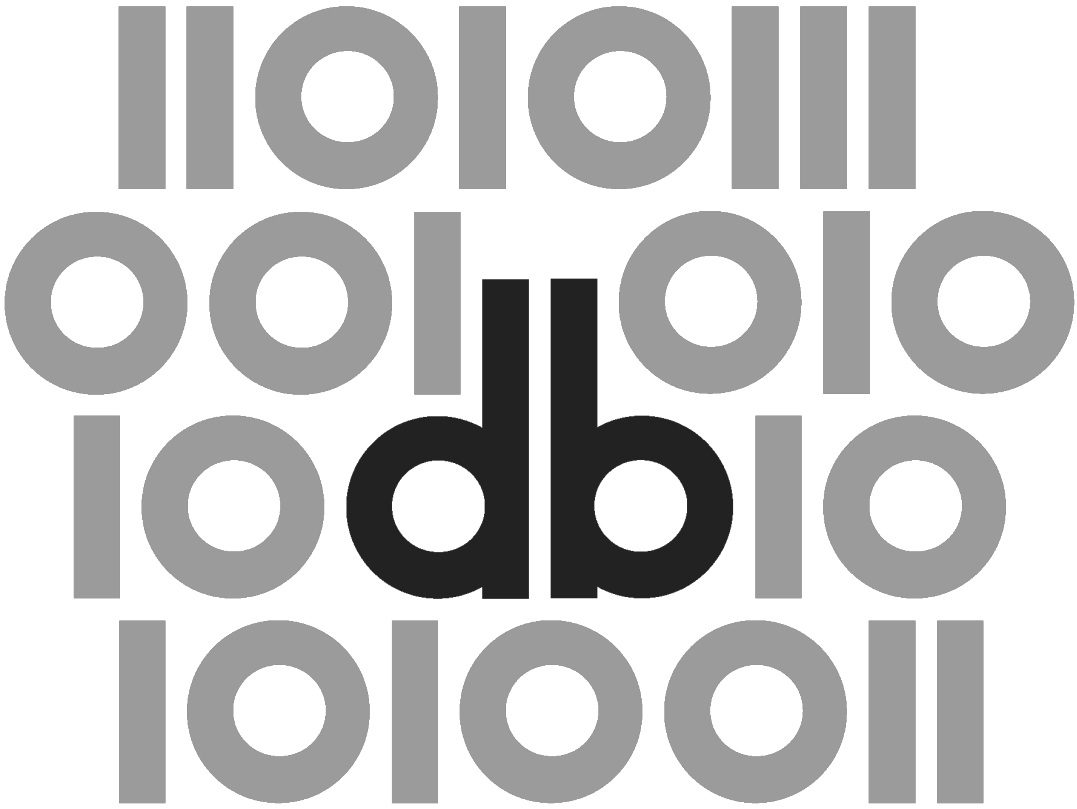
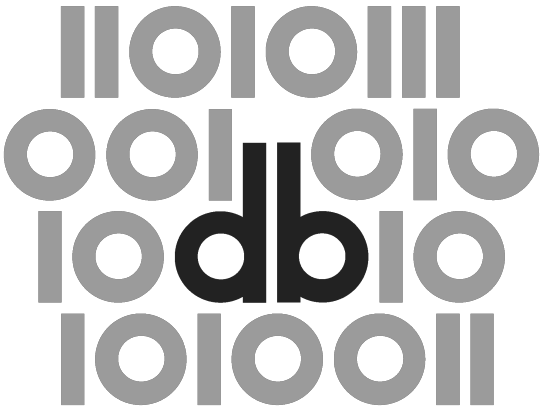


Digitális Bölcsészet
2022., hatodik szám

<DIGITÁLIS BÖLCSÉSZET>



6 (2022)

ISSN 2630-9696
DOI 10.31400/dh-hun.2022.6

Kiadja a Bakonyi Géza Alapítvány és az ELTE BTK Régi Magyar Irodalom Tanszéke (1088 Budapest, Múzeum krt. 4/A).
Felelős kiadó az ELTE BTK Régi Magyar Irodalom Tanszék vezetője.

Megjelenik az Open Journal Systems (OJS) v. 3. platformon, melynek működtetését az ELTE Egyetemi Könyvtár- és Levéltár biztosítja.

Honlap: `http://ojs.elte.hu/digitalisbolcseszet`
Email cím: dbfolyoirat@gmail.com

Olvasószerkesztő: Bucsics Katalin
Tördelés: Hegedüs Béla
Grafika: Hegyi Gábor

<TANULMÁNYOK>

Simon Gábor 0000-0001-5233-6313

Eötvös Loránd Tudományegyetem

simon.gabor@btk.elte.hu

A megszemélyesítés korpuszvezérelt vizsgálata a magyarban*

Egy *pilot* korpusz elemzésének tanulságai

Az utóbbi évek kutatásai alapján meglehetősen sokat tudunk a megszemélyesítő jelentés fogalmi és nyelvi szerveződéséről, ám lexikális és grammatikai mintázatait még nem ismerjük kellőképpen. Ez különösen igaz a magyar nyelvet tekintve. A tanulmány e kutatási hiány betöltésének első lépéseit teszi meg egy kognitív nyelvészeti, korpuszvezérelt elemzés bemutatásával. A kutatás egy félig automatizált annotálású kutatói korpusz (a PerSE korpusz) tesztváltozatának adataira épül, amely online autóteszteket tartalmaz, és amelyben a megszemélyesítő szerkezetek manuálisan annotáltak. A korpusz szövegeinek előfeldolgozását az *e-magyar* eszközlánc segítette, a kézi annotálás a MIPVU protokoll adaptált és kiegészített változatával történt. A tanulmány a korpusz felépítése mellett bemutatja az annotálás szintjeit és a folyamatát is. Ezt követően áttekintést nyújt az annotált korpusz fő adattípusaira: a megszemélyesítések lexikális mintázataira, a grammatikai jellemzőikre és a korpuszban megfigyelhető konstrukciószerű viselkedésükre.

Kulcsszavak:
megszemélyesítés, korpusz, annotálás, elemzés

1. Bevezetés

A megszemélyesítés „egy nem emberi fogalom vagy jelentés bemutatása oly módon, mintha az emberi lenne”,[1] például egy *autó* leírható erősként vagy olyan entitásként, amely érzékeny valamire. Az elmúlt évtizedekben a megszemélyesítés kutatása nem

* A tanulmány elkészítését a Magyar Tudományos Akadémia Bolyai János Kutatási Ösztöndíja, valamint az Innovációs és Technológiai Minisztérium ÚNKP-21-5 Új Nemzeti Tehetségprogramja támogatta a Nemzeti Kutatási, Fejlesztési és Innovációs Alapból. Ezúton köszönöm a tanulmány két lektorának hasznos javaslataikat mind a tanulmány szövegére, mind a kutatás folytatására vonatkozóan.

1 Joanna Thornborrow and Shan Wareing, *Patterns in Language: An Intorduction to Language and Literary Style* (London / New York: Routledge, 1998), 191, https://doi.org/10.4324/9780203979747.

bővelkedett a szisztematikus korpuszvizsgálatokban: miként Dorst 2011-ben megjegyezte, „alig történtek empirikus vizsgálatok arra vonatkozóan, milyen különböző megvalósulásai módjai vannak a megszemélyesítésnek a diskurzusban", és így nagyrészt „tisztázatlan maradt, miként lehet a megszemélyesítést megbízható módon azonosítani és elemezni".[2] Öt év elteltével sem változott érdemben a kutatás állása: a témáról szóló nemzetközi tanulmánykötet szerkesztői szerint a megszemélyesítés „kognitív formája és funkciója, retorikai és képi hatásai ritkán vonják magukra a kutatói figyelmet", ami azzal is jár, hogy a megszemélyesítés „kommunikatív eszközként magától értetődővé vált vagy pedig puszta konvencióként figyelmen kívül maradt".[3] Noha Dorst és munkatársai meglehetős figyelmet szenteltek mind a perszonifikáció fogalmi komplexitásának, mind pedig nyelvi változatosságának az angol nyelvben,[4] ezeknek az ígéretes kezdeti kutatásoknak a kiterjesztése nagyobb léptékű korpuszvizsgálatok irányába a mai napig várat magára. A jelen tanulmány célja az első lépések megtétele e kiterjesztés felé, a kognitív nyelvészet és a korpusznyelvészet elméleti és módszertani keretei között, egy épülő korpusz első verziójának (online autótesztek előfeldolgozott szövegeinek) elemzésével.

A nyelvi jellemzők vizsgálatát jól mutatja a szófaji kategória példája: empirikus tesztek során megfigyelhetővé vált ugyanis, hogy a szófaj szignifikáns szereppel bír a megszemélyesítő jelentés kialakulásában, arról azonban nincsenek adataink, hogy diskurzusainkban milyen arányban fejezzük ki a megszemélyesítő jelentést az egyes szófaji csoportokhoz tartozó kifejezésekkel. Egy másik példával élve, Dorst és munkatársai azt találták egy kísérletük során, hogy a laikus tesztalanyok által azonosított megszemélyesítések többsége (egészen pontosan 62%-a) többszavas kifejezés volt,[5] mégsem vizsgálták ezidáig sem a perszonifikáció konstrukciós viselkedését, sem idiomatikus természetüket.

Ha a magyar nyelv megszemélyesítő kifejezéseit tekintjük, a kibontakozó kép még kevésbé alapul empirikus vizsgálatokon. A perszonifikációt alakzatként értelmező áttekintő szakszócikk[6] az alábbi tényezők mentén jellemzi általában véve a jelenséget:

- a megszemélyesített entitás ontológiája (például absztrakt entitás, természeti jelenség, fizikai tárgy, állat vagy csoport);
- a megszemélyesítés módja (például cselekvés megvalósítása, érzelmek vagy más mentális állapotok tulajdonítása, emberi test megjelenítése);
- a megszemélyesítés grammatikai szerkezete (például igei predikátum, birtokos szerkezet, nominális és adverbiális elemek, vokatívusz);

[2] Aletta G. Dorst, „Personification in Discourse: Linguistic Forms, Conceptual Structures and Communicative Functions," *Language and Literature* 20, 2. sz. (2011): 113–114, https://doi.org/10.1177/0963947010395522.

[3] Walter S. Melion and Bart Ramakers, „Personification: An Introduction," in Walter S. Melion and Bart Ramakers, eds., *Personification: Embodying Meaning and Emotion* (Leiden / Boston: Brill), 1.

[4] Dorst, „Personification in Discourse;" Aletta G. Dorst, Gerben Mulder, and Gerard J. Steen, „Recognition of Personification in Fiction by Non-expert Readers," *Metaphor and the Social World* 1, 2. sz. (2011): 174–201, https://doi.org/10.1075/msw.1.2.04dor.

[5] Dorst, Mulder and Steen, „Recognition of Personification," 192.

[6] Sájter Laura, „Megszemélyesítés," in Szathmári István, főszerk., *Alakzatlexikon: A retorikai és stilisztikai alakzatok kézikönyve* (Budapest: Tinta Könyvkiadó, 2008), 383–388.

 DOI: 10.31400/dh-hun.2022.6.4409

– a megszemélyesítés regiszterspecifikussága (például beszélt nyelvi, tudományos, sajtónyelvi vagy irodalmi).

Két probléma is felmerül e megközelítés kapcsán. Egyfelől a fenti tényezőket semmilyen empirikus vizsgálat nem támasztja alá (és nem is részletezi), így szükségképpen a professzionális intuíció és nem szisztematikusan gyűjtött, szemléltető szövegpéldák támasztják alá a fontosságukat. A megszemélyesítések azonosítására kidolgozott és tesztelt módszer ugyanakkor nagyban növelné az elemzésbe vont nyelvi források körét. Másfelől a fenti tényezők kiterjednek a perszonifikáció fogalmi és nyelvi apparátusára is, azok empirikus mérése (operacionalizálása) azonban további kérdéseket vet fel a kutató számára. Míg az első és a második szemponthoz jól kidolgozott főnévi és igei ontológiákra van szükség, továbbá támaszkodhatunk nyelvspecifikus érzelemmegnevezésekre is (amennyiben a megszemélyesítő jelentés érzelmi állapotokat tulajdonít egy nem humán entitásnak), addig a grammatikai elemzés részben vagy teljesen automatizálható, ám a regiszterhez kötöttség vélhetően csak humán elemzők bevonásával állapítható meg. Másként fogalmazva, az iménti lista meglehetősen heterogén taxonómiához vezet el, amely ugyan alkalmas kiindulópont egy alapos nyelvészeti elemzéshez, de nem lehet rá egységesített és általános annotálási eljárást építeni.

A magyar megszemélyesítő szerkezetek korpuszvezérelt,[7] empirikusan tehát megalapozott elemzéséhez szükséges mindenekelőtt (i) egy korpusz, amelyben sok perszonifikáló adat figyelhető meg, továbbá (ii) olyan eljárás kialakítása, amely a korpusz szövegeiről kellő mennyiségű grammatikai információt nyújt, végül (iii) egy annotálási séma a megszemélyesítő szerkezetek azonosításához. Az itt bemutatni kívánt PerSE korpusz olyan nyelvi erőforrást jelent majd hosszabb távon, amely megbízható módon nyújt nagy mennyiségű adatot a magyar nyelv megszemélyesítő kifejezéseiről. (Innen ered a projekt elnevezése is, mely az angol *personifying structures encoded* kifejezésből előálló betűszó.) A jelen tanulmány a korpuszépítés és -elemzés kezdeti fázisát ismerteti, egy kis léptékű tesztkorpusz annotálásán keresztül. A PerSE korpusz tesztverziója online autóteszteket tartalmaz. A bevezetést (1) és a kutatás elméleti hátterének bemutatását (2) követően az alábbiakban tárgyalom a korpusz kialakításának menetét (3): a szöveganyagot, annak automatikus előfeldolgozását és a manuális annotálás protokollját. Ezt követően ismertetem a tesztkorpusz elemzésének előzetes eredményeit, részletezve a megszemélyesítések lexikális mintázatát, grammatikai jellemzőit és lehetséges konstrukcióit (4). A tanulmány rövid összefoglalással és kitekintéssel zárul (5).

2. Elméleti háttér

A megszemélyesítés kategóriája első ránézésre egyértelműnek tűnik, hiszen az a humán és nem humán entitások megkülönböztetésén, illetve megkülönböztethetőségén

[7] Elena Tognini-Bonelli, *Corpus Linguistics at Work* (Amsterdam, Philadelphia: John Benjamins, 2001), https://doi.org/10.1075/scl.6. Lásd még Simon Gábor, „Az igei jelentés metaforizációjának mintázatai: Nyelvtan- és korpuszvezérelt esettanulmányok," *Jelentés és Nyelvhasználat* 5 (2018): 1–36, https://doi.org/10.14232/jeny.2018.1.1.

alapuló jelentésalkotási mód. A kognitív nyelvészet perspektívájából tekintve azonban, amely a jelentésképzés fogalmi motiváltságának feltérképezésében érdekelt, a kép már jóval összetettebb, ugyanis a kibontakozó megszemélyesítő jelentésben több mentális művelet is szerepet játszhat. A hagyományos kognitív nyelvészeti megközelítés a perszonifikációt olyan fogalmi metaforaként elemzi, melynek forrástartománya az emberi test és elme, a céltartománya pedig egy nem humán entitás.[8] E megközelítés alapján a megszemélyesítés két fogalmi tartomány közötti megfeleléseken nyugvó reprezentációs struktúra, amely a főnévi megszemélyesítések (például A KÁBÍTÓSZER ELLENSÉG)[9] adekvát modellje.

Van azonban alternatív metaforikus modellje is a perszonifikációnak a kognitív nyelvészetben: Lakoff és Turner javaslata[10] szerint a megszemélyesítő jelentés hátterében egy generikus fogalmi mintázat, AZ ESEMÉNYEK CSELEKVÉSEK metafora áll, következésképpen a nyelvileg reprezentált esemény központi résztvevője (illetve absztrakt entitása) a metaforikus cselekvés cselekvőjeként konceptualizálható. E modell értelmében a leképezések nem két tartomány között bontakoznak ki, hanem e tartományok értékei (elemei) között, amely az igei megszemélyesítések jelentésére jellemző.[11]

Kétféle következtetés is levonható ebből a vázlatos áttekintésből. Egyrészt ezek a javaslatok nem annyira egymás alternatívái, mint inkább egymást kiegészítő modellek: míg az első perceptuális megszemélyesítések (például emberi testhez történő hasonlítás), valamint emocionális vagy mentális folyamatokra kiterjedő megszemélyesítések esetében tűnik hatékonynak, addig az utóbbi a nem humán (illetve tágabban a nem élő) entitások ágenciájának magyarázata. Másrészt e két modell ráirányítja a figyelmünket arra, hogy a megszemélyesítő jelentés nyelvi megvalósulása nem másodlagos jelentőségű (szemben a kognitív metaforaelmélet hagyományos, a fogalmi struktúrát előnyben részesítő magyarázataival), hiszen a grammatikai szerveződés orientálja a konceptualizálót a jelentés kialakításában. Következésképpen a megszemélyesítések sokféleségének feltárását célszerű a nyelvi szerkezet alapos vizsgálatával kezdeni, egy megbízhatóan annotált korpusz pedig alkalmas kiindulópont lehet a fogalmi aspektus vizsgálatához is.

Napjaink kognitív nyelvészete tehát a megszemélyesítés fogalmi hátterének összetettségét hangsúlyozza, amelyben a különböző fogalmi metaforák mellett a metonimikus jelentésalkotás is fontos szerepet játszik. Jóllehet Graham Low még metaforikus megszemélyesítések és metonímiák körültekintő megkülönböztetése mellett érvel (például *a tanulmány arra következtet* kifejezés metonimikus olvasatot kezdeményez anélkül, hogy emberi jellemzőket tulajdonítana a szóban forgó tanulmánynak, és így Low szerint legfeljebb „gyenge" megszemélyesítésnek tekinthető),[12] Dorst és munka-

[8] Zoltán Kövecses, *Metaphor: A Practical Introduction* (New York: Oxford University Press, 2010), 39, 56.

[9] Dorst, „Personification in Discourse," 119.

[10] Lásd George Lakoff, „The Contemporary Theory of Metaphor," in Dirk Geeraerts, ed., *Cognitive Linguistics: Basic Readings* (Berlin, New York: Mouton de Gruyter, 2006), 185–238.

[11] Dorst, „Personification in Discourse," 120, https://doi.org/10.1515/9783110199901.185.

[12] Graham Low, „»This Paper Thinks…«: Investigating the Acceptability of the Metaphor AN ESSAY IS A PERSON," in Lynne Cameron and Graham Low, eds., *Researching and Applying Metaphor* (Cambridge: Cambridge University Press, 1999), 221–248 https://doi.org/10.1017/CBO9781139524704.014.

 DOI: 10.31400/dh-hun.2022.6.4409

társai már inkább átfedést figyelnek meg metonímia és megszemélyesítés között, és ennek alapján a metonimikus megszemélyesítéseket specifikus alkategóriaként kezelik.[13] Kísérletük alapján a metonimikus megszemélyesítések az újszerű perszonifikációkhoz hasonlítanak a felismerési tesztek során, ennek lehetséges magyarázata, hogy ezek a jelentések egyaránt ágencia tulajdonításán alapulnak. Azzal a lényegi különbséggel, hogy míg a metaforikus megszemélyesítések tartományközi leképezéseken alapulnak, addig a metonímiákban tartományon belüli figyelmi váltás történik.[14] Ezért célszerű a metonimikus megszemélyesítéseket külön azonosítani, hogy ezáltal elemezhetővé váljon a jelentés fogalmi háttere is mindkét esetben.

További fogalmi modellt kínál a perszonifikációhoz Long, aki az úgynevezett fogalmi integráció műveletével írja le a megszemélyesítő jelentéseket.[15] Ebben a megközelítésben két mentális tér egyesül egy integrált *(blended)* térben, ám a teljes hálózat motiválja a figuratív jelentést, nem pedig annak egyes összetevői. Long tehát nem csupán a jelentésképzésbe bevont fogalmi struktúrák sokféleségét hangsúlyozza, de a megszemélyesítések többszavas jellegét is: a megszemélyesítés a diskurzusban „egy kiterjesztett jelentésegység […], elemei a csomópontként funkcionáló szó, annak kollokációi, kolligációi szemantikai preferenciája és szemantikai prozódiája."[16] A blendre építő modell tehát egyaránt hangsúlyozza a megszemélyesítés fogalmi és nyelvi összetettségét: „a jelentésbeli inkonzisztencia [amely a perszonifikáció sajátja ebben a modellben – S. G.] alapvetően a csomópont és a kollokáltja közötti inkongruenciában ölt testet."[17] Noha a kollokáció terminust kissé lazán alkalmazza a szerző, mindazonáltal ráirányítja a figyelmet a megszemélyesítő kifejezés nyelvi komponenseinek visszavisszatérő jellegére.

Összességében tehát egyetérthetünk Dorst állításával: „a megszemélyesítés azonosítása és elemzése eltérő problémákhoz vezet az elemzés különböző szintjein, és a kérdés, hogy mi számít megszemélyesítésnek, eltérő válaszokhoz vezethet az egyes szinteken."[18] E tanulmányban ugyanakkor szeretném meghaladni az elemzés szintjeinek puszta megkülönböztetését, hiszen egy korpusz, amelyben a grammatikai és szemantikai jellemzők párhuzamosan vannak annotálva a megszemélyesítő szerkezetek címkézésével, új nyelvi erőforrásként szolgálhat a kognitív szemantikai elemzés számára, ezáltal pedig empirikusan is megalapozza a további elméleti modellalkotást.

A vázolt kutatási eredmények alapján legalább két általános jellemzőt szükséges annotálni egy korpuszvizsgálat során: a szófaji kategóriát (az ugyanis szoros kapcsolatban áll a fogalmi szerveződéssel), valamint a morfoszintaktikai szerveződést (az ugyanis megfigyelhetővé teszi a visszatérő grammatikai, más terminussal *kolligációs* viszonyokat). Az elemzés további lexikális szemantikai dimenziója a konvencionalitás: a megszemélyesítő jelentés/használat lexikalizálódottságának a mértéke. Dorst

[13] Dorst, Mulder, and Steen, „Recognition of Personification."

[14] Lásd Klaus-Uwe Panther and Linda L. Thornburg, „Metonymy," in Dirk Geeraerts and Hubert Cuyckens, eds., *The Oxford Handbook of Cognitive Linguistics* (New York: Oxford University Press, 2007), 236–263.

[15] Deyin Long, „Meaning Construction of Personification in Discourse Based on Conceptual Integration Theory," *Studies in Literature and Language* 17, 1. sz. (2018): 21–28.

[16] Uo., 25. Long ezen a ponton Sinclair meghatározására épít.

[17] Uo.

[18] Dorst, „Personification in Discourse," 114.

és munkatársai szótárra alapozott elemzésükben négy kategóriát különítenek el.[19] „Újszerű" megszemélyesítések esetében az adott szó szócikke elsődleges jelentésként humánspecifikus jelentést ad meg, ugyanakkor nem tartalmazza a nem humán entitásokra vonatkozó alkalmazást, mint például az *őrködik az elektronika* kifejezés esetében, ahol az őrködik ige nem vonatkozik konvencionálisan nem emberi, esetleg nem élő ágensekre.[20] E kategória ellentéte a „konvencionális" megszemélyesítés, amelynél a szótári jelentésleírás aljelentésként magában foglalja a megszemélyesítő (nem humán entitásra kiterjesztett) használatát a szónak. Ilyen eset áll fenn az *erős autó* kifejezésnél, az *erős* melléknév ugyanis alábbi jelentéssel is bír: 'egy eszköz vagy gép, amely a maga területén nagy hatékonysággal működik', azaz a melléknév konvencionálisan használatos megszemélyesítésként (noha elsődleges jelentése az emberi fizikai, testi erőre vonatkozik). Bár az úgynevezett „alapbeállítású" *(default)* megszemélyesítésnél a szótár nem utal explicit módon humán cselekvőre/entitásra, a kifejezés értelmezése során azonban jellemzően emberi figurát azonosítunk. Például a *megbújik* ige jelentése a *két kipufogóvég bújik meg* kifejezésben a következőképpen adható meg a szótár alapján: 'rejtekhelyen meghúzódik, meglapul', és mivel ilyen tevékenységet állatok is végrehajthatnak, a kifejezés megszemélyesítő használata leginkább implicit, alapbeállítású. Másként fogalmazva: jellemzően, tipikusan emberi aktorként dolgozzuk ki a megbújás eseményének főszereplőjét, ám ez nem kizárólagos. A konvencionalizáltsági skála negyedik kategóriáját a metonimikus megszemélyesítések alkotják: ezeknél a perszonifikáló használat nem konvencionalizált, nem is alapbeállítású, hanem metonímiaként magyarázható. A *Mercedes megcsinálja a […] ferdehátúját* szerkezetben például a *Mercedes* az autógyártó vállalat mérnökeire utal metonimikusan. Fontos megjegyezni, hogy a megszemélyesítő jelentés konvencionalitása nem magának a grammatikai szerkezetnek az ismertségéből következik (noha a nyelvi szerkezetek konvencionalizálódása sok esetben a jelentés nyelvközösségbeli elterjedtségével is összekapcsolódhat), így e skála az elemzés új tényezőjeként vonható be a vizsgálatba. Másfelől ez a szempont lehetővé teszi a nyelvek közötti összehasonlítást, amely a kiterjedt, korpuszokra épülő kutatások esetében kifejezetten előnyös.

A szavak jelentésének vizsgálata mellett fontos továbbá a több szóból álló kifejezések belső szemantikai szerveződésének feltérképezése is, amelyhez különösen a kognitív nyelvtan[21] kínál alkalmas perspektívát. E nyelvleírás szerint az igék általánosságban egy vagy több résztvevőt feltételező, időbeli folyamatokat fejeznek ki. E résztvevők sematikus (azaz nem részletezett, nem kidolgozott) figurákként gondolhatók el az ige jelentésében: az elsődleges figura (trajektor) jellemzően a folyamat ágense, míg a másodlagos figurák (landmarkok) főként a folyamat elszenvedőit, eszközeit, egyéb résztvevőit vagy körülményeit képviselik az ige szemantikai szerkezetében. Mivel a jelentés konstruálása során ezeket a figurákat általában nominális kifejezések

[19] Dorst, Mulder, and Steen, „Recognition of Personification," 178.

[20] A konvencionalitás meghatározásához és címkézéséhez e tanulmányban és a teljes annotálás során a következő szótárt alkalmaztam: *Magyar értelmező kéziszótár,* főszerk. Pusztai Ferenc (Budapest: Akadémiai Kiadó, 2003).

[21] Ronald W. Langacker, *Essentials of Cognitive Grammar* (New York: Oxford University Press, 2013). A magyar nyelvre vonatkozóan lásd továbbá *Nyelvtan,* szerk. Tolcsvai Nagy Gábor (Budapest: Osiris Kiadó, 2017).

