A szavak véletlenszerű megjelenésén alapuló modellek és az irodalmi művek közötti eltérések magyarázata

Csernoch Mária

Angol-Amerikai Intézet 4010 Debrecen, Egyetem tér 1.
mcsernoch@hotmail.com

Abstract. Munkánkban arra vállalkoztunk, hogy kiderítsük, mi okozza egy természetes nyelvi szövegnek egy dinamikus lexikai statisztikai modelltől való eltérését. A feladat megoldásához egy saját fejlesztésű programot használtunk. A program segítségével meghatároztuk, majd ábrázoltuk a kiválasztott irodalmi művekben az újonnan megjelenő szóalakokat és vizsgáltuk azokat a pontokat, amelyek szignifikáns eltérést mutatnak a modellhez képest. Vizsgáltuk, hogy a szintaktikai szabályok befolyásolják-e, és ha igen mennyiben, a szóalakok megjelenését, illetve származhatnak-e más forrásból az eredeti és a mesterséges szöveg közötti eltérések. Az eredeti művet összehasonlítva a modellel, majd a mű fordításaival azt tapasztaltuk, hogy az eltérések nem szintaktikai, illetve szemantikai, hanem sokkal inkább szöveg szinten jelentek meg és okoztak látványos növekedést az újonnan bevezetésre kerülő szóalakok számában.

Bevezetés

Mindannyiunk által ismert és elfogadott tény, hogy irodalmi művekben a szavak nem egymástól függetlenül követik egymást. Ezzel látszólagos ellentmondásban a korábban megépített akár statikus, akár dinamikus modellek valamilyen szinten mind feltételezték a szavak egymástól független megjelenését az irodalmi művekben. Ez a feltételezés egy nyilvánvaló leegyszerűsítése a problémának, aminek következtében a felhasznált módszertől függően különböző mértékű, esetleg nagyságrendű eltérések tapasztalhatóak az eredeti mű és a modell között, valamint az eredeti mű tulajdonságait leírni szándékozó konstansok és a mért eredmények között (Baayen, 1993, 1996, 2001; Hoover, 2003). Ugyanakkor napjainkra az is elfogadott, hogy a szavak függetlenségét feltételező modellek segítségével a szöveg bizonyos tulajdonságait leírni képes formulákat sikerült találni.

Vizsgálva a szavak nem-véletlenszerűségének forrásait azonban azt tapasztalták (Baayen, 1996, 2001), hogy bár a mondaton belüli kötöttségek a legnyilvánvalóbbak, mégsem ezek a legfőbb forrásai a különböző szóalakok nem-véletlenszerű megjelenésének. Sokkal inkább meghatározónak bizonyult az, hogy milyen szerkezetű a mű.

Ezt bizonyítandó Baayen (Baayen, 1996, 2001) azt a módszert használta, hogy egy olyan mesterséges szöveget állított elő, amelyben véletlenszerűen összekeverte a szöveg mondatait meghagyva azonban a szavak mondaton belüli sorrendjét. Azt tapasztalta, hogy az így előállított mesterséges szöveg és az általa használt modell között

látványosan csökkentek, esetenként eltűntek azok az eltérések, amelyek az eredeti szöveg szóalakjainak száma és a modell által számolt szóalakok között még jelen voltak.

Vizsgálatainkban annak az állításnak a bizonyítására, hogy a véletlen szóhasználattól csak szöveg szinten térnek el az írók egy olyan módszert használtunk, ahol nem volt szükség mesterséges szöveg előállítására, hanem az eredeti művet hasonlítottuk össze az általunk használt dinamikus modellel (Csernoch, 2003; Csernoch és Hunyadi, 2003).

