huzamba állító regény lépcsőfoknak tekinthető afelé a teljes elfogadás felé, amely például Rusvai Mónikának a *Lavondyss* című regényről adott elemzésében tűnik fel: „ha a *más*hoz közelebb kerülünk, valójában mindig önmagunk egy addig ismeretlen részét látjuk meg benne". (Kiemelés az eredetiben, 276)

Bölcsészettudományi hátterű tanulmánykötettől szokatlan módon mintha a hétköznapokra alkalmazható üzenettel, tanulságokkal bírna a gyűjtemény, és sokszor az egyes szövegek is: hamarosan muszáj lesz megbarátkoznunk a mesterséges intelligenciával (33); a másság megismerése önismerethez vezet (276); szörnyek lehetnek „az én határain belül" is (296); illetve általában is olvasható a kötet a másság elfogadására való buzdításként. Az írások témáinak választott alkotások között akadnak itthon kevéssé elemzettek, de az ismertebb és többet kutatott művek értelmezései is szolgálnak újdonságokkal filológiai alapossággal (Panka Dániel) vagy újszerű kontextualizációjuk (Kérchy Anna) révén. A szerzők nagyrészt angol–amerikai intézményes hátterének előnye az e területen megszokott szerkezeti tisztaság, az egyértelmű célok kijelölése, a szövegek könnyű érthetősége — ezek pedig sokszorosan fontosak egy nem csak akadémiai olvasóközönségre számító antológia esetében. A szerkesztési–korrektúrázási hibák sajnos a kötet vége felé exponenciálisan szaporodnak, mintha a szerkesztő idő közben elfáradt volna, illetve az illusztrációként használt képek logikája sem mindig érthető — bizonyos grafikák teljesen fölöslegesnek tűnnek, míg a vizuális alkotásokat (filmet, képregényt, sorozatot) elemző tanulmányokhoz sosem tartozik kép. A szépirodalmi kánonból még mindig kiszoruló műfajoknak és írásmódoknak a *Rémesen népszerű*ben megjelenő értő és szakmai olvasatai azonban fontosabbak annál, hogy ilyesfajta hiányosságok az olvasó kedvét szegjék.

SZEMES Botond

A digitális bölcsészet mint kódfejtés

*A nyelv statisztikai vizsgálatának hagyományáról**

Mikor Claude E. Shannon 1948-as korszakalkotó cikkében az információ mérhetőségének és továbbíthatóságának problémáját eloldotta az üzenetek jelentésének vizsgálatától,[1] nemcsak a digitális hírközlés elméleti alapjait fektette le, hanem egyben a nyelvstatisztikai elemzés kultúrában betöltött szerepének évszázados hagyományát újította meg. Ez a lappangó hagyomány ugyanis a könyvnyomtatás, a gyors- és távírás, a kódfejtés és a nyelvtanítás területein már mindig is a kulturális működés alapjaira vonatkozott: a szövegek létrehozása, megfejtése és az azokat felépítő kódrendszer elsajátítása mind statisztikai műveletek mentén alakultak az európai modernségben. Ezek a műveletek mégsem képezték a kulturális önértésünk részét — egészen a legújabb időkig, amikor a digitális eszközök elterjedésével maguk a művészeti–kulturális folyamatok leírása vált lehetségessé statisztikai számításokra alapozva. Az alábbiakban ezeket, a kódfejtéstől a digitális bölcsészetig húzódó történeti összefüggéseket kívánom röviden felvázolni. Shannon elméletének kiindulópontja, hogy a kommunikáció során küldött jelek információértéke azok statisztikai valószínűségük alapján meghatározható: minél váratlanabb, azaz valószínűtlenebb egy jel előfordulása, annál nagyobb az információértéke, vagy Shannon terminusával: entrópiája.[2] Könnyen belátható, hogy ha biztosan mindig ugyanazt a jelet küldjük (minimális entrópia és maximális redundancia), akkor nincs is szükség kommunikációra, hiszen a címzett számára a jel nem hordoz semmiféle információt, mivel előre tudhatta, hogy az fog megérkezni. Az információérték maximumáról ezzel szemben akkor beszélhetünk, ha egy jelrendszerből a különböző jelek egyenlő valószínűséggel kerülhetnek kiválasztásra, azaz ha a legnagyobb fokú a bizonytalanság abban, hogy milyen jel érkezik a feladótól. Ahhoz, hogy ezt a bizonytalanságot — az üzenet/jel információértékét — meg tudjuk adni, ismernünk kell tehát a jelek előfordulásának gyakoriságát, hiszen csak ezen keresztül következtethetünk előfordulásuk valószínűségére. Ennek a statisztikai tudásnak nem csak elméleti, hanem fontos gyakorlati következményei is vannak,[3] ahogyan azt a különböző mesterségek már jóval az információelmélet megszületése előtt is tudták. Shannon például a távírásra hivatkozik, ahol a betűk előfordulásának gyakorisága szabja meg a hozzájuk rendelt jelek bonyolultságát; hiszen a gyakran használt betűkkel érdemes egyszerűbb jelekkel kódolni:

> Ezt bizonyos mértékig meg is valósították a Morse-távírónál, ahol a leggyakrabban előforduló angol betű az E csatornaszimbólumát egy pont jelzi, míg a kevésbé gyakori betűket, pl. a Q, X, Z-t pontok és vonalak hosszabb sorozata jelképezi. Ezt az elvet bizonyos kereskedelmi kódoknál még tovább fejlesztették és itt gyakori szavakat és kifejezéseket 4–5 betűs kódcsoportokkal jelölnek, ezáltal jelentősen lerövidítve az átlagos átviteli időtartamot. A manapság szabványosított üdvözlő és évfordulói táviratoknál ezt az elvet odáig fejlesztették, hogy egy vagy két mondatot viszonylag rövid számsorban kódolva visznek át.[4]

Ez a nyelvstatisztikai tudás nem csak az üdvözlőlapok hatékony táviratozását teszi lehetővé. Ugyanezen a felismerésen alapul a gyorsírás technikájának vagy a titkosított szövegek megfejtésének kora újkori gyakorlata is. Az előbbi esetben a leggyakrabban használt betűket és betűkombinációkat kell a legegyszerűbben és leggyorsabban leírható jelekkel helyettesíteni az írásfolyamat gyorsításának érdekében, ahogyan azt Timothy Brigth, a modern gyorsírás megalapítója már 1588-as, *Charactery: An Art of Short, Swift, and Secret Writing by character* című könyvében is kifejtette.[5] A kódfejtés egy ennél összetettebb folyamatot jelöl. Szövegek titkosításának az ókorban kialakult művelete szerint, ha minden betűt egy másik, az ABC-ben *x* távolsággal követő betűre cserélünk, akkor egy olyan értelmetlennek tűnő szöveget hozhatunk létre, amelyből rekonstruálható az eredeti üzenet, amennyiben ismerjük a betűk eltolásának (*x*) mértékét (azaz a titkosítás „kulcsát"). Feltörhető azonban egy ilyen titkosított szöveg a kulcs előzetes ismerete nélkül is; elég ehhez csupán az adott nyelvre vonatkozó betűk gyakoriságát ismerni. Hiszen a titkosított szöveg leggyakoribb betűi a nyelvben előforduló leggyakoribb betűket fogják helyettesíteni, amely megfeleltetés alapján már könnyedén meghatározható az eltolás mértéke is. Ezek a nyelvstatisztikai ismeretek a titkosított üzenetek feltörésén túl az adott nyelv szerveződésébe is bepillantást nyújthatnak: „Azok, akik titkosírással foglalkoznak, jól tudják, hogy a »w« előfordulása egy (a betűk felcserélése nélküli vagy e felcseréléstől már megtisztított szövegben) rejtjelezett francia nyelvű üzenetben csaknem biztosan egy idegen szó jelenlétét jelzi"[6] — jegyzi meg például Abraham Moles is az információelmélet és az esztétikai élmény összefüggéseit tárgyaló könyvében. Friedrich Kittler *Könyv és perspektíva* című tanulmányában a kódfejtés e technikáját Leon Battista Albertinek, a lineáris perspektíva elméletét és gyakorlatát 1436-ban összefoglaló tudósnak tulajdonítja. Az európai modernitás kialakulását a perspektivikus ábrázolás és a könyvnyomtatás „médiumszövetségéből" levezető tanulmány szerint Európának a földgolyóra kiterjesztett hatalma és szellemi–technikai fejlődése egyaránt a dolgok helyiértékét meghatározó találmányoknak köszönheti: amíg a perspektivikus rajz mint rácsozat a látvány elemeit rendezi el a felület síkján,[7] addig a könyvnyomtatás a térközzel elválasztott diszkrét betűk egymáshoz viszonyított helyét jelöli ki a papírlapon.[8] Ezek a gyakorlatok az elemek címezhetőségét (pozíciójuk meghatározását és viszonyát), megszámlálhatóságát, valamint pontos reprodukcióját tette lehetővé, egy olyan médiarendszert alkotva, amelyben széles körben váltak reprodukálhatóvá a technikai ismeretek (geometriai–szerkezeti rajzok és a hozzájuk tartozó leírások formájában).[9] Ez a médiaszövetség (nyomtatott könyv és perspektivikus rajz kölcsönhatása) érhető tetten Alberti munkásságán belül is, aki ugyanis a perspektivikus szerkesztés meghatározása mellett a modern kódfejtés megalkotójaként „nem tett mást, mint hogy alkalmazta a titkosírások elemzésére Gutenberg betűszekrényeinek elementáris elvét, mely szerint a gyakoribb betűkből többet kell készenlétben tartani, mint a ritkábbakból, s ennyiben már eleve betűgyakorisági analízisek."[10] Azaz egy nyelvben gyakrabban előforduló betűket gyakrabban kell használni a nyomdai szedés során, ezért azokból egyszerűen több darabra van szükség a szedőszekrényben, mint a ritkábban előfordulókból — így a szövegek létrehozása a nyomtatott kultúrában már mindig is az Alberti által a titkosírás feltörésére alkalmazott nyelvstatisztikai elemzésekre van utalva. Shannon — aki maga is dolgozott kódfejtőként — tanulmányában a nyelv statisztikai alapokon történő szerveződését az alábbi kísérlettel szemlélteti.

