A Szeged Treebank függőségi fa formátumban

Vincze Veronika[1], Szauter Dóra[1], Almási Attila[1], Móra György[1],
Alexin Zoltán[2], Csirik János[3]

[1] Szegedi Tudományegyetem, Informatikai Tanszékcsoport
{vinczev, szauter, gymora}@inf.u-szeged.hu, vizipal@gmail.com
[2] Szegedi Tudományegyetem, Szoftverfejlesztés Tanszék
alexin@inf.u-szeged.hu
[3] MTA-SZTE, Mesterséges Intelligencia Tanszéki Kutatócsoport
csirik@inf.u-szeged.hu

Kivonat: Az előadásban a Szeged Treebank függőségi fa formátumra történő átalakításának folyamatát mutatjuk be. Az eredetileg frázisstrukturált treebankből automatikus konverzió eredményeképpen létrejött függőségi fákat kézi úton ellenőriztük és javítottuk, létrehozva ezzel az első magyar nyelvű kézzel annotált dependenciakorpuszt. Jelenleg az üzleti híreket, újsághíreket és jogi szövegeket tartalmazó alkorpuszok annotációja fejeződött be, de terveink között szerepel a teljes korpusz átalakítása függőségi fa formátumra. Az elkészült adatbázis hasznosítható többek között az információkinyerésben és a gépi fordításban is.

1 Bevezetés

A Szeged Treebank függőségi fákat tartalmazó szintaktikai annotációjának célja az első, teljes egészében kézzel annotált magyar nyelvű dependenciakorpusz létrehozása. Az adatbázis számítógépes szempontú hasznosíthatósága többrétű, hiszen a gépi fordításban való felhasználás mellett az információkinyerés részterületein is számos alkalmazásban töltheti be a tanító adatbázis szerepét. Az előadásban ismertetjük a korpuszépítési munkafolyamatokat, a konverzió és az annotáció során felmerült problémákat és az azokra született megoldásokat, továbbá a korpusz statisztikai adatait, végül szót ejtünk a korpusz hasznosíthatóságáról is, illetve nemzetközi kontextusban is elhelyezzük a létrehozott adatbázist.

2 Függőségi nyelvtanok

A Szeged Treebank eredetileg frázisstrukturált formában kódolja a mondat összetevői közti szintaktikai viszonyokat. A frázisstrukturált korpuszban a mondatok tagmondatokból felépülő hierarchikus struktúrát alkotnak: a mondat összetevői (konstituensei) konstituensfákká szerveződnek. Maguk a tagmondatok igékre, az igék vonzataira (ezek névszói szerkezetek) és egyéb alkotóelemekre bonthatók, amelyek az egyes szinteken belül azonban nem alkotnak hierarchiát. A mondat szavai a

konstituensfa levelein helyezkednek el, a fa egyéb csomópontjai absztrakt szerveződési egységeket jeleznek (frázisstruktúra-címkékkel ellátva).

A függőségi fa formátum ettől abban tér el, hogy a fában minden egyes csomópont a mondat egy szavának felel meg. A mondatfa csúcsán egy mesterséges gyökérelem található, amelynek alárendeltjei lesznek a mondatban előforduló szavak, vagyis a gyökérelemen kívül nem találhatók absztrakt csomópontok a fában. A mondatban minden egyes szó szigorúan egy másik szó alárendeltségében van: egy szónak csak egy fölérendeltje lehet, egy csomópont alá azonban tartozhat több szó is, például az ige csomópontja alá sorolható be az ige összes bővítménye. A függőségi fában szereplő csomópontok között többféle kapcsolat is lehetséges, ezeket általában különféle címkékkel látják el, amelyek a kapcsolat jellegére utalnak.

Az első függőségi nyelvtannak Tesnière könyve [20] tekinthető, mely lefekteti a rendszer alapjait. Híres hasonlata szerint a mondatnak az ige a központi eleme, mely egész kis drámát fejez ki: a dráma szereplői az ige bővítményei, melyeket Tesnière aktánsoknak nevez. A mondatban így tehát alárendelt elemek és fölérendelt elemek szerveződnek egységbe.