Módszerek

A szövegek elemzéséhez készítettünk egy programot (Csernoch és Hunyadi, 2003; Csernoch, 2004), amely megszámolta és tárolta a szövegben megjelenő szavakat. A tárolt adatok alapján, többek között, a program meghatározta az újonnan megjelenő szóalakok számát, az egyes szóalakok gyakoriságát, az egyszer előforduló szavak számát, stb. További vizsgálatainkhoz megszámoltuk, hogy száz-szövegszó-hosszúságú intervallumokban (blokkokban) hány új szóalak (y_i, $i = 1, \ldots, n$, ahol n a blokkok száma) jelenik meg az előzőekhez képest és az így kapott értékeket ábrázoltuk. A függvény monoton csökkenő tendenciáját megtörő kiugrások (1-3. ábra; pontok) a szövegben jelenlévő trendek és szezonalitások következményei. A trendek jelenlétére utaló kiugrásokat elsődleges, míg a szezonalitások következtében megjelenőket másodlagos kiugrásoknak nevezzük (1. ábra). Szezonalitások alatt értjük azokat az eseményeket, amelyek nem logikus következményei az előzményeknek és ennek következtében bevezetésükhöz kiugróan magas számú új szóalakra van szükség. A program az egyes blokkokban újonnan megjelenő szóalakok számának meghatározása (y_i) után elvégzi a függvény ábrázolását. A grafikonról az esetek többségében jól leolvasható, hogy melyek azok a pontok, ahol ezek a rendkívüli események bekövetkeznek, de a grafikon alapján nehéz megmondani, hogy mely változások tekinthetők szignifikánsnak. További feldolgozásra volt szükség tehát annak eldöntésére, hogy az újonnan megjelenő szavakat leíró görbe mely csúcsai tekinthetők elsődleges, illetve másodlagos kiugrásnak.

Elsőként a mért adatok alapján kapott görbe simítását kellet elvégezni, az így kapott értékek (y'_i) az f' simított görbe függvényértékei. A száz szövegszó hosszúságú blokkok ugyanis kellően rövidek ahhoz, hogy visszaadják a szöveg finomabb változásait is, de éppen e miatt a jelentéktelen változásokra is érzékenyek. Amennyiben a szövegben bekövetkezett változás jelentéktelen, csak abban az egy blokkban érezteti hatását, úgy az a simítás során eltűnik, ugyanakkor a jelentős változások a simítás után is megjelennek a görbén (1/A-3/A. ábra; folytonos vonal).

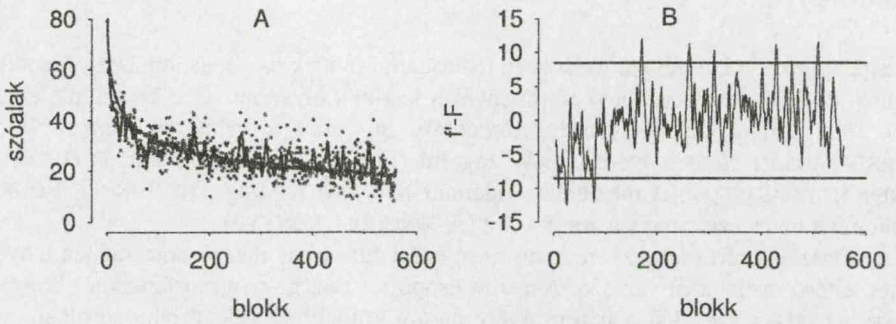

Fig. 2. Kertész Imre SORSTALANSÁG című művének elemzése. Az újonnan megjelenő szóalakok száma, a simított görbe és száz modell átlaga (A). A simított görbe és az átlag függvény közötti eltérés (B).

Ezt a simított görbét hasonlítottuk a modell által előállított mesterséges szöveg szóalakjait leíró görbék sorozatához (fp_k, k = 1, ..., 100), ahol yp_{ki} jelöli a k. függvény i. blokkjában megjelenő szóalakok számát. A modell alapján előállítottunk száz mesterséges szöveget, megszámoltuk ezen szövegekben az újonnan megjelenő szavak számát a száz szövegszó hosszúságú blokkokban és vettük az így kapott függvények átlagát (F) (1/A-3/A. ábra) (Ashby, 1972).

$$F = \frac{\sum_{k=1}^{100} yp_{ki}}{100} \tag{1}$$

A következő lépésben vettük a simított függvény és az átlag függvény különbségét (Δf, melynek függvényértékei Δy_i)

$$\Delta f = f' - F, \tag{2}$$

$$\Delta y_i = f'_i - F_i,\ i = 1, \ldots, n, \tag{3}$$

majd a különbségek átlagát (M) és szórását (σ) (Hajtman, 1971; Nemetz és Kusolitsch, 1999; Solt, 1971; Yule, 1950)

$$M = \frac{\sum_{i=1}^{n} \Delta y_i}{n},\ \sigma = \sqrt{\frac{\sum_{i=1}^{n} (\Delta y_i - M)^2}{n}}. \tag{4}$$

Azokat az eltéréseket tekintettük szignifikánsnak, amelyek az átlagtól legalább 2σ-val eltérnek. A 1/B-3/B ábrák mindegyikén tisztán kivehetőek a Δf görbének azok a pontjai, amelyek az $M \pm 2\sigma$ tartományon kívül esnek. Arra voltunk kíváncsiak, hogy milyen események következtek be a szövegben, amelyek ezeket a kiugrásokat eredményezték a görbéken.

