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.2022.35

ELTE Népdalkorpusz – magyar népdalok gépileg annotált adatbázisa

Horváth Péter
Eötvös Loránd Tudományegyetem, Digitális Örökség Nemzeti Laboratórium
horvath.peeteer@gmail.com

Kundráth Péter
peter.kundrath@gmail.com

Palkó Gábor
Eötvös Loránd Tudományegyetem, Digitális Örökség Nemzeti Laboratórium
palko.gabor@btk.elte.hu

ELTE Folk Song Corpus is a database that stores Hungarian folk songs with automatically generated annotations of the folk songs' structural units, grammatical features and sound devices. In the annotation process we followed the workflow developed for the ELTE Poetry Corpus. The corpus has an open access online query tool with several search functions. Besides the annotation process and the query tool of the corpus, the paper presents a quantitative comparison of the ELTE Folk Song Corpus and the ELTE Poetry corpus on the basis of lexical features, rhyme patterns and metrical properties.

Keywords: folk songs, corpus, automatic annotation, ELTE Poetry Corpus

1. A korpusz főbb jellemzői és létrehozásának lépései

Az ELTE Népdalkorpusz az ELTE Verskorpusz mintájára létrehozott, magyar népdalokat tartalmazó, automatikusan annotált adatbázis. A korpusz forrását az Ortutay Gyula szerkesztésében és Katona Imre válogatásával megjelent Magyar népdalok című gyűjteményes mű második, 1976-os kiadása adta, amely megtalálható a Magyar Elektronikus Könyvtár adatbázisában. A korpuszban szereplő népdalok száma 2390, a korpusz szavainak száma 110 ezer, a tokenek száma pedig 150 ezer. A korpusz a népdalok szövegei mellett három annotációs réteget tartalmaz: annotáltuk a népdalok szerkezeti egységeit, a szavak grammatikai tulajdonságait, valamint a hangzásjellemzőkhöz kapcsolódó poétikai jellemzők bizonyos típusait.

A korpusz létrehozása során az ELTE Verskorpusz létrehozásához kidolgozott módszertant követtük (Horváth et al. 2022).[1] Első lépésben egy szkript segítségével a Magyar Elektronikus Könyvtár oldaláról letöltött HTML fájlokat átalakítottuk olyan TEI XML fájlokká, amelyek tartalmazzák a szövegek szerkezeti egységeinek, azaz a címeknek, a strófáknak és a soroknak az annotációit. Ezt követően az e-magyar emtsv változatával tokenizáltuk a szövegeket, és annotáltuk a szavak szótári alakját, szófaját és morfoszintaktikai jellemzőit (Váradi et al. 2017, Indig et al. 2019). A korpusz létrehozásának harmadik lépése a hangzásjellemzőkhöz kapcsolódó poétikai tulajdonságok automatikus annotálása volt, amelyhez az ELTE Verskorpusz hangzásjellemzőinek az annotálására fejlesztett programot használtuk (Horváth 2020). A hangzásjellemzők annotálása a versszakok rímképletének, a rímpárt alkotó szavaknak, a sorok időmértékes ritmusának, az alliterációknak és a szavak fontosabb fonológiai

1 https://github.com/ELTE-DH/poetry-corpus

jellemzőinek, a szótagszámnak, a hangrendnek és a fonológiai szerkezetnek a felismertetésére terjedt ki. Természetesen az időmértékes ritmusnak népdalok esetében nincsen túl nagy relevanciája. Az erre vonatkozó annotációkat azért nem töröltük, mert egyelőre nem akartunk eltérni az ELTE Verskorpusz esetében alkalmazott formátumtól. Végezetül egy XSLT stíluslap segítségével elvégeztünk a TEI XML fájlokon egy formátumátalakítást, azaz bizonyos XML-elemeket és -attribútumokat áthelyeztünk, illetve átneveztünk, valamint további, nagyrészt a különböző szerkezeti egységek szó- és szótagszámára vonatkozó annotációkkal bővítettük a fájlokat. Az így előállt XML-fájlok bár TEI-közeliek, de nem felelnek meg a TEI által megadott szabványnak. A formátumátalakítással és az annotációk bővítésével az volt a célunk, hogy a korpuszhoz a lekérdezéseket egyszerűbben meg lehessen írni, és gyorsabban le lehessen futtatni.

