A duplakocka modell és az igei szerkezeteket kinyerő „ugrik és marad” módszer nyelvfüggetlensége, valamint néhány megjegyzés az UD annotáció univerzalitásáról

Sass Bálint

MTA Nyelvtudományi Intézet, ELTE BTK
sass.balint@nytud.hu

Kivonat Jelen tanulmány egy módszernek a magyartól különböző nyelvekre való alkalmazhatóságát vizsgálja. A (Sass, 2019) tanulmány egy valódi igei szerkezetek kinyerésére szolgáló eljárást mutat be magyar nyelvre, és két állítást fogalmaz meg mellékesen: (1) a módszer tetszőleges nyelvre alkalmazható; (2) a módszer alkalmazásához szükséges adatok függőségileg elemzett korpuszból könnyen származtathatók. E két állítást vesszük górcső alá. Adatként universal dependencies (UD) korpuszokat használunk fel. Az UD-nek köszönhetően annotációs különbségek elvileg nincsenek nincsenek a különféle nyelvű korpuszok között, csak a nettó nyelvi különbségek láthatók. Ezzel kapcsolatban gyakorlati megfigyeléseink alapján kritikát fogalmazunk meg. Bár az ige és közvetlen bővítményei közötti viszonyokat különböző nyelvek különböző eszközökkel fejezik ki, a vizsgált nyelvekre ezek a nyelvi eszközök néhány általános módon megragadhatók: esetrag, elöljáró/névutó (esetraggal vagy anélkül), szórend. Az említett eljárás működésének egyetlen feltétele az igék és közvetlen bővítményeik közötti viszonyok leírása, a fentiek alapján tehát működtethető az algoritmus. Eredményként valódi igei szerkezeket kapunk, azaz az eredmények igazolják sejtésünket, az eredeti cikk állításai megállják a helyüket.
Kulcsszavak: igei szerkezet, valódi igei szerkezet, duplakocka, korpuszháló, ugrik és marad, többnyelvű, universal dependency

1. Motiváció – kiáltvány a szerkezetekért

A nyelv alapegységei nem a szavak, hanem a szerkezetek. A szó csak a szerkezet szélső esete: olyan szerkezet, ami egy elemből áll.

A legegyszerűbb egyszavas kifejezés esetében is nagyon gyakran előfordul, hogy egy másik nyelven a megfelelője többszavas. Azt gondolnánk, hogy ‘*krumpli*’ minden nyelven egy szó, franciául mégis ‘*pomme de terre*’ (földi alma). A szenegáli wolof anyanyelvű beszélők hihetik, hogy az olyan köznapi dolgokra, mint a ‘*gëmm*’ nyilván minden nyelv külön szót használ, magyarul mégis így mondjuk ezt: ‘*behunyja a szemét*’. Azt mondhatjuk, hogy szerencse, ha valamire épp van egyszavas kifejezés egy nyelvben, tetszőleges nyelven lehetséges, hogy a szóban forgó dologra csak többszavas egység, szerkezet létezik.

Másfelől, olyan is gyakran előfordul, hogy egy szó megfelelője egy másik nyelven kötött morféma, ahogy ezt az angol '*in*' és a magyar '*-bAn*' rag egyszerű példája mutatja. Ennek megfelelően a '*believe in*' többszavas szerkezet, míg a '*hisz -bAn*' -ről ez nem mondható el a szó szoros értelmében. Utóbbi inkább csak másfél szavas.

Sőt, bizonyos szerkezeti elemek egyáltalán nem is jelennek meg a felszínen, miközben nagyon is fontosak az adott kifejezés szempontjából. Az angol ditranzitív szerkezetek három eleméről – az alanyról, a direkt tárgyról és az indirekt tárgyról – kizárólag a szórendből tudjuk meg a szerkezetben betöltött szerepüket. Innen nézve a '*give*' egy összetett, négy elemből álló szerkezet: '*give SUBJ OBJ IOBJ*'.

