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.2022.21

Az ELTE *Drámakorpuszának* létrehozása és lehetőségei

Szemes Botond, Bajzát Tímea, Fellegi Zsófia, Kundráth Péter,
Horváth Péter, Indig Balázs, Dióssy Anna, Hegedüs Fanni, Pantyelejev Natali,
Sziráki Sarolta, Vida Bence, Kalmár Balázs, Palkó Gábor
ELTE BTK Digitális Bölcsészet Központ
palko.gabor@btk.elte.hu

Absztrakt

Az ELTE DH Drámakorpusz egy folyamatosan bővülő adatbázis, amely a magyar drámairodalom szövegeit teszi elérhetővé és kereshetővé a felhasználók számára. A szövegeket TEI XML formátumba kódolva, részletes annotációval ellátva tesszük közzé, amely formátum nemcsak a szövegek felhasználóbarát megjelenítését teszi lehetővé, hanem egyben az ehhez tartozó keresőfelület létrehozásának is az alapja. Ez utóbbi segítéségével a drámák nyelvi-grammatikai elemeinek eloszlását tudjuk kvantitatív alapon vizsgálni, ami a szövegek stilometriai elemzése számára jelent elengedhetetlen kiindulópontot. Hiszen így a drámák vagy akár a szereplők megszólalásainak nyelvi felépítése válik összehasonlíthatóvá mind stiláris, mind tematikus szempontból.

A Drámakorpusz ezen túl részét képezi a DraCor nemzetközi adatbázisnak is HunDraCor néven. Ennek köszönhetően a fenti jellemzőkön túl elérhetők a drámák karakterhálózatának, valamint a jelenetekben megszólalók arányának szintén az annotált változatokból automatikusan felrajzolt vizualizációi is. A karakterhálózatokban az egyes karakterek mint csomópontok, a közöttük lévő interakciók mint élek szerepelnek, ami által – a hálózatelmélet belátásait kamatoztatva – egy átfogó kép adható egy dráma szereplőinek egymáshoz fűződő kapcsolatáról. Ekkor a grammatikai elemek gyakoriságvizsgálatával szemben elsősorban nem a szereplők nyelvi kidolgozottságáról, hanem dramaturgiai funkciójáról tudhatunk meg többet.

Abstract

The ELTE Drama Corpus is a continuously expanding database that makes texts from Hungarian drama history available and searchable for users. The texts are encoded in TEI XML format with detailed annotation, which not only allows for a user-friendly presentation of the texts, but also provides the basis of the searches. This allows us to analyze the distribution of linguistic/grammatical features quantitatively, which is an essential starting point for stylometric analysis of the plays, i.e. to compare the linguistic structure of the dramas or even of the characters' utterances from both stylistic and thematic point of view.

The Drama Corpus is also part of the DraCor international database. In addition to the above-mentioned features, here you can find visualizations of the character networks and other metrics regarding to the proportion of speakers in the scenes. In the character networks individual characters are represented as nodes and the interactions between them as edges, which provide a comprehensive picture of the interrelationships in a drama. In contrast to the analysis of the frequency of grammatical elements, with the help of these metrics and visualizations the dramaturgical structure of a play could be grasped computationally.

1. A korpusz létrehozásának szempontjai, általános jellemzők

Az ELTE *Drámakorpusz* az elmúlt években létrehozott *Verskorpusz* és *Regénykorpusz* mintájára egy folyamatosan bővülő adatbázist és egy hozzá tartozó keresőfelületet foglal magában.[1] Az adatbázis a magyar dráma történetének reprezentatív darabjait gyűjti egybe – egyelőre kizárólag a public domain körébe tartozó alkotásokat. A korpusz jelenlegi gyűjteménye két fázisban jött létre: első lépésként az online, a Magyar Elektronikus Könyvtár felületén is elérhető művekből hoztunk létre TEI XML kódolású fájlokat; második lépésként pedig online nem elérhető szövegek digitalizált és optikai karakterfelismeréssel (OCR) feldolgozott változataiból alkottuk meg a TEI XML fájlokat. Az első fázis során 58 drámát tettünk elérhetővé a *Drámakorpusz* felületén. A válogatást két szempont határozta meg: egyrészt a magyar drámairodalom kanonikus műveit kívántuk egybegyűjteni, másrészt törekedtünk arra, hogy egy szerzőtől több alkotást is tartalmazzon a korpusz, hogy ezáltal ne csak az egyes szövegek és műfajok, hanem drámaírói életművek is összehasonlíthatók legyenek.