A fent bemutatott lépések során előálló, egyre több annotációs réteget tartalmazó különböző verziók a korpusz github oldaláról[2] szabadon letölthetők. A github oldalon és az ELTE Verskorpuszt bemutató tanulmányban (Horváth et al. 2022) részletes leírás olvasható a korpusz egyes verzióiban szereplő XML-elemekről és -attribútumokról.

2. A korpusz lekérdezőfelülete

A korpuszban való keresésekhez egy szabadon elérhető online lekérdezőeszközt fejlesztettünk a Verskorpusz lekérdezőeszközének mintájára. A korpusz szövegeit és annotációit tartalmazó XML-fájlokból egy MariaDB-alapú SQL-adatbázist hoztunk létre. Ebben keres a https://verskorpusz.elte-dh.hu/nepdal címen elérhető lekérdezőeszköz. A lekérdezőeszköz felületén számos keresési lehetőség közül választhatunk. Kereshetünk szóalakokra, lemmákra, morfoszintaktikai jellemzőkre, szótagszámra, hangrendre, szótagok hosszúságára, fonológiai szerkezetre, valamint ezek tetszőleges kombinációira. Úgyszintén kereshetünk ugyanezen jellemzők alapján több szóból álló szerkezetekre. Generálhatunk gyakorisági listákat is szavakra vagy szószerkezetekre vonatkozóan, szóalakok vagy lemmák formájában, valamint szűrhetjük a népdalokat rímképletek alapján. A keresési eredményeket letölthetjük TSV-formátumban, amely a legtöbb táblázatkezelő programban megnyitható. Ha megnyitunk egy népdalt, akkor a felületen láthatjuk a népdal összes annotált jellemzőjét is.

3. A Népdalkorpusz és a Verskorpusz néhány lexikai jellemzője

Az alábbiakban a Népdalkorpusz és a Verskorpusz különböző lexikai jellemzőinek a kvantitatív összevetését mutatjuk be, érzékeltetve a Népdalkorpusz vizsgálatában rejlő lehetőségeket. Hangsúlyozandó, hogy míg az ELTE Verskorpusz a 20. század első feléig bezárólag tartalmazza a magyar kanonikus költészet legnagyobb részét, addig az ELTE Népdalkorpusz forrásául szolgáló gyűjteményes kötet anyaga egy önkényes, szubjektív szempontokat érvényesítő válogatás eredményeként jött létre. Mindez azt jelenti, hogy a Népdalkorpuszból kinyert kvantitatív adatokon alapuló következtetéseket inkább erős hipotézisekként érdemes kezelni. Az 1. táblázat a Verskorpusz és a Népdalkorpusz leggyakoribb tíz főnévi és igei lemmáját mutatja be az előfordulási számokkal. Félkövérrel emeltük ki azokat a lemmákat, amelyek az első tízben csak az egyik korpuszban fordulnak elő. A gyakorisági listákat a lekérdezőfelülettel generáltuk.