Mel'čuk [17, 18] függőségi nyelvtana az Értelem ↔ Szöveg Elméleten belül jött létre. Nála a függőségi viszony lineáris relációként jelenik meg a szavak között. Mélyszintaktikai szinten 12 viszonytípust feltételez, ebből 6 az ige és különféle bővítményei (aktánsai) között létezik, a többi viszony pedig mellérendelést és különféle módosító szerepet jelez. A Mel'čuk-féle függőségi nyelvtan különlegessége, hogy a mellérendelést is egyfajta alárendelésként fogja fel: a mellérendelés első tagjához kapcsolódik a kötőszó, illetve utóbbihoz a mellérendelés további tagja(i) speciális (COORD) viszonnyal. Egy másik érdekesség, hogy bizonyos esetekben a nyelvtan engedélyezi absztrakt, azaz a mondatban fonetikailag meg nem jelenő nyelvi elemet jelző csomópontok felvételét a függőségi fába: ilyen eset például az egyes szám harmadik személyű jelen idejű létige az oroszban (és a magyarban is), amely fonetikailag nem ölt testet a mondatban, azonban absztrakt szinten mégis jelen van, hiszen múlt és jövő időkben megjelenik testes formában.

A magyar nyelvre alkalmazott függőségi nyelvtanokról [16] és [19] nyújt áttekintést, illetve saját, morféma alapú függőségi nyelvtanuk rövid vázlatát mutatják be a szerzők. Modelljükben a függőségi fák alapelemei a morfémák, mivel agglutináló nyelvekben nem (csak) a szavak, hanem a morfémák képesek a különböző grammatikai viszonyok kifejezésére. Ez a megoldás megkönnyíti a különböző típusú nyelvek függőségi fái közti leképezéseket, mert például az angol *may* segédige csomópontjának a magyar fában a *–hat* morféma csomópontja felel meg. Ezt az eljárást alkalmazva a függőségi fákon alapuló számítógépes fordítórendszerek hatékonysága jelentősen megnövekedhet.

3 Más nyelvű dependenciakorpuszok

A világ számos nyelvére fejlesztettek már ki dependenciakorpuszt. Ezek közül az egyik leghíresebb a cseh nyelvre épített Prague Dependency Treebank [1], mely morfológiai, szintaktikai és tektogrammatikus szintű annotációt is tartalmaz. Ugyanez a műhely angolra és csehre is kifejlesztett egy párhuzamos, dependenciaannotációt

tartalmazó korpuszt [2, 3], illetve arab nyelvű dependenciakorpusz is fűződik a nevükhöz [4]. A fentieken kívül számos európai (többek között svéd [5], görög [6], orosz [7] és szlovén [8]) és Európán kívüli nyelvre (japán [9], kínai [10]) építettek már dependencia treebanket, illetve még holt nyelvekre is: egy latin nyelvű korpusz már létrejött, és alkotói egy ógörög korpuszon dolgoznak jelenleg [11]. Az első magyar nyelvű dependenciakorpusz létrehozásával ehhez a vonulathoz kívánunk csatlakozni.

4 A korpuszépítés folyamata

Ahhoz, hogy az eredetileg frázisstrukturált treebankből dependenciakorpuszt tudjunk készíteni, először is szükség van egy konverziós lépésre, melynek során a konstituensfák függőségi viszonyokká alakulnak át. Mivel az automatikus gépi konverziótól nem várhatunk tökéletes és hiba nélküli eredményt, ezt a munkafázist egy kézi ellenőrzési folyamat követi, melynek során nyelvészek átnézik a fájlokat, és a szükséges esetekben módosítják azokat.

Noha a korábbi szakirodalomban megtalálható a magyarra alkalmazott függőségi nyelvtan rövid vázlata [16, 19], a Szeged Treebank függőségi fa formátumra történő átalakításakor mégsem követjük teljes egészében ezt a modellt. Ennek az a magyarázata, hogy az említett modell morféma alapú, azaz a függőségi fa csomópontjaiban nem szóalakok, hanem morfémák szerepelnek. Ahhoz azonban, hogy a szintaktikai fákat morfémákból építhessük fel, szükség lenne egy jól működő morfológiai elemzőre, mely a Szeged Treebank szóalakjait morfémákra bontaná. Mivel a Szeged Treebank MSD-kódjai a képzéseket nem jelölik, például a műveltető és ható igék képzőit a szótő részeként kezeli a rendszer, vagyis nem lenne képes külön morfémát, azaz külön csomópontot rendelni a képzőkhöz. A morféma alapú függőségi fákra történő konverzió választása még további munkaigényes feladatokkal járna (többek között az MSD-kódrendszer átalakítása úgy, hogy lehessen jelezni a képzéseket, a szóalakok újrakódolása a korpuszon belül, jól működő morfológiai elemző kialakítása a korpuszra stb.). Emiatt csupán a szóalakok közti függőségi viszonyok bejelölésére vállalkoztunk.

