Dodé Réka – Ludányi Zsófia – Falyuna Nóra – Kuna Ágnes

Poétika és korpusz. Hogyan nyújthat segítséget a korpusznyelvészet poétikus szövegek vizsgálatához?[1]

1. Bevezetés

A nyelvleírás, valamint a nyelvészeti kutatások egyre szélesebb körben támaszkodnak elektronikus korpuszokra. Ez megfigyelhető a szótárkészítésben, a nyelvtanuló és fordítástámogató programokban, a helyesírási szótárak anyagának összeállításában és még számos területen. Jelen tanulmányunkban arra keressük a választ, hogy miképpen hozható összefüggésbe a nyelvtechnológiai értelemben használt korpusz és a poétikai kutatás. Cikkünk célja, hogy áttekintést adjon a nemzetközi poétikai korpuszokról, azok típusairól, valamint az annotálás szempontjairól és gyakorlatáról.

A célkitűzésnek megfelelően a tanulmányban jelen bevezető részt (1.) követően bemutatjuk röviden a korpusz értelmezési lehetőségeit és típusait (2.). Majd körvonalazzuk egy tervezett magyar lírai korpusz összeállításának néhány kérdését és a kognitív funkcionális elméleti háttér szerepét (3.1.). Ezt követően mutatjuk be részletesen a nemzetközi irodalomban megtalálható poétikai korpuszok típusait és számos hozzájuk kapcsolódó kutatást (3.2.). Mindezt azzal a céllal, hogy az áttekintés kiindulási alapként, illetve egyfajta referenciapontként szolgáljon a Stíluskutató csoport által tervezett magyar nyelvű lírai korpusz kialakításához és a korpusz nemzetközi pozicionálásához is (Domonkosi et al. 2018).

Jelen áttekintő tanulmány célja tehát, hogy a tervezett Magyar lírakorpusz építésének módszertani elveinek kidolgozásához hozzájáruljon, és a tervezett korpuszt a magyar és nemzetközi viszonylatban is elhelyezze.

Mivel a poétikai vonatkozású korpuszok tipikusan lírai szövegekkel dolgoznak, tanulmányunk is ezekre fókuszál. Ez az oka, hogy – némi leegyszerűsítéssel élve – a dolgozat a poétikai, illetve lírai korpusz terminusokat szinonim értelemben használja.

2. A korpuszok

A nyelvészetben a *korpusz* szó több jelentésben is használatos. Az Idegen szavak szótára (Tolcsvai Nagy 2007) – számos más területen használt jelentés mellett – az alábbi két értelmezését adja meg:

[1] Köszönettel tartozunk Simon Eszternek és Simon Gábornak a tanulmány elkészítésében nyújtott segítségükért, értékes tanácsaikért és észrevételeikért.

1. ’meghatározott módszerrel és előismeretekkel összegyűjtött nyelvi vagy irodalmi adatmennyiség, amely a tudományos kutatás vagy vizsgálat alapja’
2. ’számítógépre vitt és elemző programokkal előzetesen feldolgozott, további kutatásokra (pl. gyakorisági vizsgálatokra) alkalmas különböző szövegtípusokból gyűjtött szövegmennyiség’

Jelen tanulmány mindenekelőtt a második értelemben használja a korpuszt, amennyiben eltér ettől, azt egyértelművé teszi.

Mivel a tervezett Magyar lírakorpusz is nyelvtechnológiai értelemben használt *korpusz*ként készül, így ezt az értelmezést, illetve a hozzá kapcsolódó alapfogalmakat tisztázzuk.

Korpusznak nevezzük tehát az olyan – ténylegesen előforduló írott vagy lejegyzett beszélt nyelvi – szövegek elektronikus formájú gyűjteményét, amelyet előre meghatározott külső szempontok szerint válogattak össze a célból, hogy forrásként szolgáljon egy nyelv vagy nyelvváltozat nyelvészeti tanulmányozásához (Sinclair 2005). A korpuszok nem csupán magukat a szövegeket tartalmazzák, hanem feltüntetik azoknak a szerkezeti egységeit is (bekezdés, mondat), valamint egyéb releváns információkat is (pl. szófaji kód). Beszélhetünk ún. általános korpuszokról**,** amelyek egy adott nyelv állapotát reprezentálják, és nagyméretűek (ilyen például a Magyar nemzeti szövegtár[2] [Oravecz–Váradi–Sass 2014], amely jelenleg 1,04 milliárd szövegszót tartalmaz). Továbbá léteznek ún. speciális korpuszok, amelyeknél az adott vizsgálat tárgyának és céljának megfelelően válogatják össze a szövegeket, például egy műfaj vizsgálatakor. E besorolás alapján tehát a készülő lírakorpusz is speciális korpusznak tekinthető.

Az általános korpuszok esetén kiemelkedően fontos szempont a reprezentativitás: nem véletlenszerűen összeválogatott szövegek gyűjteményéről van szó, hanem tudatos tervezésről; a cél, hogy az adott nyelv többféle területi és társadalmi nyelvváltozata megfelelő arányban legyen jelen, ezzel – ideális esetben – az adott nyelv mintegy „miniatürizált változatáról” beszélhetünk (Szirmai 2005). Egy adott nyelv reprezentatív korpuszát elkészíteni azonban nem egyszerű, lényegében lehetetlen feladat két okból is: egyrészt egy dinamikus célpontot kíván reprezentálni (nyelvhasználat); másrészt a nyelvhasználatról a korpusz révén tudunk információkat szerezni (vö. Bieber 1993). Ez többek között problémát jelenthet akkor, ha az általános korpuszt referenciakorpuszként kívánjuk használni (mint ahogy jelen tanulmányban utalni is fogunk rá). Ilyen értelemben azt állítani a referenciakorpuszra alapozva, hogy a poétikai korpuszban fellelhető jegyek különböznek a sztenderd nyelvtől, túlzásnak hathat. Így a kapott eredményeket is ezeket a tényezőket figyelembe véve, némi kritikával szükséges kezelni.

[2] http://corpus.nytud.hu/mnsz/

A speciális korpuszok ezzel szemben nem az adott nyelv egészét kívánják reprezentálni, hanem egy meghatározott területhez kötődnek, adott szövegtípust, stílusréteget, nyelvváltozatot reprezentálnak, az adott terület tipikus szövegeit tartalmazzák. Sokszor a kutató maga készíti el a korpuszt egy probléma vizsgálatához. Ilyenkor a korpusz (ideális esetben) a specifikus nyelvhasználati színtérre nézve reprezentatív, nem pedig az egész nyelvre. Speciális korpusznak nevezzük például a hongkongi beszélt angol nyelv korpuszát (Hong Kong Corpus of Spoken English)[3] vagy a BEA magyar nyelvű spontánbeszéd-adatbázist (Gósy 2008) vagy a Történeti magánéleti korpuszt (Novák et. al 2017).

A korpuszok méretét a bennük található szövegszavak (tokenek) számával lehet megadni: ez a fogalom a korpuszban található összes szót jelenti, vagyis minden szóközzel határolt szó, illetve punktuációs jel egy szövegszó, függetlenül attól, hogy hányszor szerepel. A korpuszban előforduló különböző szavakat szóalaknak (type) nevezzük (Szirmai 2005).

A korpuszokhoz szorosan kötődő alapfogalom az annotáció is. Annotációnak nevezünk minden olyan információt és jelet, amelyet a korpusz eredetileg nem tartalmazott, de a készítés során belekerült a szövegekbe (Szirmai 2005). Egy szöveget többféle szinten annotálhatunk, így például a fonetikai, morfológiai, szintaktikai, szemantikai, pragmatikai stb. tulajdonságok szintjén.

3. Poétikai korpuszok

A poétikai korpuszok, mivel egy speciális nyelvi jelenségre és csak bizonyos műfajokra, illetve szövegtípusokra koncentrálnak, speciális korpuszoknak tekinthetők. A poétikai jellegű korpuszokat jellemzően két egymással szoros összefüggésben álló tényező hívhatja létre: egyrészt az összeválogatott szövegek lehetnek poétikusak, jellemzően, de nem kötelezően lírai szövegek (pl. dalok, versek, rapszövegek); másrészt a szövegekben annotált jelenségek kötődhetnek szorosabban a poétikai funkcióhoz. Így a szófaji, morfológiai elemzésen túl olyan nyelvi jelenségek annotációja is szerepet kaphat, amelyek tipikusan a poétikai funkció kidolgozásához járulnak hozzá (pl. verslábak, aposztrofé).