2 https://github.com/ELTE-DH/folk-song-corpus

	FŐNEVEK				IGÉK			
	Verskorpusz (2,7 millió szó)		Népdalkorpusz (114 ezer szó)		Verskorpusz (2,7 millió szó)		Népdalkorpusz (114 ezer szó)	
1	szív	8235	isten	506	van	28515	van	1750
2	**lélek**	7461	**rózsa**	450	lesz	10940	lesz	582
3	**szem**	7204	szív	387	**lát**	7985	**ad**	481
4	isten	6610	**baba**	356	tud	6706	nincs	408
5	**élet**	6599	**úr**	304	**él**	5346	megy	407
6	**ég**	6251	**anya**	286	nincs	5213	**szeret**	370
7	**világ**	5773	**lány**	277	**néz**	5112	**jár**	365
8	**föld**	5544	**legény**	271	megy	4819	jön	344
9	nap	5520	**ló**	266	jön	4796	tud	323
10	**szó**	5000	nap	242	**mond**	4734	**kell**	318

1. táblázat. A Verskorpusz és a Népdalkorpusz leggyakoribb főnévi és igei lemmái

A táblázatból látható, hogy a Verskorpusz leggyakoribb főnevei között nagyrészt absztraktabb, azaz nem tapintható, és sok esetben nem is látható entitásokra vonatkozó szavak jelennek meg, míg a Népdalkorpusz esetében előtérbe kerülnek a konkrétabb jelentésű főnevek. Ez utóbbiak sok esetben személyek által betöltött, a legalapvetőbb emberi kapcsolatok kontextusában értelmezhető szerepre vonatkoznak (*baba*, *anya*, *lány*, *legény*). Érdekes, hogy a Verskorpusz esetében egy személyre vonatkozó főnév sem szerepel a leggyakoribb tíz között. Az igék esetében a legszembetűnőbb különbség az, hogy míg a Verskorpusz esetében két perceptuális jelentésű, látáshoz kapcsolódó ige is megjelenik (*lát*, *néz*), addig a Népdalkorpusz esetében egy ilyet sem találunk a leggyakoribb tíz ige között. Mindebből arra következtethetünk, hogy a magyar kanonikus költészetben a megfigyelő, szemlélődő attitűd erőteljesebben jelenik meg, mint a magyar népdalokban.

A két korpusz szókincsét nemcsak gyakorisági listák alapján, hanem kulcsszavak alapján is összevethetjük. Kulcsszavak azok a szavak, amelyek egy adott korpuszban nagyobb arányban fordulnak elő, mint egy másik korpuszban. A kulcsszavak tehát olyan szavak, amelyek egy adott korpuszt a legjobban jellemeznek egy másik korpuszhoz képest. Kulcsszavak kinyerésére számos számítás létezik, mi a Kilgarriff-féle *simple math* nevű eljárást implementáltuk. Az eredeti módszertől annyiban eltértünk, hogy a lemmák relatív gyakoriságát nem az alapján számoltuk, hogy egy szó összesen hányszor szerepel a két korpuszban, hanem az alapján, hogy azok hány darab versben, illetve népdalban fordulnak elő. A számítást olyan módon alkalmaztuk, hogy a nagyobb gyakorisággal megjelenő szavakra fókuszálva kérdezhessünk le kulcsszavakat (Kilgarriff 2009).[3] A 2. táblázat mutatja be az alkalmazott számítás alapján a Verskorpusz és a Népdalkorpusz tíz legnagyobb kulcsszóértékkel rendelkező szavát, vagyis a nagyobb gyakoriságú szavak közül azokat, amelyek a Verskorpuszban arányaiban jelentősen több szövegben fordulnak elő, mint a Népdalkorpuszban, illetve amelyek a Népdalkorpuszban arányaiban jelentősen több szövegben fordulnak elő, mint a Verskorpuszban. A kulcsszavak mellett feltüntettük, hogy azok hány versben, illetve népdalban fordulnak elő.

3 A kulcsszavak számítása során 1000 versenkénti előfordulásra normalizáltuk a szóelőfordulásokat. A simple math paramétert 100 értékben határoztuk meg, aminek köszönhetően a nagyobb gyakoriságú, azaz a sok versben és népdalban megjelenő szavakra fókuszálva tudtuk elvégezni a kulcsszavak kinyerését.

