DOI: 10.31915/NWS.2021.7

Az ELTE DH Regénykorpusz és lehetőségei

Bajzát Tímea Borbála
ELTE DH
timimimi@student.elte.hu

Szemes Botond Bálint
ELTE DH
boboelte@student.elte.hu

Szlávich Eszter
ELTE DH
szlavich.eszter@btk.elte.hu

The corpus of the Hungarian novel created at the ELTE Department of Digital Humanities offers new methodologies for the philological research and "distant reading" approaches by providing a digitized, annotated and searchable database of freely accessible novels from the Hungarian literary history. The database fits organically into the international collection of the ELTeC COST Action Project (European Literary Text Collection https://www.distant-reading.net/eltec/), since the first 100 novels of the database are part of the Hungarian sub-corpus of that collection. Beside the description of the corpus and the aspects of the selection, the paper also reports in detail on the possibilities of the quantitative analysis of the novels. In doing so, we want to present what kind of knowledge of the Hungarian literary history can be produced by applying statistical and linguistic approaches, and what role these methods can play in the process of the interpretation of the texts. Through the visualized tendencies, a new history of the style of the Hungarian prose can be outlined, while the peculiarities of some texts in relation to others can lead to the description of the poetics of the given authors and their novels.

Keywords: distant reading, corpus linguistics, literary corpus, Hungarian novels

1. Bevezetés

A tanulmány célkitűzése az ELTE BTK Digitális Bölcsészet Tanszék által létrehozott és fejlesztett Regénykorpusz[1] projektum[2] összegző bemutatása. Az ELTE BTK Digitális Bölcsészet Tanszék által szolgáltatott Regénykorpusz olyan folyamatosan bővülő, innovatív kezdeményezés,[3] amely egyedüliként biztosítja magyar nyelvű regények annotált, kereshető elérését, ezáltal alkalmas a nagy mennyiségű adatokon végzett korpuszalapú és korpuszvezérelt szövegvizsgálatokra. A Regénykorpusz aktuálisan 100 magyar nyelven írt regényt tartalmaz, összesen 81 szerzőtől, mérete megközelítőleg 7 000 000 token. A Regénykorpusz ezen gyűjteménye szervesen illeszkedik abba a nemzetközi COST Action kutatási együttműködésbe, amely a Distant Reading for European Literary History[4] nevet viseli. A kutatási projektum célkitűzései közé tartozik egy kutatói hálózat megteremtése a projektben részt vevő kutatóintézetek között. Az együttműködés célja az, hogy olyan források és eszközök válnak széles körben hozzáférhetővé, melyek elősegítik azt, hogy az európai irodalomtörténet-írás és irodalomelméleti paradigma, valamint az ezekhez fűződő fogalmi apparátus újraértelmezhetővé váljon. A projektum a kutatás módszertanában a távoli olvasás metódusát érvényesíti. Valamint célja az, hogy adatvezérelt, koherens, számítógépes eljárást alakítson ki és tegyen tesztelhetővé az irodalmi szövegek komparatív elemzésére.[5] A projekt keretében jött létre, és dinamikusan fejlődik az European Literary Text Collection (ELTeC)[6], amely a projektben részt vevő európai nyelvek regénykorpuszait tartalmazza,[7] ennek része a jelen tanulmányban bemutatandó Regénykorpusz is.

2. A magyar regény korpusza

A Regénykorpusz jelenlegi státuszában az ELTeC, vagyis a European Literary Text Collection gyűjteményének részét képezi, ezáltal mind a gyűjteménybe kerülő szövegek kiválasztási metódusa, mind pedig az anyagon alkalmazott kódolási séma ezen

1 https://regenykorpusz.elte-dh.hu/?lang=hu-HU (utolsó elérés: 2021.05.26.)

2 A projektum létrehozásán és fejlesztésén dolgozó kutatók és az ellátott feladatkörük a következők: Dr. Palkó Gábor (projektvezető és az alkalmazott TEI XML specifikáció készítője), Fellegi Zsófia (az alkalmazott TEI XML specifikáció készítője), Takács Emma (jelölőnyelvi kódolás), Vétek Bence (jelölőnyelvi kódolás), Dr. Kundráth Péter (a Regénykorpusz lekérdező felületének létrehozása), Dr. Horváth Péter (a Regénykorpusz lekérdező felületének létrehozása), Szemes Botond (a lekérdező funkciók bővítése), Szlávich Eszter (lekérdező funkciók bővítése) és Bajzát Tímea Borbála (jelölőnyelvi kódolás, metaadatolás, lekérdező funkciók bővítése).