Annak idején az első szótárírók mégis a szavakat kezdték el listázni, első ránézésre a szavak tűntek természetes alapegységnek. Ez a hagyomány azóta is él. A szótárakban címszavakat találunk akkor is, ha egyre inkább teret kap a címszavakhoz kapcsolódó különféle típusú szerkezetek, frázisok bemutatása (Atkins és Rundell, 2008).

Amellett érvelünk tehát, hogy az lenne az üdvös, ha nem szótárakat, hanem *szerkezettár*akat hoznánk létre. A szerkezetek legtöbbször egyértelműsítik a bennük szereplő szavakat, de legalábbis csökkentik a többértelműségüket (Yarowsky, 1993; Pustejovsky, 1995). Ahhoz képest, hogy egy forrásnyelvi igéhez felsorolunk 8-10 igét a célnyelven (Kilgarriff, 1997), sokkal hasznosabb, ha az ige szerkezeteit vesszük számba, és a megfelelő szerkezeteket adjuk meg a másik oldalon.

Az angol '*go*' esetében első körben (például egy kezdő nyelvtanuló számára) elegendő az alábbi három szerkezet ismerete:

- '*go to NOUN*' = megy valahová
- '*going to VERB*' = fog csinálni vmit
- '*go ADJ*' = válik vmilyenné

Az, hogy a szerkezeteket tekintjük alapelemnek az első lépés a címszavak helyett „címszerkezeteket" tartalmazó szerkezettárak megalkotása felé.

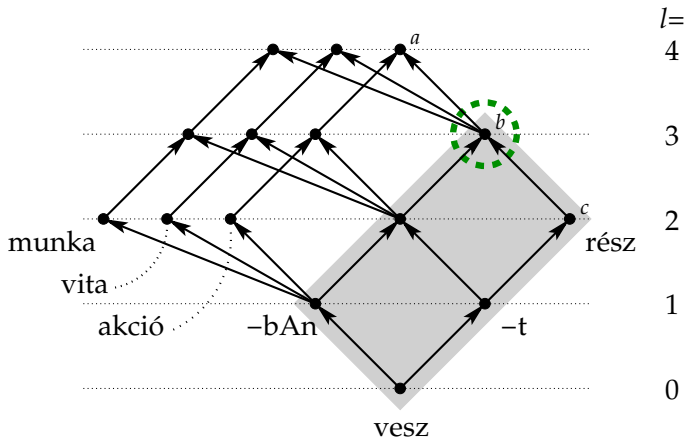
2. Korábbi munka

Kiindulópontunk (Sass, 2019), mely egy függőségileg elemzett korpuszból valódi igei szerkezeteket kinyerő algoritmust mutat be magyar nyelvre. A módszer alapját képező ún. „duplakocka" modellt (Sass, 2018) írja le részletesen.

A valódi igei szerkezet fogalma itt a lehető legáltalánosabb módon magában foglalja a lexikográfiailag hasznos valamennyi igei szerkezetet: a vonzatos igéket ('*vki hisz vmiben*'), az igei szókapcsolatokat ('*süt a nap*') és a vonzatos komplex igéket ('*vki részt vesz vmiben*') is. Éppen ezek azok az egységek, melyeknek egy ige vonatkozásában egy szerkezettárban szerepelniük kell, ezért hasznos ez az összegyűjtésükre szolgáló automatikus módszer.

A modell két kulcseleme az egy tagmondatot és a benne rejlő igei szerkezeteket (azaz az igét és a mellette lévő helyeket és kitöltőket, azaz slot-okat és filler-eket) reprezentáló duplakocka, valamint a korpusz ugyanazon igét tartalmazó összes tagmondatát egyben reprezentáló korpuszháló, mely a duplakockák

egyfajta egymásra vetítésével, kombinálásával áll elő. A korpuszháló megjeleníti az adott ige mellett előforduló összes hely és kitöltő gyakorisági viszonyait (1. ábra).



