Információkinyerés igeneves szerkezetekből

Gábor Kata, Héja Enikő, Mészáros Ágnes*

2004. augusztus 29.

Kivonat

Előadásunkban a NewsPro információkinyerő rendszer egy továbbfejlesztési lehetőségét mutatjuk be. A NewsPro egyik hiányossága, hogy csak igei állítmánnyal kifejezett eseményeket ismer fel, az igenévvel kifejezett eseményekre nem tudja illeszteni a szemantikai kereteket. Így a felhasználó az információ egy részéhez nem fér hozzá, valamint – mivel mondatonként csak egy eseményt ismer fel a rendszer – az események közti összefüggések is gyakran rejtve maradnak. Ennek kiküszöbölésére egy előfeldolgozó modult fejlesztettünk ki, mely az igeneves szerkezeteket teljes propozícióvá alakítja, így a szemantikai keretek minden további átalakítás nélkül illeszthetők ezekre.

1. Bevezetés

Az alábbiakban egy olyan nyelvészeti témájú alkalmazott kutatást szeretnénk bemutatni, melynek célja, hogy a szabályalapú információkinyerés hatékonyságát növelje. Munkánk az NKFP 2/017/2001 projektumban a MorphoLogic Kft., a Szegedi Egyetem Informatikai Tanszékcsoportja és az MTA Nyelvtudományi Intézet Korpusznyelvészeti Osztálya által elkészített NewsPro információkinyerő rendszer [1] továbbfejlesztését célozza. A NewsPro rendszer a bemeneti szövegen részleges szintaktikai elemzést hajt végre, majd előre definiált szemantikai kereteket, azaz eseménymintákat illeszt a szövegre. Sikeres illesztés esetén az eseményminták a szöveg elemeivel feltöltődnek, így a kimenet azonosítja a hírben szereplő eseményt, valamint annak szereplőit, attribútumait és körülményeit. A rendszer fejlesztésekor a vállalati rövidhírekre összpontosítottunk, így az eseménysablonok ezt a területet fedik le, de természetesen a program alkalmassá tehető tetszőleges tematikájú hírek kezelésére. A vállalati rövidhírek az MTI archívumából származnak. Egy hír általában egy mondatból áll. A hírekre illesztett eseményminták központjában ragozott igék állnak, melyek bővítményei képviselik az ige által kifejezett esemény szereplőit, körülményeit, attribútumait. A mintaillesztés tehát a szintaktikai elemző által állítmányként megjelölt igéből, illetve annak vonzatkeretéből indul ki. Emögött az az implicit feltételezés áll, hogy a hírben egy igei állítmány fejezi ki a fő eseményt. Ez a

*MTA Nyelvtudományi Intézet, Budapest VI. Benczúr u. 33. Pf. 701/518 H-1399, e-mail: {gkata,eheja,magnes}@nytud.hu

megközelítés, bár a hírek nagy részénél hatékonyan működik, gyakran azzal az eredménnyel jár, hogy a másodlagosnak, ismertnek feltételezett információkat, melyek többnyire a fő esemény előzményeként, okaként vannak feltüntetve, kihagyja a mintaillesztésből. Ezek a másodlagos információk ugyanis nem ragozott igék, hanem igéből képzett főnevek vagy igenevek formájában szerepelnek a szövegben. Például:

(1)

<u>A gyártók által tegnap bejelentett árcsökkentések</u> és a hitelkamatok mérséklése nyomán megnőtt a kereslet az új autók iránt.

Noha a fenti mondat központi információja a kereslet növekedése, hírértékkel bírhat az árcsökkenés is. Előfordulhat, hogy a felhasználó nem olvasta a korábbi híreket, vagy kíváncsi az események közti összefüggésekre, melyek akkor tárhatók fel, ha a rendszer képes egy hírben több eseményt is elemezni. A megoldandó feladat fontosságát jelzi, hogy az MTI rövidhírekből álló 25,902 mondatos korpusz összesen 6,567 folyamatos vagy befejezett melléknévi igeneves szerkezetet tartalmaz.