3 A projektet a Felsőoktatási Intézményi Kiválósági Program támogatta, jelenleg a Digitális Örökség Nemzeti Laboratórium keretei között végezzük a korpusz fejlesztését.

4 COST Action CA16204 (utolsó elérés: 2021.05.26.)

5 Vö. Christof Schöch, Maciej Eder, Arias Rosario, François Pieter, Antonija Primorac, *Foundations of Distant Reading. Historical Roots, Conceptual Development and Theoretical Assumptions around Computational Approaches to Literary Texts*. 2020, hozzáférés: 2021.05.26. https://dh2020.hcommons.org/

6 Carolin Odebrecht, Lou Burnard, Christof Schöch, *COST Action Distant Reading for European Literary History (CA16204)*, 2021, hozzáférés: 2021.05.26. doi: https://doi.org/10.5281/zenodo.4662444

7 A projekthez lásd még: https://www.distant-reading.net/eltec/ (2021.05.23), amelynek magyar nyelvű adatbázisa elérhető: https://github.com/COST-ELTeC/ELTeC-hun (2021.05.23).

projektum előírásaihoz illeszkedik.[8] Ennek alapelveit követve törekedtünk a korpusz változatosságának maximalizálására, tehát a különféle terjedelmű és kanonizáltságú művek, valamint a különböző nemű szerzők arányos eloszlására.

A Regénykorpusz gyűjteményét olyan művek képezik, amelyeknek az első nyomtatott kiadásuk az 1840-től az 1920-ig tartó periódusra datálható az Országos Széchényi Könyvtár internetes katalógusa szerint.[9] Ezen 80 évből álló időszak további négy alperiódusra oszlik fel (lásd 1. ábra), tehát az alkorpusz a 19. századi és 20. század eleji magyar regényirodalomból nyújt mintát a vizsgálatokhoz. Ugyan a szövegek kiválasztásánál ezen alperiódusokat vettük figyelembe, de a keresőfelületen az első kiadás évszáma alapján végezhetünk szűréseket. Az alperiódusok pontos mérete a következő: A T1 alkorpusz 22 regényt, a T2 21 regényt, a T3 27 regényt, a T4 pedig 30 regényt tartalmaz. A regények között nem szerepelnek fordítások, tehát mindegyike magyar nyelven íródott. A periódusoknak megfeleltethető alkorpuszokra azonban nem csupán mennyiségi megkötést alkalmaztunk, hanem minden algyűjteménynek minimum 10%-át kellett kitenniük a női szerzők által írt műveknek, ami így alkorpuszonként legalább három női szerző által írt szöveget eredményezett. A változatosság maximalizálása miatt pedig a teljes gyűjteményre érvényes volt az a szabály, hogy szerzőismétlődés legfeljebb tizenegyszer fordulhatott elő és ugyanattól a személytől legfeljebb három regény kerülhetett beválogatásra.

Az ELTeC által szabott kritériumoknak eleget téve csak olyan szöveget vettünk fel a korpuszba, amely legalább 10 000 szó terjedelemben íródott. A terjedelmi kategóriákat tekintve rövid prózának címkéztünk minden olyan művet, amelynek mérete 10 000 és 49 999 token közé esett, közepes méretűnek számítottak azon szövegek, amelyek 50 000 és 99 999 közötti szövegszót tartalmaztak és a hosszú regények kategóriába eső műveknek pedig a 100 000 szó felettiek számítottak. A teljes gyűjteményre vonatkozóan minimum 20, az előbbiek alapján hosszúnak számító regény került be a korpuszba (a Regénykorpuszban összesen 22 hosszú regény található).

A kanonikusságot tekintve a válogatási kritérium az volt, hogy a gyűjtemény minimum egyharmadát kell azon szövegeknek képezniük, amelyek magas kanonicitásúnak számítanak. Az ELTeC előírásai szerint azok a művek tartoznak ebbe a kategóriába, amelyek 1979 után minimum 2 új kiadással rendelkeznek, tehát a kanonikusság meghatározása ebben a kritériumrendszerben alapvetően a kiadástörténethez rendelődik.

8 A kiválasztási kritériumok forrásaként lásd: https://distantreading.github.io/sampling_proposal.html (hozzáférés: 2021.05.23.)