Eredmények

Vizsgálatainkban különböző nyelveken írt irodalmi művek összehasonlítását végeztük. Ahhoz, hogy összehasonlítható eredményeket kapjunk olyan műveket kerestünk, amelyek több különböző nyelven is elérhetőek. Így esett a választás Kertész Imre SORSTALANSÁG című művére, amely angolul (FATELESS) is és németül is (ROMAN EINES SCHICKSALLOSEN) megjelent, valamint Rudyard Kipling THE JUNGLE BOOKS című műveire és ezek magyar fordítására (A DZSUNGEL KÖNYVE).

A választás azért esett ezekre a művekre és fordításaikra, mert szerkezetében lényegesen eltérő nyelvekről van szó. Aszerint csoportosítva, hogy a morfémákból hogyan képzi a nyelv a szavakat a három nyelv három különböző kategóriába sorolható. A német a flektáló, a magyar az agglutináló nyelvek csoportjába tartozik, míg az angol több különböző kategória eszközeit is felhasználja, így igazán egyikbe sem illik bele (O'Grady, 1993; É. Kiss, 1998; Kiefer, 1998; Kugler, 2000; Laczkó, 2000; Quirk et al., 1995; Uzonyi, 1996) (1. táblázat). A kérdés az volt, hogy a mondatok belső kohéziója, tehát a szintaktikai szabályok befolyásolják-e, s ha igen mennyiben az új szóalakok megjelenését, illetve származhatnak-e más forrásokból az eredeti és a mesterséges szöveg közötti eltérések.

Table 1. Kertész Imre SORSTALANSÁG, a mű angol és német nyelvű fordítása, Rudyard Kipling THE JUNGLE BOOKS és magyar fordítása. A második oszlopban a szöveg hosszát (szövegszó = 100 * blokk), harmadik oszlopban a különböző szóalakok, negyedik oszlopban pedig az egyszer előforduló szavak számát tüntettük fel az egyes művekben. A kapott értékeket összehasonlítva látható, hogy az angol szövegekben fordul elő a legkevesebb különböző szóalak és ezzel párhuzamosan a legkevesebb hapax legomena.

	blokk	**szóalak**	**hapax legomena**
Sorstalanság	561	14740	10253
Fateless	716	6710	3186
Roman eines Schicksallosen	719	9992	6043
The Jungle Books	1171	7452	3124
A dzsungel könyve	922	20362	13372

Az 1. táblázat értékei mutatják, hogy az egyes nyelvek sajátosságaiból, valamint a fordításból adódóan a szövegszók, a különböző szóalakok és az egyszer előforduló szavak száma között lényeges eltérések mutatkoznak az egymásnak megfelelő szövegek esetén. Ha azonban elemezzük a grafikonok kiugrásait (1-3. ábra), megkereshetjük az eredeti szövegben azokat a szakaszokat, amelyekben az újonnan megjelenő szavak lényegesen magasabbak, mint az a modell alapján várható lenne. A kérdés az volt, hogy mivel magyarázhatóak ezek a kiugrások, tehát a mi indokolja a különböző szóalakok szokatlanul magas számát és találunk-e olyan jellemzőjét a szövegnek, amelylyel leírhatóak ezek a hirtelen változások.

A kiugrások pontos helyének, a blokkok sorszámának meghatározását MS Excellel, azoknak az n*100 szövegszó hosszúságú szövegrészeknek a meghatározását, amelyekben ezek a kiugrások megjelentek pedig a szövegfeldolgozásra használt saját programmal végeztük. A szövegrészt ismerve vissza tudtuk azt keresni az eredeti

műben és magyarázatot tudtunk adni arra, hogy miért növekedett meg hirtelen az újonnan bevezetett szavak száma.

Table 2. Azoknak a blokkoknak a sorszáma, amelyekben az újonnan bevezetett szóalakok száma magasabb, mint az a modell alapján várható volt.

