Többszintű szintaktikai reprezentáció kialakítása a Szeged FC Treebankben

Simkó Katalin Ilona[1], Vincze Veronika[2], Farkas Richárd[1]

[1]Szegedi Tudományegyetem, TTIK, Informatikai Tanszékcsoport,
Szeged, Árpád tér 2.
kata.simko@gmail.com
rfarkas@inf.u-szeged.hu
[2]Magyar Tudományos Akadémia, Mesterséges Intelligencia Kutatócsoport,
Szeged, Tisza Lajos körút 103.
vinczev@inf.u-szeged.hu

Kivonat Napjainkban a két leggyakrabban használt szintaktikai reprezentációs elmélet a konstituens és a függőségi nyelvtan. A Szeged Treebank mondatai mindkét leírással manuális annotáltak. E cikkben beszámolunk egy olyan automatikusan átalakított, többszintű reprezentáció kialakításának munkálatairól, amely e két elemzés előnyös tulajdonságait egyesíti a mondatok szintaktikai leírásában.

1. Bevezetés

A létező szintaktikai elméletek közül jelenleg a két leginkább elterjedt a konstituens és a függőségi szintaxis. A Szeged Treebank mondatai is ezen reprezentációs elméleteknek megfelelően rendelkeznek manuális konstituens [1] és függőségi [2] elemzésekkel. Mindkét reprezentációnak megvannak az előnyei és a hátrányai is. A kétféle elemzés előnyeinek kihasználása céljából készül jelenleg automatikus átalakítással a Szeged Treebank leírására egy, a konstituens és függőségi fák, valamint a szavak morfológiai elemzéseit felhasználó, összetett szintaktikai reprezentáció. A reprezentáció kialakításakor hangsúlyozottan törekszünk arra, hogy a magyar nyelv szintaktikai sajátságait a lehető legnagyobb mértékben szem előtt tartsuk, ugyanakkor kiemelt szempontként kezeljük azt is, hogy a létrejövő treebank alkalmas legyen magyar nyelvű statisztikai szintaktikai elemzők betanítására is.

Ebben a munkában részletesen ismertetjük a többszintű szintaktikai reprezentáció kialakítása során követett irányelveket. Példákon keresztül megmutatjuk, hogyan kezelünk egyes nyelvi jelenségeket, valamint kitérünk arra is, hogy elemzésünk miben különbözik a Szeged Treebank eddigi változataiban követett függőségi, illetve konstituens alapú megközelítésektől, illetve szót ejtünk arról is, hogy elemzésünk hogyan viszonyul a szintén több nyelvi elemzési szinttel operáló LFG nyelvelméleti kerethez [3].

2. Konstituens és függőségi nyelvtanok

Bár a konstituens és a függőségi nyelvtanoknak is megvannak a hátrányai, mégis ezek a legelterjedtebben használt szintaktikai reprezentációk.

A konstituens reprezentáció a mondatokat összetevőkre bontja, amik összefüggő, jelentéssel bíró alkotóelemei a mondatnak. Tagmondatokra, azokon belül pedig igékre és bővítményeikre osztja a mondatokat. A szigorú konstituens elemzési elméletben az összetevők nyelvtani szerepére csak a szórendből következtethetünk, ami kötött szórendű nyelveknél, mint az angol jól működhet, de a magyar esetében kevésbé működőképes megoldás. A számítógépes nyelvészetben léteznek megoldások, amelyek az argumentumok felcímkézésével jelzik a nyelvtani szerepet, de ezek a konstituens nyelvtan szigorúan vett elméleti nyelvészeti hátterébe nem illenek bele. Nehezen elemezhetőek a nem folytonos konstituensek is, azaz azok az egybe tartozó elemek, amelyek nem egymás mellett jelennek meg a mondatban, mint például egyes mondatokban a genitív esetű birtokos és a birtoka.