9 Lásd http://nektar1.oszk.hu/librivision_hun.html (hozzáférés: 2021.05.23.)

T1 (1840–1860)			T3 (1880–1900)		
22 db			27 db		
női szerző által írt	férfi szerző által írt		női szerző által írt	férfi szerző által írt	
3 db	19 db		4 db	23 db	
alacsony kanonicitás	magas kanonicitás		alacsony kanonicitás	magas kanonicitás	
14 db	8 db		21 db	6 db	
rövid terjedelmű	közepes terjedelmű	hosszú terjedelmű	rövid terjedelmű	közepes terjedelmű	hosszú terjedelmű
6 db	7 db	9 db	20 db	5 db	2 db
T2 (1860–1880)			**T4 (1900–1920)**		
21 db			30 db		
női szerző által írt	férfi szerző által írt		női szerző által írt	férfi szerző által írt	
6 db	15 db		7 db	23 db	
alacsony kanonicitás	magas kanonicitás		alacsony kanonicitás	magas kanonicitás	
17 db	4 db		18 db	12 db	
rövid terjedelmű	közepes terjedelmű	hosszú terjedelmű	rövid terjedelmű	közepes terjedelmű	hosszú terjedelmű
9 db	9 db	9 db	12 db	10 db	8 db

1. ábra. A Regénykorpusz algyűjteményei

A Regénykorpuszba csak szabadon elérhető szövegeket használtunk fel, a szövegek elsődleges forrása a Magyar Elektronikus Könyvtár,[10] de ahhoz, hogy a gyűjtemény megfelelhessen az ELTeC által támasztott válogatási kritériumoknak, kettő regény a Google Books szabadon hozzáférhető adatbázisából származik.[11] A leválogatásnál elsősorban arra törekedtünk, hogy olyan szövegekkel dolgozzunk, amelyek RTF formátumban elérhetők a MEK felületén, mert ezek olyan jó minőségben tárolják a munkaanyagot a számítógépes feldolgozás számára, hogy további munkálatokat nem igényelnek a kódolás során, azonban filológiai szempontból további kérdések fogalmazhatók meg velük kapcsolatban. Azonban a MEK-ről vételezett RTF dokumentumok önmagukban nem bizonyultak elegendőnek a kritériumok teljesítéséhez, így a mintavételezést kibővítettük a MEK-es és a Google Books-os anyag kétrétegű PDF-ben tárolt dokumentumaira is, amelyeken újra OCR-t (optikai karakterfelismerést) végeztünk el az ABBYY FineReader 14 szoftver alkalmazásával, majd a tipikus OCR hibákat kézzel javítottuk. A regények metaadatait (az első kiadás éve, kiadások száma) az OSZK katalógusának internetes keresőjéből és a Magyar Országos Közös Katalógusból (MOKKA)[12] gyűjtöttük össze.

A korpuszba kerülő szövegek alapvető kódolási formátuma a TEI XML jelölőnyelv,[13] amely mind az ember, mind pedig a gép számára olvasható metanyelv. Előnye, hogy eszköz- és rendszerfüggetlen, valamint a kódolt szövegtestek együttesen tárolhatók azok metaadataival. Ezen keresztül olyan irányelvek gyűjteménye, amely segítségével lehetővé válik a strukturált szöveg és információ megjelenítése a böngészőben, illetve más szövegformátummá konvertálható a feldolgozott anyag, ezen kívül a felhasználás célkitűzéseinek teljesítésére alkalmas annotációval láthatjuk el a szövegeket.[14] Az ELTeC projektum specifikus standardizációt alkalmaz a regények kódolásához, amely

10 https://mek.oszk.hu/ (hozzáférés: 2021.05.23.)

11 https://books.google.hu/ (hozzáférés: 2021.05.23.)

12 http://www.mokka.hu/ (hozzáférés: 2021.05.23.)

13 Lásd https://tei-c.org/about/history/ (hozzáférés: 2021.05.23.)

14 Kalcsó Gyula, „A TEI-XML felhasználása magyar nyelvű korpuszok építésében", in Boda István, Mónos Katalin szerk., *MANYE XX. Az alkalmazott nyelvészet ma: Innováció, technológia, tradíció*, (Debrecen: MANYE, Debreceni Egyetem), 67–68. 2011.

olyan specifikációja a TEI XML-nek,[15] mely lehetővé teszi a fejlécben a projektum szempontjából releváns metaadatok jelölését (lásd 2. ábra) az XML fejlécben.

```
            </respStmt>
            <bibl type="printSource"><title>Egy régi udvarház utolsó gazdája</title><author>Gyulai Pál</author><publisher>Interpopulart</publisher><pubPlace>Budapest</pubP
               <ref target="http://mek.oszk.hu/00600/00666/00666.pdf"/></bibl>
            <idno>ISBN 963 613 127 9</idno>
         </bibl>
         <bibl type="firstEdition">
            <title>Egy régi udvarház utolsó gazdája : Regény</title>
            <author>Gyulai Pál</author>
            <date>1857</date>
         </bibl>
      </sourceDesc>
   </fileDesc>
   <encodingDesc n="eltec-0">
      <p/>
   </encodingDesc>
   <profileDesc>
      <langUsage>
         <language ident="hu">Hungarian</language>
      </langUsage>
      <textDesc>
         <authorGender xmlns="http://distantreading.net/eltec/ns" key="M"/>
         <size xmlns="http://distantreading.net/eltec/ns" key="short"/>
         <canonicity xmlns="http://distantreading.net/eltec/ns" key="high"/>
         <timeSlot xmlns="http://distantreading.net/eltec/ns" key="T1"/>
      </textDesc>
   </profileDesc>
```