1. ábra: Az ábra a duplakocka és a korpuszháló fogalmát illusztrálja, valamint bemutatja az „ugrik és marad" módszer működését. Az ábrán látható kicsi korpuszháló a '*részt vesz munkában*', '*részt vesz vitában*' és '*részt vesz akcióban*' tagmondatok duplakockájának kombinációjaként áll elő. A gyakoriságot képviselő f függvény értéke a szürke hátterű csomópontok esetében 3, a többi csomópont esetében 1. Az l az adott csomópontokhoz tartozó szerkezetek hossza. A könnyebb áttekinthetőség kedvéért az alanyi dimenziót nem ábrázoljuk.

A modellen dolgozó algoritmus feladata kettős: meg kell állapítania, hogy mely helyek részei a szerkezetnek és ezek közül hol inherens elem a kitöltő is. A '*vki részt vesz vmiben*' valódi igei szerkezethez 3 hely (alany, tárgy és -bAn) valamint a tárgyi helyet fixen kitöltő '*rész*' elem tartozik elengedhetetlenül hozzá. Tekintsük az adott csomópont által képviselt igei szerkezet korpuszgyakoriságát megadó f függvényt a korpuszhálón. A kidolgozott „jump and stay" („ugrik és marad") módszer arra a megfigyelésre épül, hogy a valódi igei szerkezeteket képviselő csomópontokra igaz az, hogy az f függvény értéke a csomópont fölötti élen jelentősen megnő, a csomópont alatti élen pedig nem változik. A 1. ábrán látható példán azt látjuk hogy az a-b élen a függvény értéke 1-ről 3-ra nő, a c-b élen pedig nem változik. Ez jelzi, hogy a b pontban egy valódi igei szerkezetet találunk. Megfigyelhető, hogy az „ugrás" (a-b) során egy esetleges elemet ('*akció*') hagyunk el, a „maradás" (c-b) során pedig egy szükséges elemet ('*-bAn*') veszünk hozzá. Így a köztes b pont éppen a megkívánt elemeket fogja tartalmazni, így valódi igei szerkezetet kapunk. Vegyük észre, hogy az ábra tetszőleges csomópontjából indulva minden esetben a bekarikázott – helyes – csomóponthoz vezetnek az ugrás+maradás lépéssorozatok.

3. UD korpuszok előfeldolgozása

A módszert eddig kizárólag magyar nyelvű szövegen használták. Az említett cikk két állítást fogalmaz meg: (1) a módszer tetszőleges nyelvre alkalmazható; (2) a módszer alkalmazásához szükséges adatok függőségileg elemzett korpuszból könnyen származtathatók. Jelen tanulmány e két állítást vizsgálja: igyekszik megmutatni, hogy a modell és az algoritmus is nyelvfüggetlen, valamint felméri, hogy mennyi munkával lehet előállítani függőségileg elemzett korpuszból a kívánt bemenetet. Ha reményeink beigazolódnak, az megnyitja az utat tetszőleges, akár kisebb, kevesebb erőforrással bíró nyelvek alapvető szerkezeteinek számbavétele előtt. (Kis nyelvekre természetesen kisebb az esélye, hogy függőségi elemző vagy függőségileg annotált korpusz rendelkezésre áll. De ez nem is elengedhetetlen feltétel. A függőségileg elemzett korpusz kényelmes lehetőséget biztosít a szükséges bemenet előállítására, de egy egyedi szabályalapú eljárás is megfelelő lehet erre a célra.)

Azért bízhatunk a nyelvfüggetlenségben, mert lényegében pusztán arra van szükség, hogy az adott nyelvben legyenek predikátumok, a predikátumoknak argumentumai, és a kettő között valamiféle megragadható viszony. Arra pedig, hogy az inputot egyszerűen elő tudjuk állítani, a szabadon hozzáférhető, egységes annotációval bíró, kézzel annotált, gold sztenderd UD korpuszok (Nivre és mtsai, 2019) adnak lehetőséget[1].

