Főnévi csoport annotációja a CLaRK rendszerrel

Váradi Tamás

[3] MTA Nyelvtudományi Intézet
1068 Budapest, Benczúr u 33
varadi@nytud.hu

Kulcsszavak: felszíni szintaktikai elemzés, NP annotáció, lépcsős reguláris grammatika

Absztrakt. A magyar mondat szerkezetének leírásában kiemelt szerepet játszik a főnévi csoport. E dolgozat keretében beszámolunk arról a folyó munkáról, amely véges állapotú grammatika alkalmazásával megkísérli főnévi csoportok lehető legteljesebb felszíni leírását. Az ún. lépcsős reguláris grammatika (Abney 1996) kifejlesztése a CLaRK rendszerrel történt, melynek bemutatása szintén, melynek bemutatása szintén célja a jelen dolgozatnak.

1 Bevezetés

A dolgozat célja, hogy betekintést adjon a főnévi csoport automatikus felismerését célzó munkálatokba. A kutatások jelenleg is folynak, ezért az itt közzétett eredmények csak közbenső jelentésnek tekinthetők. A főnévi csoport annotációt szabályokra épülő rendszerben, a lépcsős reguláris grammatika módszerével (Abney 1996) végezzük. A fejlesztői keretrendszerül a CLaRK rendszert használjuk (Simov 2001), amely hatékonyan támogatja a kézi grammatikafejlesztést. A dolgozat felépítése a következő: az 2. részben ismertetjük a magyar főnévi csoport gépi feldolgozás szempontjából releváns sajátosságait, a 3. rész bemutatja a feldolgozott adatok szerkezetét és annotációjukat. Ezt követi a CLaRK rendszer rövid áttekintése a 4. részben, mely után ismertetjük a főnévi csoport felismerésére kifejlesztett szabályrendszer fő elveit. Az 6. rész tartalmazza magukat a szabályokat, melyek értékelését a 7. részben találjuk.

2 A kiinduló nyelvi tények rövid jellemzése

A magyar nyelvet közkeletű felfogás szerint szabad szórendű nyelvnek tekintik. Pontosabban fogalmazva, a magyarban a mondatszintű szintaktikai összetevők (szintagmák) viszonylag szabad sorrendben helyezkedhetnek el. Lényeges azonban látnunk, amint azt É. Kiss (1994) Brassai nyomán hangsúlyozza, hogy a mondatok szórendjét a topic-comment szerkezet határozza meg elsősorban, amelyet viszont a közlés

illetve a mondatokon átívelő szöveg kommunikációs sajátosságai szabnak meg. A szintagmákon belül az összetevők sorrendje kötött.

A viszonylag szabad szórendet a rendkívül gazdag alaktan teszi lehetővé, ugyanis a szintaktikai szerepeket a szintagmák főtagjának ragja jelzi. Ebből fakad az a sajátosság, hogy az egyszerű magyar mondatok döntő többségét egy ige és a körülötte található ragos főnévi csoportok alkotják. Esetragos főnévi csoportokkal fejezünk ki olyan viszonyokat, amelyeket más nyelvekben prepozíciós kifejezésekkel vagy határozószókkal fejtünk ki. Ez a tény ad kitüntetett jelentőséget a főnévi csoportok vizsgálatának.

A főnévi csoportok belső szerkezetének sajátosságaiból csak néhányat emelünk ki, amelyek megnehezíthetik az automatikus felismerést. Az első tény, amit megjegyezhetünk, hogy sajnos nincsenek olyan egyértelmű támpontok, amelyek minden helyzetben jelölnék a főnévi csoportok határait. A ragos főnevektől várhatnánk, hogy egyben a főnévi csoport jobb szélét is jelölnék de a birtokos és az igeneves szerkezetek miatt ez gyakran nincs így, másrészt a főnévi csoportból hiányozhat is maga a főnév, mely esetben a jelző veszi át a szerepét és egyben toldalékait. A főnévi csoport kezdetét egy determináns elem jelölheti ugyan, de ezek jelenléte még kevésbé feltételezhető, mint a főnévi fejé, másrészt a rekurzív beágyazódásból és az igenes szerkezetek bővítményeiből az is következik, hogy nem egyszerű feladat a determináns elem hovatartozását megállapítani.