	Sorstalanság	Roman eines Schicksallosen	Fateless
Vili bácsi			43
csepeli üzem			71
indulás a vonattal		209	156
Auschwitzba érkezés	170		215
Buchenwaldba érkezés	262	337	329
reggeli készülődés, üzem	310	398	392
testének leírása			447
lelkiállapotának leírása		518	
kórház	429	552	
Pjetyka főz	459	587	
napi menetrend			618
haza indulás	510	651	

A SORSTALANSÁGban hat szignifikánsnak tekinthető eltérést találtunk és néztünk meg részletesen. Ezek a kiugrások valamennyien olyan esetben jelentek meg, amikor a szöveghez nem szervesen kapcsolódó, a korábbi eseményektől függetleníthető leírás jelent meg a szövegben. Ez a hat esemény a megjelenés sorrendjében a következő volt: megérkezés a koncentrációs táborba, megérkezés a második táborba, reggeli események és az üzem leírása, kórház leírása, Pjetyka főz, haza indulás (1. ábra, 2. táblázat).

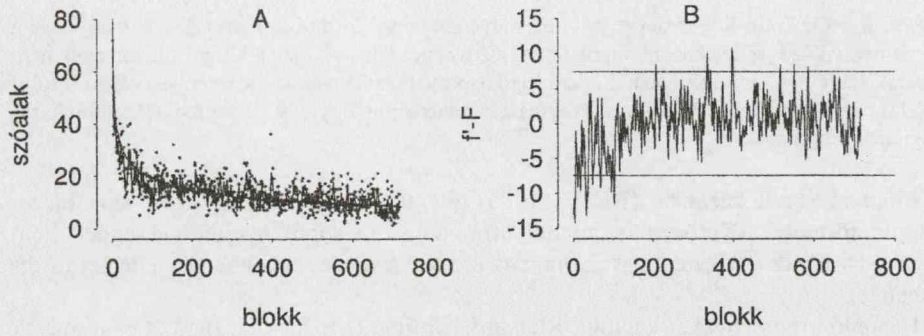

Fig. 3. Kertész Imre SORSTALANSÁG című művének német fordítása: ROMAN EINES SCHICKSALLOSEN. A német és a magyar nyelvű szövegben apró eltérésektől eltekintve a szövegnek ugyanazon a pontján emelkedett meg az újonnan bevezetett szóalakok száma.

A német nyelvű szövegben hét kiugrás található (2. táblázat, 2. ábra), amelyek közül az első nem a táborba érkezést, hanem egy korábbi eseményt, a vonatra szállást írja le. Várhatóan azért nem kaptunk a német szövegben újabb kiugrást a táborba érkezéskor, mert a vonatra szállás, a vonat leírására használt szavak nagyban fedik a

tábor jellemzésére használt szavakat. A második és a harmadik kiugrás ugyanannál a szövegrésznél következett be, mint a magyar szövegben. A német szövegben akkor jelenik meg a negyedik kiugrás, amikor egy leírás következik a főszereplő pillanatnyi lelkiállapotáról. Ez a leírás a magyar szövegben nem eredményezett szignifikáns eltérést. Végül az utolsó három kiugrás újra teljes egészében megegyezik a magyar szöveg kiugrásaival. (A német szöveg utolsó kiugrása még éppen az elfogadhatósági intervallumon belül esik, de ez várhatóan annak tudható be, hogy a digitalizálás során egy ének a magyar szövegben szótagolva került be, míg a német szövegben egybe írva.)

Az angol szöveg elemzésekor is hasonló eredményeket kaptunk (2. táblázat, 3. ábra). Olyan helyeken jelentkeztek a görbén kiugrások, ahol a műbe egy hosszabb lélegzetű leírás került. Ezek nagy része most is megegyezett a magyar (német) szöveg kiugrásaival, annyiban történt változás, hogy az angol szövegben összesen nyolc csúcs tekinthető lényeges eltérésnek a szöveg megszokott menetéhez képest. A magyar és a német nyelvű szöveghez képest megjelent a szöveg elején két kiugrás, amely további részletes leírást ad. A középső négy kiugrás megegyezik a másik két szöveg kiugrásaival, míg a két utolsó olyan leírás, amely csak az angol szövegben okozott szignifikáns eltérést.

