Kötőszók korpuszalapú vizsgálata

Gábor Kata[1], Héja Enikő[1], Mészáros Ágnes[1]

[1] MTA Nyelvtudományi Intézet, Korpusznyelvészeti Osztály
1068 Budapest, Benczúr u. 33
{gkata,eheja,magnes}@corpus.nytud.hu

Kulcsszavak: kötőszó, tagmondathatár, szintaktikai elemzés, MNSZ

Az előadásunkban bemutatni kívánt munka célja egy, a magyar nyelv felszíni szintaktikai viszonyait felismerni képes elemző létrehozása. Az általunk alkalmazott szintaktikai elemzés a jelenségeknek alapvetően kétféle megközelítését alkalmazza: a kötött szórendű, reguláris kifejezéssel viszonylag egyszerűen leírható frázisokat zárójelezéssel különíti el, az így felismert összetevők között pedig dependencia-viszonyokat kódol.

Az elemzéshez a Max Silbeztein által kifejlesztett Intex nevű eszközt használjuk. Az Intex olyen fejlesztői környezet, mely természetes nyelvek formális leírását, illetve nagy méretű szövegek gyors feldolgozását teszi lehetővé. A nyelvleírás eszközei az elektronikus szótárak, a nyelvtani szabályokat tartalmazó véges állapotú transzducerek, illetve a táblázatokban kódolt lexikai információból kompilált lexikai transzducerek. Az Intex két fontos jellegzetessége, hogy véges állapotú transzducerei tartalmazhatnak önbeágyazást (melyet a program kompiláláskor tízszeres rekurzióvá alakít), valamint képes megjegyezni a nyelvtan által felismert szekvenciákat, és lehetővé teszi, hogy változókkal hivatkozzunk rájuk.

Az elemző első lépésben a kötött szórendű frázisokat (NP, AP, névutós szerkezet) ismeri fel és látja el cimkézett zárójelekkel. Az elemzés későbbi műveletei az így megjelölt összetevőket ezután már egy egységként kezelik. A következő lépésben a címkézett frázisokat olyan jegyekkel látjuk el, melyek az összetevők közötti viszonyok feltérképezéséhez (a frázishatárok megjelölésén túl) szükségesek lehetnek: az NP megkapja a fej esetragját, az NP és az AP pedig fráziskénta a fejének a szótárban kódolt szemantikai jegyeit. Az elemzési folyamat második fázisa az összetevők mondattani szerepének felismerése. Mondattani szerep alatt dependencia-viszonyokat értünk. Kiindulópontul az igei állítmány vonzatkerete szolgál, melyet a rendelkezésre álló lexikai adatbázis alapján azonosítunk. Az argumentumszerkezetet kódoló adatbázis esetragos névszói, esetenként igei, illetve tagmondat kategóriájú vonzatokat tartalmaz. A magyar nyelv sajátosságai miatt a lehetséges argumentum-szerepű frázisok kiválasztásánál nem kívánunk a szórendi pozícióra hivatkozni: a táblázatban kódolt igei szubkategorizációnak megfelelő frázisokat akkor akarjuk ellátni a von-

zatszerepnek megfelelő annotációval, ha ugyanabban a *tagmondatban* vannak, mint a lehetséges régensük. Ez azt jelenti, hogy a vonzatkeret illesztését meg kell hogy előzze egy olyan lépés, mely kijelöli a *tagmondathatárokat.*

Hogy mit értünk tagmondat alatt, azt nagyban meghatározza az elemzés választott kerete. Mivel munkánk során az igei argumentumok azonosítására törekszünk, és a tagmondathatár jelöli ki azt a területet, amin belül az igei argumentumokat keressük, célszerű elfogadnunk az alábbi hipotézist:

Egy mondatban két finit ige között mindig van tagmondathatár.

Ahhoz azonban, hogy eldöntsük, melyik ponton található, több információra van szükségünk. Előadásunkban arra szeretnénk kitérni, hogyan használhatók a kötőszók a tagmondathatár pontos helyének azonosítására olyan esetekben, amikor tudjuk, hogy egy adott tartományban valahol lennie kell tagmondathatárnak.

Bár természetesen nem feltételezhetjük, hogy minden tagmondathatáron találunk kötőszót, azt érdemes megvizsgálnunk, hogy ahol (két finit ige között) olyan szót találunk, mely adatbázisunkban kötőszóként szerepel, ott mindig van-e tagmondathatár. Ha a kötőszót olyan szófajként definiáljuk, mely mindig két azonos típusú összetevőt köt össze, akkor a kötőszók összes előfordulása közül azokat az eseteket kell kiszűrnünk, amikor a kötőszó bal és jobb oldalán egyaránt tagmondat kategóriájú elemet találunk. Mivel azonban a (tag)mondat szerkezetét dependencia-viszonyokkal és nem zárójelezéssel akarjuk ábrázolni, ráadásul a szerkezet feltárásához már korábban fel akarjuk használni a tagmondathatárokat, ez nem járható út. Ha viszont a tagmondatok elhatárolására akkor kerül sor, amikor reguláris nyelvtanokkal már felépítettük az NP, AP, illetve PostP kategóriájú frázisokat, melyek tartalmazhatnak kötőszókat, akkor csak azokat a kötőszókat kell vizsgálnunk, melyek a megjelölt frázisokon kívül esnek.

Lexikai adatbázisunk létrehozásakor már megfigyeltük, hogy a kötőszóként felvett elemek eltérően viselkednek abból a szempontból, hogy a tényleges tagmondathatárhoz képest hol helyezkednek el. Egyes kötőszók mindig az első helyet foglalják el a tagmondatban (pl. “hanem”), mások jellemzően a határ utáni első összetevőt követik (pl. “pedig”), néhány kötőszó pedig szinte bármilyen pozícióban megjelenhet (pl. “azért”). Ezt a tulajdonságot lexikai jellemzőnek tekintettük, és kódoltuk az adatbázisban. A kötőszók korpusz-alapú vizsgálata azonban azt mutatta, hogy az adatbázisunkban szereplő 71 kötőszó legtöbb előfordulásában még a lehetséges pozíciójukra vonatkozó információ birtokában sem elégséges a tagmondathatár kijelöléséhez. Ennek oka egyrészt az, hogy a hagyományosan kötőszónak tekintett szavak nem alkotnak egységes csoportot, hanem több, eltérő disztribúciójú alcsoportra oszthatók (a különböző alcsoportokba tartozó elemek esetenként együtt is előfordulhatnak), másrészt pedig az, hogy a kötőszóként funkcionáló elemek között nagyon sok az olyan alak, mely más szófaji elemzést is kaphat (pl. gyakori a kötőszó-adverbium homonímia). Előadásunkban a kötőszók disztribúciójának a Magyar Nemzeti Szövegtár adatain alapuló vizsgálatát kívánjuk bemutatni, melynek eredményeképp javaslatot teszünk a kötőszók egy olyan csoportosítására, amely eredményesen hasznosítható a tagmondathatár felismeréséhez.