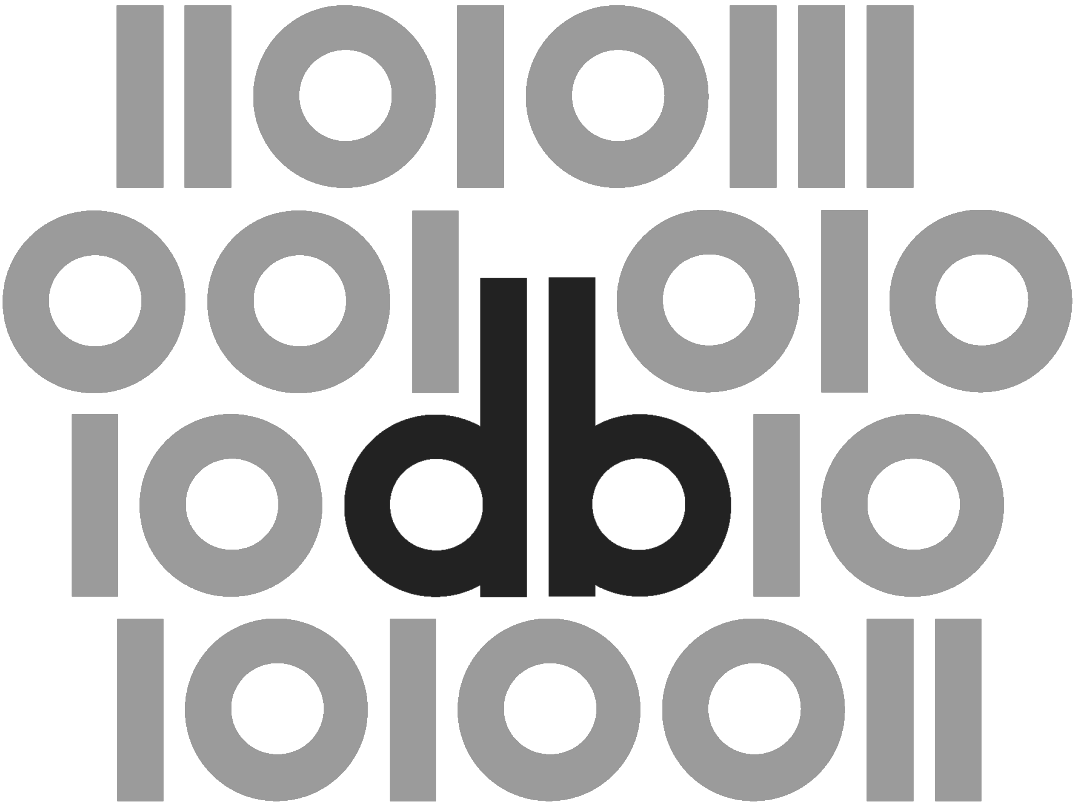
5 (2021) <DIGITÁLIS BÖLCSÉSZET>

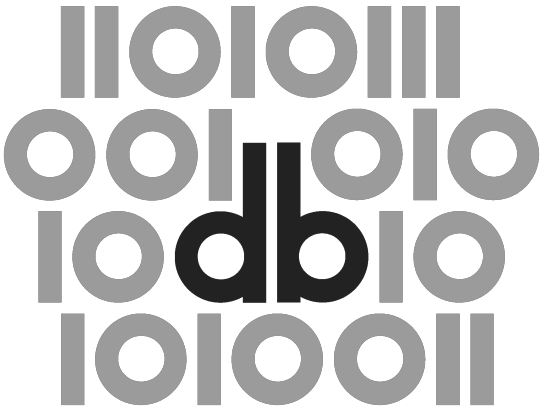
A krakkói Computational Stylistics Group (Különszám)



5 (2021) </DIGITÁLIS BÖLCSÉSZET>

Digitális Bölcsészet
2021., ötödik szám
A krakkói Computational Stylistics Group (Különszám)

<DIGITÁLIS BÖLCSÉSZET>



5 (2021)

A krakkói
Computational Stylistics Group

(Különszám)

A különszámot Szemes Botond szerkesztette.

ISSN 2630-9696
DOI 10.31400/dh-hun.2021.5
Kiadja a Bakonyi Géza Alapítvány és az ELTE BTK Régi Magyar Irodalom Tanszéke (1088 Budapest, Múzeum krt. 4/A).
Felelős kiadó az ELTE BTK Régi Magyar Irodalom Tanszék vezetője.

Megjelenik az Open Journal Systems (OJS) v. 3. platformon, melynek működtetését az ELTE Egyetemi Könyvtár- és Levéltár biztosítja.

A különszám megjelenését a Wacław Felczak Alapítvány támogatta.

Honlap: `http://ojs.elte.hu/digitalisbolcseszet`
Email cím: dbfolyoirat@gmail.com

Olvasószerkesztő: Bucsics Katalin
Tördelés: Hegedüs Béla
Grafika: Hegyi Gábor

Tartalom

Artjoms Šeļa 0000-0002-2272-2077

Instytutu Języka Polskiego PAN

artjoms.sela@ijp.pan.pl

Boris Orekhov 0000-0002-9099-0436

Higher School of Economics, Moscow

borekhov@hse.ru

Roman Leibov 0000-0002-5521-2954

Tartu Ülikool

hv.dekanaat@ut.ee

Gyenge műfajok
A költői versmérték és a jelentés közötti kapcsolat modellálása az orosz költészetben*

A dolgozat egy már meglévő, „a versmérték jelentésmezőjeként" ismert költészetelmélet formalizálását kísérli meg, amely elmélet azt állítja, hogy a modern líra különböző metrikai formái bizonyos jelentésbeli asszociációkat halmoznak fel és őriznek meg. Az LDA témamodellező *(topic modelling)* algoritmussal vizsgáltuk az orosz költészet tág korpuszát (1750–1950), hogy ezáltal minden egyes verset egy tématérben, a versmértékeket pedig a témák valószínűségének eloszlása szerint reprezentáljunk. Nem felügyelt osztályozást és kiterjedt mintavételt alkalmazva megmutatjuk, hogy a verselési formákon belül és között erős a forma és a jelentés kapcsolata: ugyanahhoz a versmértékhez tartozó két minta sokszor nagyon is hasonlóként tűnik fel, és ugyanannak a családnak két verselési formája legtöbbször szintén egy klaszterbe kerül. Ez a kapcsolat akkor is kimutatható, ha a korpusz kronológiai szempontból ellenőrzött, és nem következménye a populáció méretének. Amellett érvelünk, hogy hasonló megközelítést nyelvek és költészeti hagyományok szemantikai mezőinek összehasonlításakor is alkalmazni lehet, amelynek révén az irodalomtörténet legalapvetőbb kérdéseire adhatók releváns válaszok.

Kulcsszavak:
költészet, szemantika, versmértékek, témamodellezés, klaszterezés

* Eredeti megjelenés: Artjoms Šeļa, Boris Orekhov and Roman Leibov, „Weak Genres: Modeling Association Between Poetic Meter and Meaning in Russian Poetry," in *CHR 2020: Workshop on Computational Humanities Research*, 2020, `http://ceur-ws.org/Vol-2723/long35.pdf`.

1. Bevezetés

A költői forma és annak jelentése közötti kapcsolat talán triviálisnak tűnhet. Történetileg a metrikai megkülönböztetés vezetett a műfajok és a költői beszéd típusainak elkülönítéséhez, egészen az indoeurópai epika „hosszú" és a lírai vers „rövid" soraiig.[1] Általában nem várunk önelemző meditációt egy limericktől, míg egy szonettől talán számítunk rá. A daktilikus hexameter európai imitációi vagy az elégikus disztichon a modern verselési rendszerekben tematikusan kötődnek klasszikus kori forrásaikhoz. Vajon a versforma és annak szemantikája közötti kapcsolat érvényes a versmértékek „általános használatára" a modern költészeti hagyományban is, ahol a műfaj és a forma közötti normatív kapcsolatok gyorsabban elenyésznek? A válasz, amellyel mindenki egyetért: igen.

A versmértékek képessége, hogy az idő múlásával felhalmozzanak és megőrizzenek különböző szemantikai jellemzőket, „a versmérték szemantikai mezőjeként" is ismert a kvantitatív metrikai tudományok orosz iskolájában.[2] A kezdeti megfigyelések azonban csupán egy-egy költő metrumhasználatán[3] vagy anekdotikus bizonyítékokon alapultak (nevezetesen néhány időben elszórt, trochaikus pentameterben megalkotott költeményre). A korai kutatók mindazonáltal a versmérték-jelentés kapcsolatot organikusnak tekintették, vagyis úgy gondolták, hogy a ritmus néhány belső tulajdonsága formálja a vers jelentését.[4] Mikhail Gasparov több ezer 19. századi költemény szoros olvasására alapozva demonstrálta, hogy ezek a kapcsolatok történeti jellegűek, amelyeket a versmérték helyi hagyományai és a későbbi alkalmazásai határoznak meg, ami egy szétszóródott, de mégis megkülönböztethető szemantikai profil létrejöttét eredményezi.[5]

[1] Mikhail Leonovich Gasparov, *A History of European Versification,* trans. Gerald Stanton Smith and Leofranc Holford-Strevens (Oxford–New York: Clarendon Press–Oxford University Press, 1996), https://doi.org/10.1093/acprof:oso/9780198158790.001.0001; Antoine Meillet, *Les origines indo-européennes des mètres grecs* (Paris: Les Presses universitaires de France, 1923), hozzáférés: 2021.12.07, https://archive.org/details/lesoriginesindoe00meiluoft.

[2] Maxim Iljics Shapir, „'Semanticheskii oreol metra': termin i poniatie," in Maxim Iljics Shapir, *Universum versus: iazyk, stikh, smysl v russkoi poezii XVIII–XX vekov,* Vol. 2., 395–404 (Moszkva: Iazyki slavianskoi kul tury, 2015); Marina Tarlinskaja and Naira Oganesova, „Meter and Meaning: The Semantic Halo of Verse Form in English Romantic Lyrical Poems (Iambic and Trochaic Tetrameter)," *The American Journal of Semiotics* 4, 3–4. sz. (1986): 85–106, https://doi.org/10.5840/ajs198643/422; Mikhail Trunin, „Towards the Concept of Semantic Halo," *Studia Metrica et Poetica* 4, 2. sz. (2017): 41–66, https://doi.org/10.12697/smp.2017.4.2.03.

[3] Grigoriy Vinokur, „Vol'nye iamby Pushkina," in *Pushkin i ego sovremenniki: Materialy i issledovania,* Vol. 38–39, 23–26 (Leningrad: 1930).

