Magyar nyelvű webes szövegek morfológiai és szintaktikai annotációja

Vincze Veronika[1,2], Varga Viktor[1], Papp Petra Anna[1],
Simkó Katalin Ilona[1], Zsibrita János[1], Farkas Richárd[1]

[1]Szegedi Tudományegyetem, Informatikai Tanszékcsoport
Szeged, Árpád tér 2.
{vinczev,zsibrita,rfarkas}@inf.u-szeged.hu,
{viktor.varga.1991,papp.petra.anna,kata.simko}@gmail.com
[2]MTA-SZTE Mesterséges Intelligencia Kutatócsoport
Szeged, Tisza Lajos körút 103.

Kivonat Cikkünkben bemutatjuk az első magyar, kézzel annotált, webes szövegeket tartalmazó korpuszt, melyet tesztadatbázisnak szánunk a webes szövegekre optimalizált nyelvi elemzőink fejlesztéséhez. A korpusz morfológiai és (összetevős és függőségi szemléletű) szintaktikai elemzést, valamint szemantikai és diskurzusbeli bizonytalan kifejezések annotációját tartalmazza. Beszámolunk a magyarlanc elemző webes szövegekre történő adaptálási kísérleteiről is.

1. Bevezetés

Az interneten fellelhető szövegek mint könnyen hozzáférhető és már-már kifogyhatatlan adatforrás jelentősége már régóta a köztudatban van, sok kutatás irányította rá a tárgyát. Az utóbbi években a webes szövegek legnagyobb részét már maguk a felhasználók adják közre, legyen az a produktum blogbejegyzés, fórumokon való beszélgetés, bejegyzésekhez tartozó véleménynyilvánítás (komment) vagy mikroblogbejegyzés (pl. Twitter). Egyértelmű tehát, hogy a „webes szöveg" korántsem homogén szövegtípus, felhasználónként, csoportonként és műfajonként változik a stílus, a megszerkesztettségre való igény és ebből adódóan – a számítógépes nyelvészet szempontjából nézve – a feldolgozás nehézsége is. A már létező nyelvi elemzők viszonylag jól működnek a sztenderdhez közelebb álló szövegek (pl. blogbejegyzések) esetén, a kommentek és mikroblogbejegyzések nyelvi feldolgozása viszont zajossága és nem sztenderd formájuk miatt nehezen megvalósítható a meglevő eszközök adaptálása nélkül.

Más nyelvekre már készültek internetes szövegeket tartalmazó annotált korpuszok, l. például az angolra [1] és a franciára [2]. Célunk az volt, hogy webes szövegekből (nyilvános kérdés–válasz párokból és bejegyzésekre érkezett kommentekből) olyan kézzel ellenőrzött, helyes morfológiai és szintaktikai annotációval ellátott magyar nyelvű korpuszt hozzunk létre, amelyen a jövőben természetesnyelv-feldolgozásra kifejlesztett elemzőket lehetséges tesztelni és segítségével optimalizálni. Cikkünkben e korpusz létrehozásának folyamatát ismertetjük, az

annotáció szintjeinek bemutatásával, továbbá ismertetjük a magyarlanc elemzővel [3] elért morfológiai és szintaktikai elemzés végeredményét is. A korpuszt oktatási és kutatási célokra szabadon elérhetővé kívánjuk tenni.

2. A korpusz összetétele

A szövegek összeválogatása során szempont volt, hogy a kommunikáció szóbeli jellegzetességeit mutassák, mind stílusban, mind formában (kérdés–válasz párok, beszélgetések). A korpusz 2014 augusztusában gyűjtött nyilvános Facebook-kommentekből, valamint a gyakorikerdesek.hu oldalon feltett kérdésekből és válaszokból áll. A gyűjtés során a teljesség és visszakereshetőség követelményének megfelelően egész thread-eket mentettünk el, elérési útvonallal és időbélyeggel együtt. A bejegyzések nagyrészt hobbival, személyes érdeklődési körökkel és életmóddal kapcsolatosak.Megemlítendő, hogy a Szeged Korpuszban is találhatóak többé-kevésbé hasonló hangvételű és stílusú írások (szépirodalmi művek és általános iskolai fogalmazások formájában), bár esetükben a zajosság nem (illetve jóval kevésbé) jelenik meg.

