Névelem-felismerés magyar nyelvű jogi szövegeken

Üveges István

SZTE BTK Nyelvtudományi Doktori Iskola
uvegesistvan898@gmail.com

Kivonat: A jelen tanulmányban a névelem-felismerés hatékonyságának elemzésére teszek kísérletet jogi szövegek területén. A vizsgálat során részletesebb elemzésnek vetem alá a két elemzőt: a magyarlanc nyelvi elemzőt és a szintén az MTA-SzTE Mesterséges Intelligencia Kutatócsoport fejlesztette korábbi tulajdonnév-felismerő kimenetét. Elsőként röviden ismertetem a jelen elemzés szempontjából lényeges szakirodalmi hátteret. Ezt követően a vizsgálat tárgyát képező adatok kvantitatív elemzésére térek ki bővebben. A tanulmány következő részében néhány reprezentatív, problémás esetet és ezekre vonatkozó megoldási javaslatot ismertetek, amelyeket végül a további kutatási irányok meghatározása követ.

1 Bevezetés

A névelem-felismerés a szövegbányászat, ezen belül is az információkinyerési feladatok egyik alterülete. A szövegbányászat maga „szöveges adatokon végzett feldolgozási és elemzési tevékenység, melynek célja a dokumentumokban rejtetten meglévő új információk feltárása, azonosítása és elemzése" (Tikk et al. 2006: 22). Ezzel összefüggésben információkinyerésen „a szövegbányászati feladatok egy speciális esetét értjük, ahol a cél az adott feladat szempontjából fontos szövegrészek (információk, tények) kigyűjtése a dokumentumokból, azaz strukturálatlan szövegekből strukturált információ előállítása" (Tikk et al. 2006: 81). A névelem-felismerés a számítógépes nyelvészetben egészen a korai 1990-es évektől kezdve van jelen, és azóta is fontos feladatnak és megoldandó problémának számít. A jelen tanulmányban a hangsúly a jogi szövegekben fellelhető névelemek felismertetésén van, ahol a dokumentumok (félig) automatikus anonimizációja, valamint az informatívabb és hatékonyabb keresőeszközök kifejlesztése iránt napjainkban is folyamatos az érdeklődés. A jogi doménben névelemek alatt nem pusztán emberek vagy szervezetek neveinek említéseit érthetjük, de számításba kell vennünk például törvények neveit is.

A nemzetközi szakirodalomban számos olyan projekttel találkozhatunk, amelyek a névelemek felismerését tűzték ki célul az angolszász jogrendszer dokumentumaiban (pl. Lenci et al. 2009; Quaresma–Gonçalves 2010; Surdeanu et al. 2010), de a névelemek felismerésének kérdése megjelenik a hazai szakirodalomban is (Móra et al. 2011; Simon 2008; Simon 2017; Vincze–Farkas 2012 stb.).

Üveges István: Névelem-felismerés magyar nyelvű jogi szövegeken. In Váradi Tamás (sorozatszerkesztő), Ludányi Zsófia, Gráczi Tekla Etelka (szerkesztő): *Doktoranduszok tanulmányai az alkalmazott nyelvészet köréből 2019. XIII. Alkalmazott Nyelvészeti Doktoranduszkonferencia.* Budapest: MTA Nyelvtudományi Intézet. 2019. 126–137.
DOI: 10.18135/Alknyelvdok.2019.13.10

Gyakorlati szempontból a névelemek felismerésével és kategorizálásával a jogi információkinyerés elősegíthető mind a jogászok, bíróságok, valamint egyéb kormányzati szervek, mind pedig a joghoz laikusok számára.

2 Módszertan

Ebben a fejezetben röviden ismertetem a vizsgálat alapjául szolgáló korpuszt, illetve a kutatás metodológiáját.

A magyar nyelvre jelenleg három módon lehetséges a tulajdonnevek automatikus annotációja: tulajdonnév-felismerő algoritmussal, szófaji címke szintjén történő megkülönböztetéssel, valamint szintaktikai szintű címkézéssel. Utóbbi kettőre példa a jelen tanulmányban is elemzett magyarlanc működése, mivel ez szófaji szinten megkülönbözteti a tulajdonneveket, a szintaxis szintjén pedig jelöli a többtagúakat.

A tulajdonnév-felismerő algoritmusok megkeresik az adott szövegben a tulajdonneveket, majd azokat valamilyen kategóriába sorolják. Az [1] esetében például az algoritmus személynévi (PER) címkével látta el az adott tokent.