[4] Roman Jakobson, „Toward a Description of Mácha's Verse," in Roman Jakobson, *Selected Writings, Vol. 5: On Verse, Its Masters and Explorers,* eds., Stephen Rudy and Martha Taylor, 433–485 (The Hague–Paris–New York: Mouton Publishers, 1979), https://doi.org/10.1515/9783110803068.433; Kiril Taranovskii, „O vzaimootnosheniah stikhotvornogo metra i tematiki," *American Contributions to the Fifth International Congress of Slavists, Sofia, September 1963: Vol. 1: Literary contributions,* 287–332 (The Hague: Mouton and Co., 1963).

[5] Mikhail Gasparov, *Metr i smysl: ob odnom iz mekhanizmov kulturnoi pamiati* (Moscow: Izdatelskii tsentr RGGU, 1999).

 DOI 10.31400/dh-hun.2021.5.3145

Az ilyen megállapítások vonzereje ellenére a szemantikai mező koncepciója a formalizálás hiányában könnyen kritizálható és nehezen védhető elképzelés. Még ha egyes konkrét „mezők" nem is egyszerű mintavételi hiba termékei, magára a mechanizmusra és a metrikus formák közötti kapcsolatok szerkezetére vonatkozó általánosítások továbbra is megfoghatatlanok maradnak. Ugyanakkor néhány korábbi empirikus kísérlet, amely az orosz[6] és a baskír[7] költészetben a versmérték-jelentés viszony megközelítésére irányult, szólisták összehasonlítására támaszkodva, egészen jól körül tudta írni a metrikus formák közötti lexikális különbségeket, amely eredmények számunkra is fontos kiindulási pontot biztosítanak az alábbiakban.

Ez a dolgozat ugyanis a szemantikus mező jelenlétét igyekszik megvizsgálni az orosz költészetben, absztrakt szemantikus jellemzők (témák) alapján, amelyek minden egyedi verset egységes módon képesek leírni. Azáltal, hogy a szövegeket egyetlen modellen belül helyezzük el, rugalmasan tudunk teszteket végezni és osztályozási algoritmusokat használni a tudományos feltételezések kifejtésére és ellenőrzésére. Ennek során a hierarchikus klaszterezés módszerére támaszkodunk, hogy a versmértéken belüli szemantikus hasonlóságok szintjét (hasonlítanak-e a versmértékek önmagukra) és a versmértékek közötti kapcsolatokat (kapcsolódnak-e egymáshoz az egy családhoz tartozó, különböző versmértékek) felmérjük. Az elemzést követően tárgyaljuk, hogy a versmérték szemantikus mezőjének a formalizálása miként járulhat hozzá a metrumnak mint kulturális átadásnak *(cultural transmission)* a megértéséhez, és hogy miként lehetne hasonló megközelítést alkalmazni a témamezők különböző nyelveken és hagyományokon átívelő vizsgálatára.

2. Korpusz

A kutatásban felhasznált adatok az Orosz Nemzeti Korpusz Költészeti algyűjteményéből[8] származnak, amely a 18. századtól a 20. századig terjedő korszakban született szövegeket tartalmaz. Nagyjából tehát lefedi a modern orosz versmérték egész történetét, amely a német időmértékes verselés 1730-as megjelenésével kezdődött. A korpusz már elgondolásában is egyértelműen elfogult a kánon iránt: csak azok a 18–19. századi szövegek kerültek a gyűjteménybe, amelyek elérhetők a 20. századi kritikai kiadásokban.[9] Ezért nem vesz figyelembe sok korábbi, akadémiai kánonon kívüli költészeti teljesítményt, és egyenlőtlenséget teremt a költemények kronológiai eloszlásában is: a szövegeknek több mint 75%-a a 20. századból való. Ráadásul egyáltalán nem egységes a merítés, mivel 1917-tel kezdődően az orosz költészet három, általában véve elszigetelt hagyománnyá válik szét: szovjet, emigrációs és nem hivatalos underground. Mivel nem áll módunkban automatikusan elkülöníteni őket, a korpusz határát 1950-ben állapítjuk meg, ami kizárja a legtöbb underground művet, és megállítja az órát, mielőtt

[6] Alexander Piperski, „Semantic Halo of a Meter: A Keyword-Based Approach," in *Kompiuternaia lingvistka i intellectualnyie tekhnoloigii, Vol. 2: Kompiuternaia lingvistika: lingvisticheskiie issledovaniia,* 342–354 (Moscow: RGGU, 2017).

[7] Boris Orekhov, *Bashkirskii stikh XX veka: Korpusnoe issledovanie* (St. Petersburg: Aletejia, 2019).

[8] Orosz Nemzeti Korpusz (2003), hozzáférés: 2021.12.07, `https://ruscorpora.ru/new/en/index.html`.

[9] Kirill Korchagin, „Poezija XX veka v poeticheskom podkorpuse Natsional'nogo korpusa russkogo iazyka: problema reprezentativnosti," *Trudy instituta im. V. V. Vinogradova* 6 (2015): 235–256.

megkezdődne az észrevehető sodródás a nem klasszikus versmértékek felé. Miután minden felosztási művelet és előzetes feldolgozási lépés (lásd alább, a 3. részben) megtörtént, 47804 szöveg (2275233 szó) maradt a korpuszban.

Jelen dolgozat legfőképpen a szótagszámláló hangsúlyos verselésű *(accentual-syllabic, AS)*[10] – és általában rímes – költészetre összpontosít, ami sokkal tovább fennmaradt az orosz lírában, mint a szabadvers felé forduló nyugati tradíciókban.[11] Az orosz szótagszámláló hangsúlyos verselési rendszerek a hangsúlyok és a soron belüli szótagok számának szigorúbb behatárolásán alapszanak a pusztán hangsúlyos (ahol csak a hangsúlyok száma fontos) vagy a pusztán szótagoló (ahol csak a szótagszám számít) verseléshez képest. Ebben a verselési módban a versmértékek a ritmus visszatérő kisebb egységeire épülnek – verslábakra, amelyek mintákba rendezik a hangsúlyos és nem hangsúlyos szótagokat, általában kettőt vagy hármat (bináris vagy hármas láb). Mivel a metrikai séma a versritmus absztrakciója, és folyamatosan módosul (a várt hangsúlyos pozíciók hangsúlytalanok maradnak vagy fordítva), általában erős versus gyenge pozíciókról beszélünk „hangsúlyos" vagy „hangsúlytalan" helyett. A *B.3. táblázat* összegzést nyújt a klasszikus időmértékes formákról, amelyeket ebben a dolgozatban használunk. Kivételt képeznek az úgynevezett „dolnikok", amelyek a szótagszámra vonatkozó szabályok meglazításával eltávolodnak a szótagszámláló hangsúlyos verseléstől, ám mégsem lehet őket figyelmen kívül hagyni, olyan nagy számban vannak jelen a 20. században.

Annak érdekében, hogy – ha lehetséges – minden verset egyetlen, egyértelmű versképlettel írhassunk le, a korpusz metaadatait használtuk fel, amelyek a versformára vonatkozó annotációkat tartalmazzák. A korpuszannotációt intézményes keretek közt végezték a nyelvészet és a prozódia területén jártas szakemberek felügyeletével, ugyanakkor nem jegyezték a hibaszázalékot, vagy hogy mekkora volt az egyetértés az annotációt végzők körében.[12] Mindazonáltal nagyon magasra becsüljük a munka pontosságát, különösen a klasszikus formák tekintetében, amelyek még minimális képzettséggel is könnyen megkülönböztethetők. Megkértünk három irodalomtudóst, hogy igazoljanak 100 eredeti, a versformára vonatkozó korpuszannotációt: átlagosan 97,7% címkét jelöltek meg „igazként", a köztük lévő egyetértés mértéke pedig 96,6% volt (az alacsony mértékű egyetértés azokban az esetekben volt gyakori, amikor a címkét „hamisnak" tekintették).

[10] Az angol terminológia az angol és az ebből a szempontból hasonló orosz verselést tükrözi, amely a hangsúlyos és hangsúlytalan szótagok szabályos váltakozásán alapul (és ahol a hangsúlyos szótagok pozíciója nem kötött egy szón belül, mint például a magyarban). A klasszikus görög-latin időmértékes verselés rendszere adaptálható a hangsúlyos verselés viszonyaira, azonban ebben az esetben a verslábakat nem a rövid és hosszú, hanem a hangsúlytalan és hangsúlyos szótagok hozzák létre. Ez egyben egy kötött szótagszámú („szótagszámláló") verselést eredményez, hiszen a verslábak a szótagok számát is meghatározzák. A verstani kérdésekben nyújtott segítségért hálával tartozunk Ferencz Győzőnek – *a szerk.*

[11] Mikhail Gronas, *Cognitive Poetics and Cultural Memory: Russian Literary Mnemonics* (New York: Routledge, 2010), https://doi.org/10.4324/9780203842430.

[12] E. Grishina et al, „Poeticheskii korpus v ramkah NKRIA: obschaia struktura i perspektivy ispolzovania," in *Natsionalnii korpus russkogo iazyka: 2006–2008. Novye rezultaty i perspektivy,* 71–113 (St. Petersburg: Nestor-Istoria, 2009).

 DOI 10.31400/dh-hun.2021.5.3145

A metaadatokkal való címkézéskor meglehetősen konzervatívok voltunk, előnyben részesítettük a homogén metrikai lejegyzéseket, és kizártuk a polimetria vagy egyéb heterogén formák összetett eseteinek nagy részét. Szintén egyszerűsítéseket végeztünk a versszak tekintetében, és csak az általánosan elterjedt rímképletekre támaszkodtunk. E dolgozatban az orosz versmértékrendszerből származó metrikai lejegyzést használjuk, vagyis Jambus-4-fm a négyes jambust jelenti, amelyben rendszeresen váltakoznak a nőrímet és a hímrímet tartalmazó (akatalektikus vagy katalektikus) sorok.

A metrikus kifejezésnek tehát három szintje különböztethető meg egyetlen metrikai formulából:

> 1. A metrikai mintázat általános *család*ja (pl. a trocheus olyan versmérték, amely bináris verslábon alapul, erős pozícióval az első szótagon);
>
> 2. A *versmérték* a lábak számán alapul (pl. az ötös trocheus [Trocheus-5], 5 trochaikus verslábból álló trochaikus pentameter);
>
> 3. A versmérték katalektikus *variáns*a, amely az utolsó hangsúlyos szótag után álló nem hangsúlyos szótagok mintázatát írja le (pl. Trocheus-5-fm; f – nőrím (Xu), m – hímrím (X), d – daktilikus (Xuu) sorvég).