 DOI: 10.31400/dh-hun.2022.6.4409

jelenítik meg az elemi mondatban, a trajektor/landmark megoszlás és annak kifejezési módjai nem csupán az igét, de az ige köré szerveződő konstrukciót is jellemzik. Következésképpen a konstrukción belüli szemantikai viszonyok elemzése és címkézése új aspektusát jelenti a perszonifikáció kognitív nyelvészeti elemzésének, mert lehetővé teszi annak a megfigyelését, milyen szerepet tölt be a megszemélyesített entitás egy tágabb fogalmi jelenetben. Ha ez a szerep a trajektoré, akkor az entitás a metaforikus forrástartomány nagyfokú ágenciával bíró centrális figurája, míg ha landmark szerepű, akkor az adott entitás hozzájárul a megszemélyesítés kibontakozásához, de nem ágensként.

Másként fogalmazva, a kognitív nyelvtani elemzéssel nagyobb pontossággal lesz megragadható a megszemélyesítés konstrukciós viselkedése a magyarban. A forma–jelentés párokként[22] értelmezett konstrukciókból a megszemélyesítésre irányuló korábbi kutatás elsősorban a formai oldalt helyezte előtérbe: Long[23] például olyan összetett grammatikai mintázatokkal jellemzi az angol perszonifikációk nyelvi szerveződését, mint „nem humán alany + (csak humán létezőkre használt) igei állítmány + egyéb mondatrészek", vagy „egyéb mondatrészek + (csak humán létezőkre használt) igei állítmány + nem humán tárgy + egyéb mondatrészek", ám az efféle sémák alulspecifikáltak (például mit jelent az „egyéb mondatrészek" kategória a megszemélyesítő jelentés szempontjából?), másfelől túlságosan specifikusak (mennyire fontos például a komponensek sorrendje?). A magyar nyelvben számos eltérő mintázat elképzelhető (részben a gazdag morfológiai rendszer révén), ezért ezek a sablonok nem alkalmasak összehasonlító vizsgálatra. A megszemélyesítés szemantikai pólusát tekintve Dorst és munkatársai[24] a következő alapséma mellett érvelnek: az igei, melléknévi vagy adverbiális komponens kezdeményezi a megszemélyesítő jelentés fogalmi keretét, az ezekhez kapcsolódó főnév pedig a megszemélyesített entitást jeleníti meg. Ez utóbbi leírás ugyan kellően általános ahhoz, hogy a grammatikai és a szemantikai szerveződésre egyaránt kiterjeszthető legyen, a megszemélyesített entitás és a jelentésképzés során aktivált fogalmi keret pontos kapcsolatát azonban nem mutatja. Ezért a séma jellemzését ki kell terjeszteni a komponensek közötti jelentésbeli kapcsolatokra is, a kognitív nyelvtan korábban bemutatott kategóriái pedig éppen ezt a kiterjesztést alapozzák meg.

A megszemélyesítő szerkezetek nyelvspecifikus jellemzőiről a korábbi kutatások nem fedtek fel részleteket a magyarban. A megszemélyesítés általános igényű tárgyalása[25] ugyan hasznos áttekintést nyújt, ám nem a kognitív nyelvészet kiindulópontját érvényesíti, ezért csak részben egyeztethető össze a jelen kutatással. A magyar megszemélyesítések kognitív nyelvészeti elemzésének is vannak természetesen előzményei: egy korábbi kutatás[26] részletesen vizsgálta a perszonifikáció eseményszer-

[22] Adele Goldberg, *Constructions at Work: The Nature of Generalization in Language* (Oxford, New York: Oxford University Press, 2006).

[23] Lásd Long, „Meaning Construction," 23.

[24] Dorst, Mulder, and Steen, „Recognition of Personification," 192–193.

[25] Sájter Laura, „Megszemélyesítés".

[26] Simon Gábor, „A megszemélyesítés szemantikai sémái József Attila leíró költeményeiben," *Magyar Nyelvőr* 142, 3. sz. (2018): 328–354. Lásd még Simon Gábor, „The Event Structure of Personification in the Poetry of Attila József," in József Tóth and László Szabó V., eds., *Ereignis in Sprache, Literatur und Kultur* (Berlin: Peter Lang, 2021), 67–79.

kezetét József Attila költészetében, feltérképezve a grammatikai jellemzőket, a trajektor/landmark megoszlást, valamint a megszemélyesítés fő fogalmi kategóriáit egy kis terjedelmű poétikai korpuszban. Egy másik jelenleg is zajló kutatás az érzékszervi tapasztalatok nyelvi reprezentálását, és ebben a megszemélyesítés szerepét vizsgálja, szisztematikus korpuszelemzéssel, a kiválogatott kulcsszókra nézve reprezentatív mintákon, nyelvközi összehasonlítással.[27]

Egy nagyobb léptékű empirikus vizsgálatnak tehát megvannak már az alapjai, ám a korábbi kutatás a mintavételezések szűk köre, valamint a specifikus kutatói korpuszok miatt nem teszi lehetővé a nagyobb mértékű generalizálást. Továbbá az idézett vizsgálatok a szótáralapú azonosító eljárás sikeres adaptálása mellett is inkább a kvalitatív feltérképezést valósították meg, noha kvantitatív elemzési lépéseket is magukban foglaltak. Ily módon a korpuszépítés és az annotálás elméleti és gyakorlati kérdései napjainkig részben megválaszolatlanok maradtak. Az itt bemutatott PerSE korpusz az általános jellegű, nyelvspecifikus megszemélyesítéskorpusz kialakítása felé tett következő lépésnek tekinthető.

3. Anyag és módszer

Az előző szakasz a megszemélyesítések azonosításának és annotálásának elméleti kihívásaiba engedett betekintést. A tanulmány jelen része a gyakorlati megoldási kísérleteket veszi sorra: a PerSE korpusz tervezett struktúrájától és jelenlegi verziójától kezdve a korpuszba kerülő szövegek előfeldolgozásán át a manuális annotálás folyamataig.

3.1. A PerSE korpusz és annak tesztverziója

Egy nyelv megszemélyesítő szerkezeteinek feltárásához olyan átfogó szövegbázisra van szükségünk, amely poétikus szövegeken túl sokféle diskurzustípust tartalmaz. Ezt a sokféleséget képezi le a PerSE korpusz tervezett felépítése: egy irodalmi, egy tudományos, egy publicisztikai és egy hétköznapi alkorpuszból áll majd. Mindegyik alkorpuszba online elérhető szövegek kerülnek: az első esetben regény- és drámarészletek, valamint versek, a második alkorpuszban különböző tudományterületekről származó tanulmányok, a publicisztikai alkorpuszba a már most feldolgozott autótesztek mellett külpolitikai hírek, tudósítások, végül a hétköznapi korpuszba főként blogbejegyzések, fórumszövegek, kommentek. A regiszterek és műfajok változatossága nem csupán a perszonifikációról kialakítani kívánt általános leírás számára fontos, hanem mert ezáltal az empirikus vizsgálatban is érvényesíthető a kognitív nyelvészet egyik alaptétele, mely szerint a figurativitás nem korlátozódik a szépirodalmi diskurzusokra.[28]

Könnyű belátni azonban, hogy a korpuszépítés kezdeti fázisában nincs szükség a teljes annotálni kívánt korpusz anyagára, sokkal inkább egy kis terjedelmű tesztkor-

[27] Galac Ádám, *Megszemélyesítő konceptualizációk a látás, hallás és szaglás fogalmi tartományában: kontrasztív empirikus vizsgálat,* kézirat, 2022.

[28] A kutatás jelenlegi szakaszában a korpusz végleges mérete legfeljebb tervezhető: alkorpuszonként 15 000 – 20 000 tokennel számolva megközelítheti a 80 000 szövegszónyi terjedelmet. Fontos eredmény lehet azonban a korpusz kialakítása során annak a megállapítása mekkora a minimális szövegterjedelem a magyar nyelv megszemélyesítő szerkezeteinek általános leírásához.

 DOI: 10.31400/dh-hun.2022.6.4409

puszra, amely kezelhető mennyiségű szöveget tartalmaz, és amelyen az annotálás menete kialakítható, illetve ellenőrizhető. Az előfeldolgozás és a manuális elemzés elveinek rögzítését követően e tesztkorpusz bővíthető lesz egészen addig, amíg el nem éri a tervezett végső terjedelmet.

Így tehát az első lépésekhez olyan szövegekre volt szükség, amelyek megfelelnek két alapvető kritériumnak: (i) online elérhető írott szövegek legyenek (hogy az átírás és a digitalizálás ne nehezítse a kezdeti adatfeldolgozást), továbbá (ii) kellő számú megszemélyesítést tartalmazzanak. Az online sajtóban megjelenő autótesztek ilyen szövegtípusnak bizonyultak: e szövegek nem csupán műszaki leírást adnak a bemutatott autómodellekről, de azok részletes értékelését is elvégzik, kiemelve az autók előnyeit és hátrányait, bemutatva teljesítményüket, ezáltal ajánlva azokat jövőbeli tulajdonosaiknak. A profitorientáltság ellenére ezek az autótesztek bizonyos fokú professzionalizmussal közelítenek a tesztelt modellekhez, ami a technikai adatok részletezésében és az autógyártókkal (illetve a termékeikkel) szembeni, gyakran kritikus hangvételben is megmutatkozik. A szórakoztató jellegű információs tartalom *(infotainment)* igényéhez igazodva pedig a nyelvi megformálás széles skáláját vonultatják fel a távolságtartó és objektív hangnemtől egészen a szubjektív és értékelő nyelvhasználatig. Éppen ezért e szövegekben tipikusnak mondható az autókra (vagy a cégekre) emberi lényekként utalni, részben, hogy ezzel fokozzák az olvasó személyes bevonódását és elkerüljék a formális-formalizáló attitűdöt, részben pedig, hogy nyelvileg is megformálják a tesztelő professzionális identitását. Noha mindezek alapján a megszemélyesítések alkalmazása általános jellemzőnek tűnik a szövegtípus esetében, egyúttal az is megfigyelhető, hogy minél szubjektívebb és lazább nyelvhasználatra törekedik a cikk szerzője, annál gazdagabb lesz a szöveg perszonifikációkban.

A kellő nagyságú minta eléréséhez (és a kellő mértékű generalizáció lehetővé tételéhez) hat autótesztet[29] válogattam be a PerSE korpusz első verziójába, amely összesen 10486 szövegszót jelent. A szövegek három különböző szerzőtől származnak, így a megszemélyesítés korpuszbeli mintázata nem egyéni nyelvi preferenciák következménye, noha természetesen korlátozott mértékű általánosításokat tesz csupán lehetővé, és a tesztkorpusz semmiképpen sem tekinthető reprezentatívnak.

3.2. A szövegek előfeldolgozása, a projekt infrastrukturális háttere

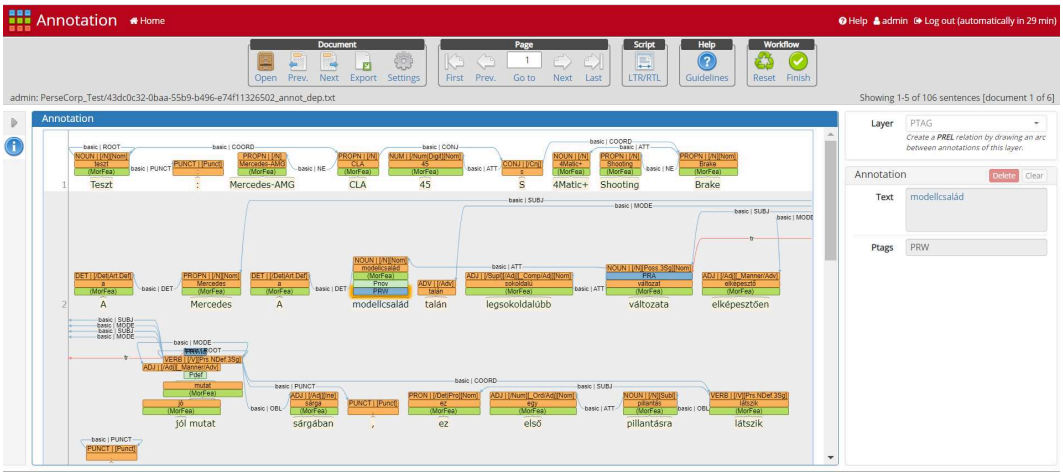
A kézi elemzés megkezdése előtt a korpusz szövegeit az *e-magyar Digitális Nyelvfeldolgozó Rendszer* segítségével elemeztem.[30] A teljes nyelvi anyag tokenizáláson,

[29] A primer szövegek az alábbi url-címeken érhetők el, hozzáférés: 2022.06.01, `https://totalcar.hu/tesztek/2021/07/01/mercedes-amg_cla_45_s_4matic_shooting_brake_teszt/`; `https://totalcar.hu/tesztek/2021/09/10/mercedes-benz-c-300-limousine-amg-line--w206/`; `https://totalcar.hu/tesztek/2021/08/02/bemutato_hyundai_kona_n_2021/`; `https://totalcar.hu/tesztek/2021/07/02/hyundai-ioniq-5-teszt/`; `https://totalcar.hu/tesztek/2021/07/05/skoda_kodiaq_rs_2.0_tsi_dsg_4x4_facelift_bemutato_menetproba/`; `https://totalcar.hu/tesztek/2021/07/27/porsche_cayenne_turbo_gt_teszt_bemutato/`.

[30] Váradi Tamás et al., „E-magyar: A Digital Language Processing System," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)* (Miyazaki: European Language Resources Association, 2018), 1307–1312. A rendszer elérhető az alábbi linken, hozzáférés:

lemmatizáláson, szófaji címkézésen, morfológiai és szintaktikai elemzésen esett át, ezek eredményét CONLLU formátumban nyertem ki az elemzőrendszerből, amely alkalmasnak bizonyult további manuális annotálás elvégzésére.

Ez utóbbi munkafolyamathoz a Webanno online annotáló felületet[31] használtam. Az első ábra szemlélteti a kézi annotálás folyamatát a platformon.



1. ábra. Az előfeldolgozott szöveg annotálása a Webanno felületén

Az automatikus elemzés által felkínált címkéken túl a kézi feldolgozás két további szinttel bővítette az annotálást: a *ptags* készlet a megszemélyesítő kifejezések komponenseinek jelölésére szolgál, míg a *pqual* készlet a konvencionalizáltság tartományait fedi le. Vagyis minden token megkapta a szótő címkéjét, a szótő szófaji kategóriájáét, valamint a szóalak morfológiailag egyértelműsített elemzését, és ehhez járult további két opcionális címke, amennyiben a token megszemélyesítő jelentés kialakításában vesz részt. A szintaktikai függőségi viszonyokat a felület nyilakkal és az azon szereplő címkékkel jelöli. Ezzel megegyező módon, ám manuálisan lehet jelölni a megszemélyesítés komponensei közötti szemantikai viszonyokat (például trajektor és landmark címkéjű nyilakkal). Ily módon a platform lehetővé teszi az automatikus elemzés és a kézi feldolgozás egyesítését, azonos formátumú, mégis elkülönítetten megvalósuló tárolását.

A szótáralapú jelentésegyértelműsítés megvalósításához a *Magyar értelemző kéziszótár* második kiadását használtam, amely az egyetlen jelenleg elérhető, átfogó és legalább részben korpuszelemzésre alapozott szótára a magyar nyelvnek (amennyiben a szógyakorisági adatok a *Magyar Nemzeti Szövegtár* korábbi változatának feldolgozásán alapulnak). A korpuszban a többszavas megszemélyesítések potenciális idioma-

2022.06.01, `https://e-magyar.hu/hu/`. Ezúton köszönöm Indig Balázs technikai segítségét az adatok előfeldolgozásában.

[31] Richard Eckart de Castilho et al., „A Web-based Tool for the Integrated Annotation of Semantic and Syntactic Structures," in *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)* (Osaka: The COLING 2016 Organizing Committee, 2016): 76–84. A platform elérhető az alábbi linken, hozzáférés: 2022.06.01, `https://webanno.github.io/webanno/`.

 DOI: 10.31400/dh-hun.2022.6.4409

tikusságát is jelöltem, ehhez a Hungarian Web 2012 (huTenTen12)[32] korpuszban végeztem kollokációs méréseket, a logDice asszociációs érték[33] mentén. (A kollokálódás küszöbértékének a 6-os logDice pontot tekintettem.)

Az annotálás eredményét TSV3 formátumban exportáltam a Webanno felületről, minden további elemzést *MS Excel* programmal végeztem.

3.3. A megszemélyesítések manuális annotálásának protokollja

A megszemélyesítések azonosításának eljárása Dorst és munkatársainak módszertani javaslatát követi.[34] Ez az eljárás voltaképpen a MIPVU nemzetközi metaforaazonosító protokoll[35] sajátos adaptációja, amelynek a magyar nyelvre kidolgozott változatát[36] is alapul vettem a manuális annotálás mentének kidolgozása során. Egy szó megszemélyesítő használatának azonosítása voltaképpen szótáralapú jelentésegyértelműsítő eljárás: az elemző mindenekelőtt meghatározza a szó szótári alapjelentését és a szövegbeli kontextuális jelentését. Míg az előbbi (jellemzően a szótár által elsőnek megadott jelentés) jellemzően humán figurára utal és konkrét jellegű, addig a második jellemzően absztraktabb és – perszonifikáció esetében – nem humán entitásra vonatkozik. Ha a két jelentés egybeesik, nincs szükség értelemszerűen semmilyen címke kiosztására. Ha azonban a nem humán jellegű kontextuális jelentés összefüggésbe hozható a humánorientált alapjelentéssel, a lexikális elem megszemélyesítésként jelölhető.

3.3.1. Az annotálás címkekészlete

Az eredeti módszer arra ad lehetőséget, hogy a megszemélyesítéseket a lexikális elemek szintjén azonosítsuk, a több szóból álló perszonifikációk belső szerveződését azonban nem tárja fel. Ezért az adaptálás során különböző szinteket alakítottam ki az annotáláshoz, megkülönböztetve a két korábban említett címkekészletet, és a szerkezet komponenseire vonatkozó címkéket a viszonyok jelölésével kiegészítve.

A ptags (komponens-)címkekészlet az alábbi kategóriákat tartalmazza.

- PRW (*[personification-related word],* megszemélyesítéshez kapcsolódó szó): A szónak megszemélyesítő kontextuális jelentése van a korpuszban. Például összetartozó autómodellek csoportjára a *modellcsalád* kifejezéssel utal a szöveg, amely önmagában megszemélyesítő, más lexikális elemek kontextuális hozzájárulása nélkül.

[32] A korpusz elérhető az alábbi linken, hozzáférés: 2022.06.05, `https://app.sketchengine.eu/#dashboard?corpname=preloaded%2Fhutenten12_hp2`. A TenTen korpuszcsaládról lásd Miloš Jakubíček et al., „The TenTen Corpus Family," in Andrew Hardie and Robbie Love, eds., *Proceedings of the 7th International Corpus Linguistic Conference CL* (Lancaster: UCREL, 2013), 125–127.

[33] Pavel Rychlý, „A Lexicographer-friendly Association Score," in Petr Sojka and Aleš Horák, eds., *Proceedings of Recent Advances in Slavonic Natural Language Processing RASLAN* (Brno: Masaryk University, 2008), 6–9.

[34] Dorst, Mulder, and Steen, „Recognition of Personification."

[35] Gerard J. Steen et al., *A Method for Linguistic Metaphor Identification: From MIP to MIPVU* (Amsterdam / Philadelphia: John Benjamins, 2010).

[36] Simon Gábor et al., „Metaforaazonosítás magyar nyelvű szövegekben: egy módszer adaptálásáról," *Magyar Nyelvőr* 143, 2. sz. (2019): 223–247.

- PRA (*[personification-related argument],* megszemélyesítéshez kapcsolódó argumentum): a szó közreműködik egy megszemélyesítő jelentés kialakulásában, de önmagában nem perszonifikáció. Jó példa erre az *Így tol ki […] 387 lóerőt* kifejezés főnévi összetevője *(lóerőt):* a kitol ige alapjelentése humán jellegű ('tolva kívülre juttat vagy mozdít'), ám itt a motor teljesítményére vonatkozik. Ezért a nominális egy megszemélyesítő kifejezés argumentumaként azonosítható.
- PRWid (*[idiomatic personification-related word],* megszemélyesítéshez kapcsolódó idiomatikus szó): a szó önmagában megszemélyesítésként azonosítható; ugyanakkor kollokációs viszonyban áll egy vagy több további szóval a referenciakorpuszban megfigyelhető mintázatok alapján, és e szavakkal együtt alkot idiomatikus kifejezést. Ilyen például a *ki lehet hozni a sodrából* szerkezet, amelyben a *kihoz* ige ('kint lévő helyre hoz') kontextuális jelentése 'nyugodt lelkiállapotából kizökkenti', amely ez esetben egy autó „provokálására" utal. Az ige továbbá erősen (logDice=10,8) asszociálódik a *sodrából* főnévi komponenssel, így egy megszemélyesítés idiomatikus csomópontjaként azonosítható.
- PRAid (*[idiomatic personification-related argument],* megszemélyesítéshez kapcsolódó idiomatikus argumentum): a szó hozzájárul a megszemélyesítő jelentés kialakulásához, mégpedig egy másik szóval alkotott idiomatikus szerkezet tagjaként. Az iménti példa főnévi összetevője *(sodrából)* ilyen idiomatikus argumentumként azonosítható az adatok alapján.
- PRWimp (*[implicit personification-related word,]* megszemélyesítéshez kapcsolódó implicit szó): a szó (a magyarban általában névmás) koreferens viszonyban áll a szöveg egy másik, megszemélyesítésként azonosított kifejezésével. Példaként tekintsük az alábbi mondatot: *Érezhetően tudna az okos C-osztály magától közlekedni a gondosan felfestett és kitáblázott utakon, ha megengedné neki a jogi környezet.* A *C-osztály*ként megnevezett entitás (a Mercedes márka egyik modellje) megszemélyesítve jelenik meg a szövegben (lásd *okos, tudna […] magától közlekedni*); így a főnévre visszautaló *neki* névmás implicit megszemélyesítésként azonosítható.

A szerkezeti komponensek felcímkézése lehetővé tette, hogy a közöttük kibontakozó szemantikai viszonyokat is jelölté tegyük. A prel címkekészlet a következő viszonytípusokra terjed ki.

- tr (*[trajector],* elsődleges figura): az argumentum (PRA vagy PRAid címkével jelölve) az igével jelölt folyamat elsődleges sematikus figuráját (vagyis az ágensét) specifikálja. A *Mercedes megcsinálja a […] ferdehátúját* szerkezetben az autómárkát megnevező főnév az igei folyamat elsődleges figuráját dolgozza ki, így a két token között trajektor viszony létesíthető az annotálás során.
- lm (*[landmark],* másodlagos figura): az argumentum (amely PRA vagy PRAid címkét kapott) az igével jelölt folyamat másodlagos sematikus figuráját (azaz a páciensi, experiensi, recipiensi, instrumentumi vagy egyéb tematikus szerepű résztvevőjét) specifikálja. Az előbbi példában landmarkviszony létesíthető a *megcsinálja* és a *ferdehátúját* tokenek között ennek alapján.
- poss (*[possessive],* birtokviszony): ez a szemantikai viszony jellemzően a testrészmegszemélyesítéseknél adatolható (például a *repülő hátán* szerkezetben), me-

 DOI: 10.31400/dh-hun.2022.6.4409

lyeknél az emberi test alakja (vagy annak egy része) jeleníti meg a fizikai objektumot (vagy annak egy részét). E viszony sajátossága, hogy nem argumentumokra terjed ki, hiszen a birtokviszony szemantikailag referenciapont-szerkezetként modellálhat a kognitív nyelvtanban,[37] amelynek tagjai nem argumentumai egymásnak. Ezért e viszony két PRW-ként címkézett token között létesíthető.

– r (*[relation],* nem specifikált szerkezeti viszony): ez a viszonytípus akkor használható az annotálás során, ha a több szóból álló kifejezés komponensei egymástól elkülönítve jelennek meg (akár közbeékelődő tokenekkel) a magyar nyelv szórendi mintázatai (inverzió, segédigék beférkőzése) következtében. E címke csupán technikai célokat szolgál: egy nem kontinuus kifejezés elemeinek a kapcsolata jelölhető vele, minden további specifikáció nélkül.

A szerkezetre vonatkozó címkék mellett átvettem a konvencionalitás kategóriáit Dorst és munkatársainak korábbi kutatásából,[38] ezek alkotják a pqual címkekészletet. (A kategóriák részletes tárgyalása megtalálható a 2. szakaszban.) A pnov címke jelöli az újszerű megszemélyesítéseket, míg a pconv vonatkozik a konvencionalizálódottakra. Az alapbeállítású megszemélyesítések jelölője a pdef, a metonimikusaké pedig a pmet a sémában.

3.3.2. Az annotálás folyamata

Az alábbiakban lépésről lépésre összefoglalom a manuális annotálás folyamatát.

I. Keressünk megszemélyesítéshez kapcsolódó szavakat (PRW) vagy argumentumokat (PRA) a szövegben, szóról szóra haladva.

   1. Ha a szó alapjelentése emberi lényre vonatkozik, de a kontextuális jelentése nem humán entitásra, jelöljük a kifejezést a PRW címkével.
   2. Ha erős asszociatív viszony figyelhető meg a referenciakorpuszban (logDice ≥ 6) egy másik szóval, jelöljük a kifejezést PRWid címkével.
   3. Ha a szó közreműködik megszemélyesítő jelentés kialakításában argumentumként, jelöljük PRA címkével.
   4. Ha erős asszociatív viszony figyelhető meg a szó és egy másik, PRWid címkével jelölt szó között, jelöljük az adott kifejezést PRAid címkével.
   5. Ha a szó koreferens viszonyban áll a szöveg egy másik, PRW-vel címkézett szavával, jelöljük a kifejezést PRWimp címkével.

II. Jelöljük a szemantikai viszonyokat a PRW/PRWid és PRA/PRAid címkékkel jelölt tokenek között.

   1. Ha az argumentum egy másik kifejezés jelentésének elsődleges figuráját dolgozza ki, létesítsünk tr viszonyt közöttük.

[37] Lásd Ronald W. Langacker, *Essentials of Cognitive Grammar* (New York: Oxford University Press, 2013), 83–85.

[38] Dorst, Mulder, and Steen, „Recognition of Personification."

2. Ha az argumentum egy másik kifejezés jelentésének másodlagos figuráját dolgozza ki, létesítsünk lm viszonyt közöttük.
3. Ha birtokviszony áll fenn két kifejezés között, és a megszemélyesítő jelentés e birtokviszonyon alapul, létesítsünk poss viszonyt a két kifejezés között.
4. Ha egy megszemélyesítő kifejezés összetartozó komponensei nem kontinuusak egymással, létesítsünk r viszonyt közöttük.

III. Értékeljük a megszemélyesítő jelentés konvencionalitását a szótári jelentésadás alapján, és jelöljük a PRW-vel címkézett tagon e konvencionalitás kategóriáját (a pnov, pdef, pconv, pmet címkék egyikével).

