Domének közti hasonlóságok és különbségek a szófajok és szintaktikai viszonyok eloszlásában

Vincze Veronika[1,2]

[1] MTA-SZTE Mesterséges Intelligencia Kutatócsoport
[2] Universität Trier, Linguistische Datenverarbeitung
vinczev@inf.u-szeged.hu

Kivonat: Ebben a cikkben a szófajok és szintaktikai relációk eloszlását vizsgáljuk különböző doménekben. A vizsgálat alapjául a Szeged Dependencia Treebank szolgál. Eredményeink alapján a szövegek témája (doménje) befolyásolja a szófajok, illetőleg a szövegszavak közti szintaktikai relációk eloszlását, így a domének között hasonlóságok és különbségek figyelhetők meg e téren, ami jelentőséggel bír különféle számítógépes nyelvészeti alkalmazásokban is, például a szófaji egyértelműsítők és a dependenciaelemzők hatékonyságának növelésében.

1 Bevezetés

A különféle nyelvi jelenségek eloszlásának kvantitatív vizsgálata nagy figyelmet kapott az utóbbi években. Számos nyelvben vizsgálták a szófajok és morfológiai jegyek eloszlását, l. például [3,9,12,13,14], mindemellett a szintaktikai viszonyok eloszlását is elemezték a kvantitatív szintaxis mint elmélet keretein belül [4,5,6,7,8].

A ragozó nyelvekben általában megfigyelhető a morfológia és a szintaxis szoros összefonódása, hiszen a nyelvtani relációk nagy részét morfológiai eszközök segítségével lehet kifejezni. Így e nyelvek kitűnő táptalajt biztosítanak a kvantitatív morfológiai és szintaktikai vizsgálatok számára. Például Köhler [6] a Szeged Treebank egy részén vizsgálta a nyelvtani viszonyok eloszlását, Väyrynen, Noponen és Seppänen [10] pedig a finnben elemzik a szemantikai viszonyokat.

Ebben a munkában a szófajok és szintaktikai viszonyok eloszlását vizsgáljuk különböző doménekhez tartozó magyar szövegekben. A vizsgálat alapjául a Szeged Dependencia Treebank [11] szolgál, amely hat különböző tématerületről tartalmaz kézzel annotált szövegeket: üzleti rövidhírek, újságcikkek, iskolai fogalmazások, szépirodalom, jogi és számítógépes szövegek. Kiinduló feltételezésünk szerint a szövegek témája (doménje) befolyásolja a szófajok, illetőleg a szövegszavak közti szintaktikai relációk eloszlását, így a domének között hasonlóságok és különbségek figyelhetők meg e téren, ami jelentőséggel bír különféle számítógépes nyelvészeti alkalmazásokban is, többek között a szófaji egyértelműsítők és a dependenciaelemzők hatékonyságának növelésében.

Vizsgálataink során az alábbi kérdésekre keressük a választ:

- Milyen jellemző eloszlási minták találhatók a magyar nyelvben a szófajokra és a szintaktikai viszonyokra nézve?
- A fenti eloszlások mennyire tekinthetők doménfüggőnek, illetve általánosnak?

A statisztikai adatok bemutatása és értelmezése mellett az eredmények nyelvészeti indoklására is törekszünk a cikkben.

2 A vizsgált korpusz

Vizsgálataink alapjául a Szeged Dependencia Treebank [11] szolgál. 82 000 mondatot, 1,5 millió szövegszót és 230 000 írásjelet tartalmaz hat doménből (iskolai fogalmazások, számítógépes szövegek, irodalom, jogi szövegek, újságcikkek és üzleti rövidhírek). A korpusz kézzel ellenőrzött morfológiai és szófaji, valamint szintaktikai (függőségi) elemzést is tartalmaz. A korpusz adatait az 1. táblázat foglalja össze.

