INNOVATÍV TECHNOLÓGIÁK ALKALMAZÁSA A SZÖVEGELEMZÉSBEN

Sándor István

magyar és angol nyelv szakos tanár, Kisdobronyi Középiskola;
PhD-hallgató, Debreceni Egyetem, Irodalomtudomány Doktori Iskola
e-mail: sandoristi88@gmail.com

Jelen dolgozat témája interdiszciplináris: a nyelv- és irodalomtudomány, illetve a számítástechnika közös pontjaira épül. A kutatómunkám során a magyar próza sajátosságaival foglalkozom, a szöveg vizsgálatánál viszont a számítógépes szövegfeldolgozás módszerét használom. A szövegbányászat hatalmas mennyiségű adatokban rejlő információk félautomatikus feltárása különféle algoritmusok alkalmazásával. A nyelvtechnológia maga is ezen a határsávon halad, de jelen feladatnál ez különösen is igaz. Számos kutatás és nyelvészeti munka alapvető alapanyaga a nagyméretű szövegkorpusz. Egyrészről e nagy mennyiségű szöveg automatikus, építése és feldolgozása során szükség van az informatika módszereire és eszközeire, másrészről a korpuszon nyelvészeti kutatásokat szeretnénk végezni. Ebben a kutatómunkámban segítségemre lesz a magyar NoojApp. Arra vállalkozom, hogy bemutassam, hogyan lehet automatizáltan, digitalizált forrásból nagy korpuszokat építeni, hogy új szemléletet hozzon a szövegelemzésbe.

Kulcsszavak: *innováció, szövegelemzés, adatbányászat, szövegbányászat, NoojAPP, kriptografikus, mesterséges intelligencia, gépi elemzés*

Abstract

Тема цієї публікації міждисциплінарна: вона базується на точках дотику мовознавства, літературознавства та інформатики. У ході дослідження висвітлено особливості угорської прози, а під час опрацювання тексту застосовано метод комп'ютерної обробки. Інтелектуальний аналіз тексту – це напівавтоматичне виявлення величезної за обсягом інформації за допомогою різних алгоритмів. Мовні технології теж крокують цією лінією, однак це особливо актуально для такого завдання. Матеріалом численних досліджень та мовознавчих робіт є великий за обсягом текстовий корпус. З одного боку, ці великої кількості тексти автоматизовані, під час їх побудови та обробки застосовані необхідні методи та засоби інформатики, з іншого боку – на корпусі заплановано робити мовознавчі дослідження. У цьому досліді мені стане в нагоді угорський NoojApp. Метою дослідження є представити, як можна автоматизовано з оцифрованих джерел побудувати великі корпуси, щоб внести нове бачення в аналіз тексту.

Ключові слова: *інновація, аналіз тексту, інтелектуальний аналіз даних, інтелектуальний аналіз тексту, NoojAPP, криптографічний, штучний інтелект, машинний аналіз*

Jelen dolgozat témája interdiszciplináris: a nyelvtudomány és a számítástechnika közös pontjaira épül. A kutatás témája többszörösen is időszerű. Két távolálló tudományágat is összekapcsol. A kutatómunkám során a magyar próza sajátosságaival foglalkozom. A szöveg vizsgálatánál viszont a informatika segítségét is felhasználom. Ez alatt azt értem, hogy az informatikában népszerű adatelemzés módszerével elemzem a szövegeket. Az adatbányászat hatalmas mennyiségű adatokban rejlő információk félautomatikus feltárása különféle algoritmusok alkalmazásával.

Mivel a téma komplexitása adott, és a tudományágak nem kapcsolódnak szorosan egymáshoz, felmerül több kérdés is. Az irodalomtudomány fejlődése majd újra értelmezése alapjaiban változott az évtizedek alatt. Az irodalomtudomány alakulását a XX. század közepétől számítva több évtizeden át befolyásolta a strukturalizmus, a szemiotika és a