A korpusz főbb adatait az 1. táblázat foglalja össze. A korpuszt – méreteinél is fogva – elsődlegesen benchmark adatbázisként kívánjuk hasznosítani, mely a különféle nyelvi elemzők webes doménre történő adaptálását teszi lehetővé.

1. táblázat. A korpusz mérete.

Típus	**Facebook**	**Gyakorikérdések**	**Összesen**
Bejegyzések	879	258	1 137
Mondatok	1 208	728	1 936
Tokenek	8 739	9 880	18 620
szó	7 171	8 236	15 106
írásjel	904	1 551	2 455
emotikon	674	91	756

3. Morfológia

A webes szövegek elemzésének nehézségei már az előfeldolgozás, azaz a mondatszegmentálás és a tokenizálás során jelentkeztek. A sztenderd szövegen tanult magyarlanc hangulatjelekkel, extrém írásjelhasználattal és helytelen egybe- és különírással nem találkozott, így – mint az várható is volt az előzetes kutatás [4] után – a folyamat nem volt megoldható automatikusan. Sok esetben a tagmondat–mondat viszony és különbség nem volt tökéletesen megállapítható. Ezek a problémák a központozás elhagyása vagy szokatlan használata miatt adódtak (pl. *Péter én meg a gépem xD majdnem szét vertem, hogy amikkor oda*

ülnék tankolni persze akkor fagyok ki:D). Megjegyzendő, hogy a szóbeli kommunikációban sokszor előforduló néma szünet gyakran megjelenik mintegy gondolategységeket elválasztó írásjelként (többnyire „..." formában), azonban használata korántsem következetes.

A szegmentálás után a korpuszban előforduló szóalakokat összegyűjtöttük, majd a magyarlanc felhasználásával és kézi kiegészítéssel megadtuk az összes lehetséges elemzésüket. Következő lépésben előállítottunk egy, a Szeged Korpusz formátumához hasonló szerkezetű szövegkorpuszt, amelyben az annotátorok egy erre a célra kifejlesztett szoftverrel kézileg egyértelműsítették a szóalakokat.

A webes szövegekben előforduló morfológiai jellegű hibákat, pontosabban a sztenderd nyelvhasználattól való eltéréseket egy korábbi kutatásban [4] már felvázoltuk, a típushibák és a lehetséges automatikus megoldások gyűjtése a jelenlegi folyamatnak is részét képezte. A legtöbb tévesztés ékezetekkel, egybeírással és egyéb helyesírási hibákkal kapcsolatos. A morfológiai annotáció során megtartottuk az eredeti – hibásan írt – szóalakot, és ehhez a kontextusnak megfelelő helyes elemzést (lemmát és morfológiai kódot) rendeltük hozzá, a Szeged Korpusz 2.5-ben [5] használt eljáráshoz hasonlóan. Speciális esetekben, például ékezetek vagy betűk elhagyása miatt felmerülő poliszémia esetén (pl. *joban – jóban* vagy *jobban*; *tok – tök* vagy *tudok*) minden lehetséges (ill. gyakori, a szövegben előforduló) helyes kódot és lemmát felvettünk, míg a tévesen egybeírt alakok a szövegben előforduló, esetleg ékezetesített lemmát kapták meg, konszenzus alapján eldöntött kóddal ellátva (pl. *jolesz* esetében igeivel).

A 2. táblázatban a tartalmas (azaz nem írásjel és hangulatjel) tokenek mondatbeli eloszlása látható. A legszembetűnőbb eltérés, hogy míg a Gyakorikérdések alkorpusznak majdnem fele több mint 10 tokenből áll, a Facebook esetén ez a nagyságrend a 3-6 tokenes kategóriában jelenik meg. A két domén átlagos mondatonkénti tokenszáma is hasonló arányokat mutat, az előző sorrendet követve 11,07 és 5,84 szóalak/mondat (összes tokenre nézve 11,19 és 6,39).

2. táblázat. Tokenszám mondatonként.

