VALÓS TÉRBEN – AZ ONLINE TÉRÉRT

Networkshop 31: országos konferencia

2022. április 20–22.
Debreceni Egyetem

Szerkesztette: Tick József, Kokas Károly, Holl András

HUNGARNET Egyesület
Budapest, 2022

NETWORKSHOP 2022

A kötet megjelenését támogatta az
Energiaügyi Minisztérium

Szerkesztette: Tick József, Kokas Károly, Holl András

Tipográfia és tördelés: Vas Viktória

Networkshop
2022. április 20–22. Debreceni Egyetem, konferencia előadásainak közleményei

ISBN 978-615-82243-0-7
DOI: 10.31915/NWS.2022

Kiadja a HUNGARNET Egyesület
az MTA Könyvtár és Információs Központ közreműködésével
Budapest
2022

Borítókép: freepik.com

TARTALOMJEGYZÉK

DOI: 10.31915/NWS.2022.41

Régi magyar verskorpusz összeállítása stilometriai és számítógépes metrikai kutatásokhoz[1*]

Maróthy Szilvia
ELTE BTK Magyar Irodalom- és Kultúratudományi Intézet OTKA 135631 kutatócsoport
mthy.szilvi@gmail.com
ORCID: 0000-0003-2558-9504

Seláf Levente
ELTE BTK Magyar Irodalom- és Kultúratudományi Intézet
selaf.levente@btk.elte.hu
ORCID: 0000-0002-6052-9841

Vigyikán Villő
ELTE BTK Magyar Irodalom- és Kultúratudományi Intézet OTKA 135631 kutatócsoport

The project "Computerized metrical and stylometric study of early modern Hungarian poetry" (OTKA, 2020–23) aims to computationally process and analyse Hungarian historical songs (epic poetry from the 16–17th century). Our paper introduces the methodology of building a corpus of 174 poems (length between 50 and 1000 strophes). The corpus mostly contains OCRd and corrected texts from printed critical editions, but born digital scholarly editions (in PDF, HTML, XML, DOC formats) are also represented. We had to make further changes to the texts to make them suitable for NLP tools: create modernized transcription, encode hiatus or non-metrical paratexts (titles, arguments). The research is assisted by the *Répertoire de la poésie hongroise ancienne* (https://f-book.com/rpha/) database, which serves the literary historical and metrical metadata for the analysed texts. During the mentioned OTKA research project a new version of the database is being developed, and content update is in progress.

Keywords: stichometry, stylometry, literary corpus, epic song, digitization

Bevezetés

A 2020-ban indult „A régi magyar költészet számítógépes metrikai és stilometriai vizsgálata" kutatási projekt három területre fókuszál: a *Répertoire de la poésie hongroise ancienne* adatbázis fejlesztésére; a 16. század második felében kibontakozó szerelmi költészetre; valamint a 16. század egyik legnépszerűbb műfajára, a históriás énekre.

A *Répertoire de la poésie hongroise ancienne*,[2] azaz a régi magyar versek adatbázisa (továbbiakban RPHA) egy 1976 óta mind technológiájában (először lyukkártyás, majd sorban nagygépes, CDS/ISIS, webes verziók), mind tartalmában fejlődő kutatói adatbázis, mely a kezdetektől az 1600-ig tartó időszak magyar nyelvű verseinek és azok forrásainak jegyzékét, részletes irodalomtörténeti és poétikai leírását adja strukturált rendszerben. A jelenleg zajló OTKA projektben készül az adatbázis hetedik verziója, mely számos webtechnológiai újításon túl

1 Az 135631 számú projekt az Innovációs és Technológiai Minisztérium Nemzeti Kutatási Fejlesztési és Innovációs Alapból nyújtott támogatásával, az OTKA-K pályázati program finanszírozásában valósult meg.

