és pszichológiai hatásoknak mint a kifáradás, vagy a Halo-effektus. Ezáltal hosszú ideig képes objektív módon, konzekvensen értékelni. A fő célja jelenleg még nem az emberi szerepvállalásnak az értékelésből való teljes kiküszöbölése, hanem inkább az értékelés objektívabbá tétele a folyamat megbízhatóságának növelésével. A legtöbb automatizált esszéértékelő program esetében minden esszétémához megfelelő mennyiségű – általában több száz – modell esszére van szükség, amelyeket gyakorlott vizsgáztatók pontoztak. Ezek statisztikai elemzésével lehet megállapítani azokat a jegyeket, amelyek képessé teszik a programot a emberi értékelők pontjainak előre jelzésére.

Az automatizált szövegértékelő programok leggyakoribb felhasználási lehetőségei a következők, példákkal: 1) Az online használható Criterion[iii] diagnosztikus szövegelemző program és a Pearson Test of English esetében elsődleges értékelő forrásként használják. 2) A TOEFL, a GMAT[iv] és GRE[v] vizsgák esetében egy emberi értékelő pontozása mellett második értékelőként funkcionál (Deane, 2013) Formatív értékelést nyújt online feltöltött szöveg különböző szempontú elemzésére (például a Coh-Metrix program). 4) A felsőoktatásban tananyagok pontos elsajátítását értékeli tartalmi elemzéssel (Intelligent Essay Assessor).

A gyakorlott vizsgáztatók megfigyelése egyértelműen hatott a számítógépes esszéértékelés programjainak kialakítására. A emberi vizsgáztatók a szövegeket kognitív módon dolgozzák fel, felhasználva ehhez a nyelvről, a szövegszerűségről és az esszéről mint társadalmi konstruktumról meglévő ismereteiket, számításba véve az adott szöveg stilisztikai jellemzőit és kreatív elemeit is. Ezzel ellentétben az automatizált esszéértékelő programok vagy előre definiált jellegzetességeket keresnek, majd ezt követően megtanulják a modell esszék mintázatait, vagy egyszerre történik a jellemzők és azok mintázatának kivonatolása és megtanulása (Zhang 2013, idézi Shin 2019). Más szavakkal, újabban a gépi tanulás két módon megy végbe az automatizált esszé értékelő rendszereknél: manuális beavatkozással, illetve mély tanulásos (deep learning) algoritmussal (Shin 2019).

Az alábbiakban három szövegértékelő program működését fogom bemutatni, azokra az elemekre koncentrálva, amelyek nyelvpedagógiai szempontból relevánsak lehetnek. A programok kiválasztásában fő szerepet játszott, hogy egyrészt széles körben régóta használt alkalmazások legyenek, másrészt részletes leírás álljon rendelkezésre működésükről és az azzal kapcsolatos tudományos eredményekről. Mint látni fogjuk, a három program különböző módon állapítja meg a szöveg minőségét, és különböző funkciókra lett kidolgozva.

E-Rater[vi]

Az E-Rater programot 1999 óta használják esszék holisztikus értékelésére. A program 2-es verziója a következő szövegjellegzetességeket figyeli: nyelvtan, nyelvhasználat, helyesírás, stílus, szerkezet, témafejlődés, lexikai komplexitás, és témaspecifikus szóhasználat (Attali, Burnstein 2006). Minden terület tovább bontható ún. mikrojellemzőkre. A nyelvhasználaton belül például ilyen mikrojellemzők a rossz névelőhasználat, nem megfelelő szavak, helytelen szóalakok vagy rossz középfokú szerkezet (Ben-Simon, Bennett 2007; Quinlan és társai 2009). A nyelvtan, nyelvhasználat és helyesírás esetében a jó és hibás alakok aránya alapján pontoz a program, idiomatikus szerkezetek használatakor pedig pozitívan értékel. A Criterion online szövegértékelő program az E-Rater programon alapulva nyúlt diagnosztikus visszajelzést az írás minőségéről.