4. Eredmények és diszkusszió

4.1. Az eredmények áttekintése

Összesen 958 komponenscímke került kiosztásra aPerSE korpusz tesztverziójában, azaz 9,15%-os relatív gyakorisága van a megszemélyesítő címkéknek a korpusz tokenszámához viszonyítva. Mivel azonban egy szövegszó több címkét is kaphat az annotálás során (hiszen az argumentumok egynél több igéhez vagy igéből képzett névszóhoz is tartozhatnak, és egyes argumentumok saját jogon is megszemélyesítésként azonosíthatók),[39] a megszemélyesítés korrigált gyakorisága a korpuszban 7,81% (összesen 818 annotált tokennel). Másként fogalmazva, közel 8%-a a tesztkorpusz szóelőfordulásainak (tehát átlagosan minden tizenkettedik szövegszó) kezdeményezi megszemélyesítő jelentés kialakulását, vagy legalábbis közreműködik annak kibontakozásában. Noha megfigyelhetők eltérések az egyes szövegeket tekintve, az összkép nem nagyon különbözik a szövegekre történő ráközelítés során sem, miként ezt az *1. táblázat* mutatja.

1. táblázat. A ptags címkék gyakorisága a korpusz szövegeiben

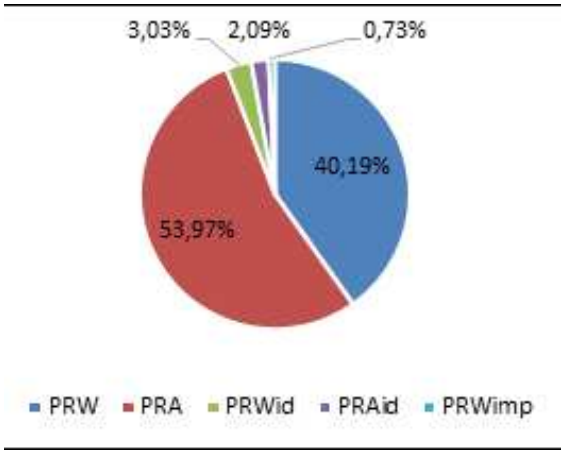
Az autóteszt sorszáma	Az autóteszt terjedelme (tokenszám)	Címkézett tokenek száma (db)	Relatív gyakoriság (%)
T1	2190	152	6.94
T2	1577	145	9.19
T3	1536	152	9.90
T4	2148	144	6.70
T5	1535	111	7.23
T6	1482	114	7.69

A ptags címkekészleten belüli megoszlást tekintve megállapítható, hogy a PRA címkék aránya a legmagasabb a mintában. A második leggyakoribb kategória a PRW címke, ugyanakkor ez utóbbi kategória darabszáma nem éri el a címkézett argumentumok

[39] Ilyen esetre példa a következő szöveghely: *a hátsó futómű […] követi a kocsi orrát,* itt ugyanis az *orrát* nominális önmagában megszemélyesítésként jelölhető (elsődleges jelentésében az emberi szaglószervre utal). Emellett ugyanakkor argumentuma a követi igealaknak, ezért további címkét kapott az annotálás során.

 DOI: 10.31400/dh-hun.2022.6.4409

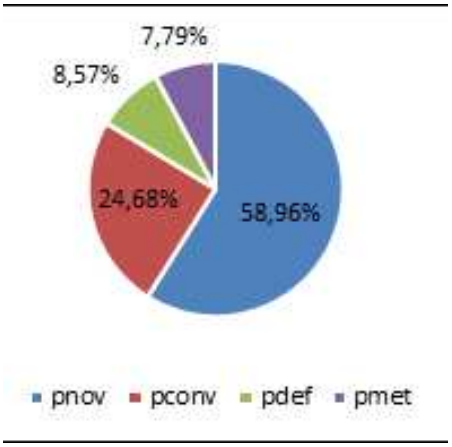
mennyiségét. Megállapítható tehát, hogy átlagosan minden PRW-vel jelölt tokenre esik legalább egy argumentum.[40] Ez a megfigyelés alátámasztja a szakirodalom azon állítását, hogy a nyelvi megszemélyesítések tipikusan egynél több szóból álló kifejezések. Az idiomatikus megszemélyesítések jóval csekélyebb arányban fordulnak elő a korpuszban (mindösszesen 5% körüli gyakorisággal), az implicit perszonifikációk előfordulása pedig még ennél is ritkább (nem éri el az 1%-át sem az annotált tokeneknek). A második ábra részletezi a címkék megoszlását.



2. ábra. A ptags címkék megoszlása a korpuszban

A konvencionalitás skálája kapcsán megfigyelhető az újszerű megszemélyesítések dominanciája a mintában: az azonosított megszemélyesítések több mint fele ebbe a kategóriába tartozik. Ezzel szemben a konvencionális megszemélyesítések csupán az összes kiosztott címke negyedét teszik ki. Az alapbeállítású megszemélyesítések jóval ritkábbak az első két kategóriánál, végül a metonimikus megszemélyesítéseknek van a legalacsonyabb aránya. A harmadik ábra szemlélteti a szemantikai címkék pontos megoszlását a korpuszban.

[40] Természetesen ez nem jelenti azt, hogy ténylegesen minden önálló megszemélyesítésnek van argumentuma. Egy nagyobb mintán célszerű lesz azt is vizsgálni, az átlag mennyire megbízható jellemzője az argumentumok valós megoszlásának.



3. ábra. A pqual címkék megoszlása a korpuszban

Ezek az eredmények nem csupán azt mutatják, hogy a vizsgált online autótesztek bővelkednek megszemélyesítő kifejezésekben, hanem azt is, hogy az esetek többségében ezek a figuratív szerkezetek kreatív, nem konvencionális nyelvi megoldások.[41]

4.2. A megszemélyesítések lexikális mintázata

Az annotálás során előálló következő adattípus azoknak a lexikális egységeknek a gyakorisági listája, amelyek gyakran azonosíthatók megszemélyesítésként a korpuszban. A második táblázat az első húsz leggyakoribb lemmát tartalmazza, bemutatva korpuszbeli gyakoriságukat (Freq), szövegeken belüli gyakoriságukat (FreqT, azaz hány szövegben válnak megszemélyesítéssé a korpuszban), és szemantikai minőségüket (pqual). (Ez utóbbi a kifejezések poliszém jellege miatt eltérő lehet az egyes kontextusokban, ezért egyazon lemma mellett több különböző címke is szerepelhet. Amennyiben a lemma argumentumát adja egy perszonifikáló kifejezésnek, szemantikai címkét nem kapott.)

2. táblázat. A leggyakoribb megszemélyesítő lemmák a korpuszban

Lemma	Freq (db)	FreqT (db)	pqual
tud	20	6	pmet, pnov
ki (igekötő)	10	4	–
motor	9	5	–
erős	8	5	pconv
meg	8	5	–
tart	8	5	pmet, pnov
segít	7	3	pnov

[41] Ezen a ponton természetesen szükséges tekintetbe venni azt a tényt is, hogy a szemantikai minőség meghatározása a kéziszótár jelentésleírásaira épült. Egy részletezőbb, ugyanakkor hasonlóan kurrens szótári adatbázis (például a *Nagyszótár* anyaga) minden bizonnyal precízebb annotálást tesz majd lehetővé a jövőben. Másfelől lényeges ismét hangsúlyozni, hogy ezek az eredmények egyetlen diskurzustípus vizsgált mintájára vonatkoznak; a nyelv egészére irányuló generalizációt a jövőbeni teljes korpusz elemzése teszi majd lehetővé.

 DOI: 10.31400/dh-hun.2022.6.4409

dolgozik	6	5	pconv, pnov
autó	6	3	–
maga	5	4	pconv
csinos	5	3	pconv
minden	5	3	–
orr	5	3	pconv
ő	5	3	pdef, pmet
tesz	5	3	pconv, pmet, pnov
okos	4	4	pnov
el	4	3	–
fenék	4	3	pnov
lóerő	4	3	–
rendszer	4	3	–

Nem meglepő, hogy az *autó* főnév és a hozzá tartozó entitások *(motor, rendszer, lóerő)* szerepelnek a listán, hiszen ezek a megszemélyesítés elsődleges „célpontjai" a vizsgált szövegekben. Ami sokkal inkább figyelemre méltó, hogy a legmagasabb gyakorisággal a *tud* igei lemma rendelkezik, amely a járművek technológiai lehetőségeit humán (mentális) kapacitásokként és/vagy képességként jeleníti meg. Ezt a csoportot gazdagítja az *okos* melléknév is, amely az autóra (vagy annak egy részére) mentális ágensként referál. Vannak továbbá olyan visszatérő igék is a mintázatban, amelyek nagyon általános folyamatok *(dolgozik, tart, segít)* alanyaként jelenítik meg az autókat: ezek leginkább ágenciát tulajdonítanak a járműveknek, de nem specifikálják konkrét cselekvésként azok működését. Végül érdemes kiemelni az autókat emberi testként[42] reprezentáló lemmák csoportját is: ide tartoznak az *erős* és a *csinos* melléknevek (fizikai izomerővel és emberi küllemmel ruházva fel a gépeket), valamint az *orr* és a *fenék* főnevek (amelyek a jármű részeire emberi testrészekként utalnak). Ha ezeknek a mintázatképző lemmáknak a konvencionalitását is megnézzük, azt találjuk, hogy a mentális ágencia tulajdonítása rendre újszerű megszemélyesítésként elemezhető, ezzel szemben az emberi testrészek megjelenítése meglehetősen konvencionális. Az általánosabb igei folyamatok egyaránt lehetnek újszerű és konvencionális megoldások, míg a metonimikus és az alapbeállítású megszemélyesítő szerkezetek nem tűnnek jellemzőnek a centrális lexikai mintázatban.

4.3. A megszemélyesítések grammatikai jellemzői

A korpusz szövegeinek automatizált előfeldolgozása alapos grammatikai elemzéseket is lehetővé tesz. Terjedelmi okokból ezúttal csupán a szófaji kategóriák és a manuálisan annotált címkék viszonyára térek ki, valamint a szemantikai viszonyok megoszlására, ez utóbbi ugyanis a megszemélyesítő kifejezések konstrukciószerű viselkedésére is következtetni enged.

[42] Természetesen ezekben az esetekben úgynevezett *default* interpretációról van szó, azaz arról, hogy prototipikusan e testrészek elsődleges referenciális tartománya az emberi test. Tekinthetjük ugyanakkor ezeket az eseteket a tágabb zoomorfizáció/biomorfizáció eseteinek is. A jelen kutatásnak nem célja e konceptuális kategóriahatárok alapos feltérképezése.

Érdekes eltérésre figyelhetünk fel, ha a ptags és a pqual címkék szófaji mintázatait vizsgáljuk. Miközben az előbbi címkék 39,04%-át nominális token kapta meg a korpuszban (összesen 24,22% került igére), ellenkező arányt látunk a második címkekészletnél: 51, 95% igealakhoz rendelődött (miközben csupán 14,29%-ban minősítenek főnevet, és 23%-ban melléknevet). Ezen a ponton érdemes ismét figyelembe venni, hogy csupán azok a tokenek kaphatnak az annotálás során pqual címkét, amelyek önmagukban (tehát nem argumentumként) tekinthetők megszemélyesítőnek (azaz a PRW, PRWid vagy PRWimp kódúak). Ez az eredmény tehát arra enged következtetni, hogy a leggyakrabban címkézett megszemélyesítő komponensek nominális kifejezések, ugyanakkor igék és melléknevek alkotják a megszemélyesítő szerkezetek csomópontját az esetek nagy többségében (74,29%-ban mindösszesen). És miközben a korpusz összes főnevének 15,86%-a kapott megszemélyesítő címkét, az igék esetében ez az arány 19,27%. A pqual címkékre áttérve megállapítható, hogy csupán a főnevek 2,33%-a részesült ilyen címkében a korpusz egészét tekintve, ám az igék esetében ez az arány 16,61%. (A melléknevek részesülése viszonylag alacsony mindkét esetben: 6,81% vált megszemélyesítő komponenssé, és 5,3% kapott szemantikai minősítést is.) Ezek az eredmények ismét alátámasztják a megszemélyesítések többszavas kifejezés jellegét a magyarban, másrészt hosszabb távon segíthetik egy szófaji elemzésen alapuló, félig automatizált megszemélyesítéseket annotáló eljárás kialakítását.

Mindezek alapján korántsem meglepő, hogy a PRW címkét kapó tokenek körében az igék a leggyakoribbak, 48,83%-kal). Második helyen a melléknevek állnak e kategóriában (23,12%), majd a főnevek (17,66%). Ezzel szemben a PRA kategóriában a főnév dominál (56,09%), amelyet a névmások (17,19%) és a tulajdonnevek (12,38%) csoportja követ. Az idiomatikus kifejezések körében is hasonló mintázattal találkozunk: miközben az igék (65,52%), a melléknevek (17,24%) és az adverbiálisként elemzett tokenek (6,9%) bizonyultak a leggyakoribb PRWid adatoknak, az ennek megfelelő PRAid címke jórészt főnevekhez (80%), névmásokhoz (15%) és adverbiális elemekhez (5%) került. Értelemszerűen az implicit megszemélyesítések alapvetően valamilyen névmási alakként jelentek meg (85,72%-ban). Egyszerűbben fogalmazva, az igék és a melléknevek tekinthetők a leginkább feltűnő megszemélyesítéseknek a korpuszban, míg a főnevek, névmások és tulajdonnevek jellemzően az argumentumai a megszemélyesítő szerkezeteknek.

Az igei és melléknévi perszonifikációk valószínűsíthető feltűnősége mellett egy további adatsor is felhozható érvként. A konvencionalitási skála mindegyik csoportjában e két szófaji kategória emelkedik ki arányaiban, miként ezt a harmadik táblázat is mutatja.

3. táblázat. A szófaji kategóriák megoszlása a pqual annotálási szintjén

Szófaji kategória	pnov (%)	pconv (%)	pdef (%)	pmet (%)
ige	58.15	42.11	39.39	50
melléknév	17.18	32.63	30.30	26.67
főnév	13.22	18.95	15.15	6.67

A grammatikai elemzés utolsó aspektusa a csomópont és az argumentum(ok) közötti szemantikai viszonyokat érinti. Mivel a birtokviszony meglehetősen ritka a korpusz-

 DOI: 10.31400/dh-hun.2022.6.4409

ban (mindössze 17 előfordulással), a továbbiakban a trajektor és a landmark viszonyokra fókuszál az elemzés. Az előbbit 236 esetben létesítettem az annotálás során, míg az utóbbira összesen 198 példa van. Az elsődleges figura kidolgozásának nagyobb arányára plauzibilis magyarázattal szolgálhat a melléknevek és az adverbiális elemek száma: ezek a tokenek (melyek 15,56%-át teszik ki az összes kiosztott komponenscímkének) jelzőként vagy határozóként funkcionálnak az elemi mondatban, így a jelzett szó, illetve a határszóval specifikált folyamat figurája az argumentuma lesz a szerkezetnek, és jellemzően a melléknévi/adverbiális jelentés elsődleges figuráját dolgozza ki szemantikailag. Következésképpen, a melléknévi, valamint adverbiális megszemélyesítések száma növeli egyúttal a trajektorviszonyok mennyiségét is.

A szemantikai viszonyok disztribúciójából három tipikusnak mondható konstrukció rajzolódik ki a korpuszban. Az első középpontjában egy megszemélyesítő igealak áll, amelynek elsődleges fokális figuráját egy főnév dolgozza ki (jellemzően ez lesz a megszemélyesített entitás), majd egy vagy több argumentum specifikálja az ige eseményszerkezetének további összetevőit. Ilyenre példa: [a biztonsági rendszer] *mindenre halálosan figyel* kifejezés. A második konstrukciótípus két összetevőből áll: egy melléknévből vagy egy adverbiális elemből (amely a megszemélyesítő jelentés fogalmi keretét aktiválja) és egy nominális argumentumból (amely e keret centrális szereplőjeként perszonifikálja a szövegvilágbeli entitást). Ez a konstrukció valósul meg a *cinikus reménytelenség ül a vállamon* kifejezés első felében. A harmadik típusba a nominális megszemélyesítések tartoznak, melyek általában testrésznevet tartalmaznak, és amelyekben a két főnév birtokviszonyban áll egymással, mint például *a repülő hátán* szerkezetben. A legkevésbé gyakori eset, amikor a megszemélyesítő kifejezés önmagában áll, és további nyelvi komponensek bevonása nélkül kezdeményez perszonifikáló konstruálást (például egy autó vagy mechanikus rendszer *erős* entitásként való jellemzése).

5. Összegzés és kitekintés

Ez a tanulmány a PerSE korpusz megalapozásának és aktuális verziójának részletes bemutatására vállalkozott. Általánosabb célja a megszemélyesítés szisztematikus elemzésének megvalósítása (de legalábbis a megkezdése) volt a magyar nyelvre vonatkozóan, korpusznyelvészeti eszközökkel. Az elemzések alapjául szolgáló korpusz egyaránt tartalmaz általános nyelvi információkat (szófaji és morfoszintaktikai elemzés), valamint a megszemélyesítő kifejezések kézi annotálását. A korpusz a jövőben további alkorpuszokkal egészül majd ki.

A jelenlegi tesztváltozat több mint 10 000 szövegszót tartalmaz online elérhető autótesztekből. A szövegek tokenizálása, lemmatizálása, grammatikai előfeldolgozása az *e-magyar* digitális eszközzel történt. A manuális annotáláshoz külön protokollt dolgoztam ki a magyarra (a korábbi nemzetközi javaslatra építve), két külön címkekészlettel (a megszemélyesítő szerkezetek nyelvi komponenseinek és a megszemélyesítő jelentés konvencionalizáltságának jelöléséhez), majd implementáltam az eljárást a Webanno felületen, kiegészítve a platfrom nyújtotta lehetőségekkel (a szemantikai viszonyok jelölésével). Az annotálás eredményeként előálló korpuszban mind a lexikális mintázatokat, mind a grammatikai jellemzőket megvizsgáltam.

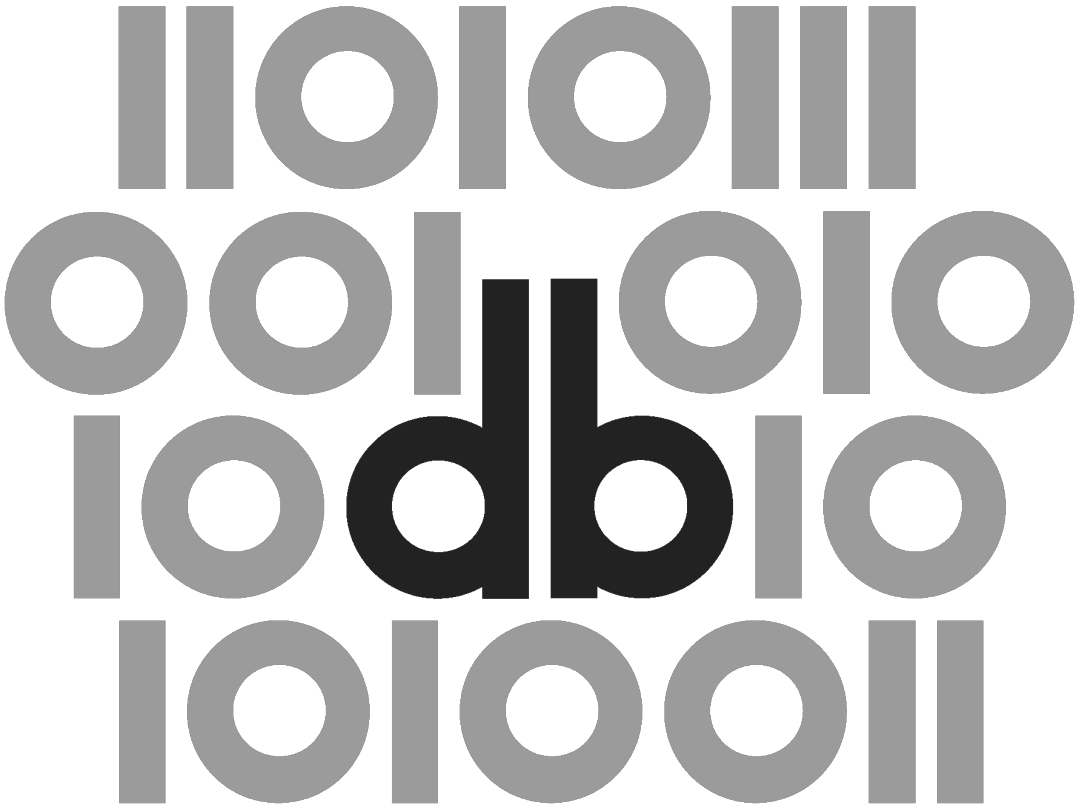
A PerSE korpusz továbbfejlesztésére több lehetőség is kínálkozik. Mindenekelőtt szükséges lesz további annotátorok bevonására, és ezen keresztül az azonosítási protokoll megbízhatóságának a felmérésére, illetve magának a protokollnak a szükség szerinti finomítására. A morfoszintaktikai elemzés további kiaknázása is lehetséges, finomabb elemzésekkel minden bizonnyal specifikusabb konstrukciós minták is kinyerhetők. Emellett a korpusz feldolgozása kiterjeszthető a megszemélyesítő jelentés fogalmi tartományának elemzésére is, létező lexikális szemantikai adatbázisokra (például Wordnet) vagy a keretszemantikai adatbázis (FrameNet) adaptálására építve. Természetesen a fejlesztés fő iránya a korpusz méreteinek növelése, újabb szövegek, szövegtípusok bevonása az annotálásba. A megszemélyesítésről így felhalmozódó nyelvészeti ismeretek pedig várhatóan segíteni fogják a nyelvtechnológiai kihívások (mint amilyen a megszemélyesítés automatikus azonosítása) jövőbeni teljesítését is.

The Corpus-driven Investigation of Personifications in Hungarian: The PerSE corpus

Despite the recent findings on the conceptual and linguistic organization of personification, we have relatively little knowledge about its lexical patterns and grammatical templates. It is especially true in the case of Hungarian which has remained an understudied language regarding the constructions of figurative meaning generation. The present paper aims to provide a corpus-driven approach to personification analysis in the framework of cognitive linguistics. This approach is based on the building of a semi-automatically processed research corpus (the PerSE corpus) in which personifying linguistic structures are annotated manually. The present test version of the corpus consists of online car reviews written in Hungarian (10468 words altogether): the texts were tokenized, lemmatized, morphologically analyzed, syntactically parsed, and PoS-tagged with the *e-magyar* NLP tool. For the identification of personifications, the adaptation of the MIPVU protocol was used and combined with additional analysis of semantic relations within personifying multi-word expressions. The paper demonstrates the structure of the corpus as well as the levels of the annotation. Furthermore, it gives an overview of possible data types emerging from the analysis: lexical pattern, grammatical characteristics, and the construction-like behaviour of personifications in Hungarian.

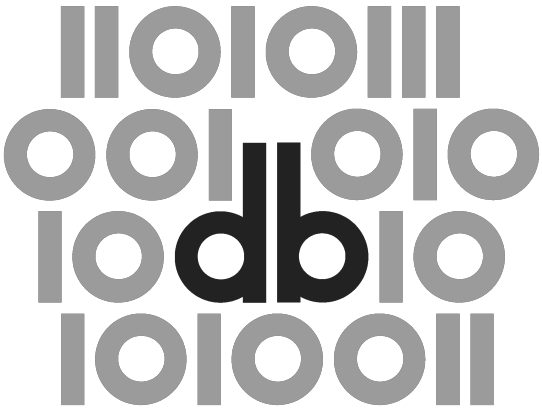
Keywords:
personification, corpus, annotation, analysis

 DOI: 10.31400/dh-hun.2022.6.4409



Digitális Bölcsészet
2022., hatodik szám

<DIGITÁLIS BÖLCSÉSZET>



6 (2022)

ISSN 2630-9696
DOI 10.31400/dh-hun.2022.6

Kiadja a Bakonyi Géza Alapítvány és az ELTE BTK Régi Magyar Irodalom Tanszéke (1088 Budapest, Múzeum krt. 4/A).
Felelős kiadó az ELTE BTK Régi Magyar Irodalom Tanszék vezetője.

Megjelenik az Open Journal Systems (OJS) v. 3. platformon, melynek működtetését az ELTE Egyetemi Könyvtár- és Levéltár biztosítja.

Honlap: `http://ojs.elte.hu/digitalisbolcseszet`
Email cím: dbfolyoirat@gmail.com

Olvasószerkesztő: Bucsics Katalin
Tördelés: Hegedüs Béla
Grafika: Hegyi Gábor

<MŰHELY>

Knap Árpád 0000-0002-4290-6025

ELTE Társadalomtudományi Kar

knap.arpad@tatk.elte.hu

Tóth Tímea Emese 0000-0002-3584-118X

ELTE Társadalomtudományi Kar

toth.emese@tatk.elte.hu

Rakovics Zsófia 0000-0002-9903-9348

ELTE Társadalomtudományi Kar

zsofia.rakovics@tatk.elte.hu

Humán annotált emóciókorpusz létrehozása aktorokhoz köthető érzelmek detektálására

Tanulmányunkban egy olyan kutatási projektet mutatunk be, amelyben egy aktorokhoz (pl. intézményekhez, személyekhez) kapcsolódó, szentimentek és konkrét érzelmek klasszifikációjára képes nyelvi modell létrehozása a célunk. A modell tanítóadatbázisát egy tízezer cikkből álló, online újságokból származó, statisztikai mintavétel segítségével összeállított, humán annotált szövegkorpusz jelenti. Az annotálás során két lépcsőben először az előforduló névelemeket, illetve aktorként funkcionáló közneveket, majd ezt követően a névelemek szövegkörnyezetében megtalálható szentiment- és érzelmi tölteteket annotáljuk. Az annotált szövegek adatbázisa jó bemeneti adatot jelenthet felügyelt klasszifikációs modellek létrehozásához. Cikkünkben ismertetjük a projekt korpuszát, a felügyelt és nem felügyelt szövegklasszifikációs eljárások sajátosságait, valamint a szentiment- és érzelemdetektálás lehetséges módszereit. Ezt követően bemutatjuk a kutatásunkban alkalmazott kétlépcsős annotálási módszertant, az ennek kialakítása során felmerült problémákat és kihívásokat, illetve azokat a kutatói döntéseket, amelyeket a létrehozni kívánt modell társadalomtudományos felhasználhatóságának érdekében hoztunk meg.