2 Horváth Iván és mtsai., „Répertoire de la poésie hongroise ancienne, v. 7.0", 1979. 2022, https://f-book.com/rpha/v7/.

jelentős tartalmi gazdagodáson is keresztülmegy, a versek metaadatai mellé nagy számban felkerülnek azok szövegei is. Az RPHA emellett a POSTDATA nemzetközi verstörténeti adatbázis projekthez is kapcsolódik, melynek keretében mind az RPHA metaadat állománya, mind az adatbázisban már elérhető szövegek lekérdezhetők lesznek a POSTDATA közös felületein is, és az ahhoz kialakított ontológia alapján kerülnek leírásra.[3]

A kutatásnak több része van. Az egyik a kora újkori magyar szerelmi, s különösen a felsőbb (ún. arisztokratikus) regiszterbe sorolt udvari szerelmi líra kibontakozása, mely elsősorban Balassi Bálint és követőinek költészetéhez fűződik. Kutatásunkban a líra és az epika nyelvi és metrikai kapcsolódási pontjait vizsgáljuk a szerelmi költészetben. Milyen művek hathattak Balassi költészetére, s milyen a Balassi-versek utóélete lírában és epikában (a szerelmi tárgyú históriákban)? Stilometriai és számítógépes metrikai eszközökkel finomíthatjuk-e a Balassi-követők és Balassi műveinek kapcsolatáról alkotott képünket?

A projekt másik fontos része a históriás ének műfajának vizsgálata. Ide olyan, általában hosszabb epikus költemények tartoznak, melyek megtörtént eseményeket vagy bibliai, ill. fiktív (szerelmi, mitológiai) történeteket beszélnek el. A műfaj a 15. század végén jelenik meg, a 16. században teljesedik ki, s a 17. század elejére már csak kis számban van jelen. Kutatásunkban a műfaj metrikai és stilisztikai egységét, sokszínűségét, valamint az oralitáshoz való kapcsolatát vizsgáljuk számítógépes metrikai és stilometriai eszközökkel. Az alábbiakban az OTKA kutatás ezen alprojektjének első eredményeiről számolunk be részletesebben.

1. A históriás énekek alprojekt

A kora újkori magyar költészet egyik legfontosabb műfaja a históriás éneké, arányuk a 16. századi magyar versek között jelentős: számukat tekintve több mint 12%, a terjedelem szempontjából, a sorok számát nézve pedig több mint 50% tartozik ide. A kevés kortárs irodalomelméleti és esztétikai reflexió szerint is a magyar kora újkor reprezentatív műfaját képviseli. A kor legfontosabb alkotói közül Tinódi Sebestyén vagy Bogáti Fazakas Miklós is számos históriás énekeket írt különféle tematikával, történelmi, bibliai témákban.

Kutatásunk arra irányul, hogy az énekeket jellemző, sokak által monotonnak, túl egyszerűnek tekintett verselés pontos jellemzőit meghatározzuk, és megvizsgáljuk, egy bő évszázad alatt hogyan alakult át a magyar verselés logikája, jellege. Mennyire volt elterjedt az önrím, a ragrím, mennyire tiszták a rímek vagy például a szótagszámok és a sormetszetek szabályosságára mennyire ügyeltek a költők, esetleg felfedhetők-e eddig nem ismert strófaszerkesztési szabályok (pl. sormetszetek változása[4]). Stilometriai és sztichometriai módszerek segítségével anonim vagy bizonytalan szerzőségű versek alkotóit is megpróbáljuk azonosítani a kiterjedt verskorpusz segítségével (például az *Eurialus és Lucretia* című szerelmi históriáét).

A históriás énekkorpusz a 15. század végéről 4 éneket, a 16. századból 180 éneket (ebből 10 szövegét nem ismerjük), a 17. századból kb. 23–25 éneket számlál (a korpusz jelenleg a 16. század végéig teljes). Műfajtörténeti áttekintésre a jelen írás keretei nem adnak lehetőséget,[5]

---

3 L. ennek kifejtését: Horváth Andor, „Poetry Database Connector (PDC)", *Digitális Bölcsészet* 6. sz., 5 (2022): megjelenés alatt. További információ a POSTDATA nemzetközi költészettörténeti projektről: Poetry Standardization and Linked Open Data, hozzáférés: 2022. június 18., https://postdata.linhd.uned.es.