generatív nyelvelmélet. Kosztolányi a nyelv fontosságára hívta fel a figyelmet azért, hogy emlékeztessen, minden tudomány alapja a nyelv, amelyen megalkotják a saját tudományukat.[1] A nyelvtudomány felhívta az irodalom figyelmét a nyelvközpontúságra, viszont az arról szóló tájékoztatás, hogy a nyelvtudományban fejlődést hozó elméleteket alkalmazzanak az irodalomtudományban, nem volt megfelelő. Tehát a nyelv és az irodalom vizsgálatánál nem lehet ugyanazokat a strukturális elemzési szempontokat figyelembe venni, mint a nyelvtudománynál. Szegedy-Maszák Mihály elkülöníttette az elbeszélő szöveg rétegeit, de óva intett mindenkit attól, hogy valamelyik réteget elsőbbrendűnek tartsa a másiknál, mivel a szöveg egészében hat, nem pedig széttagoltan elemeire bontva.[2] Ebből következik, hogy a narratológia alapvetően, mint strukturalista irányzat is csak korlátozottan használható. R. Barthes a szöveg „életéről" beszél több tanulmányában. Barthes úgy véli, a szöveg a szerzőtől is valamelyest független komplex, egész.[3] Az irodalmi szöveg többértelműséggé stilizáltsága a mondatok megszerkesztettsége megnehezíti az elemzést és sokszor nem az egyértelműségre törekszik. Tehát az irodalomtudományban nehezen alkalmazhatóak azok a módszerek, amelyek a nyelvtudományban sikert értek el, ezért különös odafigyeléssel kell megközelíteni az irodalmat, hogy megmaradjon mibenléte, és hogy valós tudományos eredményeket érjen el. Ezeket a szempontokat figyelembe véve kutatásom elsődlegesen informatikai irányvonalon történik, de a kutatás tárgyát maga az irodalom képezi.

De vajon akkor miért alkalmazzák az innovációs technológiát? – merülhet fel a kérdés. Fentebb kifejtettem, hogy az irodalom nem egyenes értelembe vett szövegalkotás. Gépintelligenciával még az egyszerű szövegek tartalmát is nehezen lehet értelmezni. Mi van akkor, ha irodalmi szöveget próbálunk meg elemezni? Sokan értelmetlennek vélik a strukturális alapokon való gépi szövegelemzést, és nem értenek egyet, a gépintelligenciával való feldolgozással. Az irodalom az művészet.

1. Az adatbányászat mibenléte

Az adatbányászat (data mining) feladata: a tudásfeltárás és az adatbányászat felhasználási területei. Az adatbányászat különböző tudományterületek „keresztezéséből" jött létre: a matematika, ezen belül a statisztika és a mesterséges intelligencia módszereit használja. Mint különálló terület, az 1980-as években jött létre. A 2000 évek elején az internet térhódítását követően robbanásszerűen fejlődni kezdett. Mivel az internet egy hatalmas adatbázis és tartalma növekszik, ezért hatalmas lehetőség az adatbányászat előtt. [4]Alkalmazási területei: kereskedelem, pénzügy, telekommunikáció, oktatás.

1.1 A szövegbányászat alapjai

Az adatbányászati algoritmusok, technikák és alkalmazások rohamos fejlődésének köszönhetően egyre nagyobb az igény egy magyar nyelvű, naprakész és lehetőség szerint az adatbányászathoz kapcsolódó témák minél szélesebb körét átfogó jegyzetre.[5] Jelen munkánkkal erre az igényre kívánunk választ adni. Mivel Holl András is így ír az MTA egyik cikkében: *„Az MTA Nyelvtudományi Intézetének egyik projektje a Magyar Nemzeti Szövegtár. A nagyméretű szövegkorpusz (1.5 milliárd szövegszó) különböző forrásokból, eltérő stílusrétegekből épült fel."* Ebben az esetben

[1] Dobos István (2002): *Az irodalomértés formái.* Csokonai, Debrecen
[2] Szegedy-Maszák Mihály (2011): *Az újraolvasás kényszere.* Kalligram, Pozsony
[3] Dobos István (2015): *Az olvasás esemény.* Kalligram, Budapest
[4] Kecskeméti, Gábor (2014): Electronic Textual Criticism. In Dávidházi, Péter (ed.): *New Publication Cultures in the Humanities.* Amsterdam: Amsterdam University Press. Website: http://www.oapen.org/search?identifier=515678 (Download: 2019.06.25.)
[5] Weboldal: http://www.cs.bme.hu/nagyadat/bodon.pdf (Letöltve: 2019.06.05.)