A jelenség kezelését a NewsPro rendszerben egy előfeldolgozó modul feladataként képzeltük el. A modul az igeneves szerkezeteket ragozott igét tartalmazó mondattá alakítja. Az így átalakított szövegen a nyelvtani elemzés és a szemantiaki keretek illesztése külön változtatás nélkül alkalmazható. Első lépésben csak a főnévi csoportokon belül előforduló befejezett melléknévi igenevekkel foglalkoztunk. Feltételeztük, hogy az igéből képzett melléknévi igenevek átalakíthatók ragozott igét tartalmazó propozícióvá, mert az igenév megőrzi az alapige jelentését, és argumentumai (legalábbis azok egy része) levezethetők a főnévi csoport szerkezetéből. A befejezett igenév mindig előidejű, így az ige múlt idejű lesz. Az átalakított mondatokon a mintaillesztés várhatóan még nagyobb pontossággal működik, mint az érintetlenül hagyott szövegrészeken, mivel a transzformáció során lehetőségünk van meghatározni a kimeneti mondatban a mondatrészek sorrendjét[1]. Ez pedig megkönnyíti a szintaktikai elemzést és az erre épülő eseménysablon-illesztést.

Az előfeldolgozó transzformáció sikere természetesen nem csak azon múlik, hogy hogyan sikerül az igeneves szerkezet szintaxisából levezetni a propozíciót, hanem azon is, hogy mekkora információtartalma van az így képzett mondatoknak. Kísérletet tettünk arra, hogy kialakítsunk egy olyan algoritmust, mely kizárólag szintaktikai információ alapján kiszűri a vélhetően informatív szerkezeteket.

A következő bekezdésekben először bemutatjuk azt a korpuszfeldolgozó eszközt, melyet a szabályok megírásához és teszteléséhez használtunk (2.). Ezt követően leírjuk az informatív szerkezetek kiszűrésére használt algoritmust (3.), majd részletesen ismertetjük a szabályokat (4.), végül kitérünk a szabályok tesztelésének eredményére (5.).

[1] A kimeneti mondatok elemeinek toldalékolásával egyelőre nem foglalkoztunk, ám - mivel kevés morfológiai változtatásra van szükség - ezt viszonylag rövid időn belül megoldhatónak gondoljuk.

2. A korpuszfeldolgozó eszköz

A transzformációt végző szabályok elkészítéséhez és teszteléséhez, valamint a szöveg szükséges előfeldolgozásával kapcsolatos valamennyi feladathoz az Intex nevű, kutatási célokra szabadon használható korpuszannotáló szoftvert [2] használtuk. Az Intex alapvetően lexikalista megközelítésű nyelvelemzésre alkalmas, alappillére az erre a célra kialakított szótár, mely egy szinten kódolja a morfoszintaktikai és a szemantikai információt, így az a nyelvtani elemzés minden szintje számára hozzáférhető. Ez nagy előnyt jelentett számunkra a transzformációt végző nyelvtan írásakor, hiszen – amint a következő fejezetekből kiderül – nyelvtanunknak hivatkoznia kellett az igenevek alapigéjére (amit szintén a szótárban kódoltunk), valamint az alapige szintaktikai jegyeire is.

3. Melyek az informatív igenevek?

Az adatok vizsgálata során kérdésként merült fel, hogy mikor érdemes egy befejezett melléknévi igenevet igévé alakítani. Egyfelől nem elhanyagolható a főnévi csoport által hordozott információtartalom, amely annál nagyobb, minél több bővítménye van jelen az igenévnek. Már ez elégséges indok arra nézve, hogy csak a bővítménnyel rendelkező igenevekkel foglalkozzunk. Azonban a fentieken túl sokkal komolyabb problémák is felmerülnek a bővítménnyel nem rendelkező igenevek igévé alakítása kapcsán. Ezt illusztrálják az alábbi NP-k és a szabályaink kimeneteként kapott mondatok:

(2)

a **jegyzett** tőke	[particip Valaki jegyzett tőke -t]
a **nyomott** hangulatot	[particip Valaki nyomott hangulatot -t]
a **mérsékelt** PC-chip kereslet	[particip Valaki mérsékelt PC-chip kereslet-t]
a **nyomtatott** sajtóban	[particip Valaki nyomtatott sajtóban -t]
a **ragozott** szóalakokból	[particip Valaki ragozott szóalakokból -t]
a **kerekített** euróárak	[particip Valaki kerekített euróárak -t]
a **használt** ingatlanok	[particip Valaki használt ingatlanok -t]

A fenti igenevek esetén a rövidesen ismertetésre kerülő átalakítási szabályok nem jól működnek. Ennek okát abban látjuk, hogy a szóbanforgó kifejezések valójában nem igenevek, hanem melléknevek, és a szófajváltással együtt a vonzatstruktúrájuk és a jelentésük is megváltozott. Így a kapott propozíciók helytelenségére két – egymástól nem független – magyarázatot adhatunk. Ha a jelentésváltozás egyértelmű (pl.:'*nyomott hangulat*'), a kiinduló ige jelentése nem releváns az NP jelentése szempontjából, így az eredeti igével való behelyettesítés szemantikailag helytelen mondatokat eredményez. Azon melléknevek esetében, ahol a jelentésváltás kevésbé éles, az igévé való visszaalakítás után azért kapunk szemantikailag helytelen mondatokat, mert – feltételezésünk szerint – a melléknévvé válás során az eredeti ige teljes vonzatstruktúrája törlődik . Így tehát az eredeti ige alanyi argumentumhelyén megjelenő főnévnek nincs szemantikai szerepe a melléknevet tartalmazó NP-ben. Azaz '*a nyomtatott sajtó*' esetén nem az a fontos, hogy valaki kinyomtatta azt a sajtóterméket, hanem az, hogy ez most már ilyen állapotban található. Hasonló a helyzet a '*kerekített euróárak*'-kal, a '*ragozott szóalakok*'-kal és a '*használt ingatlanok*'-kal is.

Az általunk kifejlesztett szabályrendszer alapja az a hipotézis, hogy csak a bővítménnyel rendelkező '*ige+(t)t*' alakú kifejezéseket tekintjük igeneveknek és a hasonló képzővel ellátott, ám bővítmények nélküli igéket mellékneveknek. Bővítmények alatt a kötelező vonzatokat vagy a szabad határozókat értjük. Így a (2)-ben szereplő NP-k kívül esnek vizsgálódásunk körén. Az alábbiakban felsorolunk néhány kritériumot, amelyek lehetővé teszik a melléknevek és az igenevek elkülönítését[3]:

(a) Predikatív helyzetben, fokozott formában csak melléknév fordulhat elő. Ezen teszt alapján levonhatjuk azt a következtetést, hogy példamondatainkban melléknevek szerepelnek[2]. Sok esetben ugyan – szemantikai okok miatt – nem fokozhatóak (pl.: **nyomtatottabb*), de minden esetben kerülhetnek állítmányi pozícióba[3].

(b) Továbbképzéssel csak melléknévből képezhető határozószó. A lexikalizálódott alakoktól eltekintve az összes (2)-ben szereplő kifejezésből képezhető határozószó[4]. Így ez a kritérium is azt támasztja alá, hogy a szóbanforgó esetekben melléknevekről van szó.

(c) Csak az igenevek előtt van elváló igekötő, a melléknevekben található igekötők nem válhatnak el[5].

Bár ez utóbbi szempont vajmi keveset árul el az eddig tárgyalt kifejezések szófajáról, mivel egyikőjük alapigéje sem rendelkezik igekötővel, ez a kritérium nemsokára még nagy segítségünkre lesz. Azt állítjuk, hogy ha egy megfelelő formájú ige környezetében azt módosító szabad határozót találunk, ez már elégséges alapot nyújt arra nézve, hogy az adott kifejezést igenévnek tekintsük[6], azaz nem szükséges vonzat megőrzése az igenéviséghez. Ezt az elgondolásunkat (a), (b), (c) disztribúciós feltételek alátámasztják:

(a') *„*A múlt héten mérsékeltebb PC-chip kereslet*" és *„*A PC-chip kereslet [a múlt héten mérsékelt] volt*" [7].