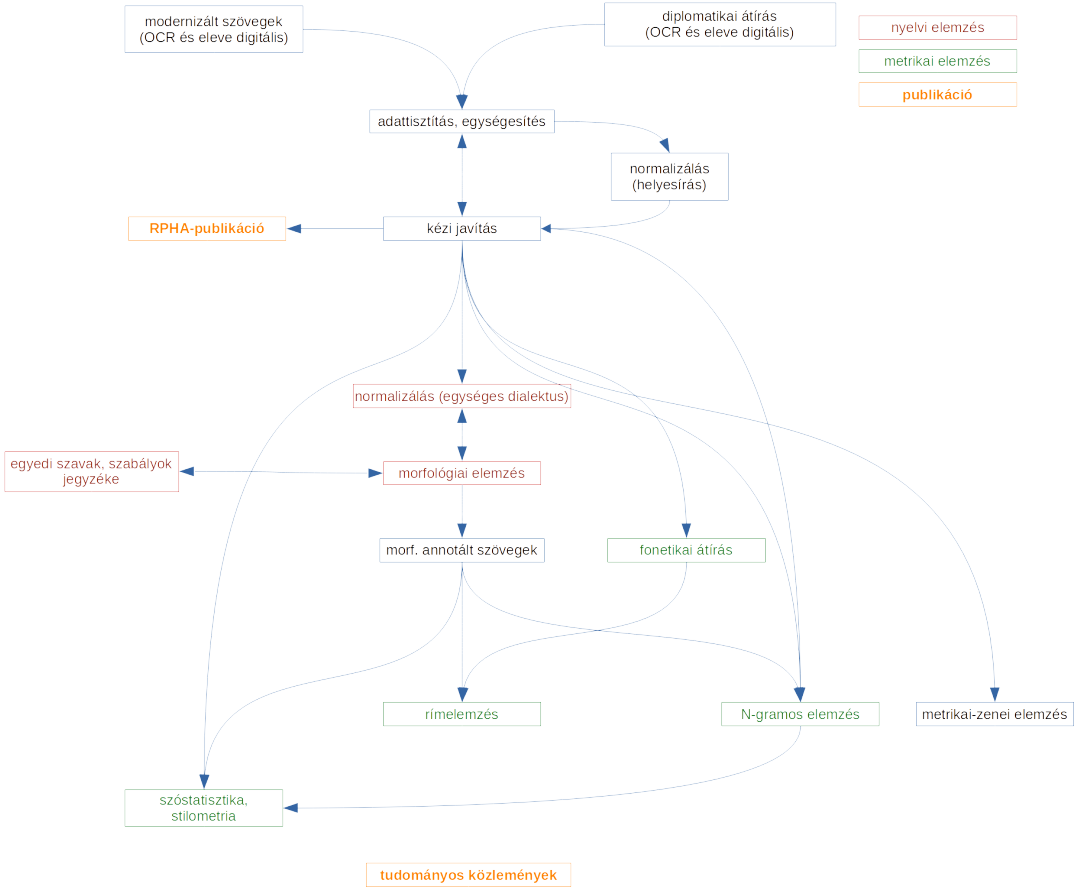
4 Erre az esetre szolgáltatott jó példát újabban Szatmári Áron előadása, mely Bogáti Fazakas (RPHA 2013) és Csanádi Demeter (RPHA 0230) egy-egy históriás énekének közös verselési szabályát fedte fel. Eszerint a 13 (7,6) sorok sormetszete a strófák harmadik sorában (az énekek második felére) nagy számban következetesen hiányzik. Szatmári Áron, „7+6, avagy Demeter király esete a kanásztánccal", *A históriás ének: poétikai és filológiai kérdések* műhelykonferencia, ELTE BTK Budapest, 2022. június 10.