Az igeneves szerkezetek elemzése különleges nehézséget jelent. A problémát az okozza, hogy a folyamatos vagy befejezett igenév (melynek szófaji besorolása szintén nem egyszerű feladat, hiszen az gyakran megkívánja a szintaktikai szerep elemzését is) olyan elem, amelyik gyakran hozza magával a bővítményei egész sorát mintegy beágyazott tagmondatot alkotva a főnévi csoporton belül. Egyéb nyelvekben a főnévi fejet követő prepozíciós szerkezettel fejezzük ki mindezt, itt tehát ugyanazzal a problémával találkozunk a magyar főnévi csoporton belül, amelyet a prepozíciós szerkezettel bíró nyelvekben a PP csatolás nehézségei címszó alatt tartanak számon.

3. Az adatok

A főnévi csoportok annotációját megelőzi a szöveg morfoszintaktikai elemzése. Ez arra a technológiára épül, amellyel a Magyar Nemzeti Szövegtár elemzett és egyértelműsített változata készült. A jelen kísérlethez az MNSZ morfoszintaktikai annotációjának némileg leegyszerűsített xml változatát használtuk. Az egyszerűsítés nem érintette a szavakhoz társított nyelvi elemzést. Minden szóalak (token) egy <w> elemen belül fordul elő és három attributum tartozik hozzá, melyek a lemmát, a morfoszintaktikai jellemzőt (msd) és a korpusz tag-et tartalmazzák.

A szintaktikai elemzés minőségét nagyban meghatározza a morfoszintaktikai annotáció és az egyértelműsítés pontossága. Az MNSZ annotációs rendszere alapvetően a HUMOR rendszer (Prószéky és Tihanyi 1996) jelkészletét használja, bár annak kimenetét további szűrésnek veti alá a párhuzamos elemzések kiszűrése és a lemma megállapítása céljából. Az egyértelműsítés pontossága eléri a 98%-ot (Oravecz és Dienes 2002).

A feldolgozott szövegeket a *Heti Világgazdaságból* merítettük. A választás szándékosan azért esett erre a folyóiratra, mert benyomásunk szerint a cikkek olyan kimunkált, időnként már-már mesterkélt stílusban íródtak, amelyek nagy számban tartalmaznak rendkívül összetett NP szerkezeteket. Bízvást állíthatjuk tehát, hogy ez a szöveg igazán próbára teszi az annotáló rendszert. Ugyanakkor azonban ezt a tényt érdemes figyelembe venni az eredmények értékelésekor.

4 A fejlesztő eszköz

Az NP annotálási szabályok fejlesztését a CLaRK rendszer (Simov et al. 2002) segítségével végezzük. A CLaRK rendszer egy XML alapú korpuszfeldolgozó eszköz, amely három technológia egyesítésével biztosítja a hatékony szövegannotációt: az Xpath mechanizmus biztosítja a szöveg tetszőleges részének elérését, a beépített véges automata dolgozza fel a reguláris kifejezésekkel definiált nyelvtant, és az ú.n. megszorítás (constraint) szabályok alkalmazásával növelhetjük az XML technológia rugalmasságát.

A legalsó szinten egy tokenizáló modul bontja fel a szöveget a kívánt egységekre. A tokenizáló szabályok tetszés szerint definiálhatók, lépcsősen egymásra épülnek, és akár minden szabályhoz külön-külön is hozzárendelhetők. A szöveg feldolgozásának központi eleme a lépcsős reguláris grammatika, amelynek szabályaihoz az Xpath kifejezések segítségével definiáljuk a szabályok hatókörét és a szöveg feldolgozandó elemeit. A nyelvtani szabályok meghatározásakor módunk van a reguláris kifejezések bal és jobb oldalán lévő szövegkontextus definiálására. A szabályok kimenete egy XML annotáció, amelyet általában arra használunk, hogy a szabályra illeszkedő szövegrész köré XML kódokat ültessünk. A nyelvtan lépcsős jellegét az biztosítja, hogy az egyes szabályok kimeneteként előállt egységek szerepelhetnek a későbbi szabályok bemenetében. Az XML annotáció jól illeszkedett a nyelvtan hierarchikus szerkezetéhez és az Xpath kifejezések valamint a constraint szabályok alkalmazásával könnyen meg lehetett fogalmazni olyan szabályokat, mint például a head jegyeinek perkolációját a legfelsőbb kiterjesztési szintre még akkor is, amikor az összetett NP struktúra miatt a két pont igen távol esett egymástól.