[2]Bár bizonyosos esetekben lexikalizálódott kifejezésekkel van dolgunk, amelyek a kritériumoknak nem megfelelően viselkednek (pl.: *'*jegyzettebb tőke*')

[3]Pl.: '*A hangulat nyomott volt.*', '*A PC-chip kereslet mérsékelt*', '*A magyar sajtó zöme nyomtatott*' (és nem elektronikus).

[4]'*A mérsékelten csökkenő PC-chip kereslet*' vs. *'*Az EU által mérsékelten csökkenő PC-chip kereslet*'; '*A használtan vásárolt ingatlanok*' vs. *'*Az árusításra használtan vásárolt ingatlanok*'.

[5]*'*A budai áruházak fel nem újítottak*' vs. '*az állam által fel nem újított utak*'.

[6]Itt Komlósy(1992) nézetével vitatkozunk, aki szerint az igenéviséghez szükséges az alapige vonzatainak megőrzése.

[7]Szerkezeti homoníma elkerülése érdekében ahol szükséges szögletes zárójellel jelöltük az összetevőket. Ha '*a múlt héten*' a '*mérsékel*' módosítója, akkor (a') (b') kritériumok valóban azt mutatják, hogy szabad határozóval módosított ige+t formájú kifejezés igenév.

(b') *„*A [múlt héten mérséklet]en csökkenő PC-chip kereslet*" vs „*A múlt héten mérsékelten csökkenő PC-chip kereslet*".

Tehát (a) és (b) alapján beláttuk, hogy jogos befejezett melléknévi igenévnek tekinteni minden '*ige+(t)t*' formájú kifejezést, ha bármilyen bővítményét (kötelező vonzat, szabad határozó) azonosítani tudjuk. Továbbá – ha közvetve is – de (c) is ezt támasztja alá; ha belátnánk, hogy az igekötők szabad határozók, akkor ennek egyik – szükséges – alapja az a megfigyelés lenne, hogy az esetek többségében az igekötő az igétől viszonylag függetlenül mozog. Mivel az igekötő az igenevek esetében válik el, ekkor viselkedik szabad határozóként. Mivel megfigyeléseink szerint a szabályok igekötős, vagy egyéb bővítménnyel rendelkező igenevek esetében működtek jól, ez fenti állításunk közvetett bizonyítékát jelenti. Így ebben a részben már csak egy feladatunk maradt, indokokkal szolgálni arra nézve, hogy miért tekintjük az igekötőket szabad határozóknak. Első pillantásra furcsának tűnhet, hogy miért jogos az igekötőket az ige bővítményei közé sorolni, hiszen az igekötő és az alapige egy lexikai tételt,[8] és ha az igekötő közvetlenül az ige előtt van, akkor egy fonológiai szót is alkotnak. Azonban a *lexikai integritás* elve alapján nincsen olyan szintaktikai szabály, amelynek bemenetéül egy szó részei szolgálnának[4]. Ezzel szemben az igekötők egy mondaton belül viszonylag függetlenül mozoghatnak az igétől, tehát vannak olyan szintaktikai szabályok, amelyeknek a bemenetét igekötők képezik. Ebből következik, hogy az igekötős ige nem lehet összetett szó. Továbbá, az igekötőkhöz disztribúciós szempontból hasonlóan viselkedik a bővítményeknek egy szintaktikailag nem egységes osztálya[9]. Ez arra utal, hogy az igekötőnek vagy vonzatnak kell lennie, vagy szabad határozónak. Most már csak az a kérdés, hogy melyiknek tekintsük őket. Komlósy(1992) szerint ha az igekötő az igének vonzata, akkor – igaz ugyan, hogy egy függvény-szerű kifejezésből függvény-szerű kifejezéseket kapunk – az igekötő maga nem lehet függvény, de azok a kifejezések, amelyek nem függvények, mindig (individuumra, tényállásra) referáló kifejezések kell, hogy legyenek, valamint formailag mindig maximális főkategóriákkal vannak kifejezve.
Az igekötőkre ezek egyike sem áll. Ezek tehát azok a megfontolások, amelyek alapján úgy döntöttünk, hogy

1. Igenévnek tekintünk minden nemcsak vonzattal rendelkező megfelelő formájú kifejezést, hanem azokat is, amelyek környezetében csak szabad határozó van jelen.