* Az Innovációs és Technológiai Minisztérium ÚNKP-20-3 kódszámú Új Nemzeti Kiválóság Programjának a Nemzeti Kutatási, Fejlesztési és Innovációs Alapból finanszírozott szakmai támogatásával készült.

[1] „Az üzeneteknek gyakran jelentésük van; ez azt jelenti, hogy valamely — bizonyos fizikai vagy fogalmi dolgokkal jellemzett — rendszerre vonatkoznak, illetőleg aszerint korreláltak. A hírközlés elméletének e szemantikai vonatkozásai közömbösek a műszaki probléma szempontjából. A lényegi kérdés az, hogy a tényleges üzenet, egy sor lehetséges közül kiválasztott egyetlen üzenet." Claude E. Shannon – Warren Weaver: *A kommunikáció matematikai elmélete*, ford.: Tompa Ferenc, Országos Műszaki Információs Központ, Budapest, 1986, 45.

[2] *Uo.*, 47–50.

[3] Az információelmélet legfontosabb gyakorlati következménye a biztonságos és hatékony kommunikáció meghatározása. Vö.: Szemes Botond: *Ottlik Budája és az ideális kódolás = A mindenség ernyőjére kivetítve. Hatvanéves az Iskola a határon,* szerk.: Osztruloczky Sarolta, Kortárs, Budapest, 2021, 234–235.

[4] Shannon–Weaver, *A kommunikáció matematikai elmélete*, 54.

[5] vö.: Hajdicsné Varga Katalin, *Információközvetítés gyorsírással,* Új Jel-Kép, 2014/3, http://communicatio.hu/jelkep/2014/3/hajdicsne_varga_katalin.htm (utolsó megtekintés: 2021. 09. 16.).

[6] Abraham A. Moles: *Információelmélet és esztétikai élmény,* ford.: Pléh Csaba – Vajda András, Gondolat, Budapest, 1973, 57. Vö.: „A köznapi angol nyelv redundanciája, a kb. 8 betűnél nagyobb távolságokra nem véve figyelembe a statisztikus szerkezetet, durván 50%. Ez azt jelenti, hogy amikor angol nyelven írunk, az írott szöveg felét a nyelv szerkezete határozza meg, míg a másik felét szabadon választjuk." Shannon–Weaver, *A kommunikáció matematikai elmélete*, 73.

[7] Vö.: Bernhard Siegert: *Cultural Techniques. Grids, Filters, Doors and Other Articulations of the Real,* ford.: Geoffrey Winthrop-Young, Fordham, New York, 2015, 121–147.