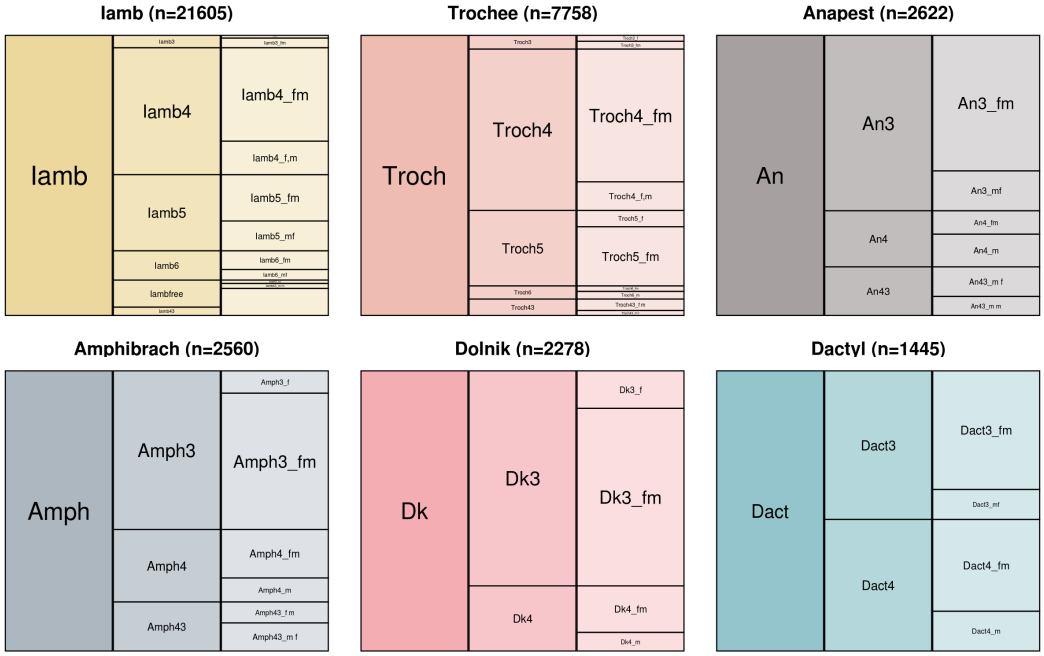
Az *1. ábra* a verseknek ezt a három szintű elosztását mutatja be a korpusz hat leggyakoribb metrikai családjához viszonyítva (versmértékenként csak a két leggyakoribb variáns látható), és megadja a versek családon belüli abszolút számát is. A jambikus versmértéknek, különösen pedig a négyes jambusnak mint az orosz időmértékes verselés „normatív versmértékének" túlsúlya egyértelmű. Azért, hogy kezeljük a versformák ennyire szélsőséges egyenlőtlenségét, a továbbiakban erősen támaszkodunk a véletlenszerű mintavételezésre és az iteratív kísérletekre.

3. A szemantika modellezése

Az egyedi versek szemantikai jellemzői révén igyekszünk modellezni a versmérték és a jelentés közötti kapcsolatot. Ennek érdekében az LDA (Latent Dirichlet Allocation) témamodellező algoritmusát[13] alkalmaztuk a teljes korpuszon a versek csoportosítása nélkül, a metrikai címkéket és az egyéb metaadatokat a dokumentumok nevében feltüntetve.

A témamodell az információt kinyerő algoritmusok egy nagy családjának együttes elnevezése, amelyek az egymás közelében előforduló elemek csoportjait keresik egy dokumentumgyűjteményben. Ezeket a csoportokat témának nevezzük (hiszen az eredeti cél a szövegbányászat volt), de a modellek alkalmazhatók például molekulák-

[13] D. M. Blei, Andrew Y. Ng and Michael I. Jordan, „Latent Dirichlet Allocation," *Journal of Machine Learning Research* 3 (2003): 993–1022.



1. ábra. A metrikai formák aránya a családokon belül. Minden egyes, legalább 200 vershez tartozó versmérték esetében a két leggyakoribb variánst ábrázoljuk. Az abszolút versszámok azonban minden verset magukba foglalnak.

ra,[14] zenére[15] és génekre[16] is, vagy bármilyen feladatra, amely megkívánja a hasonló viselkedésű csoportok kinyerését nagyszámú tulajdonság alapján (szavak, akkordok, gének, kémiai elemek stb.). A témamodellezést ma már széles körben alkalmazzák a bölcsészet- és társadalomtudományokban szövegbányászat és -osztályozás céljából;[17] és már az is többször bebizonyosodott, hogy az LDA kisebb költői szövegek korpuszára is alkalmazható.[18]

[14] Michał Woźniak et al., „Linguistic Measures of Chemical Diversity and the 'Keywords' of Molecular Collections," *Scientific Reports* 8 (2018), https://doi.org/10.1038/s41598-018-25440-6.

[15] Matthias Mauch et al., „The Evolution of Popular Music: USA 1960–2010," *Royal Society Open Science* 2, 5. sz. (2015): https://doi.org/10.1098/rsos.150081.

[16] Manuele Bicego et al., „Investigating Topic Models' Capabilities in Expression Microarray Data Classification," *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 9, 6. sz. (2012): 1831–1836, http://doi.org/10.1109/TCBB.2012.121.

[17] David Hall, Daniel Jurafsky and Christopher D. Manning, „Studying the History of Ideas Using Topic Models," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing. (EMNLP '08)*, 363–371 (Stroudsburg, PA: Association for Computational Linguistics, 2008), https://doi.org/10.3115/1613715.1613763; Paul DiMaggio, Manish Nag and David M. Blei, „Exploiting Affinities Between Topic Modeling and the Sociological Perspective on Culture: Application to Newspaper Coverage of U.S. Government Arts Funding," *Poetics* 41, 6. sz. (2013): 570–606, https://doi.org/10.1016/j.poetic.2013.08.004; Christof Schöch, „Topic Modeling Genre: An Exploration of French Classical and Enlightenment Drama," *Digital Humanities Quarterly* 11, 2. sz. (2017), https://doi.org/10.5281/zenodo.166356.

[18] Ehsaneddin Asgari, Marzyeh Ghassemi and Mark Alan Finlayson, „Confirming the Themes and Interpretive Unity of Ghazal Poetry Using Topic Models," in *Proceedings of the Neural Information Processing Systems (NIPS) Workshop on Topic Models* (Lake Tahoe, NV, December 2013); Borja Navarro-Colorado, „On Poetic Topic Modeling: Extracting Themes and Motifs From a Corpus of

 DOI 10.31400/dh-hun.2021.5.3145

Ígéretesnek tűnik továbbá a témamodellek alkalmazása a kultúrtörténet általános kérdéseinek modellezésére is: megragadhatóvá vált például a popzenében a változás mértéke,[19] az információ tudományos kutatásának módjai[20] vagy az innováció és a korábbi témák napirenden tartása közti különbség a történeti-politikai diskurzusban.[21] Ezekben az esetekben az entitások témareprezentációja pusztán a „tartalom" megragadására irányul. A költői nyelv hasonlóan absztrakt reprezentációját célozzuk meg mi is, tétován utánozva az irodalmárokat, akik olyan magas rendű szemantikai címkéket használtak a versmérték-specifikus jelentések leírásához, mint az este, az út vagy a halál (olyan témák, amelyek Gasparov szerint együttesen fejezik ki a trochaikus pentameter legfőbb szemantikai irányait az orosz költészetben).[22]

Az LDA egy generatív valószínűségmodell, amely néhány nagyon fontos feltételezésen alapul: 1) a gyűjtemény minden szövege *k* számú témából áll; 2) minden egyes téma leírható az összes rendelkezésre álló jellemző (jelen esetben: szavak) valószínűségi eloszlásával (ahol a legtöbb jellemző nagyon valószínűtlen az adott témára). Az LDA a szövegeket *k* téma eloszlása mentén határozza meg, így minden dokumentumot lényegében egyenlő méretű vektorként lehet leírni egyetlen „tématérben". Más szavakkal, az LDA megpróbál az együttesen előforduló szavak bizonyos számú csoportjára automatikusan következtetni; ennek köszönhetően minden egyes dokumentum ezeknek a csoportoknak a kombinációjaként lesz értelmezhető. A témamodellek használatát döntő fontosságúnak tekintjük, mert 1) az LDA lehetővé teszi az egyes versek szintjén az egységes szemantikai absztrakciót; 2) a dokumentumokat potenciálisan kis számú és jól értelmezhető dimenzióval fejezi ki; 3) a versekben a témák valószínűségei lehetővé teszik a lényegre törő utólagos elemzést; 4) a témamodellek függetlenítik a megközelítésünket a nyelv- és egyéb specifikus szakterületektől.

A modell betanítása előtt a korpusz előzetes feldolgozását az alábbi lépésekben végeztük el:

Spanish Poetry," *Frontiers in Digital Humanities* 5 (2018), https://doi.org/10.3389/fdigh.2018.00015; Thomas N. Haider, „Diachronic Topics in New High German Poetry," in *Proceedings of the International Digital Humantities Conference. Utrecht, 8–12 July* 2019, https://dev.clariah.nl/files/dh2019/boa/1031.html; Petr Plechac and Thomas N. Haider, „Mapping Topic Evolution Across Poetic Traditions," in *arXiv:2006.15732* [cs. stat], August 2020, hozzáférés: 2021.12.07, https://arxiv.org/abs/2006.15732.

[19] Mauch et al., „The Evolution of Popular Music."

[20] Jaimie Murdock, Colin Allen and Simon DeDeo, „Exploration and Exploitation of Victorian Science in Darwin's Reading Notebooks," *Cognition* 159, február (2017): 117–126, https://doi.org/10.1016/j.cognition.2016.11.012.

[21] Alexander T. J. Barron et al., „Individuals, Institutions, and Innovation in the Debates of the French Revolution," *Proceedings of the National Academy of Sciences of the United States of America* 115, 18. sz. (2018), 4607–4612, https://doi.org/10.1073/pnas.1717729115.

[22] Gasparov, *Metr i smysl.*

1. Minden szöveget lemmatizáltunk a *mystem* 3.1-et használva.[23]

2. A korpuszra egy általános stopszólistát alkalmaztunk (eltávolítottuk a kötőszókat, a partikulákat, az elöljárószókat, a névmásokat és a számneveket).