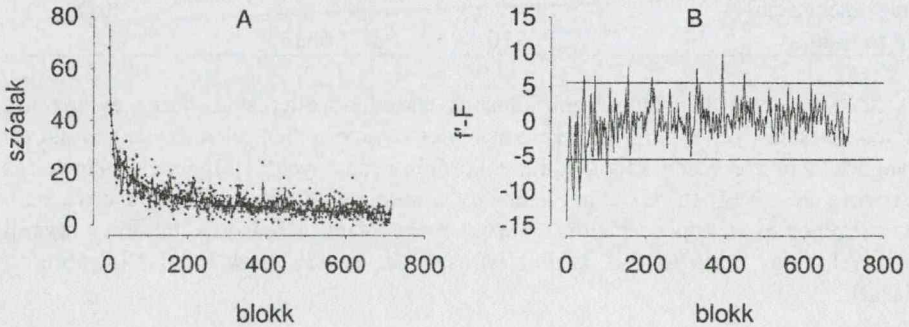

Fig. 4. Kertész Imre SORSTALANSÁG című művének angol fordítása: FATELESS. A magyar és a német nyelvű szöveghez hasonlóan olyan eseményeknél jelentek meg a kiugrások, amelyek nem képezik szerves részét a szövegnek, nem logikus következményei az előzményeknek, és a folytatáshoz sem kötődnek. Ezek a szövegrészek rendszerint egy-egy hosszabb lélegzetű leírás megjelenését jelentik.

Előzetes várakozásokkal (Balázs, 1985) ellentétben ezek a kiugrások nem fejezet határon történtek. Különös tekintettel arra, hogy az angol nyelvű szövegben nem ugyanott vannak a fejezet határok, mint az eredeti magyar szövegben és a német fordításban.

Hasonló eredményeket kaptunk Rudyard Kipling THE JUNGLE BOOKS és a művek magyar fordításának elemzésénél. Nem feltétlenül az újabb mese kezdetekor növekedett meg az újonnan bevezetett szóalakok száma, hanem sokkal inkább akkor, amikor egy hosszabb lélegzetű leírás jelent meg a műben. Ennek megfelelően egyes nem a dzsungelben játszódó történetben (The White Seal, Rikki-Tikki-Tavi, Toomai of the Elephants, The Miracle of Purun Bhagat, Quiquern), mivel színhelyük és témájuk rendkívül változatos a kiugrások egy-egy részletes leírás eredményei. A dzsungelről szóló történetekben is találtunk két lényeges kiugrást, de egyiket sem az adott mese

kezdeténél, hanem egyszer a királyi palota, míg a másik alkalommal a kincstár leírása okozta a szóalakok számának hirtelen emelkedését.

Az említett kiugrások tehát akkor következnek be, amikor a soron következő mondatok sem az előzményekhez nem kötődnek, sem a későbbiekhez való szervest kapcsolódást nem készítik elő. Olyan szövegrészek, amelyekhez nem találni olyan témát, amelyhez a bennük foglaltak kapcsolódnának. Az 1-3. ábrákon jól látható kiugrásokon túl ugyanezt támasztja alá az egyszer előforduló szavak vizsgálata is (4. ábra). Ugyanazokon a helyeken növekedett meg az egyszer előforduló szavak száma, ahol az eredeti műben szintén magas volt az újonnan bevezetett szavak száma. Ez a megfigyelés is arra enged következtetni, hogy a görbéken található kiugrások a szöveghez szervesen nem kapcsolódó részeknél jelennek meg.

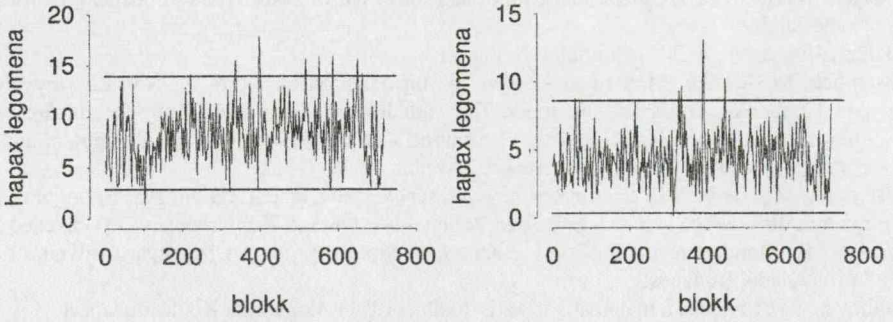

Fig. 5. Az egyszer előforduló szavak megjelenése ROMAN EINES SCHICKSALLOSEN (balra) FATELESS (jobbra) művekben. Az ábrán az átlag ±2*szórás jelző vonalakat a hapax legomena binomiális eloszlását feltételezve húztuk meg. A kiugrások azokon a helyeken jelentek meg, ahol az eredeti szövegben megnövekedett az újonnan bevezetett szóalakok száma.