[8] Friedrich Kittler: *Könyv és perspektíva,* ford.: Adamik Lajos = *Médiatörténeti szöveggyűjtemény,* szerk.: Peternák Miklós – Szegedy-Maszák Zoltán, Magyar Képzőművészeti Egyetem, Intermédia Tanszék Budapest, 2011, 9–11.

[9] *Uo.*, 13–14.

[10] *Uo.*, 10.

Az ezt a szerveződést leíró ismeretek vonatkozhatnak betűk és szavak gyakoriságára, sőt ezeknek kombinációira is: digramok esetében egy betű/szó előfordulásának valószínűségét a megelőző betű/szó határozza meg, trigramok esetében a megelőző két elem, és így tovább. Shannon kísérlete arra irányult, hogy milyen mértékben generálható értelmes szöveg csupán ezekre a statisztikákra hagyatkozva. Szavak digramjának gyakoriságából kiindulva egy olyan félig-meddig értelmes mondatot alkotott meg, amely éppen a szöveget hagyományos úton létrehozó írók ellen intézett támadásról ad hírt: „THE HEAD AND IN FRONTAL ATTACK ON AN ENGLISH WRITER THAT CHARACTER OF THIS POINT IS THEREFORE ANOTHER METHOD FOR THE LETTERS THAT THE TIME OF WHO EVER TOLD THE PROBLEM FOR AN UNEXPECTED."[11] Ez a „betűkre vonatkozó másik módszer" új fénytörésbe helyezi a nyelvet, amely így már nem csak az értelmetlen, nem-emberi zajok leválasztásának eredményeként jön létre, hanem puszta valószínűségek megvalósulásának is a terepe: „ezután a betűk nem részesülnek jobb bánásmódban, mint a számok a maguk korlátlan maniplulálhatóságában."[12]A szövegek ilyen irányú tanulmányozása mégis távolt állt a hagyományos irodalomtudomány működésének fősodrától, és inkább a nyelvészet területén található költők és írók stílusának kvantitatív szempontú leírása vált jellemzővé.[13] Zsilka Tibor *Statisztika és stilisztika* című, 1974-ben megjelent könyve például gyakoriságelemzéseken keresztül világít rá stílustörténeti, vagy adott szövegek stílusára vonatkozó jellegzetességekre — bár az elemzett szövegek terjedelméhez képest gyakran általánosítónak, vagy túlzónak hathat egy-egy kijelentése.[14] Könyvének központi kérdése „a szöveg stílusának és hírértékének az összefüggései, valamint a szépirodalmi stílus információs tartalmának a matematikai vizsgálata,"[15] amely során az információelméletből merít ösztönzést, hogy a különböző elemek frekvenciáját kiszámítva irodalmi művek esztétikai karakterét ragadhassa meg. Az irodalomtudományban ezzel szemben a szavak előfordulásának gyakoriságelemzése inkább az ismétlés poétikai szerepének nagyhatású leírásaiban öltött testet, amelyekben a redundanciát egyrészt mint a művészi anyag „formává szerveződésének az elvét",[16] másrészt mint a többletjelentéseket létrehozó retorikai figurát tárgyalták, hiszen semmilyen szemantikai ismétlődés nem identikus ismétlést jelent, hanem új kontextusokban új összefüggések létrehozását.[17] Ezek a megközelítések ugyan hangsúlyozzák, hogy a művészet a váratlan és az ismétlődő elemek kettősségében alakul,[18] ám kevésbé érvényesítik Shannon matematikai belátásait, hanem inkább a rendszerelmélet és a kibernetika felől írják le az irodalom és a művészet (ön)szerveződését.[19] Holott még a 20. század első felében a szoros olvasásnak mint az értelmezéseket előállító, elsődleges irodalomtudományos gyakorlatnak a kialakulásakor is felfedezhetjük a gyakoriságelemzés különböző módozatait. Sőt az Új Kritika korai időszaka „a szoros olvasás és a statisztikai analízis kereszteződéséniek történeteként"[20] is elmesélhető, amennyiben rávilágítunk a központi szerzők pedagógiai tevékenységeinek sokszínűségére. I. A. Richards és C. K. Ogden munkássága például a *close reading* módszerének kifejlesztése mellett az ún. *Basic English* létrehozást is magában foglalta.