[1] Péter I-PER

A [2] a) egy névelem szófaji megkülönböztetését, a [2] b) pedig a többtagú névelemek szintaktikai elemzését szemlélteti. Utóbbi esetében a névelem tagjait összekötő NE (Named Entity) él mutatja meg az érintett tokeneket.

[2]

a)	7 Dóm	Dóm	PROPN	Case=Nom|Number=Sing	8	ATT
b)	6 Dóm	Dóm	PROPN	Case=Nom|Number=Sing	7	NE
	7 utcai	utcai	ADJ	Case=Nom|Degree=Pos|Number=Sing	8	ATT

A továbbiakban elsőként a vizsgált adatokra térek ki, majd a fent említett három eljárás általam elvégezett vizsgálatát ismertetem röviden.

2.1 Adatok

A vizsgálatot a Miskolc Jogi Korpuszon (Vincze 2018) végeztem, amely jogász, nyelvész és informatikus szakértők közreműködésével készült el annak érdekében, hogy megnyissa az utat a jogi nyelv (akár korpusznyelvészeti) egzakt tanulmányozása előtt. Összeállítása során fontos célkitűzés volt, hogy a létrejövő szövegállomány a magyar jogi nyelv minél szélesebb szegmensét lefedje, így hat különböző forrásból tartalmaz szövegeket (vö. Vincze 2018):

- 5 magyar törvény teljes szövege (a továbbiakban: Törvények),
- jogi témájú fórumok szövegei (a továbbiakban: Fórumok),
- bírósági tárgyalások és rendőrségi kihallgatások átiratai (a továbbiakban: Átiratok),
- jogszabályok miniszteri indoklásai és jogi egyetemek tankönyveinek szövegei,
- bírósági és törvényszéki ítéletek szövegei,
- törvényekből és jogszabályokból kiválasztott szövegrészek.

A jelen elemzés bázisát a Törvények, Fórumok és Átiratok részkorpusz első, megközelítőleg 6000 tokenje adta (1. táblázat).

A vizsgált szövegtípusok kiválasztásakor a fő szempont a minél inkább heterogén tipográfiai megjelenés (tagolás, központozás, bekezdések stb.), valamint a jogi nyelv intuitívan és korábbi doménhasonlósági vizsgálatok alapján (vö. Vincze 2018) is leginkább eltérő aspektusainak reprezentálása volt. Ennek megfelelően a metanyelvi szövegeket a fórumbejegyzések, a „klasszikus" jogi szövegeket a törvények, míg a beszélt nyelvi jogi témájú szövegeket az átiratok reprezentálták.

Részkorpusz	Tokenszám	Szószám
Törvények	6014	4660
Fórumok	6041	4718
Átiratok	6010	4594

1. táblázat. A vizsgált részkorpuszok jellemzése

Az elemzés során fontos kérdés volt, hogy a jogi nyelv eltérő forrásai eltérő kezelést igényelnek-e a névelemek felismerése szempontjából. Ennek megválaszolására a kvantitatív vizsgálat három szinten valósult meg:

- Az első lépést a kiválasztott szövegrészletek manuális annotálása jelentette a bennük előforduló névelemekre.
- Ezt követően a standardként szolgáló manuális annotációs eredmények összevetése történt meg a magyarlanc szófaji címkézésének eredményével, valamint a tulajdonnév-felismerő kimenetével.
- Végül a magyarlanc függőségi nyelvtani elemzésében megjelenő (vagy éppen hiányzó) NE (Named Entity) címkék felmérése következett a többszavas névelemek esetében.

Az elemzés kvalitatív részében az automata elemzők leggyakoribb hibaforrásainak felmérése történt meg. Az így szerzett adatok betekintést engednek a jogi szövegek néhány olyan sajátosságába, amelyek a kiválasztott szövegek esetében megnehezítették az automatikus elemzést, így segíthetnek a jövőbeni tulajdonnév-felismerő elemzők pontosságának javításában is.

2.2 Manuális annotáció

A kontroll adatokat a kézi annotálás adta, amely nagyban támaszkodott a HunNER korpusz annotálása során alkalmazott irányelvekre (Simon et al. 2006).