1. táblázat: A Szeged Dependencia Treebank adatai.

	iskolás	számítógép	irodalom	jog	újság	rövidhír	összesen
Mondat	24 720	9 627	18 558	9 278	10 210	9 574	81 967
Írásjel	59 419	31 241	47 990	33 515	32 880	25 712	230 757
Szövegszó	283 591	183 562	189 751	225 207	190 406	201 527	1 504 801
Átlagos mondathossz	11,472	19,067	10,225	24,273	18,649	21,049	18,359

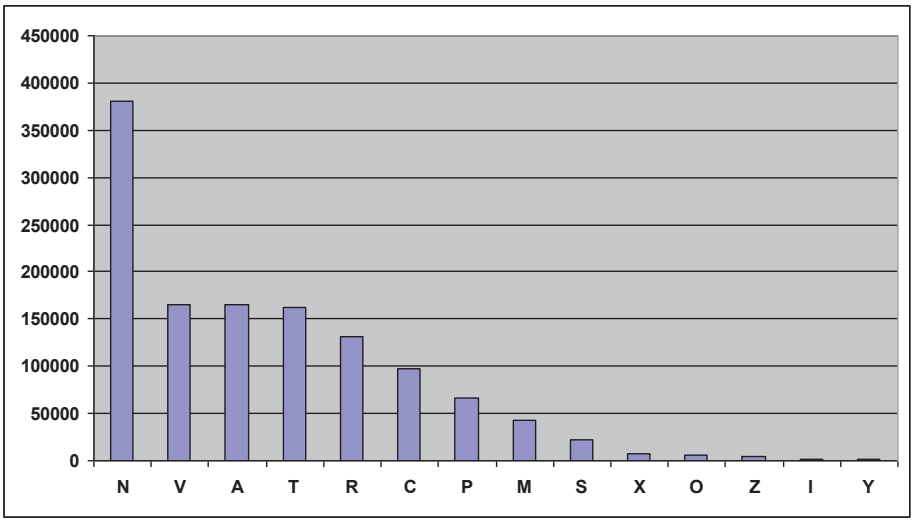
A következőkben a szófajok és függőségi viszonyok eloszlását vizsgáljuk meg a korpuszban és a különböző doménekben.

3 A szófajok eloszlása

A Szeged Dependencia Treebank az MSD morfológiai kódrendszert [2] használja a szófajok kódolására. Segítségével lehetőség nyílik mind a szófajok, mind a morfológiai jellemzők (például idő, mód, szám, személy, esetrag stb.) kódolására. Ebben a munkában kizárólag a fő szófaji információkra összpontosítunk, tehát minden token esetében csak a fő szófajt (főnév, ige, melléknév stb.) vesszük figyelembe, a finomabb morfológiai megkülönböztetésektől most eltekintünk, illetve az írásjeleket sem vonjuk be vizsgálataink körébe.

3.1 A szófajok eloszlása a teljes korpuszban

Az 1. ábrán látható a szófajok teljes korpuszbeli eloszlása. Az x tengely mutatja a szófajokat, az y tengelyen pedig a gyakorisági értékek láthatók.



1. ábra: A szófajok eloszlása a Szeged Dependencia Treebankben.

Amint az ábra is mutatja, a leggyakrabban előforduló szófajok a főnév, ige és melléknév. Ez összhangban van azzal az elvárással, hogy a szemantikai jelentéssel bíró lexikális elemek a leggyakoribbak. A névelők szintén gyakoriak, ami feltehetőleg annak köszönhető, hogy a főnevek igen gyakran szerepelnek névelő kíséretében a treebankben. Az ismeretlen szavak, rövidítések, helyesírási hibás szóalakok és a nyílt tokenosztályba tartozó szavak (X, Y, Z és O kódok) viszonylag ritkán fordulnak elő a korpuszban: összesítve a szavak 5,89%-át alkotják.

3.2 A szófajok eloszlása az egyes doménekben

A szófajok doménenkénti eloszlása a 2. táblázatban látható. A domének közti hasonlóságok részletesebb vizsgálatához felállítottuk az egyes szófajok gyakorisági rangsorát is, melyet a 3. táblázat szemléltet.

A szófajok eloszlásának vizsgálatához, illetve a domének közti hasonlóságok és különbségek megállapításához a Kendall-együtthatót (W) alkalmaztuk, amely a vizsgált elemek, jelen esetben a szófajok gyakorisági rangsorát felállítva mutatja meg, mennyire homogének a vizsgált szövegek. A Kendall-együttható értéke alapján a szövegek homogének (W = 0.9248), az eredmények szignifikánsak (DF=13, χ^2 = 154.571429). Azonban különbségek is megfigyelhetők az egyes domének között: míg a főnevek és igék az első két helyen szerepelnek az iskolás és az irodalmi szövegekben, addig a többi doménen nagy különbségek figyelhetők meg, lévén a főnév a leggyakoribb szófaj, ám az ige csak a negyedik-ötödik a gyakorisági rangsorban. A határozószavak viszonylag gyakran fordulnak elő az irodalmi és az iskolás szövegekben, ezzel szemben a melléknevek kevésbé gyakoriak, különösképpen a többi doménnel összevetve, ahol is a második vagy harmadik leggyakoribb szófajnak tekinthetők.