Intelligent Essay Assessor[vii]

Az Intelligent Essay Assessor (Landauer, Foltz, Laham 1998) elnevezésű programcsomag fő funkciója, hogy kikérdezi a megtanult tananyagot és értékeli a választ. Fő megfigyelési területe a tananyag minél pontosabb reprodukciója, ezért a szöveg "minősége" alatt főként annak tartalmát érti, bár az érvelést, szerkezetet, helyesírást, jó mondatszerkesztést, szóhasználatot és a stílust is figyeli. A Latent Semantic Analysis (LSA), magyarul rejtett szemantikai analízis módszerére épül, de tartalmaz Natural Language Processing (NLP) jellemzőket is. A szövegben megjelenő szavakra és szókapcsolatokra fókuszál. Statisztikai módszerekkel összegzi a szavak közötti relációkat. Minél jobban hasonlít a szöveg lexikailag a betáplált tankönyv szövegéhez, illetve a modell esszék szövegéhez, annál relevánsabbnak értékeli a tartalmat, és annál magasabb pontot ér el a vizsgázó. A fentiekből következik, hogy az IEA mást ért az esszé minőségén, mint az E-Rater. Kevesebb területet ölel fel, inkább a tartalomra, és sokkal kevésbé a megformáltságra fókuszál.

Coh-Metrix

A Coh-Metrix (Graesser és társai 2004) több mint 200 írásbeli változó, például a helyi és globális kohézió számszerű és szöveges elemzését biztosítja. Az egymást követő mondatokat, illetve minden mondat egymáshoz viszonyított jellemzőit figyeli. A modulok képesek a lexikon, a szintaxis, valamint a szöveg mondatok feletti szintű tulajdonságainak értékelésére. A beépített olvashatósági képletek a szövegnehézség fokát is meg tudják határozni. LSA módszert is használ az értékeléskor. Visszajelzést képes adni arról, mennyire tartalmaz a szöveg narratív vagy kifejtő esszé jellemzőket.

4. Az automatizált esszéértékelő programok érvényessége és megbízhatósága

A számítógépes és a emberi értékelés összehasonlításának megbízhatósági indexe az egyik leginkább kutatott és legkevésbé egyértelmű téma az automatizált esszéértékelésen belül. Haswell és társai (2012) összefoglalója alapján létezik olyan tanulmány, amely szerint az ember és program közötti megbízhatósági index 0.90 felett van, tehát a gépi értékelés nagyon jónak tekinthető. Azonban a kutatások egy részénél felmerül az elfogultság gyanúja, mivel a publikációkat maguk a programot készítő cégek teszik közzé (például az ETS, Educational Testing Services). Sokkal több független kutatásra van tehát szükség a megbízhatóság kérdésében. A kutatók maguk is megállapítják, hogy a statisztika alapú értékelési modellek esetében problémát jelent az értékelés belső struktúrájának leírása és

magyarázata a felhasználók felé. Ez a kommunikációs nehézség az automatizált esszéértékelő programok felszíni érvényességét (face validity) veszélyezteti (Attali, Burnstein 2006).

5. Az esszé minőségének automatikus értékelésével kapcsolatos ellentmondásos nézetek

Tanulmányok sora bizonyítja, hogy sokan szkepticizmussal fogadják az automatizált esszéértékelést. Page és Petersen 1995-ös írása humanista, defenzív, és konstruktum-ellenes kritikákra osztja fel az ellenérzéseket. A humanista kritika döntő érve, hogy az esszé társadalmi produktum, melynek célja az emberek közti kommunikáció, így a nem emberi értékelés ezt az alapvető célt hiúsítja meg. Ebből következően a számítógép nem értékelheti az írott szöveget az emberhez hasonló módon. A defenzív álláspont szerint a számítógépes értékelés csak akkor működőképes, ha „jó szándékú" esszéíróval találkozik, vagyis értelmes, vizsgahelyzetnek megfelelő szöveget tud csak elemezni. Több kísérlet született a program megtévesztésére (például Powers és társai 2002), amelyek célja az volt, hogy az automatizált értékelés korlátait felmérje. Szándékosan értelmetlen, vagy valamilyen szempontból eltúlzott jellegzetességekkel felruházott esszéket elemeztettek a programmal, majd megvizsgálták, tudta-e a feladatot emberi értékelőkhöz hasonlóan teljesíteni. A legújabb ezek sorában Farag, Yannakoudakis, Briscoe (2018) kísérlete, akik a nyelvtanilag helyes, de nem koherens egymást követő mondatokat tartalmazó szövegekkel tesztelték a programot. A lokális koherencia nem megfelelő mérését megállapítva egy új, neurális elemző modellt próbálnak kidolgozni. Page és Petersen a harmadik csoportba azokat az érveket sorolta, amelyek szerint maga a konstruktum nem megfelelő, vagyis a program nem azt méri, ami valójában fontos az esszé minőségénél. Attali és Burnstein (2006) szerint viszont manapság a programok már egyre jobban képesek az írásművek belső lényegét megragadni.

