Egy nagyobb magyar UD korpusz felé

Novák Attila[1,2], Novák Borbála[1,2]

[1]Pázmány Péter Katolikus Egyetem, Információs Technológiai és Bionikai Kar
[2]MTA-PPKE Magyar Nyelvtechnológiai Kutatócsoport
Budapest, Práter u. 50/a.
{vezetéknév.keresztnév}@itk.ppke.hu

Kivonat A cikkben egy olyan munkálat eredményeiről számolunk be, amelynek keretében a teljes Szeged Dependencia Treebanket megpróbáljuk a Universal Dependencies projekt keretében megfogalmazott annotációs elveknek megfelelő annotációjú korpusszá átalakítani, miközben az eredeti korpuszban szereplő annotációs hibákat, illetve következetlenségeket is igyekszünk feltárni, kijavítani, illetve megszüntetni.
Kulcsszavak: függőségi annotáció, Szeged Dependencia Treebank, Universal Dependencies,

1. Bevezetés

Két évvel ezelőtt egy olyan munkálatról számoltunk be (Novák és mtsai, 2019), amelynek keretében egy olyan függőségi alapú annotációs séma kialakítására tettünk kísérletet, amely a magyarra (illetve általában) korábban használt sémáknál jóval részletgazdagabb elemzést tartalmaz. Konkrétan az volt a célkitűzésünk, hogy az annotáció alkalmas legyen arra, hogy releváns kérdéseket lehessen a felhasználásával az adott szöveggel kapcsolatban feltenni. Így az annotációban használandó megkülönböztetések létjogosultságát is alapvetően az határozza meg, hogy az adott konstrukcióval kapcsolatban milyen kérdéseket lehet feltenni.

Akkor kiindulási anyagként a Universal Dependencies (UD) korpusz (Nivre és mtsai, 2020) 1800 mondatból (42000 token) álló magyar alkorpuszát használtuk(Vincze és mtsai, 2017), hogy nemzetközi szinten elfogadott annotációs sémából induljunk ki. Az UD korpusz nagyjából egységes elvek és kategóriák felhasználásával sok nyelv szövegeire tartalmaz morfoszintaktikai és szintaktikai függőségi elemzést. Mivel a magyar alkorpuszban szereplő annotáció sok szempontból nem felelt meg az érvényes UD specifikációnak, illetve sok véletlenszerű annotációs hibát találtunk, ezért már akkor foglalkoztunk a hibajavítással. A szerkezetek egy részét (pl. **appozitív szerkezetek** (*Katona Kálmán közlekedési minisztert*), **egyeztetett predeterminánst** tartalmazó szerkezetek (*azt a kutyát*), **birtokos szerkezetek**, **névutós szerkezetek**) programozottan alakítottuk át. Más esetekben (pl. a harmadik személyű **névszói állítmányt** tartalmazó tagmondatok annotációjában az alany és az állítmány sok esetben meg volt cserélve) félig manuális módszerrel tudtuk javítani az annotációt: kézzel jelöltük meg a hibás mondatokat, ahol aztán az alany és állítmány annotációját programozottan javítottuk.

Miután a magyar UD alkorpusz annotációját kibővítettük a Novák és mtsai (2019)-ben említett annotációs elemekkel (a vonzatok és a szabad határozók egy meghatározott körének tematikus szerepére vonatkozó annotációval) kiegészítettük, megnéztük, hogy ezen az annotáción egy korszerű neurális függőségi parsert betanítva milyen teljesítményt tudunk elérni (Dozat és mtsai (2017)). Az eredmény kiábrándító volt (LAS=0.57). Ugyanakkor a korpusz mérete (a tanítóanyag mindössze 900 mondat) az UD specifikációban definiált alaprelációk tekintetében (ha olyan megkülönböztetéseket nem vesszük figyelembe, mint pl. a vonatkozó mellékmondatok megkülönböztetése `acl:relcl` az `acl` reláción belül) sem tesz lehetővé LAS=0.8 fölötti eredményt a legkorszerűbb parserek esetében sem, illetve a csak a tartalmas szavakat és a morfológiai annotációt is figyelembe vevő MLAS pontszám esetében MLAS=0.67 a legjobb eredmény (Zeman és mtsai, 2018). Ez az eredmény indított arra, hogy belekezdjük a teljes Szeged Dependencia Treebank Vincze és mtsai (2010) UD-vel kompatibilis formára hozásába.

2. Problémák és megoldások

A teljes Szeged Dependencia Treebank (SZDT) mintegy 82000 mondatból (1,5 millió tokenből) áll, így ígéretes méretű anyagnak tűnik egy viszonylag pontos függőségi elemző betanításához. Ugyanakkor számos az eredeti korpusszal, illetve az abban szereplő annotációval kapcsolatos probléma nehezíti a megfelelő minőségű UD-kompatibilis korpuszváltozat létrehozását. Alább áttekintjük ezeket a problémákat, illetve hogy milyen megoldást próbáltunk alkalmazni rájuk.

2.1. Hibás szóalakok

Nincs elemzés A helyesírási hibát tartalmazó szövegrészek nincsenek morfológiailag annotálva – legalábbis ahol a hibát az annotációs folyamat során észlelték. Ezekben az esetekben az morfológiai annotáció mindössze annyit tartalmaz, hogy a szó ‘hibás’ ((1) 2. token) vagy ‘az adott kontextusban hibás’ ((1) 3-4. token), de se elemezve, se lemmatizálva nincs.