2. táblázat: A szófajok eloszlása doménenként.

	iskolás	irodalom	jog	újság	rövidhír	számítógép	összesen
főnév	56106	44737	78546	61902	79591	60201	381083
ige	58702	34805	15557	20751	16913	18958	165686
melléknév	20500	18403	40698	26955	32124	25887	164567
névelő	31253	19793	31495	25196	29027	26160	162924
határozószó	47322	29233	12725	17988	9760	14934	131962
kötőszó	29322	17348	15854	13695	7135	13522	96876
névmás	21081	14516	9549	8916	3620	9072	66754
számnév	7000	2374	6859	7032	14556	4817	42638
névutó	3286	2487	4268	3593	4933	2928	21495
ismeretlen	1026	1532	871	659	1297	2066	7451
nyílt tokenosztály	150	40	3663	284	779	827	5743
helyesírási hibás	2470	398	515	135	336	156	4010
indulatszó	738	814	6	135	5	114	1812
rövidítés	304	141	885	35	8	90	1463

3. táblázat: A szófajok gyakorisági rangsora doménenként.

	iskolás	irodalom	jog	újság	rövidhír	számítógép
ige	1	2	5	4	4	4
főnév	2	1	1	1	1	1
határozószó	3	3	6	5	6	5
névelő	4	4	3	3	3	2
kötőszó	5	6	4	6	7	6
névmás	6	7	7	7	9	7
melléknév	7	5	2	2	2	3
számnév	8	9	8	8	5	8
névutó	9	8	9	9	8	9
helyesírási hibás	10	12	13	12	12	12
ismeretlen	11	10	12	10	10	10
indulatszó	12	11	14	13	14	13
rövidítés	13	13	11	14	13	14
nyílt tokenosztály	14	14	10	11	11	11

Az egyes domének közti hasonlóságok és különbségek további vizsgálatához minden egyes doménpárra kiszámoltuk a Kendall-együttható értékét. Az eredményeket a 4. táblázat mutatja (minden eredmény szignifikáns).

4. táblázat: A domének hasonlósága a szófajok eloszlása terén.

	újság	rövidhír	számítógép	irodalom	iskolás	jog
újság		0,9802	0,9978	0,9626	0,9363	0,9758
rövidhír	0,9802		0,9780	0,9319	0,9055	0,9626
számítógép	0,9978	0,9780		0,9648	0,9429	0,9736
irodalom	0,9626	0,9319	0,9648		0,9824	0,9253
iskolás	0,9363	0,9055	0,9429	0,9824		0,9033
jog	0,9758	0,9626	0,9736	0,9253	0,9033	

A fenti eredmények alapján a szófajok eloszlása terén a két leghasonlóbb domén az újságcikkek és a számítógépes szövegek (W = 0.9978). A hasonlóságot magyarázhatja az a tény, hogy a számítógépes szövegek egy része valójában egy számítógépes magazinból származik, így egyaránt rendelkezik a számítógépes szövegekre, illetve az újságcikkekre jellemző sajátosságokkal. Az iskolai fogalmazások a szépirodalmi szövegekhez hasonlítanak leginkább (W = 0.9824), azonban eltérnek az üzleti hírektől és a jogi szövegektől. A legnagyobb különbséget a jogi és iskolás szövegek között figyelhetjük meg, amely feltehetőleg a domének között húzódó alapvető stilisztikai különbségeknek köszönhető. Az iskolai fogalmazásokban a mondatok jóval rövidebbek, továbbá többnyire a szerzőjükkel megtörtént eseményeket írnak le, az események leírása pedig az igék fokozottabb használatát követeli meg. Ezzel szemben a jogi szövegek nem annyira eseményeket, hanem inkább tényeket és állapotokat írnak le, így kevesebb igét is tartalmaznak.

4 A szintaktikai viszonyok eloszlása

A Szeged Treebank 2.0-ban található igék és bővítményeik közti szintaktikai viszonyokat először automatikusa konvertálták függőségi viszonyokra [1], majd ezeket kézzel ellenőrizve és javítva állt elő a Szeged Dependencia Treebank [11]. A függőségi viszonyok eloszlását a szófajokéhoz hasonlóan elemezzük a következőkben.