A magyar nyelv tekintetében eddig nem készült ilyen jellegű korpusz, ezért a Stíluskutató csoport munkája során megfogalmazódott 2016-ban egy olyan szövegtár összeállításának tervezete, amely egyrészt lírai diskurzusokat dolgoz fel, másrészt a poétikussághoz hozzájáruló nyelvi megoldások annotációját is tartalmazza (Domonkosi et al 2018). Ebben a fejezetben a csoport munkájára építve röviden ismertetjük a tervezett Magyar lírakorpusz[4] elméleti keretét és alapkoncepcióját.

[3] http://rcpce.engl.polyu.edu.hk/HKCSE/

[4] A Magyar lírakorpusz elkészítése egy tervezett pályázati munka célkitűzése, elnevezése – az

Továbbá bemutatunk számos a nemzetközi szakirodalomban megjelent lírai korpuszt és az azokhoz kapcsolódó kutatásokat.

3.1. A Magyar lírakorpusz elméleti kerete: kognitív funkcionális megközelítés

A korpusz összeállítását és annotálását is meghatározza az, hogy milyen elméleti háttérfeltevésekből indulunk ki. A magyar lírakorpusz alapvetően funkcionális kognitív megközelítésben készül. A kognitív poétika az irodalmat az emberi megismerésmód egy speciális esetének tartja. Központi kérdése az, hogy miként formálódik a műalkotás jelentése a nyelvi szerkezetek (konvencionális vagy attól fokozatokban eltávolodó) megformálásának hatására, a befogadás diszkurzív kontextusában, illetve hogyan kerül interakcióba a műalkotás nyelvi szerkezete és a befogadói elme. Ebben a felfogásban a hétköznapi és a szépirodalmi nyelvhasználat között nincs éles határ, a poétikusság kontinuum jellegű, és ennek megfelelően a különböző szövegtípusokban, műfajokban eltérő prototipikalitással jelenik meg (l. bővebben Simon 2015, 2016). A bemutatott kiindulási pont nagyban meghatározza a poétikai korpusz létrehozását, az alkorpuszok kialakítását is. Ennek megfelelően a Stíluskutató csoport terveiben négy, szövegtípus-, illetve műfajalapú alkorpusz létrehozása szerepel hozzávetőlegesen 2 000 000 szövegszóval, a későbbiekben a korpusz bővítését tervezve (Domonkosi et al 2018). Az alkorpuszok a következők:

(i) A kanonikus szépirodalom lírai szövegeinek alkorpusza
Az alkorpusz az iskolai oktatás kánonjában szereplő válogatott lírai szövegeket tartalmazza az 1900-as évektől napjainkig. Ezek elsődleges forrásaként az elmúlt évtizedek irodalmi szöveggyűjteményei szolgálnak.

(ii) A kortárs líra alkorpusza
Ezt az alkorpuszt a kortárs líra alkotásai alkotják, amely elsősorban a Digitális Irodalmi Akadémia[5] lírai szövegeire támaszkodik.

(iii) A dalszövegek alkorpusza
A populáris és alternatív dalszövegekből válogatott szövegek, amelyekhez forrásul a legnagyobb magyar dalszöveggyűjtő portál[6] szolgál.

(iv) A slam poetry alkorpusza
Végezetül a negyedik a slamszövegek alkorpusza, amelyhez forrásként a slammozgalom fő magyarországi weboldalát[7] használjuk.

operacionalizáció folyamatának részeként – jelenleg még munkacímként értelmezhető.

[5] https://pim.hu/hu/dia

[6] www.zeneszoveg.hu

[7] slampoetry.hu

A korpusz felépítésén túl a poétikusság kognitív funkcionális megközelítése az annotációban is megmutatkozik. A nyelvészeti feldolgozás során ugyanis a morfológiai és szófajtani elemzésen túl a Stíluskutató csoport egyéb, a líraisággal szoros összefüggésben álló nyelvi jelenségek annotálását is megcélozza. Így például kiemelt szerepet kap az aposztrofé, amely a lírai diskurzusokban és a funkcionális kognitív elméleti keretbe ágyazott vizsgálatokban központi jelentőségű (vö. Simon 2015; Tátrai 2012; Domonkosi et al 2018).

3.2. A poétikai korpuszok a nemzetközi gyakorlatban

A nemzetközi szakirodalomban számos poétikai korpusz, illetve ehhez kapcsolódó vizsgálat jelenik meg, amelyek különböző kutatási kérdésekre keresik a választ. Így például találkozhatunk nyelvészeti (morfológiai, szintaktikai stb.), nyelvhasználati, szociológiai, pszichológiai jellegű kérdésfelvetésekkel a poétikusság és a korpuszok kapcsán. Minden esetben az a cél, hogy a kérdésekre a nyelv terén megfogalmazható módon lehessen válaszokat keresni. Ebben az alfejezetben számos tanulmány alapján végigvesszük a megjelenő korpuszalapú vizsgálati és korpuszépítési módszereket, a tisztán automatikustól a manuális (időigényesebb) módszerek felé haladva. A fejezet első részében (3.2.1.) egy olyan módszert mutatunk be, amely strukturálatlan szöveggyűjtemény alapján hivatott poétikusságra utaló elemeket (verssorokat) azonosítani. Ezt követően (3.2.2.) szót ejtünk azokról a korpuszokról, amelyek bár poétikai korpuszként tekintenek magukra, nyelvtechnológiai értelemben nem tekinthetők korpusznak. Tekintettel arra, hogy a nemzetközi szakirodalomban számos kutatás kiemelten a dalszövegekhez kapcsolódik, illetve ez a készülő Magyar lírakorpuszban is fontos szerepet kaphat majd, így jelen tanulmányban is hangsúlyosan jelenik meg.

A szakirodalom áttekintése alapján a poétikai korpuszok vizsgálatában háromféle módszer elkülönítésére van lehetőség – természetesen a határvonal nem éles. Az első a csak statisztikai alapú, szógyakorisági adatok alapján történő elemzés (3.2.3). Ez esetben a következtetések levonásához referenciakorpusz szükséges. A második elemzésnek az alapja a szó- és kifejezésszámlálás előre meghatározott szólisták alapján (3.2.4). Ebben az esetben morfológiai elemzésre van szükség, mivel a listaillesztést szótöveken érdemes végrehajtani (nem toldalékolt szóalakokon), a tövesítéshez pedig elengedhetetlen a morfológiai elemzés. A harmadik módszerben (3.2.5) az általános annotáción kívül egyéb, kimondottan poétikai jegyekre irányuló annotáció is készül (pl. metaforák címkézése). Ahogy fentebb is említettük, a kategóriák sokszor nem válnak el élesen, a módszerek keveredhetnek – és általában keverednek is.

3.2.1. Strukturálatlan szövegek kezelése

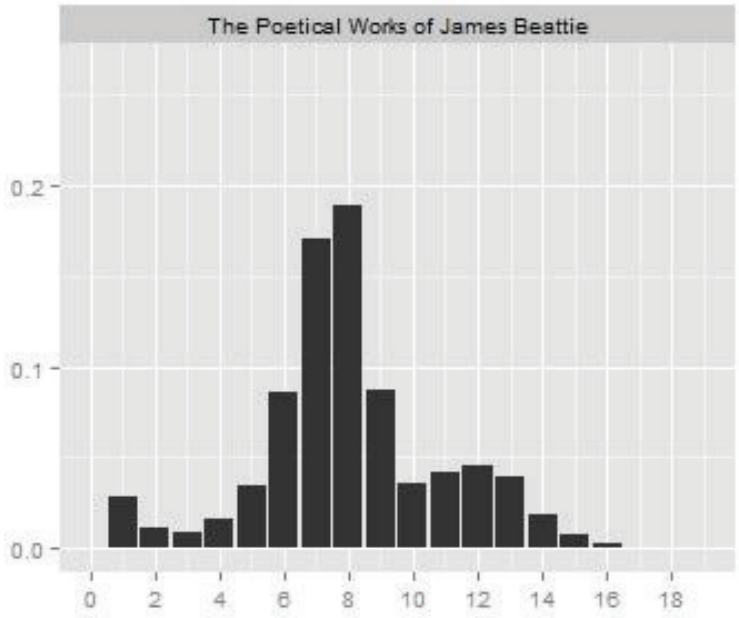
Amikor a kutatás témája a poétikusság, vizsgálhatók a poétikusság jegyei az összegyűjtött poétikus szövegekben. Illetve a vizsgálat a másik irányból is történhet, ti. hogy az előzetesen megfigyelt, egyedi jegyekkel próbáljuk azonosítani a poétikus tartalmat a strukturálatlan szövegekben.

A korpuszépítés első lépése a szövegek kiválogatása, összegyűjtése. Számos szépirodalmi szöveg elérhető a világhálón, de előfordulhat, hogy a szöveggyűjtemény anyaga csak olyan formában tölthető le, melyből hiányoznak a metaadatok, továbbá a különféle műfajok (líra, próza) keverednek, vagyis teljes mértékben strukturálatlan szöveghalmazról van szó. Ilyen esetekben gyors és hatékony megoldás lehet a poétikus szövegek kiválogatásának automatizálása.

