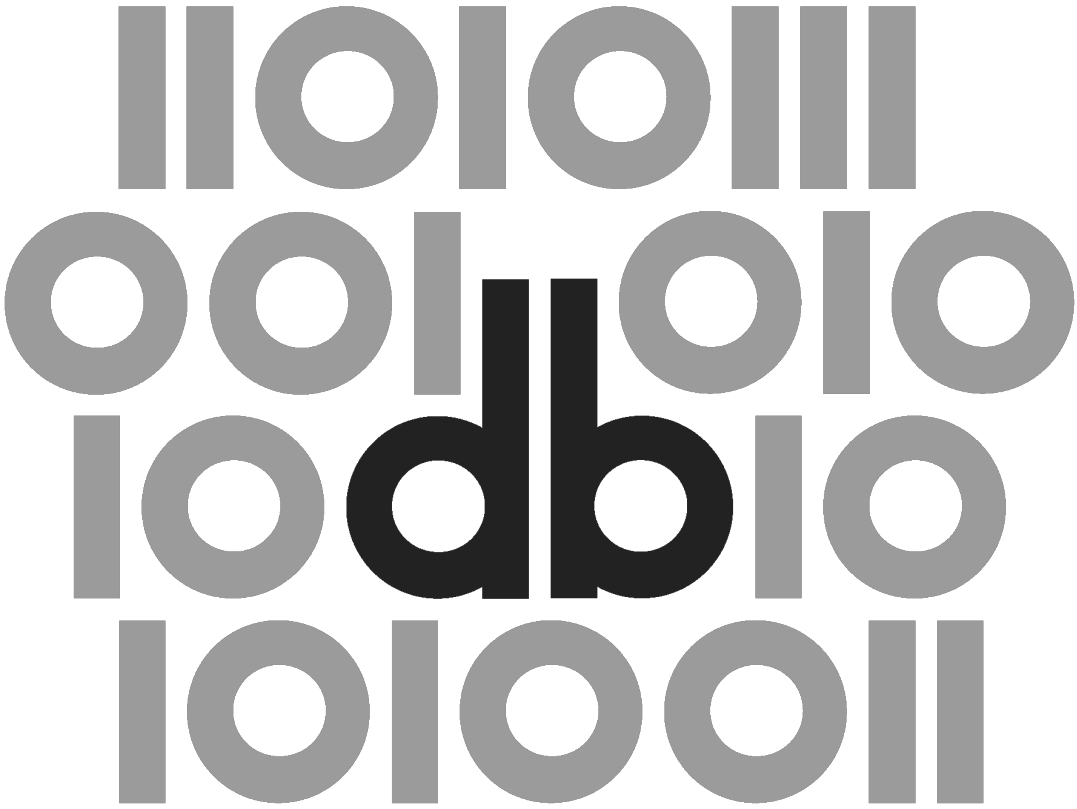
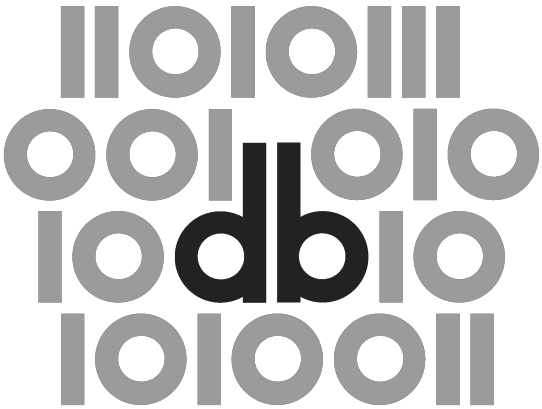