a szövegbányászat célja a szavak összegyűjtése – a korpusz és a ráépülő szolgáltatások nyersanyagul szolgálnak további kutatásokra, nagyszótár létrehozására.[6]

Ugyancsak az MTA Nyelvtudományi Intézetének projektje a MATRICA (Magyar Társadalomtudományi Citációs Adatbázis). A projekt keretében mintegy 190 hazai szakfolyóirat több éves teljes anyagának feldolgozásával nyelvtechnológiai eszközöket fejlesztettek irodalmi hivatkozások kinyerésére. Az eljárás lényege a hivatkozás szövegben való felismerése, majd a hivatkozott mű bibliográfiai adatainak azonosítása. A cél bibliometriai adatok szolgáltatása olyan tudományterületeken, ahol nemzetközi citációs adatbázisokra nem támaszkodhatunk. Mivel már az MTA Nyelvtudományi Intézet több projektje is elindult ilyen típusú vizsgálat felé, úgy érzem, hogy a fontossága megkérdőjelezhetetlen, ezért célul tűztem ki az irodalmi szövegek ilyen típusú feldolgozását.[7] A szövegbányászat az adtabányászat része. Nagy hardveres központokat hoztak létre adatbányászat szempontjából. Ahogyan föntebb is írtam a szövegbányászatot, a Magyar Tudományos Akadémia is fontosnak tartja. A Google hatalmas humán és informatikai erőforrásbázisa nagyon gyorsan halad a mesterséges intelligencia fejlesztésében.

2. A szövegbányászat módszerei

A szövegbányászat feladata az lenne, hogy új szemléletet hozzon a szövegelemzésbe. Célom, hogy a szövegekben feltárjam a rejtett „információt". Különböző dokumentumokból gépi intelligenciával való kigyűjtése és reprezentációja a magyar irodalomtudományban egyedülállónak mondható. Ezért jelen tanulmány az irodalomtudományhoz és a számítástechnikához kapcsolódik. A szövegek összehasonlítása és csoportosítása, algoritmusok segítségével történő feldolgozása egyik állomása a szövegelemzésnek, hiszen a szövegelemzésnél szükség van ellenőrzésre is. Mivel a szövegbányászatban strukturálatlan szöveges állományok képezik az alapját az adatnak, ezért elkerülhetetlenné teszi az emberi ellenőrzést.[8]

A szövegbányászat a következő folyamatokat tartalmazza:

- A szövegek előfeldolgozása;
- Osztályozás;
- Szövegklaszterezés;
- Kivonatolás.

Ezek mellett kutatómunkám során így egészítettem ki:

1. Szövegek összegyűjtése;
2. Szövegek digitalizálása;
3. Gépi értelmezés;
4. Szövegek előfeldolgozása;
5. Osztályozása;
6. Feltételek megadása;
7. Szövegklaszterezés;
8. Kivonatolás;
9. Értelmezés.

Elsősorban szépirodalmi szövegek értelmezése lenne a feladat. Nagy segítségemre volt már eddig is az MTA Sztaki Kopi program,

[6] Holl András (2015): Szövegbányászat, adatbányászat, ismeretfeltárás. Új lehetőségek a tudományos kommunikációban. *Magyar Tudomány*, 176. évf. 6. sz. 680–686. Weboldal: http://real.mtak.hu/24408/1/TDM.pdf (Letöltve: 2019.06.28.)

[7] Holl 2015

[8] Oravecz, Csaba – Váradi, Tamás – Sass, Bálint (2014): *The Hungarian Gigaword Corpus*. In: Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, Stelios Piperidis (eds.): *LREC 2014*. Ninth International Conference on Language Resources and Evaluation. Reykjavik. Website: http://www.lrec-conf.org/proceedings/lrec2014/pdf/681_Paper.pdf (Download: 2019.06.05.)