A Szeged Treebank 2.0 függőségi fa formátumra való átalakítása első lépésének a 2007-es CoNLL konferencia szervezőbizottsága által kiírt nemzetközi versenyfeladat [12] tekinthető, amikor is a tesztadatbázis elkészítésére való felkérésnek köszönhetően megtörtént a korpusz HVG- és Népszabadság-cikkeket tartalmazó részének konvertálása [13], majd ennek nyomán a teljes korpusz átalakítása.

A Szeged Treebank 2.0-ban az ige és vonzatai közti nyelvtani viszonyok jelölve voltak. Ezeket a viszonyokat kellett függőségi viszonyokká átalakítani. A konverzió során automatikusan, gépi úton történt a viszonyok átcímkézése nyelvészek által előzetesen meghatározott szabályok alapján. A lehetséges függőségi viszonyok az alábbiak:

APPEND – a mondatba szervesen nem illeszkedő mondatrészek
ATT – főnév és jelző, névutó és főnév, főnév(i módosító) és főnév közti viszony
AUX – ige és segédige közti viszony
AUXS – a mondat értékű elem
CONJ – kötőszó
COORD – mellérendelés
DAT – nAk ragos főnévi vonzat
DET – főnév és determináns közti viszony
FROM – honnan? kérdésre válaszoló határozószó, illetve névutós szerkezet
INF – főnévi igenév
LOCY – hol? kérdésre válaszoló határozószó, illetve névutós szerkezet
MODE – egyéb határozószavak, illetve névutós szerkezetek
NEG – tagadószó
OBJ – ige és tárgy közti viszony
OBL – ige és egyéb főnévi bővítménye közti viszony
PRED – ige és névszói állítmány közti viszony
PREVERB – ige és igekötő közti viszony
PUNCT – írásjel
QUE – kérdőszó
ROOT – a mondat fő eleme
SUBJ – ige és alany közti viszony
TFROM – mikortól? kérdésre válaszoló határozószó, illetve névutós szerkezet
TLOCY – mikor? kérdésre válaszoló határozószó, illetve névutós szerkezet
TO – hova? kérdésre válaszoló határozószó, illetve névutós szerkezet
TTO – meddig?, mikorra? kérdésre válaszoló határozószó, illetve névutós szerkezet

A gépi úton előállt fájlokat nyelvészek ellenőrizték, és ha kellett, javították. A javítási munkálatokhoz az erre a célra kifejlesztett, és a magyar nyelv sajátosságainak megfelelően testre szabott TrEd szerkesztőprogramot [14] használtuk.

4.1 Típushibák

A kézi ellenőrzés során elsődleges feladat a gépi konverzió átnézése, szükség esetén javítása volt. A javításra szoruló legtipikusabb hibák két kategóriába estek: (1) a csomópont rossz helyen volt a fában; (2) a csomópont és fölérendeltje nem a megfelelő viszonyban állt.

A hibák nagy része abból fakadt, hogy a frázisstrukturált korpuszban nem minden nyelvtani viszony volt jelölve, például a névelők, számnevek és jelzők a főnévi csoporton belül szerepeltek, és a főnévhez fűződő viszonyuk külön nem volt feltüntetve. A konverzió során automatikusan a főnév alá lettek bekötve ATT viszonnyal mindezen elemek, a mondatban található egyéb elemek pedig az ige alá kerültek be MODE viszonnyal. Ezeket szükség szerint javítani kellett a megfelelő függőségi viszonyra, illetve áthelyezni a megfelelő felettes (anya)csomópont alá.

Az átcímkézést igénylő leggyakoribb esetek a következők voltak:

- **jelzős szerkezeten belüli ragozott főnév**
  A konvertálóprogram a fenti okokból kifolyólag ATT címkével látott el minden főnevet, amely AP (melléknévi csoport) tagja volt, például *a ténylegesnél 1,9_milliárd dollárral magasabb árbevételt* szerkezetben a *ténylegesnél* és a *dollárral* is ATT címkét kapott a helyes OBL helyett, így ezt kézi úton kellett javítani.

- **NE-k kezelése**
  A tulajdonnevek az esetek nagy többségében ATT címkét kaptak a konverzió során, ezeket természetesen javítottuk az adott kontextusnak megfelelő címkére.

- **alárendelő mellékmondatok fő elemének címkéje**
  Az alárendelő mellékmondatokat a Treebankben annak megfelelően címkézték, hogy milyen szerepet tölt be a főmondatban az utalószó (és az utalószó alá is kötötték be, amennyiben volt ilyen a mondatban, l. alább). A dependenciakorpuszban ettől eltérően csak annyit jelölünk, hogy alárendelésről van szó, azaz ATT címkével látjuk el a mellékmondat fő elemét.