5 Műfajtörténeti áttekintéshez l. Pap Balázs, *Históriák és énekek*, Pannónia könyvek (Pécs: Pro Pannonia, 2014).

így csak röviden utalunk arra, hogy a művek kiválasztásakor a 15–16. századi anyag esetében az RPHA műfaji besorolása szerint gyűjtöttünk (*vallásos>história* vagy *világi>história* műfaji besorolású művek).[6] A 17. századi anyag összegyűjtése a releváns szakirodalom alapján történik és még folyamatban van. Gyűjtésünk az RMKT 17. századi sorozatában kiadott műveket tekinti magnak, azonban a barokk epikus költeményeket, melyek már távol állnak a históriás ének műfajától, nem tartalmazza. Ilyenek például Zrínyi Miklós *Obsidio Szigetiana* című eposza vagy Gyöngyösi István epikus költeményei. A projekt keretében a 17. századi énekek RPHA-szerkezetnek megfelelő bibliográfiai, irodalomtörténeti és poétikai leírása is elkészül, így válik majd a korpusz a szöveg és a metaadatok szintjén is együtt elemezhetővé. (Az RPHA eredeti gyűjtőköre az 1600-as évvel zárul, az OTKA projekt keretében azonban ilyen jellegű tartalmi bővítéseket is tervezünk.)

2. A korpuszépítés lépései

A históriás énekkorpusz összeállításának és elemzésének lépéseit az alábbi folyamatábrán foglaltuk össze. A jelen írás az előkészítő fázist mutatja be, mely során a nyomtatott, digitalizált vagy eleve digitális forrásokból előállt a nyelvi annotációkat tartalmazó, számítógépes elemzésre alkalmas, megbízható tudományos minőségű szöveganyag.



1. ábra: Folyamatábra a munka menetéről

Fontos kiemelni, hogy a régiség szöveganyagával való munka számos nehézséget rejt magában, hiszen különösen jellemző rá az egyedi és nem következetes ortográfia, a különféle nyelvjárások megléte, az archaikus grammatikai paradigmák és lexikális elemek nagy száma.

6 A lekérdezés linkje: https://f-book.com/rpha/v7/results.php?boole1=EMPTY&field1=M%C5%B1faj&value1=049&method1=LIKE&boole2=OR&field2=M%C5%B1faj&value2=002&method2=LIKE&boole3=AND&field3=A+szereztet%C3%A9s+ideje&value3=&method3=LIKE&boole4=AND&field4=R%C3%Admk%C3%A9plet&value4=&method4=LIKE.

Ezekre az eddig fejlesztett nyelvi elemzők nincsenek felkészülve, így az előfeldolgozás nagy kihívások elé állított minket.

A korpusz alapját zömmel nyomtatott szövegforrások, a Régi Magyar Költők Tára sorozat már bescannelt és OCR-ezett, online elérhető,[7] valamint az általunk újrascannelt és OCR-ezett kötetei adták. Ezen felül kerültek bele digitális nyomdai fájlok (PDF), tudományos igényű internetes szövegkiadások (HTML) és kutatók által modernizált szövegek digitális kéziratban[8] (DOC, DOCX), valamint előfordultak egyéb, például folyóiratbeli szövegközlések is.

A szövegek kinyerését Abbyy OCR szoftverrrel végeztük. Mivel az interneten elérhető különféle források scan és OCR minősége igen heterogén volt, a szövegfelismertetést sok esetben újra el kellett végezni. Az így kapott nyers szöveget több lépcsőben kézi és automatizált javításoknak vetettük alá. Az automatizált javításokat regexszel végeztük, ez elsősorban a nem releváns alfanumerikus elemek kiszűrését (pl. a sorszámozás és a fejlécek törlése), valamint a szöveg szegmentálását jelentette (sor és strófahatárok egységes jelölése). A kézi javítás során az így előfeldolgozott szöveget olvastuk össze a digitalizált forrással, az OCR (betűtévesztési és esetleges értelmi) hibáit szűrve ki. A kimenet egy egyszerű szöveg lett minimális, MarkDown-szerű szintaxissal (sor- és strófahatárok, hiányok, fejezethatárok, paratextusok jelölésével/elkülönítésével). Ez a többszörös ellenőrzési folyamat jóval időigényesebb, azonban ettől lesz megbízhatóbb a nyelvi korpusz, ellentétben olyan heterogén forrású, szövegkritikailag nem ellenőrzött digitális bölcsészeti korpuszokkal, melyek a webes szövegforrásokat kontroll nélkül inkorporálják.