2. ábra. A TEI XML specifikációja a <header>-ben, amely a metaadatok tárolására alkalmas

Ahhoz, hogy az ELTE Digitális Bölcsészet Tanszék által összeállított Regénykorpusz kereshető legyen, akár például a morfológiai kódok alapján, tehát teljesítse azokat az elvárásainkat, amelyeket egy annotált korpusz felé támasztunk, szükség volt a szöveg elemeinek lemmatizációjára és morfológiai, valamint szófaji elemzésére. Ezek eléréshez az MTA Nyelvtudományi Intézetben fejlesztett e-magyar automatikus elemzőlánc emtsv verzióját alkalmaztuk[16], úgy, ahogy a szintén a Tanszéken fejlesztett Verskorpusz projektum esetében is.[17] Az e-magyar segítségével így lehetővé vált a szövegek tokenizálása, lemmatizálása, morfológiai és szófaji elemzése is. A következő (3.) fejezetben útmutatót adunk a Regénykorpuszban[18] való keresés lehetőségeinek használatához.

3. Keresés a korpuszban

A Regénykorpuszhoz elérhető egy nyilvános online lekérdező felület is, amely funkcióival és arculatával illeszkedik az ELTE Digitális Bölcsészeti Tanszék többi szolgáltatása közé (az ELTE DH elérhető, kereshető szövegkorpuszai: Verskorpusz és Cikk-kereső).

15 Vö. Lou Burnard, Christof Schöch, Carolin Odebrecht, In Search of Comity: TEI for Distant Reading, 2019, https://doi.org/10.5281/zenodo.3552489 (utolsó elérés: 2021.05.23.) 65–72., https://github.com/COST-ELTeC/ELTeC-hun (utolsó elérés: 2021.06.23.)

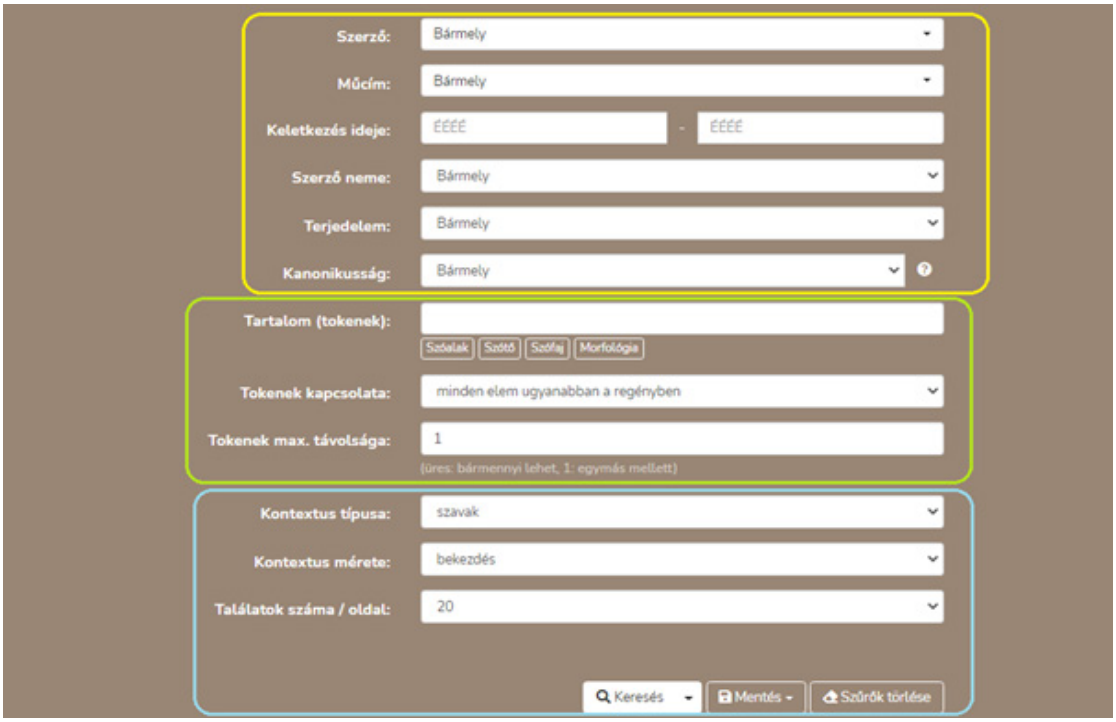
16 Indig Balázs, Sass Bálint, Simon Eszter, Mittelholcz Iván, Kundráth Péter, Vadász Noémi, „emtsv – egy formátum mind felett", in Berend Gábor, Gosztolya Gábor, Vincze Veronika szerk., *XV. Magyar Számítógépes Nyelvészeti Konferencia*. (Szeged: Szegedi Tudományegyetem TTIK, Informatikai Intézet), 235–247. 2019. Mittelholcz Iván, „emToken: Unicode-képes tokenizáló magyar nyelvre", in Vincze Veronika szerk., *XIII. Magyar Számítógépes Nyelvészeti Konferencia*. (Szeged: Szegedi Tudományegyetem, Informatikai Intézet) 61–69. 2017.