- **mellérendelések második, harmadik… tagja**
  A Treebankben a mellérendelések a frázisstruktúra-nyelvtanokban szokásos megoldásnak megfelelően kívülről kaptak egy közös címkét, melynek típusa megegyezett a mellérendelés tagjainak saját címkéjével: tehát két egymás mellé rendelt főnévi csoport (NP) egy külső NP címkével is rendelkezett, mely mindkettőt magában foglalta. Mivel a dependencia-nyelvtanokban nincsenek mesterséges csomópontok, ez az eljárás nem bizonyult követhetőnek, így a Mel'čuk-féle megoldást követtük a mellérendelések elemzésénél, l. lejjebb.

- ***ez/az* mutató névmások**
  A mutató névmások ATT címkét kaptak, ha mutató névmás + névelő + főnév konstrukcióban (*ez a ház*) fordultak elő. Alanyesetű előfordulásukkor DET, azaz determinánsi címke járt nekik, ha pedig esetragot viseltek (pl. *ebben a házban*), akkor az adott esetnek megfelelő címkére kellett javítani (jelen példában OBL-ra).

A csomópontok áthelyezése a fában az alábbi esetekben volt a legszükségesebb:

- ***alárendelő mellékmondatok***
  Amint már fentebb utaltunk rá, a kötőszó nem képezte az alárendelő mellékmondatok részét a Szeged Treebank frázisstrukturált változatában. Ennek eredményeképpen a konverzió után a főmondat fő eleméhez kapcsolódott a kötőszó és a mellékmondat fő eleme is (külön-külön). A kézi ellenőrzés folyamán a nyelvészek a kötőszóhoz kötötték hozzá a mellékmondat fő elemét, így teremtve meg a kapcsolatot a két összetevő között.

- ***birtokos szerkezetek***

  A birtokos szerkezetek két része, a birtokos és a birtok gyakran nem kapcsolódott össze a korpuszban. Különösen érvényes volt ez a *–nAk* ragos birtokosra, főleg, ha nem a birtok melletti pozíciót foglalta el a mondatban. A dependenciakorpuszban a birtokost mindig összekötöttük a birtokkal, még akkor is, ha ezzel keresztező függőségek álltak elő, azaz a fa két éle metszi egymást. (Ez a frázisstruktúra-nyelvtanokban szigorúan tilos, mivel ott lehetségesek a mozgatások, dependencia-nyelvtanokban azonban elfogadott a keresztezések léte.)

- ***mellérendelés***

  Amint már az átcímkézési eseteknél említettük, mellérendelésnél nemcsak a csomópontok címkéit, hanem a helyzetüket is módosítani kellett. A gépi elemzés során általában a kötőszó funkcionált a szerkezet fejeként, és a mellérendelés tagjai vele álltak függőségi viszonyban. A Mel'čuk-féle megoldásnak megfelelően azonban a szerkezet első tagja funkcionál fejként, ez alá kell kötni a kötőszót (amennyiben volt) CONJ viszonnyal, majd a mellérendelés többi tagja következik COORD viszonnyal kapcsolva az előző elemhez.

- ***főnévi igenevek és igekötők***

  Ha a mondatokban szerepelt egy olyan (segéd)ige, amelynek főnévi igenév vonzata volt (*szeret*, *kíván*, *fog*, *kell*…), akkor a gépi elemzés a főnévi igenév esetleges igekötőjét a főigéhez társította. Ezt a hibatípust is kézzel javították a nyelvészek az ellenőrzés során.

4.2 Mellérendelés

A mellérendelés kérdése problémákat vet fel a legtöbb szintaktikai elmélet számára: egyes elméletek hívei azt a megoldást tartják jónak, hogy a kötőszó a koordináció feje, mások pedig amellett érvelnek, hogy a szerkezet feje a mellérendelés egyik tagja. Vizsgáljuk meg ezeket az elképzeléseket külön-külön!

Tegyük fel, hogy a kötőszó a szerkezet feje. Felmerül azonban a kérdés, hogy mit lehet tenni a direkt koordináció eseteiben, amikor nincs az elemek között kötőszó. Ha nincs kötőszó, akkor fel kell tételezni egy virtuális csomópontot, amely képes fejként funkcionálni. Az elképzelésnek azonban más hátulütője is van: ha több mellérendelt elem van, akkor nem tudjuk megkülönböztetni az „A és B és C" típust az „A, B és C" típustól. A problémát meg lehetne úgy kerülni, hogy felveszünk egy absztrakt „és"-t az „A" és „B" fölé, de akkor a „B" egyidejűleg két csomóponthoz (egy virtuális *ÉS* és egy valós *és*) kapcsolódna, ez pedig szigorúan tilos. További hátránya az elgondolásnak, hogy ha például a mellérendelt frázis a mondat alanya, akkor a kötőszó és az ige közt lenne SUBJ viszony, ez pedig igen kevéssé lenne szokványos.