	Verskorpusz (13 064 vers)		Népdalkorpusz (2390 népdal)	
1	s	10429	baba	247
2	és	9113	édesanya	169
3	amely	3418	legény	208
4	e	3932	kislány	105
5	lélek	3642	rózsa	295
6	mint	6225	szerető	174
7	ég	3966	ló	170
8	élet	3671	katona	108
9	minden	4478	mög	44
10	álom	2118	Kossuth[4]	47

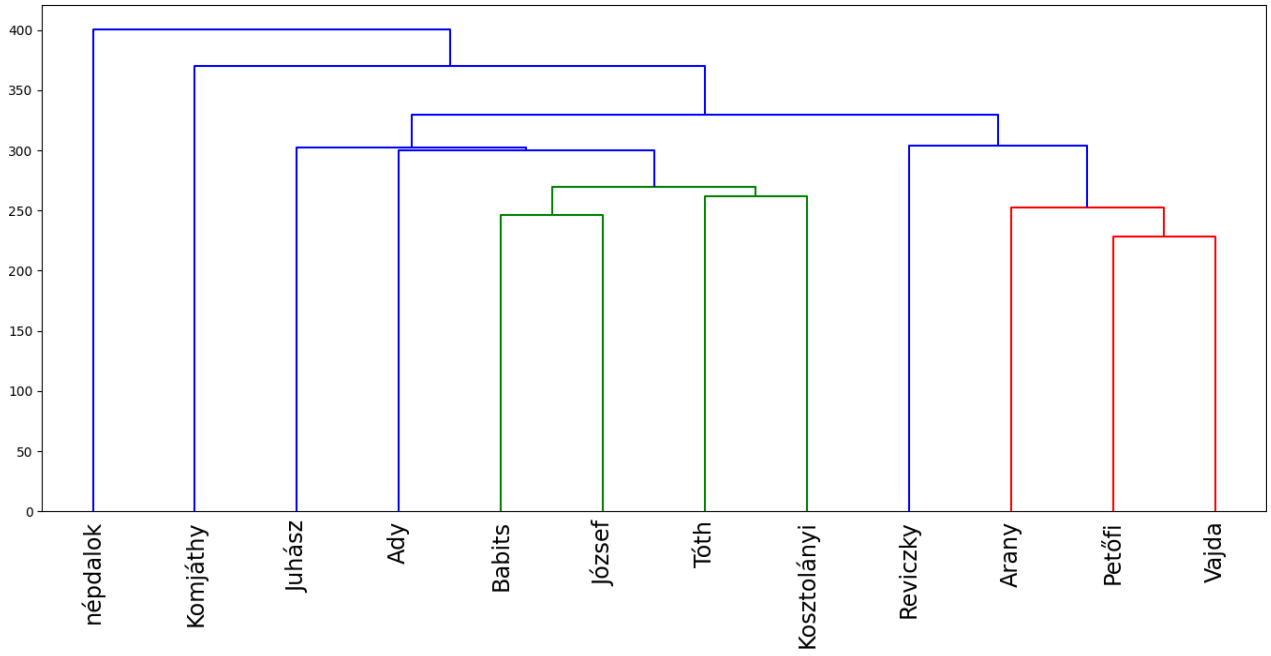
2. táblázat. A Verskorpusz és a Népdalkorpusz kulcsszavai

A két korpusz tíz legnagyobb kulcsszóértékével rendelkező szavának a listája hasonló tendenciát mutat, mint a főnevek gyakorisági listái. A táblázatból látható, hogy míg a Verskorpusz kulcsszavai között megjelenő főnevek (*lélek*, *ég*, *élet*, *álom*) absztrakt jelentésűek, addig a Népdalkorpusz első tíz kulcsszava nagyrészt konkrét jelentésű szó, amelyek többnyire valamilyen személy által betölthető szerepre utalnak. A táblázatból azt is láthatjuk, hogy a Verskorpuszban számos grammatikai szó is kulcsszóként jelenik meg, amelyek utalhatnak az összetettebb, illetve retorikusabb szövegszerkesztésre, például a mondaton belüli felsorolások és mellérendelő tagmondatkapcsolatok (*s*, *és*), az alárendelő tagmondatok (*amely*), a retorikai kérdések és rámutatások (*e*) vagy a hasonlatok (*mint*) nagyobb mértékű jelenlétére.