Az UD korpuszok közül a vizsgálatainkhoz cseh, német, angol, finn, magyar, holland, norvég, török és wolof nyelvű korpuszt választottunk[2]. Az elvégzett munka legnagyobbrészt a korpuszok előfeldolgozásából állt. A megfelelő bemenet előállítása után eredeti formájában futtattuk az „ugrik és marad" módszert a valódi igei szerkezetek kinyerésére.

Az előfeldolgozás feladata tehát az igék, valamint az ige közvetlen bővítményeit képviselő helyek és kitöltők meghatározása volt. Ez nagyon hasonló a „konstituensfa felsőszintű szintaktikai elemei"-hez („top level syntactic sequence of the constituent tree") (Shi és mtsai, 2016), azzal a különbséggel, hogy az elemek sorrendjét mi nem vesszük figyelembe.

Az igét és közvetlen bővítményeit a tagmondat tartalmazza, de tagmondatra bontásra a függőségi elemzésnek köszönhetően nem volt szükség. Sőt, úgy döntöttünk, hogy nemcsak a tagmondatok főigéjét, hanem minden egyes igei alakot (UD: UPOS=`VERB`) gyökérnek tekintünk, így a potenciális igei szerkezetek száma megnőtt, és adott igei alak két szerkezetnek is része lehet, az egyiknek gyökérként, a másiknak bővítményként. A '*He didn't think he needed to know anything about South Asia.*' mondatban a '*need*' és a '*know*' is ilyen kettős szerepű ige.

[1] A használt UD terminusok feloldása a `http://universaldependencies.org` oldalon található meg.

[2] Konkrétan az UD 2.4-es verziójából vett alábbi fájlokkal dolgoztunk: cs_pdt-ud-dev.conllu, de_hdt-ud-dev.conllu, en_ewt-ud-train.conllu, fi_tdt-ud-train.conllu, hu_szeged-ud-train.conllu, nl_alpino-ud-train.conllu, no_bokmaal-ud-train.conllu, tr_imst-ud-train.conllu, wo_wtb-ud-train.conllu. Ezek nagyjából egyforma méretű 2-300000 szavas korpuszok.

Ennek köszönhetően tehát az igei szerkezetekben főnévi igenévi bővítmények is megjelennek.

1. hely (slot) megállapítása ♦ Az ige és közvetlen bővítményei közötti viszonyok a vizsgált nyelvekre néhány általános módon megragadhatók: esetrag, elöljáró/névutó (esetraggal vagy anélkül), szórend. Az ige közvetlen dependenseként megjelenő `nsubj`, `obj`, `iobj`, `obl`, `case` és `xcomp` relációval kötődő elemeket vesszük tekintetbe, valamint azokat, melyek tetszőleges reláció mellett rendelkeznek `Case` feature-rel. (Az `xcomp` a fent említett esetet jelenti, mikor a bővítmény egy tagmondat, így saját igéje van általában főnévi igenévi alakban.) Fontos kiemelni, hogy ezen kívül figyelembe vesszük ezen dependensek dependenseként megjelenő elöljárók/névutók (UD: UPOS=`ADP`) lemmáját is. Példa: az '*Acc=in*'[3] olyan közvetlen bővítményi helyet jelöl, amely tárgyesetben áll és van egy '*in*' elöljárója.