(1)
```
# sent_id = 10elb.ud - 11
# text = De végülis oda értünk, mert jött az egyik osztálytársam ...
1 De de CONJ CONJ _ 4 CONJ _ _
2 végülis végülis X X _ 4 MODE _ _
3 oda oda X X _ 4 PREVERB _ _
4 értünk értünk X X _ 0 ROOT _ _
...
```

Sok esetben a hiba nem is az eredeti szövegből származik, hanem a feldolgozási folyamatba csúszott hiba eredménye (pl. az *1984* alkorpuszban minden kurzív szövegrész kezdetén álló szó a feldolgozás során egybeíródott az előző szóval, és ennek a szisztematikus alkorpusz-specifikus tokenizálási hibának a következményeit később nem javították ((2) 2. és 10. token).

(2) # sent_id = 1984.ud - 683
text = Winston arégi számok-at tárcsázta a teleképen, és kérte aTimes megjelölt számait.
1 Winston Winston PROPN PROPN Case=Nom|Number=Sing 4 SUBJ _ _
2 arégi arégi X X _ 3 ATT _ _
3 számok-at számok-at X X _ 4 OBJ _ _
...
10 aTimes aTimes X X _ 12 ATT _ _
...

A *Piszkos Fred* alkorpusz esetében „az író szándéka" vezetett rengeteg agrammatikus szöveg létrehozásához, ez az anyag azonban nem különösebben hasonlít ahhoz, ahogy valódi emberek hibáznak, tehát az sem világos, hogy a parser mit is tanulhatna pontosan ezekből a szövegrészekből.

A probléma nagyságrendjét érzékelteti, hogy a korpusz mondatainak több mint 10%-a tartalmaz hibásként annotált, morfológiailag nem elemzett szóalakokat.

Zárólbe tett hibák Az iskolások fogalmazásainál az adatbevitel során is torzult az anyag (azon túl, amikor a diákoknak nem sikerült azt leírniuk, amit akartak). Egyrészt bizonyos szövegrészeket nem sikerült az adatbevivőknek elolvasni, és itt néha értelmezhetetlenül hiányos vagy torz mondatok kerültek a korpuszba (3a). A tanulók által zárójelbe tett szövegrészek ezzel szemben mind bekerültek a korpuszba, bár a diákok a zárójelbe tétellel a legtöbb esetben azt jelölték, hogy a szövegrész törlendő (3b). Később az annotátorok kénytelenek voltak ezekkel a részekkel is kezdeni valamit. AZ SZDT függőségireláció-készletében nincs olyan elem, ami kifejezetten a hibás elemek megjelölésére szolgálna. A zárójelbe került kifejezés feje 'hibás' „szófajcímkét" kapott, és az `APPEND` relációval csatolták a mondat többi részéhez (pl. a (3b) mondatban a *nemcsa*, az *Én* ennek az alanyaként van megjelölve). Az `APPEND` reláció ugyanakkor a legtöbb esetben valóban a szöveg részét képező elemeket jelöl, sokszor még a 'hibás' „szófajcímkéjű" és `APPEND` relációval csatolt szavak esetében is. A 'hibás' „szófajú" `APPEND` relációval csatolt fejű zárójelbe tett szövegrészeket azonban nagy biztonsággal törölni lehet a szövegből. Legjobban ebben az esetben járunk, mert a 'hiba' mint szófajcímke megtartásának nincs értelme, és az iskolai kontextuson kívül nem fordul elő, hogy a zárójelbe tett részek azt jelentenék, hogy úgy kell érteni, mintha az a szövegrész oda se lenne írva, ezért nem igazán van értelme erre betanítani egy parsert.

(3) a. (10erv.ud - 3158) Továbbá még véleményezem azt is, hogy a reális **XXX** sem **XXX** tantárgyakból kevesebb ill. a humán **XXX** fordítva.

b. (10elb.ud - 7819) **(Én nemcsa)** Egy vasárnapi nap volt.

Nem annotált hibák Az elírások egy részét nem vették észre az annotáció során. Ezekben az esetekben a tokenizálás, illetve a lemma hibás.

Szétvágott és egyben maradt mondatok A korpusz elírásokból és az automatikus mondatrabontás hibáiból fakadóan 95 olyan mondatnak annotált egységet tartalmazott, amely valójában több mondatból állt. Emellett néhány olyan esetet is találtunk, ahol egy összefüggő mondat szakadt meg egy szó közepén.

Megoldás A hibás szóalakokkal kapcsolatos problémák megoldására azt láttuk célravezetőnek, hogy létrehozzuk a hibás szövegek javított változatát, és ezt annotáljuk (nyilvántartva, hogy mit javítottunk). A hibásnak annotált szóalakot tartalmazó mondatokat kigyűjtöttük a forrásfájl és a mondatazonosító megjelölésével, és létrehoztuk az eredeti mondat normalizált változatát olyan formában, hogy abból rekonstruálható, hogy az eredeti mondatban mely tokeneket módosítottuk, töröltük, vontuk össze, választottuk részekre, illetve hova szúrtunk be esetleg új tokent. A javított/normalizált változat kezdeti verzióját a gyakori hibákhoz készített automatikus hibajavítási lista alapján generáltuk az eredetiből (tehát ez már tartalmazott lényegében biztosra vehető javításokat). A hibásnak jelölt szóalakok ki voltak emelve, de a mondatokban észrevett egyéb hibákat is korrektúráztuk. Ebben a fázisban a mondathatárok javításán túl 9400 javítást vezettünk be az anyagba (1367 betűköztörlés, 1051 szó/tokentörlés, 1382 betűközbeszúrás, 321 szóbeszúrás, 5281 szójavítás).