A kanonikusság szempontját elsősorban színház- és irodalomtörténeti munkák alapján érvényesítettük, mindenekelőtt Kékesi Kun Árpád *Színházi kalauz* című munkájára és a Gintli Tibor főszerkesztésében 2010-ben megjelent *Magyar irodalom* című kötetre támaszkodva.[2] Elképzelésünk szerint azok a drámák, amelyek ezekben a munkákban megjelennek, kanonikusnak számítanak. Ezt a belátást némileg módosította az online elérhetőség kérdése, amely bizonyos értelemben szintén mint kanonizációs folyamat értelmezhető. Ebben az esetben a számunkra elérhető adatbázisok digitalizációs tevékenysége számít a kanonizáció aktusának, az adatbázis mögötti szempontrendszer pedig a kánon keretrendszerének. Ennek tanulságát fontos a saját projektünkre is levonnunk: a *Drámakorpusz* felülete szintén egyfajta kanonizációs gesztusként is érvényesíthető, amely különösen a korpusz bővítésének második fázisára igaz, amely során olyan szövegeket teszünk online elérhetővé, amelyek mindeddig nem rendelkeztek digitális kiadással. Meg kell továbbá említenünk, hogy az online elérhetőség kérdése inkább bővítette és nem szűkítette a színház- és irodalomtörténeteken alapuló címlistát. A korpuszépítés során a Magyar Elektronikus Könyvtár mellett az MTA Irodalomtudományi Intézetének Szövegtárát használtuk a szövegek forrásaiként.[3]

Az első fázis 74 drámája 31 szerző között oszlik el; a művek keletkezési ideje a 16. századtól egészen a 20. századig terjed. A legtöbb, öt vagy hat szöveggel Csiky Gergely, Madách Imre és Jókai Mór, Hunyady László szerepel, míg Babits Mihály, Balázs Béla, Szigligeti Ede és Vörösmarty Mihály négy drámája érhető el a felületen. Mindez összesen 1 141 490 db

---

1 Horváth Péter, *„Az ELTE Verskorpusz automatikus annotációs eljárásai révén nyerhető kvantitatív adattípusok”* in *Nyelvtan, diskurzus, megismerés*, szerk. Simon Gábor, és Tolcsvai Nagy Gábor, 313–331 (Budapest: Eötvös Kiadó, Budapest, 2020).
Horváth Péter, Kundráth Péter, Indig Balázs, Fellegi Zsófia, Szlávich Eszter, Bajzát Tímea Borbála, Sárközi-Lindner Zsófia, Vida Bence, Karabulut Aslihan, Timári Mária, Palkó Gábor, *„ELTE Verskorpusz: a magyar kanonikus költészet gépileg annotált adatbázisa”*, in *XVIII. Magyar Számítógépes Nyelvészeti Konferencia*, szerk. Berend Gábor, Gosztolya Gábor és Vincze Veronika, 375–388 (Szeged: Szegedi Tudományegyetem, Informatikai Intézet, 2022).
Bajzát Tímea, Szemes Botond, Szlávich Eszter, *„Az ELTE DH Regénykorpusz és lehetőségei”*, in *Online térben az online térért: Networkshop 30: országos online konferencia. 2021. április 6-9.*, szerk. Tick József, Kokas Károly és Holl András, 63–72 (Budapest: HUNGARNET Egyesület, 2021).

2 Kékesi Kun Árpád, szerk., *Színházi kalauz* (Budapest: Saxum, 2008).
Gintli Tibor, szerk., *Magyar Irodalom*, (Budapest: Akadémiai Kiadó, 2010).