Erre példa Duhaime (2014) próbálkozása, amely olyan számítógépes módszerek fejlesztését tűzte ki célul, amelyek képesek strukturálatlan szövegben azonosítani a poétikusságra utaló elemeket, illetve megállapítani, hogy az adott mű milyen műfajhoz tartozik. Első lépésként a korai angol könyvek korpuszából (Early English Books, EEBO) (1475–1700-ig terjedő művek) és a 18. századi művek korpuszából (Eighteenth Century Collections, ECCO) kiszűrte az összes <l> 'line of verse' címkével megjelölt sort. Az eredményül kapott 16 571 szövegfájl azonban sok feliratot (epigráfiát) tartalmazott, amelyek nem relevánsak a kutatás szempontjából, ezeket eltávolította. Az EEBO-ból kiszűrt szövegfájlok a 16–17. századi irodalmi trendeknek egészen reprezentatív mintáját adják, míg a 18. századi szövegek esetében kevésnek bizonyult az eredmény: az automatikusan kiválogatott szövegek csupán 1%-át teszik ki a fennmaradt nyomtatott angol nyelvű szövegeknek, így további 18. századi szövegekkel kellett kiegészíteni a gyűjteményt. Ehhez a Gutenberg Project[8] szövegeit használták fel: letöltötték az összes angol nyelvű művet. Eredményül egy teljes mértékben strukturálatlan szöveghalmazt kaptak, amely semmilyen metaadatot nem tartalmazott (például a szerző neve, a mű címe), így metaadatok nélkül nehézséget okozott a poétikus szövegek kiválogatása.

Emiatt valamilyen más megoldást kellett találni a poétikus, illetve nem poétikus szövegek egymástól való elkülönítésére. Duhaime (2014) hipotézise az volt, hogy a poétikus szövegek több sortörést tartalmaznak, mint a prózai művek, továbbá az egy sorban található szavak száma is jóval kevesebb, mint a prózai alkotások esetében. Ennek alapján szövegfájlonként az egy sorban található szavak számát vette mértékül. A kiválogatott szöveghalmazból véletlen mintát vett, és megvizsgálta, hogy néz ki az egyes művek szó/sor profilja.

[8] https://www.gutenberg.org/



1. ábra: J. Beattie poétikus műveinek szó/sor profilja (Duhaime 2014)

A grafikon x tengelyén a szövegfájl egy sorában található szavak száma, míg az y tengelyen az x szót tartalmazó sorok relatív gyakorisága látható, vagyis az olvasható le belőle, hogy hány szóból álló sorok relatív gyakorisága a legmagasabb az adott szövegfájlban.

Példaként nézzük meg részletesebben a 2. ábrán látható grafikont, amely James Beattie poétikus műveinek szógyakoriságát ábrázolja. A sorok ~5%-ában 12–13 szó van, míg csaknem 20%-ukban 7–8 szó, vagyis a 7–8 szavas sorok dominálnak Beattie poétikus műveiben. A többi szerző esetén is hasonló eredményt kapott. Duhaime több, ugyanebből az időszakból véletlenszerűen kiválogatott prózai szövegre is készített hasonló grafikonokat, majd összevetette az eredményeket. A kétféle grafikon között jelentős különbség volt: míg a poétikus szövegeknél a 7–8 szavas sorok voltak túlsúlyban, a prózai műveknél a 11–12 szavas sorok domináltak. A hipotézis tehát beigazolódott: mivel a prózai szövegek az egész oldalt kitöltik, míg a lírai szövegek sorokba vannak tördelve, az utóbbiak értelemszerűen kevesebb szót tartalmaznak.

Összefoglalva tehát: Duhaime (2014) a poétikus szövegeknek strukturálatlan szöveghalmazból való kinyeréséhez kiszámította az adott szöveg szó/sor profilját, melynek segítségével el lehetett dönteni, hogy a szöveg költői vagy prózai kategóriába tartozik-e. Ennek segítségével meghatározta, hogy a Gutenberg-korpuszban 3150 poétikus szöveg van, ebből néhány száz való a vizsgált korszakból.

Duhaime módszerét azért szükséges mindenképpen említenünk, mert jó kiindulás lehet egy későbbi poétikai korpusz elkészítéséhez: az interneten elérhető, nem feltétlenül strukturált szépirodalmi szövegek gyűjteményéből automatikus módszerekkel könnyen kiszűrhetők azok a műalkotások, amelyek feltételezhetően nem prózai, hanem lírai szövegek. Az ily módon kiválogatott szövegek alapul szolgálhatnak egy későbbi poétikai korpusz létrehozásához.

3.2.2. Online szöveggyűjtemények

Ahogy korábban már említettük, a korpuszt alkotó szövegeket valamilyen előre meghatározott szempont alapján válogatják, majd annotálják. Vannak azonban olyan elektronikusan elérhető szöveggyűjtemények, amelyeket ugyan valamilyen kritérium szerint válogattak össze, de a metaadatokon kívül (pl. cím, szerző, keletkezés dátuma) nem tartalmaznak további információkat, és annotációval sem látták el őket. Ahogy jelen tanulmány elején erről esett szó, a *korpusz* terminust ezekre a gyűjteményekre is szokás alkalmazni, fontos azonban tisztázni, hogy nyelvtechnológiai értelemben véve nem korpuszok. Ezeken az általunk esetenként **vizsgálati korpusz**nak is nevezett szöveggyűjteményeken egy-egy adott kutatási kérdés céljából valamilyen szempont alapján összegyűjtött szövegek összességét értjük.

Nézzünk két példát az ilyen értelemben vett online szöveggyűjteményre!

Az első az óangol költeményeket tartalmazó The Online Corpus of Old English Poetry, OCOEP[9], amely tehát nyelvtechnológiai értelemben véve nem korpusz. Az a kritérium azonban igaz rá, hogy egy adott szempont alapján összeválogatott szövegekből állították össze. Ez a gyűjtemény az összes fellelhető óangol költeményt, illetve töredéket tartalmazza, amelyek különféle kutatások alapjául szolgálhatnak.

A szövegek egy részéhez glosszárium is kapcsolódik: a szavakra kattintva megjelenik az adott szó mai angol megfelelője, emellett jegyzetekkel is ellátták. A projekt célja, hogy az érdeklődő hallgatók, kutatók számára széles körben elérhetővé tegyék az egyébként nehezen hozzáférhető óangol költeményeket. Bár ez az elektronikus szöveggyűjtemény adott szempont szerint összeválogatott szövegeket tartalmaz, és nevében szerepel a *korpusz* szó, nem feltétlenül tekinthetjük szűkebb értelemben vett korpusznak, hiszen – a glosszáriummal ellátott csekély számú költeménytől eltekintve – nem tartalmaz annotációt, és keresőfelület sincs hozzá.

Az előzőhöz hasonlóan inkább szöveggyűjteménynek, semmint korpusznak tekinthető a New Northvegr Center[10] projekt, amely – többek között – az Eddát, valamint óizlandi nyelvű sagákat tesz közzé. Az említetteken kívül angolszász költemények gyűjteménye is megtalálható az oldalon (The Complete Corpus of Anglo-

[9] http://www.oepoetry.ca/
[10] http://northvegr.org

Saxon Poetry), amely – az elnevezés ellenére – nem tekinthető poétikai korpusznak, de még csupán korpusznak sem.

A szöveggyűjtemény és a korpusz között azért nem teljesen éles a határvonal; vannak komplexebb szöveggyűjtemények is. Erre jó példa a középangol nyelvű szövegeket tartalmazó Middle English Corpus[11], amely az említetteknél jóval összetettebb keresőfelülettel rendelkezik. A gyűjteményt az oxfordi szövegarchívum anyagaiból állították össze. A felület többféle keresést tesz lehetővé: egyszerű keresés, összetett keresés az ÉS, VAGY, NEM operátorok segítségével, kereshetünk megadott méretű szövegkörnyezetben, szerző és cím alapján. Vizsgálható két vagy három szó, illetve frázis együttes előfordulásának gyakorisága. A következő alfejezetekben különféle korpuszvizsgálati módszereket mutatunk be.

3.2.3. A statisztikai alapú módszer

A korpuszon alapuló vizsgálatok lehetnek statisztikai alapúak, főként szógyakorisági adatokra épülők. Ez a módszer a vizsgálathoz összeállított (jelen esetekben) poétikai korpusz mellett egy referenciakorpuszt is igényel, mivel a különböző korpuszokon számolt szógyakorisági adatokat veti össze egymással, és abból von le következtetéseket a hasonlóságokra, különbözőségekre vonatkozóan. Az egyik itt bemutatott módszer a ROLC (Rock Lyrics Corpus)[12] korpuszon alapul.