Ezekhez képest a lényegi eltérést a tag-for-tagging elv alkalmazása jelentette, azaz, hogy a keresett kifejezések nem az aktuális, szövegeli szerepük, hanem lexikális jelentésük alapján lettek kategorizálva, mintegy egyszerűsítve ezzel az automatikus elemzők kimenetének kiértékelését. A névelemek annotálandó kategóriáihoz az ACE 2006 annotálási útmutató szolgált alapul (Linguistic Data Consortium 2006), mindazonáltal az ott felsorolt jelentős számú kategória közül csak a név szerinti említések (name mentions), a helyek nevei (locations) és a szervezetek nevei (organizations) kerültek be a végleges, annotálandó kategóriák közé. Így tehát az annotálás során keresett három alapkategória a személynevek (PER), helynevek (LOC) és szervezetek (ORG) megnevezései voltak.

Emellett említést érdemelnek még a törvények, rendeletek nevei is (pl. Ptk., Tht.) amelyek szintén jelölésre kerültek. Az annotált kifejezéseket a 2. táblázat mutatja be részletesen.

Fontos még említést tenni az alapvető különbségről az angol névelem-felismerés (Named Entity Recognition – NER) és a magyar tulajdonnév-felismerés között. A névelem-felismerés egy alapvetően tágabb vizsgálati tartományt jelöl ki, amelybe beletartoznak numerikus kifejezések, dátumok és minden olyan kifejezés (akár azonosítók, telefonszámok és e-mail címek is), amely a világ valamely entitására egyedi módon (unikálisan) referál (vö. Tikk et al. 2006: 90–98). A szűkebb értelemben vett tulajdonnév-felismerés, ahogyan a neve is sugallja, kizárólag a tulajdonnevekre koncentrál. Többen is írtak már arról, hogy a számítógépes nyelvészet szempontjából pontosan mi tekinthető tulajdonnévnek (pl. Vincze–Farkas 2012). A jelen tanulmányban azonban csak a fentebb említett négy „típust" (személynevek, helynevek, szervezetek nevei és törvények, rendeletek nevei) tekintettem annotálandónak.

2.3 Automatikus névelem-felismerés

A kiválasztott szövegeknek a magyarlanccal és a tulajdonnév-felismerővel való elemeztetése után a következő lépés annak eldöntése volt, hogy az egyes tokenekhez megtörtént-e a megfelelő címkék hozzárendelése.

A magyarlanc esetében az elvárt szófaji címke a tulajdonnév (PROPN) volt, míg a tulajdonnév-felismerő esetében egy bármilyen tulajdonnévi kategória (PER, ORG stb.) jelenléte. Utóbbinál a kategóriák szerinti besorolás helyessége nem képezte a vizsgálat tárgyát.

Az eredmények ismertetése előtt fontos röviden kitérni az elemzők eredeti tanítókorpuszára, mivel a névelem-felismerés erősen doménspecifikus feladat, az elemzők betanítása során használt szövegek tehát erősen kihatnak azok későbbi eredményeire is.

Korpusz	Névelemként annotált tokenek száma	Névelemek száma	Többszavas névelemek	Kategória
Átiratok	56	29	23	PER
	41	22	11	LOC
	69	33	22	ORG
	25	11	7	REG[1]
	191	95	63	A részkorpuszban
Fórumok	95	51	19	PER
	4	4	0	LOC
	6	5	1	ORG
	6	6	0	REG
	111	66	20	A részkorpuszban
Törvény-szövegek	0	0	0	PER
	5	3	2	LOC
	14	8	4	ORG
	6	1	1	REG
	25	12	7	A részkorpuszban
Összesen:	327	173	90	

2. táblázat. Annotált kifejezések ([1]REG – Regulations: törvények, rendeletek nevei)

A magyarlanc eredeti tanítókorpusza (a Szeged Treebank) hat eltérő részből épült fel; célja szerint a lehető legkülönbözőbb tematikájú szövegtípusokat volt hivatott reprezentálni (Csendes et al. 2004). A tanításhoz használt korpusz ez esetben tartalmazott jogi szövegeket, bár azoknak csak egy speciális esetét; törvények szövegeit. A tulajdonnév-felismerő, bár ugyanezen a korpuszon lett betanítva, de nem a teljes szövegtesten, csak annak egy részhalmazán: az üzleti rövidhíreken. Ugyanezért tehát a jogi szövegek a tanítókorpusznak nem képezték részét. Ennek az elemzőnek az eredetileg mért pontossága az akkor vizsgált négy kategóriára (PER, ORG, LOC, MISC) együttesen 94,77% volt (Szarvas et al. 2006).

A fentiek tükrében az előzetes várakozás szerint a magyarlanc nagyobb eséllyel volt hivatott megfelelően címkézni a névelemeket a jogi domén szövegeiben. Ezen belül is a legpontosabb eredményeket a törvényszövegek címkézése során vártam.