17 Horváth Péter, „Az ELTE Verskorpusz automatikus annotációs eljárásai révén nyerhető kvantitatív adattípusok", in Simon Gábor, Tolcsvai Nagy Gábor szerk., *Nyelvtan, diskurzus, megismerés*, (Budapest: Eötvös Kiadó), 313–331, 2020.

18 https://regenykorpusz.elte-dh.hu/ (utolsó elérés: 2021.06.23.); https://github.com/ELTE-DH/regenykorpusz (utolsó elérés: 2021.06.23.)

A korpusz létrehozásáról és a kereső részletes leírásáról a Súgó menüpontban lehet tájékozódni. A keresőfelület elérhető angol nyelven is.

A regénykorpusz honlapján elérhető részletes keresőfelület funkcióit három fő kategóriába lehet sorolni: a felület egyrészt alkalmas alkorpuszok létrehozására, másrészt lehet rajta tokenekre és tokenkapcsolatokra keresni, és végül vannak funkciók az adatok feldolgozási módjaira is.



3. ábra. A korpusz keresőfelülete

A keresőfunkciók első nagyobb csoportja a szövegek metaadatait szűri, azaz a szövegtestek körét szűkíti le a kereséshez: bevonható és/vagy kizárható egy-egy műcím vagy szerző; szűrhetjük a szövegeket a szerzők neme alapján; megadható a keletkezési időszak intervallumosan; valamint a regények terjedelme is specifikálható egy háromfokozatú skálán. Fontos és újszerű funkció a Regénykorpuszban az irodalmi kanonikusság címkézése, így szűrhetővé válnak a kevésbé vagy jobban kanonikus regények.

A tokenekre, illetve tokenkapcsolatokra való szűrés lehetővé teszi konkrét szóalakok vagy szótövek keresését. Az annotált korpuszok azonban alkalmasak arra is, hogy különböző morfológiai jellemzők alapján úgy keressünk különböző elemekre vagy elemkapcsolatokra, hogy a konkrét szóalakot nem adjuk meg. Mivel az egyes tokenek több szempontú annotációval rendelkeznek, szűrhetők a szóalakok, azaz a szövegbeli előfordulásuk alapján, szótövük alapján, illetve megadható a szófaj és a morfológiai jellemzők is. A kívánt szemponthoz tartozó jellemzőket legördülő menüben választhatjuk ki. Több tokenre való keresés esetén megadható a tokenkapcsolat és a tokentávolság is, hogy egyes nyelvi szerkezetek teljes köre kinyerhető legyen a korpuszból.

A harmadik nagyobb csoport a találatok megjelenítéséhez, az eredmények feldolgozásához kapcsolódik. Egyrészt megadható, hogy a találatokat mekkora kontextusban szeretnénk lekérni, illetve tetszőlegesen választható, hogy mekkora egységet szeretnénk egy oldalon megjeleníteni (5–500 találat/oldal). Végül a *Mentés* gomb legördülő menüjén kiválaszthatjuk, hogy a találatok listáját, gyakorisági listát, statisztikát, vagy a kiválasztott regények metaadatait szeretnénk-e menteni. Az így előállított listák (tsv formátum) minden statisztikai alapokon nyugvó digitális bölcsészeti kutatás kiindulópontját jelentik; a Regénykorpusz egyik fontos jellemzője, hogy ezeket pár kattintással, komolyabb technikai ismeretek nélkül elérhetővé teszik a felhasználók számára.