Falk (2012) kutatásának célja az volt, hogy kvantitatív és kvalitatív elemzéssel azonosítsa a rockzene szóhasználati sajátosságait, stilisztikai jellegzetességeit, illetve összevesse több másik műfajjal és az általános szóhasználattal. Ennek érdekében egy 300 szövegből és 52 907 szövegszóból álló korpuszt hozott létre, a ROLC korpuszt. A korpusz anyagát az 1950-es évektől az 1990-es évekig írt angol nyelvű rockszámok szövegei alkotják (ehhez Falk különböző oldalak adatait felhasználva összegyűjtötte az egyes évtizedek legnépszerűbb zeneszámait), a dalszövegek elemzése az ingyenesen letölthető AntConc[13] korpuszelemző eszközzel történt. Annak eldöntéséhez, hogy a rockdalszövegek inkább a beszélt vagy az írott nyelvhez állnak közelebb, szükség volt a Corpus of Contemporary American English[14] (COCA) referenciakorpuszra is. Falk (2012) a gyakorisági listáknál stopszavak listáját is használta (olyan szavak listája, amelyeket a vizsgálat során nem kívánunk figyelembe venni), és szóalakok helyett lemmákat (szótöveket) vizsgált. A rockspecifikus szavak összevetéséhez a GBoP korpuszon (Giessen–Bonn Corpus of Popular Music [Kreyer–Mukherjee 2007], részletesebben l. 3.2.6 fejezet) végzett vizsgálatok eredményeit használta

[11] http://quod.lib.umich.edu/c/cme/

[12] Ahol nem tüntettünk fel URL-t, ott a korpusz a nyilvánosság számára nem elérhető.

[13] http://www.laurenceanthony.net/software/antconc

[14] http://corpus.byu.edu/coca/

fel, melyekből már a kötőszókat, determinánsokat, prepozíciókat is kihagyta. Ennek eredményeképpen megállapította, hogy van néhány műfajspecifikus szó (pl. *know, oh, love, just)*, de nehéz messzemenő, általános következtetéseket levonni. Néhány jelenség azonban megfigyelhető, például a rockban több az akcióorientált kifejezés (vagyis több igét tartalmaz), mint a popzenében, illetve vannak szavak, amelyek a popdalokban nemigen fordulnak elő, míg a rockban igen (pl. *get, will, can)*. A rockszövegekre jellemző még a felkiáltások, sóhajtások viszonylag magas száma *(oh, ah, yeah)*, illetve a nem sztenderd alakok *(gonna, wanna)* előfordulása. A nem sztenderd alakok a beszélt nyelv sajátosságai is. Ennek ellenére sem a beszélt, sem az írott nyelvhez képest nem szignifikáns a különbség. Falk (2012) a szinkrón elemzés mellett diakrón elemzést is végzett, melynek eredménye, hogy bár sok gyakori szó megjelenik több évtized rockszámaiban is, de az eredményekben bizonyos szavak gyakoriságában eltérések mutatkoznak az évtizedek között (például *baby, oh, love*).

Hasonló módszert alkalmaztak a Bollywood Lyrics Corpuson végzett elemzés esetében is. Behl és Choudhury (2011) a hindi popszámok szövegeiből állított össze korpuszt, a Bollywood Lyrics Corpust. A mumbai hindi moziipar évente körülbelül 800 filmet forgat, ezek többnyire zenés filmek, amelyekben számos dal is elhangzik. A *Bollywood* terminust ezeknek a filmeknek az összességére használják. A dalok általában hindiül íródnak, de ez a hindi nyelvváltozat némileg eltér a sztenderd hinditől, továbbá urdu és perzsa szavakat is tartalmaz. Behl és Choudhury (2011) a webről összegyűjtött szövegeiből végül – a duplumok kiszűrése után – egy 6529 dalt és körülbelül félmillió szövegszót tartalmazó korpuszt hozott létre. Minden szöveghez az alábbi metaadatokat tüntették fel: szerző, dalszövegíró, filmcím, évszám. A megjelenés éve a dalszövegek fejlődésének vizsgálatához nyújt alapot. A morfológiai elemzést követően a korpusz körülbelül 12 300 szóalakot tartalmazott.

A korpuszalapú vizsgálatok sztenderd statisztikai metódusokkal és komplexhálózat-elemzési technikákkal történtek. A statisztikai számításokhoz a kutatók a sztenderd hindi nyelvet reprezentálni hivatott referenciakorpuszt használtak, mellyel összevetve kimutathatók a különbségek a Bollywood és a sztenderd nyelvváltozat között. A kutatáshoz használt másik fő elméleti keret a komplexhálózat-elmélet volt, melyben a csomópontok nyelvészeti entitásokat jelölnek, az élek pedig a közöttük fennálló viszonyt. Ennek segítségével könnyebben felfedezhetők és értelmezhetők az általános nyelvészeti jelenségek. A komplexhálózat-elméletnek egy speciálisabb fajtája a kollokációs háló, amelyben a csomópontok az egyedi szavakat reprezentálják, az élek pedig mondatbeli együttes előfordulásokat.

A felhasznált általános statisztikai módszer a gyakorisági eloszlás volt. A Zipf-törvény vagy Zipf-eloszlás szerint egy szó előfordulási gyakorisága fordítva arányos a gyakorisági listában elfoglalt helyével, tehát a leggyakoribb szó körülbelül kétszer gyakoribb, mint a második leggyakoribb szó, háromszor gyakoribb, mint a harmadik, és így tovább. A tökéletes Zipf-eloszlás azonban csak abban az esetben lehetséges,

ha a bővülő korpusz szókincse potenciálisan végtelen (Kornai 2002). A kutatás egyik eredménye, hogy a Bollywood-korpusz szókincse (a szóalakok száma) rendkívül korlátozott, annak ellenére, hogy a dalok és így a szavak (a korpusz szövegszóit értjük ez alatt) száma évről érve növekszik. Csupán 17 000 szóalak van egy félmilliós korpuszban, és minden hozzáadott dal csak átlag 1,02 új szóval bővíti a listát.

A hálózatmodellel folytatott vizsgálat alapján a Bollywood-korpusz körülbelül 1000 funkciószót tartalmaz, valamint közel 17 000 kevéssé használt szót, ún. perifériaszót (vö. Saha Roy et al. 2011), melyek aránya egymáshoz így 17. Ez az arány az átlaghoz képest jóval kisebb (a sztenderd angolban ez az arány 85), magyarázata pedig szintén a limitált szókincs és a szóhasználat változatosságának hiánya, hiszen egy-egy szóalakot jóval többször figyelhetünk meg kontextusában. Ezenkívül megfigyelhető a kreatív szóhasználat – a szóegyüttállási hálózat szerint –, továbbá a klaszterezettség alapján kimutatható (vö. Watts–Strogatz 1988), hogy a várttal ellentétben – miszerint a dalszövegek szintaktikailag szabadabbak – a dalszövegek szintaktikailag se nem kötöttebbek, se nem szabadabbak, mint a sztenderd nyelvi korpusz szövege.

E kutatás tehát tisztán statisztikai módszerek segítségével von le következtetéseket az általános nyelvhasználat és a dalszövegek nyelvhasználata közti különbségekkel kapcsolatban.

Ezzel a módszerrel – azon túl, hogy a sztenderd nyelvvel hasonlítjuk össze az adott nyelvváltozatot – diakrón vizsgálatokat is végezhetünk. Ebben az esetben a referenciakorpusz nem az általános nyelvhasználatot reprezentáló korpusz, hanem korokat (éveket, évtizedeket) reprezentáló alkorpuszok gyűjteménye. Ehhez metaadatokra van szükség, amelyek segítségével összeállíthatók az alkorpuszok. Ezenkívül egyéb referenciakorpuszokkal is összevethető (más zenei stílus korpuszával, amelyből a műfajspecifikus különbségeket lehet kimutatni).

A referenciakorpuszt – főként amennyiben az a sztenderd nyelvhasználatot hivatott reprezentálni – és az ebből született eredményeket a korábban említett okok (l. 2. fejezet) miatt érdemes kritikával kezelni.

3.2.4. A szólistán alapuló módszer

Ebbe a kategóriába az olyan vizsgálati módszereket soroljuk, amelyeknél az általános statisztikai módszereken túl úgynevezett szólistákat is alkalmaznak, továbbá a korpusz szövegszavai tövesítésen kívül egyéb annotációval is meg vannak jelölve (például szófaji címkékkel). A szólistákat előre meghatározott jelenséghez készítik, amely szavakat aztán vizsgálják a korpuszban.