Gondot jelent, hogy a német korpuszban az elöljáró+névelő kontrakciók (pl.: '*am*'='*an*'+'*dem*', '*ins*'='*in*'+'*das*') lemmája sajnos megegyezik a szóalakkal, ahelyett, hogy az eredeti elöljáró lenne a lemma. Itt egyedi eljárással mappelni kellett a kontrakciókat az elöljárókra, hogy a szerkezetekben ne váljon ketté a sima elöljáró és a kontrahált forma. Egy másik probléma a főnévi igenevekhez kapcsolódik. Bizonyos nyelvekben a főnévi igenévhez tartozik egy elöljárókhoz hasonló plusz szócska: angolban például '*to*', hollandban '*te*', wolofban '*ci*'. Ez a nagyon specifikus elem az UD annotációban összemosódik más jellegű elemekkel: szófaja partikula (UPOS=`PART`), hasonlóan a cseh '*je*' (csak), norvég '*ikke*' (nem) szóhoz vagy a magyar '*meg*' igekötőhöz; függőségi relációja `mark`, ami pedig az összes alárendelő tagmondatot jelölő elem közös kódja. Emiatt ezek az elemek végül is csak nyelvfüggő módon, a szóalakjuk alapján ragadhatók meg. A főnévi igenév jelölőszócskája egy olyan egyedi elem, amelynek érdemes lenne bevezetni egy külön egyedi szófajt/kódot, amit nagyon jó lenne az összes korpuszban egységesen használni.

2. kitöltő (filler) megállapítása ♦ A kitöltő az ige közvetlen dependensének kisbetűsített lemmája lesz. Sass (2019) említi, hogy a névmások, mivel nagyon gyakoriak, hajlamosak megjelenni kitöltőként, pedig általában nincs idiomatikus jelentésük. A cikk javaslata szerint a névmásokat (UD: UPOS=`PRON`) az előfeldolgozás során töröljük, kivétel ezalól a '*maga*' és az '*egymás*'. A '*maga*' megfogható a `Reflex=Yes`, az '*egymás*' pedig a `PronType=Rcp` UD feature alapján.

Gondot jelent, hogy a '*maga*' esetén a német '*sich*' annotációja eltér ettől, így külön kell kinyerni a lemmája alapján. Az '*egymás*' esetén összetettebb a helyzet, a cseh '*navzájem*', a német '*einander*' és a török '*birbiri*' esetén különféle eltérő módokon jelölik az annotációban a korpuszok ezt a szót. Az '*einander*' elöljárós alakjai további problémát jelentenek: ezeket az alakokat egybeírjuk ('*miteinander*', '*zueinander*'), a korpuszban úgy döntöttek, hogy ezeket az egybeírt alakokat adják meg lemmaként (!) ahelyett, hogy az eredeti szó lenne a lemma, és

[3] '*Acc=in*': itt az egyenlőségjel két elem összetartozását jelöli, azt, hogy az adott helyet két elem együtt reprezentálja.

külön elemként kapcsolódna hozzá az elöljáró – függetlenül attól, hogy egybeírjuk. Még további problémát jelent az angol '*each other*', ez esetben a külön két szóba írás a gond. Ez a kifejezés teljesen önálló kódot kap (`DET+ADJ`), külön egyedi megoldással lehet csak rátalálni. A főnévi igenév jelölőszócskájához hasonlóan az '*egymás*' is tipikus esete az olyan különleges szónak, aminek saját szófaj, saját kód dukál, amit aztán az összes korpuszban egységesen lehet használni.

3. ige megállapítása ♦ Az ige kisbetűsített lemmája elé kapcsoljuk az esetleges elváló igekötőt. Az igekötő-ige kapcsolatot általában a `compound:prt` UD reláció jelzi. Az egységesség kedvéért minden nyelvben igekötő–ige sorrendben szerepeltetjük az igekötős igék elemeit, ez az angolban '*upbreak*', '*inturn*' alakokat eredményez.