Kulcsszavak:
humán annotáció, szentimentdetektálás, érzelemdetektálás, szövegklasszifikáció, felügyelt modellek

1. Bevezetés, a projekt célkitűzései

A politikai és közéleti diskurzusok elemzése során szerzett tapasztalataink[1] azt mutatják, hogy a szövegek affektív, érzelmi töltetének automatizált meghatározása kulcsfontosságú elemzési eszközt jelentene a nagy mennyiségű szöveget értelmezni kívánó kutatók számára. A konkrét entitásokhoz, tehát személyekhez, intézményekhez vagy akár eseményekhez köthető érzelmek klasszifikálása az ilyen módszerek kiterjesztéseként fogható fel, amely segítségével az érzelmeket hordozó szavak tárgya is azonosíthatóvá válik. A jelenleg szabadon elérhető eszközkészlettel egy ilyen jellegű kutatás úgy valósítható meg, hogy névelem-felismerés segítségével azonosítjuk a kérdéses szereplőket, ezt követően a szereplők szövegkontextusát valamilyen módon definiálva leválasztjuk, majd egy szentiment- vagy érzelemszótár segítségével hozzárendeljük az affektív töltetet az egyes szövegrészekhez. Jelenlegi kutatási projektünk elsősorban a szótáras megoldások inherens gyengeségeire reflektál, illetve arra kínál megoldási javaslatot, hogy a nyelvészeti névelemfogalom kiterjesztésével a köznevekkel jelölt szereplők azonosítása is lehetővé váljon az elemzett szövegekben, amely társadalomtudományos célzatú kutatásokban kulcsfontosságú. Kutatásunkban tehát olyan, gold standard korpusz létrehozását tűztük ki célul, amely bemeneti adatként szolgálhat felügyelt nyelvi modellek tanításához. Projektünk végső célja egy klasszifikációs modell létrehozása, amelyhez transzformer-alapú nyelvi modellt tervezünk alkalmazni.

Jelen írásunk lényegében beszámoló a fenti céllal elindított kutatásunkról, amelyben bemutatjuk a projekt célkitűzéseit, módszertanát, a névelemek és az érzelmek kódolása során alkalmazott kétlépcsős annotálási folyamat logikai struktúráját, valamint reflektálunk azokra a megfontolásokra, amelyek a modell társadalomtudományos felhasználásának igényéből fakadnak. Kutatásunk jelenleg is zajlik: a 2023. januári állapot szerint a névelemeket az általunk összeállított, tízezer cikkből álló korpuszban két független kódoló annotálta, az érzelemannotálás pedig hozzávetőlegesen 20 százalékos készültségi szinten áll.

2. Korpusz és adatok

Mivel tehát a projekt céljaként meghatározott modellt elsősorban társadalomtudományos kutatásokban való felhasználásra kívánjuk optimalizálni, ezért nem csupán nyelvészeti szempontokat érvényesítettünk a korpusz kiválasztása és a módszertan összeállítása során. Az annotálandó szövegek kiválasztásakor fontos volt számunkra,

[1] Lásd például Ildikó Barna és Árpád Knap, „Antisemitism in Contemporary Hungary: Exploring Topics of Antisemitism in the Far-Right Media Using Natural Language Processing," *Theo-Web* 18, 1. sz. (2019): 75–92, https://doi.org/10.23770/TW0087; Knap Árpád, Bartha Diána, és Barna Ildikó, „Trianon és a holokauszt emlékezetpolitikai jellegzetességeinek elemzése természetesnyelv-feldolgozás használatával," *Szociológiai Szemle* 31, 4. sz. (2021): 28–62, https://doi.org/10.51624/SzocSzemle.2021.4.2; Ildikó Barna és Árpád Knap, „Analysis of the Thematic Structure and Discursive Framing in Articles about Trianon and the Holocaust in the Online Hungarian Press Using LDA Topic Modelling," *Nationalities Papers,* 2022. május 16., 1–19, https://doi.org/10.1017/nps.2021.67; Kmetty Zoltán és Knap Árpád, „Trágárság mint érzelmi válasz a COVID-19 járvány idején," in Szabó Gabriella, szerk., *Érzelmek és járványpolitizálás: Politikai érzelemmenedzserek és érzelemszabályozási ajánlataik Magyarországon a COVID-19 pandémia idején,* 173–190 (Budapest: ELTE Eötvös Kiadó, 2022).

 DOI: 10.31400/dh-hun.2022.6.4576

hogy a lehetőségekhez mérten időben és tartalmilag is változatos korpusz képezze a projekt alapját, és így a modellhez felhasználandó tanítóadatbázist. A változatosság azért kiemelt fontosságú, hogy a modell többféle stílusú, tematikájú, illetve eltérő időpontokban keletkezett szövegek klasszifikációja során is megfelelő teljesítményt nyújtson. Ezért a Digitális Örökség Nemzeti Laboratórium és az ELTE Research Center for Computational Social Science (ELTE RC2S2) kutatócsoport együttműködésében megvalósuló Webaratás projektben[2] 2021 júniusáig legyűjtött portálok tartalmaiból választottuk ki a korpuszt, rétegzett mintavétel segítségével, az alábbiaknak megfelelően.

A korpuszban négy weboldal, az *Abcúg,* a *Magyar Idők,* a *Válasz.hu* és a *VS* cikkei szerepeltek, összesen 307915 tétel. A mintavételezés során a rétegképző szempontok a megjelenés éve, a cikk rovata és annak forrása voltak, tehát itt is arra törekedtünk, hogy a megjelenés ideje és témája is – a lehetőségekhez mérten – változatos legyen. A megjelenés éve szerint három kategóriára osztottuk a dokumentumokat: 2001–2012, 2013–2016 és 2017–2020. A rovatokat hét féle kategóriára egyszerűsítettük az eredeti rovatok alapján: (1) belpolitika, közélet, vélemény; (2) színes, egészség, életmód; (3) gazdaság; (4) kultúra; (5) külpolitika, külföld vegyes; (6) sport; (7) rovat nélkül. Kizártuk a mintavételezésből azokat a rekordokat, ahol nem szerepelt a megjelenés éve (26 cikk), illetve a *Magyar Idők* és a *Válasz.hu* portálokról azokat a cikkeket, ahol nem volt rovat (9 cikk). A mintában az *Abcúg*ról 94, a *Magyar Idők*ről 5378, a *Válasz.hu*ról 2816, a *VS*-ről pedig 1714 darab cikk származik. A korpuszban szereplő cikkek végleges elemszáma a kerekítések miatt: 10002.

3. Felügyelt és nem felügyelt módszerek

A szöveganalitikában a klasszifikáció kiemelten fontos feladat. A klasszifikációs eljárások között megkülönböztetünk felügyelt *(supervised),* és nem felügyelt *(unsupervised)* algoritmusokat. A felügyelt modellek figyelembe veszik a bemenő adathalmaz metaadatait, amelyek lehetnek például emberek által megadott címkék, kategóriák – összefoglaló néven annotációk. Ebben az esetben az adathalmazt tanító illetve teszthalmazra bontjuk szét. A klasszifikációs modellt az adathalmazunk felcímkézett (tanító) részén hozzuk létre, és a teszthalmazon értékeljük ki. Gépi tanulásra építő, például neurális hálókon alapuló modellek esetében ilyenkor az algoritmus a bizonyos metaadatokkal, címkékkel rendelkező tartalmak nyelvi tulajdonságait, sajátosságait veszi figyelembe a klasszifikáció során. A létrehozott modell teljesítménye pedig elsősorban keresztvalidációval mérhető.[3]

[2] Balázs Indig, et al., „The ELTE.DH Pilot Corpus – Creating a Handcrafted Gigaword Web Corpus with Metadata," in *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC 2020),* 33–41 (Marseille: European Language Resources Association, 2020).

[3] Németh Renáta, Katona Eszter Rita, és Kmetty Zoltán, „Az automatizált szövegelemzés perspektívája a társadalomtudományokban," *Szociológiai Szemle* 30, 1. sz. (2020): 44–62, `https://doi.org/10.51624/SzocSzemle.2020.1.`; Renáta Németh and Júlia Koltai, „The Potential of Automated Text Analytics in Social Knowledge Building," in Tamás Rudas and Gábor Péli, eds., *Pathways Between Social Science and Computational Social Science,* Computational Social Sciences, 49–70 (Cham: Springer International Publishing, 2021), `https://doi.org/10.1007/978-3-030-54936-7_3`; R. Sathya and Annamma Abraham, „Comparison of Supervised and Unsupervised Learning Algorithms

A nem felügyelt algoritmusok kizárólag az adathalmazban rejlő látens struktúrákra, mintázatokra hagyatkoznak, előzetesen hozzáadott címkéket nem vesznek figyelembe. Az ilyen nem felügyelt módszerek közé tartozik például a topikmodellezés vagy a szóbeágyazási modellek alkalmazása, amelyek képesek szövegkorpuszok látens tematikus struktúrájának, illetve látens szemantikai struktúrájának feltárására.[4] A nem felügyelt és a felügyelt algoritmusok között átmenetet képeznek a félig felügyelt *(semi-supervised)* módszerek, például a Seeded Latent Dirichlet Allocation topikmodell.[5]

Ebben a projektben, az előzetes tesztelések alapján a vektortérmodellek legújabb generációját, a transzformereken alapuló, úgynevezett kontextualizált vektortérmodelleket tervezzük alkalmazni. A korábbi, statikus vektorterekhez képest (pl. Word2vec, fastText, GloVe), ahol minden szónak egy vektortér-reprezentációja áll elő kontextustól függetlenül, a kontextualizált modellek az adott kontextushoz kötik, hogy milyen vektort kap egy szó. A transzformereken alapuló nyelvi modellek közös jellemzője, hogy nem alkalmaznak statikus beágyazást, illetve rendelkeznek bekódoló *(encoder)* valamint kikódoló *(decoder)* elemmel, amelyek különböző neurális hálókból épülnek fel. A kontextualizált modellek egyik legújabb családját a BERT (Bidirectional Encoder Representations from Transformers[6]) jelenti, amely kiugróan magas teljesítményt mutat a legtöbb természetesnyelv-feldolgozással kapcsolatos feladatban. Ennek egyik oka, hogy a modell a korábbiakhoz képest nem egy meghatározott irányban (pl. balról jobbra) olvassa be a szöveget, hanem a teljes szósorozatot egyszerre dolgozza fel, teljes környezetüket figyelembe véve tanulja meg a szavak kontextusát a, amely különösen a többértelmű kifejezések esetében hasznos. BERT-alapú modell már létezik

for Pattern Classification", *International Journal of Advanced Research in Artificial Intelligence* 2, 2 sz. (2013): 34–38, https://doi.org/10.14569/IJARAI.2013.020206.

[4] Lásd például David M. Blei, Andrew Y. Ng, and Michael I. Jordan, „Latent Dirichlet Allocation," *Journal of Machine Learning Research* 3, January (2003): 993–1022; David M. Blei and John D. Lafferty, „Topic Models," in *Text Mining,* 101–124 (Chapman and Hall/CRC, 2009); Tomas Mikolov, et al., „Efficient Estimation of Word Representations in Vector Space," arXiv, 2013. szeptember 6., http://arxiv.org/abs/1301.3781; Armand Joulin, et al., „Bag of Tricks for Efficient Text Classification," arXiv, 2016. augusztus 9., http://arxiv.org/abs/1607.01759; Piotr Bojanowski, et al., „Enriching Word Vectors with Subword Information," arXiv, 2017. június 19., http://arxiv.org/abs/1607.04606; Jeffrey Pennington, Richard Socher, and Christopher Manning, „Glove: Global Vectors for Word Representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP),* 1532–1543 (Doha, Qatar: Association for Computational Linguistics, 2014), https://doi.org/10.3115/v1/D14−1162; Kmetty Zoltán, „Szóbeágyazási vektortérmodellek társadalomtudományi alkalmazása," *Statisztikai Szemle* 100, 2. sz. (2022): 105–136, https://doi.org/10.20311/stat2022.2.hu0105.

[5] Bin Lu et al., „Multi-Aspect Sentiment Analysis with Topic Models," in *2011 IEEE 11th International Conference on Data Mining Workshops,* 81–88 (Vancouver, BC: IEEE, 2011), https://doi.org/10.1109/ICDMW.2011.125.

[6] Jacob Devlin, et al., „BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding," arXiv, 2019. május 24., http://arxiv.org/abs/1810.04805.

 DOI: 10.31400/dh-hun.2022.6.4576

magyar nyelvre is, mint például a huBERT,[7] amelynek a szentimentek és érzelmek klasszifikációját illető használatában már van előzménye.[8]

4. Szentiment- és érzelemdetektálás

A szöveganalitikai kutatások irányainak egyik fő területét a szövegben rejlő affektív, érzelmi tartalmak automatizált felderítésével kapcsolatos munkák jelentik,[9] amelyek két alapvető iránya a szentimentelemzés és az emócióelemzés. A szentimentelemzés jellemzően negatív–semleges–pozitív tengelyen helyezi el a szöveget, míg az emócióelemzés konkrét érzelmeket különböztet meg egymástól (pl. öröm, megvetés, undor[10]). Az iránytól függetlenül, a besorolás végezhető szótáralapon, tehát a szövegben az adott szentimentkategóriával vagy konkrét érzelemmel azonosított szavak és kifejezések megkeresésével. Ez a típusú megközelítés amellett, hogy nagy arányban eredményez hamis találatokat (mivel például egyáltalán nem, vagy rossz megbízhatósággal ismeri fel a tagadást, az iróniát vagy a szarkazmust), nyelvünk agglutináló jellege miatt sem alkalmazható kielégítő teljesítménnyel magyar nyelvű szövegekre. A szótáralapú megoldás, ahogy a nevéből is következik, nagy méretű és megfelelő minőségű szentimentszótár meglétét igényli. Bár léteznek ilyen szótárak a magyar nyelvre is,[11] azonban kutatói tapasztalataink azt mutatják, hogy az általunk vizsgált korpuszok esetében ezek nem adnak megfelelő eredményeket. Az ugyancsak nehezíti a szótáralapú megközelítést, hogy bizonyos szavak jelentése szövegkörnyezettől függően nagyon eltérő lehet (a „balos" szó jelentése például egészen más politikai kontextusban, mint egy nyílászárókkal foglalkozó szakmai fórumon).

A másik megközelítést a szövegek humán (azaz emberek által végzett) annotálása, majd ezt követően felügyelt modellekkel történő klasszifikációja jelenti. Ilyenkor a klasszifikációhoz használt algoritmus, más felügyelt modellekhez hasonlóan, az egyes szentiment- vagy érzelmi kategóriákba tartozó szövegek nyelvi sajátosságait veszi figyelembe. Más nyelveken számos kutatásban sikerrel alkalmazták ezt a megközelítést. Duwairi és Qarqaz arab nyelvű tweeteken és kommenteken alkalmazott sikerrel Naive Bayes, SVM és k-legközelebbi szomszéd eljárásokat.[12] Habernal, Ptáček és Steinberger cseh nyelvű közösségimédia-tartalmakon kísérletezett különböző előfeldolgozási és

[7] Dávid Márk Nemeskey, „Introducing huBERT," in *XVII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2021),* 3–14 (Szeged: Szegedi Tudományegyetem, Informatikai Intézet, 2021).

[8] Zoltán Kmetty, et al., „Miniszterelnöki csata az online térben" in Böcskei Balázs and Szabó Andrea, eds, *Az állandóság változása: Parlamenti választás 2022,* 141–161 (Budapest: Gondolat Kiadó, MTA Társadalomtudományi Kutatóközpont Politikatudományi Intézet, 2022).

[9] Bing Liu, *Sentiment Analysis and Opinion Mining* (New York: Springer Cham, 2012), `https://doi.org/10.1007/978-3-031-02145-9`.

[10] Lásd például Alan S. Cowen and Dacher Keltner, „Self-Report Captures 27 Distinct Categories of Emotion Bridged by Continuous Gradients," *Proceedings of the National Academy of Sciences* 114, 38. sz. (2017): E7900–7909, `https://doi.org/10.1073/pnas.1702247114`.

[11] Lásd például Szabó Martina Katalin és Vincze Veronika, „Egy magyar nyelvű szentimentkorpusz létrehozásának tapasztalatai," in *XI. Magyar Számítógépes Nyelvészeti Konferencia,* 219–226 (Szeged: Szegedi Tudományegyetem Informatikai Tanszékcsoport, 2015).

[12] Rehab M. Duwairi and Islam Qarqaz, „Arabic Sentiment Analysis Using Supervised Classification," in *2014 International Conference on Future Internet of Things and Cloud* (FiCloud), 579–583 (Barcelona: IEEE, 2014), `https://doi.org/10.1109/FiCloud.2014.100`.

klasszifikációs eljárásokkal, azzal a céllal, hogy a tartalmak szentiment töltetét kategorizálják, munkájuk eredményeképpen pedig egy bárki által szabadon hozzáférhető, tízezer Facebook-bejegyzést tartalmazó annotált korpuszt is közzétettek.[13] Shi és Li kínai nyelvű szállodaértékeléseket tartalmazó korpuszon alkalmaztak sikerrel SVM modelleket a szövegek szentimentpolaritásának automatizált értékelésére.[14] Számos olyan kutatás is készült, amelyben az elérhető attribútumszelekciós és gépi tanulási algoritmusok teljesítményét vizsgálják különböző szentimentanalízishez kapcsolódó problémákon.[15]

Az utóbbi néhány évben indultak ilyen típusú, magyar nyelvre irányuló, részben társadalomtudományi vonatkozású kutatások, amelyekből nyilvánosan elérhető eredmények is születtek.[16] Projektünkkel elsődleges célunk ehhez a kutatói munkához csatlakozni, ezen túl pedig általánosságban is vizsgálni a felügyelt klasszifikációs eljárások olyan alkalmazási lehetőségeit, amelyek nem csupán a szentiment- és emócióelemzés feladata során segíthetnek, hanem más, társadalomtudományi szempontból lényeges kutatást is támogatni képesek. Mivel jelenleg még nem létezik szabadon elérhető, mindenki által használható, szentimentek és emóciók klasszifikálására alkalmas, magyar nyelvű szövegekre készített modell, kutatásunkba éppen egy ilyen algoritmus létrehozásának céljával kezdtünk bele.

5. A kétlépcsős annotálási folyamat

A projektben a korpuszban található újságcikkek szövegét kétlépcsős annotálás során látjuk el először névelemcímkékkel, majd a szöveget kisebb egységekre bontva, az annotált névelemek környezetét szentiment- illetve érzelem szempontból értékeljük.

[13] Ivan Habernal, Tomáš Ptáček, and Josef Steinberger, „Supervised Sentiment Analysis in Czech Social Media," *Information Processing & Management* 50, 5. sz. (2014): 693–707, https://doi.org/10.1016/j.ipm.2014.05.001.

[14] Han-Xiao Shi and Xiao-Jun Li, „A Sentiment Analysis Model for Hotel Reviews Based on Supervised Learning," in *2011 International Conference on Machine Learning and Cybernetics* (ICMLC), 950–954 (Guilin: IEEE, 2011), https://doi.org/10.1109/ICMLC.2011.6016866.

[15] Lásd például Yang Liu, Jian-Wu Bi, and Zhi-Ping Fan, „Multi-Class Sentiment Classification: The Experimental Comparisons of Feature Selection and Machine Learning Algorithms," *Expert Systems with Applications* 80 (2017. szeptember): 323–339, https://doi.org/10.1016/j.eswa.2017.03.042; Ajay Deshwal and Sudhir Kumar Sharma, „Twitter Sentiment Analysis Using Various Classification Algorithms," in *2016 5th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)* (ICRITO), 251–257 (Noida: IEEE, 2016), https://doi.org/10.1109/ICRITO.2016.7784960; Jeremy Barnes, Lilja Øvrelid, and Erik Velldal, „Sentiment Analysis Is Not Solved! Assessing and Probing Sentiment Classification," *ArXiv:1906.05887 [Cs],* 2019. június 13., http://arxiv.org/abs/1906.05887.

[16] Ilyen a szabadon hozzáférhető OpinHuBank, amelyben bizonyos entitásokat tartalmazó mondatok, illetve az egyes mondatokra vonatkozó szentimenttöltetek annotációja szerepel (Miháltz Márton. *OpinHuBank: Szabadon hozzáférhető annotált korpusz magyar nyelvű véleményelemzéshez* [Szeged: MTA Nyelvtudományi Intézet, 2013]). Kutatásunkban az OpinHuBank adatbázisához hasonló módszertant követünk, azzal a kiegészítéssel, hogy mi konkrét érzelmeket is annotálunk a szentimentértékek mellett. Szintén megemlítendő a HunEmPoli korpusz, amelyben emóciókategóriákat annotáltak egy speciális nyelvezettel rendelkező, parlamenti beszédeket tartalmazó korpuszon (Ring Orsolya, et al., „HunEmPoli: Magyar nyelvű, részletesen annotált emóciókorpusz," in *XIX. Magyar Számítógépes Nyelvészeti Konferencia,* 187–201 (Szeged: Szegedi Tudományegyetem, 2023).

 DOI: 10.31400/dh-hun.2022.6.4576

Az annotálás mindkét fázisában két független kódoló munkája alapján kódoljuk a szövegeket, a kérdéses esetekben pedig supervisor annotátor dönt a helyes kódolásról.[17] Az aktorokhoz kötődő szentiment- és érzelemdetektálásra alkalmazott nyelvi modell létrehozásához készülő szövegkorpusz megalkotása több ponton is kihívások elé állított minket. A kihívások egy része elméleti, másik része technikai jellegű volt. Fontosnak tartjuk, hogy az alábbiakban reflektájunk az ilyen jellegű tapasztalatainkra annak érdekében, hogy a hasonló projekteken dolgozó kutatói közösség munkájához hozzájárulhassunk.

5.1. Névelemek, aktorok annotálása

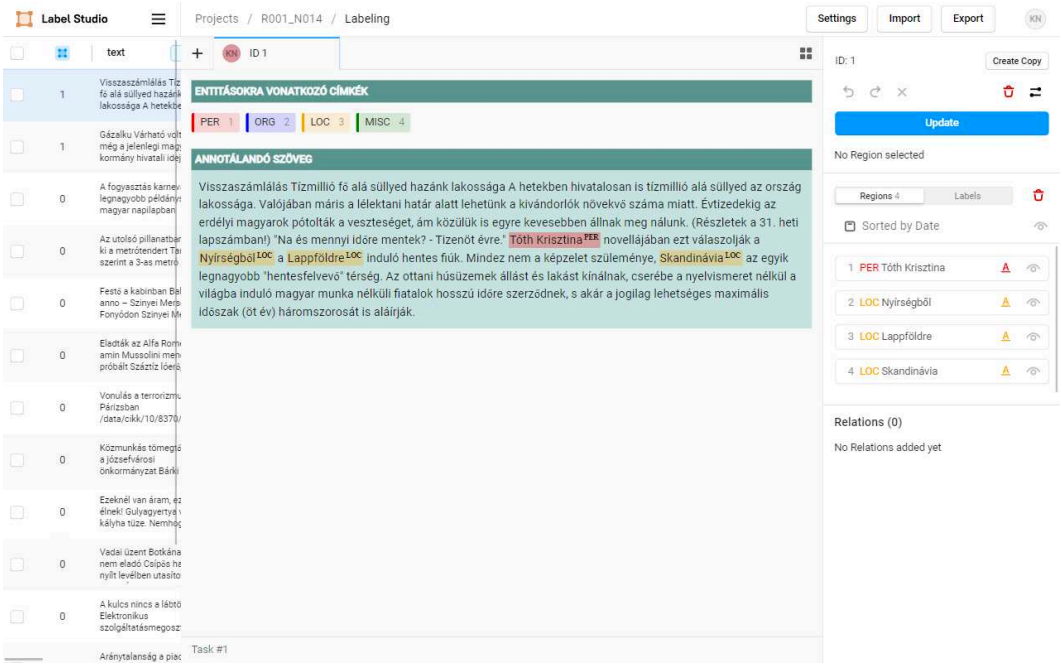
A névelemek annotálását a Simon Eszter és Vadász Noémi által összeállított NerKor annotálási útmutató[18] jelen korpuszra és feladatra kiegészített, példákkal ellátott változata alapján végezzük a *Label Studio*[19] nevű célszoftverben. A névelemek annotálása során elsősorban tulajdonneveket jelölünk a szövegekben, és az alábbi típusokat különböztetjük meg egymástól.

- PER (személynevek): valós és kitalált személyek nevei, becenevek, művésznevek, álnevek.
- ORG (szervezetnevek): intézmények, vállalatok, kormányzati hivatalok, sportcsapatok, múzeumok, egyetemek, szervezett struktúrával rendelkező szervezetek, üzletek stb. nevei.
- LOC (helynevek): országok, városok, földrészek, hegyek, folyók és tavak, tengerek, óceánok, ember alkotta építmények, például repterek, utak, gyárak nevei. A helyek egyaránt lehetnek földrajzilag vagy politikailag definiáltak.
- MISC (egyéb nevek): a fenti csoportok egyikébe sem tartozó nevek, például könyvek és festmények címei, kiállítások, konferenciák, újságok, online hírportálok és médiumok, márkák, televízió- és rádióállomások, ünnepek, programozási nyelvek, kereskedelmi útvonalak, járműmodellek nevei.

[17] Fontosnak tartjuk megjegyezni, hogy a létező automatizált eszközök alkalmazásával szemben azért döntöttünk a névelemek manuális annotálása mellett, mert úgy találtuk, hogy az automatizált megoldások nem teljesítenek megfelelően a projekt célkitűzéseihez, illetve egy *gold standard* annotált névelemkorpuszt is létre kívántunk hozni a kutatásunk során. Ezen kívül, ahogyan azt a tanulmány további részében részletesen ismertetjük, nem csupán tulajdonneveket, hanem társadalomtudományos elemzésekhez rendkívül fontos, a szövegekben aktorokként funkcionáló közneveket is annotálunk, amely a „hagyományos", nyelvészeti névelemértelmezés kiterjesztéseként értelmezhető.

[18] Eszter Simon and Noémi Vadász, „Introducing NYTK-NerKor, A Gold Standard Hungarian Named Entity Annotated Corpus," in Kamil Ekstein, Frantisek Pártl, and Miloslav Konopík, eds., *Text, Speech, and Dialogue – 24th International Conference, TSD 2021, Olomouc, Czech Republic, September 6–9, 2021, Proceedings,* Lecture Notes in Computer Science, 222–234 (New York: Springer Cham, 2021), https://doi.org/10.1007/978-3-030-83527-9_19.

[19] Maxim Tkachenko et al., „Label Studio: Data Labeling Software," 2020, hozzáférés 2023.03.04, https://github.com/heartexlabs/label-studio.



1. ábra. A *Label Studio* szoftverben kialakított, névelem-annotálásra használt felület

A névelemek annotálását minden esetben két, egymástól független kódoló végzi. Ez a gyakorlatban úgy történik, hogy a 10000 cikket tartalmazó korpuszt véletlen sorrendbe rendeztük, ezt követően 50 cikkből álló szeletekre, „pakkokra" bontottuk fel, majd ezeket a pakkokat adjuk ki annotálni a kollegáknak, akik a saját gépükön futó (tehát mások által nem elérhető), nyílt forráskódú *Label Studio* programban végzik az annotálást, majd eredményeiket egy meghatározott felhőtárhelyre töltik fel, ahol kizárólag az adott kódoló és a kutatás vezetői érik el azt. Ilyen módon tehát az annotálást végzők nem látnak rá egymás munkájára, ami a kódolások esetleges torzulását okozhatná.