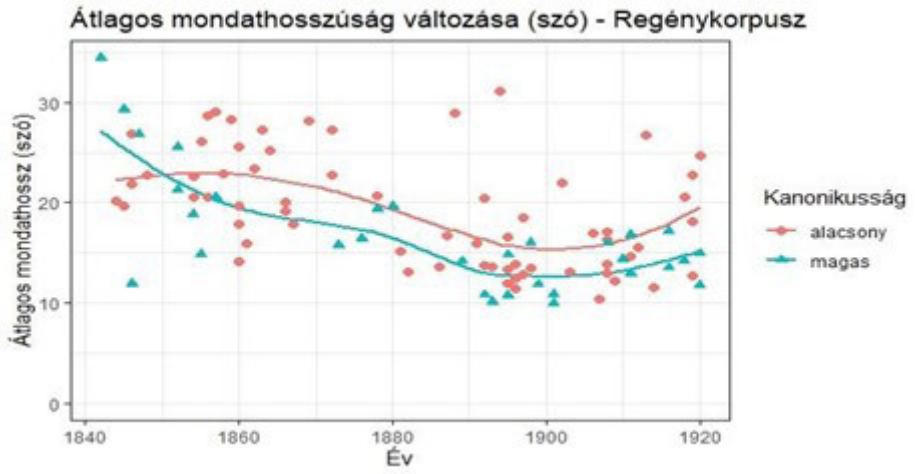
4. Két példa a keresőfelület által létrehozott adatok irodalomtudományos hasznosíthatóságára

A keresőfelület segítségével létrehozott gyakorisági listák nemcsak a legtöbb nyelvstatisztikai megközelítés alapját jelentik, hanem fontos tanulságokkal szolgálhatnak az irodalomtudomány és az irodalomtörténet-írás számára is. Segítségükkel ugyanis a szövegek nyelvi szerveződésének mélystruktúráiba nyerhetünk bepillantást, valamint olyan összefüggéseket tárhatunk fel, amelyek az olvasás folyamatában reflektálatlanok maradnak. Hiszen ezek a keresések a nyelvi működés egy más szintjére vonatkoznak, mint amelyek a befogadás során a figyelem előterébe kerülnek. Ezáltal egyrészt az egyes szövegekre vonatkozó értelmezéseink egészülhetnek ki új szempontokkal, másrészt olyan irodalom- és stílustörténeti folyamatokat tehetünk láthatóvá, amelyek a nagy mennyiségű adatból kiindulva az irodalomtörténet-írás egy reprezentatívabb megvalósulását ígérhetik, amennyiben az így felvázolt folyamatok nem csak egy szűk kánonhoz tartozó szerzők történetére vonatkoznak, hanem az irodalmi termelés egészére kívánnak rálátást biztosítani. A feltárt összefüggések sokszor valóban új tudás létrehozásához járulnak hozzá, azaz újraírhatják eddigi, kizárólag az olvasás tapasztalatán nyugvó megállapításainkat. Más esetekben a statisztikai elemzés eredményei alátámasztják korábbi történeti, vagy a szövegek belső szerveződésére vonatkozó elképzeléseinket – ám ekkor ugyanolyan fontos és hasznos a kereséseket elvégezni, hiszen a korábbi hipotézisek és intuíciók ezáltal statisztikailag alátámasztható, adatolható visszaigazolást nyernek.

Az alábbi egyszerű példák a Regénykorpusz keresőfelületén lekérdezett adatok feldolgozását és vizualizációját foglalják magukban, és az eredmények irodalomtudományos hasznosíthatóságát hivatottak szemléltetni. A példák egyszerűsége jól mutatja, hogy még a meglehetősen banális szempontok mentén létrehozott adatok is a regények elrendezését, összehasonlítását és sajátosságaik kiemelését segíthetik elő.

A 4. ábrán a 100 regény mondatainak a szavak számában megadott átlagos hosszúsága látható egy idővonal mentén. Mivel a keresés lehetővé teszi a kanonikusság szerinti csoportosítását, így az ábrán a magas és alacsony kanonikusságú szövegek összehasonlítása is látható. Az ábrán tisztán kirajzolódik a mondatok hosszúságának csökkenése a 19. század során, ami a magyar prózahagyomány stiláris átalakulásán túl a