Gondot jelent, hogy a magyar korpuszban erre a relációra eltérő jelölést (`compound:preverb`) használnak. Szintén probléma, hogy az angol korpuszban ugyanarra a jelenségre többféle annotáció használatos: a '*stir up*' például helyesen `compound:prt`, a '*get through*' vagy a '*go away*' viszont `advmod` (`ADV` szófajjal). A vizsgált 9 nyelv közül háromban (cseh, finn, török) nem találtam elváló igekötőt. Lehetséges, hogy van, csak ismét más a jelölés. További eltérés, hogy a magyarban + kapcsolja össze az egybeírt igét és igekötőt, más nyelvekben (pl.: német) nem jelzi ezt semmi. A holland '*plaatsnemen*' igekötős ige, míg formai és egyben tartalmi angol megfelelője '*take place*' ige+tárgy szerkezetű.

A universal dependency treebank-ek kiváló erőforrások, bár éri némi kritika is őket (Osborne és Gerdes, 2019). A fentiek alapján megállapíthatjuk, hogy nem felelnek meg maradéktalanul annak az alapvetőnek vélt követelménynek, hogy ugyanazon jelenséget mindig ugyanúgy jelöljünk, eltérő jelenséget pedig mindig eltérően jelöljünk („use the same term ... for the same function") (Croft és mtsai, 2017). Azaz mindig minden egységesen, ugyanúgy működjön, hogy amennyire csak lehet, ne kelljen nyelvfüggő lépéseket végezni. Ez a hiányosság legfőképpen azért baj, mert veszélyezteti a „minden találatra szükség van" elvet. Eszerint mindenfajta korpuszkereséskor – az igék, helyek és kitöltők fent részletezett megállapítása is ilyen – a felhasználó mindig az összes találatot szeretné látni, azaz a recall az, ami itt kiemelten fontos. Az UD treebank-ek ezzel együtt nagyon jól használhatók, az előfeldolgozás során a fent részletezett problémákat megoldva igyekeztünk a találatvesztés esélyét a lehető legkisebbre szorítani.

Úgy is fogalmazhatunk, hogy valójában nem „formai" hanem „funkcionális" függőségekre van szükségünk az igei szerkezetek megragadásához, és a treebankekre épülő eljárásokban általában is ezek tűnnek igazán hasznosnak. A fenti átalakító lépések mindegyike tekinthető egy ebbe az irányba – a funkcionális függőségek felé – tett lépésnek, ahol az azonos *funckiójú* elemek, szavak, illetve relációk kapnának azonos jelölést.

A korpuszokban a legalább 20× előforduló igéket vizsgáltuk. Az előfeldolgozó szkriptek és az eredményfájlok elérhetők a `https://github.com/sassbalint/double-cube-jump-and-stay-multilingual` címen. Jelen cikk az `5dde1d7` commit azonosítójú verzióval készült.

4. Eredmények

Az eredményül kapott szerkezetek túlnyomó többsége megfelelő valódi igei szerkezet. Az 1. táblázatban egy mutatvány látható különféle szerkezetekből.

#	nyelv	igei szerkezet	magyar megfelelő
1.	cs	'být SUBJ:rozdíl mezi'	(van különbség vmi között)
2.	cs	'investovat do'	(befektet vmibe)
3.	cs	'stát se OBJ'	(válik vmivé)
4.	de	'fallen SUBJ:aktie auf'	(esik részvény vmire)
5.	de	'finden sich SUBJ:information auf'	(megtalálható információ vhol)
6.	de	'handeln sich um'	(arról van szó)
7.	en	'do IOBJ OBJ:favor'	(szívességet tesz vkinek)
8.	en	'get in touch with'	(kapcsolatba lép vkivel)
9.	en	'make sure'	(meggyőződik)
10.	en	'take OBJ:care of'	(vigyáz vmkire)
11.	fi	'ottaa Ill:huomio OBJ'	(figyelembe vesz vmit)
12.	fi	'ottaa Ill:käyttö OBJ'	(használatba vesz)
13.	fi	'ottaa Ill:käsi OBJ'	(kézbe vesz vmit)
14.	hu	'lesz SUBJ:szükség -rA'	
15.	hu	'tesz lehetővé -t'	
16.	nl	'zien OBJ:kans te'	(lát lehetőséget vmit csinálni)
17.	no	'få OBJ med seg'	(magával visz vmit)
18.	no	'få OBJ:gjennomslag i'	(áttörést ér el vmiben)
19.	no	'få OBJ på seg'	(felvesz vmit (ruhaféleséget) magára)
20.	no	'få OBJ:tillit'	(önbizalma lesz)
21.	no	'få OBJ i løp'	(futtat vmit (szoftvert))
22.	no	'ha OBJ på seg'	(vmi (ruhaféleség) van rajta)
23.	wo	'am OBJ:kàttan ci'	(van energiája vmit csinálni)
24.	wo	'wax IOBJ OBJ'	(mond vkinek vmit)