5. Az NP annotáció általános elvei

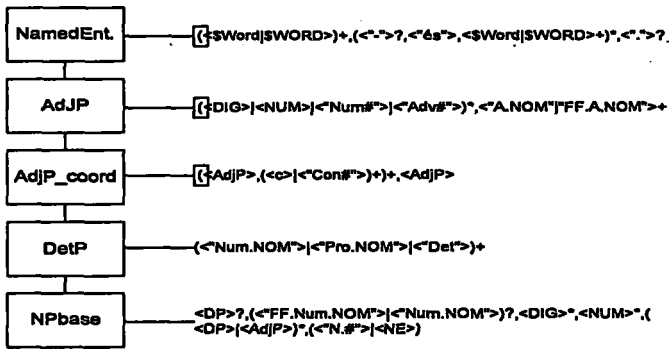
A 2. részben ismertetett sajátosságokat figyelembe véve a következő elvekre építettük a főnévi felismerő szabályainkat. Mivel a magyarban a főnévi csoport belső szerkezete balra rekurzív, az NP bal szélső eleme az NP feje, amit alapfeltevésként azaz a szabályok első körében egy N tölt be. A leghosszabb illeszkedő mintát használtuk a reguláris kifejezésekben. Az NP-n belül szerepelhet módosító szerepben N is, de csak nominatívusz esetben. A teljes NP annotációs nyelvtan két szakaszra bomlik: az elsőben meghatározzuk azokat az egyszerű NP-ket, amelyeknek a feje N vagy tulajdonnév

(NamedEntity). Amint az az 1 . ábrán látható, ez a szakasz is lépcsősen egymásra hivatkozó szabályokból épül fel.

Arra való tekintettel, hogy a magyarban a főnévi csoportok fejének szerepét a főnevek hiányában módosító elemek is átvehetik, az elemzés további szakaszában a „depth-first" stratégiát követtük, vagyis először az N fejű összetett főnévi csoportok szerkezetét határoztuk meg egészen addig, amíg a szabályok már nem találtak illeszkedő adatokat, majd ezután következett a nem N-fejű NP-k feldolgozása. Még itt is két szakaszt kellett elválasztanunk, először ugyanis csak olyan NP-ket határoztunk meg, melyekben a fej szerepét nem igenév tölti be, majd csak ezek kimerítő lefedése után engedhettük meg az igeneveket fej szerepben (ld. NP2 és NP3 a 2. ábrában). Külön problémát jelentett az igeneves szerkezetek előtt álló módosító elemek jobb szélének a meghatározása. Jobb híján kénytelenek voltunk megengedni tetszőleges NP bővítményt, ami kétségtelenül a túlgenerálás egyik forrása lehet.

6. Az NP felismerő szabályrendszer

A kidolgozott szabályokat az 1. és a 2. ábra tartalmazza. Amint látható a szabályok egyaránt hivatkoznak szintaktikai osztályokra (<DP>), msd attributumokra (<"FF.Num.NOM">) és szóalakokra (<"és">). A reguláris kifejezések sajátos notációjának leírását a CLaRK rendszer leírásában találhatjuk (Simov 2001).



1. ábra. Az alapszintű NP-k szabályrendszere

7. Eredmények

A szabályrendszert 100 kézzel azonosított mondaton (gold standard) teszteltük. A tesztelésben a legfelsőbb szintű NP-k helyességét vizsgáltuk. A 2537 tokent tartalmazó tesztszöveg összesen 488 mondatszintű NP-t tartalmazott. Két mérőszámot is célszerűnek tartottunk alkalmazni, az egyik a szerkezetekre vonatkoztatva mutatja a pontosság és lefedettség számait, a másik az érintett szövegszavakban méri ugyanezt.