A javítási lista alapján programozottan vezettük vissza a korpuszba a javításokat. A szétvágott elemeket jobb fejűnek feltételeztük és az elemek között alapesetben `NE` , illetve `ATT` relációt feltételeztünk attól függően, hogy névnek vagy más elemnek tűnt az eredeti token. A módosított tokeneket morfológiailag elemeztük, és a korpusz többi részén betanított PurePos taggerrel (Orosz és Novák, 2013) egyértelműsítettük. Az automatikus morfológiai annotáció és a függőségi címkék kézi átnézése/javítása folyamatban van.

2.2. Az igék, igenevek és ragozott névmások annotációja

Igenevek Az igenevek minden esetben melléknévként vagy határozószóként vannak annotálva. Az eredeti SZDT-ben az annotációból az sem derül ki, hogy igenévről van szó. A magyar UD korpuszban az igenév típusára utaló jegy szerepel morfoszintaktikai jegyek között. Ez sem volt azonban elegendő, amikor a bevezetésben említett kutatásunkban (Novák és mtsai, 2019) igei vonzatkereteket próbáltunk a szövegre illeszteni. Az illesztéshez az igenév igei tövére is szükség van az igenév típusa mellett a megfelelő vonzatkeret azonosításához. Természetesen problémát jelent, hogy számtalan lexikalizált melléknév van, amelyek formailag azonosak valamely igenévvel, azonban ezek az igenevektől eltérően általában predikatívan is használhatóak (pl. *derült, tartózkodó* stb.) Ezen elemek nem predikatív előfordulásainak egyértelműsítése azonban sajnos csak manuális ellenőrzéssel valósítható meg.

„Ható", műveltető és gyakorító igealakok Bizonyos teljesen produktív igei végződések (az egyértelműen inflexiós jellegű *-hAt* végződés mellett pl. a műveltető- és a gyakorítóképző) nincsenek leválasztva a lemmáról, és nem jelennek meg a morfológiai elemzés szintjén (4).

(4) adhat adhat VERB VERB Definite=Ind|Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin|Voice=Act

Ez az igenevek elemzetlenségéhez hasonlóan szintén problémát jelent a vonzatkeret-illesztés szempontjából. A *-hAt* végződés le nem vágása a lényegében önállósult lexikai elemmé vált *lehet* igétől eltekintve nem igazán tűnik indokoltnak. A gyakorító-, illetve műveltetőképzős alakok között ennél jóval több lexikalizált elem található (*mosogat, beírat stb.*), tehát ezek egyben tartása az eredeti SZDT-ben motiváltabb döntés volt.

Ragozott névmások A ragozott névmások annotációja nem felel meg sem a magyar morfológiai elemzők által adott, sem az UD specifikációban szereplő névmás-annotációs elveknek sem a lemma, sem a morfoszintaktikai jegyek tekintetében (pl. nem derül ki az eset).

(5) nekem neki ADV ADV Number=Sing|Person=1|PronType=PrsPron

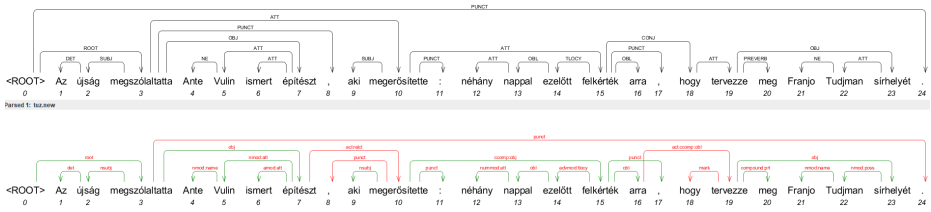
Egyéb szófajproblémák Sok funkciószó szófaji annotációja nem felel meg az intuíciónknak, illetve a morfológiai elemző által adott elemzéseknek (főleg a névmási jellegű és a kötőszószerű elemek, illetve egyes névszói eredetű névutószerű vonzatos határozószók esetében).

Megoldás A korpusz anyagát morfológiailag újraelemeztük, és a meglévő annotációban az egyes tokenekhez rendelt lemma és morfoszintaktikai annotáció alapján kiszámítható annotációhoz leghasonlóbb, illetve azzal kompatibilis (ideális esetben azzal azonos) elemzést választottuk. Sajnos vannak olyan lényeges többértelműségek, amelyeket az eredeti annotáció neutralizál (pl. melléknévként lexikalizálódott igenevek, névutószerű vonzatos határozószóként vagy időhatározóként lexikalizálódott ragozott főnevek stb.) Sajnos a szótípusok szintjén is tízezres nagyságrendű listát kell átnézni, ez jelenleg folyamatban van. (Annak idején az e-magyar projekt (Váradi és mtsai, 2017) keretében készült egy konverzió az SZDT anyagáról az emMorph (Novák és mtsai, 2017) elemző által használt formátumra, azonban abban sok a hiba, sok esetben nem is a korpuszban szereplőnek megfelelő elemzésre történt a leképezés, ezért nem ebből az anyagból indultunk ki). A nem egyező elemzések fele a névelemek annotációjához köthető (l. a 2.4. részben).

2.3. Függőségi relációk

Egy-több megfeleltetés a relációk között A UD-ben és a SZDT-ben használt függőségi relációk halmaza nem feleltethető meg egymásnak egyértelműen. Az alap UD készlet pl. számos tagmondatok közötti relációs viszonyt megkülönböztet (az alárendelt tagmondat mondatbeli szerepétől függően), ugyanígy a frázisszintű módosítók típusait is a módosító szófaja szerint megkülönbözteti. Ezeken belül pedig opcionálisan további altípusokat lehet megkülönböztetni. Az SZDT-ben minderre egyetlen reláció szolgál, az `ATT`.