Szólistákat használtak kutatásukhoz DeWall és munkatársai (DeWall et al. 2011). A kutatás során amerikai dalszövegeket vizsgáltak. Céljuk az volt, hogy a kulturális változás jelenségét tanulmányozzák, megértsék a kultúrák és emberek közötti különbségeket a dalszövegeken keresztül. Véleményük szerint a popszámok szövegei

kulturális műtermékek, amelyek hatással vannak érzelmeinkre, gondolatainkra, viselkedésünkre és vice versa. DeWall és munkatársai a következő szociológiai jelenségeket és érzelmi állapotokat vizsgálták: az énközpontúságot, a társadalmi elszigetelődés jelenségét, a dühöt és az antiszociális viselkedést, valamint a depressziót.

A korpusz összesen 88 621 szövegszót tartalmaz és az 1980 és 2007 közötti dalokat tartalmazó Billboard Hot 100 segítségével állították össze, majd a Linguistic Inquiry Word Count (LIWC) (Pennebaker et al. 2007) program segítségével elemezték.

A LIWC program egy olyan eszköz, amely az előre meghatározott (szó) kategóriák alapján azt nézi, hogy a szöveg szavai milyen mértékben kötődnek az adott szókategóriához. A különböző szókategóriák (nyelvészeti tulajdonságok, mint például többes szám / egyes szám; pszichológiai, szociális és kognitív állapotok, mint például pozitív emóciókhoz kapcsolódó szavak, szociális interakciós kifejezések) pszichometrikus tulajdonságokkal rendelkeznek. Ezek a legtöbbször egyszerű szólisták, de egy szó több kategóriához is tartozhat, több címkét is kaphat (például az *I cried* szerkezetet az alábbi címkékkel látja el a program*:* 1st pers singular, Past focus, Sadness stb.). DeWallék (2011) diakrón szempontú vizsgálatot végeztek, az idő függvényében tanulmányozták az egyes jelenségeket. A kutatás fő kérdései és az ehhez számolt adatok a következők voltak:

- Énközpontúbbak, narcisztikusabbak lettek-e a dalszövegek az idő folyamán? Ennek nyelvi megjelenései az E/1. és a T/1. névmások. Raskin és Shaw vizsgálata kimutatta, hogy spontán beszédbeli monológokban a narcisztikus személyiségjegyek pozitívan korrelálnak az E/1. névmásokkal, míg a T/2. névmásokkal negatívan (Raskin–Shaw 1988). DeWall és munkatársai a szóban forgó névmásokat az idő függvényében számoltak a LIWC segítségével. Az eredmény szerint az évek múltával valóban megnövekedett az E/1. és a T/1. névmások használata.
- Megfigyelhető-e társadalmi elszigetelődés a dalszövegekben az idő folyamán? (A tanulmányból nem derült ki egyértelműen, hogy az elszigetelődés mint téma lesz gyakoribb – vagyis a szöveg az elszigetelődésről szól, vagy pedig mint az interakció hiányáról van szó, amikor a megnyilatkozó maga szüntet be kapcsolatokat, és erre utal nyelvi jelekkel.) Ehhez a kérdéshez kapcsolódóan az olyan, a „szociális interakció" szókategóriához tartozó szavakat számolták, mint a *beszélgetni* vagy a *megosztani.* A hipotézis itt is beigazolódni látszott: az ilyen típusú szavak száma csökkenőben van.
- Megfigyelhető-e a növekvő düh és antiszociális viselkedés a dalszövegekben az idő folyamán? Ennek a kérdésnek a megválaszolásához a dühhöz és az antiszociális viselkedéshez kapcsolódó kifejezéseket számolták (pl. *ölni, gyűlölet, kurva)*. A hipotézis itt is beigazolódott.

- Megfigyelhető-e a pozitív érzelmek csökkenése a dalszövegekben az idő folyamán? A pozitív érzelmi töltetű szavak *(öröm, boldogság)* összeszámlálása azt mutatja, hogy megfigyelhető a pozitív érzelmek csökkenése, a hipotézis tehát ez esetben is beigazolódott.

A fent említett LIWC programot alkalmazták egy másik kutatásban is. Petrie és munkatársai (Petrie et al. 2008) vizsgálata azért különösen érdekes, mert nem egy műfajra, hanem egy konkrét együttes dalszövegeire fókuszál, illetve azért is, mert az összeállított dalszövegkorpusz pszichológiai kutatáshoz készült. A vizsgálat Beatles-dalok szövegeinek elemzésével igyekezett feltárni a zenekaron belül történt változásokat, valamint a főbb dalszövegírók (John Lennon és Paul McCartney) személyiségjegyeit. A figyelem az emocionális hangra, a kognitív dinamikára, a szociális attitűdre és identitásra, valamint az időorientációra irányult. Ezek mentén két dimenzió különült el, amelyekkel jellemezni, kategorizálni lehetett a dalszövegírók személyiségjegyeit: közvetlen és különbséget tevő. A közvetlenséget jellemzi például a rövid szavak, a jelen idő, az egyes szám első személy használata. A különbségtevésre pedig a tagadás, a kizáró szavak *(but, except),* a valószínűségre utaló szavak *(perhaps, maybe),* a bennefoglalást jelentő szavak *(and, with)* használata jellemző.

A kutatók a dalszövegeket két weboldalról (egy, csak Beatles-dalok szövegeit tartalmazó honlapról és egy általános dalszövegoldalról) töltötték le. A letöltött szövegeket összehasonlították, majd csak a Beatles által írt és előadott dalok kerültek be az adathalmazba. Végül összesen 185 dalszöveg került bele a korpuszba, ebből 78 Lennon-, 67 McCartney-, 25 Harrison- és 15 Lennon–McCartney-szerzemény.

Az amerikai angol helyesírással átírt szövegekből a háromszor vagy többször előforduló frázisokat törölték, minden dalt egy szövegfájlba töltöttek, majd a LIWC program segítségével a szövegbeli szavakat a fentebb említett kategóriákba sorolták. A szerzők egy újabb kategóriával bővítették az eddigieket: a szexualitással kapcsolatos szavakkal. A vizsgálat eredményeként megfogalmazódott, hogy az előzetes feltevésekkel ellentétben (például az, hogy Lennon szövegei intellektuálisabbak, míg McCartney szövegei szentimentálisabbak) McCartney szövegei intellektuálisan összetettebbnek mutatkoztak (a komplexitás mércéjének a hosszú szavak használatát vették), míg Lennon dalai sok negatív érzést tartalmaznak, és énközpontúságot mutatnak. Az elemzés az együttes dinamikájával és struktúrájával kapcsolatosan is engedett következtetéseket levonni: Lennonnak nagy befolyása volt a zenekarra. A korai dalokra a pozitív érzelmek, a közvetlen hangvétel és a közeli emberi kapcsolatok, míg a későbbi dalokra a kevésbé közvetlen hang és kevesebb pozitív érzelem, viszont erősebb reflektív hatás és összetettség a jellemző (Petrie et al. 2008).

A szólistás módszer – mint például az említett LIWC program – hátránya, hogy nem képes komplex elemzésekre, nem tudja kezelni a nem elsődleges jelentést, az iróniát, a konnotációt. Ennek oka, hogy nem tartalmaz szintaktikai, szemantikai, pragmatikai

stb. annotációt. A módszer hibái továbbá abból fakadnak, hogy a sokszor manuálisan készített listák ritkán tudnak teljesek lenni, illetve ezen listák összeállításánál a szubjektivitás is közrejátszik.

3.2.5. Egyéb, speciális annotációval ellátott korpusz és vizsgálata

A legidőigényesebb, és ennek köszönhetően legkomplexebb módszer, amikor a korpuszt az általános annotációkon túl (morfológiai elemzés) egyéb, speciális annotációval is ellátjuk. Ahogy azt fentebb említettük, a szövegek többszintű annotációt (fonetikai, morfológiai, szintaktikai, szemantikai, pragmatikai stb.) tartalmazhatnak. Annotálhatjuk például a hangsúlyokat, a magánhangzókat, mássalhangzókat, a szintaktikai egységeket, a tulajdonneveket, a tematikus szerepeket, a terminusokat vagy akár a metaforikus kifejezéseket. A poétikai korpuszok a különféle poétikai eszközök szintjén is annotálva lehetnek (például metaforák, verslábak, aposztrofé). Most ezekre nézünk néhány példát.