3 Eredmények

A 3. táblázat a fontosabb tokenszintű mérőszámokat ismerteti. Az adatok számítási alapját a magyarlanctól PROPN, a tulajdonnév-felismerőtől pedig I-PER, I-ORG, I-LOC vagy I-MISC címkét kapott tokenek adták. Nem volt ugyanakkor kritérium, hogy egy adott token mind a két elemző által felismert legyen, a két szoftver kimenetét tehát ebből a szempontból egymástól függetlenül értékeltem.

Az adatokból leolvasható, hogy az elemzők közül a tulajdonnév-felismerő konzekvensen jobb eredményeket ért el a mért számok mindegyikében, valamennyi részkorpusz elemzése során. Az előzetes várakozásokkal ellentétben a törvényszövegek bizonyultak a legkevésbé pontosan elemzettnek, míg a skála másik végpontját az Átiratok adták.

A lényegesebb jellemző hibaforrásokat a továbbiakban az egyes részkorpuszokra lebontva tárgyalom.

Részkorpusz		Névelem- felismerő	magyarlanc morfológia
Fórumok	Pontosság	83,10	69,51
	Fedés	51,75	50,00
	F-érték	63,78	58,16
Átiratok	Pontosság	**94,48**	63,22
	Fedés	70,26	56,41
	F-érték	80,59	59,62
Törvényszövegek	Pontosság	63,33	**26,67**
	Fedés	73,08	61,54
	F-érték	67,86	37,21

3. táblázat. Az elemzők által elért tokenszintű eredmények

3.1 Fórumok

Az internetes jogi fórumok bejegyzéseiben előforduló névelemek közül a nicknevek bizonyultak a leginkább problémásnak, mivel alakjuk szerint (kis- vagy nagybetűs írásmód, kiterjedés stb.) nehezen megjósolhatók, éppen ezért potenciálisan nagy kihatásuk lehet az elemzők pontosságára.

A [3] néhány tipikus előfordulást szemléltet olyan nicknevekből, amelyek egy tokent tartalmaznak, a [4] pedig néhány olyat, amelyek több tokenné lettek szegmentálva az automatikus elemzés során.

Habár a fenti példák esetében a besorolás nem minden esetben volt megfelelő, azt fontos hangsúlyozni, hogy az elemzők eredeti tanítókorpuszai nem tartalmaztak olyan szövegeket, amelyekből azok elsajátíthatták volna a nicknevek felismeréséhez szükséges mintákat.

A nickneveknek névelemként való kezelése nyelvészeti szempontból is érdekes probléma. Egyrészről az informális szövegek (jelen esetben a fórumbejegyzések) automatikus feldolgozása már kiindulásként is sokkal összetettebb feladat, hiszen azok gyakran nem tesznek eleget minden szigorú grammatikai konvenciónak, több nyelvtani és helyesírási hibát tartalmazhatnak, ami megnehezíti az elemzők munkáját (vö. Einat et al. 2005). Jogi fórumok esetében pedig, mivel a célközönség erősen limitált, nagy eséllyel fordulnak elő csak a szaknyelvre jellemző terminusok, fordulatok, rövidítések (bár ez utóbbi mind a három vizsgált részkorpuszról elmondható).

[3]

Token	Magyarlanc által társított szófaji címke	Tulajdonnév-felismerő címkéje
55teki55	PROPN	O
heidi1115	NUM	O
ObudaFan	PROPN	I-ORG

[4]

Token	Magyarlanc által társított szófaji címke	Tulajdonnév-felismerő címkéje
Dr.	NOUN	I-PER
Attika	NOUN	I-PER
Kovács	PROPN	I-PER
_	X	I-PER
Béla	PROPN	I-PER
_	X	I-PER
Sándor	PROPN	I-PER

Másrészről a nicknevek nem tartoznak a klasszikus értelemben vett tulajdonnevek közé. A névelem-felismerési feladatoknak lehetnek alanyai, hiszen ez egy sokkal tágabb definíció, amely éppen ezért megengedőbb is a tárgykörébe tartozó, felismerendő kifejezések körével kapcsolatban. A nemzetközi szakirodalomban például előfordul a márkanevek mellett akár konkrét termékek azonosítási kísérletével foglalkozó cikk is (Yangjie–Aixin 2016), így ebbe a tárgykörbe a nicknevek könnyen beletartozhatnak.