Az alárendelő mellékmondatok csatolása Az SZDT-ben minden alárendelő mellékmondat az azt tartalmazó tagmondat fejéhez van csatolva (a vonatkozó mellékmondatok is). Ez legalábbis a vonatkozó mellékmondatok tekintetében határozottan nem felel meg az UD-ben megfogalmazott elveknek. Ráadásul semmi nem utal az annotációban arra, hogy a főmondatban sok esetben jelen lévő utalószónak és a hozzá tartozó alárendelő mellékmondatnak bármi köze lenne egymáshoz. Bár az UD elvek szerint az utalószók esetében ennek így is kellene lennie, és valószínűleg inkább expletív névmásokként kellene őket annotálni, végső soron nem lehet megúszni, hogy ha bármire is akarjuk használni az annotációt, összekapcsoljuk megfelelő elemeket egymással. Megoldásunkat és az előbbi *egy-több megfeleltetés*-problémát az 1. ábra illusztrálja. Az ábrán látható, hogy az SZDT 'all-in-one' `ATT` relációja a belőle automatikusan generált UD annotációban számtalan különböző reláció alakját ölti. A ábrán látható rövid mondatban az eredeti mondat `ATT` típusú dependensei 6 merőben különböző szerepet kapnak: van itt melléknévi jelző (`amod:att`), NP módosító (`nmod:att`), vonatkozó mellékmondat (`acl:relcl`: ezt a konverzió során automatikusan helyesen átkötjük az általa ténylegesen módosított NP fejére), számnévi módosító (`nummod:att`), tárgyi alárendelő mellékmondat (`ccomp:obj`), oblikvuszi alárendelő mellékmondat (`acl:ccomp:obl`: ezt is automatikusan átkötjük a vonzat esetét hordozó megfelelő utalószóra) és birtokos (`nmod:poss`).



1. ábra. Az SZDT 'all-in-one' `ATT` relációja 6 különböző relációvá lényegül át egy röpke mondat UD-beli reprezentációjában.

A relációk iránya és topológiája A relációk feje, illetve a komplexebb szerkezetekhez (pl. a többszörös mellérendeléshez) rendelt annotáció topológiája szempontjából is sok alapvető eltérés van az UD és az SZDT között. Az UD-ben *a tartalmas szó a fej*-elv érvényesül, így pl. a kopulás és a névutós szerkezetek feje is a névszó. A kopulás, illetve a névutós szerkezetek átalakítása általában viszonylag problémamentesen megoldható, a mellérendeléssel kapcsolatos problémákra alább részletesebben kitérünk (2.5).

Üres elemek Az SZDT beszúrt üres elemekkel (`VAN, ELL`) operál a zérókopula és az ellipszis tekintetében is. Ez mindenképpen problémát jelent a parserek működése szempontjából, hiszen a nyers szövegből nyilvánvalóan hiányoznak ezek

a testetlen elemek. Az elemzés folyamán valamikor valahogyan a szövegbe kellene kerülniük ahhoz, hogy SZDT-kompatibilis elemzés jöhessen létre. Bár azzal kapcsolatban a magyar beszélőknek világos intuíciója van, hogy hol lenne a zérókopula helye egy konkrét magyar nyelvű tagmondatban, vagy hogy honnan hiányzik egy elliptált elem, a treebankbe ezek az elemek teljesen véletlenszerű helyekre lettek beszúrva, néha nem is abba a tagmondatba, ahova tartoznak. Sajnos az ellipszis jelölése az SZDT-ben nemcsak a törölt elem helyének jelölése tekintetben következetlen: sok ténylegesen elliptikus szerkezetben egyáltalán szerepel az `ELL` elem.

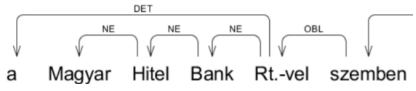
A zérókopula kezelését az UD megnyugtatóan megoldja. Ugyanakkor az elliptikus szerkezetek annotációjára az UD 2-es verziójában elfogadott megoldás (a törölt elem legprominensebb dependensének fejjé való előléptetése, és a többi elem hozzácsatolása az `orphan` reláció használatával) önmagában nem teszi lehetővé a tényleges viszonyok visszafejtését az annotációból. Ráadásul ha csak egy konstituens marad az elliptikus tagmondatban, egyáltalán nem derül ki az annotációból, hogy ellipszisről van szó.

Megoldás A többértelmű relációk konverziójánál az adott token és konstrukciót alkotó egyéb elemek morfológiai/lexikai tulajdonságaira támaszkodva egyértelműsítjük a szerkezetet, és ezután alakítjuk át a megfelelő topológiájúvá az annotációt (fej-dependens viszony megfordítása, az eredeti fej nem-lokális dependenseinek átkötése az új fejre (vagy az összesé a konstrukciótól függően), adott esetben láncolt szerkezetek átalakítása lapos szerkezetté). Az átalakítás az egyértelmű (pl. zérókopulás/névutós/mutató determinánsos) szerkezetek esetében hasonlóan megy. Az alárendelő mellékmondatok átcsatolására készített megoldásunk eredményét az 1. ábrán szemléltettük. A mellékmondatok típusának azonosításához a mellékmondatot tartalmazó mátrixtagmondatban keresünk utalószót (itt számos esetben a mellékmondat kötőszavára támaszkodhatunk, pl. *akkor-amikor, annál-minél, ott-ahol,* stb.), illetve az utalószóként szóba jöhető deiktikus névmási elem és a tagmondat távolságát és sorrendjét is figyelembe vesszük az alkalmazott heurisztikákban. Vonatkozó mellékmondatok esetében a legközelebbi névszói fejet célozzuk meg, ha utalószóhoz való csatolás nem lehetséges. Tárgyi alárendelő mellékmondatot feltételezünk a *hogy* kötőszós mellékmondatoknál, ha a mátrix ige definit ragozású, és nincs explicit tárgya. Egyébként a megtalált utalószó esete, illetve a kötőszó alapján döntünk a mellékmondat típusáról.