amellyel a szövegeket egyeztettem.[9] Kovács Vilmos *Holnap is élünk* című regényével kapcsolatban használtam a szövegegyeztetési módszert. Mivel a regény három kiadásban jelent meg a cenzúra nyomása miatt, ezért a regényből kimaradtak fontos részletek. Az első regény is többszöri módosításon esett át. A *Holnap is élünk* meghurcoltatások után 1965-ben mégis megjelenhetett könyvalakban. A szövegen a cenzúra hatására történt néhány változtatás, amire Kovács Vilmos öccséhez írt levelében így reflektált: „*A regényen különben csekély, a lényeget nem érintő szerkesztést végeztem. Remélem, ezzel sikerült becsapnom az illetékteleneket.*" Azóta két újabb kiadás látott napvilágot. A kéziratot gondozó és a regény második kiadását szerkesztőként felügyelő M. Takács Lajos így kategorizálja a Kovács által 1975-ben tett változtatásokat:(ez miért dőlt?) „*Sok bekezdésnyit kihúzott hőse, Somogyi Gábor művészetről és életről filozofáló „belső monológjaiból", de megkurtította a párbeszédeket is. Elhagyott sok fölöslegesnek ítélt jelzőt és szóismétlést.*" A regényből kimaradtak a kárpátaljai magyar iskolák helyzetéről, a moszkvai időszámítás kényszerű alkalmazásáról, a déli harangszó betiltásáról szóló részletek, amelyek az eredeti változatban gazdagították a regény valóságtartalmát. Szoftveres szövegfeldolgozással könnyen megtaláltam a különbségeket.

Tehát a szövegelemzést felgyorsítja a különböző algoritmusok futtatása. A Kopi program segítésével nagyon gyorsan lehet kiszűrni a különbségeket.[10]

Az alábbi szűrőket alkalmazom a szövegek vizsgálatához:

- A névmásokat és azok használatát;
- Szófajok típusait;
- Mondatrészek típusait;
- Utalószó és kötőszó kapcsolatát;
- Mondatok típusait.

2.1 Szövegelemző szoftverek

Szoftveres szempontból nagyon érdekes a kérdés. Egy új szoftver kifejlesztésében részt venni hatalmas feladat lenne. Több idegen nyelvű szoftver is jóval előrébb jár, mint a magyar nyelvű szoftverek. Ez természetesen érthető, mivel a magyar nyelv összetettsége megnehezíti ezt. Viszont sok kezdeményezés indult például Prószéky Gábor, a Magyar Tudományos Akadémia Nyelvtudományi Intézetének igazgatója, a Morphologic szoftvercég alapítója a számítógépes nyelvészet alapjait tette le.[11] Több nyilvános forráskód is továbbfejleszthető interneten szabad fejlesztői hozzáféréssel. Így vélekedik a számítógépes nyelvelemzéséről Prószéky egyik cikkében: „*Chomsky elméleti szempontból közelített a nyelvhez, megengedhette magának azt a luxust, hogy univerzumának határai egyetlen mondat két végén legyenek. A mondatfüzérek sokasága a chomskyánus elmélet alapján nem értelmezhető, ezzel a gyakorlatban nem megyünk sokra.*"

Chomsky elméleti megközelítésével szemben az analitikus grammatikáé a nyelvi egységeket nem előállító, vagyis generáló, hanem azokat elemző program létrehozása. Ez sokkal gyakorlatiasabb cél. Ezek a gondolatok bátorítanak és vezetnek munkám során.

Programozói szempontból több lehetőség van magyar nyelven a szövegelemzéshez. A **Neticel Semantic API**[12] és a **Magyar Nooj**[13] szoftverek képezik alapját a kutatómunkámnak. Ezek a

[9] Pataki Máté – Micsik András – Kovács László – Szabó Mihály (2014): KOPI-Fotó: Plágiumkeresés egy lefotózott oldal alapján. In Kunkli Roland, Papp Ildikó, Rutkovszky Edéné (szerk.): *Informatika a felsőoktatásban – 2014 konferencia.* Debrecen: Debreceni Egyetem, Informatikai Kar. Weboldal: http://eprints.sztaki.hu/8019/ (Letöltve: 2019.06.05.)

[10] Pataki–Micsik–Kovács–Szabó 2014

[11] Holl András (2013): Információáradat és hullámlovaglás. *Magyar Tudomány*, 174. évf. 4. sz. 473–478. Weboldal: http://www.matud.iif.hu/2013/04/13.htm (Letöltve: 2019.06.28.)

[12] Weboldal: https://api.neticle.hu/ (Letöltve: 2019.06.05.)