Érdekes kérdés az is, hogy a népdalok szókincse és bizonyos költők szókincse hogyan viszonyul egymáshoz, vajon vannak-e olyan költők, akiknek a szókincse közelebb áll a magyar népdalokéhoz, mint más költőkéhez. Ehhez elvégeztünk egy agglomeratív hierarchikus klaszterelemzést, amelyhez alkorpuszként a Népdalkorpuszt, valamint a Verskorpuszból kiszedett 11 költő összes versét használtuk, akik a következők voltak: Arany János, Petőfi Sándor, Vajda János, Reviczky Gyula, Komjáthy Jenő, Ady Endre, Juhász Gyula, Babits Mihály, Kosztolányi Dezső, Tóth Árpád, József Attila. A klaszterezéshez egy saját szkriptet használtunk, amely a leggyakoribb 1000 lemma alapján, a Burrows' delta eljárást alkalmazva adta meg az 1. ábrán látható kimenetet (Burrows 2002).[5]

4 A *Kossuth* tulajdonnév azért szerepel ennyi szövegben, mert a korpusz forrásául szolgáló gyűjteményes kötet tartalmaz egy „48-as dalok" tematikájú részt.

5 A 12 alkorpusz távolságának a kiszámításához használt Burrows' delta eljárás a szavak (jelen esetben lemmák) normalizált előfordulásainak standardizált értékei által kijelölt, az alkorpuszokat reprezentáló pontok közötti Manhattan-távolságon alapul. A klaszterek nagyobb klaszterekbe való sorolásához a klasztereket alkotó adatpontok közötti átlagos távolságot használtuk.



1. ábra A Népdalkorpusz és a Verskorpusz 11 költőjének hierarchikus agglomeratív klaszterelemzése

Az 1. ábrán látható, hogy szemben a megfogalmazott hipotézissel, egyik költő sem került a népdalokkal egy közös, alacsonyabb szintű klaszterbe. A klaszterelemzés eredményéből úgy tűnik, hogy a népdalok és a magyar kanonikus költészet szókincse markánsan elválik egymástól. Még az olyan, népies megszólalásmódokkal kísérletező költők, mint Petőfi vagy Arany szókincse is jobban hasonlít más költők szókincsére, mint a Népdalkorpusz szókincsére. Megjegyzendő, hogy a klaszterelemzés Komjáthyt leszámítva jól el tudta különíteni egymástól a 19. és a 20. századi költőket.

4. A Népdalkorpusz és a Verskorpusz néhány hangzásjellemzője

Mivel a Népdalkorpuszban és a Verskorpuszban nemcsak a szavak grammatikai jellemzőit, hanem a szövegek bizonyos hangzásjellemzőit is annotáltuk, ez utóbbiak alapján is összevethető a két korpusz. A 3. táblázat a két korpusz tíz leggyakoribb rímképletét mutatja be. Egy adott rímképlet előfordulásába csak azokat a verseket, illetve népdalokat számoltuk bele, amelyeknek mindegyik strófája ugyanazzal a rímképlettel lett annotálva.