1. táblázat. Mutatvány a kapott szerkezetekből. A kitöltetlen alanyi helyet nem tüntetjük fel.

Jópár igénél az összes kijövő szerkezet jó. '*look*' → '*look*', '*look good/great*', '*look like*'; a '*deal*' egyetlen szerkezete a '*deal with*'; a '*go*' esetén lényegében azok a szerkezetek jöttek ki, amelyeket korábban említettünk (2. oldal).

Az eredményekből kontrasztív tanulságokat is le lehet vonni. A 2. táblázatban a '*beszél -rÓl*' szerkezet látható az egyes nyelveken. A magyar kivételével valamennyi szerkezetet az „ugrik és marad" algoritmus futtatásának eredményeként kaptuk. A cseh '*čekat*' (vár) összes szerkezete helyes: '*čekat*', '*čekat Acc=na*' (vár -rA), '*čekat OBJ*' (vár -t), '*čekat se*' (várandós). A neki megfelelő német '*warten*' igénél hasonló szerkezeteket kapunk: '*warten*', '*warten Acc=auf*'. Érdekes kérdés, hogy mennyire feleltethetők meg egymásnak a különböző nyelvek preopozíciói. A cseh '*mít OBJ Dat=k:dispozice*' és a német '*stehen OBJ Dat=zu:Verfügung*' (rendelkezésre áll) párhuzama arra utal, hogy a '*Dat=k*' és

hu	beszél	-rÓl
cs	mluvit/hovořit	o
de	sprechen	von
en	talk	about
fi	puhua	Ela
nl	praten	over
no	snakke	om

2. táblázat. A '*beszél -rÓl*' szerkezet megfelelői. Látjuk, hogy a két finnugor nyelv eseteket használ, az indoeurópai nyelvek pedig különféle elöljárókat.

a '*Dat=zu*' megfelel egymásnak. Ezt alátámasztja a '*patřit Dat=k*' – '*gehören Dat=zu*' (tartozik vmihez) pár is.

Az UD korpuszok viszonylag kis méretűek, ezáltal sok esetben nem kiegyensúlyozottak. A német korpusz szerkezetei arra utalnak, hogy a szövegei főként számítástechnikai területről származnak: '*arbeiten unter windows*', '*laufen mit mhz*', '*laufen unter mac*', '*laufen auf system*', '*laufen unter windows*'.

Az implementált névmástörlés jól működik, nyilvánvalóan nincsenek személyes stb. névmást tartalmazó szerkezetek, ugyanakkor a visszaható névmásos szerkezetek jelentős számban megjelennek (ld. pl. az 1. táblázat 3., 5., 6., 17., 19. és 22. szerkezetét). Az „ugrik és marad" módszer ismert korlátai megjelennek: előfordulnak a szerkezetekben gyakori, jellegzetes de nem idiomatikus szavak kitöltőként. A norvég '*få øye på*' (pillantást vet vmire, meglát vmit) kifejezés például '*få øye på løve*' (meglátja az oroszlánt) formában jelenik meg elsősorban a kicsi, az eredeti cikkhez képest két nagyságrenddel kisebb korpuszméret miatt.