A tulajdonnevekkel szemben támasztott egyik fontos követelmény az identifikáló funkció, vagyis, hogy egy adott név különböző nyelvekben egyformán szerepelhet, egyformán azonosíthatja jelöltjét, nem fordítható (Farkas 2007: 167). A nicknevek a Kripke által megfogalmazott merev jelölő definíciójának (egy merev jelölő kifejezés minden lehetséges világban ugyanazt a dolgot jelöli, amennyiben az a dolog létezik a kérdéses lehetséges világban) is megfelelnek (vö. Kripke 1980), még ha a lehetséges világ terminusa nehezen is hozható összefüggésbe a konkrét fórumbejegyzésekkel.

A legfőbb indok azonban, ami miatt a nicknevek tulajdonnévként azonosíthatók, azt főként az a fajta használatuk, hogy a fórumok alkotta környezetben tulajdonneveket helyettesítenek, azok szerepét töltik be. Használatuk célja, hogy az online környezetben azonosítsák viselőjüket, annak valós nevét helyettesítsék, amely szintén tulajdonnév. Ilyen értelemben tehát maguk is unikusan, egyedi módon referálnak a világ valamely entitására.

Egy másik kérdéses pont lehet, hogy miképpen ítéljük meg a korpuszban az egyes szervezetek név szerinti említéseit. Számos olyan eset fordult elő a szövegben, ahol az ugyanarra a szervezetre referáló kifejezés kétféleképpen fordult elő; egy olyan változatban, ahol a név kezdőbetűje kisbetű, és egy olyanban, ahol nagy kezdőbetűs írásmód érvényesült, például:

[5] *...ez volt a legfőbb érve a **törvényszéknek**, hogy szabálytalanul lett kézbesítve az idézés.*

[6] *...a végzés ellen fellebbezést nyújtsak be a várossal egy megyében található **Törvényszéknek** címezve 3 példányban.*

Az ilyen esetekben a két megjelenési forma közül kizárólag a nagy kezdőbetűs írásmóddal rendelkezőt tekintettem tulajdonnévnek (a [6] esetében például konkrétan a Szegedi Törvényszékre történő hivatkozásnak), a kisbetűvel írt változatot az intézmény köznevesült említéseként tartottam számon, tehát nem is került annotálásra felismerendő kifejezésként, nem lett része a statisztikai adatoknak.

Ennek alapját az a feltételezés adta, hogy a „beszélői" szándék szerint, amennyiben a fórumozó konkrét intézményt említ meg, akkor annak teljes nevét, vagy legalábbis nagy kezdőbetűs írásmódját alkalmazza, amennyiben viszont csak az adott intézmény típusára, szerepkörére (iskola, bíróság, rendőrkapitányság stb.) akar utalni, azt kis kezdőbetűs írásmód alkalmazásával teszi.

3.2 Átiratok

A tárgyalások, kihallgatások írott változatai esetében a legjellemzőbb problémák a mondatkezdő pozícióhoz voltak köthetők a magyarlanc kimenetében. Ezen belül két típushiba volt a legjellemzőbb.

A leiratokban a diskurzusok szegmentálásának fontos eszköze a megszólaló személyének rögzítése minden beszélőváltást követően. Ezek a jelölések tipikusan a megszólalások elején helyezkednek el, továbbá az adott beszélőnek a konkrét perben vagy eljárásban betöltött szerepét rögzítik. Ilyen rövidítés volt a korpuszban például a *V.*, amely a vádlott helyett állt, a *B.* amely a bírói szerepkörre utalt, illetve az *Ü.* vagy éppen *Ü / Ügyv.* amely az ügyvéd megjelöléseként került alkalmazásra. A [7] tipikus és rövid példája annak, amikor a rövidítés nem megfelelően került elemzésre.

A magyarlanc által a részkorpuszban tévesen PROPN címkével ellátott (fals pozitív) esetek jelentős többsége (60.93%, ami 39 esetet fedett le az összes 64-ből) ebből a speciális esetből következett.

A további, nem helyesen megjósolt címkék változatosabb okokra voltak visszavezethetők. Ezek között előfordultak könnyebben magyarázhatók, mint például a szintén mondatkezdő pozícióban elhelyezkedő *Bíró*, amely ugyan a magyar vezetéknévvel való analóg formai megjelenése miatt szintén PROPN címkét kapott, habár a jelen esetben csak az illető perben betöltött szerepére utalt vele a leiratozást végző. Ugyanakkor néhány nehezebben interpretálható címke is felbukkant, mint például a [8] esetében.