3 Kecskeméti Gábor, Mészáros Tamás, Bucsics Katalin, Kiss Margit és Markó Veronika, *Nemzeti klasszikusok kritikai kiadásai – A BTK Irodalomtudományi Intézet textológiai portálja*, v. 1.0 (2021. január 1.), szolgáltatja a BTK Irodalomtudományi Intézet,https://szovegtar.iti.mta.hu/. (Utolsó elérés: 2022. 06. 06.)

szóelőfordulást (token) jelent, valamint 871 542 lemmát.[4] A második fázis során az ELTE Egyetemi Könyvtárában elérhető kötetek közül válogattunk a fent bemutatott szempontok szerint, majd a Könyvtár munkatársai által készített szkennelt fájlok OCR-ezését és TEI XML kódolását végeztük el. Az OCR-ezést és a két fázis kódolási munkálatait az ELTE BTK hallgatói végezték el. A második fázis bővítése egyaránt vonatkozik drámaírók gyűjteményes köteteire (ilyen például Bessenyei György és Hunyadi Sándor), valamint válogatáskötetekre, amelyekben több szerző művei találhatók.

Fontos továbbá kiemelni, hogy az ELTE BTK *Drámakorpusza* egyben a DraCor nemzetközi adatbázisának magyar nyelvű alkorpuszát is képezi HunDraCor néven. A DraCor felülete (https://dracor.org/) nem egyszerűen drámák gyűjteményét jelenti, hanem a szövegek elemzéséhez fontos eszközöket is kínál. Ezek közül is kiemelkedik a drámákhoz tartozó karakterhálózatok automatikus felrajzolása, amely hálózatokban az egy jelenetben szereplő karakterek mint csomópontok között húzódik él (ennek vastagsága a közös jelenetekben való előfordulások száma szerint alakul). Az ilyen karakterhálózatok az olvasás vagy az előadás nézésének szükségszerűen időben lezajló folyamatához képest szimultán teszik átláthatóvá a szereplők közötti viszonyrendszereket, az összetartozó csoportokat és az egyes figurák dramaturgiai funkcióit. Mindezen túl lehetőséget biztosítanak arra is, hogy hálózatelméleti mérőszámokkal írjuk le a felrajzolt ábrát (például milyen sűrű az adott hálózat), ami így a szövegek felépítésének számszerű összehasonlítására teremthet alapot.[5] A *Drámakorpusz* felületére készített TEI XML kódok megfelelnek a DraCor által érvényesített specifikációnak, így a két korpusz egységesítése probléma nélkül megtehető.

2. A drámaszövegek jelölőnyelvi kódolása

A korpuszba kerülő szövegek elsődleges forrásai tehát a Magyar Elektronikus Könyvtár adatbázisának szabadon hozzáférhető (public domain) magyar drámaszövegei voltak. Az elérhető fájlformátumok közül elsődlegesen az RTF-kiterjesztésű fájlokat preferáltuk a korpusz létrehozásakor, de azokban az esetekben, amikor ezek elérésére nem volt lehetőségünk, a HTML-formátumot választottuk, illetve némely mű esetében a kétrétegű PDF fájl volt az egyetlen elérhető forrás a jelölőnyelvi kódolást végző annotátorok számára.

A drámakorpusz szövegeinek a kódolási formátuma a TEI XML jelölőnyelv[6] szabványain alapuló és annak eleget tevő specifikáció, amely egyaránt megfelel a DraCor formai követelményének. A TEI XML egyik előnye, hogy eszköz- és rendszerfüggetlen, továbbá a jelölőnyelvi kódolással ellátott szövegek együttesen tárolhatók a metaadatokkal,[7] valamint lehetővé teszi a szövegek olyan strukturálását és annotációját, amelyek hozzájárulnak ahhoz, hogy a szövegek bizonyos nyelvi, tartalmi és metaadatszintű kereshetősége megvalósulhasson.