[13] Weboldal: http://www.nooj-association.org/ (Letöltve: 2019.06.05.)

programok kiválóan elemzik a szöveget, pedig még próbaverzióban futnak. A korpuszok alkalmazása, mint például a Magyar Webkorpusz több mint 50 millió szót tartalmaz, amelynek segítségével a mondatrészek és az általuk kifejezett szófajokkal a stílusvizsgálatot végeztem. A kutatás első részét a *hunpost* és a *hunmorph* szoftver alkalmazásával kezdtem, amely igencsak érdekes eredményeket hozott. Az alábbi következtetésekre jutottam: 1. Az író gyakran használ névmásokat, kerüli a neveket. 2. A szófajoknak különös szerepe van a szóképeknél. 3. A legújabb kutatások irányvételére a nyelvi-poétikai megalkotottság vizsgálata a jellemző. Az utóbbi nézőpontból a kárpátaljai magyar irodalom „újraolvasása és újraértelmezése" fontos, időszerű tudományos feladat.

2.2 Magyar NooJ

A NooJ rendszer (1. ábra) egy nagyon gyors, hatékony szövegelemző rendszer, amelynek célja, hogy támogatást nyújtson a magyar nyelv korszerű technológiával történő, empirikus kutatásához.[14]

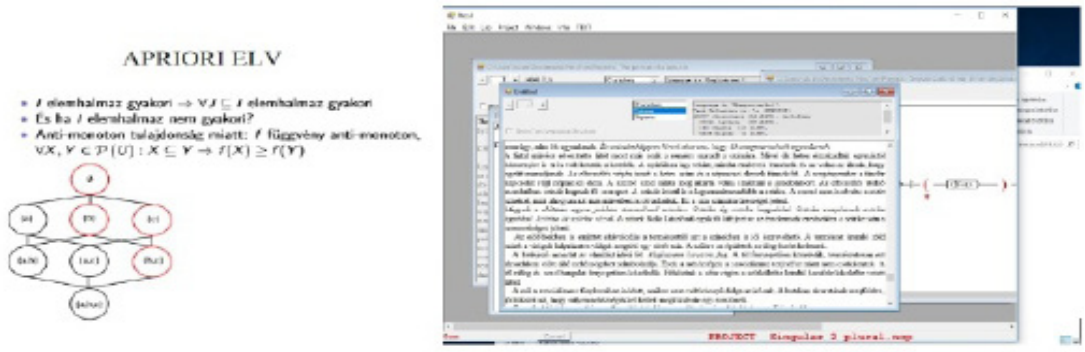
NooJ egy ingyenesen letölthető szövegelemző program. Letöltés után csak indítani kell, és hozzáadni a különböző szótárakat.

3. Irodalom és a számítógépes szövegelemzés

A XX. század egyik leghíresebb prózáját, az Édes Anna című regényt választottam a számítógépes szövegelemzésem bemutatójához. A program megnyitása után ki kell választani a nyelvet és betallózni a szöveget. A TEXT fülre kattintva elérhető a Linguistic Anlysis parancs, amely megállapítja a szöveg nyelvét és elvégzi a szövegfelismerést. Az Édes Anna például 340 ezer karaktert tartalmaz, de a program listázza nekünk az elemzés eredményeit (1. ábra).

Mindezek mellett hozzáadhatók szótárak, amelyek segítségével több elemzést is elvégezhetünk. Magyar szótárak: összes képzett és ragozott szóalak nagyméretű szólisták alapján Windowsra, 2015.05.18. frissítéssel telepíthető.

1. ábra. ***NooJ-program***



Forrás: http://www.nooj-association.org/ 2019.06.05.

A program továbbfejlesztésében vettem részt. A szótármodulok folyamatosan bővülnek, amelyet saját fejlesztésű forráskóddal kiegészítve alkalmaztam munkám során. A korpusz ténylegesen előforduló írott, vagy lejegyzett beszélt nyelvi adatok gyűjteménye. A

- 16000 elemű magyar szólista
- 72000 elemű magyar szólista.

Az MNSZ 2 és fél millió szóalakja morphdb.hu-val elemezve.

A NooJ-szövegelemző a következőket tartalmazza:

[14] Weboldal: http://corpus.nytud.hu/nooj/ (Letöltve: 2019.06.05.)