6. Az automatizált esszéértékelés jelenlegi korlátai

A hatékony használat feltételezi, hogy megismerjük az automatizált esszéértékelés korlátait. Powers és társai (2002) tanulmányában azokat az eseteket vizsgálta meg, amikor jelentős eltérést tapasztaltak a gépi és emberi esszéértékelés között. A program a legmagasabb pontszámot adta, a vizsgáztató viszont nagyon alacsonyra pontozta azokat az írásokat, amelyek a megadott cím nagy részét bemásolták a szövegbe, vagy amelyek elemzés helyett egyetértettek a megadott érvekkel. Ebből az következik, hogy a program a szöveg egyezőségét preferálja, tehát egy bizonyos ponton ezt a paramétert a fejlesztőknek pontosítaniuk kell. Azokat az esszéket, amelyekben irodalmi/kulturális utalás szerepelt, vagy az érvek közötti átmenet nem volt explicit módon jelölve, a program alacsony pontszámmal büntette. Powers ebből azt a következtetést vonta le, hogy az E-Rater program emberi értékelő nélkül nem alkalmas a feladat ellátására. A tanulmány másik érdekes konklúziójaképpen ráterelődött a figyelem azokra a szövegjellemzőkre, amelyeket gépi eszközökkel nem lehetett mérni, és amelyekre így – valószínűleg nehezen megragadható voltuk miatt – különösképpen fel kell hívni a vizsgázók figyelmét.

7. Következmények a nyelvpedagógia számára

Nem minden tanár fogadta egyértelműen pozitívan az automatizált esszéértékelés megjelenését. Ez bizonyos mértékben annak a következménye, hogy a kutatásokban használt zsargon és módszertan nem vagy alig segíti elő a téma jobb megértését. A nyelvtanárnak ugyanakkor tisztában kell lennie a vizsgakövetelményekkel és a pontozás részleteivel, enélkül nem tudja a tanulókat hatékonyan felkészíteni. Ha tudja tehát, hogy a TOEFL iBT esszét egy program javítja, és ez a javító program lepontozza az irodalmi/kulturális utalásokat, az iróniát

vagy a kiugróan eredeti érvelést, ugyanakkor többletponttal jutalmazza az idiomatikus szerkezeteket, akkor ezekre fel tudja hívni a vizsgára készülő figyelmét. Az automatizált esszé értékelés megértése tehát segíti a vizsgára való felkészítést.

Érdekes helyzet állhat elő azonban, amennyiben a vizsgára felkészítő tanár a tanuló konformista viselkedését erősíti. Ez akkor fordulhat elő, ha „betanítja" az esszékben általában előforduló alaki jellemzőket és a hozzájuk tartozó nyelvi eszközöket, esetleg még bátorítja is a nyelvtanulót, hogy memorizáljon bizonyos fordulatokat (esetleg egész bekezdéseket) a modell esszékből, hogy azok minél jobban hasonlítsanak egy „ideális" szövegre, azt gondolván, hogy a 100 százalékos egyezés hozza majd a legjobb eredményt. Ha a program nem képes ezeket az előre gyártott részleteket kiszűrni az adott írásból, akkor nemcsak az értékelés validitása kérdőjeleződik meg, hanem az írás kommunikációs célja is végképp elvész.

Összefoglalva, az automatizált esszéértékelő programok egyre terjednek, és a nyelvtanároknak a pedagógiai folyamat hatékonyabbá tétele érdekében meg kell ismerkedniük azok módszertani alapjaival. A kutatók és programfejlesztők részéről több és közérthetőbb leírásra van szükség, hogy az információk nagyobb közönséghez is eljussanak és hasznukról társadalmi konszenzus jöhessen létre az összes érdekelt fél bevonásával.

Irodalom

Attali, Y., Burstein, J. 2006. Automated Essay Scoring With e-rater® V.2. *Journal of Technology, Learning, and Assessment, 4*(3). http://www.jtla.org

Ben-Simon, A., Bennett, R. E. 2007. Toward More Substantively Meaningful Automated Essay Scoring. *The Journal of Technology, Learning, and Assessment, 6*(1). 1–47.

Brandon, R. D., Crossley, S. A., McNamara, D. S. 2012. A Linguistic Analysis of Expert-Generated Paraphrases. *Proceedings of the 25th International Florida Artificial Intelligence Research Society Conference, FLAIRS-25,* 268–271.

Burstein, J., Marcu, D., Knight, K. 2003. Finding the WRITE stuff: Automatic Identification of Discourse Structure in Student Essays. *IEEE Intelligent Systems: Special Issue on Natural Language Processing, 18*(1), 32–39.