Egy másik elképzelés szerint azonos szinten szerepelnek a koordinált elemek és a kötőszó, de nincsenek összekapcsolva, például a *Jancsi és Juliska mézeskalácsháza* szókapcsolatban a *mézeskalácsháza – Jancsi*, *mézeskalácsháza – és*, valamint *mézeskalácsháza – Juliska* viszonyok állnak fönn. Ez esetben az jelenti a problémát, hogy noha *Jancsi* és *Juliska* összetartozását az azonos címkéjű (ATT) viszony még vala-

hogy tudná jelölni, de eléggé kérdéses, hogy milyen viszonyban állna a *mézeskalács-háza* és az *és*, arról nem is beszélve, hogy eléggé szokatlan, hogy a koordináció két tagját nem kapcsoljuk össze.

A fenti megoldások egyike sem nyújt kielégítő választ a felmerülő problémákra, éppen ezért a korpusz átalakítása során a koordináció esetén a Mel'čuk-féle elképzelést [17, 18] követjük, ahol is a mellérendelés egyfajta „alárendelés". Mindig a koordináció első eleme a fej, mert az tud az egész frázis helyett állni. Vegyük a következő példákat:

Elmentem a boltba Józsival és Kival.
Elmentem a boltba Józsival.
**Elmentem a boltba Józsival és.*
**Elmentem a boltba és Kival.*

A második, illetve a harmadik és negyedik mondat közti különbség mutatja, hogy a koordináció nem bontható fel két egyenrangú részre, hiszen ha a *Józsival* és az *és Kival* elemek egyenértékűek lennének, akkor elfogadhatónak kellene lennie az utolsó mondatnak. A *Józsival* az *és* elemmel sem tartozik szorosan össze, hiszen akkor a harmadik mondat is jó lenne. A megoldás az, hogy három részt feltételezünk a koordinációban: az első elem a fej, ehhez kapcsolódik a kötőszó CONJ viszonnyal, illetve a kötőszót követi a második mellérendelt tag COORD viszonnyal:

Józsival
| CONJ
és
| COORD
Kival

Ez ábrázolás szempontjából igaziból „alárendelés", és így szerkezetben nem lesz különbség mellé- és alárendelés között: csak a viszonyok (ATT, illetve COORD) jelzik, hogy melyikről van szó.

4.3 Predikatív névszók

A magyar nyelv sajátságaiból adódóan a predikatív névszót tartalmazó mondatokban a létige kijelentő mód jelen idő E/3. alakja nem jelenik meg a felszínen, szemben a más módú, idejű vagy számú, illetve személyű formákkal:

*András katona (*van).*
András legyen katona!
András katona lesz.

A mellérendeléshez hasonlóan, jelen problémánál is kétféle megoldási lehetőség létezik. Az első lehetőség szerint a mondat fő elemének a predikatív névszót tekintjük, ez alá csatoljuk az alanyt, és nem veszünk fel virtuális csomópontot. Azonban ennek a megoldásnak az a hátránya, hogy teljesen más szerkezetet tulajdonítunk

ugyanannak a mondatnak jelen és például múlt időben, ami megkérdőjelezhető, mert az egyik esetben a predikatív elem és az alany között közvetlen, másik esetben pedig közvetett kapcsolat van:

A másik megoldás fenntartja az azonos szerkezetet a mondat bármely előfordulása esetén, igaz, ennek az az ára, hogy fel kell tételeznünk egy virtuális csomópontot a létige kijelentő mód jelen idő E/3. alakja számára (VAN). Így a következőképpen alakulnak a függőségi fák:

További érv a virtuális csomópont alkalmazása mellett, hogy szintaktikai szinten mindenképpen jelen van a VAN, hiszen a többi igealak/igeidő/igemód esetében testes morfémaként jelenik meg. Az már másodlagos (morfológiai) kérdés, hogy jelen idő E/3-ban miért zéró morféma az alakja (vö. [18]). Előnyt jelenthet a virtuális csomópont alkalmazása a korpusz nemzetközi felhasználhatóságában is, hiszen például egy függőségi fákra épülő fordítóprogram jóval hatékonyabb működésre képes, ha azonos struktúrájú fát kell leképeznie a másik nyelvre, szemben azzal, ha még ráadásul külön transzformációs lépéseket is be kell iktatnia a fordítás folyamatába.