---

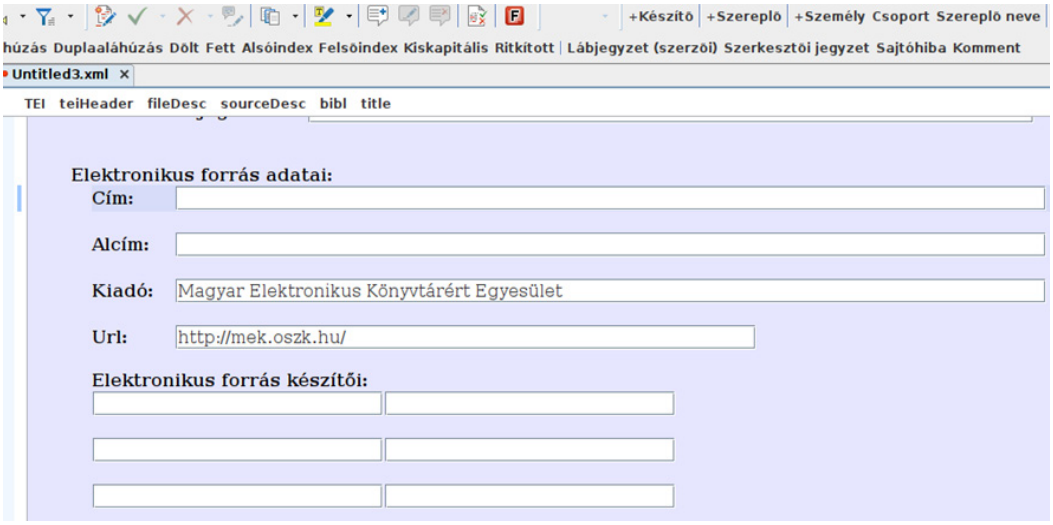
4 A tokenizálás a számítógépes szövegfeldolgozás fontos előkészítő lépése, amely során a szöveget a későbbi strukturálás alapjául vett tokenekre, egységekre bontjuk. A drámakorpusz esetében a token szóalakokat jelöl.

5 Frank Fischer, Ingo Börner, Mathias Göbel, Angelika Hechtl, Christopher Kittel, Carsten Milling, Peer Trilcke, *„Programmable Corpora: Introducing DraCor, an Infrastructure for the Research on European Drama"*, in *Proceedings of DH2019: „Complexities"*, 1–6(Utrecht: Utrecht University, 2019) doi:10.5281/zenodo.4284002.

6 TEI Consortium, eds., *TEI P5: Guidelines for Electronic Text Encoding and Interchange* http://www.tei-c.org/Guidelines/P5/ (Utolsó elérés: 2022. 06. 06.).

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

A jelölőnyelvi kódolás megkönnyítése, illetve a kódolás során potenciálisan fellépő szintaktikai és validitási hibák redukálása érdekében olyan digitális környezetet készítettünk, amely a kódolási munkafázis során vizualizációjában elrejtette az XML jelölőnyelvi struktúrát, és a szabályrendszernek megfelelő címkéket kattintással lehetett a megfelelő szöveghelyekhez rendelni. A <teiHeader> – a drámaszövegek metaadatait – tartalmazó fejlécként funkcionáló kódrészlet pedig adatlapszerűen jelent meg, ahol a megfelelő szövegdobozokat kitöltve lehetett felvinni a művek metaadatait. A kódolást az Oxygen XML Editor[8] programmal végeztük, a „kódmentes" munkakörnyezet elkészítésére is ez a program adott teret, Oxygen Framework formában. A framework egyszerre érvényesíti az általunk írt specifikációnak a validálási szabályait, továbbá felel a kódolási felület vizualizációjáért. A specifikációnak megfelelő validálást egy DTD-formátumú dokumentum végzi el, a fent részletezett elvek alapján. A címkékbe kerülő sztringek (például dátumok) jól formáltságát Schematron ellenőrzi, és hiba esetén üzenetet küld az annotátornak. A vizuális megjelenítés mögött egy CSS-fájl található. A címkék ikonszerű megjelenítését és a munkakörnyezet beállításait egy framework-kiterjesztésű dokumentum tárolja, és ennek szerkesztése ad lehetőséget a munkakörnyezet finomhangolására.



1. ábra. Részlet az Oxygen Framework által biztosított munkakörnyezetről.