Függőségi elemzésben a mondat szavai közvetlenül egymáshoz kapcsolódnak absztrakt csomópontok nélkül. Ezzel jól reprezentálhatóak a nyelvtani szerepek a mondatban és a nem folytonos összetevők kezelése is egyszerű feladat, elveszítjük viszont az összetartozó szavak egységként való kezelésének lehetőségét. Mindemellett a tagmondatok és mellérendelések kezelése például kevésbé intuitív, mint a konstituens elemzésben.

Mivel mindkét reprezentáció tartalmaz fontos információkat a magyar és a hasonlóan gazdag morfológiájú nyelvek szintaxisára vonatkozóan, nem eldöntött, hogy melyik a jobb leírás az ilyen nyelvek esetében. Hasonlóan, léteznek mind konstituens, mind függőségi elemzők a magyar nyelvre, melyek a Szeged Treebank különböző változatain lettek betanítva [4], azonban az automatikus elemzések kiértékelése során használatos mutatók sem teszik le egyértelműen a voksot egyik reprezentáció mellett sem. Ezen okokból döntöttünk egy olyan szintaktikai reprezentáció létrehozása mellett, amely egyesíti a két elmélet által kódolt információkat.

A Szeged Treebank mondatai kézzel annotált konstituens és függőségi elemzéssel is el vannak látva. A kétféle reprezentáció részben megegyező, részben az adott reprezentációnak megfelelő információkat kódol a mondat szintaktikai szerkezetével kapcsolatban. Ezeket az információkat egyesítjük egy új, többszintű szintaktikai leírásban.

3. Többszintű szintaktikai reprezentáció

A Szeged Treebank többszintű szintaktikai reprezentációja a lexikai funkcionális grammatika [3] elmélethez hasonló szerkezetű és a már létező, kézzel annotált konstituens és függőségi elemzések és morfológiai kódok felhasználásával jön létre. Az LFG-hez hasonlóan a különféle nyelvtani jellemzőket különböző szinteken jelenítjük meg.

A LFG reprezentációk több különböző struktúrát rendelnek a mondatokhoz. Ezek különböző szintaktikai szerkezeteken kívül szemantikai, fonológiai és egyéb nyelvi szintekhez kapcsolódó információkat is hozzákapcsolnak a mondat kifejezéseihez. A struktúrák egy többszintű reprezentáció alkotórészeit képezik ebben a keretben, egy-egy kifejezéshez a leírás több különböző szintjéről másmás információk társulnak és ezek együtt, egymással összekapcsolva alkotják az LFG elméletbeli reprezentációját az adott mondatnak.

Az LFG struktúrái közül a szintaktikai szempontból legalapvetőbb c- és f-struktúrák létrehozása mellett döntöttünk. A c-struktúra a mondat felszíni szerkezetét tükrözi, azt összetevőkre bontja. Az f-struktúrában a mondat argumentumszerkezete, illetve morfológiai információk jelennek meg attribútum-érték párokként. A két szerkezet szavai és nagyobb összetevői egymással összeindexelve, közösen alkotják ezt a többszintű modellt.

A magyar nyelv bizonyos jelenségeinek ebben a modellben való elemzéséről már nagyon sok cikk született [5,6], de a magyart általánosan leíró LFG nyelvtan legjobb tudomásunk szerint nem létezik. Jelen átalakítás alapelveinek lefektetésekor egy átfogó jellegű szabályrendszert igyekeztünk létrehozni, és a kisebb számban előforduló speciális nyelvi jelenségek kezelésére átvesszük a Szeged Treebank előző verzióiban kifejlesztett megoldásokat.

4. Átalakítás

4.1. C-struktúra

A c-struktúra átalakítása a Szeged Treebank konstituens elemzéséből indul ki. Ez az átalakítás viszonylag kevés módosítással jár. Megtartjuk a kézzel annotált frázisokat és hozzájuk adunk egy-egy indexet, ami összekapcsolja őket az f-struktúra megfelelő részeivel.

Így a konstituensnyelvtan előnye, az összetevős struktúra megmarad ebben az új modellben is, az ebben nehezen reprezentálható nyelvtani szerepek pedig más szinten vannak kezelve.