2. Szabad határozónak tekintjük az igekötőket is, így a csak igekötővel rendelkező formák is a szabályok bemenetét képezik.

Így alátámasztottnak tekintjük kiinduló hipotézisünket, mely szerint csak igekötővel vagy egyéb bővítményekkel rendelkező kifejezéseket tekintünk igeneveknek és azokat, amelyek környezetében ezek egyike sem fordul elő, mellékneveknek. Mivel

[8] Az igekötő-ige egység együtt képezi szóképzés bemenetét.

[9] '**Pirosra** *festi a kerítést*'; '*Péter* **ügyesen** *vezeti a labdát*'; '*Péter* **okosnak** *tartja Marit*'; '*Péter* **úszni** *akar*'. Ezeknek az eltérő szófajú szavaknak egy része az ige vonzata, egy másik része pedig szabad határozója.

eredeti célunk az volt, hogy kiszűrjük az informatív szerkezeteket, azt kell megvizsgálnunk, hogy a szintaktikai krtitériumok által elkülönített két csoport hogyan állítható párhizamba az informatív – nem informatív csoporttal. Azt látjuk, hogy az általunk informatívnak tartott szerkezetek egybeesnek a fenti szintaktikai kritériumokkal definiált igeneves szerkezetekkel.
A következő pontban a szabályokat fogjuk részletesen bemutatni.

4. A nyelvtan

Az NP-n belüli melléknévi igeneves szerkezetek transzformációs szabályainak kialakításakor az alábbi alapfeltételezésekkel éltünk:

(a) melléknévi igenevet tárgyas és tárgyatlan igéből is lehet képezni,

(b) tárgyatlan ige esetén az NP fejét alkotó főnév a melléknévi igenév alapigéjének alanya,

(c) tárgyas ige esetén az NP fejét alkotó főnév a melléknévi igenév alapigéjének tárgya; ebben az esetben az alapige ágens alanyú,

(d) a melléknévi igenév előtt megjelenhetnek az alapige vonzatai és szabad határozói (ragos NP-k, főnévi igenév, melléknévi csoport, határozószók stb.),

valamint – bár nem feltételezhetjük, hogy minden, igenevet tartalmazó NP elején áll determináns – a kezelni kívánt főnévi csoportok körét leszűkítettük a determinánssal kezdődő NP-kre. Erre azért volt szükség, mert a melléknévi igenév előtt megjelenő, igétől örökölt vonzatok igen sokfélék lehetnek, így determináns nélkül rendkívül nehéz lenne az igenevet tartalmazó főnévi csoport bal szélét pontosan meghatározni (a szerkezeti homonímia gyakorisága miatt ez világismeret nélkül gyakran lehetetlen). Így azonban feltételezhetjük, hogy minden, a determináns és az igenév között megjelenő elem az igenév alapigéjének bővítménye, míg az NP fejét képviselő főnév saját bővítményei az igenév mögött találhatók. Például az " a*kulcsfontosságúnak* tekintett *német* eladásoknak" főnévi csoportban a *kulcsfontosságúnak* az igenév, a *német* a főnévi fej módosítója.

A fenti általánosítások alapján tehát először két csoportot különítettünk el: a tárgyas és a nem tárgyas igékből képzett igenevet tartalmazó NP-ket. A transzformációt végző lokális nyelvtanok olyan szótárra támaszkodnak, melyben kódolva van az ige tárgyas ill. tárgyatlan volta[10](tárgyasnak tekintettünk minden olyan igét, melynek lehet tárgyas előfordulása).