A kézi és automatizált javítások következő fázisa a modernizálás vagy normalizálás. Modernizálásra esetünkben két okból volt szükség. Egyrészt a korpusz jelentős része eleve modernizált kiadásban jelent meg az RMKT köteteiben, így a homogén korpusz előállításához minden szövegnél modernizálni kellett. Maga a több mint egy évszázadra visszatekintő RMKT sorozat is tükrözi a magyar helyesírás változásait, ezekre is tekintettel kellett lennünk (pl. *c* hangérték *cz* betűvel való jelölése a régebbi kötetekben). A modernizálás másik oka, hogy a nyelvi elemző programok a különféle ortográfiájú, nyelvjárású szövegeket jelentős hibaszázalékkal ismerik csak fel, s ezért eredménytelenek vagy fals eredményeket hoznak. A projektben elsősorban nem olyan jellegű kutatásokat végzünk, ahol az ortográfiai jellegzetességeknek jelentősége lenne (vö. némely stilometriai, szerzői attribúciót célzó kutatással), így az eredeti ortográfia az elemzés szempontjából nem releváns. Fontos azonban hangsúlyozni, hogy a munkafolyamat során létrejött különféle szövegátiratokat megtartjuk, azokra a munkafolyamat különböző állomásain szükség van. Például az eredeti, kritikai kiadáshoz közel álló szöveg az RPHA adatbázis felületén jelenik meg, a már csak verses szövegelemeket tartalmazó, de nem modernizált szöveg fonetikai annotációt kap, a számítógépes elemzésre szánt modernizált szöveg pedig részint nyelvi, részint metrikai annotációt stb. (Lásd a folyamatábrán fentebb.)

A kézi modernizálás számára közös szabályrendszert alkottunk, az egyes eseteket példákkal is illusztrálva. Noha törekedtünk az egységességre, a gyakorlat sokszor felülírta a szabályokat, következetlenségeket idézett elő, hiszen egyszerre kellett figyelembe venni a nyelvjárási vagy verstani szempontból fontos alakok megtartását, s ugyanakkor a (gépi) érthetőséget is szem előtt kellett tartanunk. A modernizáláshoz szükséges idő nagyon eltérő volt, attól függően, milyen a 16. századi nyomtatványok, kéziratok helyesírása. A 20. század második felében elterjedt, a régi helyesírást, sőt néha betűképet is pontosan tükröző kritikai kiadások esetében

7 *Nemzeti klasszikusok kritikai kiadásai*, BTK Irodalomtudományi Intézet, hozzáférés: 2022. 06. 18, https://szovegtar.iti.mta.hu/hu/sorozatok/rmkt-xvii-szazad/.

8 Ezúton mondunk köszönetet Pap Balázsnak és az elhunyt Vadai István örököseinek a rendelkezésünkre bocsátott szövegekért.

ez meglehetősen időigényes munka. Soronként gyakran 8–10 változtatásra is szükség volt, ha például az eredeti kiadás nem használt egyáltalán ékezeteket (vagy csak az a/á, e/é betűpároknál), vagy ha fonetikus és nem szóelemző helyesírással éltek a nyomdászok és másolók. Egy kb. 100 strófás vers modernizálásához korábbi filológiai tapasztalat birtokában is minimum 4–5 munkaórára van szükség.