(i) szerkezeti mutatószámok:

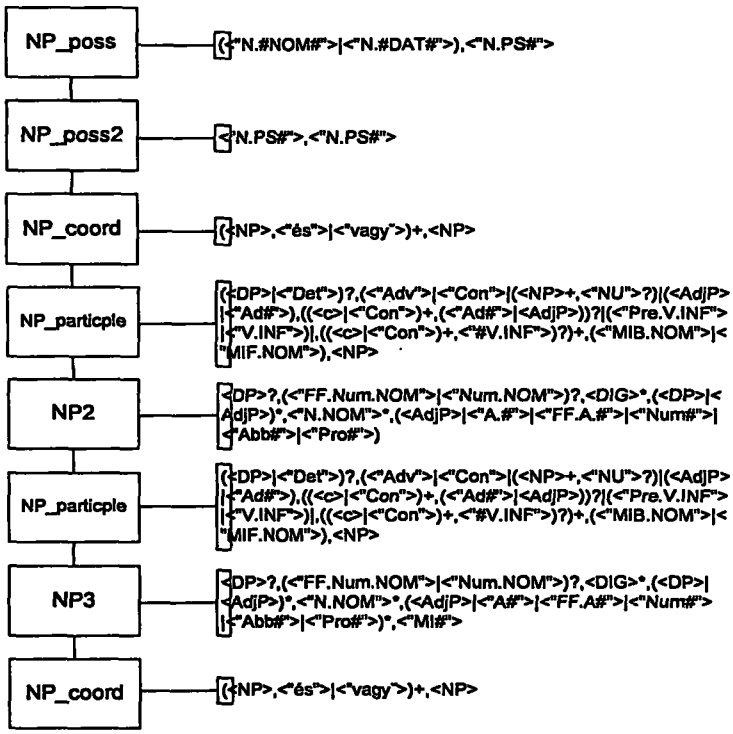
- pontosság: a kézzel ellenőrzött és a mintában egyaránt szereplő NP-k száma/ a mintában szereplő NP-k száma

- lefedés: a kézzel ellenőrzött és a mintában egyaránt szereplő NP-k száma / a kézzel ellenőrzött anyagban szereplő NP-k száma

(ii) szóalak mutatószám : Ugyanaz a két arány, mint (i)-ben, de nem NP-k ben, hanem a szóalakok számában meghatározva.

Az FB1 értékeket a szokásos módon, az alábbi képlet szerint számoltuk: FB1=2*pontosság * lefedés/(pontosság + lefedés).

Az eredményeket az 1. és a 2. táblázatban foglaltuk össze.



2. ábra. Az összetett NP szerkezeteket előállító szabályrendszer

A számszerű mutatókat az érintett szavak tekintetében elfogadhatónak tekinthetjük. A nyelvtan kimenetét minőségileg vizsgálva a benyomásaink kedvezőbbek annál, mint amit a számok tükröznek. Az eltérést részben az indokolja, hogy, amint azt a 3. részben említettük, a feldolgozott szöveg a főnévi csoportok szempontjából több tekintetben is extrémnek tekinthető. A szabályrendszer kimenetének hibaelemzése jelenleg is folyik. A további munka kereteit egyértelműen kijelölik a jelenlegi szabályok által lefedett jelenségek ismert korlátai.

NP szám gold standard-ban:	488
NP szám a mintában:	611
Helyes NP-k száma	323
NP pontosság:	52.87%
NP lefedettség:	66.17%
FB1:	58.78%

1. táblázat NP szerkezeti mutatószámok

Szószám a gold standard-ban:	1660
Szószám a mintában::	1577
Szószám a helyes NP-kben:	1511
Szószám pontosság:	95.81%
Szószám lefedettség:	91.02%
FB1:	93.36%

2. táblázat Szóalak szerinti mutatószámok

Hivatkozások

Abney S 1996 Partial Parsing via Finite-State Cascades In *Proceedings of the ESSLLI'96 Robust Parsing Workshop,* pp 1 – 8

É. Kiss K 1994 Sentence structure and word order. In: Kiefer–É. Kiss (eds): *The Syntactic Structure of Hungarian.* San Diego, Academic Press. 1–90.

Oravecz Cs, Dienes P 2002: Efficient stochastic part of speech tagging for Hungarian. In *Proceedings of the Second International Conference on Language Resources and Evaluation,* Las Palmas, pp 710—717

Prószéky G, Tihanyi L 1996 "Humor -- a Morphological System for Corpus Analysis." *Proceedings of the first TELRI Seminar in Tihany.* Budapest, pp 149-58.

Simov K 2001 CLaRK – an XML-based System for Corpora Development in *Proceedings of the Corpus Linguistics 2001 Conference,* Lancaster, pp 553-560.

Simov K et al. 2002 CLaRK System: Construction of Treebanks in *The First Workshop on Treebanks and Linguistics* Theories Sopozol: LML CLPP Bulgarian Academy of Sciences 183-199.