5 Statisztika

A Szeged Treebank 2.0 állománya 82.000 mondatot, 1,2 millió szövegszót és 250 ezer írásjelet tartalmaz. A szövegek hat különböző témakörből kerültek ki, témakörönként ~200 ezer szó terjedelemben. A témakörök a következők:

- Szépirodalom
- 14-16 éves korú tanulók fogalmazásai
- Újságcikkek (Népszabadság, Népszava, Magyar Hírlap, HVG)
- Számítástechnikai szövegek
- Jogi szövegek
- Gazdasági és pénzügyi rövidhírek

2009 novemberéig a gazdasági és pénzügyi rövidhíreket tartalmazó alkorpusz, az újsághírek és a jogi szövegek dependenciaelemzése készült el teljes egészében, illetve a számítógépes témájú szövegek elemzése zajlik jelenleg. Az eddig elkészült korpusz statisztikai adatai a következő táblázatban foglalhatók össze:

1. táblázat: A korpusz statisztikai adatai.

	newsml	**újsághírek**	**jogi szövegek**	**összesen**
Mondatok	9574	10210	9278	29062
Szavak	186030	182172	220069	588271
Írásjelek	25712	32880	33515	92107

Az annotációs munkálatok várhatóan 2010 elején fejeződnek be.

6 A korpusz hasznosíthatósága

A számítógépes nyelvészet több területén is haszonnal bírhat a függőségi fák alkalmazása: mind a gépi fordításban, mind az információkinyerésben sikeresen felhasználhatók a függőségi fa formátumú korpuszok.

6.1 Gépi fordítás

A szintaktikai transzformáción alapuló gépi fordítási eljárások alapvetően két forrásra építenek: vagy a forrásnyelvi konstituensfákat képezik le a célnyelvi konstituensfára, vagy pedig függőségi fákkal dolgoznak. A konstituensfákat alkalmazó módszer előnyei közé tartozik, hogy rokon nyelvek gépi fordítására jól alkalmazható, hiszen a rokon nyelveknek többnyire hasonló a szintaxisa, továbbá az eltérő szórendből adódó problémákat is elfogadható mértékben oldja meg. A módszer hátránya viszont, hogy rendkívül bonyolult és költséges transzformációs szabályokat kell bevezetni a rendszerbe, ráadásul ha a mondatnak teljesen eltérő szintaktikai szerkezete van a forrás-, illetve a célnyelvben, a fordítás teljesen elfogadhatatlanná válik.

Gyakori hiba továbbá a konstituensfákat használó fordítórendszerekben, hogy az elemző gyakran hibás szerkezetet tulajdonít a fának, felesleges címkéket szúr be vagy rossz csomópontokat feleltet meg egymásnak. A mesterséges csomópontokból adódó hibák kiküszöbölését sikeresen oldják meg a függőségi fákra alapuló fordítórendszerek, hiszen a függőségi fákban nincsenek absztrakt (mesterséges) csomópontok. A fa minden csomópontja így egy természetes nyelvi elemnek feleltethető meg a mondatban, a fa nem tartalmaz szintaktikai csomópontokat, a nyelvek közti szintaktikai különbségek így eltűnnek. A gépi fordítási eljárás során minden csomópont lefordítódik, és ha szükséges, akkor a csomópontok újrarendeződnek bizonyos előre megadott valószínűségek mentén. A függőségi fákat alkalmazó gépi fordítási eljárás különösen a nem rokon vagy eltérő szintaxisú nyelvpárok esetén lehet gyümölcsöző.

6.2 Információkinyerés

A számítógépes nyelvészet egy más területén, az információkinyerésben is hasznosíthatók a függőségi fák. A szintaktikailag annotált korpuszok igen fontos szereppel bírnak az automatikus információkinyerés területén, ugyanis nem elégséges csak azt tudni, hogy milyen szavak, illetve kifejezések szerepelnek az adott szövegben, annak is nagy jelentősége van, hogy ezek egymással milyen viszonyban állnak. Például gazdasági jellegű szövegekben a különböző tranzakciókról szóló információk között szerepelnie kell annak is, hogy ha A és B cég vett részt egy adásvételi folyamatban, akkor melyik cég vásárolta fel a másikat (azaz melyik a *felvásárol* ige alanya és tárgya). Ahhoz azonban, hogy ezt nagy biztonsággal meg lehessen állapítani, szintaktikai viszonyokat is tudni kell elemeznie az információkinyerő rendszernek. A szintaktikai viszonyok tanításában rendkívüli szereppel bírnak a szintaktikailag annotált korpuszok.