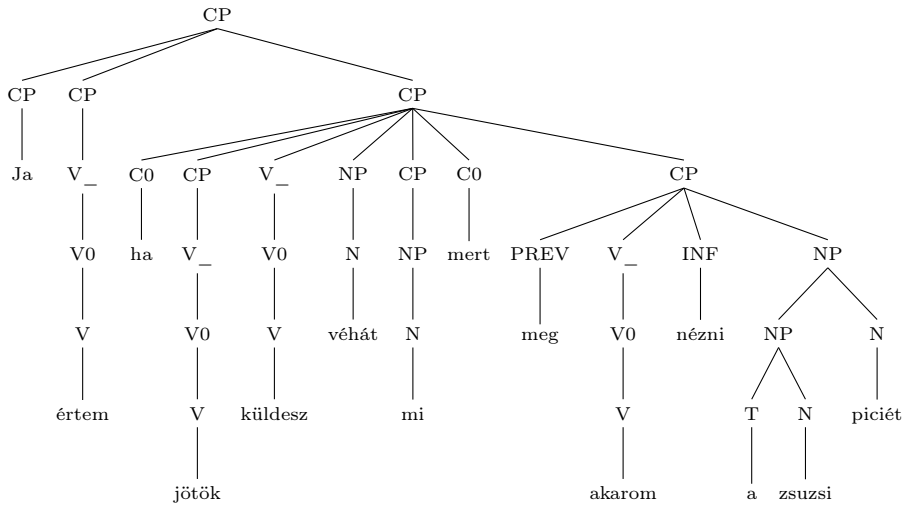
Tokenszám	**Facebook**	**%**	**Gyakorikérdések**	**%**	**Összesen**	**%**
1-2 token	33	4,53	94	7,78	127	6,56
3-6 token	233	32,01	685	56,71	918	47,42
7-10 token	128	17,58	199	16,47	327	16,89
10+ token	334	45,88	230	19,04	564	29,13
Összesen	728	100,00	1208	100,00	1936	100,00

4. Szintaxis

A korpusz szövegeit konstituens- és függőségi elemzéssel is elláttuk. A következőkben ezeket a munkafolyamatokat részletezzük.

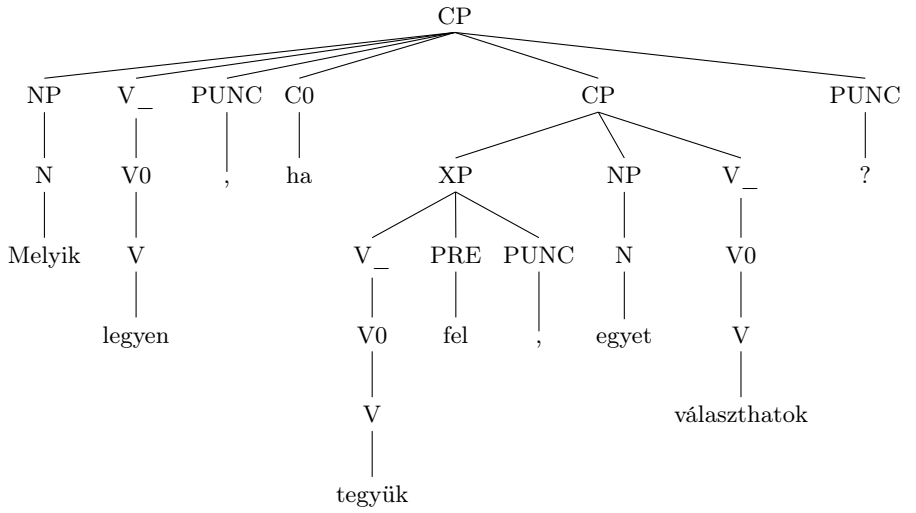
4.1. Összetevős elemzés

A mondatokat a morfológiai annotálás és hibaelemzés után a nyelvész szakértők először összetevős (azaz konstituens-) elemzéssel látták el, a morfológiához hasonlóan kézzel. Az annotáció során törekedtünk arra, hogy a Szeged Treebankhez hasonló módon történjen, kiegészítve a webes szövegek annotációja során felmerült megoldásokkal. Ilyen módosítás például, hogy a hibásan különírt szavak és szócsoportok (tipikusan toldalékolt szavak és szóösszetételek) egy összetevőbe kerültek. Az emotikonok a mondathoz szorosan nem tartozó összetevőként (azaz XP-ként) lettek felvéve.