Az automatizált előnormalizálás vagy modernizálás[9] során a nyelvjárásában, ortográfiájában homogénebb szövegcsoportokra a tipikus esetek figyelembe vételével szabályokat alkottunk, ezek mentén cserélte ki az előnormalizáló program a szövegek betűkészletének adott részeit. Önmagában ez a módszer sem nyújt megoldást, hiszen a kézi modernizálással szemben „túlságosan" is következetes, így nem minden esetet fed le és használatával a javításkor néha sajnos rontunk is. Például egy ö-ző szövegben az *ö*-k cseréje *e*-re az *öröm* (/N+Nom) szót az *erem* (/N+Poss.1Sg+Nom) szóra cseréli. Így mindkét esetben szükség van kézi ellenőrzésre is; a legcélravezetőbb a két módszer kombinálása és iteratív alkalmazása a megfelelő szövegállapot eléréséig.

A munkafolyamat következő lépése a szövegannotáció. A nyelvi elemzéshez az E-magyart, illetve az emMorph elemző ó- és középmagyarra alakított változatát, az emMorphOMH-ot használtuk.[10] Utóbbi mellett azért döntöttünk, mert jóval alacsonyabb hibaszázalékkal ismeri fel a nyelvi elemeket, hiszen a régies nyelvi paradigmák, lexémák jó részét tartalmazza, melyeket az E-magyar nem.

A rímek elemzését a RhymeTagger[11] programmal végezzük, melynek implementálásában a készítő, Petr Plecháč segíti kutatócsoportunkat. A program bemenete a versszövegek IPA szabvány szerinti fonetikus átírása, melynek előállítását az Epitran és az Espeak programokkal végeztük tokenizálást követően.

Az így összeállt históriás énekkorpusz mérete számokban: 174 vers (a fennmaradó 11 vers töredék, melyet csak említésből ismerünk), 25 007 versszak, 98 141 verssor és 527 598 szövegszó. Összehasonlítás végett más történeti korpuszok mérete: az Ómagyar Korpusz[12] 3,2 millió szövegszót, a Történeti Magánéleti Korpusz[13] 1 millió 112 ezer elemzett szövegszót, a Magyar Történeti Szövegtár pedig 3 millió szövegszót tartalmaz. A kutatásaink során létrejött

9 Ezen munkafolyamatban kutatócsoportunk tagja, Simon Eszter volt segítségünkre.

10 Novák Attila, „Milyen a jó Humor? [What is good Humor like?]", in *I. Magyar Számítógépes Nyelvészeti Konferencia* (Szeged: SZTE, 2003), 138–44; Attila Novák, „A New Form of Humor – Mapping Constraint-Based Computational Morphologies to a Finite-State Representation", in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, szerk. Nicoletta Calzolari és mtsai. (Reykjavik, Iceland: European Language Resources Association (ELRA), 2014), 26–31; Attila Novák, Borbála Siklósi, és Csaba Oravecz, „A New Integrated Open-Source Morphological Analyzer for Hungarian", in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, szerk. Nicoletta Calzolari és mtsai. (Paris, France: European Language Resources Association (ELRA), 2016), 23–28; Váradi Tamás és mtsai., „Az e-magyar digitális nyelvfeldolgozó rendszer", in *XIII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2017)*, szerk. Berend Gábor, Gosztolya Gábor, és Vincze Veronika (Szeged: Szegedi Tudományegyetem Informatikai Tanszékcsoport, 2017), 49–60; Tamás Váradi és mtsai., „E-Magyar – A Digital Language Processing System", in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, szerk. Nicoletta Calzolari (Conference chair) és mtsai. (Miyazaki, Japan: European Language Resources Association (ELRA), 2018). Ezúton köszönet Sass Bálint és Novák Attilának, hogy rendelkezésünkre bocsátotta az elemzőt.