A tulajdonnév-felismerő fals pozitív címkéit megvizsgálva szintén változatos esetekkel találkozunk [9]. A [9] a) esetében a téves címke egyértelmű, a [9] b) és c) példák érdekesebbek néhány szempontból.

[7]

1	Ü	Ü	PROPN	Case=Nom|Number=Sing	0	ROOT
2	/	/	PUNCT	_	1	PUNCT
3	Ügyv	Ügyv	PROPN	Case=Nom|Number=Sing	1	COORD
4	:	:	PUNCT	_	1	PUNCT
5	Nem	nem	ADV	PronType=Neg	1	NEG

[8]

1	Öö ö	Öö ö	PROPN	Case=Nom|Number=Sing	6	SUBJ
2	amikor	amikor	ADV	PronType=Rel	4	TLOCY
3	azt	az	PRON	Case=Acc|Number=Sing|Person=3|PronType=Dem	4	OBJ
4	felvetettünk	felvet	VERB	Definite=Ind|Mood=Ind|Number=Plur|Person=1|Tense=Past| VerbForm=Fin|Voice=Act	6	ATT

[9]

a)	.	I-ORG
b)	Urat	I-PER
c)	Interneten	I-ORG

Az *internet* esetében az angolban, ahonnan a szó ered, sokáig két írásmódja volt használatos, a kisbetűs *internet* egyszerűen számítógépek, vagy egyéb informatikai eszközök belső hálózatát jelentette, míg a nagybetűs *Internet* a mai értelemben a World Wide Web megfelelője volt, arra mint egyedi, elvont fogalomra utalt (Simpson–Weiner 1989). A magyar helyesírásban a szó eredetileg mindkét változatban elfogadott volt, azaz tulajdonnévi és köznévi használata is „megengedettnek" számított (Deme et al. 1999). Ezt később egyértelműsítették (Laczkó–Mártonfi 2004), és már csak kisbetűs, köznévi használata volt megengedett. Ahogyan ezt Vincze–Farkas (2012: 100–101) több más példával is alátámasztva részletesen kifejti, ez jól mutatja, hogy a tulajdonnévség kérdése túlmutat a formai vagy helyesírási kérdéseken. Emellett arra is jó példa, hogy a nyelv különböző diakrón állapotai során is eltérhet egy-egy konkrét szó tulajdonnévi vagy köznévi megítélése (pl. köznevesülés).

Hasonlóan problémás eset lehet a [9] b), ahol az *urat* megjelölés a tulajdonnév részének tekinthető, bizonyos körülmények között. Ilyen eset lehet például, ha egy terem-

ben több Kovács vezetéknevű ember is tartózkodik. Ilyenkor a név jelentése önmagában homályos lehet, szükség van pontosításra, hogy a jelölet egyértelműen azonosítható legyen. Ha a jelenlevők közül csak egy férfi van, és őt akarjuk megszólítani, az *úr* mint titulus akár tekinthető a tulajdonnév részének.

3.3 Törvények

Habár az átiratokban is megjelent, a törvényszövegek részkorpuszára volt igazán jellemző a jogszabályhelyek tételes említése, mint például:

- *Btk. 236 § (1)* (2012. évi C. törvény a Büntető Törvénykönyvről)
- *Ptk 6: 494§ (2)* (2013. évi V. törvény a Polgári Törvénykönyvről)
- *Tht 1§ (2)* (2003. évi CXXXIII. törvény a társasházakról)

Ezekben az esetekben az adott törvény, rendelet vagy jogszabály neve (*Btk*. stb.) volt tulajdonnévként jelölve, a pontos szakasz, bekezdés, pont stb. nem, mivel az csak a jogszabályi hivatkozások belső struktúrája szempontjából fontos részinformáció.

A részkorpuszon elért alacsonyabb F-érték a magyarlanc esetében két fő hibatípusra vezethető vissza. Az egyik hibatípus a jogszabályok tagolása során alkalmazott megoldások téves besorolása volt, amelynek a [10] jellemző esete.

A névelem-felismerő esetében a legszembeötlőbb hiba egyértelműen az *1952. évi III. Törvény a Polgári perrendtartásról* esete volt, ahol egyetlen tokent sem jelölt névelemként a szoftver. Habár az elnevezés szerkezete meglehetősen tipikus a törvények nevezéktanában (YYYY évi RÓMAI SZÁM törvény a … -ról/ről), amely akár reguláris kifejezéssel is leírható, ám ha a rendszer eredeti tanítókorpuszában nem szerepelt hasonló adat, akkor nem bejósolható.