[21] A *Basic English* (British American Scientific International and Commercial English) egy olyan 850 szót tartalmazó lista, amelynek segítségével bármely angol szöveg értelmesen visszaadható egyszerűsített formában. Ez a lista a korabeli szógyakorisági listák alapján jött létre, amelyek a *Basic Englis*-hez hasonlóan pedagógiai célokat szolgáltak és a legtöbbet használt angol kifejezéseket tartalmazták, amelyeket egy végzős diáknak tanulmányai végére ismernie kell. Ogden és Richards vállalkozása a nemzetközi és tudományos kommunikáció általános nyelvének létrehozására (amelyet manapság leginkább *Erasmus English*-nek nevezhetnénk), a nyelvtanítás elősegítésére, valamint a költői szövegek „üzenetének" könnyebb megragadhatóságára irányult.[22] De ugyanígy fellelhetők hasonló műveletek Richards legnagyobb hatású tanítványának, William Empsonnak a munkásságában is. Az *Othello* értelmezésekor például abból von le interpretatív következtetéseket, hogy a *honest,* illetve *honesty* ('becsületes/őszinte', illetve 'becsületesség/őszinteség') szavak 52-szer fordulnak elő a szövegben, ami más Shakespeare-drámákhoz képest kiugró értéknek tekinthető.[23] Szintén a digitális irodalomtudomány módszereit idézi központi eljárása, amely során egy költemény szavainak lehetséges jelentéseit szótárak segítségével gyűjti össze, majd ütközteti egymással — a vektortér alapú szemantika nagyon hasonló módon jár el, amikor szavak jelentéseit eloszlásuk, azaz a környezetükben gyakran előforduló más szavak jelentései mentén határozza meg, és amely eljárás Empson módszeréhez hasonló, a szavak poliszémiáját kibontó értelmezések létrehozását teszik lehetővé.[24] Napjaink digitális irodalomtudománya szinte minden esetben a szövegek részeinek összeszámlálásából indul ki. Ez leginkább a szerzőattribúciós kutatások területén, azaz az ismeretlen szerzőségű szövegek alkotójának meghatározásakor szembetűnő. Ezek a kutatások ugyanis több ízben bizonyították már, hogy a szövegek tematikus szintje alatt létezik egy ún. „szerzői ujjlenyomat", amely a leggyakrabban és ezért nem tudatosan használt szavak, főként konkrét jelentés nélküli funkciószavak (névelők, kötőszók stb.) eloszlására vonatkozik, és amely eloszlás a szerzők különböző időszakban írt, különböző műfajú szövegeiben is hozzávetőlegesen állandó.[25] A leggyakrabban használt szavak olyan mennyiségű adatot szolgáltatnak a matematikai számításokhoz, amelyek alapján lehetőség van az egy alkotóhoz tartozó műveket elkülöníteni másokétól. Ez nem csak a vitatott szerzőségű szövegek esetében jár fontos következménnyel, hanem a stílus fogalmának gyökeres megváltozását is maga után vonja, amely innentől kezdve statisztikai módon, egyszerű gyakoriságelemzés útján válik leírhatóvá. Ez természetesen csak a digitális lehetőségnek köszönhetően „egyszerű", hiszen az elemek összeszámlálása számítógépes kapacitással és parancssorok segítségével gyorsan és hatékonyan végezhető el. Bár ennek a módszernek is található előzménye a digitalitást megelőző korokból. Wincenty Lutoslawski lengyel filozófus például már a 19. század végén a szavak frekvenciájának és eloszlásának számítását hívta segítségül, hogy meghatározza Platón dialógusainak kronologikus rendjét.[26] Statisztikai alapokon nyugvó módszerét a digitális stíluskutatásban is átvett „stilometriának" keresztelte el, és az így elért eredményeivel komoly hatást gyakorolt a 20. század filozófiatörténeti és klasszika-filológiai munkáira is. „A Mester mosolygott volna azokon, akik szövegeiben a szavakat számolják. De ha a modern mechanika a Platón számára még ismeretlen módszereket alkalmazva a bizonyosságnak arra a fokára lépett, amely alapján az emberi lélek bármely vizsgálatánál egy egzaktabb tudománynak tarthatjuk, akkor nem engedhetjük Platón nyelvi szkepticizmusának, hogy távol tartson bennünket stílusának ilyen irányú elemzéséről."[27]