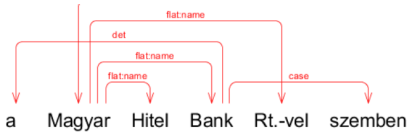
2.4. Névelemek

Az SZDT-ben minden névszerű kifejezés (beleértve a művek stb. címeit is) minden eleme tulajdonnév szófajcímkével van címkézve (a funkciószavak is), és a kifejezés elemei koordináció kezeléséhez hasonlóan láncba vannak fűzve, csak ebben az esetben a lánc jobb fejű (2a ábra). Az UD 2 specifikáció szerint a belső szerkezet és fej nélküli névkifejezéseket a koordinációhoz hasonlóan fixen bal fejű kvázilapos szerkezetként kell ábrázolni (más persze a függőségi viszony, mint a

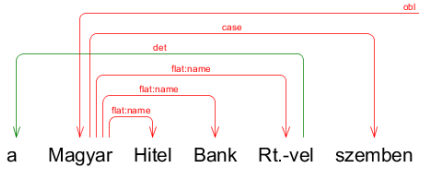
koordinációnál: (2c ábra)). Mivel a magyarban az NP-k minden esetben a leghatározottabban jobb fejűek, ezért ez a javasolt annotáció ilyen formában biztosan nem használható (pl. a név esetét adó névutó, amely maga instrumentalist vonz, az alanyesetben álló első szóhoz kapcsolódik, ami nonszensz), helyette például valamilyen jobb fejű lapos annotáció használata lenne indokolt (2d ábra)). Jelenleg az UD Treebankben egyébként a 2b ábrán látható teljesen zűrzavaros szerkezet szerepel. Ezt feltehetőleg az UD 1.0-s változatban szereplő eredetileg helyesen jobb fejű szerkezeteket tartalmazó annotáció programozott „megrongálásával" hozta létre az UD treebankek valamelyik magyarul nem tudó adminisztrátora. Ugyanakkor az UD 2 specifikáció szerint a világos belső szintaktikai szerkezettel és függőségi viszonyokkal rendelkező névelemek annotációjának ezeket a viszonyokat kellene tükröznie (2e ábra)).



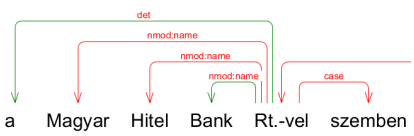
(a) SZDT változat



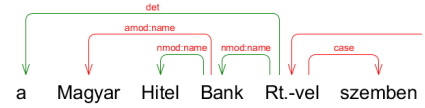
(b) Az UD Treebankben jelenleg fent lévő zűrzavaros változat



(c) A lapos névszerkezetet feltételező bal fejű „UD-kompatibilis" változat



(d) Jobb fejű lapos változat



(e) A névelem belső szerkezetét ténylegesen tükröző UD-kompatibilis változat

2. ábra. Egy névelem lehetséges (d,e) (és nemkívánatos (b,c)) ábrázolásmódjai

A világos belső szintaktikai szerkezettel rendelkező nevek, pl. a művek címeinek egy része a magyar esetében egyébként problémát jelent abból a szempontból is, hogy míg maga a cím jobb fejűnek tekinthető (lévén kívülről nézve egy NP), a belső szerkezet feje egész más lehet (pl. egy ige), és így a jobb szélen álló elem akár egyszerre két esetben állhat (6a,6c). Tulajdonképpen ezekben az esetekben egyfajta lexikai ellipszisről van szó (6b,6d).

(6) a. Stohl András párja lesz Éder Enikő a *Hegedűs a háztetőn*-ben.

b. Stohl András párja lesz Éder Enikő a *Hegedűs a háztetőn* című musicalben.

c. A HBO Max levette az *Elfújta a szélt/szelet.*

d. A HBO Max levette az *Elfújta a szél* című filmet.

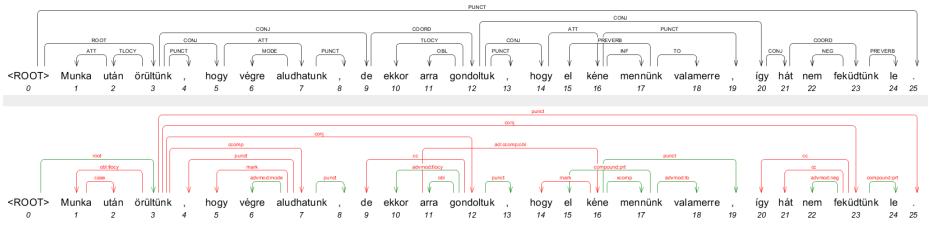
Megoldás Az névelemek szétválogatása és a fenti szempontok szerinti újraelemzése mindenképp jelentős erőfeszítést igényel. Az ehhez szükséges munkába egyelőre nem vágtunk bele: a jelenleg változatban a jobb fejű lapos annotációnál maradtunk. A típusok szintjén a morfológiai elemző elemzéseitől való eltérések fele a névelemek annotációjából adódik.