sajtó szerepének megerősödésével, az új íróeszközök elterjedésével, valamint az írás és az olvasás oktatására vonatkozó iskolai reformokkal hozható összefüggésbe.[19] Ezeknek a folyamatoknak az elemzése, valamint a mondatok rövidülésének stílustörténeti értékelése már eddig is az irodalomtörténet-írás részét képezték,[20] ám a kvantitatív módszerrel elért eredmények egyrészt adatolható módon igazolják vissza az eddigi belátásokat, másrészt az ábra – mint előre nem értelmezett viszonyrendszereket bemutató képi reprezentáció – lehetőséget teremt a különböző területek (stílus-, sajtó-, oktatás-, médiatörténet) összekapcsolására is a kirajzolódó tendenciák értelmezése során. Érdemes kiemelni továbbá, hogy a magas és alacsony kanonikusságú szövegek egyaránt követik ezt a tendenciát – – érdekes azonban, hogy míg a 19. század első harmadában a magas kanonikusságú szövegekhez tartozik egy inkább hosszú mondatos prózastílus, addig a század többi részében ez az arány megfordul. Ennek a megállapításnak az érvényességét egy nagyobb korpuszon végzett kutatás igazolhatja: a Regénykorpusz adatbázisának bővülése ennek lehetőségét is megteremtheti. Ez azért is lenne különösen fontos, mert a kanonikus és nem kanonikus szövegek összehasonlító elemzése (azaz annak a kérdésnek a vizsgálata, hogy a kánonképződés során kiválasztott szövegek rendelkeznek-e sajátos, elkülöníthető textuális jellemzőkkel a különböző történeti korokban) ezidáig az irodalomtörténetírás perifériájára szorult.



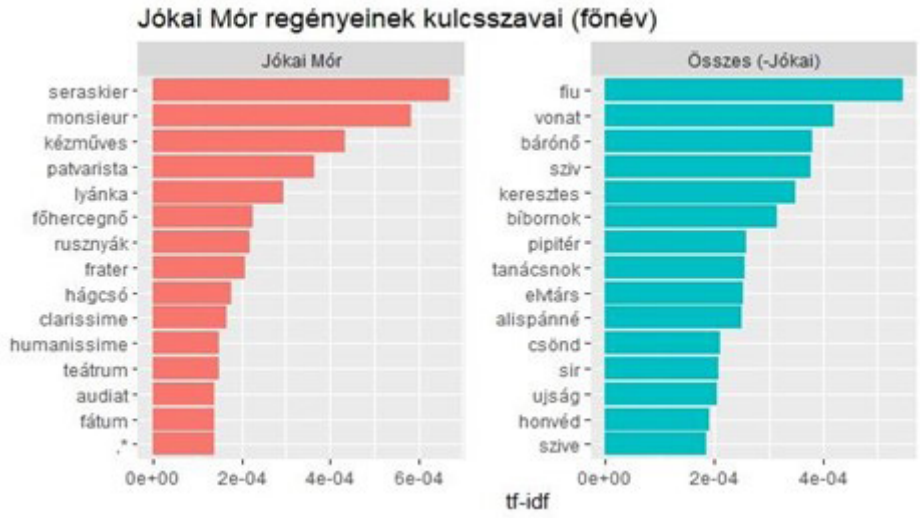
4. ábra. Átlagos mondathosszúság a Regénykorpuszban

Az 5. ábra a Jókai-alkorpusz, valamint a korpusz további 97 regényének összehasonlítását mutatja. Ehhez a Regénykorpusz felületén lekérdezett gyakorisági listákat az ún. „kifejezés gyakoriság–fordított dokumentum gyakoriság" (term-frequency-inverse document frequency, TF-IDF) módszerével hasonlítottuk össze. A módszer lényege, hogy a korpuszok egymáshoz képesti kulcsszavait tudjuk feltárni: kulcsszónak azok a kifejezések minősülnek, amelyek az egyik korpuszban magas számban, míg a

19 A mondathosszúságok változásának részletesebb elemzését lásd: Szemes Botond, Mondathosszúság és irodalomtörténet. 100 magyar regény szövegstatisztikai elemzése, *Literatura* (2020):3: 335–367.

20 Pl. Herczeg Gyula, *A XIX. századi magyar próza stílusformái*, Budapest, Tankönyvkiadó, 1981.

másikban ritkán fordulnak elő – és viszont.[21] Az ábra Jókai Mór tárgyalt regényeinek egy karaktertípusára és a hozzá tartozó nyelvhasználatra hívja fel a figyelmet: a franciás kifejezések a korpusz többi regénye esetében nem tekinthetők jellemzőnek. Az egyes karakterekhez kötődő francia nyelvhasználat már az olvasás tapasztalatában is reflektálttá válik, ugyanakkor annak kimutatásához, hogy ez mennyiben tekinthető a Jókai-próza sajátosságának, elengedhetetlen a fent vázolt módszer alkalmazása.