Összegzés

Irodalmi művek statisztikai elemzésénél arra kerestük a választ, hogy mi okozhatja az eredeti mű és a szavak véletlenszerű megjelenését feltételező modellek közötti eltérést. A modelleknél azt tapasztaltuk, hogy eleinte felül becsülik az eredeti művet, míg vannak olyan helyek, amelyeken a modell által generált mesterséges szövegben kevesebb szóalak jelenik meg, mint az eredeti műben. Előzetes várakozásaink szerint ezek az eltérések egyrészt a szintaktikai szabályok következetes használatából adódhatnak, esetleg új fejezetek kezdetén növekedhet meg a szóalakok száma. Ezek a változások valóban megjelennek az újonnan bevezetett szóalakokat ábrázoló görbén, de csak kisebb, elsődleges kiugrásokat eredményeznek, ami nem nagyobb mint a zaj a görbén. Megválaszolatlan maradt azonban az a kérdés, hogy a másodlagos kiugrásokat mi okozza. Azt tapasztaltuk, hogy a másodlagos kiugrások szöveg szinten jelennek meg, akkor amikor a szöveg olyan leírásokat tartalmaz, amely sem az előzményekhez, sem a következő részekhez nem kapcsolódnak logikusan. Kétféle módszert is használtunk ennek igazolására. Első lépésként elemeztük a művek fordításait is, így rá tudtuk mutatni, hogy a mondaton belüli kohézió nem okozhatja ezeket a másodlagos kiugráso-

kat. Ezt követően megvizsgáltuk az egyszer előforduló szavak eloszlását, és az előzőhöz hasonló módon azt tapasztaltuk, hogy azokon a helyeken magas ezen szavaknak a száma, ahol a szöveghez alig kapcsolódó szövegrész jelenik meg.

Irodalom

Ashby, W. R. Bevezetés a kibernetikába (1972) Akadémiai Kiadó, Budapest

Baayen R. H.: Statistical Models for Word Frequency Distributions: A Linguistic Evaluation. Computers and the Humanities 26. (1993) 347-363.

Baayen R. H.: The Effect of Lexical Specialization on the Growth Curve of the Vocabulary. Computational Linguistics 22. (1996) 455-480.

Baayen, R. H. Word Frequency Distributions (2001) Kluwer Academic Publishers, Dordrecht, Netherlands

Balázs, J. A szöveg (1985) Gondolat, Budapest

Csernoch, M. Another Method to Analyze the Introduction of Word-Types in Literary Works and Textbooks, Conference Abstract, The 16th Joint International Conference of the Association for Literary and Linguistic Computing and the Association for Computers and the Humanities (2004) Göteborg University, Sweden

Csernoch, M; Hunyadi L. Szótípusok bevezetésének szabályszerűsége magyar és angol nyelvű nyomtatott szövegekben, Magyar Számítógépes Nyelvészeti Konferencia (2003) Szeged

É. Kiss, K. Mondattan. In É. Kiss, K., Kiefer, F. Siptár, P. (eds), Új magyar nyelvtan (1998) Osiris Kiadó, Budapest

Hajtman, B. Bevezetés a matematikai statisztikába (1971) Akadémiai Kiadó Budapest

Hoover D. L.: Another Perspective on Vocabulary Richness. Computers and the Humanities 37 (2003) 151-178.

Kiefer, F. Alaktan. In É. Kiss, K., Kiefer, F. Siptár, P. (eds), Új magyar nyelvtan (1998) Osiris Kiadó, Budapest

Kugler, N. Alaktan. In Balogh, J., Haader, L., Keszler, B., Kugler, N., Laczkó, K. Lengyel, K. (eds), Magyar grammatika (2000) Nemzeti Tankönyvkiadó, Budapest

Laczkó, K. Alaktan. In Balogh, J., Haader, L., Keszler, B., Kugler, N., Laczkó, K. Lengyel, K. (eds), Magyar grammatika (2000) Nemzeti Tankönyvkiadó, Budapest

Nemetz, T.; Kusolitsch, N. Guide to the empire of random (1999) TypoTEX, Budapest

O'Grady, W., Dobrovolsky, M. and Aronoff, M. Contemporary Linguistics, An Introduction (1993) New York: St. Martin's Press.

Quirk, R.; Greenbaum, S.; Leech, G.; Svartvik, J. A Comprehensive Grammar of the English Language (1995) Longman Group UK Limited, London and New York

Solt, Gy. Valószínűségszámítás (1971) Műszaki Könyvkiadó, Budapest, Hungary

Uzonyi, P. Rendszeres német nyelvtan (1996) AULA Kiadó Budapest, Hungary

Yule, G. U. An Introduction to the Theory of Statistics (1950) Charles Griffin & Company Limited, London, UK