4.1 A szintaktikai viszonyok eloszlása a korpuszban

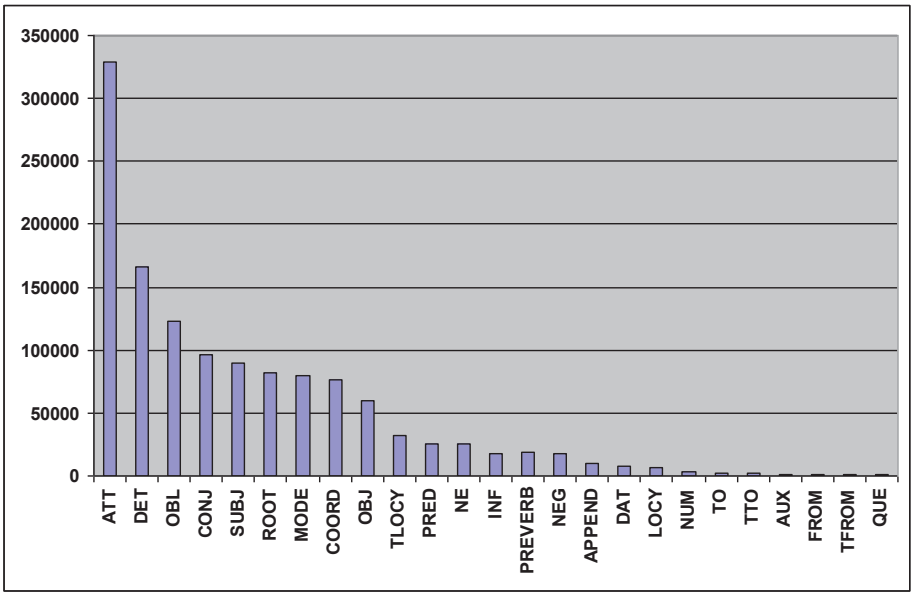
A szintaktikai viszonyok eloszlásának tanulmányozásához először is megszámoltuk, milyen viszonyok húzódnak az egyes szavak között a korpuszban (az írásjeleket ismét figyelmen kívül hagyva). Az eredményeket az 5. táblázat mutatja.

5. táblázat: A szintaktikai viszonyok eloszlása doménenként.

	iskolás	irod.	jog	újság	rövidhír	sz.gép	összesen	%
ATT	46046	35386	80603	51891	66889	48221	329036	25,82
DET	32044	20477	32187	25536	29192	26618	166054	13,03
OBL	23502	15771	23182	18044	25143	17168	122810	9,64
CONJ	29236	17108	15584	13569	7114	13296	95907	7,53
SUBJ	20650	15784	12094	14144	14181	12615	89468	7,02
ROOT	24723	18564	9284	10210	9577	9658	82016	6,44
MODE	24253	15320	11051	11432	7533	10345	79934	6,27
COORD	20553	12852	11533	11600	7959	12073	76570	6,01
OBJ	14077	9547	9609	9553	7180	9433	59399	4,66
TLOCY	12533	5983	2257	4349	3854	2897	31873	2,50
PRED	8014	5045	3655	3348	1696	3949	25707	2,02
NE	764	1187	1344	4535	11820	5447	25097	1,97
INF	7847	2796	2437	1877	618	2513	18088	1,42
PREVERB	4357	3214	2698	2620	2668	2719	18276	1,43
NEG	5734	4009	2788	2406	862	1722	17521	1,38
APPEND	1157	913	2769	1248	1188	2156	9431	0,74
DAT	2259	1401	1365	1397	759	1108	8289	0,65
LOCY	2912	2041	237	779	328	616	6913	0,54
NUM	5	0	99	593	2277	365	3339	0,26
TO	1198	764	66	278	245	142	2693	0,21
TTO	654	379	130	315	166	177	1821	0,14
AUX	318	476	36	146	25	53	1054	0,08
FROM	285	262	91	165	32	126	961	0,08
TFROM	213	243	13	214	197	84	964	0,08
QUE	202	209	106	155	20	104	796	0,06