4.2. F-struktúra

Címkék. Az f-struktúra a mondat argumentumszerkezetét tükrözi. Ezen a szinten találhatóak a kifejezésekhez tartozó nyelvtani szerepek, és a nem folytonos összetevők elemzése is megoldható. Leginkább a függőségi nyelvtanban kódolt információval feleltethető meg, ezért a Szeged Dependencia Treebank és a mondatok szavaihoz rendelt morfológiai kódok átalakításával hozzuk létre.

Ezen a szinten a szintaktikai információ attribútum-érték párokból álló szerkezetben jelenik meg. Minden kifejezés f-struktúrájában megtalálhatóak a hozzátartozó releváns morfológiai adatok és a kifejezés különböző vonzatainak f-struktúrái. A függőségi nyelvtanban található relációk címkéit itt attribútumok címkéiként jelennek meg, az ezekhez kapcsolódó érték a kapcsolódó kifejezés f-struktúrája.

A mondat PRED jegye alatt megtaláljuk a fő elemet és a vonzatait zárójelben. A mondatok fő eleme a függőségi nyelvtan ROOT eleme, vonzatai a függőségi nyelvtanban hozzá csatlakozó szavak. A PRED jegy után a releváns morfológiai jegyek találhatóak, amelyeket a szavak morfológiai kódjából nyerünk.

Ezután a predikátum argumentumai következnek a nyelvtani szerepüknek megfelelő címkével. A függőségi nyelvtan SUBJ (alany) és OBJ (tárgy) relációi azonos nevű címkék lesznek az f-struktúrában. A kötelező vonzatok, a függőségi nyelvtanban DAT (részes eset) és OBL (egyéb eset) relációban állók egy közös, OBL címkét kapnak, míg a különböző határozói szerepű vonzatok (MODE, LOCY, FROM, TO, TLOCY, TFROM, TTO függőségi reláció) ADJ (adjunktum) címke alá kerülnek. Az INF, PA és AUX relációkkal rendelkező főnévi igenevek, melléknévi igenevek és segédigék szintén megtartják a függőségi relációjuk nevét az f-struktúra-beli címkéjükben.

A vonzatok f-struktúrája hasonló felépítésű: a PRED jegy az adott kifejezést jelöli, utána a vonzatait, módosítóit találjuk. Ezután a szófajának megfelelő morfológiai jegyek értékei következnek. A vonzatokat OBL vagy DAT függőségi relációval módosító, kötelező bővítmények itt is OBL címke alá kerülnek. Az ATT és MODE viszonyúak ADJ címkét kapnak. A névszókat módosító birtokosok POSS címkével kerülnek a birtok f-struktúrájába. A határozott és határozatlan névelők DEF=+ és DEF=- jegyekként jelennek meg a szerkezetben.

A névszói predikátumok függőségi PRED relációját az LFG elméletnek megfelelően [7,8] PREDLINK címkével jelöltük az f-struktúrákban. Ennek mintájára a többszavas névelemek kezelésére a függőségi NE viszonyt NELINK-ké alakítottuk, az összetett számnévi kifejezések NUM relációját pedig NUMLINK-ké.

Összetett mondatok. Az összetett mondatok kezelésében szintén az LFG-ben használt megoldást választottuk. Alárendelő szerkezetek és vonatkozói mellékmondatok esetén a főmondat PRED elemének egy vonzata az alárendelt mondat fő eleme, a beágyazott mondat f-struktúrája COMP címkével jelenik meg a főmondat f-struktúrájában. Mellérendelés esetén a mellérendelt kifejezések f-struktúrái egymás mellett jelennek meg. A kifejezéseket összekapcsoló esetleges kötőszavak alárendelés esetén az alárendelt mondat f-struktúrájában, mellérendelés esetén a mellérendelt tagok f-struktúrái alatt, CONJ-FORM címke alatt találhatóak.

Kötelező jegyek. Az f-struktúrában az egyes kifejezések alatt megtalálható kötelező morfológiai jegyeket az adott kifejezés morfológiai kódjából nyerjük ki. Az, hogy egy szónál milyen jegyeknek kell kötelezően megjelenni, a szó szófajától függ.