A két független annotálás eredményét ezt követően a kutatás vezetői automatizált eljárásokkal hasonlítják össze. Ahol nincsen eltérés, tehát a két kódoló ugyanazon szövegrészletet annotálta, és ugyanazt a címkét használta a fentiek közül, azt az annotálást elfogadjuk. Ahol a szöveghatárokban vagy az alkalmazott címkékben eltérés mutatkozik, ott egy supervisor annotátor dönti el, hogy melyik a megfelelő a két verzió közül, illetve ahol egyik sem, ott harmadik, javított annotálást jegyez fel.

Fontos megemlíteni, hogy ezzel a folyamattal párhuzamosan egy módszertani kísérletet is végzünk, amelynek során a kódolóink egy részét arra kértük, hogy ne csupán tulajdonneveket, hanem olyan közneveket is jelöljenek a szövegekben, amelyek aktorokként funkcionálhatnak, és érzelmek kapcsolódhatnak hozzájuk. Az annotált köznevek eltérő címkéket (K_PER, K_ORG, K_LOC, K_MISC) kapnak a tulajdonnevekhez képest, hogy könnyen megkülönböztethetők legyenek a továbbiakban. Ez a gyakorlatban olyan köznevek annotálását jelenti, mint például: főpolgármester, miniszterelnök, államtitkár, szövetségi kapitány, rektor stb. Azt mondhatjuk tehát, hogy projektünkben tartalmi tekintetben általánosabban, módszertanilag specifikusabban szeretnénk az aktorokat megtalálni és hozzájuk érzelmeket társítani. Például azt szeretnénk elérni,

 DOI: 10.31400/dh-hun.2022.6.4576

hogy az általánosabb „józsefvárosi önkormányzat" kifejezés is felismerhető legyen, valamint társíthassunk hozzá szentimentértékeket és érzelmeket, annak ellenére, hogy az entitás hivatalos elnevezése „VIII. kerület Józsefvárosi Önkormányzat Polgármesteri Hivatala". Ezzel a tevékenységgel az a célunk, hogy egy kellően nagy számú, elsősorban politikai és közéleti diskurzusok elemzésekor használható köznévlistát állítsunk össze, amelynek segítségével tehát nem csupán a tulajdonnevekhez köthető, hanem az aktorokként funkcionáló köznevekhez kapcsolódó szentimenteket és érzelmeket is azonosítani tudjuk majd a modell segítségével. A szigorúan vett nyelvészeti névelemfogalommal szemben a társadalomtudományos elemzésekben ugyanis a köznevekkel jelölt, aktorként funkcionáló entitások felismerése is kulcsfontosságú, ezért szükséges egy lehetőség szerint minél bővebb köznévlistát létrehozni. A folyamat során ilyen jellegű példák nyomán elengedhetetlen volt egy olyan szabályrendszer[20] megalkotása, amely alapján pontosan eldönthető, kell-e jelölni az adott köznevet vagy sem. Ennek értelmében az a köznév annotálandó, amely egyértelműen utal valamely, az adott szövegben szereplő entitásként azonosított tulajdonnévre; illetve globálisan felismerhetőnek bizonyul és valamilyen politikai, közéleti entitást jelöl. Emellett rendkívül fontosnak tartottuk, hogy a politikai szférában előforduló, különböző politikai oldalakhoz köthető kifejezések is annotálva legyenek, így a „szocialisták", „fasiszták", illetve azok, amelyek valamilyen médiummal állnak együtt, például a „ballib sajtó". Ez utóbbi kifejezések kulcsfontosságúak olyan társadalomtudományos elemzések esetében, amelynek során például bizonyos politikai oldalak narratíváját vizsgáljuk.

5.2. Érzelem- és szentimentannotálás

A névelemek annotálását követően minden kódolt névelem esetében automatizált eljárás segítségével kiválasztjuk az adott névelem, aktor szövegkontextusát. Az erre vonatkozó előzetes vizsgálataink alapján úgy találtuk, hogy leggyakrabban az adott névelemre a +/- 1 mondatos környezetben vonatkoznak információk, érzelmek, ezért a szövegkontextust úgy definiáltuk, hogy az adott névelem előtti mondatot, azt, amelyben a névelem szerepel, illetve az azt követőt emeljük ki a teljes szövegből.[21] Ezzel a folyamattal együtt vizuálisan megjelöljük az aktuálisan annotált névelemet a szövegben, hogy az érzelemannotálás során akkor is egyértelmű legyen, melyik aktor szempontjából kell az annotálást elvégezni, ha a szövegrészletben több névelem is szerepel. Eddigi annotálásaink során úgy találtuk, hogy egy 50 cikkből álló szövegben

[20] Ahogy említettük, csak az annotálást végzők egy részét képeztük ki a köznevek annotálására. Mivel a köznevek annotálása módszertani kísérlet volt, egyúttal törekedtünk arra, hogy olyan útmutatót írjunk ezek annotálásához, amely kellőképp egyértelmű, ugyanakkor ne járjon túl nagy kapacitással egy kétséges alkalmazhatóságú útmutató értelmezése.
A köznevek annotálásának kiértékelését, illetve az annotálást végzők közti egyetértési arányok kiszámítását a köznevek annotálási folyamatának lezárultával végezzük el.

[21] A projekt tervezése során felmerült a függőségi elemzés alkalmazása az aktorok szövegkontextusának kinyerése érdekében, azonban úgy találtuk, hogy az általunk feldolgozott korpusz esetében a +/- 1 mondatos környezet megbízhatóan működik, és a függőségi elemzés implementálása az alkalmazott munkafolyamatba nagyobb költséggel járna, mint amekkora hasznot jelentene az adatok minősége szempontjából.

általában 200 és 900 között mozog az annotált névelemek száma, tehát ennek megfelelő mennyiségű, érzelmi szempontból annotálandó szövegkontextust generálunk minden egyes „pakkból".

Az annotálandó adathalmaz előállítása után a szentimenttöltetek és érzelmek bekódolása következik. Ehhez szintén a már korábban említett *Label Studio* programot használjuk, és ahogyan az előző fázisnál, ebben az esetben is két, egymástól független kódoló végzi minden egyes szövegrész annotálását. Az annotálási munkamenet az alábbiaknak megfelelően zajlik.

- A kódolók eldöntik, hogy az adott szövegrészt kódoljuk-e. Nincsen szükség a szövegek kódolására olyan esetekben, ahol például az adatok esetleges hibás legyűjtése miatt programkódot tartalmazó szövegrészletről van szó. Szintén nem szükséges a szöveg annotálása, ha például fotós nevével, egy fotó keletkezési helyével vagy helyszínével, vagy akár nevek hosszas felsorolásával (pl. futballcsapatok névsorával) találkoznak. Ezek az elemek, bár megfelelnek a névelemek annotálási kritériumainak, tartalmilag nem járulnak hozzá a szöveghez.
- Ezt követően a kódolók megállapítják a szövegről, hogy van-e benne érzelem, vagy nincs. Ha nincsen, akkor a „nincs érzelem" kódot alkalmazzák, ekkor az adott szövegrésszel nincsen további teendőjük. Fontos továbbá, hogy nem jelölnek az annotátorok érzelmeket arra a személyre vonatkozóan, aki valamilyen semleges tevékenységet folytat, például nyilatkozik. A „nincs érzelem" kód alkalmazandó olyan esetekben is, amikor például a cikk címében hiányos, tehát érzelmi szempontból nem bekategorizálható kontextusban említenek egy adott entitást.[22] Fontos kitételként szerepel az is, hogy a birtokos mondatszerkezetben szereplő birtokoshoz nem rendelhető érzelem, kizárólag a birtokhoz, amennyiben van rá utalás a szövegben.
- A szövegben jelölhető érzelem vonatkozásában negatív, semleges és pozitív szentimentkategóriák közül választhatnak a kódolók. Itt természetesen egynél több kategória is hozzárendelhető ugyanazon szöveghez, amennyiben az egyszerre tartalmaz negatív és pozitív szentimenttöltetet.
- Ezt követően a kódoló azt is bejelöli, hogy a szövegben van-e irónia. Az iróniadetektálás az aktorokként azonosítható köznevek kódolásához hasonlatos módszertani kísérlet, tehát nem központi elem a projektben. Az iróniát gyakran két ember is teljesen máshogyan értelmezi, és lényegében nincsen olyan nyelvi meghatározottsága, nyelvi jegye, amely alapján nyelvi modell segítségével azonosítható lenne. Mivel azonban az ironikus tartalmak megjelölése nem okoz jelentős többletráfordítást az annotálási folyamat során, izgalmas kísérletnek gondoltuk megpróbálkozni egy erre is alkalmas modell létrehozásával.
- A kódoló ezután annak megfelelően, hogy a szöveget a negatív, semleges vagy pozitív szentimentkategóriába sorolta, a bejelölt szentimenteknek megfelelő,

[22] Például a *Vadai üzent Botkának, a DK nem eladó* cím esetében nem egyértelműen eldönthető az író szándéka szerinti érzelemtöltet, ilyen módon pedig az olvasóból kiváltott érzelem kerülne annotálásra.

 DOI: 10.31400/dh-hun.2022.6.4576

konkrét érzelmeket tartalmazó listá(ko)n bejelöli, hogy mely érzelmek vannak jelen a szövegben. A konkrét érzelmek listáját[23] az alábbi táblázat tartalmazza.

1. táblázat. Az egyes szentimentekhez tartozó konkrét érzelmek[24]

Negatív	Semleges	Pozitív
bánat	együttérzés, szimpátia	elégedettség, öröm (elragadtatás, csodálat, rajongás, szórakozás)[25]
düh	érdeklődés, érdekesség	reménykedés, bizakodás, vágyakozás
elégedetlenség[26]	nosztalgia	
félelem, rémület, szorongás	meglepődés (szokatlanság, furcsaság)[27]	
gúnyolódás, undor, megvetés		
irigység, féltékenység		
zavartság, értetlenkedés (kellemetlenség)		

[23] Az érzelmek listájának kialakításához felhasználtuk egy korábbi projekt tapasztalatait lásd Kmetty, et al., „Miniszterelnöki csata az online térben".

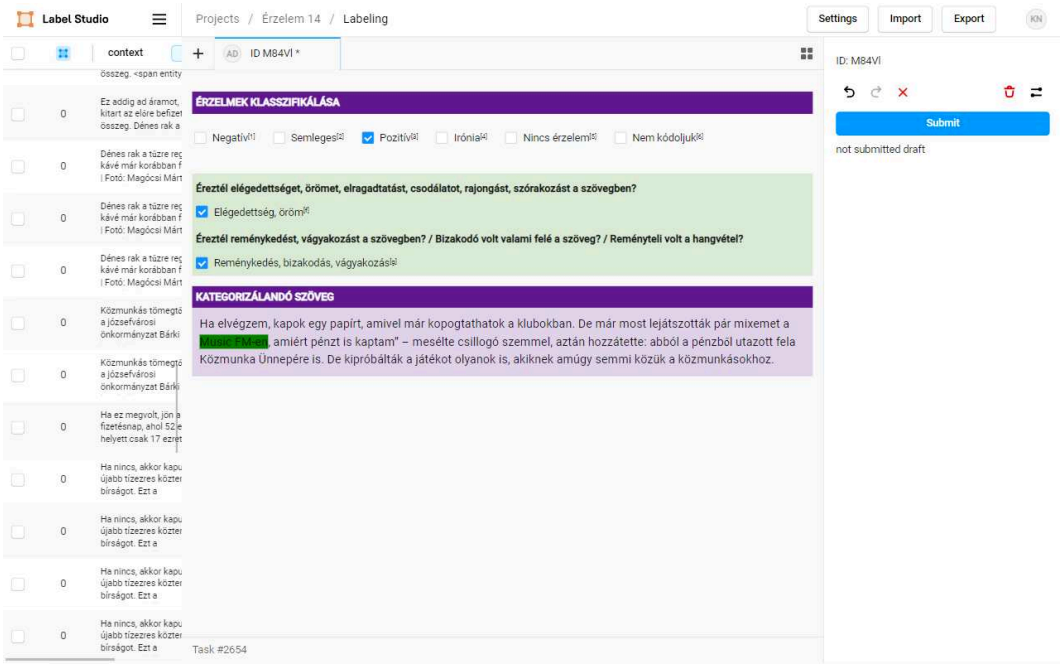
[24] Korábbi annotálási tapasztalataink alapján, amelyet magyar nyelvű internetes kommentek korpuszán végeztünk, a nemzetközileg bevett emóciókategóriák csak részben használhatók a magyar nyelvű szövegekre, a szövegek eltérő érzelmi töltete és érzelmi eloszlása miatt. Ezért ebben a projektben a korábban hivatkozott Cowen és Keltner-féle érzelmi kategóriarendszer egy módosított verzióját használjuk. Az annotálást végzők közti egyetértési arányokat az annotálási folyamat lezárultát követően tervezzük kiértékelni.
A táblázat néhány kategóriájához lábjegyzetben példamondatokat mellékeltünk.

[25] Például: „Jó erőben vagyok, mostanában minden sikerül. Ilyen egyszerű lenne? A Ferencváros már a második, megközelítette a Vasas ellen a Szusza-stadionban összeroppanó éllovas Videotont, a nemrég még sereghajtó Herczeg Andrásnak köszönhetően újra saját nevelésű játékosaira építő Debrecen*MISC* pedig sorozatban negyedik győzelmével feljött a dobogó közelébe." Ch. Gáll András, „Két hét szünet felélesztette a Fradit: Remekelt a bajnok otthonában Thomas Doll csapata, már második a tabellán," *Magyar Idők,* 2017. szept. 18., `https://www.magyaridok.hu/sport/ket-het-szunet-felelesztette-fradit-2224843/`.

[26] Például: „Tulajdonképpen a Varga Roland–Bobál Dávid-párharc döntötte el a mérkőzést, a kilenc forduló alatt tíz gólnál járó válogatott szélső kétszer is szemfülesen megelőzte a halvérű, nem a saját posztján játszó védőt, a harmadik gólt éppen a sokat szidott Botka *PER* beadásából a villámgyors Paintsil szerezte, immár a szakadó esőben, villámlás közepette." Uo.

[27] Például: „Bencsik András*PER* olyat mondott, ami két napja még elképzelhetetlen lett volna" – Vörös Szabolcs, „Bencsik András olyat mondott, ami két napja még elképzelhetetlen lett volna," *Válasz.hu,* 2018. febr. 27., `https://web.archive.org/web/20200126092543/http:/valasz.hu/itthon/bencsik-andras-olyat-mondott-ami-ket-napja-meg-elkepzelhetetlen-lett-volna-127585`.

Az annotálás során tehát hétféle negatív, négy semleges és két pozitív érzelmet különítünk el egymástól. Az érzelmi kategóriák kialakítása során figyelembe vettük korábbi tapasztalatainkat, ami a magyar nyelvű interneten elérhető szövegek – elsősorban hozzászólások – érzelmi megoszlását illeti. Ezek alapján túlnyomó a negatív érzelmek többsége a másik két kategóriához képest. Mivel a modell tanításához minden érzelemből megfelelő mennyiségű annotált szövegre van szükség, ezért bizonyos érzelmek összevonása, azokból nagyobb kategóriák létrehozása mellett döntöttünk, ahogyan az a táblázatban is látható.



2. ábra. A *Label Studio* szoftverben kialakított, szentiment- és érzelemannotálásra használt felület

Nem mehetünk el amellett, hogy az érzelmek annotálásának kihívásairól, nehézségeiről is említést tegyünk. A pszichológiában is alapvető problémakörként említik az érzelmek kategorizálhatóságát, melynek két pillére a szisztematikus kategorizáció nehézsége és az érzelmek univerzalizmusának és kultúrafüggőségének kérdésköre.[28] A pszichológia továbbá megkülönböztet alapérzelmeket és komplex érzelmeket is. Az alapérzelmek univerzálisnak tekinthetők, általában a hozzájuk társított arckifejezések okán, a komplex érzelmek észlelését és összetettségét viszont többek között a kulturális környezet és az egyén is képes befolyásolni. Lindquist (2008) szerint az érzelmek komplexitását növelheti az is, hogy milyen fogalmi kategóriákkal rendelkezik az egyén egy adott érzelem kapcsán, és hogy számára ez milyen összetevőket tartalmaz, tehát mit tud mondjuk a szorongás, vagy a félelem jelenségéről, illetve, milyen formában em-

[28] Hámori Ágnes, „Az érzelmek elemzési lehetőségei a kognitív poétikai kutatásban és korpuszfeldolgozásban," in *Nyelv, poétika, kogníció: Elmélet és módszer a poétikai kutatásban,* 139–173 (Eger: Eszterházy Károly Egyetem Líceum Kiadó, 2018).

 DOI: 10.31400/dh-hun.2022.6.4576

lékszik ezekre, hogyan használja őket.[29] Ennek okán az érzelmek felismerése egyéntől függően még az alapérzelmek vonatkozásában is nagy variabilitást mutathat, azonban annotáláskor törekednünk kell ennek standardizálására. Külön nehézséget okozhat továbbá, hogy megállapítsuk és elválasszuk az olvasás közben ránk törő érzéseket a szöveg írója által közölni kívánt érzelmektől. Ezért volt szükséges standardizálni egyrészt azt az érzelmet, amit az adott annotátornak jelölnie kell a szövegben, másrészt a csapatban dolgozók munkáját is.

Megoldásképp készítettünk egy útmutatót, amelyben rögzítettük, hogy az érzelmi töltet meghatározása a cikk írójának, szerzőjének szándéka szerint detektálandó, a legszigorúbb mondatkörnyezet alapján, vagyis az annotálás a közvetlen mondatkörnyezeti utalások figyelembevételével kell hogy történjen, nem pedig egyéni értelmezés szerint. A „cáfolata annak a balos maszatolásnak, mely szerint a Fidesz összekacsint a Magyar Gárdával" esetén a „balos maszatolás" kifejezés egyértelmű Fidesz-szimpátiát árul el, azonban arról nincs információnk, hogy a cikk szerzője a Magyar Gárdát hogyan ítéli meg, így arra a névelemre a „Nincs érzelem" címke kerül. Speciális esetnek tekinthető, amikor idéznek valakit a szövegben. Ekkor az idézett, a nyilatkozó érzelmei a mérvadóak az adott entitásra vonatkozó érzelem meghatározásához.

Végezetül fontos kiemelni, hogy a projektben elsődlegesnek tekintjük az adatok jó minőségű bekódolását, ezért számos minőségbiztosítási megfontolást is figyelembe vettünk a munkafolyamat kidolgozásakor. Ennek érdekében online oktató videókat, illetve részletes, példákkal illusztrált útmutatót készítettünk mindkét annotálási fázishoz. Szintén a minőségbiztosításhoz tartozik, hogy az annotáláson dolgozó gyakornokok bármikor elérhetik a kutatás vezetőit kérdéseikkel, a specifikus, de mindenki számára hasznos tudást közvetítő problémákat pedig megosztjuk egy közösen elérhető felületen. A folyamatok átláthatósága, követhetősége érdekében kanban rendszerű projektmenedzsment szoftvert, az adatok biztonsága érdekében felhőtárhelyet, a kommunikációhoz pedig könnyen visszakereshető munkahelyi csetplatformot használunk.

6. Összefoglalás

Ahogyan korábban már említettük, a projekthez jelenleg használt korpusz tízezer, változatos időszakokban keletkezett és eltérő témákban íródott cikkből áll. Az annotálási folyamat végén egy olyan humán annotált szövegkorpusszal fogunk rendelkezni, amely becslésünk szerint legalább 80000darab,[30] érzelmi töltetre vonatkozóan bekódolt szövegrészletet tartalmaz majd. Ezen a ponton még kérdéses, hogy ebben a 80000 szövegben pontosan milyen képet mutat majd az egyes érzelmek megoszlása. Amennyiben nagyon kedvezőtlen módon, például szélsőségesen magas lesz az érzelmeket nem tartalmazó, vagy semleges szövegek aránya, szükség lehet az annotált korpusz további

[29] Kristen A. Lindquist and Lisa Feldman Barrett, „Emotional Complexity," in Lisa Feldman Barrett, Michael Lewis, and Jeannette M. Haviland-Jones, eds., *Handbook of Emotions,* 513–530 (New York, London: Guilford Publications, 2008).

[30] Ez jelenleg természetesen csak egy becslés, amelyet a következőképpen kalkuláltunk: a jelenlegi korpusz 10000 darab cikket tartalmaz, amely 200 pakkra oszlik. Úgy találtuk, hogy egy pakkban átlagosan kb. 400 darab annotált entitás szerepel, eszerint a teljes korpuszban kb. 80000 darab entitás jelenik meg.

szövegekkel való kiegészítésére. Amennyiben erre lesz szükség, a további szövegeket olyan módon fogjuk kiválasztani, hogy egy előzetes modell segítségével azokhoz a dokumentumokhoz rendelünk majd nagyobb súlyt a mintavétel során, amelyek a modell szerint nagyobb eséllyel tartalmazzák az előzőleg nem kellő számosságban annotált érzelmi kategóriákat.

Az annotált szövegkorpusz segítségével létrehozott nyelvi modell szándékunk szerint tehát képes lesz magyar nyelvű szövegekben szentiment- és érzelmi tölteteket azonosítani. Különösen érdekesnek tartjuk olyan szentimentanalízis alkalmazását, amely névelem-felismeréssel összekövetve alkalmas a szövegben szereplő különböző aktorokhoz (pl. közéleti szereplők vagy történelmi személyiségek), eseményekhez kötődő érzelmi töltetek detektálására, tehát olyan elemzésre, amelynek során az érzelem tárgyát is lehetséges azonosítani *(entity-level sentiment analysis)*[31]. Az érzelmek detektálásának társadalmi relevanciájához, véleményünk szerint, nem fér kétség: az nemcsak politikai szövegek tartalomelemzése során, hanem a laikus diskurzusban megjelenő vélemények felderítéséhez is elengedhetetlen fontosságú eszközt jelent.

Jelen tevékenységünk célja tehát három részre osztható. Elsődlegesen cél az ismertetett nyelvi modell létrehozása, majd e modell teljesítményének lemérése, végül a projekt tapasztalatainak összegzése. A projekt jellegéből adódó célunk továbbá, hogy az említett módszereket, illetve a kutatás során gyűjtött módszertani tapasztalatainkat más kutatók számára is megismerhetőbbé, hozzáférhetőbbé tegyük, gyakorlati példákkal alátámasztva mindezeket. Nem lenne teljesíthető azonban ez a célkitűzés a létrehozott modell tartalmi fókuszú kutatásokban való, releváns elemzési eszközként való alkalmazása nélkül, amely azonban már túlmutat a jelenlegi projekt szűkebb célkitűzésein.

Köszönetnyilvánítás

Ezúton szeretnénk megköszönni a projektben dolgozó ELTE TáTK Szociológia alapszakos hallgatók munkáját, akik nélkül kutatásunk nem valósulhatna meg, illetve Dömötör Andreának az érzelemannotálási útmutató megírásában nyújtott segítségét.

Creating a Human Annotated Emotion Corpus for the Detection of Actor-related Emotions

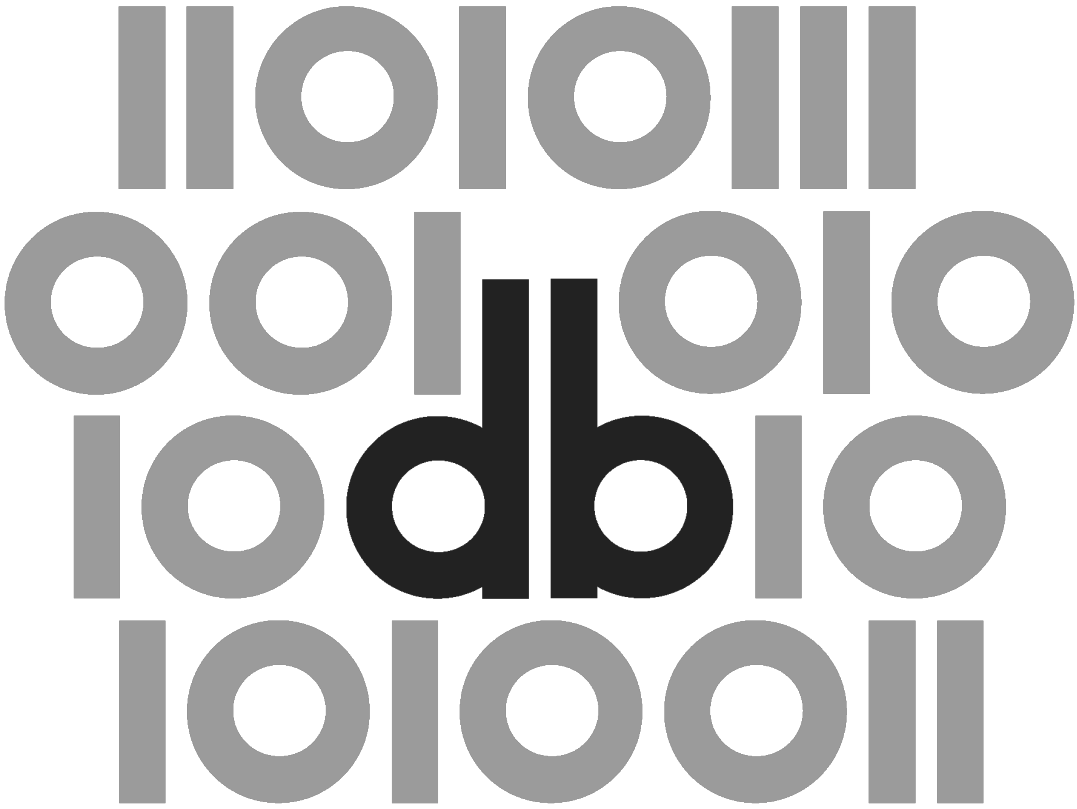
In our study, we present an ongoing research project in which our goal is to create a language model capable of classifying sentiments and specific emotions related to actors (e.g., institutions, persons). The training database of the model is a human-annotated text corpus consisting of ten thousand articles from online newspapers, compiled using statistical sampling methods. In the project, we employ a two-phase annotation design. First, we annotate named entities and common names that function as

[31] Lásd például: Jin Ding et al., „Entity-Level Sentiment Analysis of Issue Comments," in *Proceedings of the 3rd International Workshop on Emotion Awareness in Software Engineering* (ICSE '18), 7–13 (Gothenburg: ACM, 2018), https://doi.org/10.1145/3194932.3194935.

 DOI: 10.31400/dh-hun.2022.6.4576

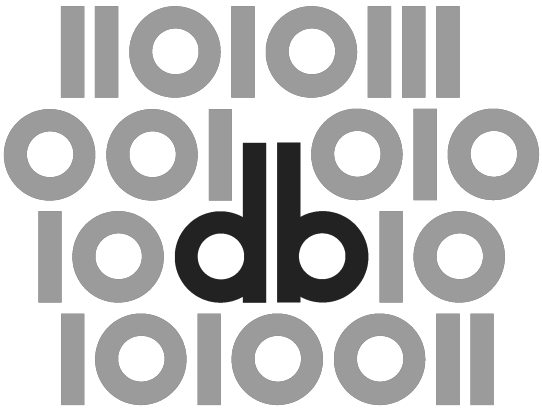
actors. Second, we annotate sentiments and specific emotions found in the context of the previously marked actors. Such a database of annotated texts can provide excellent input for creating supervised classification models. In this article, we describe the corpus of the project, the characteristics of supervised and unsupervised text classification procedures, and possible methods for sentiment and emotion detection. After that, we present the two-phase annotation methodology used in our research, the problems and challenges that arose during its development, as well as the research decisions that we made to create a model that can be used as a capable research tool in social sciences.