- Linguistic Analysis-Szintaktikai és szemantikai vizsgálata;
- Tokenizálás;
- Morfológiai információk elemzése, felismeri a szövegszavakat, a szövegszó szótövét, és azonosítja az aktuális toldalékokat.
- Konkordancia harmónia és illeszkedés a szófajok között;
- Lexical Analysis a szó lexikális elemzését különböző nyelvi szótárak, nevezetesen magyarázó, frazeológiai, antonyms, szinonimák és homonímák segítségével végzik.
- Statisztika-szavak száma, előfordulása, változatai;
- Linguistic Anlysis. Nyelvtani elemzés statisztikai értékeket jelenít meg (2. ábra).

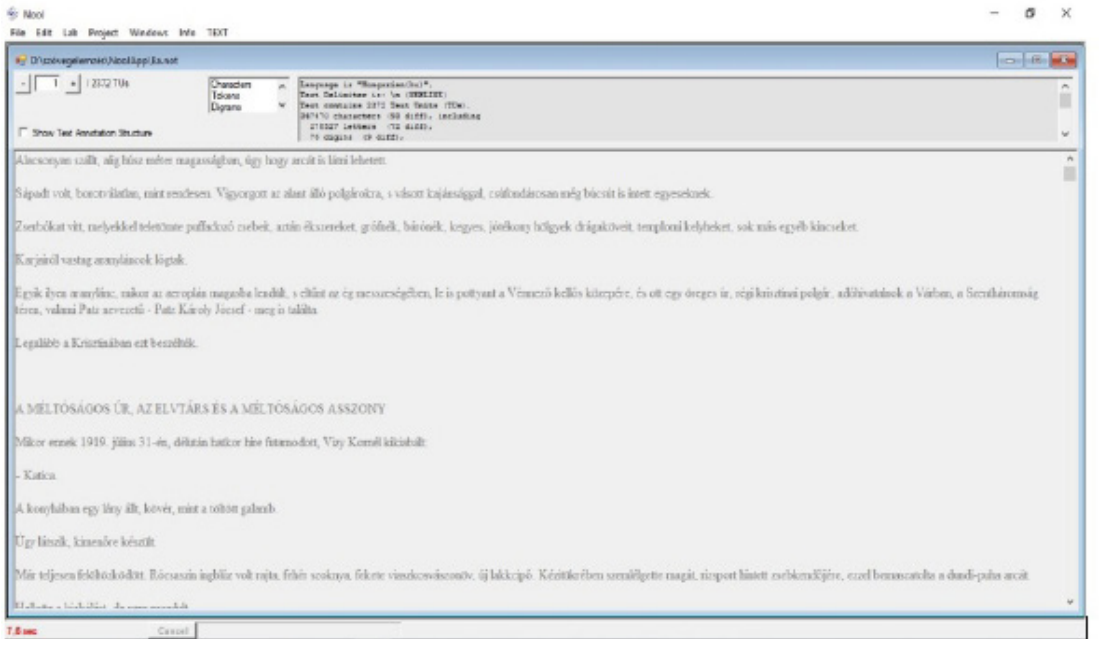
53350 blanks (2 diff), 15517 other delimiters (10 diff),64928 tokens including: 49335 word forms76 digits (9 diff)15517 delimiters (10 diff) Text contains 0 annotations (49335 different)Linguistic Resources applied to the text:

1. táblázat. ***Tokenizálás***

NooJ APP megnevezés	Magyarázat	Adatok
Language is "Hungarian(hu)"	Nyelve magyar	hu (szöveg)
Text contains	Bekezdések száma	235
characters	karakterek (93 különböző karakterek, ezek lehetnek betűk, számok, írásjelek)	347 470
letters	betűk	278 527
blanks	szavak száma	53 350
tokenek	szavak száma (előfordulásuk szerint)	49 335
digits	szókapcsolatok	15 517

Forrás: http://www.nooj-association.org/ 2019.06.05.

2. ábra. ***NooJ-program kezdőlapja***



Forrás: http://www.nooj-association.org/ 2019.06.05.

Language is "Hungarian(hu)".Text Delimiter is: \n (NEWLINE)Text contains 2372 Text Units (TUs).347470 characters (93 diff), including 278527 letters (72 diff), 76 digits (9 diff),

3.1 Lexikai analízis

A NooJ Lexical Analisys eredménye a lexikai annotálás, például a *szépbe* szövegszónak.