3. Csökkenteni akartuk a korpusz lexikális változatosságát, ezért csak az 5000 leggyakoribb szóra képeztük ki a modellt. Az LDA eredménye általában annál jobb, minél kevesebb a szórványosan előforduló adat, ezért szokás eltávolítani a ritka szavakat az előkészítés során. Ugyanakkor a költői nyelv szemantikai leegyszerűsítése érdekében különböző, ugyanezen a korpuszon betanított szóbeágyazási *(word-embedding)* modelleket is használtunk, hogy az 5000 szavas „magon" kívül eső szavakat helyettesíteni tudjuk (a word2vec implementációja a gensim nevű *Python* könyvtáron keresztül, a vektor mérete=300). Egy adott szót akkor helyettesítettünk a leggyakoribb 1000 szó valamelyikével, ha annak volt szemantikai szomszédja a hozzá kontextuálisan leginkább hasonlító 10 szó között (a megfelelő vektorok koszinuszhasonlóságában mérve). Ez az eljárás lehetővé teszi számunkra, hogy a szavakat hiponímiájukkal, gyakoribb szinonimáikkal vagy grammatikai variánsaikkal helyettesítsük (pl. kicsinyítő alakok kicserélése), és néhány esetben azt, hogy a költői nyelv tradicionális metonímiáit megmagyarázzuk (pl. a „Pontus" „óceán"-ra cserélése). Az eljárás nem volt tökéletes, némi zajjal járt, ami ugyanakkor nem volt észrevehető hatással a modellre. Azt is meg kell jegyeznünk, hogy az eredményeink akkor sem változtak szélsőségesen, ha nem hajtottuk végre a kontextuális cserét, vagy ha más felső határt szabtunk a leggyakoribb szavak kijelölésénél. A jelentéktelen hatásoktól függetlenül továbbra is a kontextuális cserével létrejött adatokra vonatkozó eredményeket közöljük, mivel hiszünk abban, hogy ez a szemantikus absztrakció felé vett irány fontos és a jövőben fejlesztésre érdemes. Az eredeti korpuszra vonatkozó főbb eredmények a mellékletben találhatók *(B.5. táblázat)*.

4. A korpuszt a szövegméret alapján is korlátoztuk, hogy az LDA-t legalább összehasonlítható szóeloszlású dokumentumokon képezzük ki. Eltávolítottuk a nagyon kicsi (kevesebb mint 4) és a nagyon nagy (több mint 100 soros) verseket, ami az összes szöveg körülbelül 95%-át meghagyta nekünk. Ezután tovább szűrtük a korpuszt a szavak száma alapján, csak a 10 és 90 százalék közötti szövegeket meghagyva (20 és 102 szavas versek, ami nagyjából megfelel a 12 és 50 soros verseknek, ha beleszámoljuk a stopszavak eltávolítását). Ezek a korlátozások mutatják, hogy a modellünk alapvetően a rövid lírai költészetre összpontosít (ami az uralkodó forma az orosz hagyományban, amelyben a vershossz összezsugorodása tapasz-

[23] Ilya Segalovich, „A Fast Morphological Algorithm with Unknown Word Guessing Induced by a Dictionary for a Web Search Engine," in Hamid R. Arabnia and Elena B. Kozerenko, eds., *Proceedings of the International Conference on Machine Learning; Models, Technologies and Applications. MLMTA'03, June 23–26, 2003, Las Vegas, Nevada,* 273–280 (CSREA Press, 2003).

 DOI 10.31400/dh-hun.2021.5.3145

> talható).[24] Úgy véljük azonban, bármilyen eredményünk is van, annak a hosszú elbeszélő költészetre is érvényesnek kell lennie, ahol a versritmus szemantikai hagyományai sokkal hangsúlyosabbnak tűnnek.[25] Az összes művelet után 47804 szöveg maradt a korpuszban (amiből 39220 versnek egyedi, a korpusz annotációjából származó, formára vonatkozó címkéje van).

Nincs általánosan elismert mód a modell számára optimális témaszám meghatározására:[26] ebben a dolgozatban a 80 témával tanított LDA eredményeiről számolunk be, ami a témakoherencia (log-valószínűség) és a témazavar (a modell „meglepetése", amikor nem látott adatot jósol meg) közötti kompromisszum középponti modellje. Más témaszámokkal (10 és 200 között) is teszteltük a főbb eljárásokat, amelyek szintén nagy teljesítményt mutattak (lásd *B.5. táblázat*ot a mellékletben). Az LDA-ra az alpha=0.1 (nem akartuk, hogy sok téma generáljon egyetlen szöveget, s hogy kezelhetetlenek legyenek az eloszlások) és a beta=0.3 (nem akartuk, hogy túl sok szó járuljon hozzá egy témához, inkább csak néhány) beállításokat alkalmaztuk.

A végső modell gyors ellenőrzéséhez összevethetjük a korábban kvalitatív módon meghatározott témákat a versmértékekhez rendelt szavak csoportjával (lásd *B.4. táblázat*ot a mellékletben). Míg néhány téma összeegyeztethetőnek tűnik a versmértékek feltételezett jelentésmezőjével, természetesen nincs közvetlen kapcsolat közöttük. A témák, ha nem is egyeznek meg az absztrahált metrikai témákkal (Gasparov sem rendszerszerűen használta őket a különböző versmértékek leírásakor), még így is jól értelmezhetők és felhasználhatók a céljainkra a versmértékek tartalmának eloszláson alapuló reprezentációjakor.

4. A mező feltérképezése

4.1. A versmértéken belüli hasonlóságok

A „versmérték jelentésmezőjének elmélete" feltételezi, hogy a jelentés nem véletlenszerűen oszlik el a metrikai formák között, azaz hogy minden egyes versmérték történeti módon egyedi szemantikai profilt épít ki. Az elmélet továbbá kumulatívnak tekinti a mezőhatást (legalábbis implicit módon): nem volnánk képesek rekonstruálni a versmérték szemantikáját egy elszigetelt versben, de sajátos mintázat tűnik fel, ha a versmérték használatának sokkal nagyobb körét vizsgáljuk egy tradíción belül. Ezeket az elveket újrafogalmazva azt mondhatjuk, hogy a jelentés-versmérték kapcsolat bizonyos önhasonlóságot feltételez egy versformán belül. Ha a mezőhatás létezik, akkor az ugyanolyan versmértékű versek két független csoportjának egymáshoz közelebb kellene lenniük a jelentés tekintetében, mint a különböző versmértékűekhez.

[24] Artjoms Šeļa and Oleg Sobchuk, „The Shortest Species: How the Length of Russian Poetry Changed (1750–1921)," *Studia Metrica et Poetica* 4, 1. sz. (2017): 66–84, https://doi.org/10.12697/smp.2017.4.1.03.

[25] Vö. Gasparov, *Metr i smysl.*

[26] Stefano Sbalchiero and Maciej Eder, „Topic Modeling, Long Texts and the Best Number of Topics: Some Problems and Solutions," *Quality & Quantity* 54, 4. sz. (2020): 1095–1108, https://doi.org/10.1007/s11135-020-00976-w.

Tegyük fel, hogy a hagyomány egészének a megfigyelői vagyunk, és a metrikai mezőkre az 1950-es évekből tekintünk (ez a korpuszunk felső időbeli határa). Ahhoz, hogy teszteljük, vajon a jelentés-versmérték kapcsolat észlelhető-e általános szinten, nem felügyelt osztályozást hajtunk végre minden, legalább 500 versre jellemző versmérték 200 példányának két véletlenszerű mintáján (visszatevés nélkül). Minden mintára kiszámítjuk a témavalószínűségeket, hogy az összesített témaeloszlást reprezentálni tudjuk az adott versmértéken belül. Mivel valószínűségek eloszlásával van dolgunk, következő lépésként a Jensen–Shannon divergenciát (amely szimmetrikus a Kullback–Leibler divergenciával[27]) számítjuk ki a minták között, és az így kapott távolságok alapján alkotjuk meg a hierarchikus klasztereket (dendrogramokat). Ezután folytatódik a mintavétel és az újraszámítás még 100-szor. A 100 klaszterelemzés információiból egy összesített, a „többségi szabály" elvével létrehozott konszenzusfa rajzolható fel:[28] az ábra akkor kapcsol össze elemeket, ha az összes dendrogramon 50%-os egyezés figyelhető meg, vagyis két ág nem kapcsolódik, ha nem tartoznak egy klaszterbe legalább a fák felében *(2a. ábra)*.

Ugyanezt az eljárást lehet alkalmazni a metrikai variánsok szintjén is. A metrikai annotációban levő szórványos adatok és zaj miatt csak a négyes jambust (Iamb-4) és azokat a variánsait használjuk, amelyek legalább 200 versre jellemzők, miközben eltávolítjuk a leggyakoribb variánst diffúz szemantikája miatt (Iamb-4-fm). Ez a négyes jambusnak mindössze négy formáját hagyja meg nekünk *(2b. ábra)*.

Anélkül, hogy ennek a megközelítésnek további komplikációit szóba hoznánk, világos, hogy a korpuszban vannak versmértéken belüli szemantikai hasonlóságok (ugyanazon mértékhez tartozó variánsok egymás mellé rendeződnek az ábrán). Természetesen a metrikai variánsok között felfedezhető szemantikai különbség is, bár az ilyen szintű részletességhez sokkal jobb annotációkra és strófainformációkra volna szükség. Mindenesetre a témainformáció önmagában elég arra, hogy két, ugyanabból a versmértékből származó, vitathatóan nagy mintát következetesen egy csoportba rendezzünk (ha a korpuszunkban egy vers méretének mediánja 50 szó, akkor a versmérték-jelentés kapcsolat 10000 szónyi mintában már meglehetősen hangsúlyos). A metrikai mező „kumulatív" hatását ellenőrizhetjük azáltal, hogy megnézzük, hogyan változik a hierarchikus csoportosítás teljesítménye a minta méretével.