A függőségi viszonyok eloszlása a teljes korpuszon a 2. ábrán látható. Amint láthatjuk, a leggyakoribb reláció az ATT, mely az összes szintaktikai viszony körülbelül negyedét adja. Az ATT általános módosító viszonynak tekinthető, melybe a jelzői és alárendelői szerepek egyaránt beletartoznak, vagyis szavakat és tagmondatokat egyaránt összekapcsolhat. Ennek a ténynek feltehetőleg fontos szerepe van a reláció gyakori előfordulásában. Mivel a főnevek általában egy névelővel együtt fordulnak elő, a DET reláció is viszonylag gyakran szerepel a szövegekben (13%). A harmadik leggyakoribb szintaktikai viszony, az OBL számos magyar esetragot foglal magába, emiatt a gyakorisága is megfelel ezen esetragok összesített gyakoriságának. A kötőszavak

is viszonylag gyakoriak a korpuszban, ami azt sugallja, hogy számos alá- és mellérendelés található a korpuszban, mind tagmondatok, mind szavak szintjén.



2. ábra: A függőségi viszonyok eloszlása a Szeged Dependencia Treebankben.

4.2 A szintaktikai viszonyok eloszlása az egyes doménekben

Szerettük volna megvizsgálni azt is, hogy az egyes doménekre nézve milyen sajátosságok mutathatók ki a szintaktikai viszonyok eloszlására nézve, illetőleg milyen hasonlóságok és különbségek figyelhetők meg a domének között. A függőségi viszonyok rangsorát a 6. táblázat szemlélteti.

A táblázatbeli adatok alapján szintén kiszámoltuk a Kendall-együttható értékét a korpuszra, és azt találtuk, hogy az adatok homogének (W= 0.9321). Az eredmények szignifikánsak (DF=25, $\chi^2 = 134.221538$).

A domének közti összehasonlítás számos érdekességet tartogat. Először is az üzleti rövidhírek bővelkednek a több tagból álló tulajdonnevekben és számnevekben, hiszen számos, cégekkel kapcsolatos pénzügyi hírt tartalmaznak. Így az NE és NUM relációk is igen gyakran fordulnak elő ezen a doménen. Másodszor, a jogi szövegekben igen nagy mennyiségben fordul elő az APPEND reláció, mely a mondatba szorosan nem tartozó közbevetéseket jelzi. A jogi szövegekben számos utalás található törvényekre, paragrafusokra, melyek nem képezik a mondat szerves részét, így az APPEND relációval kapcsolódnak a többi elemhez.

6. táblázat: A szintaktikai viszonyok gyakorisági rangsora doménenként.

	iskolás	irodalom	jog	újság	rövidhír	számítógép
ATT	1	1	1	1	1	1
DET	2	2	2	2	2	2
CONJ	3	4	4	5	10	4
ROOT	4	3	9	8	6	8
MODE	5	7	7	7	8	7
OBL	6	6	3	3	3	3
SUBJ	7	5	5	4	4	5
COORD	8	8	6	6	7	6
OBJ	9	9	8	9	9	9
TLOCY	10	10	15	11	11	12
PRED	11	11	10	12	14	11
INF	12	14	14	15	18	14
NEG	13	12	11	14	16	16
PREVERB	14	13	13	13	12	13
LOCY	15	15	18	18	19	18
DAT	16	16	16	16	17	17
TO	17	19	23	21	20	21
APPEND	18	18	12	17	15	15
NE	19	17	17	10	5	10
TTO	20	21	19	20	22	20
AUX	21	20	24	25	24	25
FROM	22	22	22	23	23	22
TFROM	23	23	25	22	21	24
QUE	24	24	20	24	25	23
NUM	25	25	21	19	13	19

A függőségi viszonyok esetében szintén a Kendall-együtthatót alkalmaztuk a domének közti hasonlóságok felderítésére, az eredmények ez esetben is szignifikánsak. A 7. táblázatban látható eredmények alapján az egymáshoz leghasonlóbb doménpárok az újságcikkek és a számítógépes szövegek, illetőleg a fogalmazások és a szépirodalmi szövegek (W = 0.9965 és 0.995, rendre). A legnagyobb eltérés pedig a fogalmazások és az üzleti rövidhírek között mutatkozik (W = 0.8973), hasonlóan a szófajok eloszlásához, ami a stilisztikai eltérésekre vezethető vissza.

7. táblázat: A domének hasonlósága a szintaktikai viszonyok eloszlása terén.