5. ábra. Jókai Mór regényeinek kulcsszavai

5. Összefoglalás és kitekintés

A tanulmány célkitűzése az ELTE BTK Digitális Bölcsészet Tanszék által létrehozott és fejlesztett Regénykorpusz projektum jelenlegi stádiumának és felhasználási lehetőségeinek összefoglaló bemutatása volt. A tanulmányban igyekeztünk felvázolni a korpuszban való keresés lehetőségeit, illetve bemutatni néhány egyszerű példán keresztül azt, hogy milyen potencialitással rendelkezik a gyűjtemény és annak keresőfelülete akár az irodalomtörténeti kérdések megválaszolásában. A Regénykorpusz esetében fontos hangsúlyozni azt, hogy a projekt korántsem tekinthető lezárt munkának. A jövőben tervezzük a korpuszban elérhető szövegek bővítését, s ezzel együtt azt, hogy lazítjuk a kiválasztási kritériumokat (például felveszünk olyan műveket is, amelyek 1840 előtt, vagy 1920 után kerültek kiadásra), ezáltal a felület alkalmas lesz akár diakrón vizsgálatokhoz való felhasználásra. Ezzel párhuzamosan tervezzük a keresési lehetőségek bővítését is. Jelenleg azon dolgozunk, hogy elérhetővé váljon egy tagmondatkapcsolat-felismerő funkció a felhasználói felületen, ami többek között mondatstilisztikai kutatások számára teremthet majd alapot.

21 Vö.: Shahzad Qaiser, Ramsha Ali, „Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents" *International Journal of Computer Applications* (0975 –8887) Volume 181. (2018):1. Eun-Soon You, Gun-Hee CHOI, Seung-Hoon KIM, „Study on Extraction of Keywords Using TF-IDF and Text Structure of Novels", *Journal of the Korea Society of Computer and Information*, 20/2, (2015): 121–129.

6. Bibliográfia

Lou Burnard, Christof Schöch, Carolin Odebrecht, *In Search of Comity: TEI for Distant Reading*, 2019, hozzáférés: 2021.05.26. https://doi.org/10.5281/zenodo.3552489

Eun-Soon You, Gun-Hee Choi, Seung-Hoon Kim, „Study on Extraction of Keywords Using TF-IDF and Text Structure of Novels", *Journal of the Korea Society of Computer and Information*, 20/2, (2015): 121–129.

Horváth Péter, „Az ELTE Verskorpusz automatikus annotációs eljárásai révén nyerhető kvantitatív adattípusok", In *Nyelvtan, diskurzus, megismerés*, szerkesztette Simon Gábor, Tolcsvai Nagy Gábor, 313–331. Budapest: Eötvös Kiadó, 2020.

Indig Balázs, Sass Bálint, Simon Eszter, Mittelholcz Iván, Kundráth Péter, Vadász Noémi, „emtsv – egy formátum mind felett", In *XV. Magyar Számítógépes Nyelvészeti Konferencia*, szerkesztette Berend Gábor, Gosztolya Gábor, Vincze Veronika, 235–247.Szeged: Szegedi Tudományegyetem TTIK, Informatikai Intézet. 2019.

Kalcsó Gyula, „A TEI-XML felhasználása magyar nyelvű korpuszok építésében", In *MANYE XX. Az alkalmazott nyelvészet ma: Innováció, technológia, tradíció*, Boda István, Mónos Katalin szerkesztette, 65–72. Debrecen: MANYE, Debreceni Egyetem, 2011.

Mittelholcz Iván, „emToken: Unicode-képes tokenizáló magyar nyelvre", In *XIII. Magyar Számítógépes Nyelvészeti Konferencia*. Vincze Veronika szerkesztette, 61–69. Szeged: Szegedi Tudományegyetem, Informatikai Intézet,2017.

Carolin Odebrecht, Lou Burnard, Christof Schöch, *COST Action Distant Reading for European Literary History (CA16204)*, 2021, hozzáférés: 2021.05.26., https://doi.org/10.5281/zenodo.4662444

Christof Schöch, Maciej Eder, Arias Rosario, François Pieter, Antonija Primorac, *Foundations of Distant Reading. Historical Roots, Conceptual Development and Theoretical Assumptions around Computational Approaches to Literary Texts*, hozzáférés: 2021.05.26. https://dh2020.hcommons.org/

Shahzad Qaiser, Ramsha Ali, „Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents", *International Journal of Computer Applications (0975–8887) Volume 181*. (2018):1.

Szemes Botond, Mondathosszúság és irodalomtörténet. 100 magyar regény szövegstatisztikai elemzése, *Literatura* (2020):3: 335–367.