A nem felügyelt osztályozás értékelésére két módszert használtunk: az egyszerű klasztertisztaságot (Cluster Purity, CP: a klaszteregyezések összege osztva az egyedi minták számával[29]) és a korrigált Rand-indexet (Adjusted Rand Index, ARI),[30] amelyeket arra terveztek, hogy összehasonlítsanak két osztályozást; az utóbbi pedig számot ad a véletlenszerű osztályozásról is (visszatérési értéke 0 körüli). Ebben az esetben is a versmértékenként 500 elérhető vers küszöbét használjuk, és az eddigi eljárásokat alkalmazzuk az egyre nagyobb mintákon (250 versig), kiszámolva a CP-t és az ARI-t

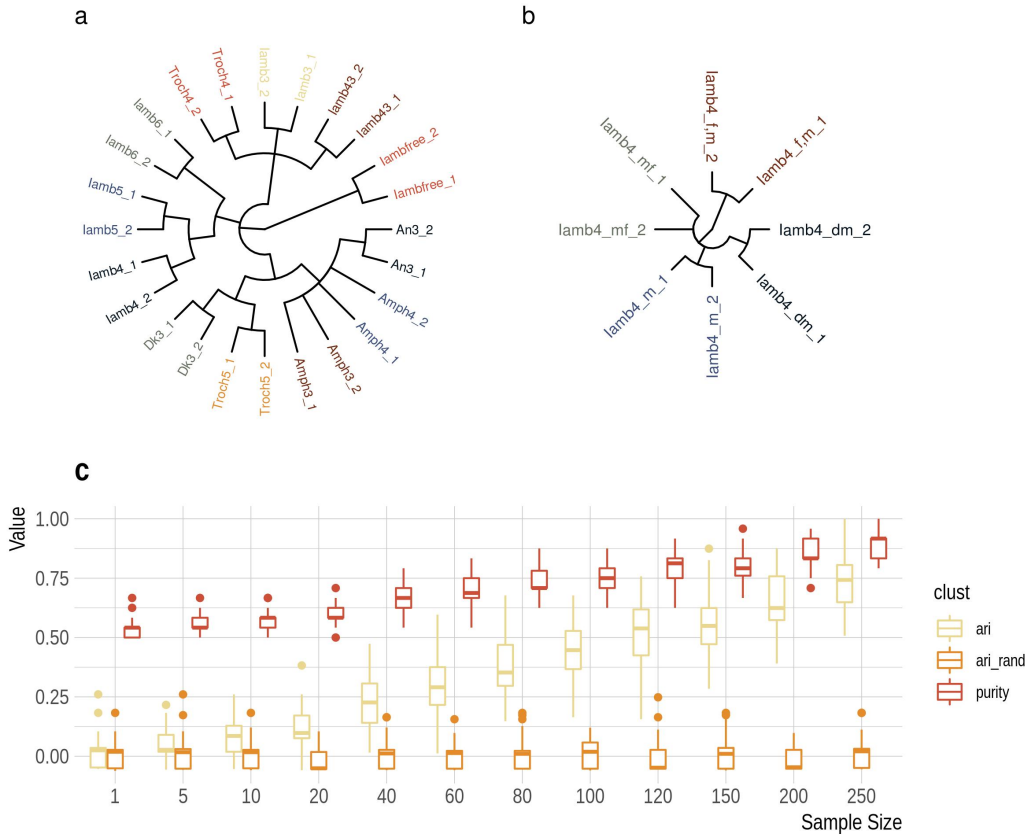
[27] Solomon Kullback and Richard A. Leibler, „On Information and Sufficiency," *The Annals of Mathematical Statistics* 22, 1. sz. (1951): 79–86, http://doi.org/10.1214/aoms/1177729694.

[28] Joseph Felsenstein, „Confidence Limits on Phylogenies: An Approach Using the Bootstrap," *Evolution* 39, 4. sz. (1985): 783–791, http://doi.org/10.1111/j.1558-5646.1985.tb00420.x.

[29] Florian Cafiero and Jean-Baptiste Camps, „Why Molière Most Likely Did Write His Plays," *Science Advances* 5, 11. sz. (2019): 2375–2548, http://doi.org/10.1126/sciadv.aax5489.

[30] Lawrence J. Hubert and Phipps Arabie, „Comparing Partitions," *Journal of Classification* 2, 1. sz. (1985): 193–218, http://doi.org/10.1007/BF01908075.

 DOI 10.31400/dh-hun.2021.5.3145



2. ábra. a) A klaszterezésbeli megegyezéseket mutató, a „többségi szabályt” alkalmazó konszenzusfa (Jensen–Shannon divergencia, teljes kapcsolat) – 100 iteráció, 9 versmérték 2 véletlenszerű mintája, mintánként 200 vers. b) A jambikus variánsok klaszterezése közötti megegyezést mutató konszenzusfa, mintánként 100 vers. c) A hierarchikus klaszterezés teljesítménye CP-ben és ARI-ban az „alapigazsággal” (metrikai címkék) szemben, a véletlenszerűen hozzárendelt klaszterekhez képest. Ugyanazokon a versmértékeken futtatva (mindegyik esetben legalább 500 verssel), mintaméreteként 100 iteráció.

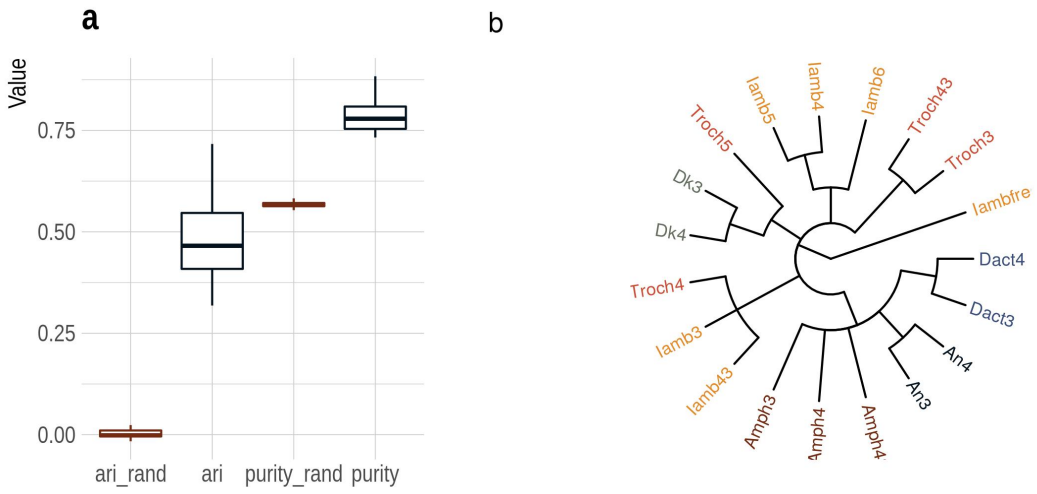
a klaszterezés minden egyes esetére. Mint várható volt, a klaszterezés pontossága a mintában lévő versek számával együtt nő, egészen az ARI=0.73 és a CP=0.90 mediánig *(2c. ábra)*. Ugyanakkor fontos, hogy a nem véletlenszerű klaszterezés hamar észrevehető, és a versmértékekben néhány szemantikus minta felismerhető már mintánként 20–40 versnél.

4.2. A versmértékek közötti hasonlóságok

A *2c. ábra* konszenzusfája rámutat az egyes családokhoz tartozó metrikai formák jelentésbeli kapcsolataira is. Több jambikus méter hajlamos egy klaszterbe kerülni, és ugyanez történik azokkal a hármas formákkal, amelyek világosan egy csoportot alkotnak (az amphibrachus formák közötti minimális eltéréssel). Nem igazán van „alapigazság” arra nézve, hogy hogyan kellene a költői formáknak egymáshoz viszo-

nyulniuk jelentéstani szempontból, kivéve a történeti alkalmazásukra és a hasonló eredetükre vonatkozó néhány megfigyelést. Ugyanakkor számíthatunk rá, hogy a jelentés legalább részben a metrikai családokhoz kötődik (pl. jambikus vagy trochaikus versek), mivel ezek nagyon hasonló ritmikus és grammatikai határokat szabnak a nyelvnek.[31] Néhány esetben az egyes családba tartozó versmértékek történeti kötődéseikben is hasonlóak. Például a legtöbb jambikus mértéket kezdetben magas presztízsű műfajokban használták: a négyes jambust [Iamb-4] az ódákban, az ötös jambust [Iamb-5] a drámában, a hatost [Iamb-6] az elégiában. Ugyanakkor a trochaikus formákat gyakran az „elitet" alkotó jambikus vers ellenpontjaként fogták fel; sőt bizonyos ritmikai vonásuk átfedésbe került a szóbeli hagyománnyal, ami folklórimitációkként határozta meg használatukat, és ennek megfelelő asszociációs mező alakult ki körülöttük.

A feltételezett „családi hatás" tesztelésére egy igen konzervatív kísérletet terveztünk: mivel az elérhető versmértékek száma nem egyezik meg családonként, csak azokat a családokat vettük figyelembe, amelyekben legalább két gyakori versmérték van (> 400 vers). Ezután családonként 20-szor véletlenszerűen veszünk két versmértéket; egy versmértékkészletben 300 véletlenszerűen kiválasztott vers szerepel; ugyanúgy számítjuk ki a klasztereket, mint ahogyan azt leírtuk a 4.1. pontban, ám ezúttal a formák klaszterezését családjuk „alapigazságával" szemben igazoljuk (jambus, trocheus stb.). A folyamatot 100-szor megismételjük a 20 versmértékkészlet mindegyikén. Ennek során jegyezzük az átlagos ARI- és CP-értékek eloszlását minden mintavételezett versmértéknél a véletlenszerű csoportosításhoz mérve.



3. ábra. a) A versmértékek kapcsolatának erőssége a nekik megfelelő családdal. A klasztereket családonként egyenlő számú versmértékkel számolva (k=6, n=12, mintaméret = 300). 100 iteráció minden 20 metrikai készletben. b) Versmértékek közötti stabil szemantikai kapcsolatokat bemutató konszenzusfa (k=6, n=19, mintaméret = 300, fák = 100).

[31] Mikhail Leonovich Gasparov and Marina Tarlinskaja, „The Linguistics of Verse," *The Slavic and East European Journal* 52, 2. sz. (2008): 198–207.

 DOI 10.31400/dh-hun.2021.5.3145

Habár az így létrejött klaszterezés teljesítménye talán nem tűnik magasnak (a medián ARI 0.44 körüli, a CP – 0.76), az értékek mégis elegek ahhoz, hogy megerősítsék, legalábbis bizonyos fokig, hogy a versmértéken belüli kapcsolatokat [vagyis az azonos mértékhez tartozó minták kapcsolata, vö. 4.1. alrész – *a szerk.*] a metrikai családok motiválják. Hogy jobban szemléltessük ezt a hatást, 100 klaszterezés eredményéből kiszámítottunk egy konszenzusfát a családonkénti versmértékek számának bármilyen korlátozása nélkül (*3b. ábra:* a jambusok, a daktilusok, az anapesztusok és néhány trocheus rendszerint következetesen egy klaszterbe kerül; a hármas formák szemantikája valamiképpen diffúz marad, de még így is egy klasztert formálnak egymással).

A „rossz attribúciók" esetei szintén informatívok lehetnek, és összhangban vannak a tárgyhoz kapcsolódó tudományos munkákkal. A hármas jambus [Iamb-3] és a négyes trocheus [Trochee-4] hasonlósága jól ismert: mindkettő a 18. századi anakreóni versből ered, és sok variánsban megegyezik a „dal" ["song"] szemantikájával.[32] A négyes és hármas jambus [Iamb-43] a különböző verslábú sorok rendszeres módosulásával szintén a lírai dalból és a balladából fejlődött ki, és a lírikus-epikus költészethez kapcsolódik.