	újság	rövidhír	számítógép	irodalom	iskolás	jog
újság		0,9762	0,9962	0,9665	0,9577	0,9723
rövidhír	0,9762		0,9708	0,9158	0,8973	0,9227
számítógép	0,9962	0,9708		0,9627	0,9565	0,9777
irodalom	0,9665	0,9158	0,9627		0,995	0,9627
iskolás	0,9577	0,8973	0,9565	0,995		0,9588
jog	0,9723	0,9227	0,9777	0,9627	0,9588	

5 Az eredmények értelmezése

Az eredmények alapján kirajzolódnak az alkorpuszok (illetve domének) közti hasonlóságok, illetve távolságok. Mind a szófajok, mind a szintaktikai viszonyok szempontjából a legnagyobb hasonlóságot az újságcikkek és a számítógépes szövegek mutatták. A hasonlóságot magyarázhatja, hogy a számítógépes szövegek jó része valójában egy számítástechnikai témájú magazinból származik, így a nyelvezetük erősen hasonlít a sajtónyelvre, csakúgy, mint az újságcikkek nyelvezete. A szépirodalmi szövegek és az iskolai fogalmazások közti hasonlóság azzal magyarázható, hogy mindkét esetben történetek elbeszéléséről van szó, tehát az elbeszélő stílus jegyei figyelhetők meg mindkét domén szövegeiben. Az üzleti hírek, illetve a jogi szövegek pedig egyedi nyelvi jellemzőkkel bírnak.

6 Az eredmények alkalmazása a számítógépes nyelvészetben

A vizsgálat eredményei számos területen hasznosíthatók a számítógépes nyelvészetben. Mivel a magyar szófaji egyértelműsítők és szintaktikai elemzők nagy része a Szeged (Dependencia) Treebanket használja tanító adatbázisként (pl. [15]), a domének közti különbségek jelentősen befolyásolhatják azt, hogy mely részkorpuszokat érdemes tanító adatbázisként kiválasztani egy adott elemzendő szöveghez. Például egy regény szintaktikai elemzésekor valószínűleg az iskolai fogalmazások és a szépirodalmi szövegek unióján tanított elemző éri el a legjobb eredményt. A domének közti hasonlóságoknak és különbségeknek a részletes elemzése megnyitja az utat a különféle doménadaptációs technikáknak a szófaji egyértelműsítésben és szintaktikai elemzésben való alkalmazása előtt is. Végül az eloszlási minták, pontosabban a domének közti hasonlóságok és különbségek elemzése a dokumentumosztályozásban is hasznosítható.

A fentiek alátámasztására végeztünk egy kísérletet. Az *Egri csillagok* című regényt morfológiailag elemeztük, szófajilag egyértelműsítettük, majd dependenciaelemzésnek vetettük alá a *magyarlanc* elemzővel [15]. Az elemzések alapján meghatároztuk a szófajok és a szintaktikai viszonyok eloszlását, majd az eloszlási mintákat összevetet-

tük a Szeged Dependencia Treebank minden egyes doménjével, és minden párra kiszámítottuk a Kendall-együtthatót. A szignifikáns eredmények a 8. táblázatban láthatók.

8. táblázat: Az *Egri csillagok* és a domének hasonlósága a szófajok és a szintaktikai viszonyok eloszlása terén.

	szófaj	dependencia
iskolás	0,9868	0,9865
irodalom	0,9934	0,9946
újság	0,9604	0,9638
számítógép	0,9670	0,9546
rövidhír	0,9341	0,9212
jog	0,9297	0,9527

Az eredmények azt mutatják, hogy mind a szófajok eloszlása, mind a függőségi viszonyok eloszlása terén az *Egri csillagok* a legnagyobb fokú hasonlóságot az irodalmi szövegekkel mutatja, illetve a második leginkább hasonló domén az iskolás szövegek. Vagyis ha nem tudnánk a szöveg műfaját, akkor is az eredmények alapján a legnagyobb valószínűséggel irodalmi szövegnek titulálnánk egy dokumentumosztályozási feladat során. Mivel tudjuk, hogy az *Egri csillagok* is az irodalmi művek sorába tartozik, így ez a döntés helytálló lenne. A fenti számok ugyanakkor megerősítik azt a korábbi eredményünket is, miszerint az iskolás és az irodalmi szövegek hasonlítanak egymáshoz.