Annak érdekében, hogy az annotált drámakorpuszon grammatikai tulajdonságok alapján keresést végezhessünk, szükséges volt a szöveg lemmatizálása, morfológiai, valamint szófaji elemzése. Ehhez az MTA Nyelvtudományi Intézete által fejlesztett e-magyar automatikus nyelvi elemzőrendszert[9] alkalmaztuk, úgy, ahogyan a bevezetőben említett *Verskorpusz* és *Regénykorpusz* esetében is: az e-magyart futtató szkript a drámák szerkezeti annotációit tartalmazó TEI XML fájlokból kiszedte a szöveget, lefuttatta rajtuk az e-magyart, majd a grammatikai annotációkat belerakta a TEI XML fájlokba.

8 https://www.oxygenxml.com (Utolsó elérés: 2022. 06. 06.)

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

3. A lekérdezőfelület

Az ELTE drámakorpuszhoz elérhető egy nyilvános online lekérdezőfelület is, amely funkcióival és arculatával illeszkedik az ELTE Digitális Bölcsészet Tanszék többi szolgáltatása közé. A kereső részletes leírásáról a *Súgó* menüpontban lehet tájékozódni. A keresőfelület elérhető angol nyelven is.

A drámakorpusz esetében elérhető részletes keresőfelület funkciói lehetővé teszik, hogy létrehozzunk alkorpuszokat az adatbázisban található metaadatok alapján. Kereshetünk tokenekre és tokenkapcsolatokra, illetve megválaszthatjuk az adatok feldolgozási és megjelenítési módjait. A *Regénykorpusz*hoz és a *Verskorpusz*hoz képest a *Drámakorpusz* specifikuma az, hogy létrehozhatunk alkorpuszokat a mű szereplőinek megszólalásaiból, de akár a szereplők neme alapján is szűkíthetjük a találatokat.

2. ábra. Az ELTE drámakorpusz keresőfelülete

Ahogyan az a lekérdezőfelületet bemutató 2. ábrából is látható, szerzők és műcímek megadásával szűkíthetjük a vizsgálati korpuszt. Továbbá megadhatunk időintervallumot a drámák keletkezési idejére vonatkozóan.

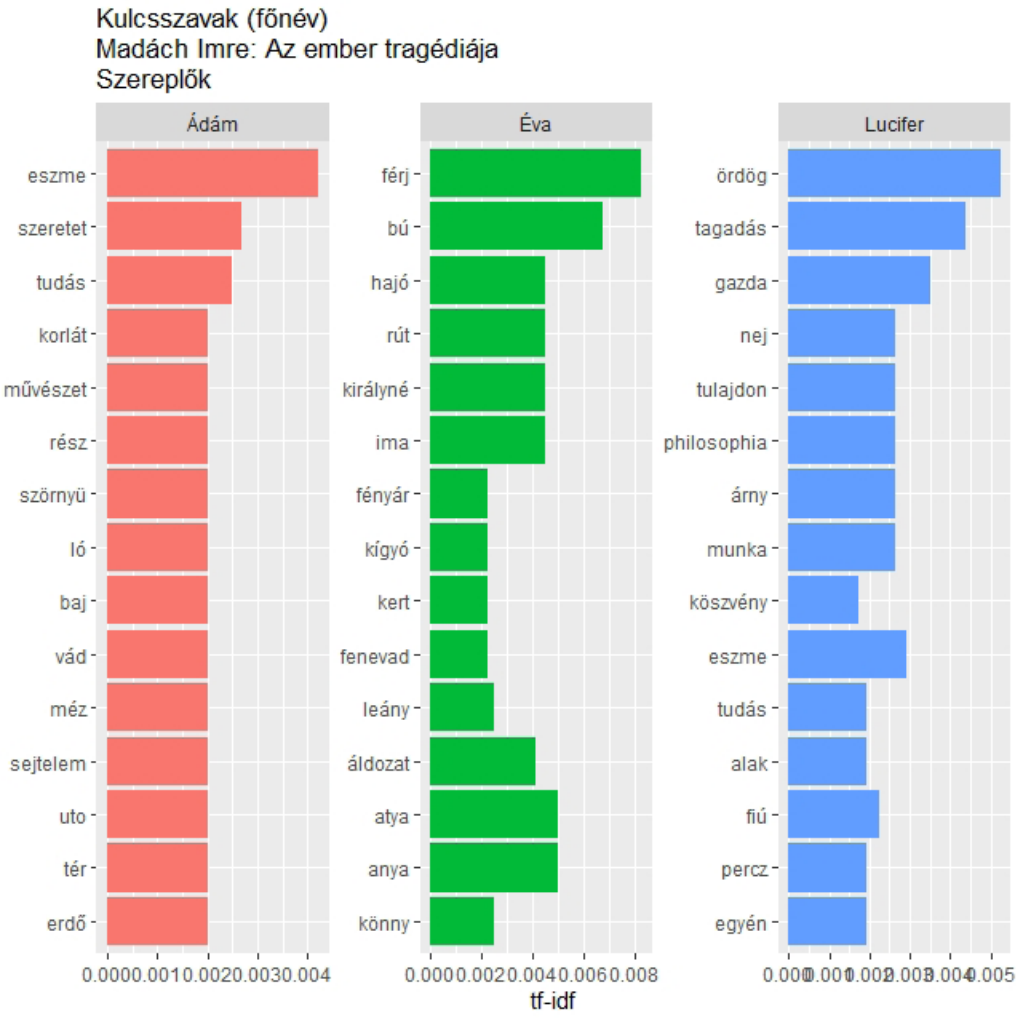
A tokenekre, illetve tokenkapcsolatokra való szűrés lehetővé teszi konkrét szóalakok vagy szótövek keresését. Nem csupán szóalakokra és szótövekre kereshetünk, hanem szófaji és morfológiai kategóriákra is, aminek révén konstrukciós mintázatok is adatolhatók. Több tokenre való keresés esetén megadható, hogy a tokenek milyen típusú szövegegységen belül, illetve egymástól maximálisan mekkora távolságra forduljanak elő.