Említésre méltó a VU Amsterdam Metaphor Corpus[15], amely a British National Corpus (BNC) „Baby" alkorpuszának a szövegeit használta fel. Fontosságát az adja, hogy a szövegek a metaforák szintjén is annotálva vannak. A kognitív metaforaelmélet szerint a metaforák nemcsak a költői nyelvhasználat sajátjai, mindennapi megnyilatkozásainkban is nagy számban használunk nyelvi metaforákat, illetve a világ megismerését is fogalmi metaforák segítik (vö. Lakoff–Johnson 1980; Kövecses 2002, 2005). Az említett korpusz tartalmaz ugyan szépirodalmi szövegekből álló alkorpuszt, de javarészt köznapi szövegekben annotálja a metaforikus kifejezéseket – tehát összességében nem tekinthető poétikai korpusznak, mivel a metaforák sem kizárólag a költői nyelvhasználat jellemzői. Ugyanakkor mégis iránymutatóként szolgálhat poétikus szövegek metaforák szerinti annotálásához. A korpusz az alábbi négy területet öleli fel: tudományos szövegek, hírek, szépirodalom, beszélt nyelvi szövegek; minden területről kb. 50 000 szövegszót tartalmaz. A metaforaszintű annotálás azt jelenti, hogy minden egyes szót az alábbi fő kategóriák egyikébe soroltak be: metaforához kapcsolódó szavak (MRW – metaphor related words), metaforát jelző szavak (MFlag), illetve olyan szavak, amelyek nem kapcsolódnak metaforához. A metaforához kapcsolódó szavak esetén különbséget tettek azok között az esetek között, amikor egyértelműen metaforáról van szó, illetve a határesetek között. A metaforákon belül továbbá megkülönböztették a direkt, indirekt és implicit metaforákat.

A több mint 300 millió szövegszóból álló Russian National Corpus[16], vagyis az orosz nemzeti korpusz rendelkezik poétikai alkorpusszal, amely ~2,5 millió tokenből áll, főként az 1750–1890 közti időszakból, kisebb részben 20. századi poétikus szövegeket

[15] http://ota.ox.ac.uk/desc/2541
[16] http://www.ruscorpora.ru/en/

is tartalmaz. A korpusz nem csupán lexikai és grammatikai (morfológiai) jegyek mentén kereshető, hanem speciális poétikai annotációval is ellátták, például jelölték az időmértéket, a különféle rímtípusokat. Lehetőség van olyan keresést végezni, amely az egyes verselési típusokra, verslábakra ad találatokat, például ha amfibrachiszokat (rövid-hosszú-rövid szótagból álló versláb) keresünk.

Az 1,8 millió tokenből álló baskír nyelvű poétikai korpusz[17] saját elmondása alapján a világon a második olyan gyűjtemény (az orosz nemzeti korpusz után), amely poétikai korpusznak nevezhető: körülbelül 500 000 verssort, 101 szerzőtől több mint 15 000 költeményt tartalmaz. A morfológiailag elemzett, részben szemantikailag is annotált szövegeket metrikai és prozódiai címkékkel látták el, így lehetőség van speciális metrikákra, rímtípusokra keresni. A baskír nyelvű szavak orosz megfelelői is fel vannak tüntetve, így a baskír nyelvet nem beszélő kutatók is tudják használni.

Összegezve: ebben az alfejezetben tehát olyan korpuszokat mutattunk be, amelyek a morfológiai elemzésen túl speciálisabb annotációval rendelkeznek, például jelölve vannak a metaforák.

3.2.5.1. A metaforák vizsgálata

A metaforák azonosítása és címkézése történhet (részben) automatikus módszerekkel. Babarczy és munkatársai (Babarczy et al. 2010; Babarczy–Simon 2012) egy olyan korpusz létrehozását tűzték ki célul, amelyben a metaforikus mondatok meg vannak jelölve azzal a címkével, hogy mely fogalmi metaforához tartoznak. A kutatók arra keresték a választ, hogy „a konceptuális metaforáknak szövegekben való automatikus megtalálása mennyire sikeres a testesültség hipotézisét alapul véve" (Babarczy–Simon 2012), vagyis abból a kognitív nyelvészeti alaptételből indultak ki, hogy az absztrakt fogalmak konkrét fogalmakra épülnek, és a konkrét fogalmak jelentése a világgal való testi tapasztalatok révén rögzül (Gibbs 2008; Kövecses 2002, 2005; Lakoff–Johnson 1980, 1999).

Bár ez a vizsgálat sem poétikus szövegekkel dolgozott, a metaforák és a metonímiák felismerésének automatizálása, a jelenségek annotálása miatt relevanciával bír a poétikai korpuszok annotálási lehetőségeit tekintve. Babarczy és munkatársai összesen 13-féle fogalmi metaforát használtak, köztük például: A VÁLTOZÁS MOZGÁS *(jön a hideg),* AZ ERŐFORRÁSOK ÉTELEK *(rengeteg áramot fogyaszt),* AZ IDŐ PÉNZ *(nem pazarolom az időmet)* stb. A metaforák azonosításához Martin (2006) módszerét alkalmazták: olyan mondatokat kerestek, amelyekben a forrás- és a céltartomány kifejezései egyaránt szerepeltek. Azt feltételezték ugyanis, hogy ha egy mondat egyaránt tartalmaz forrás- és céltartományi kifejezést, akkor az nagy valószínűséggel metaforikus mondat. A módszer alkalmazásához forrás- és céltartományi szavakat

[17] http://web-corpora.net/bashcorpus/search/index.php?interface_language=en

tartalmazó szólistákat állítottak össze asszociációs kísérletek, szinonimaszótár alapján, illetve referenciakorpusz segítségével.

A metaforákon kívül történtek kísérletek a metonimikusan viselkedő nevek automatikus kinyerésére is (Farkas et al. 2007). Az alapvető tulajdonnév-kategóriák – a személy-, hely- és intézménynevek – felismerése viszonylag hatékonyan működik, de a tulajdonnév-felismerő rendszerek jellemzően nem tesznek különbséget a metonimikusan és a literálisan viselkedő tulajdonnevek között. A GYDER egy maximum entrópián alapuló gépi tanuló rendszer, amely 80% körüli eredménnyel különíti el egymástól a metonimikus és nem metonimikus neveket angol nyelvű szövegekben. A metonimikusan viselkedő nevek felismeréséhez Farkas és munkatársai (Farkas et al. 2007) a következő jegyeket használták: szintaktikai információk (dependenciarelációk, determinánsok, többes szám), szemantikai általánosítási módszerek (Levin-igeosztályok, WordNet-hiperonimák) és a tokenek felszíni tulajdonságait kódoló felszíni jegyek.

3.2.5.2. Szentimentanalízis és érzelemvizsgálat

Ahogy azt az előző alfejezetben olvashattuk, a LIWC program különféle szókategóriákkal dolgozik. Az egyik kategória az Affect words, azon belül pedig a Positive emotion és a Negative emotion (ennek még van három alkategóriája: Anxiety, Anger, Sadness). A pozitív és negatív érzelmek, vélemények automatikus eldöntésére strukturálatlan szövegekben a szentimentanalízis ad lehetőséget. A szentimentanalízisnek nagy nemzetközi szakirodalma van, Magyarországon pedig – többek között – Szegeden foglalkoznak vele. A magyar nyelvű kézzel annotált szentimentkorpusz már elkészült (Szabó–Vincze 2015), és a releváns nyelvi elemeken kívül fragmentum- és aspektusszintű annotációval rendelkezik. A korpusz annotálási nehézségeiről és felhasználásáról Szabó és munkatársai írnak (Szabó–Vincze 2015; Szabó et al. 2016). Az annotáláskor első körben a teljes értékelő kifejezést, majd azon belül a pozitív és a negatív polaritású szentimentkifejezéseket és a shiftereket jelölték. A szentimentshifterek egyrészt azok az elemek, amelyek a „szentimentkifejezések szintaktikai kontextusában befolyásolják azok lexikális szintű, prior szentimentértékét" (pl. *a béka nem gusztustalan, a hangminőség aránylag jó*), másrészt amelyek a „prior szentimentértékeket nem változtatják meg ugyan, azonban lehetetlenné teszik az értékelést megfogalmazó szövegrész faktív olvasatát" (pl. *a hangminőség valószínűleg jó*) (Szabó–Vincze 2015: 221). Az aspektusszintű feldolgozásban az értékelést és a „feldolgozás alapegységét egy target, valamint az annak vonatkozásában kifejezett szentiment kapcsolatában határozza meg" (Szabó et al. 2016). Először a szentimentkifejezések használati sajátságairól szerettek volna pontosabb képet kapni, majd a csökkenő és növekvő intenzifikáló elemek megoszlását vizsgálták *(nagyon, kevésbé),* továbbá a polaritásváltást *(nem jó).* A kézzel készített korpuszból bizonyos

elemcsoportokra szólistákat generáltak (pozitív lexikon, negatív lexikon, entitás- és aspektusszótár, szentimentshifterek szótára), amelyek a későbbiekben felhasználhatók az automatikus szentimentelemzéshez.