A kötött szórenddel rendelkező nyelvek esetén jó alternatíva lehet a konstituensfákat használó, szintaktikailag annotált korpusz: ezekben ugyanis adott szintaktikai szerkezethez adott szintaktikai viszony társul. A függőségi nyelvtanokra épülő korpuszok azonban inkább a szabad szórendű nyelvek esetén nyújtanak nagy segítséget az információkinyerésben, hiszen esetükben a szintaktikai viszonyokat illetően nem lehet a szórendet segítségül hívni: a függőségi nyelvtanok lényege, hogy a szórendtől függetlenül képes meghatározni a mondat szintaktikai szerkezetét.

Jelen korpuszban jelölve vannak az ige és bővítményei közti alapvető viszonyok, azaz a bővítmények közül az alany, tárgy és részeshatározó szerepű argumentumok könnyen azonosíthatók (SUBJ, OBJ és DAT címkével vannak ellátva), a további bővítmények pedig OBL címkével rendelkeznek. Így az információkinyerő program is sikeresen meg tudja állapítani a következő példában rejlő szintaktikai viszonyokat:

Az E.ON_Hungária_Energetikai_Rt. 87,713 százalékra növelte részesedését a Tiszántúli_Áramszolgáltató_Rt-ben.

A kinyerhető releváns szintaktikai viszonyok a következők:

növelte - Az E.ON_Hungária_Energetikai_Rt. (alany)
növelte – részesedését (tárgy)
növelte – a Tiszántúli_Áramszolgáltató_Rt-ben (bővítmény)

A szintaktikai viszonyokból a számítógép számára is kiderül, hogy a mondatban szereplő két Named Entity viszonya milyen, azaz az E.ON rendelkezik tulajdonrészszel a Titászban, és nem fordítva, ezáltal a szintaktikai viszonyokat is felhasználó információkinyerés pontossága igencsak megjavul az azokat nem hasznosító modellekhez képest.

6.3 Többnyelvűség

A magyar nyelvű dependeciakorpusz létrehozásával lehetőség nyílik a többnyelvűséget szem előtt tartó alkalmazások fejlesztésére is. A Szeged Treebank alkorpuszai

közül a kapcsolódási pontot a többnyelvű (párhuzamos) korpuszokhoz az *1984* és a *Windows2000* szövegállományok jelenthetik, hiszen ezeknek a szövegeknek bizonyosan létezik idegen nyelvű megfelelője is. Amennyiben az idegen nyelvű verziók tartalmaznak függőségi viszonyokra alapuló szintaktikai annotációt, könnyen létre lehet hozni egy magyar-adott nyelvű párhuzamos dependenciakorpuszt. Ez nagyban elősegítené egyrészt a többnyelvű információkinyerést támogató rendszerek fejlesztését, másrészt pedig a függőségi fákon alapuló, szintaktikai módszerekre építő gépi fordítóprogramok létrehozását. A korpusz létrehozása tehát mind elméleti, mind gyakorlati szempontok alapján jelentőségteljesnek és haszonnal kecsegtetőnek nevezhető.

7 Összegzés

A tanulmányban a Szeged Treebank függőségi fa formátumra történő átalakításának folyamatát mutattuk be: ismertettük a munkafolyamatokat, a felmerült problémákat és az azokra nyújtott megoldásokat. Szót ejtettünk a korpusz gépi fordításban, illetve információkinyerésben való hasznosíthatóságáról, továbbá a kontrasztív nyelvészet és a dependeciaszintaxis kutatói is számára haszonnal bírhat az adatbázis. A későbbiekben szeretnénk továbbá kifejleszteni egy magyar nyelvű dependenciaparsert is (vagy egy már rendelkezésre álló korábbi (például a MaltParser [15]) testreszabásával, vagy pedig önálló kutatás-fejlesztés eredményeként), melyhez az elkészült korpusz tanító adatbázisként szolgálhat.

Köszönetnyilvánítás

A kutatást – részben – a TUDORKA és a MASZEKER projekt (Jedlik Ányos programok) keretében az NKTH támogatta.