A találatok megjelenítése többféle módon lehetséges, érdemes azt az opciót kiválasztani, amely a leginkább garantálja az adatok előkészítését a további feldolgozásokhoz. Egyrészt megadható, hogy a találatokat mekkora kontextusban szeretnénk lekérni, illetve tetszőlegesen választható, hogy hány találatot 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 előállított listák TSV-formátumban tölthetők le.

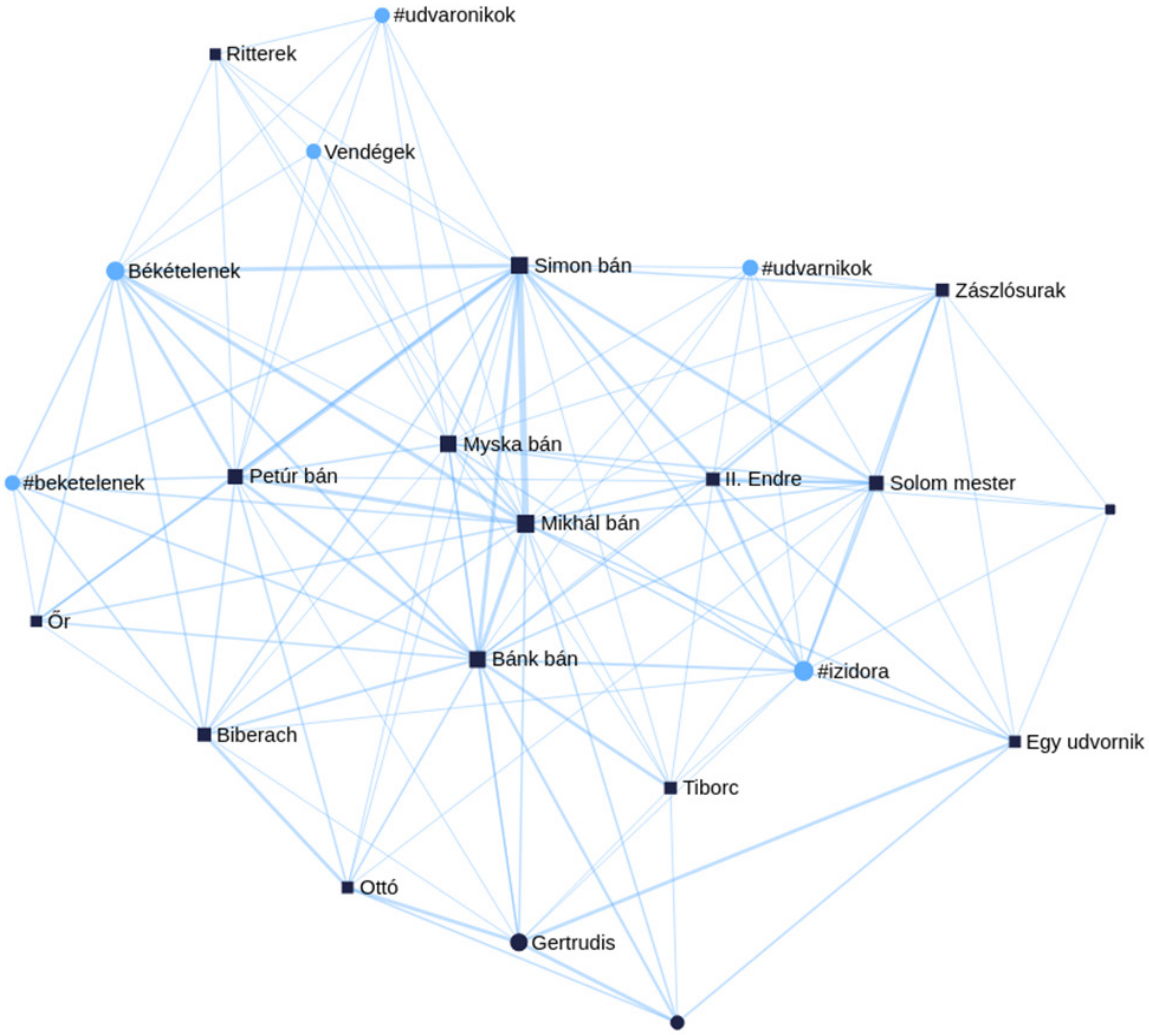
4. Lehetséges felhasználás

Azáltal, hogy nem csupán a drámák szövegét, hanem azok nyelvi elemzővel annotált változatát tárolja az adatbázis, számos, a drámák szövegére vonatkozó keresést egyszerűbben végezhetünk el, mint korábban, illetve olyan módon is kereshetünk, amelyre nem lenne lehetőség abban az esetben, ha a korpusz csupán a drámaszövegeket tartalmazná. Például ha kíváncsiak vagyunk arra, hogy az „*ördög*” szó mely művekben, milyen szövegkörnyezetben fordul elő, akkor nem kell külön rákeresnünk a szó összes alakjára (*ördög*, *ördögök*, *ördögöt*, *ördögöket* stb.), hanem a szavak szótári alakjának az adatbázisban való szerepeltetése révén egy kereséssel listázhatjuk az összes ilyen szöveghelyet. Mivel a szavak szófaját és morfoszintaktikai jellemzőit is tartalmazza az adatbázis, nem csupán szavakra, hanem grammatikai jellemzőkre is kereshetünk. Például rákereshetünk az összes olyan szöveghelyre, amely a melléknév + *ördög* szerkezetet tartalmazza, de akár a középfokú melléknév + *ördög* szerkezeteket is listázhatjuk. Ezek a keresések elsősorban olyan gyakoriságvizsgálatok számára teremtenek alapot, amelyeken keresztül az egyes művek tematikus és stiláris szerveződésére is következtetni lehet (például a leggyakrabban használt főnevek listája), és ezáltal a drámák vagy a drámatörténet kvantitatív szempontú megértéséhez járulhatnak hozzá. A korábbi korpuszokhoz képest a *Drámakorpusz* sajátossága, hogy nemcsak teljes szövegek, hanem az egyes szereplők nyelvének összehasonlítását is lehetővé teszi. A 3. ábrán *Az ember tragédiája* három szereplőjének a főnevek közül kikerülő kulcsszavait mutatjuk be példaként.