	Verskorpusz (13 064 vers)		Népdalkorpusz (2390 népdal)	
	Rímképlet	Előford.	Rímképlet	Előford.
1	abcb	520	aabb	411
2	abab	458	aa	100
3	aa	305	aaaa	87
4	aabb	300	aabc	86
5	ab	209	abcc	70
6	abcd	186	aaba	27
7	aba	115	abcb	27
8	aab	74	aabbcc	24
9	aaaa	66	aaa	15
10	abba	65	aabbc	15
Összes		4959		1318

3. táblázat. A Verskorpusz és a Népdalkorpusz leggyakoribb rímképletei

A 3. táblázatból látható, hogy míg a Verskorpusz első két leggyakoribb rímképlete a félrím és a keresztrím, addig a Népdalkorpusz esetében a páros- és a bokorrímek vannak a lista élén, és csak a lista 6. helyén jelenik meg az első olyan rímképlet, ahol az egymással rímelő sorok közé beékelődik egy azokkal nem rímelő sor (aaba). A kapott eredmények megerősítik azt a már Arany János által is hangoztatott meglátást, miszerint „[o]da törekedvén a népköltészet, hogy az értelemnek mielőbb teljességet adjon, nem fűzi hosszan gondolatját, hanem siet azt befejezni, mi külsőleg a második rím által történik. Ezért nem kaphatott lábra népdalainkban a váltogató, vagy keresztrím, mint a mely hosszabb elnyújtását a gondolatnak engedi meg” (Arany 2012, 315).

A Verskorpusz és a Népdalkorpusz jelenlegi verziója csak az időmértékes ritmusra, azaz a hosszú és rövid szótagokra vonatkozó annotációkat tartalmaz, metrumra vonatkozó annotációkat nem. Az utóbbi időben azonban elkészült egy program, amely képes az alapvetőbb metrumok automatikus felismerésére (Horváth 2021). E program futtatásával például megtudhatjuk, hogy melyek a leggyakoribb ütemhangsúlyos metrumok a Verskorpuszban és a Népdalkorpuszban. Ezt a 4. táblázat mutatja be.[6]

	Verskorpusz (13 064 vers)		Népdalkorpusz (2390 népdal)	
	Ritmus	Előford.	Ritmus	Előford.
1	6 | 6	622	4 | 4	270
2	4 | 4	305	6	160
3	6	229	4 | 4 | 3	150
4	5 | 5	185	4 | 3	105
5	5 | 6	166	6 | 6	91
6	4 | 6	138	2 | 2 | 4	61
7	4 | 4 és 4 | 3	120	4 | 6	52
8	5 | 4	79	2 | 2 | 3	45
9	5 | 3	64	4 | 2 | 2	30
10	5 | 6 és 5 | 5	57	4 | 1 | 3	27

4. táblázat. A Verskorpusz és a Népdalkorpusz leggyakoribb ütemhangsúlyos metrumai

6 A programot a következő beállításokkal futtattuk: megegyező szótagszámú sorok és ütemkezdetre eső szó eleji szótagok minimális aránya: 0,75; ütemek maximális szótagszáma: 6.

A 4. táblázatból látható, hogy a Verskorpusz esetében a leggyakoribb ütemhangsúlyos metrum a felező tizenkettes, a második leggyakoribb metrum pedig a felező nyolcas. A Népdalkorpusz esetében a leggyakoribb ütemhangsúlyos metrum a felező nyolcas, a második leggyakoribb metrumot pedig azok a szövegek adják, amelyek hat szótagos sorokból állnak, és a sorok további ütemekre nem oszthatók. Megjegyzendő, hogy ez utóbbi eset felező tizenkettesnek is tekinthető. A két korpusz leggyakoribb metrumai között két feltűnőbb különbséget találhatunk. Egyrészt a Verskorpusz esetében a hetedik és a tizedik helyen is megjelenik olyan metrum, ahol a páratlan és a páros sorokban eltérő ütemosztás található. Például a hetedik helyen szereplő 4 | 4 és 4 | 3 azt jelenti, hogy a versek páratlan sorai felező nyolcasok a páros sorok pedig kétütemű hetesek. Ilyen típusú, a páratlan és a páros sorok eltérő ütemezésére épülő metrumok a Népdalkorpusz esetében egyáltalán nem fordulnak elő a leggyakoribb tíz metrum között. Ez a különbség természetesen szorosan összefügg azzal a rímképletek kapcsán említett különbséggel, hogy a népdalok monolitikusabb rímelésével szemben a Verskorpuszra nagymértékben jellemzőek a félrímes, illetve keresztrímes megoldások. A másik szembetűnő különbség az, hogy a Népdalkorpusz esetében a leggyakoribb tíz metrumból öt három ütemből áll. A Verskorpusz esetében ilyen típusú metrum egyáltalán nem jelenik meg a leggyakoribb tíz között.