4.3. A jelentésmező időbelisége

Ideje elhagynunk annak a megfigyelőnek a pozícióját, aki „visszatekint" az egész hagyományra. A *2.* és a *3. ábra* világossá teszi, hogy a klaszterezést bizonyos mértékig erősíti az eltérő időben feltűnő versmértékek közötti különbség. Mivel az LDA algoritmus a dokumentumokon belül a szavak együttes előfordulását használja fel, természetszerűen eredményez olyan szócsoportokat, amelyek különböző időből származnak (például a „nép" témája, a „szovjet" téma vagy a naturalisztikus háború témája). Ez határozza meg a divergenciaszámításokat is: a szovjet témának közel nulla a valószínűsége a 18–19. századi szövegekben. Továbbá ezt láthatjuk abból is, ahogyan a dolnik és az ötös trocheus [Trochee-5] következetesen egy klaszterben tűnik fel – két nagyon népszerű 20. századi forma, amelyek korábban ritkán fordultak elő.

A „szabad" jambus jó példa lehet erre. Ezt a formát a változó verslábhosszok jambikus sorainak a szabálytalan módosulásai alkották, és majdnem kizárólag sajátos műfajokban alkalmazták: költői episztolákban, fabulákban és epigrammákban, valamint teljes mértékben elhagyták a használatát az 1850-es évek után. A szabad jambusban írt 1200 versben csak két téma fordul elő, amelyek együttesen 20% valószínűséggel bírnak (állatok és kommunikáció). Az elnevezése ellenére ez a versmérték megfagyott az időben, műfajok kombinációja nyomja rá a bélyegét, ezáltal két innen származó mintát nem nehéz egy klaszterbe rendezni. Röviden, a szemantikai absztrakciónk nem eléggé absztrakt ahhoz, hogy figyelmen kívül hagyjuk a kronológiai különbségeket.

Az idő ugyanakkor nem érvényteleníti a metrikai mező általános jelenlétét; végső soron a metrikai formák aszinkron fejlődése és divatbeli változásai alakítják az észlelt különbségeket, és határozott korokhoz kötik őket (a négyes jambus [Iamb-4-fm] az 1820-as, 1830-as évek „aranykorához", a hármas daktilus [Dactyl-3] az 1850–1880-as évek polgári és politikai érzületéhez, a dolnik a modernista költészethez). A korpusz

[32] Gasparov, *Metr i smysl.*

ellenőrzése kronológiai szempontból nemcsak a mezőhatás kisebb léptékű tesztelése szempontjából hasznos, hanem lehetőségeket teremt a versmérték-jelentés kapcsolat működésének időbeli vizsgálatához is. Mindezt azonban csak röviden érintjük, mivel ez külön probléma, amely célzott kísérleteket és adatokat érdemel.

Először is azt szeretnénk látni, hogy a versmértéken belüli szemantikai hasonlóságok jelen vannak-e, ha minden versminta ugyanabból az időből származik. Ennek érdekében a korpuszunkat 30 éves szakaszokra osztjuk fel (kizárva a 18. századot, mert ott a népszerű versmértékek változatossága nem nagy). Minden egyes időkeretből vesszük annak hat leggyakoribb versmértékét, és jelezzük az átlagos ARI-értékeket *(1. táblázat)*. Ezek az értékek nem hasonlíthatók közvetlenül össze, mivel különböző versmértékekre és mintaméretekre vonatkoznak (az alsó határ időszakonként a legritkább versmértékű szövegek számának a fele), de elég ahhoz, hogy rámutassunk: a mezőhatás észlelhető korlátozott időkereteken belül is, sőt bizonyos időszakokban kisebb mintákban is, mint várható (vö. *2. ábra*).

Másodszor arra is fel lehet használni a kronológiai információt, hogy kérdéseket fogalmazzunk meg a mezőviselkedéssel és a szemantikai felhalmozással kapcsolatban. Ha egy versmérték jelentésmezőjét történeti és nem organikus jelenségnek tekintjük, akkor arra számíthatunk, hogy a versmérték és a jelentés közötti kapcsolat gyengül az idő múlásával. Egészen pontosan arra számítunk, hogy eltérést találunk a 19. század eleji és végi költészet klaszterezésének értékelésében. Ez megerősítené a versmérték szemantikájának történetiségét, és feloldaná a forma és a műfaj közötti merev kapcsolatot, amelyet a normatív esztétika erőltetett a költészetre. Mivel az osztályozásunk függ a minta méretétől, egyszerűen elfelezzük a 19. századi adatokat, és megfigyeljük az ARI-értékek eloszlását: a klaszterezést szigorúan csak ugyanazon a versmértékkészleten és ugyanakkora mintával végezzük el mindkét időbeli csoport esetében.

Kiderült, hogy a két periódus közötti különbség jelentős *(2. táblázat)*. A 40 versből álló mintákban – a két csoport esetében a lehető legnagyobb mintaméret – jobban elkülöníthetők a versmértékek a 19. század első felében, amely tudatosabb versmértékhasználatot jelez. Ha a 19. század második felében növeljük a minta méretét, akkor az átlagos klaszterpontosság megjósolhatóan emelkedik (mintaméret=100, ARI=0.43), ami a versmértékekben a szemantikai felhalmozódás folyamatára mutat – azaz szemantikailag diffúzabbak lesznek, de nem válnak felismerhetetlenné.

1. táblázat. Versmérték-jelentés kapcsolat különböző időkeretekben, 100 iteráció periódusonként

Period	Median ARI	Poems per sample	Meters
1800-1829	0.51	30	I4 If T4 I6 I5 I3
1830-1859	0.47	50	I4 T4 I6 I5 If Amph3
1860-1889	0.23	30	I4 T4 I6 I5 An3 An43
1890-1919	0.77	270	I4 I5 T4 I6 An3 T5
1920-1949	0.58	250	I4 I5 T4 T5 Dk3 An3

 DOI 10.31400/dh-hun.2021.5.3145

2. táblázat. A 19. század első és második fele, összehangolt klaszterezés. 1000 mintavételező iteráció minden periódusban. Jelentős a különbség az ARI-k eloszlásában (t-test, t=19,6, $p < 0{,}0001$)

Period	Median ARI	Poems per sample	Meters
1800-1849	0.48	40	I4 T4 If I6 I5 Amph4
1850-1899	0.33		

4.4. A témakifejezés és a versmérték gyakorisága

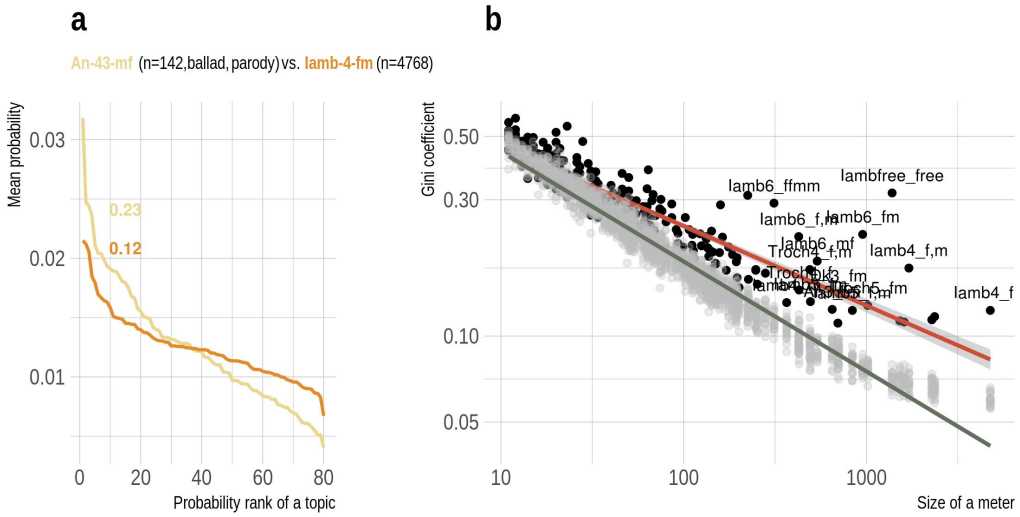
A dolgozat elejétől fogva ijesztő kérdés árnyékolja be az eredményeket: mi van, ha az egész mezőhatás abból az egyszerű tényből ered, hogy a metrikai formák népszerűsége változó? Könnyű bármilyen témakonfigurációra következtetni a négyes jambus, és nehezebb a hármas jambus esetében, míg a hármas trocheus [Trochee-3-dm] esetében szinte nem is lehet. A jelentésmező így talán a mintavételi hibáknak, az időbeli különbségeknek és az azonos metrumú verseket hasonlóan olvasó irodalmárok torzító visszaigazolásainak a kombinációjaként született.

Sőt ennél is bonyolultabb a helyzet, mivel az irodalmárok szerint számítanunk kell arra is, hogy egy mező „jellegzetessége" természeténél fogva csökken, ha nő egy versmérték gyakorisága. Egyszerűen a ritka formákban nincs helye a szemantikai variációnak. Ennélfogva feltételezzük, hogy 1) egy versmérték kifejezésének erőssége és gyakorisága között lineáris kapcsolatnak kell lennie; 2) amennyiben a jelentéstulajdonítás a versmértékek esetében véletlenszerű, másmilyen tendenciát kellene látnunk.

Ahhoz, hogy felmérjük, mennyire „jellegzetes" egyetlen versmértéken belül a jelentés, felhasználhatjuk a témavalószínűségek eloszlásának grafikonjait, és megnézhetjük őket az „egyenlőtlenség"[33] perspektívájából. A kevésbé népszerű versmértékek esetében arra számítunk, hogy kevesebb jellemző témát találunk, mint a nagyon gyakori formákban. Erre mutat példát a *4a. ábra:* a témavalószínűségeket a két metrikai variáns (a négyes jambus [Iamb-4-fm] és a négyes–hármas anapesztus [An-43-mf]) összes verséből hoztuk létre, és aszerint rendeztük el, hogy összességében hogyan járulnak hozzá a metrum szemantikájához. Minden egyes így létrejött grafikonra ki lehet számítani az úgynevezett Gini-együtthatót – amelyet eredetileg arra terveztek, hogy mérje a nemzeti egyenlőtlenséget a vagyoneloszlásban. A Gini 1-es értéket vesz fel, amikor egy disztribúció ördögi módon egyenlőtlen (egyetlen témának 100% a valószínűsége), és 0-t, amikor tökéletesen egyenlő (mind a 80 téma valószínűsége 1,25%). Nyilvánvaló, hogy ez az együttható képes megragadni, hogy mennyire koncentrált a versmérték szemantikája, legalábbis viszonylagosan; a Gini abszolút értékeit azonban befolyásolnák az LDA alapbeállításai (a 0,1 alpha magasabb fokú egyenlőtlenséget feltételez egy vers témavalószínűségeiben, mint például 0,5 alpha).