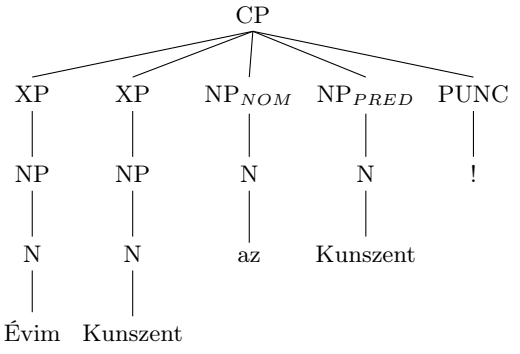
1. ábra: Központozás hiánya.

Az annotáció során felmerült további problémákat a szövegek beszélt nyelvhez közeli stílusából adódó nyelvhasználati jellemzők okozták. A kérdés–válasz struktúrának megfelelően jelentős számban fordultak elő hiányos mondatok. Gyakran okozott problémát a mondat- és tagmondathatárok megállapítása, ebben a jelentésbeli összefüggéstelenség és a központozás nem rendeltetésszerű használata is közrejátszottak (főként a Facebook alkorpuszban, 1. ábra).



2. ábra: Közbevetés.

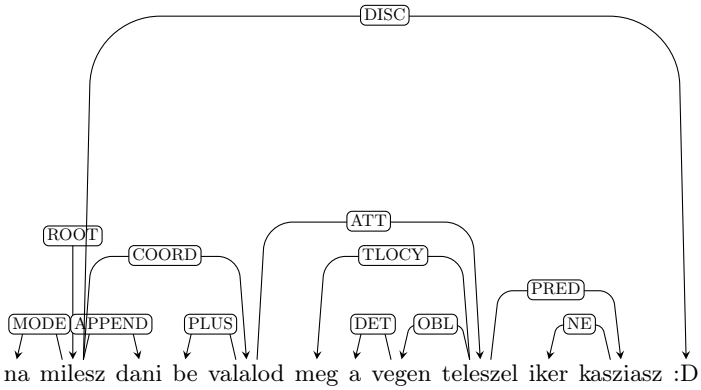
Ehhez a problémakörhöz kapcsolódik, hogy előfordulnak a mondatot megszakító, beágyazott mondatok, amelyek a mondatnak a beszélő által szükségesnek vélt kiegészítései, pl. hangsúlyozzák a szubjektivitást vagy modalitást közölnek. Ezeknek a státusza nyelvészeti szempontból sem teljesen tisztázott, közbevetésként (azaz a mondat szerkezetébe nem tartozó összetevőként) elemeztük őket (2. ábra). Egyelőre hasonlóan jártunk el a megszólítások és a nyelvészeti szakirodalomban kontrasztív topiknak nevezett jelenség esetében is (3. ábra).



3. ábra: Megszólítás és kontrasztív topik.

4.2. Függőségi annotáció

A korpusz mondataihoz kézzel hozzárendeltük azok függőségi ábrázolását is. A mondatokat a magyarlanc [3] függőségi elemző moduljával előelemeztük, majd az így kapott automatikus annotációt nyelvész szakértők kézzel kijavították. A munkálatok során alapvetően a Szeged Dependencia Treebank [6] létrehozása során alkalmazott elveket követtük, néhány változtatásra azonban szükség volt a webes szövegek sajátságainak megfelelően. Két új függőségi viszonyt vezettünk be az eddig is használatosak mellé: a DISC relációval jelöltük a szövegben előforduló diskurzusjelölőket és emotikonokat, a PLUS reláció pedig a hibásan különírt szavakat vagy betűkapcsolatokat köti össze (vö. 4. ábra).



4. ábra: Függőségi fa.

A korpuszban – a Szeged Dependencia Treebankhez hasonlóan – jelöltük az ún. virtuális csomópontokat is. Összesen 299 virtuális kopulát (kijelentő módú, jelen idejű harmadik személyű, fonológiailag üres létige) annotáltunk a korpuszban, ebből 119 a Facebook-, 180 pedig a Gyakorikérdések alkorpuszban található. Összesen 10 ellipszist találtunk a korpuszban, ami feltehetőleg annak köszönhető, hogy viszonylag kevés összetett mondat szerepel az anyagban, így a tagmondatok között ismétlődő elemek nem törlődnek.