Keywords:
human annotation, sentiment detection, emotion detection, text classification, supervised models



Digitális Bölcsészet
2022., hatodik szám

<DIGITÁLIS BÖLCSÉSZET>



6 (2022)

ISSN 2630-9696
DOI 10.31400/dh-hun.2022.6

Kiadja a Bakonyi Géza Alapítvány és az ELTE BTK Régi Magyar Irodalom Tanszéke (1088 Budapest, Múzeum krt. 4/A).
Felelős kiadó az ELTE BTK Régi Magyar Irodalom Tanszék vezetője.

Megjelenik az Open Journal Systems (OJS) v. 3. platformon, melynek működtetését az ELTE Egyetemi Könyvtár- és Levéltár biztosítja.

Honlap: `http://ojs.elte.hu/digitalisbolcseszet`
Email cím: dbfolyoirat@gmail.com

Olvasószerkesztő: Bucsics Katalin
Tördelés: Hegedüs Béla
Grafika: Hegyi Gábor

<KRITIKA>

Király Péter 0000–0002–8749–459

Göttingen eResearch Alliance Gesellschaft für wissenschaftliche Datenverarbeitung mbH

peter.kiraly@gwdg.de

Folgert Karsdorp, Mike Kestemont, and Allen Riddel. *Humanities Data Analysis: Case studies with Python*. Princeton and Oxford: Princeton University Press, 2021. 360 oldal. ISBN 9780691172361

A Folgert Karsdorp, Mike Kestemont és Allen Riddel alkotta szerzőhármas a szó szoros értelmében nehéz könyvet tett le az asztalra. A keménytáblás borító és a magasfényű papír miatt a kötet súlya 1,2 kg. Az oldalak fényvisszaverődése miatt napsütésben, tükröződő lámpafényben nehéz olvasni. Aki azonban eme nehézségeken átküzdi magát, igen színvonalas tartalommal töltekezhet. Mindemellett immár nyílt eléréssel, online könyvként is olvasható a `https://www.humanitiesdataanalysis.org/`-on, a felhasznált adatok pedig letölthetőek a *Zenodo* adatrepozitóriumból.[1]

A könyv célja, hogy tipikus esettanulmányokon keresztül bevezetést nyújtson a Python-alapú digitális bölcsészeti kutatás módszertanába, ehhez a Python programozási nyelvet választotta útitársként. A Python választása logikus, hiszen talán a legnépszerűbb nyelv ezen a területen, és viszonylag könnyű eljutni odáig, hogy a tanuló megírja és lefuttassa az első saját kódsorait. Bár a könyv a szerzők szerint igényel valamennyi Python-ismeretet, az olvasó nemcsak a digitális bölcsészetbe, de a nyelvbe is bevezetést nyer az alapoktól kezdve, még ha csak vázlatosan is (a kezdő olvasó számára megfelelő bevezető könyvet ajánlva). Az esettanulmányok kiválasztása is szilárd talajon áll, többnyire a digitális bölcsészeti szakirodalom egy-egy tanulmányát választották kiindulási pontnak, így az olvasó könnyen utánanézhet a könyv által nem tárgyalt részleteknek.

A könyv két nagy részből áll: az első négy fejezet az adatelemzés alapjait (programozási alapok, a Python által használt főbb adatszerkezetek, a legfontosabb fájltípusok és kezelésük); a többi öt fejezet pedig a haladó adatelemzés egy-egy kiemelt tárgykörét ismerteti (a statisztika alapjai, valószínűségszámítás, térképek, témamodellek [Topic Model] és stilometria). A fejezeteknek kötött dramaturgiájuk van. Bevezetéssel kezdődnek, melyek ismertetik az adott helyen érvényes kutatási kérdéseket, a feldolgozandó forrásokat, sőt néhány esetben a kérdés kutatástörténetére is kitérnek. A fejezetek végén további olvasmányokra vagy tudnivalókra hívják fel a figyelmet, illetve kezdő, mérsékelten nehéz és kihívásszámba menő gyakorlópéldák találhatók. Bizonyos, a témához tartozó, de a bevezetési szinten túli, haladóknak szánt fogalmat vagy módszert csak a példák tartalmaznak. Aki a könyvből igazán profitálni szeretne, annak azt tanácsolom, hogy oldja meg a példákat. A tanároknak pedig, akik a könyvet választják egy digitális bölcsészeti kurzus tananyagául (hiszen ez erre a legteljesebb

[1] `https://doi.org/10.5281/zenodo.891264`.

mértékben alkalmas), a feladatok lehetőséget adnak arra, hogy rávilágítsanak a Pythonban megvalósítható különböző módszerek előnyeire és hátrányaira.

Az első fejezet *(Bevezetés)* nagyon röviden áttekinti a bölcsészetben alkalmazott kvantitatív adatelemzés történetét, majd átveszi a Python legfontosabb eszközeit (változók, sorozatok indexelése, iteráció, listák, halmazok és szótárak, feltételes utasítások, külső modulok importálása, függvények definiálása, fájlműveletek). A könyv tucatnyi kódkönyvtárra támaszkodik, és sajnos elkerülhetetlen, hogy némely esetben ezek újabb változatainak funkciói nem kompatibilisek már a könyv írása idején aktuálissal, ilyen esetekben szükséges a könyvtár eredeti dokumentációját tanulmányozni, hogy a példaprogramot működésre bírjuk (pl. a PyPDF esetében). A rövid Python-bevezető után rögtön egy esettanulmányon lehet tanulmányozni a nyelv alkalmazását: Mit evett az Egyesült Államok lakossága? A cél a „feltáró adatelemzés" *(Exploratory Data Analysis)* módszerének ismertetése, amellyel behatóbban ismerhetjük meg adatainkat, jelen esetben a 18. század végétől a 20. század elejéig tartó időszakban kiadott szakácskönyveket tartalmazó gyűjteményből vett mintákat. Az adatokat egy CSV-fájlból a Pandas nevű kódkönyvtár által definiált speciális adatszerkezetbe, úgynevezett „adatkeretbe" *(Data Frame)*[2] másoljuk, ami az adatok tárolásán túl számos metódust is biztosít a bennfoglalt adatok elemzésére, például megjeleníthetjük az adatsor elejét vagy végét, egy adott oszlopban szereplő egyedi értékek listáját akár előfordulásuk mennyiségével együtt, de akár grafikonokat is létrehozhatunk. A feltáró adatelemzés során ezen módszerek segítségével meg tudjuk becsülni, hogy két változó között van-e összefüggés, és – ha van – milyen. Például hogyan változott a paradicsom konyhai szerepe 1810 és 1930 között, vagy melyek azok az alapanyagok, amelyek kezdetben főként egy adott népcsoport receptjeiben fordultak elő, majd később általános népszerűségre tettek szert.

A második fejezet címe *Strukturált adatok feldolgozása és kezelése.* Az adatok forrása valamilyen fájl – a könyv csupán említés szintjén foglalkozik más adatforrásokkal, például adatbázisokkal, példákat ezekre nem ad. Sorra veszik a bölcsészet által leginkább használt fájltípusokat: az egyszerű szövegállományt, a CSV-t (ebben az oszlopokban elrendezett adatsor mezőértékeit valamilyen határolójel, legtöbbször vessző választja el), a PDF-et,[3] a JSON-t (JavaScript-objektumjelölés), az XML-t általában és külön a TEI-t és a HTML-t. A szigorúan vett fájlfeldolgozás mellett megtudhatjuk, hogyan szűrjük az információkat, hogyan csomagoljunk ki tömörített állományokat, illetve hogyan lehet fájlt az internetről letölteni Pythonban. A fejezet bölcsészeti kérdése a drámaszövegek szereplői között fennálló interakciós hálózat kinyerése és az ebből adódó tanulságok levonása. A szakirodalomban tájékozott magyar olvasóknak Barabási Albert-László munkásságának köszönhetően ismerősek lesznek az itt olvasható

[2] Az adatkeret neve és fogalma megtalálható más adatelemző eszközökben is, például az R nyelvben, vagy az Apache Sparkban. Hogy melyikben bukkant fel először arra nem találtam adatot. Az utóbbi években ezek nagyon sokban hatottak egymásra, s ennek az egyik előnye, hogy a Pandas ismeretében legalább olvasni lehet a másik két eszközön írt adatelemző programokat.

[3] Érdemes itt is megjegyezni, hogy a PyPDF kurrens változatának (2.x) metódusnevei megváltoztak a könyvben található 1.x változathoz képes, így PDF.PdfFileReader() helyett PDF.PdfReader()-t, pdf.getPage(1) helyett pdf.pages[1]-t stb. kell alkalmazni. A kódkönyvtár alkotói szerencsére készítettek egy részletes migrációs útmutatót: PyPDF2. Migration Guide: 1.x to 2.x., hozzáférés: 2023.02.19., `https://pypdf2.readthedocs.io/en/latest/user/migration-1-to-2.html`

 DOI 10.31400/dh-hun.2022.6.5580

hálózatelméleti metrikák, és világos lesz, hogy a fejezet (miként a többi is) épphogy csak elindítja az olvasót a téma felé.

A harmadik fejezet témája a szövegjellemzők megismerése a vektortérmodell használatával. A vektortérmodell a keresőgépek révén vált széles körűen ismertté, de teret nyert más szövegfeldolgozási feladatokban is.[4] A lényege, hogy egy szövegkorpuszt egy nagy táblázatként tárolunk. Ennek a sorai az egyes dokumentumokat reprezentálják, oszlopai a korpuszban található szavakat, az egyes mezőértékek pedig adott kifejezés adott dokumentumban található előfordulási gyakoriságát tartalmazzák. A szótár előállítása általában többlépcsős normalizálási folyamaton keresztül történik: a szöveget – miután kikerültek belőle az írásjelek – szavakra bontják (tokenizálás), a szavakat kisbetűsítik, de akár szótövesíthetik *(stemming),* vagy lemmatizálhatják *(lemmatization)* is, ekkor a ragozott alak helyett valamilyen egységes szótővel vagy a szótári alakkal számolnak. Fontos tudni, hogy a modellben a szavak sorrendjét nem reprezentálja semmi, ez az információ eltűnik. Mivel a modell egy nagy, számokból álló táblázat, kiválóan lehet rajta mátrixműveleteket végezni. Ennek eredményeként ki tudjuk számolni a két vagy több dokumentum közötti távolságot, vagyis azt, hogy ezek, szókészletüket tekintve mennyire hasonlítanak egymásra. A könyv több lehetséges módszert is bemutat mind a szöveg feldolgozására, mind a távolságok kiszámítására. A módszert az automatikus műfajazonosítás feladatán keresztül ismertetik, ehhez a Paul Fièvre által gondozott, klasszikus 17. századi francia drámákat tartalmazó *Théâtre Classique* TEI-gyűjteményt használják fel.[5] A korpuszban található darabok háromféle korabeli műfajbesorolást tartalmaznak: tragédia, komédia és tragikomédia. A dokumentumtávolság elemzésével kiszűrhetjük a műfaj tipikus reprezentánsaitól nagyban eltérő darabokat. Kiviláglik továbbá, hogy összességében a tragikomédiák a két másik műfajhoz képest nem középen állnak, hanem sokkal inkább közelebb a tragédiákhoz, vagyis ezek lényegében olyan tragédiák, amelyekhez – a drámaiság élét tompítandó – némi humoros csavart adott a szerző. A fejezet továbbá tartalmaz egy kiegészítést a NumPy kódkönyvtár vektor- és mátrixműveleteiről.

Az első rész utolsó, negyedik fejezete a táblázatos adatok feldolgozásáról szól. Ennek a fő eszköze a Pandas kódkönyvtár, ezen belül is a már említett adatkeret nevű adatszerkezet. A Python kiváló lehetőségeket nyújt az adatok szűrésére, szelektálására, csoportosítására, módosítására. Az esettanulmány témája az Egyesült Államok-beli névadási szokások változásainak elemzése. Alapötlete, hogy évről évre vizsgáljuk meg, melyek voltak a legnépszerűbb keresztnevek, és vessük össze az egymásra következő időszakokat. Ennek alapján nemcsak azt állapíthatjuk meg, hogy melyek azok a nevek, amelyek újonnan lettek népszerűek, de azt is, hogy milyen gyakran váltak népszerűvé új nevek. A szerzők megvizsgálták azt is, vajon igaz-e az a feltételezés, hogy az utóbbi időben egyre többször találkozni *n*-re végződő nevekkel, illetve azt is, hogy

[4] Mártonfi Attila hívta fel a figyelmemet Jékel Pálnak és Papp Ferencnek a vektortérmodell felfutását több mint húsz évvel megelőző alkalmazását felvonultató kutatására, ami (eddig) meglehetősen visszhangtalan maradt a szakirodalomban (Jékel Pál és Papp Ferenc, *Ady Endre összes költői műveinek fonémastatisztikája* (Budapest: Akadémiai Kiadó, 1974). Kétségtelen – értékel Mártonfi –, itt a vektorteret nem a lexémák, hanem a fonémák feszítik ki (a korabeli gépi kapacitás nem is tett volna mást lehetővé); egyéb tekintetben azonban nagyon erős a hasonlóság.

[5] hozzáférés: 2023.03.17., `https://www.theatre-classique.fr/`

mi a helyzet a lányoknak és fiúknak egyaránt adott, uniszex nevekkel. A szerzők felhívják a figyelmet a forráskritika fontosságára: az alapforrás megbízhatóságát, reprezentativitását ugyanis számos ok befolyásolta az idők során (vagyis az adatok nem mindig tükrözik a teljes populáció névadási szokásait), ennek következtében bizonyos időszakokra a következtetések is szükségszerűen kevésbé szignifikánsak. Mindezeken felül e fejezetben ismerkedhetünk meg a vonaldiagram kiugró csúcsait lelapító, és így a trendeket tisztábban ábrázoló mozgóátlag-számítással, valamint néhány vizualizációs trükkel.

A könyv második része az elsőben megtanult fogalmakra épít, és míg az elsőben igyekeztek a kódokban előforduló újdonságokat a szövegben megmagyarázni, a második részben ez már ritkábban fordul elő, ott is inkább csak a főbb pontoknál (pl. leírva, hogy egy metódus mire jó, de nem kitérve egyes paramétereire). Ezért a második rész egyrészt több figyelmet igényel az olvasótól, másrészt – különösen az Allen Riddell által írt fejezetek – kicsit több odafigyelést kaphattak volna a szerzőktől is, lévén ezekben olyan kisebb-nagyobb gondolati ugrások, elmaradt magyarázatok akadnak, amelyek megnehezíthetik a tanultak alkalmazását másféle forrásokon és másféle kutatási kérdések esetében.

Az ötödik fejezet – címéhez *(A statisztika alapjai: ki olvas regényeket?)* híven – a statisztikai mérésekkel foglalkozik, s ehhez illusztrációképp az egyesült államokbeli általános társadalmi felmérést használja. A statisztikai mérőszám *(statistic)* megfigyelésekből álló adatgyűjtemény függvénye, ilyen például az összeg, az átlag, a minimum érték vagy a szórás. A fejezet sorra veszi a leíró vagy összegző statisztika fontosabb fogalmait és mérőszámait, valamint bemutatja a mennyiségi és a kategorikus változókra vonatkozó műveleteket. Itt találkozunk először matematikai modellel, amelynek a segítségével létre tudunk hozni a megfigyelt adatsorra hasonlító mesterségesen generált adatsort, a modell paramétereinek változtatásával pedig szimulálni tudunk eltérő kimeneteket is. A könyvben a háztartások jövedelemeloszlásának tanulmányozására alkalmazzák, a gammaeloszlás modelljével. Ezek a matematikai modellek jól ismert tulajdonságokkal bírnak, és a segítségükkel jobban le lehet írni az általános trendeket, vagy megkülönböztetni azokat az eseteket, amelyeknek az általános trendtől való eltérését magyarázhatja a véletlen, azoktól, amelyeknél az eltérés szignifikáns, vagyis a véletlennel nem magyarázható. A könyvben említett modellek általában valamilyen eloszlást is leírnak. A könyv kétféle eljárást ismertet annak megállapítására, hogy melyik modell illeszkedik az adatokra. Az elsőben az adatokhoz illeszkedő modellből különféle paraméterbeállításokkal ábrákat generálunk, és a szemre leginkább illeszkedőt választjuk, a másodikban pedig a Python (főként a scikit-learn kódkönyvtár által biztosított) gépi tanulási eszköztárát használjuk a modell paramétereinek kiválasztására. A fejezet utolsó része a mennyiségi és kategorikus változók közötti kapcsolatokkal foglalkozik, vagyis azzal, hogy két változó között van-e valamilyen korreláció. Végül megtudjuk, hogy a regényolvasási szokások és az USA régiói között mért 0,0069-es kapcsolat nem jelez olyan erős viszonyt, amelyet ne lehetne betudni a puszta véletlennek.

A hatodik fejezet a valószínűségszámításba, méghozzá annak is a Bayes-féle következtetésekről szóló ágába enged betekintést. A Bayes-szabály egy esemény bekövetkezésének valószínűségét számolja ki abban az esetben, ha egy ezzel összefüggő másik esemény már bekövetkezett. Például tudjuk, hogy ha véletlenszerűen kiválasztunk egy

 DOI 10.31400/dh-hun.2022.6.5580

1960 és 2010 között megjelent irodalmi művet, akkor 0,001% annak az esélye, hogy a mű szerzője Thomas Pynchon. Tegyük fel, hogy létezik olyan stilometriai alkalmazás, amely az általa írt műveket 90%-os valószínűséggel tulajdonítja helyesen Pynchonnek, és ez egy általunk vizsgált művet az írónak tulajdonít. Mennyi a valószínűsége, hogy a művet valóban ő írta? Itt a két esemény a következő: a művet Pynchon írta, és az alkalmazás neki tulajdonítja. A Bayes-szabály értelmében a válasz – elsőre talán meglepő módon – valamivel kisebb, mint 0,1%. A fejezet egy ehhez hasonló kérdést vizsgál: Ki írta *A föderalista (The Federalist Papers)*[6] vitatott cikkeit, Alexander Hamilton vagy James Madison? A szerzőség megállapításához a nem vitatott szerzőségű cikkekben szereplő kötőszavak (pl. *upon, by*) eloszlását vizsgálták (a stilometria egyik felismerése, hogy az ilyen, szinte öntudatlanul használt funkciószavak *[function words]* inkább jellemzők egy szerzőre, mint a főnevek, a melléknevek vagy az igék), majd a megfigyelt eloszláshoz egy arra illeszkedő matematikai modellt illesztettek (a negatív binomiális eloszlást). A modell előnye, hogy számszerűen össze lehet vetni az egyes szerzők szóeloszlási valószínűségeit, majd az így kapott értéket behelyettesíteni a Bayes-szabályba.

A hetedik fejezet címe: *Elbeszélés térképekkel.* Az amerikai harcmezővédelmi program nyilvántartást vezet az amerikai polgárháború jelentősebb csatáiról. Az egyes csatákról rögzítik, többek között, az időpontot, a helyszínt, a veszteségek számát, a győztes oldal nevét. A településnevek földrajzi koordinátáit geokódoló szolgáltatás segítségével nyerik ki.[7] A koordináták térképre vetítéséhez azonban kell még két összetevő: egy alaptérkép, amelynek a legelterjedtebb fájlformátuma az úgynevezett *shapefile,* valamint az adott terület sajátosságait leginkább tükröző vetület kiválasztása. A Python-térképek, amiként a Pandas grafikonjai is a Matplotlib kódkönyvtárral működnek együtt, ennek az előnye, hogy a grafikai ábrázolások kezelése itt is hasonló. A fejezet példája esetében ezt úgy aknázzák ki, hogy a polgárháború minden egyes hónapjáról készül egy – színkódokkal a csata eredményét, a pont nagyságával pedig a veszteséget jelölő – térkép, és ezek egyetlen nagyobb képbe vannak rendezve, összességében világosan megrajzolva a polgárháború hadi eseményeinek főbb tendenciáit.

A nyolcadik fejezet visszatér a stilometria és a szerzőazonosítás témájához, de egy másik módszert, a klaszterálást állítva fókuszba, és amerikaiak helyett középkori szerzőkkel: Bingeni Szent Hildegárddal, utolsó titkárával, Guibert de Gembloux-val és Clairvaux-i Szent Bernáttal. A fejezet ismerteti a „Burrows-féle Delta" nevű eljárást, amely egy gépi tanuló algoritmus, s az első, tanulási fázisban ismert szerzők műveit elemezve nyeri ki a szerzőre jellemző ismérveket, majd a második, előrejelző fázisban azt nézi meg, hogy a nem ismert szerzőjű művek ismérvei melyik szerző ismérveire hasonlítanak leginkább. Az ismérveket egy normalizált szám jelöli, a statisztikából vett *z* szám, amely azt mutatja meg, hogy az adott érték hány szórásnyira van az átlagértéktől. A szerzőazonosítás a tapasztalatok szerint akkor működik jól, ha a vizsgált szövegek hossza nagyobb egy bizonyos mennyiségnél (a könyvben idézett

[6] Magyarul: Alexander Hamilton, James Madison, és John Jay, *A föderalista: Értekezések az amerikai alkotmányról,* ford. Balabán Péter, jegyz. Magyarics Tamás (Budapest: Európa Könyvkiadó, 1998).

[7] A 232. oldal 6. lábjegyzete szerint az a Python-szkript, amely kiolvassa a forrásból a településnevet és lekérdezi a geokódoló szolgáltatást a könyv weboldalán elérhető, ez azonban sajnos nem így van. Sem ott, sem a könyv kódrepozitóriumaiban (GitLab és Zenodo) nem érhető el.

kutatás szerint az alsó határ 6500 szó), valamint a vizsgált szövegek nagyjából egyforma hosszúak. Jelen esetben a szövegeket 10000 lemmás darabokra bontották, részben azért is, hogy megnézzék, az egyes szerzők különböző szövegei valóban ugyanabba a csoportba kerülnek-e. A vizsgálandó szavak körének kijelölésére lehet gyakoriságon alapuló automatikus módszert választani a Python gépi tanulási módszereket tartalmazó scikit-learn kódkönyvtára eszköztárából, de akár saját szótárral is lehet dolgozni. A vektortérmodellt ezúttal nem a korábban ismertetett kézi módon, hanem a scikittel valósították meg.

A Burrows-féle Delta az úgynevezett felügyelt tanítás családjába tartozik, tipikusan címkét rendel a vizsgált egyedekhez attól függően, hogy a tanulás során az ismert egyedeknek milyen címkéjük volt (jelen esetben a címke a szerző neve). Létezik azonban egy másik algoritmuscsalád, a felügyelet nélküli tanulás, amely nem használ efféle címkéket. A könyv két ilyen eljárást ismertet, a hierarchikus egyesítő klaszterezést *(Hierarchical Agglomerative Clustering)* és a főkomponens-elemzést *(Principal Component Analysis, PCA)* – közös tulajdonságuk, hogy egyik sem kínál osztálynevet, és mindkettő alkalmas grafikai megjelenítésre. Az első eredménye bináris faszerkezet, ahol az ág egyszerre legfeljebb kétfelé ágazhat el, és ahol a kutató az ágrajz vizsgálata után döntheti el, hogy hány csoportot képez. Szemben a címkézéssel, amely pusztán szerzők szerint csoportosítja a szövegeket, itt az azonos szerzőhöz tartozó művek alcsoportjai is detektálhatók. A PCA úgynevezett dimenziócsökkentési eljárás. Jelen esetben a szövegek dimenziószáma a szótár nagyságával, 65-tel egyenlő (a szótár csak a funkciószavakat tartalmazza). Az egyes dimenziók azonban sokszor mutatnak valamilyen negatív vagy pozitív korrelációt egymással, vagyis az egyik ismeretében a másik értéke előrejelezhető. A főkomponensek alkotásához az algoritmus elemzi ezeket a korrelációkat, és sokváltozós egyenletet állít fel, amelyben az egyes dimenziók súlyként szerepelnek. Ezek a főkoponensek, számuk megegyezik a dimenziók számával, de magyarázó képességük erősen eltér. A gyakorlatban csak néhány főkomponens képes reprezentálni a sokaság jelentős csoportjait, a többi inkább csak finomítja a képet. A könyvben két ilyen főkomponenssel számolnak, mindkettőben külön-külön súllyal szerepelnek az egyes dimenziók; együttes magyarázó erejük közel 60%. A két dimenziót már meg lehet jeleníteni grafikonon, ezen kiválóan látszik, hogy a három szerző mennyire tér el egymástól.