Az érzelemkifejezésekhez és szavak érzelmi töltetéhez kapcsolódik Hámori (2018) vizsgálata is, amely arra tesz kísérletet, hogy adalékul szolgáljon az érzelmek poétikus szövegekben történő felismeréséhez és elemzéséhez. Az annotálási lehetőségeknél az érzelmet, az explicitséget és az érzelem kifejezésének eszközeit tekinti annotálandó elemnek, előrevetíti azonban az annotálás nehézségeit is.

3.2.6. További poétikai vizsgálati példák

A poétikus szövegeket sokféle szempontból lehet vizsgálni. Nézzünk még néhány további érdekes kutatási kérdést.

A dalszövegek elemzésére irányuló legkorábbi kutatások főleg kvalitatív jellegűek voltak, és elsősorban lexikai kérdésekre fókuszáltak. Az első olyan vizsgálat, amely a kvalitatív elemzés mellett már kvantitatív elemzést is elvégzett egy kisebb, 13 000 szövegszóból álló, az 1987-es évi 50 legnépszerűbb dalból összeállított dalszövegkorpuszon, Murphey (1990) vizsgálata volt. A kutatás fókuszában a névmáshasználat és a tér-idő referencia vizsgálata állt, célja pedig az volt, hogy a popdalok nyelvének, nyelvhasználatának elemzését integrálja a nyelvoktatásba.

Katznelson és munkatársai korpuszalapú vizsgálata (Katznelson et al. 2010) két szempontból is releváns: a szociolingvisztika oldaláról azt hangsúlyozták, hogy a különböző műfajok metaforahasználata és az egyes szavak gyakoriságának elemzése segíti a kulturális ideológiák és a sztereotípiák megértését és tudatosítását. A pedagógia oldaláról pedig azt, hogy a dalszövegek elemzése bevonható a nyelvhasználat tanításába. Hat weboldalról gyűjtötték össze a dalszövegeket, az amerikai populáris zene négy műfajára koncentrálva: pop, rock, country, hiphop. 110 rock, 102 pop, 112 country és 109 hiphop műfajú zeneszám szövege került bele a korpuszba, összesen tehát 433 dal és 178 982 szövegszó.

Kreyer és Mukherjee (2007) kortárs popzenék szövegeiből állították össze a GBoP (Giessen–Bonn Corpus of Popular Music) vizsgálati korpuszt, amelynek elemzésével popdalok szövegeinek stílusjelölőit igyekeztek azonosítani, valamint a szavak és a lexiko-grammatikai gyakorlatok kvantitatív és kvalitatív elemzésével helyesírási és lexikai kérdésekről igyekeztek számot adni. A korpusz a 2003-as US Album Charts top 30 albuma közül tartalmazott 27-et, összesen 442 dalt és hozzávetőlegesen 176000 szót.

Schneider és Miethaner (2006) a BLUR (Blues Lyrics Collected at the University of Regensburg) nevű, bluesdalok szövegeiből álló korpuszt vizsgálták azzal a céllal, hogy feltérképezzék az afroamerikai angol nyelv sajátosságait. A korpusz több mint 8000 szöveget tartalmaz a 20. század elejéről. A szerzőpáros amellett érvelt, hogy a

blueszenék szövegei nyelvi adatok értékes forrásainak tekinthetők. Számos korábbi vizsgálat van, amely intuícióalapú, nem pedig korpuszalapú. Schneider és Miethaner (2006) vizsgálata ezért azokra a szintaktikai konstrukciókra fókuszál, amelyek korábban nem lettek feltérképezve.

Összefoglalva elmondható tehát, hogy a speciális annotációval ellátott korpuszok létrehozása nagyon időigényes, viszont rendkívül komplex jelenségek vizsgálatára használható.

Továbbá láthatjuk, hogy a korpuszalapú vizsgálatok sokféle kérdésre adhatnak választ. Ehhez fontos azonban, hogy a kutatás célját a korpusz tervezése előtt pontosan definiáljuk. A következő fejezetben a Magyar lírakorpusz tervezetéről ejtünk szót a fent vázolt nemzetközi példák és az azokból fakadó tanulságok alapján.

4. Összegzés – tanulságok a készülő Magyar lírakorpuszhoz

Tanulmányunkban a megfogalmazott célkitűzéseknek megfelelően röviden felvázoltuk a készülő Magyar lírakorpusz elméleti hátterét, tervezetét, illetve az alapkoncepcióját. Ezt követően részletesen bemutattuk és csoportosítottuk a nemzetközi szakirodalomban megjelenő lírai korpuszokat, valamint ismertettük számos ehhez kapcsolódó kutatás módszereit, eredményeit és korlátait. Ahogy láthattuk, a poétikai korpuszok nagyon különbözőek lehetnek abban a tekintetben, hogy milyen anyagon alapulnak, illetve ezek milyen mértékben és milyen szempont szerint vannak nyelvészetileg feldolgozva.

A részletes áttekintés fontos lépése a Magyar lírakorpusz előkészítésének, és számos tanulsággal bír már a tervezési fázisra nézve is. A korpusz a lírai diskurzusok közé sorolható szövegekből poétikai szempontok alapján annotált korpusz létrehozását tűzi ki céljául. A tanulmányban bemutatott korpuszok rávilágítanak, hogy mind a szövegek kiválasztása, mind az annotálandó jelenségek körének behatárolása pontos előkészítést igényel a kutatási kérdések figyelembevételével. Az is megmutatkozott, hogy a feldolgozott szövegek metaadatainak kezelése szintén központi kérdés a későbbi elemzések szempontjából, így ezek feldolgozása alapvető feladat. Továbbá pontos mérlegelést igényel, hogy milyen elemzési szempontok érvényesüljenek az annotálás során a morfológiai és szintaktikai feldolgozáson túl; és hogy ezek milyen mértékben hajthatók végre automatikusan, illetőleg manuálisan. Az eredeti tervben a figurativitás, a rímszerkezetek és az aposztrofé annotálása szerepel. A figurativitás önmagában is komplex kérdéskör, így ezen belül valószínűleg olyan jelenség annotálását érdemes megcélozni, amelynek kialakult módszertana és bevett gyakorlata van, vagy viszonylag könnyen kezelhető automatikus módszerek segítségével. Ilyen lehetőség például a metaforaelemzés (vö. Steen et. al 2010), amelyben azonban korlátozott az automatikus annotálás lehetősége (vö. Babarczy et al. 2010; Babarczy–Simon 2012). A rímszerkezetek, valamint az aposztrofé jelenségéhez több lehetőség, illetve fogódzó is adódik a félig automatikus elemzéshez a szerkezeti és személyviszonyokhoz való

lehorgonyzás által. Fontos tanulság tehát, hogy az annotálás során figyelembe kell venni: hogy melyek azok a jelenségek, amelyek a poétikusság szempontjából valóban relevánsak; melyeknek van kialakult vagy kialakítható módszertana és gyakorlata; illetve melyek azok, amelyek automatikus módszerek segítségével is megközelíthetők.

Összességében elmondhatjuk, hogy jelen tanulmány a tervezett Magyar lírakorpusz építésének módszertanához kíván hozzájárulni, adalékul szolgálni, a lehetőségeket feltérképezni. Minthogy bármely kvantitatív elemzést alapos kvalitatív elemzésnek kell megelőznie, a tanulmányban felvázolt tanulságokkal, kritikai észrevételekkel feltétlenül számolni kell a kutatás kezdetén, ám ezeknek a gyakorlati megvalósítása csak jóval később, a kutatás folyamán lesz lehetséges.

Irodalom

Babarczy Anna – Bencze Ildikó – Fekete István – Simon Eszter 2010. A metaforikus nyelvhasználat egy korpuszalapú elemzése. In: Tanács Attila – Vincze Veronika (szerk.): *VII. Magyar Számítógépes Nyelvészeti Konferencia.* Szeged: Szegedi Tudományegyetem, Informatikai Tanszékcsoport. 145–156.

Babarczy Anna – Simon Eszter 2012. A fogalmi metaforák és a szövegstatisztika szerepe a metaforák felismerésében. In: Prószéky Gábor – Váradi Tamás (szerk.) *Általános Nyelvészeti Tanulmányok.* Budapest: Akadémiai Kiadó. 223–241.

Behl, Aseem – Choudhury, Monojit 2011. A corpus linguistic study of Bollywood song lyrics in the framework of Complex Network Theory. In: *Proceedings of ICON-2011: 9th International Conference on Natural Language Processing.* Macmillan Publishers. India.

Bieber, Douglas 1993. Representativeness in corpus design. *Literary and Linguistic Computing* 8: 244–257.

DeWall, C. Nathan – Pond, Richard S. – Campbell, W. Keith – Twenge, Jean M. 2011. Tuning in to psychological change: Linguistic markers of psychological traits and emotions over time in popular U. S. song lyrics. *Psychology of Aesthetics Creativity and the Arts* 5: 200–207.