Chapelle, C.A., Douglas, D. 2006. *Assessing Language through Computer Technology*. Cambridge: Cambridge University Press.

Crossley, S. A., McNamara, D. S. 2010. Cohesion, Coherence and Expert Evaluations of Writing Proficiency. *32nd Annual Conference of the Cognitive Science Society,* 984–989.

Crossley, S. A., McNamara, D. S. 2014. Does Writing Development Equal Writing Quality? A Computational Investigation of Syntactic Complexity in L2 Learners. *Journal of Second Language Writing, 26,* 66–79.

Crossley, S. A., Salsbury, T., McNamara, D. S. 2009. Measuring L2 Lexical Growth Using Hypernymic Relationships. *Language Learning 59,* 307–334.

Cumming, A., Kantor, R., Powers, D. 2001. Scoring TOEFL Essays and TOEFL 2000 Prototype Writing Tasks: An Investigation into Raters' Decision Making and Development of a Preliminary Analytic Framework. *TOEFL Monograph Series*, 1–95.

Deane, P. 2013. On the Relation between Automated Essay Scoring and Modern Views of the Writing Construct. *Assessing Writing, 18* (1), 7–24.

Dikli, S. 2006. An Overview of Automated Scoring of Essays. *Journal of Technology Learning, and Assessment, 5*(1). http://www.jtla.org

Farag, Y., Yannakoudakis, H, Briscoe, T. 2018. *Neural Automated Essay Scoring and Coherence Modeling for Adversarially Crafted Input.* https://www.aclweb.org/anthology/N18–1024

Freedman, S. W. 1979. How Characteristics of Student Essays Influence Teachers' Evaluations. *Journal of Educational Psychology, 71* (3), 328–338.

Graesser, A. C., McNamara, D., Louwerse, M. M., Cai, Z. 2004. Coh-Metrix: Analysis of Text on Cohesion and Language. *Behavior Research Methods, Instruments, and Computers, 36* (2), 193–202.

Halliday, M. A. K., Hasan, R. 1976. *Cohesion in English.* London: Longman.

Haswell, R., Donnelly, W., Hester, V., O'Neill, P., Schendel, E. 2012. An Annotated Bibliography of Writing Assessment: Machine Scoring and Evaluation of Essay-Length Writing. *The Journal of Writing Assessment, 5*(1). http://www.journalofwritingassessment.org/article.php?article=58.

Hoey, M. 1991. *Patterns of Lexis in Text*. Oxford. Oxford University Press.

Landauer, T. K., Foltz, P. W., Laham, D. 1998. Introduction to Latent Semantic Analysis. *Discourse Processes, 25*, 259–284.

Molnár Gy. 2010. Technológia-alapú mérés-értékelés hazai és nemzetközi implementációi. *Iskolakultúra*, 7–8. 22-34.

Page, E. B. & Petersen, N. S. 1995. The computer moves into essay grading: Updating the ancient test. *Phi Delta Kappan, 76*, 561–565.

Powers, D. E., Burstein, J., Chodorow, M., Fowles, M. E., Kukich, K. 2002. Stumping E-Rater: Challenging the Validity of Automated Essay Scoring. *Computers in Human Behavior 18* (2).103–134.

Quinlan, T., Higgins, D., Wolff, S. 2009. Evaluating the Construct-Coverage of the E-Rater Scoring Engine. *ETS Research Report*. Educational Testing Service.

Shin, E. 2019. *A Neural Network approach to Automated Essay Scoring: A Comparison with the Method of Integrating Deep Language Features using Coh-Metrix.* MA Thesis: Univeristy of Alberta.

Teich, E., Fankhauser, P. 2004. WordNet for Lexical Cohesion Analysis. In: Petr Sojka, Karel Pala, Pavel Smrž, Christiane Fellbaum, Piek Vossen (Eds.): *GWC 2004, Proceedings,* Masaryk University, Brno, 2003, 326–331.

Zhang, M. 2013. Contrasting automated and human scoring of essays. *R & D Connections*, *21*(2). 1-11.

[i] TOEFL iBT vizsgaleírás angolul: https://www.ets.org/toefl/

[ii] Coh-Metrix: www.cohmetrix.com

[iii] Criterion: www.criterion.ets.org

[iv] GMAT: lásd www.mba.com

[v] GRE: lásd www.ets.org/gre

[vi] E-Rater: www.ets.org/erater

[vii] Intelligent Essay Assessor: www.assets.pearsonschool.com/asset_mgr/legacy/200842/IEA_FAQ_13801_1.pdf