5. Bizonytalansági annotáció

A korpusz szövegeiben megjelöltük a bizonytalanságot jelző nyelvi elemeket is. Az annotálás során a [7] és [8] által kidolgozott, majd a [9] által magyarra alkalmazott annotációs elveket követtük, azaz szemantikai és diskurzusszintű bizonytalanságot egyaránt annotáltunk. A szövegekben található bizonytalansági kategóriák megoszlását a 3. táblázat mutatja.

Az adatokból látszik, hogy a szemantikai bizonytalanság jóval gyakoribb a Gyakorikérdések alkorpuszban, ami valószínűleg annak köszönhető, hogy itt a

3. táblázat. Bizonytalansági annotáció a korpuszban.

	Típus	**Facebook**	**%**	**Gyakorikérdések**	**%**	**Összesen**	**%**
Szemantika	Episztemikus	7	3,45	39	10,32	46	7,92
	Doxasztikus	30	14,78	59	15,61	89	15,32
	Feltételes	18	8,87	64	16,93	82	14,11
Diskurzus	Weasel	23	11,33	31	8,20	54	9,29
	Hedge	58	28,57	117	30,95	175	30,12
	Peacock	67	33,00	68	17,99	135	23,24
	Összesen	203	100	378	100	581	100

felhasználók többsége nem ismeri egymást, ezért számos feltételezéssel élnek beszélgetéseik során, amit ki is fejeznek nyelvi eszközökkel. Ezzel szemben a Facebookon az ismerősök között zajló beszélgetéseket dolgoztuk fel elsősorban, és ebben a körben a felhasználók előszeretettel állítják be tényként a nem feltétlenül objektív állításaikat, azaz a diskurzusszintű bizonytalanság elemei lesznek gyakoribbak.

6. Statisztikai adatok

A 4. táblázatban láthatjuk a szófajok eloszlását, az 5. táblázat pedig a függőségi viszonyok eloszlását mutatja a korpuszban. Látható, hogy elsősorban a DISC reláció használata, illetve a mondatszavak, indulatszavak gyakorisága mutat nagy különbséget a domének között: a Facebookról származó szövegekben sokkal gyakoribb a használatuk, mint a Gyakorikérdések oldalon. Ez valószínűleg annak köszönhető, hogy a Facebookon a felhasználók egymás közti beszélgetéseikben sokkal inkább kimutatják érzelmeiket, illetve az adott beszélgetéshez való viszonyulásukat, mint a valamivel formálisabb Gyakorikérdések oldalon, ahol többnyire ismeretlen emberekkel társalognak.

Sajátos jelenség még a központozás hiánya vagy megléte. A Facebookról származó szövegekben arányaiban jóval kevesebb írásjelet találunk, mint a Gyakorikérdések alkorpuszban. Ez a mondatok átlagos hosszával függ össze: míg a Facebookon a felhasználók általában rövid, akár 1-2 szavas megjegyzéseket írnak, addig a Gyakorikérdések oldalon a hozzászólások többnyire hosszabbak, bővebben kifejtettek.

Az összetevők eloszlása a 6. táblázatban látható. Szembetűnő, hogy a két alkorpusz közötti különbség az XP kifejezésekben, azaz az emotikonok és egyéb, mondatszerkezetbe nem beleszóló összetevőkben mutatkozik meg (az egymáshoz viszonyított arány 82 és 17%). Ez annak köszönhető, hogy a facebookos szövegekben az emotikonok használata jóval elterjedtebb, mint a sztenderd nyelvhasználathoz közelebbi fórumhozzászólásokban, ahol a kategória főként zárójeles közbevetéseket jelölt. Érthető is, hiszen a kommentek élőbeszédszerű, azaz online-ságot és gyorsaságot megkövetelő használata megkívánja a kommunikációt kisegítő multimodális eszközök (pl mimika, gesztusok) pótlását.

4. táblázat. Szófajok (POS) eloszlása.