5. Összefoglalás

Az ELTE Népdalkorpusz egy olyan automatikusan annotált, szabadon elérhető adatbázis, amely reményeink szerint mind a magyar népdalokra, illetve a magyar költészetre vonatkozó kutatásokban, mind a közoktatásban hasznosítható lesz. A korpuszhoz fejlesztett lekérdezőfelület lehetővé teszi, hogy a felhasználó mélyebb informatikai tudás nélkül is különböző, a korpuszt jellemző kvantitatív adatokhoz jusson. A népdalok szövegei és az annotációkat tartalmazó XML-fájlok szabadon letölthetőek kutatás céljából a korpusz github oldaláról. A tanulmányban a Népdalkorpuszt a szókincs, valamint a rím és a metrum különböző kvantitatív jellemzői alapján vetettük össze az ELTE Verskorpusszal, rámutatva a két szövegtípus néhány különbségére és a Népdalkorpusz használatának a lehetőségeire. Az ELTE Népdalkorpusz nem egy lezárt projekt, a jövőben szeretnénk az adatbázist további népdalokkal, valamint további annotációs rétegekkel bővíteni.

Bibliográfia

Arany János: A magyar nemzeti vers-idomról. In Arany János: *Tanulmányok és kritikák I.* vál., szerk., az utószót és a jegyzeteket írta S. Varga Pál. Második, javított kiadás. Debrecen, 2012, Debreceni Egyetemi Kiadó. 288–320.

Burrows, John: 'Delta': a measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing*, 2002. vol. 17. No. 3. 267–287. https://doi.org/10.1093/llc/17.3.267

Horváth Péter: A vershangzás jellemzőinek automatikus feltárása József Attila verseiben. *Digitális Bölcsészet*, 2020. 3. sz. M:3–M:27. https://doi.org/10.31400/dh-hun.2020.3.422

Horváth Péter: Két eljárás magyar nyelvű versek metrumának gépi felismertetéséhez. *Digitális Bölcsészet*, 2021. 4. sz. T:79–T:103. https://doi.org/10.31400/dh-hun.2021.4.2361

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*, szerkesztette Berend Gábor, Gosztolya Gábor, Vincze Veronika. Szeged, 2022, Szegedi Tudományegyetem TTIK, Informatikai Intézet. 375–388.

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. Szeged, 2019, Szegedi Tudományegyetem TTIK, Informatikai Intézet. 235–247.

Kilgarriff, Adam: Simple math for keywords. In *Proceedings of the Corpus Linguistics Conference 2009 (CL2009). Held at the University of Liverpool, UK, 20-23 July 2009*, edited by Michaela Mahlberg, Victorina González-Díaz, Catherine Smith. https://ucrel.lancs.ac.uk/publications/cl2009/171_FullPaper.doc

Váradi Tamás, Simon Eszter, Sass Bálint, Gerőcs Mátyás, Mittelholtz Iván, Novák Attila, Indig Balázs, Prószéky Gábor, Farkas Richárd, Vincze Veronika: Az e-magyar digitális nyelvfeldolgozó rendszer. In *XIII. Magyar Számítógépes Nyelvészeti Konferencia*, szerkesztette Vincze Veronika. Szeged, 2017, Szegedi Tudományegyetem, Informatikai Intézet. 49–60.