Az MSD kódban tárolt információk közül a szintaktikailag relevánsakat jelenítjük meg. Az ige altípusa, száma, személye, az igemód, igeidő és határozottság az ige f-struktúrájában jelenik meg. A névszói vonzatok esetében a szám és az eset jelenik meg kötelezően. Melléknevek esetén ezeken felül a fokozás, névmásoknál a személy.

Hely- és időhatározók. A Szeged Treebankben található három-három hely- és időhatározó típus megkülönböztetését az átalakított többszintű reprezentációba nem vettük át, mivel úgy gondoljuk, hogy ezen megkülönböztetés már túlmutat a szintaxis szintjén. Az irányhármasságot is kifejező hely- és időhatározói címkéket minden esetben ADJ jegyként kezeltük a mondatok f-struktúrájában.

A későbbiekben ezt az információt egy újabb struktúrába tervezzük felvenni, amelyben megtennénk ezt a szinte már szemantikai megkülönböztetést a hely- és időhatározók típusai között.

5. Virtuális csomópontok

A magyar LFG reprezentációjával kapcsolatban ugyanúgy felmerül a virtuális csomópontok problémája, mint a függőségi elemzésben. Mivel mindkét elmélet kerüli a fonológiailag jelen nem levő kifejezések megjelenítését a szintaktikai struktúrákban, a magyarban megjelenő kétféle virtuális összetevő kezelése nehézségeket okozhat.

A magyarban előforduló egyik ilyen meg nem jelenő összetevő a *van* ige harmadik személyű, kijelentő mód, jelen idejű alakja. A *Józsi katona* mondat esetén például nem jelenik meg az ige, ami más személy, mód vagy igeidő esetén már igen, például *Józsi katona volt*.

A másik típus az ellipszis, az a több nyelvre is jellemző jelenség, amikor egy már elhangzott szót vagy kifejezést nem mondunk ki újra, illetve a több tagmondatban ismétlődő kifejezéseket csak a tagmondatok egyikében szerepeltetjük. A ki nem mondott kifejezés lehet a tagmondat fő igéje, vagy annak bármely argumentuma, illetve az argumentum kisebb része. A *Józsi katona volt, Béla pedig pék* mondat esetén például a második tagmondatból a *volt* ige elliptálva van.

A virtuális csomópontok mindkét típusánál hasonló megoldás mellett döntöttünk. A virtuális kifejezések a mondathoz tartozó c-struktúrában nem jelennek meg, mivel az szigorúan a mondat felszíni szerkezetét rendezi frázisokba. Ezek a kifejezések csak az f-struktúrában jelennek meg, ami a szigorú LFG elméletben szintén kerüli a ki nem mondott kifejezések reprezentálását, viszont az ott megjelenített viszonyok leírásához fontos, hogy kitöltsük ezeket a csomópontokat is.

Az f-struktúrában a PRED jegyben jelöljük, hogy virtuálisról van szó: VAN vagy ELL értéket kap. A további jegyeket csak a VAN kapja meg, azok közül is csak azokat, amelyek biztosak: az igemód, igeidő és személy.

6. Eltérések az LFG-től

A Szeged Treebank átalakításakor főként az LFG elméletben [3] használt megoldásokat követtük, így a reprezentáció nagyon hasonló a lexikai funkcionális grammatika c- és f-struktúráihoz. Néhány ponton viszont eltértünk a szigorú LFG elmélettől. A következőkben ismertetünk néhányat ezen eltérések közül.

6.1. C-struktúra

Az LFG reprezentációk c-struktúrái a generatív nyelvtanokban használt bináris, X-vonás elméletnek megfelelő fákból állnak [9].

Az általunk átalakított c-struktúrák a Szeged Treebank konstituens fáihoz hasonlóan nem követik a szigorú chomskyánus nyelvtant, hanem a fő elem szófajának megfelelő frázisokra bontják a mondatokat.

6.2. Topik és fókusz pozíciók

Az LFG elemzésben a mondatok f-struktúrájában jelölve van a topik és a fókusz pozíció is, főleg a magyarhoz hasonló diskurzuskonfigurációs nyelvek szintaktikai leírása esetén.