Szófaj	Facebook	%	Gyakorikérdések	%	Összesen	%
Főnév	1401	18,17	1700	20,88	3102	19,56
Ige	1470	19,06	1510	18,54	2980	18,79
Határozószó	1364	17,69	1321	16,22	2685	16,93
Névmás	746	9,67	834	10,24	1580	9,96
Kötőszó	565	7,33	898	11,03	1463	9,23
Névelő	530	6,87	785	9,64	1315	8,29
Melléknév	467	6,06	644	7,91	1111	7,01
Indulatszó	944	12,24	153	1,88	1097	6,92
Számnév	153	1,98	206	2,53	359	2,26
Névutó	38	0,49	78	0,96	116	0,73
Nyílt osztály	15	0,19	14	0,17	29	0,18
Ismeretlen	19	0,25	0	0,00	19	0,12
Összesen	7712	100,00	8143	100,00	15856	100,00

Különbségek láthatóak továbbá a C (kötőszó), ADJP (melléknévi frázis) és a PP (névutós kifejezés) kategóriában, amelyek a Gyakorikérdések alkorpuszban fordultak elő gyakrabban. Ezek az adatok valószínűleg a mondatok összetettségében és hosszában lévő különbségekkel magyarázhatók.

7. Automatikus szófaji egyértelműsítés és függőségi elemzés

A létrejött korpusz lehetővé tette azt is, hogy kísérleteket végezzünk a magyarlanc 2.0 [3] szófaji egyértelműsítő és függőségi elemző moduljával is. Első lépésben megnéztük, hogy a Szeged Korpusz 2.5-ön [5] tanított szófaji egyértelműsítő és függőségi modell milyen eredményeket képes elérni a webes szövegeken. A kísérleteket a Facebook és Gyakorikérdések alkorpuszokon külön-külön is, továbbá a teljes szövegállományon is elvégeztük. A függőségi elemzéshez az etalon (kézzel annotált) morfológiai kódokat használtuk fel. A kiértékeléshez a szófaji egyértelműsítés esetén a pontosság (accuracy) metrikát, míg a függőségi elemzés esetén a Labeled Accuracy Score (LAS) és Unlabeled Accuracy Score (ULA) metrikákat alkalmaztuk.

Az eredményekből azt láttuk, hogy a sztenderd szövegen tanított modellek szerényebb eredményt képesek csak elérni a webes szövegeken. Ezért megnéztük azt is, hogy webes (azaz hibával terhelt) szövegen tanítva milyen eredményt tudunk elérni. Ehhez a korpusz mondatait véletlenszerűen osztottuk fel tanító és teszthalmazra: a mondatok 20%-a került a teszthalmazba, 80%-a pedig a tanító halmazba. Az összehasonlíthatóság kedvéért azt is megvizsgáltuk, hogy ugyanezen a teszthalmazon a sztenderd szövegen tanított magyarlanc milyen eredményt képes elérni. Az így kapott eredmények a 7. táblázatban láthatók.

Az eredmények azt mutatják, hogy – nem meglepő módon – a webes szövegek nyelvi elemzése nehezebb a sztenderd szövegekénél. Ugyanakkor, ha már a tanító

5. táblázat. Függőségi viszonyok eloszlása.

Függőségi viszony	**Facebook**	**%**	**Gyakorikérdések**	**%**	**Összesen**	**%**
PUNCT	898	10,28	1541	15,60	2439	13,10
ATT	785	8,98	1279	12,95	2064	11,08
ROOT	1208	13,82	727	7,36	1936	10,40
CONJ	559	6,40	894	9,05	1453	7,80
MODE	632	7,23	735	7,44	1367	7,34
DET	540	6,18	801	8,11	1341	7,20
SUBJ	570	6,52	708	7,17	1278	6,86
COORD	524	6,00	675	6,83	1199	6,44
OBL	395	4,52	550	5,57	945	5,08
DISC	689	7,88	91	0,92	780	4,19
OBJ	306	3,50	378	3,83	684	3,67
TLOCY	332	3,80	283	2,86	615	3,30
NEG	243	2,78	249	2,52	492	2,64
PRED	219	2,51	270	2,73	489	2,63
INF	134	1,53	181	1,83	315	1,69
PREVERB	133	1,52	141	1,43	274	1,47
APPEND	154	1,76	67	0,68	221	1,19
NE	110	1,26	79	0,80	189	1,02
PLUS	125	1,43	45	0,46	170	0,91
DAT	75	0,86	74	0,75	149	0,80
LOCY	47	0,54	62	0,63	109	0,59
TTO	14	0,16	16	0,16	30	0,16
TO	11	0,13	15	0,15	26	0,14
AUX	15	0,17	7	0,07	22	0,12
TFROM	9	0,10	4	0,04	13	0,07
QUE	8	0,09	3	0,03	11	0,06
FROM	3	0,03	5	0,05	8	0,04
NUM	1	0,01	0	0,00	1	0,01
Összesen	8739	100,00	9880	100,00	18620	100,00

6. táblázat. Összetevők eloszlása.