2.5. Mellérendelés

A mellérendelésről azt tanultuk, hogy exocentrikus szerkezet, és mint ilyen alapvetően problematikus a kizárólag endocentrikus szerkezetekben gondolkodó függőségi grammatika számára. Ezért bármit is teszünk, az annotáció mindenképp önkényes lesz, különös tekintettel arra, hogy a mellérendelésnek kénytelen lesz feje és iránya lenni. UD specifikációjában szereplő tartalmas fejeket összekötő kvázilapos szerkezet (minden mellérendelt elem közvetlenül az önkényesen kijelölt fejhez van kötve) elvileg kevesebb problémát okoz, mint az SZDT-ben alkalmazott vegyes láncba fűzött és a mellérendelt elemet a kötőszón keresztül csatoló megoldás. Az előbbiben ugyanis minden mellérendelt konstituens feje maximum egy lépés távolságra van a mellérendelő szerkezet önkényesen kijelölt fejétől, és így determinisztikus módon elérhetők mindegyik elemből a szükséges információk, míg a második megoldásban nem korlátos az adott koordinált elem és a valódi grammatikai szerepére vonatkozó információk helye közötti távolság.

Egyik módszer sem teljesen alkalmas a különböző egymásba ágyazott szerkezetek megkülönböztetésére. Az UD-ben alkalmazott reprezentáció esetében az A,B,C és az (A,B),C szerkezethez tartozik azonos fa, az A,(B,C) szerkezeté különbözik tőlük. Az SZDT-ben alkalmazott láncolt megoldásnál az első és a harmadik ad azonos szerkezetet, a második különbözőt. Ezért nem lehet az SZDT-ben alkalmazott láncolt szerkezeteket 'ész nélkül' az UD specifikációban ajánlott szerkezetekké alakítani. A 3. ábrán látható, hogy a valójában A,(B,C) típusú szerkezetet ábrázoló SZDT annotációt A,B,C szerkezetűnek feltételezve és azt az UD specifikáció által javasolt formára alakítva az ábrán alul szereplő annotációt kapjuk, amiről azt a badarságot olvashatjuk le, hogy a mondat szerint *Munka után örültünk, hogy végre aludhatunk, így hát nem feküdtünk le.* A korpuszban a kettőnél több tagmondatot tartalmazó konkrét tagmondatmellérendelés-példákat nézegetve nagyon hamar arra a következtetésre juthatunk, hogy az esetek nagyobb részében teljes zagyvaság jön ki, ha hozzányúlunk a tagmondatok közötti viszonyokhoz és A,B,C mellérendelést feltételezve átkötögetjük a harmadik tagmondatot az elsőre. Az SZDT-ben szereplő mellérendelés-annotáció

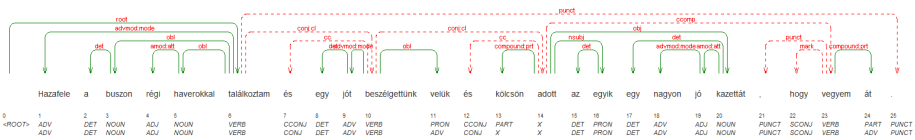
általában jól tükrözi a diskurzusszerkezetet, a tagmondatok közötti időbeli, illetve a logikai viszonyokat. Általában ténylegesen az előző tagmondathoz csatlakozik a következő mondanivalója, ha meg nem, azt ez a típusú annotáció is jól ki tudja fejezni az (A,B),C szerkezet megadásával. Tagmondatok esetében, úgy tűnik, nem sok értelme van többszörös mellérendelésről beszélni, mert még explicit kötőszó hiányában is gyakran legalábbis az események sorrendjét tükrözi a tagmondatok sorrendje. Ezért nincs értelme a tagmondat-mellérendelések annotációjába belepiszkálni.



3. ábra. Mi lesz, ha az SZDT mellérendelt tagmondatait ‘ész nélkül’ az UD specifikáció szerinti formára hozzuk: *Munka után örültünk, hogy végre aludhatunk, így hát nem feküdtünk le.*

A névszói kifejezések koordinációjánál ugyanakkor megfontolandó lehet az UD specifikációban javasolt mellérendelés-annotációra való áttérés. Az UD és az SZDT annotáció különös közös fogyatékossága, hogy a frázis- és a tagmondat-koordináció nincs megkülönböztetve: ugyanazt a relációt használják a két viszony megjelölésére. Ez például annak az egy szintaktikai annotációval szemben támasztható viszonylag alapvetőnek tűnő elvárásnak a teljesíthetőségét is megkérdőjelezi, hogy a tagmondathatárokat az annotáció alapján meg lehessen állapítani. Ugyanakkor az UD specifikáció megengedi az alaprelációkon belül altípusok létrehozását, tehát megkülönböztethetünk egy `conj:ph` frázis-mellérendelés és egy `conj:cl` tagmondat-mellérendelés relációt. Elvileg az összekapcsolt fejek kategóriája alapján meg lehet különböztetni a kettőt: a tagmondatok feje általában ige. Az UD annotációban azonban a névszói állítmányok és az ellipszis esetében ez nincs így, bár az utóbbiaknál szerencsés esetben egy `orphan` reláció jelenlétéből meg lehet állapítani, hogy ellipszist tartalmazó tagmondatról van szó (ha egynél több konstituens maradt ‘árván’ az elliptikus tagmondatban).