Az utolsó, kilencedik fejezet az Egyesült Államok Legfelsőbb Bíróságán 100 éves időtávban született döntésekhez csatolt különvélemények témamodelljét vizsgálja. Az alkalmazott matematikai modell a Dirichlet-féle rejtett elhelyezkedés *(latent Dirichlet allocation)*. A modell alapja az a feltételezés, hogy a trendszerűen egymás szomszédságában vagy egyazon dokumentumban előforduló szavak csoportja egyúttal egy témát is kirajzol. A metódushoz, amely a scikit-learn kódkönyvtár része, meg kell adnunk, hány témával szeretnénk számolni (láttuk, hogy a hierarchikus klaszterezésnél erről utólag dönthettünk). A számítás eredményeképpen megkapjuk az egyes témákhoz tartozó szavak listáját a befoglaló dokumentumok számával, illetve az egyes dokumentumokhoz tartozó témákat és azok súlyát, vagyis hogy azok mennyire relevánsak az adott dokumentumra nézve. Az egyes témáknak nincs nevük, csak azonosítójuk, azokat a kutatónak kell elneveznie vagy legalább a jelentését értelmeznie, és ez nem minden esetben egyszerű. Fontos, hogy a szavak eloszlása a témák között

 DOI 10.31400/dh-hun.2022.6.5580

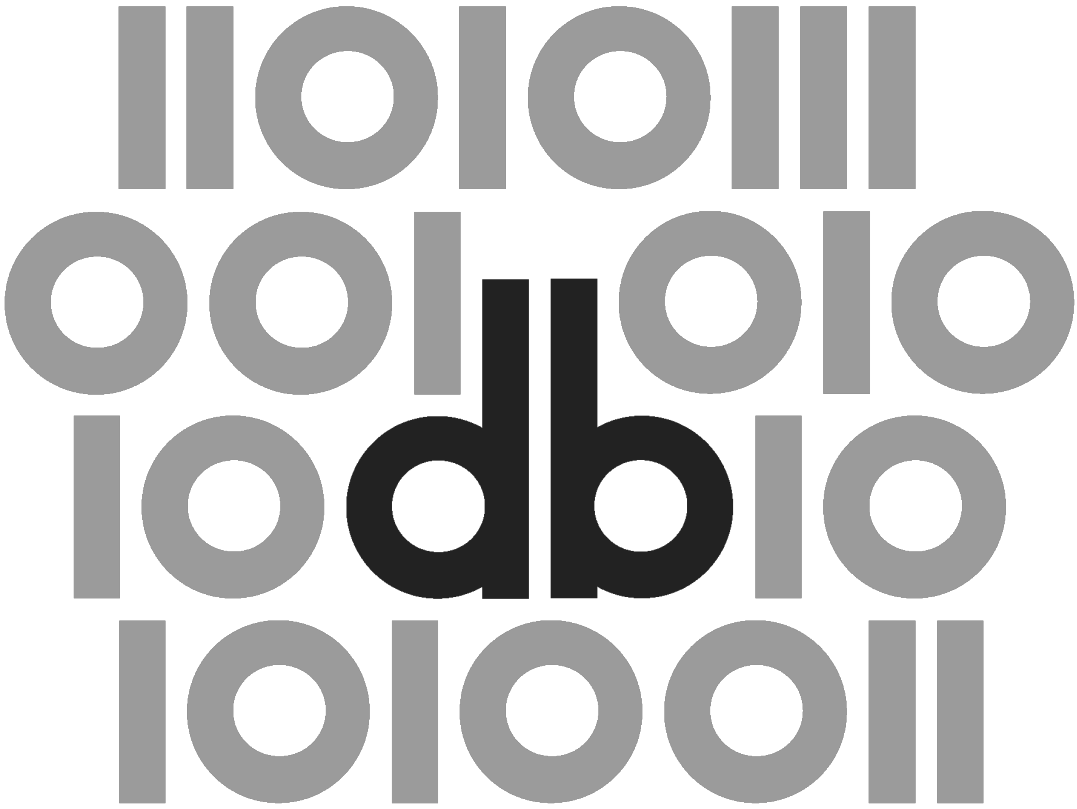
nem kizárólagos, egy-egy szó több különböző témának is része lehet, ugyanakkor érdekes módon a többjelentésű szavak különféle jelentései általában más-más témába kerülnek. Részletesen csupán a leginkább használatos témamodell leírását olvashatjuk, de a szerzők utalnak arra, hogy vannak egyéb közelítések, például olyan, amely a kronológiát is figyelembe veszi. A magyar és más agglutináló nyelvek szempontjából érdekes az a kutatás, amely a szótövezés és a stopszavak alkalmazásának hatását vizsgálja.

A könyv végén bibliográfia és jó tanácsok sorakoznak a tudományos célú informatika „elég jó gyakorlatainak" tárgyában (az adatkezelésről, a szoftverről, az együttműködésről, a munkaszervezésről, a változáskövetésről és a kéziratokról).

A könyvben a Python olyan segédeszköz, amelyet kutatási (adatelemzési) céllal használnak, ennek megfelelően a nyelvnek csak azokat a tulajdonságait érintik, amelyek ehhez a feladatkörhöz tartoznak, de még ezek közül is csak a legfontosabb összetevőket. Nincs szó például osztályokról, függvények és változók névadásáról, tesztelésről vagy akárcsak arról sem, hogy egy Python-szkriptnek mit kell tartalmaznia ahhoz, hogy le tudjuk futtatni. Ezekről a könyvben ajánlott forrásokat kell tanulmányozni. Hasznos lett volna a kutatásiszoftver-fejlesztés sajátosságainak tárgyalása is, ami a kutatásiadat-kezeléssel párhuzamoson fejlődő újabb szakterület, de erre csak az epilógusban van utalás.[8] A könyvre épülő kurzusban véleményem szerint ezeknek a témáknak a könyvnél kifejtettebb módon kellene szerepelnie, mivel e két további összetevő garantálja, hogy az általunk írott kód néhány év múlva is futtatható és érthető maradjon.

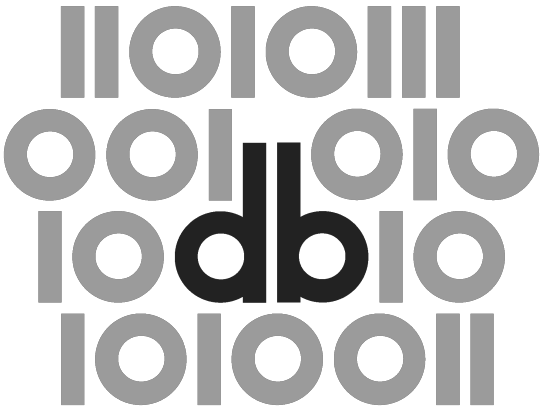
Nagy erénye a könyvnek, hogy nem akar mindentudó lenni, és hogy a műfaj legjobb hagyományai szerint a vizsgált témákról további irodalmat ajánl. Bár a bölcsészeti adatelemzés számos témáját átfogja, de – és ezt láthatjuk akár a szakterület érettségének a bizonyítékaként is – legalább ennyi minden nem fért bele e vállalkozásba. Hogy csak néhányat említsünk: adattisztítás és adatminőség; nevek, fogalmak kinyerése és a hozzájuk tartozó entitások azonosítása; nyelvészeti (pl. szófaj-, mondat-) elemzés; hipotézistesztek vagy a nem szöveges adatok (zene, kép, mozgókép, térbeli objektumok) elemzése. Nagyon remélem, hogy ezen és más hiányzó témák bemutatása céljából vagy a szerzők rugaszkodnak neki egy második kötetnek, vagy a könyvtől ihletett olvasók készítenek hasonlókat.

[8] Wilson et al., „Good Enough Practices in Scientific Computing," *Plos Computational Biology,* 2017. június 22., `https://doi.org/10.1371/journal.pcbi.1005510`. Emellett érdemes elolvasni az Európai Kutatási Infrastruktúra-szoftver Mérnökök Hálozata (EURISE) műszaki referenciadokumentumát, hozzáférés: 2023.03.17, `https://technical-reference.readthedocs.io/en/latest/`, illetve a brit Software Sustainability Institute útmutatóit, hozzáférés: 2023.03.17, `https://www.software.ac.uk/resources/get-speed`. Tananyagként pedig a következő mű használható: Damien Irving et al., *Research Software Engineering with Python: Building Software that Makes Research Possible,* hozzáférés: 2023.03.07, `https://merely-useful.tech/py-rse/index.html`.



Digitális Bölcsészet
2022., hatodik szám

<DIGITÁLIS BÖLCSÉSZET>



6 (2022)

ISSN 2630-9696
DOI 10.31400/dh-hun.2022.6

Kiadja a Bakonyi Géza Alapítvány és az ELTE BTK Régi Magyar Irodalom Tanszéke (1088 Budapest, Múzeum krt. 4/A).
Felelős kiadó az ELTE BTK Régi Magyar Irodalom Tanszék vezetője.

Megjelenik az Open Journal Systems (OJS) v. 3. platformon, melynek működtetését az ELTE Egyetemi Könyvtár- és Levéltár biztosítja.

Honlap: `http://ojs.elte.hu/digitalisbolcseszet`
Email cím: dbfolyoirat@gmail.com

Olvasószerkesztő: Bucsics Katalin
Tördelés: Hegedüs Béla
Grafika: Hegyi Gábor

<KRITIKA>

Németh Márton 0000-0003-1864-8107

Blinken OSA Archivum

nemethm@ceu.edu

Tófalvy Tamás szerk. *A magyar internet történetei*. Budapest: Typotex, 2021. 192 oldal. ISBN 9789634931485

Úttörő jellegű kezdeményezés a magyar internet múltjának egyes szeleteit taglaló kötet összeállítása. A műben szereplő, tematikus körökhöz illesztett egyes tanulmányok a „magyar internet" különféle kutatási szempontjait villantják fel. Nemzetközi távlatból nézve a digitális média tanulmányozásának mintegy három évtizedes múltja van (a kötet szerkesztője is utal erre előszavában). Az internet történetének különféle dimenzióival foglalkozó diskurzusok azonban a 2010-es években kezdtek intézményesülni, amelyből voltaképpen néhány év alatt új történeti segédtudomány formálódott. Az évtized második felében lett saját tudományos szakfolyóirata *Internet Histories* címmel (így, többesszámban, mint ahogy a most ismertetett kötet címében is szerepel), s megjelent néhány, a most ismertetett tanulmánykötethez hasonló, az internet történetével különféle nézőpontokból foglalkozó tanulmánykötet, melyek összeállításában a dániai aarhusi egyetem professzora, Niels Brügger játszott vezető szerepet. Az utóbbi években a webarchívumok tudományos hasznosítása kapcsán formálódott WARCNET hálózat különféle rendezvényein többször feltették nekem a kérdést, hogy jelentek-e már meg cikkek, tanulmányok a magyar internet múltjához kötődő különféle témákról, bekerült-e már ez a terület az itthoni tudományos vérkeringésbe. Ezért is vettem nagy elégedettséggel kezembe a szóban forgó tanulmánykötetet. A kötet, szerzői sokszínű témaválasztásának köszönhetően, méltóképpen illeszkedik az internet történeteiről szóló nemzetközi diskurzusba. Ennek kiteljesítéséhez, véleményem szerint, nagy szükség lenne a kötet angol nyelvű megjelentetésére is, bár jelen helyzetben annak is feltétlen örülnünk kell, hogy egy ilyen sokszínű szerzőgárdával bíró, igényesen szerkesztett kötet napvilágot látott e témakörben.

Mint Tófalvy Tamás, a kötet szerkesztője már a bevezetőben rámutat, az internet történetével való foglalkozás nem azoknak a kutatóknak az ideális kutatási terepe, akik nagy, átfogó narratívák felvázolásával mutatnak be egyes történeti témákat. Maga az internet természeténél fogva megfoghatatlan, komplex hálózat, melynek egyes szeleteit önkényesen kiragadhatjuk archiválás, illetve tudományos vizsgálódások céljából, de nagy narratív elbeszélések tárgyául ezek is bajosan szolgálhatnak. Egész egyszerűen nincs forrásanyag ehhez a vállalkozáshoz. Amik rendelkezésünkre állnak, azok a világháló áttekinthetetlen gazdagságából önkéntes kutatási szempontjaink szerint kiragadott források, melyeket különféle módszertani elemek és kutatási koncepciók mentén tanulmányozhatunk. Nem hiába utal Tófalvy ennek kapcsán a mikrotörténet-írásra, azokra a történeti irányzatokra, melyek alulnézetből vizsgálnak történeti jelenségeket, folyamatokat s ebből kiindulva vázolják fel következtetéseiket az általános narratívaalkotás igénye nélkül. Ebben a könyvismertetésben röviden át-

tekintem azokat a megközelítésmódokat, témaköröket, melyeket a szerzők az internet múltjának tanulmányozása kapcsán vizsgálatra méltónak tartottak.

Az internet történetei, mint a kötetből is kiderül, rengetegféle megközelítésmódból vázolhatók fel. Társadalmi megközelítésmódban tehetünk kísérletet egyes társadalmi jelenségek, társadalmi csoportok webes lenyomatának tanulmányozására (mint ahogy a kötet harmadik tanulmánya teszi a nők megjelenésével az infokommunikációs iparágakban), illetve kiemelhetünk akár egyes hibrid módon megjelent szövegeket is, mint a digitális átmenetiség kordokumentumait, ahogyan a kötet első tanulmánya sikerrel vállalkozik erre. A második tanulmány, Szakadát István révén, ehhez társulva áttekinti azokat a hagyományos és digitális formában megjelenő szövegtípusokat, s azok értelmezési tartományát, melyeket eszközként használhatunk, egyrészt a webes múlt egyes szeleteinek tanulmányozásához, másrészt a hagyományos papíralapú dokumentumtípusokhoz kapcsolódásul. Mintegy bevezetésképpen, egyfajta tágabb kontextust kínálva a különféle forrástípusok terén, bemutatja, hogy a tartalmilag megegyező szövegek különféle formákba, keretekbe ágyazva hogyan válnak egyre intelligensebbekké a technikai fejlődés eredményeként. A fejezet szerzője találóan rámutat arra is, hogy akárhány keretet s hozzá tartozó technikai hátteret vázolunk fel, a dinamikus világnak azok mindig egyfajta merev leképezései lesznek, tehát illúzió, hogy mindenre kiterjedő, végleges és érvényes általános magyarázati sémát kaphatnánk általuk.

Fontos állítása a Szijjártó Zsolt és Németh Szilvia által jegyzett tanulmánynak, hogy miközben a digitális technika gyors elterjedése új és gyors gondolatműveletek elvégzésének ad teret, érdemes számot vetni azzal is, hogy mit vesztünk el az automatizálással. A digitalizálás kora előtti adatgyűjtési, rendszerezési technikák olyan összefüggésekre utaltak (például egy hatalmi struktúra meghatározott működési logikájának kibontása), melyek lényegesek lennének továbbra is, ám a digitális korban háttérbe szorultak. Fontos kérdés, hogy a digitális eszköztár, az online tér használata miként alakítja át a társadalomtudományos megismerés kereteit. Mint ahogy arra a szerzők is utalnak, további összehasonlító kutatások szükségesek ennek feltárására.

Andok Mónika a nők digitális bevonódásáról ad körképet, a digitális megosztottság három szintjét vizsgálva (hozzáférés, használat, használat minősége). Érdekes megállapítás, hogy miközben a hozzáférésben a nők és férfiak között minimális a különbség, a használatban már eltérő minőségű a digitális jelenlét. Hiába számoljuk fel tehát a különbségeket a két nem között a hozzáférés terén, ettől még számos további kihívással szembesülünk. Abban semmi meglepő nincs, hogy mind a nők, mind a férfiak a saját értékeik, preferenciáik szerint használják a világhálót. Lényeges megállapítás, és a pozitív jövő irányába hat, hogy amint a javarészt nők által végzett feladatoknak erősödik a digitális, illetve online térben való elvégezhetősége (például a közigazgatásban az online ügyintézés erősödése során, amire a COVID-járvány is ráerősített a könyv megjelenése óta), úgy számukra is egyre hangsúlyosabbá válik az online tér használata. A szerző arra is rámutat, hogy komolyan foglalkozni kell a nőket érintő online zaklatások visszaszorításával, hiszen ez is erős gátja az online szolgáltatások szélesebb körű terjedésének. Sajnos, mint megállapítja, a nők esetében a technológia használata kapcsán továbbra is jelentős a kisebbrendűségi érzés. Tudatosítani kellene a tanulmány szerzője szerint, hogy a nőiség és a technológia nem egymásnak ellent-

 DOI 10.31400/dh-hun.2022.6.4985

mondó jelenségek, nagyon fontos lenne a készségek és képességek elsajátításának akadályaként tapasztalható, a nőket érintő specifikus gátak lebontásának elősegítése. Ennek érdekében jobban be kellene vonni a nőket az online tartalom előállításába is (például a *Wikipedia* szerkesztőinek csupán húsz százaléka nő).

A második tematikus blokk tanulmányait összekötő szempont az archiválás igényének különféle megközelítései. Golden Dániel a világhálóról eltűnt Internet Expo magyar pavilon kapcsán egy olyan forrásról értekezik, amelyről, miután eltűnt az élő webről, már csak másodlagos források segítségével kaphatunk képet. A szerző persze nem adta fel a reményt, hogy egyszer előkerül majd egy archivált példány esetleg egy elfekvő régi számítógépről, mint ahogyan az Internetto hírlevél anyagával is történt. A teljes magyar kulturális örökség webes publikációja s annak archivált rögzítése a jövő nemzedékek számára külön-külön is illúzió csupán. Az internetarchiválásnak viszont jelentős szerepe van abban is, hogy ma már hozzáférhetetlen internetes forrásokat térképezzen fel, hogy legalább képet kaphassunk arról, melyek voltak a leglényegesebb elemei elveszett webes múltunk darabjainak.

Mester Tibor a közösségi archívumokat vizsgálja tanulmányában. Közösségi archívumoknak az olyan hálózati alapú médiagyűjteményeket nevezi, melyek elsődleges motivációja a csoporton belüli megosztás és használatba vétel, emellett az archívum kialakítása és működtetése során az önkéntesek részvételére és a vélemények szabad kifejtésére is kisebb-nagyobb mértékben támaszkodnak. A belépési küszöb általában alacsony, az összegyűjtött javakhoz jellemzően bárki hozzátehet és részesedhet is belőlük. A tartalom gyűjtését és feldolgozását főleg a közösséghez tartozás motiválja. Egyszerűen, sallangmentesen működnek, a tartalomszervező eszközök használatában is ezt az alapelvet tartják szem előtt. Nincs mögöttük viszont fizikai gyűjtemény, működtetésük során az önkéntes archivisták szakértelmére vannak utalva. A globális online platformjokon futó szolgáltatások könnyen elérhetetlenné válhatnak. Külön kihívást jelent az archivált anyag szerzői jogi státusza, illetve manipulációtól mentes, tiszta státuszának garantálása (bár a szerző utal arra, hogy az ezzel kapcsolatos nemzetközi botrányokhoz hasonlóak nálunk szerencsére még nem jelentkeztek). Az eredet, az eredetiség háttérbe szorul, a hozzáférés garantálása az elsődleges. Ennek kapcsán is ki vannak szolgáltatva azonban a hálózati kultúra jellegzetességeinek. A hosszútávú megőrzés feltételei nem tisztázottak, rendszereik sérülékenyek, ki vannak téve a kisajátítás és felforgatás veszélyeinek is. Az archivált anyagok bemutatási és újrahasznosítási lehetőségei rendezetlenek, ami komoly kockázatot jelent. A hagyományos archívumok nézőpontjából radikálisan új kezdeményezést jelentett a megjelenésük. Annál is inkább, mert a közösségi archívumok jellemzően a hagyományos intézményrendszerrel szemben határozzák meg önmagukat, sokszor olyan dolgokat is őriznek, melyeket a hagyományos intézmények nem tartanak erre alkalmasnak. Igen lényegesnek látom azonban a tanulmány kapcsán megjegyezni, amire a szerző is utal, hogy a közösségi archívumok megjelenésével olyan kollaboratív archiválási szemléletmód nyert teret, mely lassanként beszivárgott már a hagyományos intézmények világába is. Erősíti azok társadalmi kapcsolódási pontjait, új szempontokkal gazdagítja az archiválási módszereket a közösségalapúság; a részvételen alapuló szempontok, melyek a közösségi archívumok működésének alapjait adják, fontos tényezői lehetnek a hagyományos archívumok tevékenységei részleges megújításának is.

Moldován István nagy ívű áttekintésében szót ejt a könyvtárak gépesítésének kezdeteiről, az első adatbázisok létrejöttéről, s az azóta bejárt fejlődési útról. Ennek jelentős állomása az integrált könyvtári rendszerek megjelenése, mely a legfontosabb munkafolyamatok digitalizálását tette lehetővé. Számos példát mutat be az első magyar könyvtári honlapokról, a világháló megjelenése kapcsán lezajlott ismeretterjesztő tevékenységekről, kampányokról. A módszertani támogatás, a tapasztalatok cseréje, az akadémiai hálózaton dolgozó szakemberek kapcsolattartásának máig működő fő fórumaként jelenik meg a Networkshop konferenciasorozat. Igen fontos feladat az eredeti helyükről jellemzően elég gyorsan eltűnő honlapok, e-folyóiratok e-könyvek megőrzése. A MEK keretében eleinte webes címek gyűjtését is megkísérelték, de ez aztán a mennyiségi korlátok s egyéb intézményközi szervezési okok miatt sem tudott folytatódni. A tanulmány figyelemre méltó fejlődési ívet vázol fel, ahogy elkezdődött a gyűjtés a könyvekkel a Magyar Elektronikus Könyvtár számára, majd amikor az Elektronikus Periodika Archívum felvállalta az elektronikus folyóiratok címeinek nyilvántartását, illetve a teljes szövegű gyűjtést, végül a web archiválását. Utóbbit sajnos csak 2017-ben, az első európai közgyűjteményi projektek megjelenése után, 15–20 év késéssel tudott elkezdődni. Ezt követően azonban már viszonylag hamar, a projekt kísérleti szakaszának végére megteremtődtek a szabályozási keretek is. Törvénymódosítás, illetve kormányrendelet tette ezt is nemzeti könyvtári alapszolgáltatássá, illetve vázolta fel az alapvető kereteket. Külön feladatot jelent a webarchiválás kapcsán a tudományos célú hasznosítás elősegítése, a partnerségi formák kiépítése a kutatói közösségek felé. Partnerintézményekkel, kutatói műhelyekkel kötött szövetségek nélkül a digitális dokumentum, illetve webarchiválás sem végezhető eredményesen. Végeredményben megállapíthatjuk, hogy a levéltárakhoz hasonlóan az új típusú digitális dokumentumfajták feltűnésével a könyvtárak is rendkívül komplex kihívásokkal szembesülnek. Mint ahogy arra a szerző is utal, a közgyűjtemények felelőssége az információk hiteles, hosszútávú megőrzésében lelhető fel. Nem kevesebb a tét, minthogy pár év, évtized múlva ne digitális középkorként tekintsünk vissza a jelen időszakunkra.

A könyv harmadik nagy egysége a popkultúra és az online tér viszonyáról kínál tanulmányokat. Az első tanulmány szerzője, Radnai Dániel Szabolcs, az Omega együttes és rajongói online térben való megjelenésével foglalkozik. A vizsgálat azzal szembesít, hogy míg a digitális korszak előtt az együttesről megjelent monográfiák, interjúkötetek alkalmasak voltak arra, hogy felállítsanak egy uralkodó narratívát, az online térben létrejövő rajongói közösségek korában ez a törekvés illúzióvá vált. Egymással versengő megközelítésmódok korában élünk, amelyeket persze az adott együttes hivatalos kommunikációja is próbál befolyásolni, de mégiscsak ránk van bízva, hogy milyen narratívákat érzünk magunkhoz közelállónak, s könnyedén bekapcsolódhatunk az ahhoz kötődő kommunikációs tevékenységekbe is. Így akár egy több mint ötvenéves rockzenekar online rajongói közössége is működhet a közösségi média logikája szerint. Az Omega együttes eredetiségének és autentikus voltának megragadásáért folytatott küzdelem, a különféle életérzések megélésének lehetőségei az online világban fogalmazódnak újra.

A zene iránti lelkesedés mellett az online térben történő kultúrafogyasztás a sportrajongás világát is erőteljesen átformálta. Fodor Péter tanulmánya ennek egy tipikusan magyar szeletét bontja ki. A TrollFoci, az új online műfajnak, a mémnek az egyik legsi-

 DOI 10.31400/dh-hun.2022.6.4985

keresebb itthoni megjelenési formájává vált. A kezdetben kizárólag mémek közlésére támaszkodó portálon keletkeztek az ott zajló közösségi tevékenységeknek egyéb leágazásai az online videók és szöveges tartalmak felé. Itt is a szabad identitásválasztás, a narratívák egyéni megválasztása a kulcsmozzanat, s miután a világ a sportközvetítések által házhoz jön, az identitás szabad megválasztása egyben transznacionálissá válik. A TrollFoci példája viszont azt mutatja, hogy egy-egy jó kiinduló ötlet, amely mögé aztán közösséget is lehet szervezni, a lokális alapú futballidentitást is átélhetővé teszi a közösségi térben.

Egy másik fajta identitásmegélésre mutat rá az önsegítő tevékenységeket, illetve azok kritikáját, parodizálását bemutató tanulmány Molnár-Kovács Dorottya jóvoltából. Voltaképpen itt is versengő narratívákról van szó. Az önsegítés különféle formái új lendületet kapnak az online térben, de vele egyidőben megjelenik a kritika, a szkepticizmus és a paródia is. Ráadásul ez utóbbi is intézményesülő műfajszerű keretekbe kerül. A 2011–2016 között itthon létező demotiváló (az önsegítés különféle módszereit kifigurázni kívánó) plakátok történetének feldolgozása egyben jó példája az archivált webes anyagok tudományos célú hasznosításának is, hiszen ezek ma már az élő weben nem érhetők el. A tanulmány a hagyományos formában létező demotiváló plakátok történetének áttekintése után tér rá ennek online világban való jelenségére. Az önsegítés kritikája is átcsúszhat egyfajta alternatív önsegítő módszer bemutatásába, az önfejlesztés üzenetei megjelennek a plakátokon, a kritikai szövegekben, és ezáltal a paródia sokszor inkább erősíti, mintsem gyengíti az eredeti mondanivalót, az önsegítés műfajtudatát az online világban is.

A világháló magyar szeletének retró oldalaival foglalkozó Márfai Molnár László által írt fejezet két webhelyet emel ki. Mindkettő tematikus rendező elvként az 1960–1990 közötti tárgyi világot mutatja be. A *retronóm* olyan közösségi fórum és közösségi archívum jegyeit egyesíti magában, ahová bárki tölthet fel, illetve kommentálhat tartalmakat. A *szétszedtem* projekt ezzel szemben egy személyes identitás egyes elemeit dokumentálja. Egyedi, fanyar humorral osztja meg az adott tárgyakról szóló emlékeket, tapasztalatokat. Mindkét közösség a személyes emlékek megosztására, a nosztalgia iránti igényre épít; az egyik közösségi jelleggel elégíti azokat ki, a másik pedig az egyéni, tapasztalt szaki nézőpontjából történő beszámolóknak, emlékeknek ad teret. A kettő szervesen kiegészíti egymást. Értékes látleletet nyújtanak arról a jelenségről, amit a retró bosszújaként is szoktak emlegetni, hogy az újfajta felületeken az online világot megelőző, teljesen másfajta koordinátarendszerben zajló élet emlékei, s az azokhoz kapcsolódó reakciók jelennek meg, amelyek természetszerűleg már magukon viselik napjaink hálózati kommunikációjának sajátosságait. Azért is fontosak ezek az oldalak, mert újfajta, alulnézetből közvetített közösségi, illetve egyéni látásmódú képet adnak a Kádár-korszakról, annak hétköznapi tárgyi kultúrájáról.

A kötet negyedik, utolsó nagy tematikus blokkja a tartalomszolgáltatásoknak ad teret. Az első példa az *Origó,* Gálik Mihály révén, aki a portál megszületését, sorsának alakulását a Deutsche Telekom által történt értékesítés lezárultáig követi nyomon. Nagy figyelmet fordít a MATÁV, illetve Magyar Telekom kapcsán a portál megszületését, működését befolyásoló üzleti környezet bemutatására is. Szembesülhetünk azzal, hogy milyen egész amatőrnek tűnő elgondolások alapján vázolták fel a leendő portál üzleti tervét és tartalomszolgáltatási céljait. A tanulmányból kiderül, hogy

az egészet átszőtte a kilencvenes évek eleji korszaknak egyfajta naivitása, amely nem számolt azzal, hogy milyen gazdasági, politikai erőtérbe kerülhet egy sikeres tartalomszolgáltatás, s azzal sem foglalkozott, aminek pedig voltak itthon példái a HVG vagy a Népszabadság esetében, hogy a szerkesztőség jogait hogyan lehet különféle statútumokkal védeni. Az egész portál eleinte az *Origó* társadalmi szerepvállalásának egyfajta kivetüléseként működött, meglehetős veszteséggel, de a média világában egyre jobban felértékelődő súllyal. Később pedig, amikor a nagy infrastrukturális beruházások eredményeként a szélessávú internet elterjedésével összhangban a tartalomszolgáltatás üzletileg is felértékelődött, már a tulajdonos, a Deutsche Telekom üzleti érdekeit kezdte keresztezni az *Origó* független hangvétele 2010 után, ezért is döntöttek a portál eladása mellett. Tanulságos történet, amely elég jól rámutat arra, hogyan tud egy tulajdonos eleinte még egyfajta hobbyprojektként elkezdeni működtetni egy hírportált, amelynek sorsa – miután egyre sikeresebbé válik – gazdasági-politikai erőtérbe kerülve megpecsételődik.