Tárgyas igék

A tárgyas alapigéből képzett igenevek átalakításához használt szabály alapja az alábbi transzformáció:

[10] A szótár kialakításához, azaz a szintaktikai viselkedést kódoló jegyekhez a Korpusznyelvészeti Osztályon készült igei vonzatkeret-adatbázist használtuk.

`Det (V_bőv) VMIB N → Valaki V_vmib Det N -t (V_bőv).`

Ahol *Det*: az NP determinánsa, *V_bőv*: az alapige bővítményei, *VMIB*: az igenév, *N*: az NP feje, *V_vmib*: az alapige, a zárójel pedig opcionalitást jelent. Ilyen átalakításra példa:

(3) *a garéi hulladéklerakó ügyében benyújtott keresetét*
particip Valaki benyújtott a kereset -t a garéi hulladéklerakó ügyében. particip

Az alapige argumentumszerkezetét tehát úgy töltjük fel, hogy a főnévi csoport fejét tekintjük tárgynak, az alanyt pedig – ami az esetek többségében nem jelenik meg a szerkezetben – „valaki" névmással töltjük ki, mivel tudjuk, hogy ágens szerepű. Természetesen van olyan eset, amikor az alany megjelenik az „által" névutóval az igei bővítmények szokásos helyén. Az ilyen szerkezeteket az alábbi szabállyal alakítjuk át:

`Det Nsubj által (V_bőv) VMIB N → Nsubj  V_vmib N -t (V_bőv).`

Például:

(4) *a bankok által felszámított túl magas hitelkamatok*
[particip bankok felszámított túl magas hitelkamatok -t . particip]

Az alapige alanya nemcsak az „által" névutós szerkezetben jelenhet meg az igenév előtt, hanem alanyesetben is, méghozzá az igenevet tartalamzó főnévi csoport fejének birtokosaként. A birtokos megjelenése önmagában nem cáfolja feltevésünket, mely szerint az igenév előtt megjelenő elemek az alapige bővítményei, hiszen a birtokost többnyire jogosan emeljük alanyi pozícióba:

(5) *a svéd Networks tervezett adósságátalakítási programjában*
[particip svéd Networks tervezett a adósságátalakítási programja –t. particip][11]

Tárgyatlan igék

Tárgyatlannak azokat az igéket tekintettük, melyeknek az igei vonzatkeret-adatbázisban egyetlen tárgyas argumentumszerkezete sem szerepel. A tárgyatlan alapigék argumentumszerkezetének meghatározása nem jelent problémát: az NP feje a tárgyatlan ige alanyával azonos, a többi bővítmény pedig – a tárgyas igéknél látottakhoz hasonlóan – az igenév előtt áll. Érdekes, hogy a rövidhír-korpuszban szereplő, tárgyatlan alapigéből képzett igenevek alapigéje mindig páciens alanyú[12].

[11] Sajnos akadnak olyan esetek is, amikor csak a világismeretünk segítségével dönthetjük el, hogy az NP fejének birtokosa azonos-e az alapige alanyával:

a cseh Komercni Banka meghirdetett 60 százalékára
[particip cseh komercni banka meghirdetett 60 százaléka-t particip]

[12] Ez nem jelenti azt, hogy más szövegben sincsenek ágens alanyú tárgyatlan igéből képzett igenevek, pl. '*a társaság lemondott elnöke*'.

Nagyrészt keletkezést, illetve állapotváltozást jelentő igéket találunk köztük. A tárgyatlan igéből képzett igeneveket az alábbi szabállyal alakítjuk át:

`Det (V_bőv) VMIB N` → `DET N V_vmib (V_bőv).`

Például:

(6) *A kereskedés utolsó perceiben bekövetkezett áremelkedés*
particip A áremelkedés bekövetkezett kereskedés utolsó perceiben. particip

Mint a fenti példából is látható, a tárgyatlan igék argumentumszerkezete maradéktalanul kitölthető az igeneves szerkezet elemeivel. Az információkinyerés szempontjából azonban ezek a transzformációk kevésbé hasznosak, kevesebb implicit információt fejtenek ki, mivel az igenevek olyan igékből származnak, melyek szemantikailag kevéssé tartalmasak: *'bekövetkezett', 'beindult', 'létrejött', 'kialakult', 'megszületett'* – így valószínűleg argumentumaik azonosítása sem nyújt többletinformációt. Ennek ellenére érdemes lehet foglalkozni velük, mivel legalább a már ismert események közti összefüggések feltárásában segíthetnek.