A Szeged Treebank átalakítása során nem használtuk az f-struktúrában a topik és fókusz pozíciókat, mivel az erre vonatkozó információ sem a meglévő konstituens, sem a meglévő függőségi treebankben nincs kódolva, és így automatikus konvertálásuk nem megoldható. A topik és fókusz jelölése egy későbbi lépésben belekerülhet az f-struktúrákba kézi annotációval.

6.3. Fonológiailag üres névmási kategóriák

Bár az LFG kerüli az üres kategóriák felvételét az elemzésbe, pro elemek mégis megjelennek ki nem mondott névmások helyén az f-struktúrában. A magyarban gyakran ki nem tett személyes névmási alany és tárgy helyére például egy pro kerül az LFG elemzés f-struktúrájába.

Mivel a Szeged Treebank egyik verziója sem jelöli a fonológiailag üres névmásokat, az átalakítás során az ehhez hasonló esetekben nem vettük fel a pro PRED jegyű elemet, az ehhez tartozó jegyeket egy szinttel feljebb jelenítjük meg. Például egy elhagyott alany esetén annak száma és személye a magyarban megjelenik az igén, így ezeket a jegyeket ott reprezentáljuk ahelyett, hogy egy pro PRED jegyű alanyt vennénk fel az f-struktúrába ezekkel a jegyekkel.

7. A Szeged FC Treebank kialakítása

A fentiekben ismertetett elveket a gyakorlatba átültetve kialakítjuk a Szeged Treebank egy újabb verzióját, a Szeged FC Treebanket. Ezt elsődlegesen automatikus konverzió segítségével állítjuk elő a meglévő konstituens- és függőségi reprezentációk alapján, minimálisra csökkentve az utólagos kézi javításokat. A létrejövő új treebank kitűnő lehetőséget teremt arra, hogy létrehozzunk egy olyan statisztikai szintaktikai elemzőt, amely kifejezetten a magyar nyelv szintaktikai sajátságaira van optimalizálva, ugyanakkor egyesíti magában a konstituens és függőségi elemzők nyújtotta előnyöket is.

A Szeged FC Treebank kialakítása a Szeged Treebank konstituens és függőségi elemzéseinek automatikus konvertálásával történt a már leírt szabályok mentén. Az alábbiakban bemutatjuk egy példán keresztül az átalakítás különböző lépéseit.

A c-struktúrát a konstituens fákból egyszerűen a nyelvtani szerepjelölések eltávolításával nyertük, l. 1. és 2. ábrák.

Az f-struktúra és a függőségi nyelvtan között már nagyobb különbség látható, vö. 3. és 4. ábrák. A példamondatban az alá- és mellérendelő szerkezeteken kívül a birtokos szerkezetek kezelése is látható a két különböző elméleti keretben.

A Szeged FC Treebank reprezentációi a Szeged Korpusz mondataihoz a fent láthatóakhoz hasonló c- és f-struktúrákat rendelnek. Ez a két leírás együtt alkotja az új treebank elemzését.

8. Összegzés

Ebben a munkában bemutattuk a készülő Szeged FC Treebank elméleti alapját képező többszintű szintaktikai reprezentációt, mely egyesíti magában a konstituens és függőségi reprezentációk előnyeit, ugyanakkor kifejezetten a magyar nyelv szintaktikai sajátságaira van szabva. Az LFG elméletéhez hasonlóan, e reprezentáció is c és f-struktúrában jeleníti meg a releváns szintaktikai információkat, azonban attól néhány fontos vonásban eltér. Az újonnan létrejövő treebank reményeink szerint egy új, a magyar nyelvet minden eddiginél hatékonyabban feldolgozni képes statisztikai szintaktikai elemző létrehozásának alapjául szolgálhat.

Köszönetnyilvánítás

A jelen kutatás a futurICT.hu nevű, TÁMOP-4.2.2.C-11/1/KONV-2012-0013 azonosítószámú projekt keretében az Európai Unió támogatásával és az Európai Szociális Alap társfinanszírozásával valósult meg.