Lexical Analysis által előállított annotált változata: <LU lexikai egység eleje LEMMA="szép"

szótő=szép CAT="A"

szófaj=melléknév FLX="N12B9"

morfológiai kód= N12B9 DOMAIN="quality"

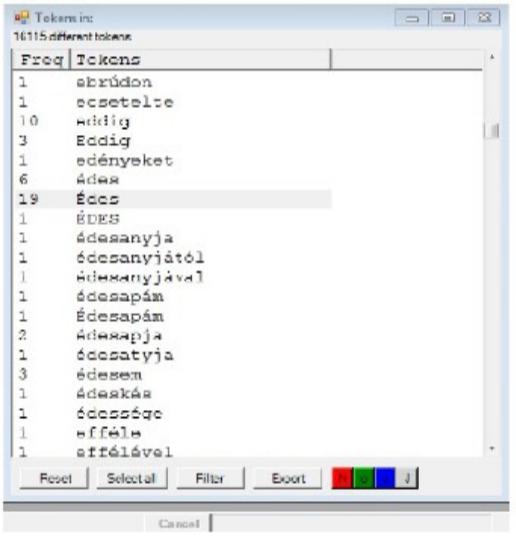
szemantikai osztály=minőség SUMO="SubjectiveAssessmentAttribute" javasolt egyesített ontológia=szubjektív értékelés

Case="ill" eset=illativusz Number="sg"> szám=egyes szám szépbe szövegszó </LU> lexikai egység vége (1. táblázat).

3.2 A konkordancia fogalma

A konkordancia a szavak illeszkedését vizsgálja. A szókapcsolatok gyakorisága látható az Édes mint főnév, tulajdonnevesítve (3. ábra). Édes Anna nevének előfordulása a szövegben.

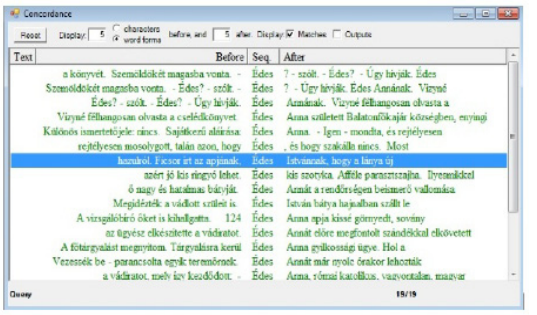
3. ábra. ***Édes, mint tulajdonnév keresési eredményei***



Forrás: http://www.nooj-association.org/ 2019.06.05.

A konkordancia lehetőséget ad a szavak illeszkedésének vizsgálatához, az következő ábrán látható (4. ábra).

4. ábra. Az Édes *szó illeszkedése a mondatokba*

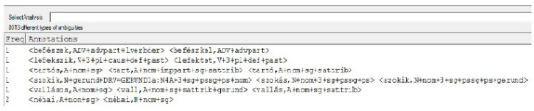


Forrás: http://www.nooj-association.org/ 2019.06.05.

3.3 Szöveg szótárlista

A program következő elemzési szempontja a szavak morfológiai vizsgálata. A szöveg szavainak a vizsgálatához a szótárakat kell használni. Ilyenkor morfológiai elemzéseket is végezhetünk a szavakon (5.ábra).

5. ábra. ***Szöveg és szóváltozatok***



Forrás: http://www.nooj-association.org/ 2019.06.05.

Néhány példát szeretnék bemutatni táblázatok formájában. *Elment* szó jelentésének és morfológiai vizsgálatának során az alkalmazás képes felismerni a szótövet, megkülönbözteti a szófajokat és a toldalékokat (2. táblázat; 3. táblázat).

2. táblázat. ***Morfológiai elemzés: „elment”***

Elment	
el (igekötő)+megy (ige)=men +t (múlt idő jele)	
el (IK)	*el* morféma, igekötő
megy(ige)=*men*	*megy* lexikai alak itt eltérő *men*
t(t1)	*t* (múlt idő jele) kijelentő mód

Forrás: saját szerkesztés

3. táblázat. ***Morfológiai elemzések***

Szó	Elemzés
lelketlenül	-etlen – denominális nomenképző, többalakú képző, fosztóképző lelketlen – relatív, lexikai tő ül – esetrag (essivusi–modálisi)
várnunk	vár – abszolút tő, lexikai tő, relatíve szabad tő, egyalakú igető -unk – többes szám első személyű, általános –n – a főnévi igenév képzője ragozású igei személyrag -unk a főnévi igenév személyragja várnunk zárt szóalak
tegyünk	*te* – abszolút tő, lexikai tő, kötött tő *sz*-es *v*-s tőtípusú igető *gy* – a felszólító mód jele *tegy* – relatív, szintaktikai tő *-ünk* – többes szám első személyű általános ragozású igei személyrag