Összetevők	**Facebook**	**%**	**Gyakorikérdések**	**%**	**Összesen**	**%**
NP	2125	23,23	2634	28,70	4759	25,97
CP	2271	24,83	1995	21,74	4266	23,28
V	1306	14,28	1302	14,19	2608	14,23
ADVP	979	10,70	1011	11,02	1990	10,86
C	568	6,21	898	9,78	1466	8,00
XP	1011	11,05	210	2,29	1221	6,66
ADJP	249	2,72	439	4,78	688	3,75
NEG	246	2,69	248	2,70	494	2,70
INF	148	1,62	204	2,22	352	1,92
PREVERB	187	2,04	143	1,56	330	1,80
PP	40	0,44	81	0,88	121	0,66
PA	16	0,17	13	0,14	29	0,16
Összesen	9146	100,00	9178	100,00	18324	100,00

7. táblázat. Szófaji egyértelműsítés és függőségi elemzés a magyarlanccal.

Tanító halmaz	**Teszthalmaz**	**Pontosság**$_{POS}$	**LAS**	**ULA**
Szeged Korpusz 2.5	100% Facebook	66,41	69,88	76,03
Szeged Korpusz 2.5	20% Facebook	64,32	75,03	80,39
80% Facebook	20% Facebook	75,11	75,43	81,89
Szeged Korpusz 2.5	100% Gyakorikérdések	79,24	80,17	82,91
Szeged Korpusz 2.5	20% Gyakorikérdések	79,18	85,11	88,36
80% Gyakorikérdések	20% Gyakorikérdések	79,54	79,57	84,96
Szeged Korpusz 2.5	100% webes szöveg	74,63	75,34	79,66
Szeged Korpusz 2.5	20% webes szöveg	71,68	80,77	84,65
80% webes szöveg	20% webes szöveg	78,97	79,9	85,01

halmazban is hibával terhelt szövegek szerepelnek, akkor sokkal jobb eredményeket tudunk elérni szófaji egyértelműsítésben, a teljes szövegállományon több mint 7 százalékpontnyi a javulás, a Facebook-szövegek esetében pedig több mint 10 százalékpont. Ez arra utal, hogy a Facebook-szövegek esetében különösen nagy jelentőséggel bír a doménadaptáció, hiszen nyelvezetük távolabb esik a sztenderd nyelvhasználattól, mint azt a Gyakorikérdések esetében láthatjuk, ahol nem számottevő a különbség a sztenderd modell és a doménen belüli modell által elért eredmények különbsége.

A függőségi viszonyokat vizsgálva valamivel árnyaltabb képet kapunk. A Facebook-szövegeken ismét látszik a tanítóhalmaz doménjének fontossága: a Facebookon tanult modell jobb eredményt ér el, mint a Szeged Korpuszon tanított modell. Ezzel szemben a Gyakorikérdések esetében számottevően jobb eredményt ér el a sztenderd szövegeken tanult modell, mint a saját doménbeli adatokon tanult modell. A különbség magyarázata feltehetőleg az, hogy a Gyakorikérdések alkorpusz – a mondatok szintaktikai szerkezetét tekintve – közelebb áll a sztenderd szövegekhez, mint a Facebook-szövegek, így a nagyságrendekkel nagyobb tanító adathalmazon (kb. 60 000 mondaton) tanított modell 3-4 százalékponttal nagyobb pontosságot képes elérni, mint a saját doménen (kb. 600 mondaton) tanított modell. A jövőben tervezett doménadaptációs kísérleteink remélhetőleg pontosabb képet nyújtanak majd a magyarlanc webes szövegekre történő adaptálási lehetőségeiről.