[10] „(3a) A (2) bekezdés szerinti nyilvántartás a (3) bekezdés c) – i) pontja szerinti adatok…”

1	(	(	PUNCT	_	0 PUNCT
2	3a	3a	PROPN	Case=Nom|Number=Sing	8 ATT
3	)	)	PUNCT	_	2 PUNCT
4	A	A	PROPN	Case=Nom|Number=Sing	2 COORD
5	(	(	PUNCT	_	6 PUNCT

3.4 Függőségi nyelvtani elemzés

A teljes kép érdekében az utolsó tervezett szint a magyarlanc függőségi nyelvtani címkézésének vizsgálata volt.

A 4. táblázat a szintaktikai viszonyokban fellelhető NE élcímke jelenléte szerint számolt tokenszintű értékeket mutatja. Ugyanakkor fontos megjegyezni, hogy a fenti értékek nem tekinthetők reprezentatívnak, mivel a kézi annotálás során a vártnál kevesebb, több tokenből álló névelemet sikerült csak találni a jelen tanulmányhoz elemzett korpuszrészben. Az 5. táblázat a kézi annotációban megjelenő névelemeket összegzi.

A második oszlop a magyarlanc által NE élcímkével összekapcsolt, több tokenből álló névelemeket jelöli, a harmadik oszlop a kézzel annotáltakat.

Az adatok ritkasága (mind a manuális, mind az automatikus annotáció alapján) egyelőre nem teszi alkalmassá azokat egy pontos statisztikai elemzésre, így e téren további vizsgálatok szükségesek, a szintaktikai elemzés hatékonysága nem ítélhető meg pontosabban.

Részkorpusz		
Fórumok	Pontosság	80,75
	Fedés	70,00
	F-érték	74,99
Átiratok	Pontosság	76,92
	Fedés	66,67
	F-érték	71,43
Törvényszövegek	Pontosság	100,00
	Fedés	57,14
	F-érték	72,73

4. táblázat. A magyarlanc szintaktikai annotációjának eredményessége a tulajdonnév-felismerésben

Részkorpusz	A magyarlanc szintaktikai elemzésében megjelenő névelemek	Manuálisan jelölt több tokenes névelemek
Átiratok	23	63
Fórumok	15	20
Törvényszövegek	7	7

5. táblázat. Referenciaadatok a szintaktikai elemzés szintjéről

4 Következtetések

A jelen tanulmányban a névelem-felismerés hatékonyságát vizsgáltam meg magyar nyelvű jogi szövegek esetében. Ennek érdekében egy tulajdonnév-felismerő szoftver és a magyarlanc kimenetét hasonlítottam össze manuális annotációval a Miskolc Jogi Korpusz kiválasztott részletein. A megvizsgált két szoftver kimenetében azonosítottam a legtipikusabb hibaforrásokat a megfelelő szófaji címkézés (magyarlanc) és a névelem-felismerés (tulajdonnév-felismerő) szempontjából.

A feladat doménspecifikusságával kapcsolatban a vizsgálatok alátámasztották, hogy mindhárom elemzett részkorpusz/szövegtípus esetében vannak olyan sajátosságok, amelyek kezelése még nem megoldott a vizsgált rendszerekben. A feltárt hibaforrások alapján a szükséges megszorítások visszaillesztésével mindkét rendszer hatékonyabbá és pontosabbá tehető, valamint esetleges későbbi névelem-felismerő rendszerek hatásfoka is javítható az eredmények figyelembevételével.

A vizsgálat során kiderült, hogy a több tokenes névelemek sokkal alacsonyabb arányban vannak jelen a kiválasztott szövegekben, mint ami egy pontos statisztikai elemzéshez szükséges. Ennek megoldásához szükséges a vizsgálat megismétlése egy nagyobb szövegtesten.

Távlati cél az elemzés kiterjesztése a Miskolc Jogi Korpusz mind a hat részkorpuszára annak érdekében, hogy a jogi szövegeken végzett névelem-felismerés azok minél szélesebb spektrumán tesztelhető legyen.

Irodalom

Csendes, D. – Csirik, J. – Gyimóthy, T. 2004. The Szeged Corpus: A POS Tagged and Syntactically Annotated Hungarian Natural Language Corpus. In: Sojka, P. – Kopecek, I. – Pala, K. (szerk.). *Lecture Notes in Artificial Intelligence* (Subseries of Lecture Notes in Computer Science) 3206: 41– 47.