11 RhymeTagger, hozzáférés 2022. június 16., https://github.com/versotym/rhymetagger. Vö. Petr Plecháč, „A Collocation-Driven Method of Discovering Rhymes (in Czech, English, and French Poetry)", in *Taming the Corpus: From Inflection and Lexis to Interpretation*, szerk. Masako Fidler és Václav Cvrček, Quantitative Methods in the Humanities and Social Sciences (Cham: Springer International Publishing, 2018), 79–95, https://doi.org/10.1007/978-3-319-98017-1_5.

12 http://omagyarkorpusz.nytud.hu

13 http://tmk.nytud.hu/

szövegkorpuszt nyílt licenc alatt tesszük közzé, így gazdagítva az elérhető magyar nyelvi korpuszok halmazát. (Részletek a kutatás folyamatosan frissülő GitHub repozitóriumában: https://github.com/versotym/oldhun.)

3. Első eredmények és a folytatás

Első kísérleti tanulmányunkban[14] 26 históriás ének számítógépes metrikai vizsgálatát és annak eredményeit mutattuk be. A kutatás a históriás énekkorpusz verstani homogeneitását vizsgálta. A tanulmány elsősorban annak a közel-kortársi vélekedésnek járt utána kvantitatív eszközökkel, miszerint a históriás énekek verselése monoton, kevés poétikai invencióval szolgál – ilyen a 17. század elején Szenci Molnár Albert „számtalan az sok vala vala vala" kijelentése is (*Psalterium Ungaricum*, 1607, Herborn). Ez a kis korpusz ugyan még korántsem volt reprezentatív, de elemzése tendenciákat így is sejtet. Eszerint az idő előrehaladtával határozottan változik a verselés technikája. Így például egyre kevesebb (a szövegekben kisebb arányú) a morfémarím, a szóismétlő rím és a rímtelen sorpár a vizsgált epikus énekekben.

Az elemzés arra is rámutatott, hogy egyes szerzők művei számítógépes metrikai eszközökkel is jól elkülöníthető csoportokat alkotnak. Az önrímek időbeli megoszlását figyelve például a históriás énekköltészetben is kiemelkedő Tinódi Sebestyén és Bogáti Fazakas Miklós énekei alkotta csoportok láthatóan elkülönülnek, a korabeli átlagtól távol esnek. Tinódinál ugyan a három vizsgált énekből egy épp az átlag fölé kerül, azaz több benne az önrím – ezt az is indokolhatja, hogy az újabb kutatások szerint a szöveg valójában egy nagyobb históriás ének töredékesen fennmaradt része, nem pedig különálló mű. Az eltérés pontos okára és további verstani és stilometriai mintázatokra a teljes Tinódi-korpusz vizsgálata deríthet majd fényt.

A históriás énekköltészetet vizsgáló projektünk jelen célkitűzése, hogy

a) a 16. századra összeállt teljes korpuszon elvégezzük a fenti vizsgálatokat;

b) az énekkorpuszt kibővítsük a 17. századi anyaggal, mind a metaadatok, mind pedig a szövegek szintjén;

c) a számítógépes metrikai elemzések mellett stilometriai vizsgálatokat végezzünk a nyelvileg annotált szövegkorpuszon (pl. szerzői és műfaji csoportok, vagy az oralitás és írásbeliség vizsgálata);

d) az RPHA szolgáltatta metaadatok segítségével a statisztikai elemzéseken túl intertextuális és interperszonális jelenségeket kutassunk, a szöveg és a metaadatok elemzését összekapcsoljuk;

e) az RPHA adatbázison további tartalmi és technológiai fejlesztéseket végezzünk.

A projektet megelőzően ilyen mélységű és ilyen változatos számítógépes elemzés még nem készült magyar verses korpuszon.

14 Szilvia Maróthy, Levente Seláf, és Petr Plecháč, „Rhyme in 16th-Century Hungarian Historical Songs: A Pilot Study", in *Tackling the Toolkit: Plotting Poetry through Computational Literary Studies*, szerk. Petr Plecháč és mtsai. (Institute of Czech Literature of the Czech Academy of Sciences, 2022), 47–62, https://doi.org/10.51305/ICL.CZ.9788076580336.04.