Az SZDT annotáció esetében az okoz problémát, hogy az ellipszis időnként nincs jelölve, illetve hogy a ‘hibás szó’ szófajjelölést kapott szavak szófaját nem lehet az annotációból kiolvasni. Szerencsés esetben azonban legalább a koordinációt kifejező `conj` reláció fejének annotációja használható, ami alapján mégis meg tudjuk különböztetni a frázis- és a tagmondat-koordinációt (így azonosíthatjuk, hogy a 4 ábrán látható mondatban a *beszélgettünk* és a hibásnak megjelölt *(kölcsön) adott* közötti viszony `conj:cl`), azaz tagmondat-mellérendelés.



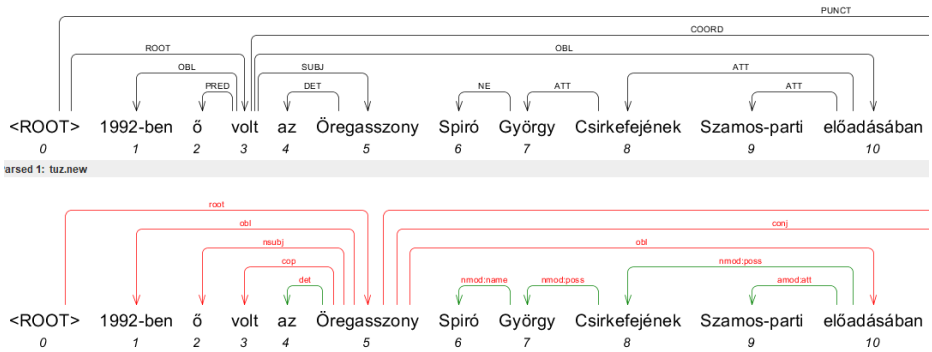
4. ábra. Tagmondat-mellérendelés azonosítása hibásnak annotált tagmondatfej esetén *(kölcsön adott)*.

2.6. Egyéb annotációs hibák és inkonzisztencia

Alany-állítmány tévesztés Az SZDT-ben a harmadik személyű **névszói állítmányt** tartalmazó tagmondatok annotációjában az alany és az állítmány szinte minden esetben fel van cserélve olyankor, ha az alany fókuszos, és az állítmány definit. Az elv láthatóan az volt, hogy ha két definit harmadik személyű NP verseng az alany és az állítmány szerepéért, akkor a fókuszos elemet jelölték állítmánynak (5. ábra). A fókusz és az állítmány azonban különböző dolog. A zavart az okozza, hogy fókuszos mondatok használatának szerencsefeltétele az, hogy tudjuk, hogy az adott állítás (az állítmány) valakire/valamire igaz, a mondatban az új információ az, hogy kire/mire (ez tűnik tehát állításnak). Az alany legnyilvánvalóbb ismérve, hogy mondatban a finit ige mindig az alannyal van egyeztetve (nyilván akkor is, ha az fókuszban áll). A kopulától különböző igét tartalmazó mondatokban fel sem merül, hogy a fókuszban álló elemet állítmánynak tekintsük. Minden esetben gyanús, ha névmás van állítmánynak jelölve. Ha az állítmánynak jelölt névmást (vagy egyéb gyanús névszói elemet) első/második személyű névmásra tudjuk cserélni, és a mondat értelmes és lényegében hasonló jelentésű marad (és persze a finit ige ilyenkor első/második személyű egyeztetésre vált), akkor egyértelműen hibás volt az eredeti annotáció.

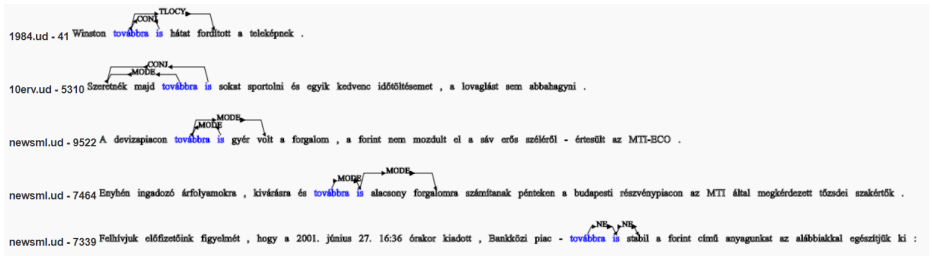
Megoldás A gyanús névszói állítmányokra (amelyeknél sokkal valószínűbb, hogy referáló kifejezés fejeként szerepel) rákeresve találtunk rá az ilyen típusú hibásan annotált mondatokra. Kézzel jelöltük meg a valóban tévesen annotált esetekben a kopulát, és ezekben a mondatokban automatikusan megcseréltük az alanyt és az állítmányt.

Inkonzisztens annotációk – ismétlődő mondatok Az SZDT-ben 1801 mondat(token) egybeesik valamelyik másik korpuszbeli mondattal (ismétlődik). Néhány esetben egy szövegrész véletlen ismétlődéséről van szó (itt az ismétlődő részt törölni kellene), de az esetek nagyobb részében nem ez a helyzet. Az 1801 ismételt mondat(token) 578 különböző mondattípust testesít meg, ezek közül 440 mondat kétszer fordul elő, és van olyan mondat is, ami ötvenkétszer *(Részletek a hétfői Befektetőben.)*. 118 ismétlődő mondattípusnak két különböző elemzése van, 8-nak három.



5. ábra. Az SZDT-ben hibásan felcserélt alany-állítmány viszony javítása a mondat UD-beli reprezentációjában.