Az *Index* portállal szintén önálló fejezet foglalkozik. Ennek esetében is fennáll az imént említett üzleti alapú befolyásrendszer. Az idők során azonban ez jóval átláthatatlanabb formát öltött, mint az előző fejezetben tárgyalt esetben. Emellett itt fokozottan megjelenik a globális médiaszolgáltató óriások hatása a hirdetési piacra és a szerkesztési stílusra is. Az állam és a közösségi média befolyásán túl a közvetett piaci befolyással fellépő politikai hatás veszélye is egyre inkább fellépett. S fokozott jelentősége van annak a kognitív környezetnek, amelyben az újságírók, illetve az egész szerkesztőség próbált reagálni az őket érintő kihívásokra. A portált piaci módon profi menedzserszemlélettel működtető, tartalmi kérdésekbe bele nem szóló Wallis után csupa olyan tulajdonos tűnt fel a színen, akik folyamatosan próbálták a szerkesztőség függetlenségét aláásni. A történteket látva nem kis bravúr, hogy végül csak 2020-ra vált végképp lehetetlenné a helyzet, akkorra vált nyilvánvalóvá, hogy autonóm, független szerkesztőségi szemlélettel adott tulajdonosi környezetben már nem működtethető tovább a portál. A média és a függetlenség különféle tényezőinek igen találó és tömör tálalása véleményem szerint a kötet egyik legsikerültebb fejezetévé teszi Tófalvy Tamás írását.

Tematikailag ide illik a kötetben utolsóként szereplő, az erdélyi online sajtó sorsát, helyzetét tárgyaló fejezet is, Botházi Mária jóvoltából. Ennek tanulságaként megfogalmazható, hogy a kisebbségi helyzet sajátosságai mellett nehéz úgy független, sokoldalú tartalomszolgáltatásokat működtetni, ha erre alig van társadalmi igény az erdélyi magyar közösségben. A védekező jellegű, elzárkózásra épülő, a közösség vélt javait szolgáló, tabukérdéseket nem tárgyaló, elsősorban helyi ügyekre koncentráló attitűdökre épült, 2010 után folyamatosan a magyar kormányzat által kialakított médiabirodalom hálózataaz erdélyi magyar online nyilvánosságban talán az itthoninál is egyoldalúbb környezetet alakított ki. Hozzá kell tenni, hogy a tanulmány megírása után is jelentős negatív változások történtek e téren, az egyetlen független, sokoldalú kulturális missziót betöltő s aktuális hírszolgáltatást nyújtó *Transindex,* az *Index*hez hasonlóan lehetetlenült el a tulajdonosi érdekek átláthatatlan hálójában (utóda a *Transtelex*). A magyar kormányzat pedig drámai mértékben kezdte el kivonni a forrásokat az uniformizált erdélyi magyar médiabirodalomból, mely a nyomtatott sajtótermékek jó részének megszűnéséhez, s az online portálok működésének takarék-

 DOI 10.31400/dh-hun.2022.6.4985

lángra kerüléséhez vezetett. A szerző által emlegetett lélegeztetőgép-effektus minden eddiginél drámaibban nyilvánult meg. Kérdés persze, hogy merre vezet a továbbiakban az út – találóan összegez a tanulmány: az újságírói függetlenséget, igényességet aláásó gazdasági, politikai érdekek, illetve a közösségi médiaszolgáltatások nyomásának helyi vetülete következtében a puszta túlélés a cél.

A politikai-gazdasági erőtér által erőteljesen befolyásolt hírportálokról szóló részek után egy olyan fejezet következik, Barkóczi Flóra vizsgálata nyomán, mely a piaci alapú megközelítéstől *ab ovo* idegenkedő művészeti tartalomszolgáltatási kezdeményezéseket tekint át a világháló magyar hőskorában *(Artpool, C3, Éjjeli Őrjárat),* melyek közül az utóbbi webes lenyomatai már csak webarchívumokban tanulmányozhatók. Kreatív és kooperatív projekteknek, művészeti kezdeményezéseknek adott teret mindegyik felület. Az internet kollaboratív és önszerveződő természetére építve járultak hozzá a képzőművészeti mező közösségformálásához, az együttműködések online tér révén formálódó fenntarthatóságához, s egyben hű lenyomatát adják a kilencvenes évek online művészi, esztétikai kultúrájának is.

Az *Artpool* erőteljesen épített az 1960–1990 közötti avantgárd művészeti tevékenységek hagyományaira, ennek online lenyomataként, továbbéléseként is működött, természetesen ezen új felület sajátosságaihoz kötődő projekteknek is teret adva. Az organikusan fejlődő webhely mára már egy hatalmas, szinte áttekinthetetlen hálózatot alkot, melyet egy tudatosan hiperlink-struktúrákkal létrehozott labirintusjelleg határoz meg. Az internetet kísérleti térként kezelve napjainkig folyamatosan alakítja saját felületét, mindenféle piaci szempont mellőzésével.

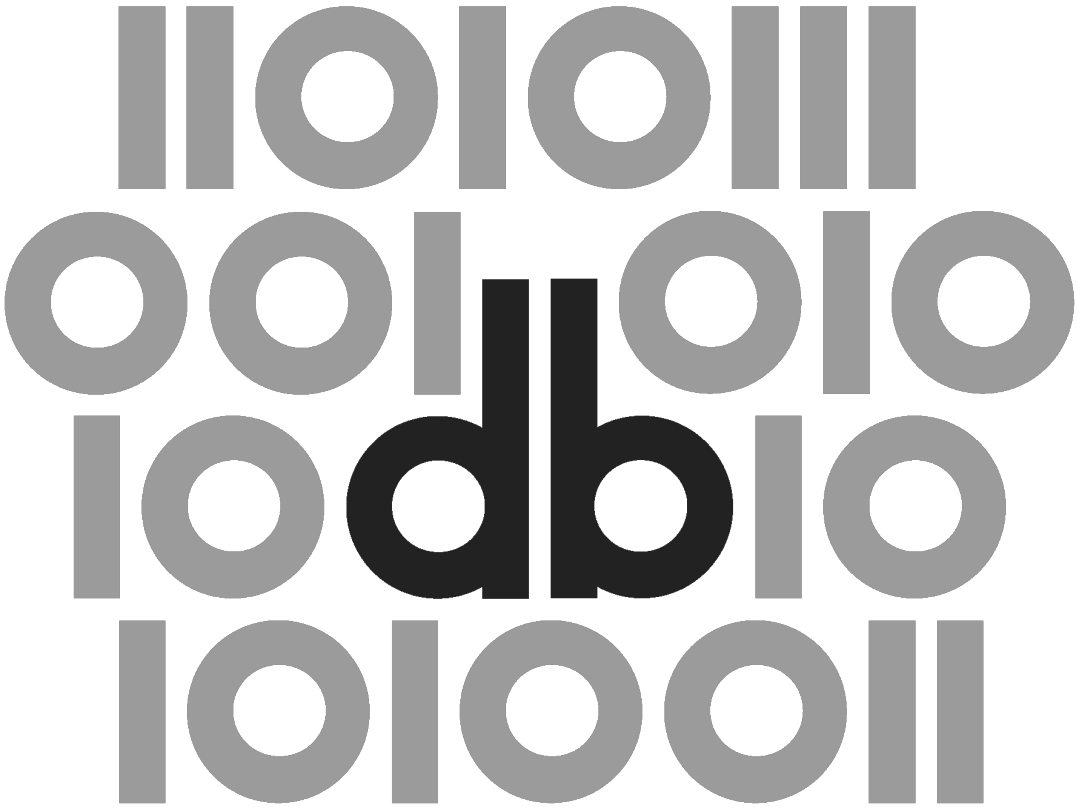
A C³ Kulturális és Kommunikációs Központ 1996-ban kifejezetten a magyar online kultúra támogatására és fejlesztésére jött létre. Az intézményi oldal máig működik, melynek komoly webtörténeti jelentősége is van. Az *Artpool*lal vagy az *Éjjeli Őrjárat*tal szemben elsősorban kommunikációs, illetve aggregátorcsatornaként működött, egyik fenntartója dokumentálta a Soros Alapítvány szervezeti tevékenységeit is, illetve eredetileg a *C³* volt a *freemail* levelező szolgáltatás fenntartója is. Ezen kívül az internetről való hazai gondolkodás és párbeszéd főszereplője volt a kilencvenes évek második felében, küldetésének része pedig egy olyan infrastruktúra létrehozása, melynek révén a hazai képzőművészeti közeg megismerkedhetett az internettel kísérletező nemzetközi képzőművészeti kezdeményezésekkel. A *c3.hu* domain számos webes művészeti projekt és portfólióoldal címeként szolgált és máig szolgál elsősorban médiaművészeti területen. A 2000-es években a források elapadása miatt azonban elveszett az a lehetőség, hogy modern művészeti centrummá váljanak a mindenkori technológia elemeit innovatívan felhasználva.

Az *Éjjeli Őrjárat* az imént említettekkel szemben eleven művészeti közösségszervező felület is volt. Önszerveződő kezdeményezésként reagált az információs társadalom új lehetőségeire. Sajnos azonban hosszabb távon épp ennek az energiája fulladt ki, s tette fenntarthatatlanná a működési modellt.

A kötetről a recenzió zárógondolataként újra meg kell állapítanunk, hogy a különféle nézőpontok gazdagsága hűen képezi le az internet történetei kapcsán zajló diskurzus igen színes voltát. Nagyon jó lenne, ha a szerzők közül minél többen bekapcsolódnának a nemzetközi internettörténettel foglalkozó szakmai fórumok munkájába. A bevezetőben szó esett már arról, hogy az internet történetének tanulmányozása

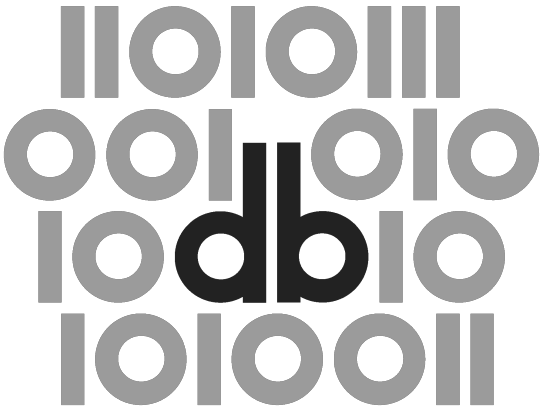
a nagy narratívák felállítása helyett jellemzően a mikrotörténet-írás terepe, ezért is lenne fontos, hogy ezek a tipikusan magyar történetek, szempontok gazdagítsák a nemzetközi tudományos diskurzust.

 DOI 10.31400/dh-hun.2022.6.4985



Digitális Bölcsészet
2022., hatodik szám

<DIGITÁLIS BÖLCSÉSZET>



6 (2022)

ISSN 2630-9696
DOI 10.31400/dh-hun.2022.6

Kiadja a Bakonyi Géza Alapítvány és az ELTE BTK Régi Magyar Irodalom Tanszéke (1088 Budapest, Múzeum krt. 4/A).
Felelős kiadó az ELTE BTK Régi Magyar Irodalom Tanszék vezetője.

Megjelenik az Open Journal Systems (OJS) v. 3. platformon, melynek működtetését az ELTE Egyetemi Könyvtár- és Levéltár biztosítja.

Honlap: `http://ojs.elte.hu/digitalisbolcseszet`
Email cím: dbfolyoirat@gmail.com

Olvasószerkesztő: Bucsics Katalin
Tördelés: Hegedüs Béla
Grafika: Hegyi Gábor

<KRITIKA>

Kuzma Gréta 0000-0002-4212-316X

PPKE Bölcsészet- és Társadalomtudományi Kar Irodalomtudományi Doktori Iskola

kgreti9307@gmail.com

Rab Virág. *Kapcsolati hálózatok a történelemben. Gerard Vissering és Hegedüs Loránt példája.* Budapest: Gondolat Kiadó, 2019. 232 oldal. ISBN 9789636939311

Az elmúlt évtizedekben a hálózatkutatás térhódítása nemcsak a természettudományok, de a humán- és társadalomtudományok esetében is egyre számottevőbbnek bizonyul. Noha nemzetközi szinten egyelőre gazdagabb szakirodalmi háttér támogatja a kutatók elmélyülését a hálózattudományokban, azonban kijelenthető, hogy egyrészt a nemzetközi kutatások eredményei egyre növekvő számban olvashatók el magyar nyelven, másrészt, és ez talán még üdvösebb, a hazai kutatók figyelmét mindinkább felkelti ez az irány.

Rab Virág első kötete 2010-ben *Diagnózisok és terápiák: Javaslatok az európai gazdaság újjáélesztésére az első világháború után* címmel jelent meg a Gondolat Kiadónál. Legújabb munkája a Kaposi Zoltánnal közösen szerkesztett *Different Approaches to Economic and Social Changes: New Research Issues, Sources and Results* (Working Group of Economic and Social History Regional Committee of the Hungarian Academy of Sciences in Pécs, 2022).

Az itt ismertetett *Kapcsolati hálózatok a történelemben: Gerard Vissering és Hegedüs Loránt példája* című kötetében, mely 2019-ben jelent meg, mélyrehatóan és új megközelítésből vizsgálja a hálózatkutatás, hálózattudomány alkalmazásának lehetőségeit a történettudományon belül. Teszi mindezt elsősorban a 20. század első felének két jelentős gazdasági szereplőjére fókuszálva, portréjuk és kapcsolati hálózatuk részletes bemutatásával, miközben az említett holland és magyar szakember sikerességének avagy sikertelenségének okait az egyéni és a környezeti tényezők figyelembevételével is vizsgálja. A kutatás újdonságának ereje abban rejlik, hogy a történettudományban eddig kevéssé alkalmazott hálózatok világát képes legitimálni. Rab Virág módszere, a két tudományterület együttes, kiegészítő használata, a szerző szándéka szerint kísérleti jellegű, célja, hogy a történettudományi kérdéseket a hálózattudomány segítségével, annak eszköztárával válaszolja meg. A fő kérdés, miszerint mit nyújt a hálózattudomány a történettudománynak és fordítva, megválaszolásra kerül, a két diszciplína metszéspontján lehetőség adódik a mikro- és makrotörténelem összekapcsolására, mindamellett további kutatási irányokra nyílik lehetőség. Ugyanakkor a szerző, bár sikerrel valósítja meg a két tudományterület együttes alkalmazását, a középpontba inkább a történettudományt helyezi, amelynek új megközelítési és értelmezési módjaként, kereteként szolgál a kapcsolati hálózatok bemutatása.

A kötet szerkezete jól felépített, a szerző szakszerűen vezeti be a témát az előzményektől kezdve a hipotézis felállításán át a kutatás megkezdéséig, az eredmények bemutatásáig és a konklúziók levonásáig. A precizitás visszaköszön a mű fő részének

megszerkesztettségében is, amely három nagyobb fejezetre bontható. Ezt előzi meg egy bevezető, amely a két tudományterület kohéziójának lehetőségeit fogalmazza meg, valamint rövid betekintést enged e kutatás történetébe.

Az első rész a hálózatok tudományát, a kutatások jelenlegi állását, kialakulásának történetét, a történettudomány szempontjából hasznos perspektívákat tárja az olvasó elé közel harminc oldalon keresztül. Ez nemcsak azért jelentős, mert a szerző így tüzetesen kontextualizálja a kutatás és a monográfia egészét, hanem mert az olvasót a hálózattudomány alapfogalmaival is megismerteti közérthető módon. A műben nem történik meg a hálózattudomány részletekbe menő bemutatása, ezt a szerző maga is elismeri. Megfogalmazott célja nem is ez, hanem a tudományterület legfőbb komponenseinek ismertetése oly módon, hogy forrásaiként az alapvető szakirodalmat, valamint az újonnan megjelent hazai és nemzetközi műveket is megjelöli. A kötet így további kutatások forrásaként szolgálhat.

A második részben Rab Virág közel nyolcvan oldalon át a hálózattudomány történettudományon belüli alkalmazásának lehetőségeit veszi sorra. Ezenfelül Gerard Visseringnek (1865–1937), a Holland Nemzeti Bank 1912-től 1931-ig soros elnökének, illetve Hegedüs Lorántnak (1872–1943), a kor magyar pénzügyminiszterének (1920–1921), gazdaságpolitikusának, írójának kapcsolatait mutatja be; családi, baráti kötelékeiket, s személyiségüket, a 20. század első felében születő kihívásokra adott válaszaikat, ún. „megküzdési stratégiáikat" mind hivatalos szerepeikben, mind magánéletükben. A szerző Hegedüs és Vissering személyes viszonyát is bemutatja, illetve összehasonlítja a két szakember többi kapcsolatával. Viszonyuk jelentőségét továbbá a magyar gazdaság alakulására nézve is megvizsgálja önmagában, valamint az európai és a világgazdaság felől értelmezve is.

A harmadik fejezetben a szerző Hegedüs Loránt családi kapcsolatait elemzi, mintegy ötven oldalon keresztül. Bár már a második fejezetben is kitér Hegedüs kapcsolati hálójára és életére, itt bővebben értelmezi mindezt a pénzügyminiszter egykori környezetének, családjának visszaemlékezései alapján. Rab Virág azért is tartja lényegesnek a történelmi szereplők magánéletének részletes megismerését és megismertetését, mert a személyiségből, az egyéni habitusból fakadó döntések, a problémamegoldó képesség, „megküzdési stratégiák" stb. mind hatással lehetnek nagyobb eseménysorokra is. Így az egyén családi kapcsolatrendszerének felvázolása (a családot mint társadalmi alrendszert értelmezve) elengedhetetlen részét képezi a kutatásnak. Hegedüs családjában betöltött szerepét leszármazottai, rokonainak visszaemlékezései, memoárjai segítségével ábrázolja: három alfejezeten keresztül ismertetve meg az olvasót a 20. század eleji magyar történelem egyik jelentős szereplőjével, annak legkisebb lánya, unokája és húga unokája nézőpontján keresztül. A könyv egyik különlegessége, hogy nemcsak írásos forrásokra támaszkodott a szerző, hanem lehetősége volt több alkalommal interjút készíteni Hegedüs Loránt unokahúgával, Tőry Magdolnával is. További érdekességként, de önmagában is izgalmas forrásként szolgált volna a visszaemlékezések, interjúk további részleteinek közlése.

A mű olvasása során részben megerősítést nyer, amit maga a szerző is hangsúlyoz már a kötet elején: egyik fő célja a hálózattudomány, a hálózatok létjogosultságának biztosítása, vagy legalábbis az erre való törekvés a történettudományon belül. Ennek értelmében a hálózatok alkalmazása, bemutatása nem puszta illusztrációs célt szolgál:

 DOI 10.31400/dh-hun.2022.6.3715

a vizuális elemek, az ábrák, táblázatok használata nem a kutatás szöveges, értelmezői összefoglalásának rovására történik, hanem mindvégig kiegészíti azt. Csakúgy, mint a kötet borítóján látható Chord diagram, amely több ízben visszatér a kötetben, az adott gondolatmenetet, hipotézist vagy kutatási eredményt segít értelmezni az olvasónak. Mindazonáltal a hálózati ábrák egyik nagy előnye, a színek alkalmazása hiányzik a könyvből, ami még inkább megkönnyíthette volna az olvasó számára az értelmezést. Ennek áthidalására megoldás lehetne esetleg egy szemléltető honlap létrehozása, amelyre egy megadott linken vagy QR-kódon keresztül lehet eljutni, s maguk az ábrák is megtekinthetők, nagyíthatók, értelmezhetők ilyenformán is. Már csak azért is, mivel érintőlegesen a szerző maga ugyancsak megemlíti, hogy a hálózati ábrák mögött komoly kutatómunka és egy online adatbázis, weblap létrehozása is áll, amelynek részletesebb bemutatása szintén értékes információkkal szolgálhatna az olvasó számára, mind hálózat-, mind történettudományi szempontból.

Maga Rab Virág is bemutatja a hálózati ábrákon keresztül saját tudományos kapcsolatrendszerét, de természetesen Gerard Vissering és Hegedüs Loránt szakmai viszonyrendszerei sem maradnak ki a könyvből. A kutatás során felvázolódik, hogy Vissering kapcsolati hálózata nem hierarchikus, hanem horizontális értelemben épült fel. Mindez azt jelenti, hogy egyenlőségen és kölcsönös tiszteleten alapult a hálózat, melyben a hierarchia legfeljebb tudásalapon jöhetett létre. Hegedüsre sem volt jellemző az alá-fölérendeltségi viszonyon alapuló rendszer kiépítése, az egyenlőség, partneri kapcsolatok kialakítása az ő esetében is elsődlegesnek mutatkozott.

Ahogy a szerző a hálózatok alkalmazásának legitimálása során több ízben is megjegyzi, a vizualizáció lényege a korábban felmerült előfeltevések igazolása, de egyben az új kérdések lehetősége, az új hipotézisek felállítása és az új konklúziók levonásának lehetősége is benne rejlik. Rab Virág fontosnak tartja kiemelni a különálló tudományterületek között létrehozható hidakat, a tudomány átjárhatóságát is. Így tehát a kötet nemcsak a történettudomány és a hálózattudomány meglévő fogalomkészletét vagy alkalmazási módszereit használja, hanem több tudományterület határmezsgyéjén keresztül keresi a választ a feltett kérdésekre, ennek köszönhetően pedig lehetőség nyílik összetett vagy új problémák megfogalmazására és megoldására, komplex rendszerek megértésére is. Ilyen eredmény lehet az is, amelyre a hálózatok alkalmazása is rávilágít, miszerint a nemzetközi és hazai gazdasági kapcsolatok egyik megkerülhetetlen tényezője az informális csatornák jelenléte. Részben ezen informális csatornák határozhatták meg Vissering és Hegedüs megküzdési stratégiáit, sőt lehetőségeiket az I. világháborút követő gazdasági káosz mérséklésére, a rend helyreállítására tett kísérleteik során. Ezt tovább segíthették avagy nehezíthették a sajátos környezeti és egyéni tényezők. Az erős és gyenge kapcsolatok kérdését is sikeresen építi be a szerző a művébe, kiemelve a gyenge kapcsolatok fontosságát, környezetformáló erejét mikro- és makroszinten egyaránt.

Rab Virág kutatása érdekes kordokumentumokkal és újszerű megközelítésekkel ismertetheti meg az olvasót abból a szempontból is, hogyan tudott a magyar gazdaság az európaihoz, majd azon keresztül a világgazdasághoz kapcsolódni az I. világháborút követően. Mindazonáltal a szerző kutatási iránya nemcsak új források megismerését teszi lehetővé, hanem korábbról ismert forrásokat más megközelítésből mutat meg. Hasonló célt szolgál a két, holland és magyar gazdasági szakember életútjának felvá-

zolása is, betekintést engedve nemcsak a Visseringet és Hegedüst ért történelmi eseményekbe és az azokra adott válaszaikba, hanem személyiségük, jellemük elemi ismertetőjegyeibe, magánéletük alakulásába (noha Hegedüsnél bővebben ismerhetjük meg ez utóbbi alakulását, mint Visseringnél), amelyek felől értelmezhetők a makroszintek eltérései is (a környezeti tényezők sajátos jellemzői mellett). Vissering szerepe ebben az időszakban különösen érdekes, hiszen az I. világháborúban betöltött pozíciójától függetlenül érintkezik minden oldallal; egyik legfőbb célja a háború által szétzilált gazdasági színtér stabilizálása. Hegedüs szintén erre törekszik magyar viszonylatok között, kérdéses, hogy erre mennyi lehetősége nyílik egy erősen eltérő kulturális és gazdasági közegben.

Rab Virág kutatási eredményeit egyszerre képes tudományos, mégis érthető és élvezhető módon bemutatni, az adott tudományággal akár csak érintőlegesen ismerkedő vagy laikus olvasó is értékes információk és következtetések birtokába juthat. Az ábrák és táblázatok további segítséget nyújtanak az értelmezésben és későbbi hipotézisek megfogalmazásában is (az adatok feldolgozása és a hálózati térképek, ábrák vizuális megalkotása Kriszbacher Gergő munkája). Rab Virág stílusa végig befogadható, olvasmányos marad, nem nehezül el, érthető módon vezeti végig kutatását a téma megszületésétől a hipotézisek megfogalmazásán és a feltáró munkák és elemzések megkezdésén át a konklúziók levonásáig. Kiemelendő ezenkívül, hogy a kutatás szempontjából a jelentős fordulópontokat akkurátus módon mutatja be, majd összegezi eredményeit. Mindez a témával vagy a tudományterületekkel behatóan foglalkozó olvasók, kutatók számára hatalmas segítség, képes további kérdések megfogalmazását indukálni, így gazdagítva a tudományos párbeszédet.

A szerző nem hagy megválaszolatlanul egyetlen kérdést sem, a történelmi szereplők kisebb-nagyobb mértékben, de minden esetben bemutatásra kerülnek. A felhasznált forrásokat, levelezéseket, irodalmat is igen gondosan állította össze. Talán még jobban ki lehetett volna térni a hálózattudomány gyakorlati részére, akár egy szemléltető táblázat bemutatásával vagy olyan kifejezések, szakfogalmak pontosításával, mint például a *source* vagy a *target id* (forráshoz és célszemélyhez kapcsolt azonosító szám), amelyek a hálózati ábrák megalkotása során kerülnek szóba, s egy laikus olvasó számára ismeretlenek lehetnek; összességében ugyanakkor a könyvben így is átlátható kép rajzolódik ki a hálózatokról, a hálózattudományról. Rab Virág művének jelentősége abban is megmutatkozik, hogy nemcsak a két fő sodor, a történettudomány és a hálózattudomány határmezsgyéjén mozogva keresi a kérdésekre a választ, de beemel számos más tudományágat is, így a szociológiát, pszichológiát vagy szociálpszichológiát. Az olvasó további segítségére szolgál a könyv végén található névmutató, valamint értékes anyagokat kínál későbbi kutatások számára a források és a felhasznált irodalmak jegyzéke is.

Összegezve, Rab Virág *Kapcsolati hálózatok a történelemben* című műve igen értékes munka mind a hálózattudomány, mind a történettudomány, de más tudományágak szempontjából is, hidat biztosítva a különböző területek közötti átjárhatóságnak. A könyv egyben a hálózatok társadalom- és humán tudományon belüli alkalmazhatóságának bizonyítéka.

 DOI 10.31400/dh-hun.2022.6.3715