5. Értékelés

A szabályok helyes működésének ellenőrzésére kétféle lehetőség kínálkozik. Egyrészt vizsgálhatjuk az igeneves szerkezetek felismerésének arányát (recall) és a kimenet helyességét (precision). Ezt a folyamatot sajnos részben sem tudtuk automatizálni, mert a tesztkorpusz rendelkezésünkre álló kézzel annotált változatában a melléknévi igenevek nincsenek megkülönböztetve a melléknevektől. Másrészt tesztelhetjük azt is, hogy a modul használata mennyivel növeli a sikeresen illesztett szemantikai minták számát. Az értékelés első lépéseként kézzel ellenőríztük a tesztszövegen kapott találatok egy részét. Ebben a részben a típushibákat mutatjuk be.

Az ellenőrzéshez összesen 7058 mondatot (a teljes korpusz 43%-át) vizsgáltunk meg. A tesztkorpuszban a rövidhírek téma szerint sorrendezve szerepelnek, ezért az ellenőrzött korpuszt úgy állítottuk össze, hogy a teljes korpuszból véletlenszerűen 15, egyenként körülbelül 500 mondatból álló részletet vágtunk ki.

Az alábbi típushibákkal találkoztunk:

1. Helytelen morfológiai elemzés, azaz szótárhiba okozta a hiányok túlnyomó többségét.

2. A nem determinánssal kezdődő NP-ket - amint azt a Bevezetésben is említettük - nem tudjuk kezelni. Szerencsére azonban az informatív (és egyben hosszabb) szerkezetek többsége tartalmaz determinánst.

3. A számneves kifejezéseket (mint például a dátum, pénzes kifejezések, mennyiségjelölők)a szabályok jelen állapotában nem kezeljük tökéletesen. E hiány korrigálására a későbbiekben teszünk kísérletet.

4. A szöveg jellegéből fakadóan sok találatban szerepelnek szokatlan NP-k (márkanevet, illetve cégnevet tartalmazó, N N szerkezetű NP-k), melyek felismerése néha problémát okoz.

5. Egyes lexikalizálódott igenevek, bát tartalmazhatnak igekötőt vagy egyéb bővítményt, inkább melléknévként értelmezendők (pl.: '*elmúlt, ismert*').

Az általunk készített modul a NewsPro rendszer hatékonyságát hivatott növelni, így ennek fényében érdemes a működését értékelni. A fent felsorolt hibák elsősorban a találati arányt rontják, viszont a találati pontosság a nyelvészeti megalapozottság miatt kielégítő. Ez utóbbit fontosabbnak tartjuk, mivel az információkinyerésben a helyes kimenet létrehozása az elsődleges, hiszen a pontatlan találat félrevezetőbb a felhasználó számára, mint a találat hiánya.

Hivatkozások

[1] Prószéky G.: Automatikus információszerzés gazdasági rövidhírekből. In: Alexin Zoltán - Csendes Dóra (szerk.): A Magyar Számítógépes Nyelvészeti Konferencia 2003 rendezvényen elhangzott előadások kötete, Szegedi Tudományegyetem Nyomdája, 2003. Szeged, 161-167.o.

[2] Silberztein,M.: Dictionnaires électroniques et analyse automatique de textes: Le systeme Intex. Masson, 1993. Paris

[3] Komlósy A.: Régensek és vonzatok. In: Strukturális magyar nyelvtan I. Akadémiai Kiadó, 1992. Budapest, 299-529.o.

[4] É. Kiss K.: Mondattan. In: Új magyar nyelvtan. Osiris, 1999. Budapest 17-184.o.