Hogy igazoljuk a mező kiterjedtségét és (nem-)véletlen természetét, először kiszámoljuk a Gini-együtthatókat minden metrikai variánsra, amely legalább 10-szer előfordul a korpuszban. Majd elvégezzük ugyanezt a számítást a szövegek újraelosztását

[33] Azaz, hogy a témák mennyire egyenlően oszlanak meg a verstípusokban – *a szerk.*



4. ábra. a) A témaegyenlőtlenségbeli különbség (Gini-együttható) az általánosan elterjedt négyes jambus (Iamb-4-fm) (0,12) és a négyes és hármas anapesztus (An-43-fm) ritka formája között. b) A szemantikai egyenlőtlenség csökkenése: a versmérték gyakorisága (fekete) versus a véletlenszerűen újraelosztott versek (szürke) alapján, 20 független újraelosztás. A két lineáris modell lejtői különbözőek (-0,28, -0,39), és a modell nagyobb variációt ír le az újraelosztott (R^2 = 0,96), mint az empirikus (R^2 = 0,81) adatban.

követően: minden egyes *n* gyakoriságú metrikai forma esetén azonos számú, *n* versnyi véletlenszerű mintát veszünk (visszatevés nélkül) a korpuszból. Végül minden verset véletlenszerű módon áthelyezünk üres „versmértékkosarakba" – ezt az újraelosztást 20 különböző alkalommal végezzük el. Ha a mezőből nem lehet véletlenszerűen mintát venni, akkor a két pontcsoport között észrevehetőnek kell lennie az eltérésnek, ahogyan az egyenlőtlenség a mintanagysággal korrelál.

A *4b. ábra* az egyenlőtlenségben felfedezhető különbségeket mutatja a véletlenszerűen összesített versek és a valódi metrikai formák közötti logaritmikus skálán. A várakozásoknak megfelelően a versmérték gyakorisága mentén csökken a szemantikai egyenlőtlenség (azaz több téma is jellemző az adott csoportra), ám árulkodók a kivételek, amelyek a gyakori metrikai formákban (mindenekelőtt a szabad jambusban) is koncentrált szemantikai mezőt jeleznek. Másfelől az egyenlőtlenség a véletlenszerűen újraosztott adatban gyorsabban csökken, és sok valódi versmértéket hagy a vonal felett. Ez arra utal, hogy míg a nagyon ritka formákban véletlenszerűen *talán* előfordulhat hasonló szintű egyenlőtlenség, nincs okunk erre számítani a jelentésmező egészében. Nagyon is valószínűtlen, hogy még a mindig semlegesként és általánosan elterjedtként számontartott négyes jambus szemantikai görbéjét is képesek legyünk véletlenszerű mintavétellel létrehozni.

 DOI 10.31400/dh-hun.2021.5.3145

5. Diszkusszió

A dolgozatban megmutattuk, hogy önmagában a témára vonatkozó információ alapján felismerhető egy versforma. Az egyazon versmértékben írt költemények szemantikailag hasonlók maradnak egymáshoz; a különböző versmértékek pedig gyakran akkor mutattak stabil viszonyt, ha egy családból származnak. Az osztályozás pontosságának történeti különbsége szintén azt sugallja, hogy a metrikai formákban szemantikai felhalmozódás történik, valamint a versmérték „jelentése" szétszóródik az idő folyamán anélkül, hogy felismerhetetlenné válna. Ezek a felfedezések, úgy hisszük, megerősítik a jelentésmező elméletét és legfőbb feltételezéseit, legalábbis egy általános szinten.

A jövőben a metrikai mező hatását jobban megértjük majd a kulturális evolúció keretében,[34] ami lehetőséget biztosít arra, hogy jobban kifejtsük azt is, miként gondolkodunk a történeti folyamatokról és a kulturális átadásról *(cultural transmission)*. A „kulturális evolúció" egy kialakulóban lévő tudományterület, amely a kulturális információ (amit általában úgy határoznak meg, hogy minden olyan információ, amelyet szociális tanulással szerzünk meg) változását, fennmaradását és szétszóródását tanulmányozza. Ez a keret sokféle diskurzuson és területen átívelő kutatást fog át: alkalmazták a régészeti leletek,[35] népmesék[36] és középkori kéziratok kulturális törzsfejlődésének rekonstruálására;[37] az emberi tanulás és a kultúra felgyülemlő erőihez való hozzájárulás megértésére;[38] a populáris zenében történő újítások mértékének vizsgálatára;[39] a nyelvfejlődés makromintáinak tanulmányozására[40] vagy a kulturális információ szétszóródásában és fennmaradásában a népesség méretének szerepét vizsgálva .[41]

[34] Alex Mesoudi, *Cultural Evolution: How Darwinian Theory Can Explain Human Culture and Synthesize the Social Sciences* (Chicago: University of Chicago Press, 2011), https://doi.org/10.7208/chicago/9780226520452.001.0001; Oleg Sobchuk, *Charting Artistic Evolution: An Essay in Theory*. PhD-thesis (Tartu: University of Tartu Press, 2018).

[35] Michael J. O'Brien and R. Lee Lyman, „Evolutionary Archeology: Current Status and Future Prospects," *Evolutionary Anthropology: Issues, News, and Reviews* 11, 1. sz. (2002): 26–36, http://doi.org/10.1002/evan.10007.

[36] Sara Graça da Silva and Jamshid J. Tehrani, „Comparative Phylogenetic Analyses Uncover the Ancient Roots of Indo-European Folktales," *Royal Society Open Science* 3, 1. sz. (2016), http://doi.org/10.1098/rsos.150645.

[37] Adrian C. Barbrook et al., „The Phylogeny of *The Canterbury Tales,*" *Nature* 394, 839. sz. (1998), http://doi.org/10.1038/29667; Joris van Zundert. „Computational Methods and Tools," in Philipp Roelli, ed., *Handbook of Stemmatology: History, Methodology, Digital Approaches,* De Gruyter Reference, 292–356 (De Gruyter, 2020), https://doi.org/10.1515/9783110684384-006.

[38] Claudio Tennie, Josep Call and Michael Tomasello, „Ratcheting Up the Ratchet: On the Evolution of Cumulative Culture," *Philosophical Transactions of the Royal Society B: Biological Sciences* 364, 1528. sz. (2009): 2405–2415, http://doi.org/10.1098/rstb.2009.0052.

[39] Mauch et al., „The Evolution of Popular Music."

[40] Russell D. Gray and Fiona M. Jordan, „Language Trees Support the Express-Train Sequence of Austronesian Expansion," *Nature* 405, 6790. sz. (2000): 1052–1055, http://doi.org/10.1038/35016575; Remco Bouckaert et al., „Mapping the Origins and Expansion of the Indo-European Language Family," *Science* 337, 6097. sz. (2012): 957–960, http://doi.org/10.1126/science.1219669.

[41] Joseph Henrich, „Demography and Cultural Evolution: How Adaptive Cultural Processes can Produce Maladaptive Losses: The Tasmanian Case," *American Antiquity* 69, 2. sz. (2004): 197–214, http://doi.org/10.2307/4128416; Adam Powell, Stephen Shennan and Mark G. Thomas, „Late

Nem túlzás azt mondani, hogy minden új vers korábbi versekből ered. Sőt azok legtöbbször az imitáció termékei: legalábbis rendkívül ritka, amikor egy költő teljesen meg tud szabadulni a hagyománytól, vagy egyedül képes megalkotni egy teljes verselési rendszert. Ha mégis megtörténik, nagy rá az esély, hogy ezek az egyéni erőfeszítések nem lesznek hosszú életűek, egyszerűen azért, mert nem lesz elég követőjük. A költői formák kitartók és konzervatívok: az olyan dolgok, mint a jambikus pentameter, a rímelés vagy a szonett mintázata, századokon át képesek túlélni. Ez azt jelenti, hogy az új verseknek elődeikkel nagyon sok közös formai jellemzőjük van – mint például a versmérték –, hatékonyan ismétlik a korábban alkalmazott formát. Érvelhetnénk amellett, hogy semmi sem állíthatja meg a lírai önkifejezés individualizált modern hagyományának költőjét abban, hogy teljesen szabadon használjon egy metrikai formát, függetlenül annak szemantikai vonatkozásaitól; de láthatjuk, hogy nem ez a helyzet.

A versmértékekre és a versformákra tekinthetünk úgy, mint amelyek a kultúra „TRIM"-jeihez (Transmission Isolating Mechanism – 'átviteli elszigetelő mechanizmus')[42] hasonlóan viselkednek. Ezek a mechanizmusok olyan kondíciók (házassági hagyományok, háztartási szerveződés stb.), amelyek fenntartják az információátvitel „vertikális" szintjét (a szülőktől az utódokig) a kultúrában, amit általában a „horizontális" kapcsolatok (kortársaktól kortársakig) kiterjedt területének tekintenek. Hasonló módon korlátozzák a versmértékek a versek szemantikai lehetőségeit, és a jelentéselőállítást homályos, mégis határozott utakra terelik. Ez biztosítja, hogy a modern költészettörténetekben a versmértékek „gyenge műfajokként" viselkednek, és jellemzők bővülő készletét termelik újra azokban a versekben, amelyek szintén hasonló formai eredettel bírnak.

Miért kell a költészetben ennek a formai elszigetelődésnek egyáltalán megtörténnie? Az egyik kézenfekvő válasz, hogy a versmérték képes hatékony mnemotechnikai rendszerként működni, amely a nyelvet egy magasabb szintű mintázatba helyezi és növeli a megjegyezhetőséget.[43] A költői formák a szóbeli hagyományból erednek, ami nagyon sok formális működésen alapult (versmérték, rím, nyelvi formulák, történetek stb.), ezek behatárolták, hogyan lehet létrehozni és újramondani egy szöveget úgy, hogy elősegítse a memorizálását és az átvitelét.[44] A versmérték emlékeztető ereje nyilvánvalóan számít az írásos tradícióban is. Nem csupán egy forma, amelyre emlékeznek és amelyet újraalkotnak; már azáltal, hogy használatban van, a versmérték továbbcipeli örökké növekvő batyuját. Az orosz költészet sok fordulatot és tudatosan irányított forradalmat foglal magába az egyes versmértékekkel kapcsolatos elvárások-

Pleistocene Demography and the Appearance of Modern Human Behavior," *Science* 324, 5932. sz. (2009): 1298–1301, http://doi.org/10.1126/science.1170165.

[42] William H. Durham, „Advances in Evolutionary Culture Theory," *Annual Review of Anthropology* 19, 1. sz. (1990): 187–210, http://doi.org/10.1146/annurev.an.19.100190.001155; Jamshid Tehrani and Mark Collard, „Do Transmission Isolating Mechanisms (TRIMS) Influence Cultural Evolution? Evidence from Patterns of Textile Diversity Within and Between Iranian Tribal Groups," in Roy Ellen, Stephen J. Lycett and Sarah E. Johns, eds., *Understanding Cultural Transmission in Anthropology: A Critical Synthesis,* 148–164 (Oxford: Berghahn Books, 2013).