Domonkosi Ágnes – Kuna Áges – Simon Gábor – Tátrai Szilárd – Tolcsvai Nagy Gábor 2018. Poétikai mintázatok kognitív stilisztikai kutatása. A Stíluskutató csoport kutatási terve. In: Domonkosi Ágnes – Simon Gábor (szerk.): *Nyelv, poétika, kogníció. Elmélet és módszer a poétikai kutatásban.* Eger: Líceum Kiadó. oldal???

Duhaime, Douglas 2014. *Identifying poetry in unstructured corpora.* Elektronikus dokumentum.
http://douglasduhaime.com/posts/identifying-poetry-in-unstructured-corpora.html

Falk, Johanna 2012. *We will rock you: A diachronic corpus-based analysis of linguistic features in rock lyrics.* Bachelor thesis. Linnaeus University, Faculty of Arts and Humanities, Department of Languages. http://www.diva-portal.org/smash/get/diva2:605003/FULLTEXT02.pdf

Farkas, Richárd – Simon, Eszter – Szarvas, György – Varga, Dániel 2007. GYDER: maxent metonymy resolution. In: *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007).* Prága: Association for Computational Linguistics. 161–164.

Gibbs, Raymond W. – Matlock, Teenie 2008. Metaphor, imagination and simulation. Psycholinguistic evidence. In: Gibbs, Raymond W. (ed.): *The Cambridge handbook of metaphor and thought.* Cambridge: Cambridge University Press. 161–176.

Gósy Mária. 2008. Magyar spontánbeszéd-adatbázis – BEA. *Beszédkutatás* 194–207.

Hámori Ágnes 2018. Az érzelmek elemzési lehetőségei a kognitív poétikai kutatásban és korpuszfeldolgozásban. In: Domonkosi Ágnes – Simon Gábor (szerk.): *Nyelv, poétika, kogníció. Elmélet és módszer a poétikai kutatásban.* Eger: Líceum Kiadó. oldal???

Katznelson, Noah – Gelman, Joseph – Lindblom, Katrin – Caput, Marie 2010. *American song lyrics: A corpus-based research project featuring twenty years in rock, pop, country and hip-hop.* San Francisco: San Francisco State University.

Kornai, András 2002. How many words are there? *Glottometrics* 4: 61–86.

Kreyer, Rolf – Mukherjee, Joybrato 2007. The style of pop song lyrics: A corpuslinguistic pilot study. *Anglia. Journal of English Philology* 125: 31–58.

Kövecses, Zoltán 2002. *Metaphor. A practical introduction.* Oxford: Oxford University Press.

Kövecses Zoltán 2005. *A metafora. Gyakorlati bevezetés a kognitív metaforaelméletbe.* Budapest: Typotex Kiadó.

Lakoff, George – Johnson, Mark 1980. *Metaphors we live by.* Chicago: The University Press Of Chicago Press..

Lakoff, George – Johnson, Mark 1999. *Philosophy in the flesh: The embodied mind and its challenge to western thought.* New York: Basic Books..

Martin, James H. 2006. A corpus-based analysis of context effects on metaphor comprehension. In: Stefanowitsch, Anatol – Gries, Stefan Thomas (eds.) *Corpus-based approaches to metaphor and metonymy.* Berlin, New York: Mouton de Gruyter. 214–236.

Murphey, Tim 1990. *Song and music in language learning: An analysis of pop song lyrics and the use of song and music in teaching English to speakers of other languages.* Frankfurt am Main: Peter Lang.

Novák, Attila – Gugán, Katalin – Varga, Mónika – Dömötör, Adrienn 2017. Creation of an annotated corpus of Old and Middle Hungarian court records and private correspondence. *Language Resources and Evaluation* 51: 1–28.

Oravecz, Csaba – Váradi, Tamás – Sass, Bálint. 2014. The Hungarian gigaword corpus. In: Calzolari, Nicoletta – Choukri, Khalid – Declerck, Thierry – Loftsson, Hrafn – Maegaard, Bente – Mariani, Joseph – Moreno, Asuncion – Odijk, Jan – Piperidis, Stelios (eds.): *Proceedings of Ninth International Conference on Language Resources and Evaluation (LREC 2014).* 1719–1723.

Pennebaker, James W. – Booth, Roger J. – Francis, Martha E. 2007. *Operator's Manual. Linguistic Inquiry and Word Count: LIWC 2007*. Austin: LIWC.

Petrie, Keith J. – Pennebaker, James W. – Sivertsen, Borge 2008. Things we said today: A linguistic analysis of the Beatles. *Psychology of Aesthetics, Creativity, and the Arts* 2: 197–202.

Saha Roy, Rishiraj – Ganguly, Niloy – Choudhury, Monojit – Singh, Navin Kumar 2011. Complex Network Analysis reveals kernel-periphery structure in web search queries In: *Proceedings of the 2nd International ACM SIGIR (Association for Computing Machinery Special Interest Group on Information Retrieval) Workshop on Query Representation and Understanding 2011 (QRU 2011).* Peking: UMass. 5–8.

Schneider, Edgar W. – Miethaner, Ulrich 2006. When I started to using BLUR: Accounting of unusual verb complementation patterns in an electronic corpus of Earlier African American English. *Journal of English Linguistics* 34: 233–256.

Simon Gábor 2015. Megszólalás és megszólítás: Az interszubjektivitás intázatai lírai diskurzusokban. In: Bódog Alexa – Csatár Péter – Németh T. Enikő – Vecsey Zoltán (szerk.): *Használat és hatás: újabb eredmények a magyarországi pragmatikai kutatásokban.* Budapest: Loisir Kiadó. 35–66.

Simon Gábor 2018. Kognitív és/vagy nyelvészeti poétika. In: Domonkosi Ágnes – Simon Gábor (szerk.): *Nyelv, poétika, kogníció. Elmélet és módszer a poétikai kutatásban.* Eger: Líceum Kiadó. oldal???

Sinclair, John. 2005. Corpus and text: Basic principles. In Wynne, Martin (ed.): *Developing linguistic corpora: A guide to good practice.* Oxford: Oxbow Books. 1–16. Elérhető: http://ota.ox.ac.uk/documents/creating/dlc/.

Steen, Gerard J. – Dorst, Aletta G. – Herrmann, J. Berenike – Kaal, Anna – Krennmayr, Tina – Pasma, Trijntje 2010. *A method for linguistic metaphor identification: From MIP to MIPVU*. Amsterdam, Philadelphia: John Benjamins.

Szirmai Monika 2005. *Bevezetés a korpusznyelvészetbe.* Budapest: Tinta Könyviadó.

Szabó Martina Katalin – Vincze Veronika 2015. Egy magyar nyelvű szentimentkorpusz létrehozásának tapasztalatai. In: Tanács Attila – Varga Viktor – Vincze Veronika (szerk.): *XI. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2015).* Szeged: Szegedi Tudományegyetem. 219–226.

Szabó Martina Katalin – Vincze Veronika – Hangya Viktor 2016. Aspektusszintű annotáció és szentimentet módosító elemek egy magyar nyelvű szentimentkorpuszban In: Tanács Attila – Varga Viktor – Vincze Veronika (szerk.): *XI. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2015).* Szeged: Szegedi Tudományegyetem. 174–182.

Tátrai Szilárd 2012. Az aposztrofé és a dalszövegek líraisága. In: Szikszainé Nagy Irma (szerk.): *A stilisztikai-retorikai alakzatok szövegés stílusstruktúrát meghatározó szerepe.* Debrecen: Debreceni Egyetemi Kiadó. 197–207.

Tolcsvai Nagy Gábor 2007. *Idegen szavak szótára.* Budapest: Osiris Kiadó.

Watts, Duncan J. – Strogatz, Steven 1998. „Collective dynamics of 'small-world' networks". *Nature* 393: 440–442.

Poetics and corpora. How can corpus linguistics help to investigate poetical texts?

The aim of this paper is to provide an overview of international poetical corpora, as well as their types, and also the aspects and methods of the annotation. This overview is the first step and reference point for the ongoing project of The Research Group of Stylistics.

In this paper we introduce the types of corpora according to the methods of the annotation (automatically or manual, semi-manual) and the level and the type of the annotation, then we outline the main points of the ongoing poetical corpus. The theoretical background is based on the functional cognitive linguistic approach. Afterwards we introduce several poetical corpora and related works.

Various problems and questions (linguistical, psychological etc.) can be solved by corpus-based approaches, however as it reveals by the end of the study, to define the aims and questions in detail before designing the corpus and the annotation is crucial.

Thus it is important to consider the followings during the annotation: what are those phenomena that are relevant in terms of poeticity; that have methodology and practices created or to be created; that can be approached automatically.