Forrás: saját szerkesztés

4. Statisztika

A statisztika az irodalomtudományban nem a legjobban alkalmazható eszköz. Mivel a szavak és a szórend a szerzők által akár fel is borulhatnak, gondolva itt a versekre, de a prózában se lehet messzemenő következtetéseket levonni. Az *Édes Anna* vizsgálata során a statisztikai szempontot csak részinformációként lehet kezelni. Tudományos munkákhoz gazdasági elemzésekhez alkalmazzák leginkább. Mivel az *Édes Anna* szövege folyamatos, mondatszerkesztése egyszerű, ezért a szófajok típusait vizsgáltam a szövegben. Leggyakrabban természetesen igéket és főneveket tartalmaz a szöveg. Ilyen típusú elemzések akkor jöhetnek jól, amikor össze szeretnénk hasonlítani két írást, és számszerűsíteni akarjuk a mellékneves összetételeteket. Az biztos, hogy statisztikai szempontból a legkiválóbb eredmények születnének, de hogy mennyire gazdagítaná ez az irodalomtudományt, arra sokan keresik velem együtt a választ.

A jövő felé haladva az irodalomtudománynak is a figyelembe kell venni a változásokat. Ezért kutatómunkámban szerettem volna rávilágítani az irodalom és számítástechnika közös pontjára, amely irodalmi szempontból is új irányt jelent a strukturális szövegelemzésben.

Irodalomjegyzék

1. Dobos István (2002): *Az irodalomértés formái.* Csokonai, Debrecen
2. Dobos István (2015): *Az olvasás esemény.* Kalligram, Budapest
3. Holl András (2015): Szövegbányászat, adatbányászat, ismeretfeltárás. Új lehetőségek a tudományos kommunikációban. *Magyar Tudomány*, 176. évf. 6. sz. 680–686. Weboldal: http://real.mtak.hu/24408/1/TDM.pdf (Letöltve: 2019.06.28.)
4. Holl András (2013): Információáradat és hullámlovaglás. *Magyar Tudomány*, 174. évf. 4. sz. 473–478. Weboldal: http://www.matud.iif.hu/2013/04/13.htm (Letöltve: 2019.06.28.)
5. Kecskeméti, Gábor (2014): Electronic Textual Criticism. In Dávidházi, Péter (ed.): *New Publication Cultures in the Humanities.* Amsterdam: Amsterdam University Press. Website: http://www.oapen.org/search?identifier=515678 (Download: 2019.06.25.)
6. Szegedy-Maszák Mihály (2011): *Az újraolvasás kényszere.* Kalligram, Pozsony
7. Pataki Máté – Micsik András – Kovács László – Szabó Mihály (2014): KOPI-Fotó: Plágiumkeresés egy lefotózott oldal alapján. In Kunkli Roland, Papp Ildikó, Rutkovszky Edéné (szerk.): *Informatika a felsőoktatásban – 2014 konferencia.* Debrecen: Debreceni Egyetem, Informatikai Kar. Weboldal: http://eprints.sztaki.hu/8019/ (Letöltve: 2019.06.05.)
8. Oravecz, Csaba – Váradi, Tamás – Sass, Bálint (2014): *The Hungarian Gigaword Corpus.* In: Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, Stelios Piperidis (eds.): *LREC 2014.* Ninth International Conference on Language Resources and Evaluation. Reykjavik. Website: http://www.lrec-conf.org/proceedings/lrec2014/pdf/681_Paper.pdf (Download: 2019.06.05.)
9. Weboldal: http://www.cs.bme.hu/nagyadat/bodon.pdf (Letöltve: 2019.06.05.)
10. Weboldal: https://api.neticle.hu/ (Letöltve: 2019.06.05.)
11. Weboldal: http://www.nooj-association.org/ (Letöltve: 2019.06.05.)
12. Weboldal: http://corpus.nytud.hu/nooj/ (Letöltve: 2019.06.05.)