[43] Gronas, *Cognitive Poetics and Cultural Memory.*

[44] David C. Rubin, *Memory in Oral Traditions: The Cognitive Psychology of Epic, Ballads, and Counting-Out Rhymes* (New York–Oxford: Oxford University Press, 1995).

 DOI 10.31400/dh-hun.2021.5.3145

kal szemben (mert ezek az elvárások léteztek), ugyanakkor úgy tűnik, hogy senki nem menekülhet meg igazán az időmértékes vers mnemonikus zsarnokságától.

Úgy gondoljuk, a szemantikai mező jelen lesz (nagyobb vagy kisebb mértékben) bármely költészeti hagyományban, bármilyen verselési rendszeren is alapuljon, amely lehetővé teszi a sajátos és stabil versformák létrejöttét az idők során. A témamodellek absztrakt, származtatott eszközök készletét nyújtják nekünk (osztályozási pontosság, egyenlőtlenség stb.), amelyek mentén különböző nyelvek és hagyományok válnak összehasonlíthatóvá. Ez egyúttal hozzáférést biztosít az irodalomtörténet általános kérdéseihez is: hogyan „buknak el" a költői műfajok az idők során, milyen mértékig maradnak felismerhetők a versmértékek (ha egyáltalán), hogyan történik az új formák feltalálása, vagy hogy mi a szerepe az egyéni költőknek és egyedi verseknek a szemantikai mező alakításában.

Fordította: Vásári Melinda

Weak Genres: Modeling Association Between Poetic Meter and Meaning in Russian Poetry

This paper aims to formalize an established theory in versification studies known as "semantic halo of a meter" which states that different metrical forms in modern poetry accumulate and retain distinct semantic associations. We use LDA topic modeling on a large-scale corpus of Russian poetry (1750-1950) to represent each poem in one topic space and then proceed to represent each meter as a distribution of aggregated topic probabilities. Using unsupervised classification and extensive sampling we show that robust form-meaning associations are present both within and between metrical forms: two samples of the same meter tend to appear most similar, while two metrical forms of the same family tend to group together. This effect is present if corpus is controlled for chronology and is not an artifact of population size. We argue that similar approach could be used to align and compare semantic halos across languages and traditions to give meaningful general-level answers to questions of literary history.

Keywords:
poetry, semantics, meters, topic modeling, clustering

Köszönetnyilvánítás

Artjoms Šeļát a „Large-Scale Text Analysis and Methodological Foundations of Computational Stylistics" (NCN 2017/26/E/ HS2/01019) elnevezésű projekt keretében a Polish National Science Centre támogatta. Szeretnénk megköszönni a két névtelen bírálónak, hogy figyelmesen olvasták és javították a dolgozatunkat. Köszönjük Joanna Byszuknak, Maciej Edernek, Antonina Martynenkónak, Vera Polilovának és Oleg Sobchuknak a hozzájárulásukat, segítségüket és támogatásukat.

Függelék A

A kódok és az adatok hozzáférhetősége

A Document-Term Matrix, a feldolgozási lépések, a végső modellek és a teljes elemzés szabadon hozzáférhető: `https://github.com/perechen/semantic_halo_rus`. Az elemzés során az *R* 4.0.2. verzióját használtuk, az LDA-implementációt a *topicmodels,*[45] a modell kimenetének kezelését a *tidytext,*[46] a számításokat a *phylentropy*[47] és a *ineq,*[48] a fák megrajzolását a *ggtree,*[49] az ábrákat pedig a *patchwork*[50] csomaggal hoztuk létre. Az ábrákhoz a MonikeMedium színpalettát a *ghibli* csomag szolgáltatta.[51]

[45] Bettina Grün and Kurt Hornik, „Topicmodels: An R Package for Fitting Topic Models," *Journal of Statistical Software* 40, 13. sz. (2011): 1–30, `http://doi.org/10.18637/jss.v040.i13`.

[46] Julia Silge and David Robinson, *Text Mining with R: A Tidy Approach* (O'Reilly Media Inc., 2017), hozzáférés: 2021.12.07, `https://www.tidytextmining.com/`.

[47] Hajk-Georg Drost, „Philentropy: Information Theory and Distance Quantification with R," *Journal of Open Source Software* 26, 3. sz. (2018), `https://doi.org/10.21105/joss.00765`.

[48] Tze-I Yang, Andrew Torget and Rada Mihalcea, „Topic Modeling on Historical Newspapers," in Kalliopi Zervanou and Piroska Lendvai, eds., *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities,* 96–104 (Portland OR: Association for Computational Linguistics, 2011).

[49] Guangchuang Yu et al., „Two Methods for Mapping and Visualizing Associated Data on Phylogeny Using Ggtree," *Molecular Biology and Evolution* 35, 12. sz. (2018): 3041–3043, `http://doi.org/10.1093/molbev/msy194`.

[50] Thomas Lin Pedersen, *patchwork: The Composer of Plots* (2020), hozzáférés: 2021.12.07, `https://CRAN.R-project.org/package=patchwork`.

[51] Ewen Henderson, Danielle Desrosiers and Michael Chirico, *ghibli: Studio Ghibli Colour Palettes* (2020), hozzáférés: 2021.07.12, `https://CRAN.R-project.org/package=ghibli`.

 DOI 10.31400/dh-hun.2021.5.3145

Függelék B

B.3. táblázat. A dolgozatban használt legfőbb versmértékek. 1 – a verslábon belül erős pozíciót jelöli (valószínű a hangsúly), 0 – gyenge. A zárójeles szótagok vagy jelen vannak, vagy nem (a sor végén; dolniknál sor közben is). A klasszikus versmértékekre angol példákat adunk.

Meter	Type	Foot	Example	Comment
Iamb	binary	01	01|01|01|01|01(00) Thus was|I, slee|ping, by | a bro|ther's hand Of life,| of crown,| of queen,| at once| dispatch'd	Iambic Pentameter
Trochee	binary	10	10|10|10|1(00) Tell me| not in| mournful | numbers, Life is | but an | empty | dream	Trochaic Tetrameter
Dactyl	ternary	100	100|100|100|1(00) Brightest and| best of the | sons of the| morning	Dactylic Tetrameter
Amphibrach	ternary	010	010|010|010|01(00) Oh, hush thee,| my baby|, thy sire was | a knight Thy mother | a lady| both lovely | and bright	Amphibrachic Tetrameter
Anapest	ternary	001	001|001(00) He is gone | on the moun|tain He is lost | to the for|est	Anapestic Dimeter
Dolnik	/	/	(00)1(0)01(0)01(00)	3-ictus Dolnik based on number of stressed positions (3) but unstressed syllable interval is limited to 1-2 syllables

B.4. táblázat. A megkülönböztető témák (a szavakat lefordítottuk) három versmértékben Gasparov leírásaihoz hasonlítva. Az átlagtól leginkább eltérő top 10 témát listáztuk. A korábban meghatározott jelentésmező szempontjából releváns témákat kiemeltük.

Meter	Halo (Gasparov)	Topic	Top words
Trochee-5-fm	Night, Landscape, Love, Death, Road	69	**to know, to live, to be, to die, nothing**
		41	war, to go, soldier, battle, bullet
		61	**goodbye, last, to go (away), hand, parting**
		25	**wind, steppe, sand, grass, desert**
		66	**garden, green, leaf, branch, linden**
		45	**train, wheel, smoke, to fly, wind**
		38	window, house, wall, room, table
		39	**water, river, shore, to swim, lake**
		31	**to go, path, road, to cross, leg**
Trochee-3-fm	Song, Road, Nature, Yearning, Love, Death	76	**to sing, song, nightingale, voice**
		77	matter, take, give, comrade, most
		23	**red, to go, oi, white, "ka" (folksong love topic)**
		43	**snow, white, ice, winter, snowy**
		51	door, house, enter, window, wait
		51	**woods, pine, green, tree**
		10	**wind, leaf, autumn, rain, autumn**
		21	**dream, to dream, night, to wake, morning**
		22	night, darkness, murk, dark
		31	**to go, path, road, to cross, leg**
Iamb-4-dm	dangerous movement through space / love	62	**city, tower, wall, stone**
		59	**horror, death, evil, blood**
		1	**star, world, sky, earth, abyss**
		6	**shade, dream, ghost, pale**
		4	**soul, dream, beauty, world, power**
		68	**hour, wait, to come, soon, or**
		12	god, temple, tsar, before, world
		61	**goodbye, last, to go (away), hand, parting**
		80	**city, road, house, light, to go**
		50	poem, write, poet, book, word

B.5. táblázat. Az Ajusted Rand Index mediánértékei a versmértéken belüli klaszterezéshez *(2c. ábra)* különböző mintában (ARI@*n* vers mintánként), LDA-modelleket használva változó *k* számú témával. Az „ARI family" oszlop ARI-értékeket tartalmaz a versmértékek közötti hasonlóságtesztekhez *(3a. ábra)*. A klaszterezés erős teljesítménye különböző modellekben azt mutatja, hogy a témák száma kevéssé van hatással a kísérleti eredményekre, és így meglepő módon nem igazán befolyásoló tényező. A 20 témával ellátott LDA az egyik legmagasabb teljesítményt mutatja a mintánkénti 250 versnél, ami még inkább kiemeli a szemantikai információ redukciójának hatékonyságát. Ugyanakkor van egy éppenhogy észrevehető kompromisszum a lokális (versmértéken belüli) és a globális (versmértékek közötti) felismerés között (úgy tűnik, a 80 és 100 témát magukba foglaló modellek szolgáltatják a legkiegyensúlyozottabb teljesítményt). További teszteket hajtottunk végre az eredeti Document-Term Matrixon tanult LDA-modelleken (a kevésbé gyakori szavaknak a gyakoribb szemantikai szomszédjaikra való cserélése nélkül – lásd a „w/o replacement" részt).

DTM	k	ARI@10	ARI@60	ARI@100	ARI@150	ARI@250	ARI family
w/ replacement	10	0.07	0.28	0.47	0.55	0.62	0.33
	20	0.09	0.28	0.45	0.57	0.76	0.40
	40	0.08	0.31	0.44	0.60	0.71	0.36
	60	0.09	0.30	0.41	0.55	0.70	0.43
	80	0.09	0.29	0.44	0.56	0.73	0.43
	100	0.09	0.31	0.45	0.60	0.76	0.39
	120	0.04	0.30	0.45	0.54	0.70	0.48
	150	0.08	0.28	0.42	0.57	0.70	0.41
	200	0.04	0.28	0.42	0.54	0.70	0.42
w/o replacement	20	0.08	0.34	0.46	0.55	0.70	0.32
	80	0.09	0.31	0.41	0.56	0.76	0.32
	150	0.08	0.27	0.45	0.60	0.73	0.45

 DOI 10.31400/dh-hun.2021.5.3145