7 Összegzés

A cikkben a szófajok és szintaktikai relációk eloszlását vizsgáltuk különböző doménekben. A vizsgálat alapjául a Szeged Dependencia Treebank szolgált. Eredményeink alapján a szövegek doménje befolyásolja a szófajok, illetőleg a szövegszavak közti szintaktikai relációk eloszlását, így a domének között hasonlóságok és különbségek figyelhetők meg e téren. Mind a szófajok, mind a szintaktikai viszonyok szempontjából a legnagyobb hasonlóságot az újságcikkek és a számítógépes szövegek mutatták. Az irodalmi és az iskolás szövegek szintén hasonlítanak egymásra, az üzleti hírek és a jogi szövegek pedig önálló sajátságokkal bírnak.

A jövőben szeretnénk megvizsgálni a szófajok és a szintaktikai jellemzők eloszlását más szövegtípusokban is, illetőleg a fenti eredmények felhasználásával szeretnénk doménadaptációs kísérleteket végezni a szófaji egyértelműsítés és dependenciaelemzés területén.

Köszönetnyilvánítás

Szeretnék köszönetet mondani Reinhard Köhlernek a munkámat segítő számos hasznos tanácsáért és értékes megjegyzéséért.

A kutatás – részben – az A/11/83421 jelű fiatal kutatói ösztöndíj keretében a Deutscher Akademischer Austauschdienst támogatásával, illetve a futurICT.hu nevű, TÁMOP-4.2.2.C-11/1/KONV-2012-0013 azonosítószámú projekt keretében az Európai Unió és az Európai Szociális Alap társfinanszírozása mellett valósult meg.

Hivatkozások

1. Alexin Z.: A frázisstrukturált Szeged Treebank átalakítása függőségi fa formátumra. In: V. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2007). Szegedi Tudományegyetem, Szeged (2007) 263–266
2. Alexin, Z., Csirik, J., Gyimóthy, T., Bibok, K., Hatvani, Cs., Prószéky, G., Tihanyi, L.: Annotated Hungarian National Corpus. In: Proceedings of EACL (2003) 53–56
3. Best, K.-H.: Word class frequencies in contemporary German short prose texts. Journal of Quantitative Linguistics, Vol. 1 (1994) 144–147
4. Cech, R., Pajas, P., Mačutek, J.:. Full valency. Verb valency without distinguishing complements and adjuncts. Journal of Quantitative Linguistics, Vol. 17, No. 4 (2010) 291–302
5. Köhler, R.: Syntactic Structures. Properties and Interrelations. Journal of Quantitative Linguistics, Vol. 6, No. 1 (1999) 46–57
6. Köhler, R.: Quantitative Syntax Analysis. de Gruyter, Berlin, New York (2012)
7. Liu, H.: Probability distribution of dependency distance. Glottometrics, Vol. 15 (2007) 1–12
8. Liu, H.: Probability distribution of dependencies based on Chinese dependency treebank. Journal of Quantitative Linguistics, Vol. 16, No. 3 (2009) 256–273
9. Tuzzi, A., Popescu, I.-I., Altmann, G.: Quantitative analysis of Italian texts. RAM, Lüdenscheid (2010)
10. Väyrynen, P. A., Noponen, K., Seppänen, T.: Preliminaries to Finnish word prediction. Glottotheory, Vol. 1 (2008) 65–73
11. Vincze, V. Szauter, D., Almási, A., Móra, Gy., Alexin, Z., Csirik, J.: Hungarian Dependency Treebank. In: Proceedings of the Seventh Conference on International Language Resources and Evaluation (2010)
12. Vulanović, R., Köhler, R.: Word order, marking, and Parts-of-Speech Systems. Journal of Quantitative Linguistics, Vol. 16, No. 4 (2009) 289–306
13. Ziegler, A.: Word class frequencies in Brazilian-Portuguese press texts. Journal of Quantitative Linguistics, Vol. 5 (1998) 269–280
14. Ziegler, A.: Word class frequencies in Portuguese press texts. In: Uhlířová, L., Wimmer, G., Altmann, G., Köhler, R. (eds.): Text as a linguistic paradigm: levels, constituents, constructs. Festschrift in honour of Luděk Hřebíček. Wissenschaftlicher Verlag Trier, Trier (2001) 295–312
15. Zsibrita J., Vincze V., Farkas R.: magyarlanc 2.0: szintaktikai elemzés és felgyorsított szófaji egyértelműsítés. In: Tanács A., Vincze V. (szerk.): IX. Magyar Számítógépes Nyelvészeti Konferencia. Szegedi Tudományegyetem, Szeged (2013) 368–374