Deme L. – Fábián P. – Tóth E. (szerk.) 1999. *Magyar helyesírási szótár*. Budapest: Akadémiai Kiadó.

Einat, M. – Richard, C. W. –William, W. C. 2005. Extracting personal names from emails: Applying named entity recognition to informal text. *Human Language Technology / Empirical Methods in Natural Language Processing:* 443–450.

Farkas T. 2007. A tulajdonnevek fordíthatóságáról és napjaink fordítási hibáiról, közszók és tulajdonnevek példáján. *Névtani Értesítő* 29: 167–188.

Kripke, S. 1980. *Naming and Necessity*. Cambridge, Massachusetts: Harvard University Press.

Laczkó K. – Mártonfi A. 2004. *Helyesírás*. Budapest: Osiris.

Lenci, A. – Montemagni, S. – Pirrelli, V. – Venturi, G. 2009. Ontology learning from italian legal texts. In: *Proceeding of the 2009 Conference on Law, ontologies and the Semantic Web: Channelling the Legal information Flood:* 75-94.

Linguistic Data Consortium. ACE (automatic content extraction) English annotation guidelines for entities. 2006. Elérhető: https://www.ldc.upenn.edu/ collaborations/past-projects/ace, Version 5.6.6 2006.08.01. (letöltés ideje: 2019. 01. 10.)

Móra Gy. – Vincze V. – Zsibrita J. 2011. Szófaji kódok és névelemek együttes osztályozása. In: Tanács A. – Vincze V. (szerk.) *VIII. Magyar Számítógépes Nyelvészeti Konferencia.* Szeged: Szegedi Tudományegyetem. 131–142.

Quaresma, P. – Gonçalves, T. 2010. Using Linguistic Information and Machine Learning Techniques to Identify Entities from Juridical Documents. In: Francesconi, E. – Montemagni, S. – Peters, W. – Tiscornia, D. (szerk.) *Number 6036 in Lecture Notes in AI.* Springer-Verlag. 44–59.

Simon E. – Farkas R. – Halácsy P. – Sass B. – Szarvas Gy. – Varga D. 2006. A HunNER korpusz. In: Tanács A. – Vincze V. (szerk.) *IV. Magyar Számítógépes Nyelvészeti Konferencia*. Szeged: Szegedi Tudományegyetem. 373–376.

Simon E. 2008. Nyelvészeti problémák a tulajdonnév-felismerés területén. In: Sinkovics B. (szerk.) *LingDok 7. Nyelvész-doktoranduszok dolgozatai*. Szeged: Szegedi Tudományegyetem, Nyelvtudományi Doktori Iskola.

Simon, E. 2017. The Definition of Named Entities. In: Gyuris, B. – Mády, K. – Recski, G. (szerk.) *K + K = 120. Papers dedicated to László Kálmán and András Kornai on the occasion of their 60th birthdays*. Budapest: Research Institute for Linguistics, Hungarian Academy of Sciences (RIL HAS): 1-18.

Simpson, J. A. – Weiner, E. S. C. 1989. *The Oxford English Dictionary*. Oxford: Clarendon Press.

Surdeanu, M. – Nallapati, R. – Manning, C. D. 2010. Legal claim identification: Information extraction with hierarchically labeled data. In: *Proceedings of the LREC 2010 Workshop on the Semantic Processing of Legal Texts (SPLeT).* Malta, May 2010

Szarvas Gy. – Farkas R. – Kocsor A. 2006. A Multilingual Named Entity Recognition System Using Boosting and C4.5 Decision Tree Learning Algorithms. In: *The Ninth International Conference on Discovery Science LNAI 4265:* 267–278.

Tikk D. – Farkas R. – Kardkovács Zs. T. – Kovács L. – Répási T. – Szarvas Gy. – Szaszkó S. – Vázsonyi M. 2006. *Szövegbányászat.* Budapest: Typotex.

Vincze V. – Farkas R. 2012. Tulajdonnevek a számítógépes nyelvészetben. *Általános Nyelvészeti Tanulmányok XXIV*: 97–119.

Vincze V. 2018. A Miskolc Jogi Korpusz nyelvi jellemzői. In: Szabó M. (szerk.) *A törvény szavai*. Miskolc: Bíbor Kiadó. 9-36.

Yangjie, Y. – Aixin, S. 2016. Mobile phone name extraction from internet forums: a semi-supervised approach. *World Wide Web* 19(5): 783–805.