Hivatkozások

1. Hajič, J., Böhmová, A., Hajičová, E., Vidová Hladká, B.: The Prague Dependency Treebank: A Three-Level Annotation Scenario. In: A. Abeillé (ed.): Treebanks: Building and Using Parsed Corpora, Amsterdam:Kluwer (2000) 103-127
2. Čmejrek, M., Cuřín, J., Havelka, J., Hajič, J., Kuboň, V.: Prague Czech-English Dependecy Treebank: Syntactically Annotated Resources for Machine Translation. In: 4th International Conference on Language Resources and Evaluation, Lisbon, Portugal (2004)
3. Čmejrek, M., Cuřín, J., Havelka, J.: Prague Czech-English Dependecy Treebank: Any Hopes for a Common Annotation Scheme? In: HLT/NAACL 2004 Workshop: Frontiers in Corpus Annotation, Boston, Massachusetts (2004) 47-54
4. Hajič, J., Smrž, O., Zemánek, P:, Šnaidauf, J., Beška, E.: Prague Arabic Dependency Treebank: Development in Data and Tools. In: Proceedings of the NEMLAR International Conference on Arabic Language Resources and Tools. Cairo, Egypt, September 2004. (2004) 110-117

5. Nivre, J.: Theory-Supporting Treebanks. In: Nivre, J. and Hinrichs, E. (eds.) Proceedings of the Second Workshop on Treebanks and Linguistic Theories (TLT 2003), Växjö University Press (2003) 117-128
6. Prokopidis, P., Desipri, E., Koutsombogera, M., Papageorgiou, H., Piperidis, S.: Theoretical and practical issues in the Construction of a Greek Dependency Corpus. In: Proceedings of the 4th Workshop on Treebanks and Linguistic Theories (TLT-2005), Barcelona (2005)
7. Boguslavsky, I., Grigorieva, S., Grigoriev, N., Kreidlin, L., Frid, N.: Dependency Treebank for Russian: Concept, Tools, Types of Information. In: Proceedings of the 18th conference on Computational linguistics. Saarbrücken, Germany (2000) 987–991
8. Džeroski, S., Erjavec, T., Ledinek, N., Pajas, P., Žabokrtský, Z., Žele, A.: Towards a Slovene Dependency Treebank. In: Proceedings of Fifth International Conference on Language Resources and Evaluation, LREC'06, 24-26 May 2006. Genoa, Italy (2006)
9. Lepage, Y., Shin-Ichi, A., Susumu, A., Hitoshi, I.: An annotated corpus in Japanese using Tesnière's structural syntax. In: Proceedings of COLING-ACL'98 Workshop on the Processing of Dependency-based Grammars, Montreal (1998)
10. Liu, H.: Building and Using a Chinese Dependency Treebank. GrKG/Humankybernetik No. 48 Vol. 1 (2007) 3–14
11. Bamman, D., Crane, G.: The Design and Use of a Latin Dependency Treebank. In: Proceedings of the Fifth International Workshop on Treebanks and Linguistic Theories (TLT 2006) (Prague) (2006) 67-78
12. Nivre, J., Hall, J., Kübler, S., McDonald, R., Nilsson, J., Riedel, S., Yuret, D.: The CoNLL 2007 Shared Task on Dependency Parsing. In: Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL 2007, Prague (2007) 915–932
13. Alexin, Z.: A frázisstrukturált Szeged Treebank átalakítása függőségi fa formátumra. In: Tanács, A., Csendes, D. (szerk.): V. Magyar Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2007). Szegedi Tudományegyetem, Szeged (2007) 263-266
14. http://ufal.mff.cuni.cz/~pajas/tred/
15. Nivre, J., Hall, J., Nilsson, J., Chanev, A., Eryigit, G., Kübler, S., Marinov, S., Marsi, E.: MaltParser: A language-independent system for data-driven dependency parsing. Natural Language Engineering, No. 13, Vol. 2. (2007) 95-135.
16. Koutny I., Wacha B.: Magyar nyelvtan függőségi alapon. Magyar Nyelv Vol. 87 No. 4. (1991) 393–404.
17. Mel'čuk, I. A.: Dependency Syntax: theory and practice. State University of New York Press, Albany, NY (1988)
18. Mel'čuk, I. A.: Levels of Dependency in Linguistic Description: Concepts and Problems. In Agel, V., Eichinnger, L., Eroms, H.-W., Hellwig, P., Herringer, H. J., Lobin, H. (eds.): Dependency and Valency. An International Handbook of Contemporary Research, vol. 1, Berlin-New York, W. de Gruyter (2003) 188-229
19. Prószéky, G., Koutny, I., Wacha, B.: Dependency Syntax of Hungarian. In: Maxwell, Dan; Klaus Schubert (eds.) Metataxis in Practice (Dependency Syntax for Multilingual Machine Translation), Foris, Dordrecht, The Netherlands (1989) 151–181
20. Tesnière, L.: Éléments de syntaxe structurale. Paris, Klincksieck (1959)