5. Összefoglalás

Jelen cikkben egy eredetileg csak magyar nyelvre alkalmazott valódi igei szerkezeteket kinyerő algoritmus – az „ugrik és marad" módszer – nyelvfüggetlenségét vizsgáltuk meg. A módszer csupán a predikátum-argumentum struktúra meglétét követeli meg, így remélhető volt, hogy szinte bármely nyelvre működőképes lesz. Nyolc európai nyelv függőségileg elemzett UD korpuszából nyertük ki az algoritmus bemenetéhez szükséges adatokat. Az UD korpuszok előfeldolgozása során jónéhány helyen ütköztünk a korpuszok nem teljesen egységes, nem teljesen univerzális annotációjából adódó problémákba. Ezeket részletesen elemeztük. Az algoritmus lefuttatása révén helyes valódi igei szerkezeteket kaptunk felügyeletlen módon. Elmondhatjuk, hogy az absztraktban felvetett mindkét állítás megállja a helyét: viszonylag egyszerűen elő lehet állítani függőségileg elemzett korpuszból az algoritmus bemenetét; valamint hogy az algoritmus valóban lexikográfiailag is hasznos valódi igei szerkezeteket szolgáltat számos nyelven, függetlenül attól, hogy az eredeti tanulmányhoz képest két nagyságrenddel kisebb korpuszokkal dolgoztunk. A jövőben tervezzük a módszer nagyobb elemzett korpuszokon való kipróbálását. Eredményeink megteremtik az

alapját annak, hogy szinte tetszőleges, akár kisebb, kevesebb erőforrással bíró nyelvek tipikus igei szerkezeteit összegyűjtsük. A kód és az eredmények elérhetők `https://github.com/sassbalint/double-cube-jump-and-stay-multilingual` címen.

6. Köszönetnyilvánítás

A kutatást az MTA Bolyai János Kutatási Ösztöndíja támogatta (ügyszám: BO/00064/17/1; időtartam: 2017-2020). Az Információs és Technológiai Minisztérium ÚNKP-19-4 kódszámú Új Nemzeti Kiválóság Programjának szakmai támogatásával készült.

Hivatkozások

Atkins, B.T.S., Rundell, M.: The Oxford Guide to Practical Lexicography. Oxford University Press (2008)

Croft, W., Nordquist, D., Looney, K., Regan, M.: Linguistic typology meets universal dependencies. In: Dickinson, M., Hajic, J., Kübler, S., Przepiórkowski, A. (szerk.) Proceedings of the 15th International Workshop on Treebanks and Linguistic Theories (TLT15). pp. 63–75 (2017)

Kilgarriff, A.: "I dont't believe in word senses". Computers and the Humanities 31(2), 91–113 (1997)

Nivre, J., Abrams, M., Agić, Ž., et al.: Universal Dependencies 2.4, LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University (2019), http://hdl.handle.net/11234/1-2988

Osborne, T., Gerdes, K.: The status of function words in dependency grammar: A critique of universal dependencies (UD). Glossa: A Journal of General Linguistics 4(1), 17 (2019)

Pustejovsky, J.: The generative lexicon. Cambridge, MA, US: The MIT Press (1995)

Sass, B.: Az igei szerkezetek algebrai struktúrája, avagy a duplakocka modell. Argumentum 14(1), 12–44 (2018)

Sass, B.: The 'jump and stay' method to discover proper verb centered constructions in corpus lattices. In: Proceedings of RANLP 2019. pp. 1076–1084. Varna, Bulgaria (2019)

Shi, X., Padhi, I., Knight, K.: Does string-based neural MT learn source syntax? In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. pp. 1526–1534. Association for Computational Linguistics, Austin, Texas (Nov 2016), https://www.aclweb.org/anthology/D16-1159

Yarowsky, D.: One sense per collocation. In: Proceedings of the workshop on Human Language Technology. pp. 266–271. Princeton, New Jersey (1993)