Inkonzisztens annotációk – ismétlődő szósorok Emellett a korpusz 27245 olyan legalább kételemű szótokensorozatot tartalmaz, amelyen belül a függőségi viszonyok legalább két különböző módon vannak annotálva, illetve 13815 olyan sorozatot, ahol a fő szófajcímkék sorozata nem azonos. Természetesen az esetek nagy részében valódi többértelműségeknek felelnek meg az annotációs különbségek, illetve az inkonzisztenciák nagy része a zérókopula és az ellipszis annotációjára használt üres tokenek teljesen következetlen helyekre történő beszúrásából ered. Ezeknek az inkonzisztenciáknak a feltárásához az *Errator* eszköz egy módosított verzióját használtuk (Wisniewski, 2018). A *továbbra is* szókapcsolat annotációjával kapcsolatos valódi inkonzisztenciákra mutatunk példákat a 6. ábrán, amely jól érzékelteti egyrészt az *is* klitikum és az egyéb funkciószavak szófaji besorolásával és szintaktikai kapcsolásával kapcsolatos bizonytalanságokat.



6. ábra. Példák a *továbbra is* szókapcsolat annotációjával kapcsolatos inkonzisztenciákra az SZDT-ben

Megoldás Az ismétlődő mondatok inkozisztens annotációja kézzel átnézhető. A rövidebb ismétlődéseknél feltárt problémák kezelésénél a gyakori típusokat

(pl. az *is* eseteit) programozott megoldással javíthatjuk. Az egyedibb esetekben valószínűleg szintén csak a kézi javítás segít.

3. Összefoglalás

Cikkünkben a Szeged Dependencia Treebanket UD-kompatibilis verziójának előállítására tett erőfeszítéseink során felmerülő problémákat, illetve az ezek megoldására tett kísérleteinket mutattuk be. Az automatikusan megoldható átalakítások nagy részét implementáltuk, és sok manuális javítást, illetve ellenőrzést is elvégeztünk. Azonban a feladat elég sok élő munkát igényel, így egyelőre folyamatban lévő munkálatról beszélhetünk. Az egyelőre félkész terméket (illetve majd remélhetőleg a „kész" változatot) az `nlpg.itk.ppke.hu/resources` címen tesszük elérhetővé.

Köszönetnyilvánítás

Jelen kutatás a Nemzeti Kutatási Fejlesztési és Innovációs Alap által biztosított támogatással az FK 125217 és a PD 125216 számú projekt keretében az FK 17 és a PD 17 pályázati program, valamint a Nemzeti Kutatási, Fejlesztési és Innovációs Hivatal Mesterséges Intelligencia Nemzeti Kiválósági Programja támogatásával a 2018-1.2.1NKP-2018-00008 azonosítójú projekt keretében valósult meg.

Hivatkozások

Dozat, T., Qi, P., Manning, Ch.D.: Stanford's graph-based neural dependency parser at the CoNLL 2017 shared task. In: Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, Vancouver, Canada, August 3-4, 2017. pp. 20–30 (2017), `https://doi.org/10.18653/v1/K17-3002`

Nivre, J., de Marneffe, M.C., Ginter, F., Hajič, J., Manning, Ch.D., Pyysalo, S., Schuster, S., Tyers, F., Zeman, D.: Universal Dependencies v2: An evergrowing multilingual treebank collection. In: Proceedings of the 12th Language Resources and Evaluation Conference. pp. 4034–4043. European Language Resources Association, Marseille, France (May 2020), `https://www.aclweb.org/anthology/2020.lrec-1.497`

Novák, A., Laki, L.J., Novák, B., Dömötör, A., Ligeti-Nagy, N., Kalivoda, Á.: Egy magyar nyelvű kérdezőrendszer. In: XV. Magyar Számítógépes Nyelvészeti Konferencia. pp. 83–95 (2019)

Novák, A., Rebrus, P., Ludányi, Zs.: Az emMorph morfológiai elemző annotációs formalizmusa. In: XIII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY2017). pp. 70–78 (2017)

Orosz, Gy., Novák, A.: PurePos 2.0: a hybrid tool for morphological disambiguation. In: Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2013). pp. 539–545. INCOMA Ltd. Shoumen, BULGARIA, Hissar, Bulgaria (2013)

Vincze, V., Simkó, K., Szántó, Zs., Farkas, R.: Universal Dependencies and morphology for Hungarian - and on the price of universality. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers. pp. 356–365. Association for Computational Linguistics, Valencia, Spain (Apr 2017), https://www.aclweb.org/anthology/E17-1034

Vincze, V., Szauter, D., Almási, A., Móra, Gy., Alexin, Z., Csirik, J.: Hungarian dependency treebank. In: Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., Tapias, D. (szerk.) Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10). European Language Resources Association (ELRA), Valletta, Malta (may 2010)

Váradi, T., Simon, E., Sass, B., Gerőcs, M., Mittelholcz, I., Novák, A., Indig, B., Prószéky, G., Farkas, R., Vincze, V.: Az e-magyar digitális nyelvfeldolgozó rendszer. In: XIII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY2017). pp. 49–60 (2017)

Wisniewski, G.: Errator: a Tool to Help Detect Annotation Errors in the Universal Dependencies Project. In: chair), N.C.C., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Hasida, K., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S., Tokunaga, T. (szerk.) Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). European Language Resources Association (ELRA), Miyazaki, Japan (May 7-12, 2018 2018)

Zeman, D., Hajič, J., Popel, M., Potthast, M., Straka, M., Ginter, F., Nivre, J., Petrov, S.: CoNLL 2018 shared task: Multilingual parsing from raw text to universal dependencies. In: Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. pp. 1–21. Association for Computational Linguistics, Brussels, Belgium (October 2018), http://www.aclweb.org/anthology/K18-2001