Léteznek továbbá a nyelv olyan matematikai megközelítései, amelyek irodalmi szövegek értelmezésében is fontos szerepet játszhatnak. Ezek az eljárások Zsilka Tibor stilisztikai kutatásaihoz hasonlatosak, amelyeket ugyanakkor a számítógépes kapacitásnak köszönhetően precízebben és a kvantitatív kritériumoknak sokkal inkább megfelelő módon tudunk elvégezni. Erre példa a szókincsgazdagság mérése, amelynek alapja típus–token arány (type–token ratio, TTR), azaz a szövegben előforduló típusok (mint szótári alakok) és a tényleges szavak számának hányadosa. Ugyanakkor ezzel a módszerrel a különböző hosszúságú szövegek nem összehasonlíthatók, hiszen minél hosszabb egy szöveg, annál redundánsabb, azaz annál gyakrabban ismétlődnek benne az egyes szavak. Zsilka még különböző képleteket alkalmaz, hogy mérései az eltérő hosszúságú szövegek esetében is használhatók legyenek, ám ezek csak hozzávetőleges és nem minden esetben megbízható eredményekre vezetnek. Megoldást jelent viszont a problémára a Georgia University kutatói által fejlesztett szoftver, ami csak egy beállítható nagyságú szakaszban vizsgálja a típus–token arányt az összehasonlítandó szövegekben, például 500 szavanként: az első 500 szó után egyetlen szót tovább lépve szintén kiszámítja ezt az arányt a 2. és az 501. szó által határolt egységben is, majd így tovább, egyesével lépegetve a szöveg végéig. Ezáltal egy mű egészét átvizsgálja a megadott lépték szerint, és az így megkapott arányszámokat átlagolva (ez a type–token arány mozgó átlaga, a MATTR, azaz a Moving–Average Type–Token Ratio) ad meg egy olyan értéket, amelyek már összemérhetővé teszik a különböző hosszúságú szövegeket, hiszen ezek az értékek ugyanakkora szövegrészek átlagát mutatják.[28]

Még inkább szoros kapcsolat létesíthető a statisztikai mérések és a szövegek értelmezése között a téma-modellezés *(topic modelling),* illetve a kulcsszó-elemzés *(keyword analysis)* során. Az előbbi a szöveget témák „keverékeként" gondolja el,

[11] „A FEJ ÉS FRONTÁLIS TÁMADÁSBAN EGY ANGOL ÍRÓVAL SZEMBEN HOGY E PONT KARAKTERE ENNÉLFOGVA EGY MÁSIK MÓDSZER A BETŰKRE NÉZVE AMI ANNAK AZ IDEJE AKI VALAHA A PROBLÉMÁT MONDTA EGY VÁRATLANRA." Shannon–Weaver, *A kommunikáció matematikai elmélete,* 60. A fordítást módosítottam — Sz. B.

[12] Friedrich Kittler: *Jel és zaj távolsága*, ford.: Lőrincz Csongor = *Intézményesség és kulturális közvetítés*, szerk.: Bónus Tibor – Kelemen Pál – Molnár Gábor Tamás, Ráció, Budapest, 2005, 462.

[13] Lásd például: Deme László: *Mondatszerkezeti sajátságok gyakoriségi vizsgálata,* Akadémiai, Budapest, 1971; Nagy Ferenc: *Kvantitatív nyelvészet*, Tankönyvkiadó, Budapest, 1972.

[14] A legrészletesebb számítások a *Mérések a szöveg fonetikai, ritmikai és morfológiai szintjén* című fejezetben található, amelyben Tóth Árpád, József Attila, Kassák Lajos és Weöres Sándor négy-négy költeményének összehasonlítását végzi el: a hangzás kapcsán fonémák (mássalhangzók–magánhangzók, rövid–hosszú magánhangzók, zöngés–zöngétlen mássalhangzók aránya), a ritmus kapcsán szótagok, az irodalomtörténeti korszakolás (impresszionalizmus vs. expresszionalizmus) és a szövegek tematikus szintje (pl. tárgyilagosság, társadalmiság) kapcsán a szófajok előfordulásának gyakoriságát számítja ki. Zsilka Tibor: *Stilisztika és statisztika*, Akadémiai Kiadó, Budapest, 1974, 46–76.