Digitális Bölcsészet
2022., hatodik szám

<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

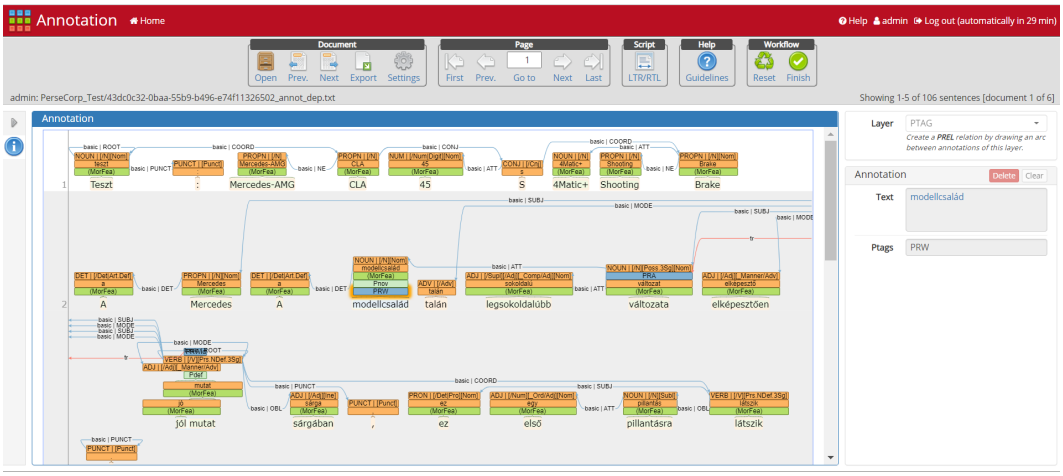
A kézi elemzés megkezdése előtt a korpusz szövegeit az *e-magyar Digitális Nyelvfeldolgozó Rendszer* segítségével elemeztem.[30] A teljes nyelvi anyag tokenizáláson,

[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

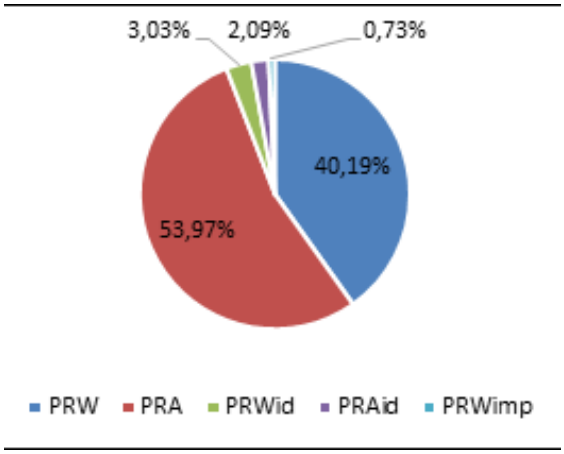
Az autóteszt sorszáma	Az autóteszt terjedelme (tokenszám)	Címkézett tokenek száma (db)	Relatív gyakoriság (%)
T1	2190	152	6.94
T2	1577	145	9.19
T3	1536	152	9.90
T4	2148	144	6.70
T5	1535	111	7.23
T6	1482	114	7.69

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

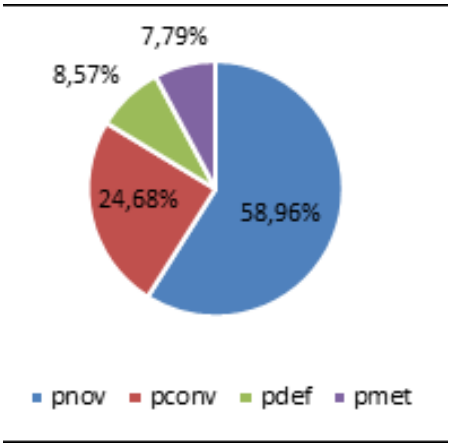
mennyiségét. Megállapítható tehát, hogy átlagosan minden PRW-vel jelölt tokenre esik legalább egy argumentum.[40] Ez a megfigyelés alátámasztja a szakirodalom azon állítását, hogy a nyelvi megszemélyesítések tipikusan egynél több szóból álló kifejezések. Az idiomatikus megszemélyesítések jóval csekélyebb 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