3. ábra: Az ember tragédiája *három szereplőjének a főnevek közül kikerülő kulcsszavai a tf-idf módszer alapján*

A DraCor korábban említett adatbázisa mindezek mellett más típusú információkat is elérhetővé tesz. A DraCor használatával elsősorban ugyanis nem a szereplők nyelvi kidolgozottságáról, hanem dramaturgiai funkciójáról tudhatunk meg többet: mennyire tekinthető egy karakter központi szereplőnek, hány másik szereplővel tartja a kapcsolatot, mennyiben nélkülözhető a hálózat stabilitásának szempontjából stb. A jelenetekben megszólalók száma a dráma felépítéséről (például tömegjelenetek a dráma elején és/vagy végén) tanúskodik, illetve a monológok, dialógok és csoportos jelenetek arányát mutatja. A 4. ábra a Bánk bán karakterhálózatát mutatja, amelyet a Gephi vizualizációs szoftverben hoztunk létre a DraCor adatai alapján.



4. ábra: A Bánk bán karakterhálózata a DraCor felületén[10]

5. Függelék

Az ELTE Drámakorpusz lekérdezőfelülete elérhető: https://dramakorpusz.elte-dh.hu/

A drámák kódolt szövegei elérhetők: https://github.com/ELTE-DH/drama-corpus

6. Bibliográfia

Bajzát Tímea, Szemes Botond és Szlávich Eszter. „Az ELTE DH Regénykorpusz és lehetőségei", In *Online térben az online térért: Networkshop 30: országos online konferencia. 2021. április 6-9.*, szerkesztette Tick József, Kokas Károly és Holl András, 63–72. Budapest: HUNGARNET Egyesület, 2021. https://doi.org/10.31915/NWS.2021.7

Fischer, Frank, Börner, Ingo, Göbel, Mathias, Hechtl, Angelika, Kittel, Christopher, Milling, Carsten and Trilcke, Peer, „Programmable Corpora: Introducing DraCor, an Infrastructure for the Research on European Drama", In *Proceedings of DH2019: „Complexities"*, 1–6. Utrecht: Utrecht University, 2019.

10 https://dracor.org/hun/katona-bank-ban

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

Horváth Péter, Kundráth Péter, Indig Balázs, Fellegi Zsófia, Szlávich Eszter, Bajzát Tímea Borbála, Sárközi-Lindner Zsófia, Vida Bence, Karabulut Aslihan, Timári Mária és Palkó Gábor, *„ELTE Verskorpusz: a magyar kanonikus költészet gépileg annotált adatbázisa"*, In *XVIII. Magyar Számítógépes Nyelvészeti Konferencia*, szerkesztette Berend Gábor, Gosztolya Gábor és Vincze Veronika, 375–388. Szeged: Szegedi Tudományegyetem, Informatikai Intézet, 2022.

Kecskeméti Gábor, Mészáros Tamás, Bucsics Katalin, Kiss Margit és Markó Veronika, *Nemzeti klasszikusok kritikai kiadásai – A BTK Irodalomtudományi Intézet textológiai portálja*, v. 1.0 (2021. január 1.), szolgáltatja a BTK Irodalomtudományi Intézet, https://szovegtar.iti.mta.hu/. (Utolsó elérés: 2022. 06. 06.)

Indig Balázs, Sass Bálint, Simon Eszter, Mittelholcz Iván, Kundráth Péter és Vadász Noémi, „emtsv – egy formátum mind felett", In *XV. Magyar Számítógépes Nyelvészeti Konferencia*, szerkesztette Berend Gábor, Gosztolya Gábor és 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*, szerkesztette Boda István és Mónos Katalin, 65–72. Debrecen: MANYE, Debreceni Egyetem, 2011.

Gintli Tibor, szerk. Magyar Irodalom, Budapest: Akadémiai Kiadó, 2010.

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.

Kékesi Kun András, szerk. Színházi kalauz, Budapest: Saxum, 2008.

TEI Consortium, *TEI P5, eds. Guidelines for Electronic Text Encoding and Interchange*. TEI Consortium. http://www.tei-c.org/Guidelines/P5/ (Utolsó elérés: 2022. 06. 06.)