[15] *Uo.*, 11. Zsilka Shannon entrópia-fogalmát és képleteit is alkalmazza — vö. pl.: „Entrópiáról a nyelvvel kapcsolatban is beszélhetünk: szemantikai szinten a szavak várható vagy váratlan előfordulásában, az adott helyen történő felhasználásuk kisebb vagy nagyobb valószínűségében gyökerezik." *Uo.,* 28.

[16] Szegedy-Maszák Mihály: *Az ismétlődés mint a művészi anyag formává szerveződésének elve* = Uő: *Világkép és stílus. Történeti–poétikai tanulmányok,* Magvető, Budapest, 1980, 367. Vö. Moles, *Információelmélet* , 92; és Umberto Eco: *A nyitott mű,* ford.: Dobolán Katalin – Mártonffy Marcell, Európa, Budapest, 2006, 146.

[17] Szegedy-Maszák, *Az ismétlődés…* 370–71.

[18] *Uo.,* 371; Eco, *A nyitott mű,* 54, 146–165.

[19] A kibernetika entrópia-fogalmához lásd: Norbert Wiener: *Cybernetics. Bevezetés,* ford.: Tarján Rezsőné = Uő: *Válogatott tanulmányok,* Gondolat, Budapest, 1974, 77; a művészetelméleti megközelítéshez: William R. Paulson: *The Noise of Culture*, Cornell UP, Ithaca–London, New York, 1988.

[20] Yohei Igarashi, *Statistical Analysis at the Birth of Close Reading,* New Literary History, 46/3 (2015), 485.

[21] *Uo.*, 485–487.

[22] *Uo.*, 492–495. A nyelv redundanciájának vizsgálatakor Shannon is a *Basic English*-re mint az egyik szélsőértékre hivatkozik: „A redundancia két szélsőséges példája az angol prózában a *Basic English* és James Joyce *Finnegan ébredése* című könyve. A *Basic English* nyelv szókészlete 850 szóra korlátozódik és redundanciája igen nagy. Ez tükröződik abban a tényben, hogy egy bekezdést *Basic English*-re lefordítva az meghosszabbodik. Másfelől Joyce megnövelte a szókészletet és — úgy tartjuk — a szemantikai tartalom tömörítését érte el." Shannon–Weaver, *A kommunikáció matematikai elmélete,* 74.

[23] William Empson: *Honest in Othello* = Uő: *The structure of complex words, Harvard UP,* Cambridge MA, 1989.

[24] Michael Gavin: *Vector Semantics, William Empson, and the Study of Ambiguity,* Critical Inquiry, 2018/44, 641–673.

[25] Lásd pl.: Harald Baayen: *Word Frequency Distributions,* Kluwer, Dodrecht, 2001.
John Burrows: *„Delta": A Measure of Stylistic Difference and a Guide to Likely Authorship*, Literary and Linguistic Computing, 2002/17, 267–287.
Patrick Juola: *Authorship Attribution,* Foundations and Trends in Information Retrieval, 2006/1, 233–334. Stilometriai kutatásokat összefoglaló tanulmány: Maciej Eder: *Style-Markers in Authorship Attribution. A Cross-Language Study of the Authorial Fingerprint,* Studies in Polish Linguistic, 2011/1, http://www.ejournals.eu/sj/index.php/SiPL/article/view/2261/0 (utolsó megtekintés: 2021. 09. 16.).

[26] „Ugyanazon szerző két azonos méretű alkotása közül az áll közelebb időben egy harmadikhoz, amely nagyobb számú stílusbeli sajátosságokon osztozik vele, feltéve, hogy figyelembe vesszük a sajátosságok eltérő fontosságát, és hogy ezek száma elegendő ahhoz, hogy meghatározza mindhárom mű stiláris jellegzetességét." Wincenty Lutoslawski: *The Origin and Growth of Plato's Logic: With an Account of Plato's Style and of the Chronology of His Writings,* Forgotten Books, London, 2018, 153.

[27] *Uo.*, 65.

[28] Michael A. Covington – Joe D. McFall: *Cutting the Gordian Knot: The Moving–Average Type–Token Ratio (MATTR)*, Journal of Quantitative Linguistics, 2010/2, 94–100. A magyar nyelv agglutináló jellege miatt érdemes ezeket a méréseket a szövegek szavainak lemmatizált formáján elvégezni.