8. Összegzés

Cikkünkben bemutattuk az első magyar, kézzel annotált, webes szövegeket tartalmazó korpuszt, melyet morfológiai és (összetevős és függőségi szemléletű) szintaktikai elemzést, valamint szemantikai és diskurzusbeli bizonytalan kifejezések annotációját tartalmazza. Ismertettük az annotáció folyamatát, illetve beszámoltunk a magyarlanc elemző webes szövegekre történő adaptálási kísérleteiről.

A korpusz méreteinél fogva nem alkalmas statisztikai elemzők tanítására, célunk egy benchmark adatbázis előállítása volt. Véleményünk szerint mivel a webes szövegek témában és műfajban is igen változatosak, nem is lenne célravezető

a felügyelt gépi tanulás paradigmáját követni, hanem doménadaptációs megoldások jelenthetik a megoldást. A jövőben tovább kívánunk foglalkozni a webes szövegekre adaptált elemzők továbbfejlesztésével, illetve terveink között szerepel a korpusz újabb annotációs rétegekkel (névelemek, többszavas kifejezések) való kézi bővítése is.

A korpusz oktatási és kutatási célokra ingyenesen elérhető a http://rgai.u-szeged.hu/SzegedTreebank oldalon.

Köszönetnyilvánítás

A jelen kutatás a futurICT.hu nevű, TÁMOP-4.2.2.C-11/1/KONV-2012-0013 azonosítószámú projekt keretében valósult meg. Vincze Veronika kutatásait a TÁMOP 4.2.4.A/2-11-1-2012-0001 azonosítószámú Nemzeti Kiválóság Program – Hazai hallgatói, illetve kutatói személyi támogatást biztosító rendszer kidolgozása és működtetése konvergencia program című kiemelt projekt támogatta. Mindkét projekt az Európai Unió támogatásával, az Európai Szociális Alap társfinanszírozásával valósul meg.

Hivatkozások

1. Mott, J., Bies, A., Laury, J., Warner, C.: Bracketing Webtext: An Addendum to Penn Treebank II Guidelines. Linguistic Data Consortium (2012)
2. Seddah, D., Sagot, B., Candito, M., Mouilleron, V., Combet, V.: The French Social Media Bank: a treebank of noisy user generated content. In: Proceedings of COLING 2012, Mumbai, India, The COLING 2012 Organizing Committee (2012) 2441–2458
3. Zsibrita, J., Vincze, V., Farkas, R.: magyarlanc: A Toolkit for Morphological and Dependency Parsing of Hungarian. In: Proceedings of RANLP-2013, Hissar, Bulgaria (2013) 763–771
4. Varga, V., Wieszner, V., Hangya, V., Vincze, V., Farkas, R.: Magyar nyelvű webes szövegek számítógépes feldolgozása. In Tanács, A., Varga, V., Vincze, V., eds.: X. Magyar Számítógépes Nyelvészeti Konferencia, Szeged, SZTE (2014) 327–331
5. Vincze, V., Varga, V., Simkó, K.I., Zsibrita, J., Nagy, Á., Farkas, R., Csirik, J.: Szeged Corpus 2.5: Morphological Modifications in a Manually POS-tagged Hungarian Corpus. In: Proceedings of LREC'14, Reykjavik, Iceland (2014) 1074–1078
6. Vincze, V., Szauter, D., Almási, A., Móra, Gy., Alexin, Z., Csirik, J.: Hungarian Dependency Treebank. In: Proceedings of LREC 2010, Valletta, Malta, ELRA (2010)
7. Szarvas, Gy., Vincze, V., Farkas, R., Móra, Gy., Gurevych, I.: Cross-Genre and Cross-Domain Detection of Semantic Uncertainty. Computational Linguistics **38** (2012) 335–367
8. Vincze, V.: Weasels, Hedges and Peacocks: Discourse-level Uncertainty in Wikipedia Articles. In: Proceedings of the Sixth International Joint Conference on Natural Language Processing, Nagoya, Japan, Asian Federation of Natural Language Processing (2013) 383–391
9. Vincze, V., Simkó, K.I., Varga, V.: Annotating Uncertainty in Hungarian Webtext. In: Proceedings of LAW VIII - The 8th Linguistic Annotation Workshop, Dublin, Ireland, Association for Computational Linguistics and Dublin City University (2014) 64–69