Technológiai fejlesztések a NooJ pszichológiai alkalmazásában

Vincze Orsolya[1], Gábor Kata[2], Ehmann Bea[3], László János[4]

[1] PTE Pszichológia Intézet
orsolyavincze@hotmail.com
[2] MTA Nyelvtudományi Intézet
gkata@nytud.hu
[3] MTA Pszichológia Intézet
ehmann@mtapi.hu
[4]MTA Pszichológia Intézet
laszlo@mtapi.hu

Kivonat: A NooJ nyelvi fejlesztő környezete egy jól kezelhető, dinamikus felületet nyújt az automatizált narratív pszichológiai szövegelemzésben. Az előadás több éves pszichológiai módszertani fejlesztés legújabb eredményeit kívánja bemutatni, különös tekintettel a NooJ nyelvi fejlesztő környezetében kialakított protézisnyelvtanra [1], amely a pszichológiailag releváns kifejezéseket (mentális állapotok, aktív-passzív igék, közelítést-távolítást jelző igék…stb) szemantikai és nyelvtani szerepük alapján összekapcsolja. Ezt megelőzően a nyers szöveg nyelvi elemzését a MorphoLogic Moose szintaktikai elemzőprogramja [2] végzi, ami előkészíti a protézisnyelvtan számára a szövegeket: a szöveget bekezdésekre, mondatokra, tokenekre bontja, elvégzi a szavak morfológiai elemzését, valamint nem csupán beazonosítja az NP és VP csoportokat, de össze is illeszti őket. Kiosztja a nyelvtani szerepeket a főnévi csoportokra és a tematikus szerepeket a vonzatokra. Ez utóbbi esetben a tematikus szerepek kiosztásához a Moose rendszer vonzatkeret-leíró formalizmusát kibővítettük *theta* jeggyel.

1 Bevezetés

A PTE Pszichológia Intézet és az MTA Pszichológiai Intézet kutatóiból álló narratív kutatócsoport hazai és külföldi nyelvtudományi, informatikai és pszichológiai kutatócsoportokkal együttműködve az elmúlt öt évben jelentős nemzetközi áttöréssel járó kutató-fejlesztő munkát végzett. A kutatások eredményeként megszületett és nemzetközi elfogadást nyert a tudományos narratív pszichológia. Az új tudományos paradigma lényege, hogy az emberek természetes közegben zajló, hétköznapi viselkedéséből és kommunikációjából tudományos eszközökkel képes személységükre, lelki állapotaikra és társas beállítódásaikra vonatkozó következtetéseket levonni. Ez úgy történik, hogy a személyes élettörténeti eseményekre, illetve a társadalmi csoportok, például a nemzetek történetére vonatkozó elbeszélések nyelvi és kompozíciós tulajdonságait tudományos eszközökkel megfeleltetjük az identitásképzés pszichológiai

folyamatainak. A nyelvi mintákat nyelvtechnológiai eszközökkel számítógépes programokká fejlesztjük, és ezekkel a programokkal elemezzük a természetes szövegeket. Ez képessé tesz arra, hogy a lelki állapotokról és tartós beállítódásokról diagnosztikus és a társas alkalmazkodás különböző formáit előre jelző eredményeket kapjunk. A tudományos narratív pszichológia fogalmai és eljárásai, amellett, hogy a személyiség és a társas élet pszichológiai folyamatainak komplex megközelítését teszik lehetővé, különösen előnyösnek bizonyultak olyan problémák vizsgálatában, ahol jelen idejű kutatásokra nincs lehetőség, például történeti szövegek esetében, illetve ahol a kérdőíves vagy teszteljárások alkalmazásának lehetősége behatárolt, például addiktológiai betegek esetében. Az alkalmazási lehetőségek köre kiterjed az űrkutatás területére is, mivel a narratív pszichológiai diagnosztikus eljárások alkalmasnak tűnnek a hosszabb űrutazáson részvevő személyek pszichológiai állapotának monitorozására is.

Jelen dolgozat célja, hogy áttekintést nyújtson az automatikus narratív pszichológiai eljárás újabb technikai fejlesztéseiről.

2 Narratív pszichológiai modulok

A kutatócsoportunk által kidolgozott automatikus tartalomelemző eljárás pszichológiailag releváns nyelvi változók köré csoportosuló modulokba rendeződik, mint például az aktivitás-passzivitás [3], érzelem [4], kognitív [5], értékelés [6], intencionalitás [7], idői modulok [8], pszichológiai perspektíva [9].

A pszichológiai modulok több almodulból tevődnek össze, amelyek az elemzés szintjén a pszichológiai jelentés és a technikai kivitelezés tekintetében is különböző komplexitásúak. Ugyanakkor a tartalomelemző algoritmusok működése bizonyos tekintetben azonos: szó- és mondatszintű elemzést végeznek. Ezeken belül azonban eltérések mutatkozhatnak az egyes modulok között a tekintetben, hogy milyen morfológiai vagy szintaktikai megszorításokat alkalmaznak.

2.1 NooJ nyelvi fejlesztő környezet alkalmazása az automatikus pszichológiai tartalomelemzésben

Az egyes modulok automatikus tartalomelemző algoritmusai a NooJ nyelvi fejlesztő környezetében kerültek kidolgozásra [10], ami dinamikus felületet biztosít, lehetővé téve a szoftver biztonságos és rugalmas kezelését nem nyelvészek számára is.

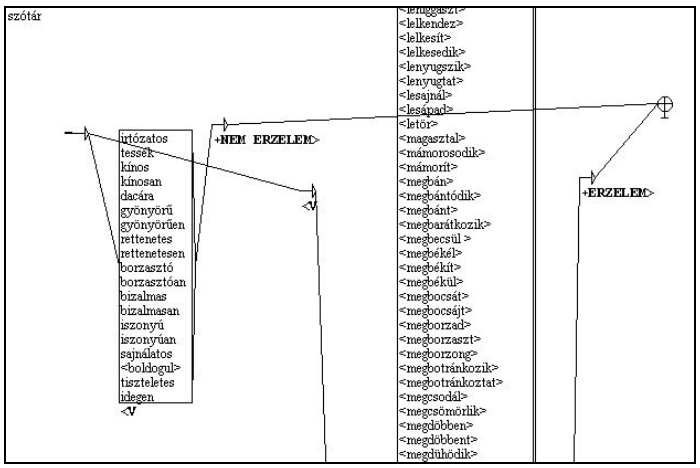
A szoftver központi eleme a szótár, aminek szókincsét egyfelől a magyar írott nyelv általános szókincsét reprezentáló szövegkorpuszokból (Magyar Nemzeti Szövegtár [11], Szeged Korpusz [12]), másfelől specifikus pszichológiai szövegekből álló korpuszból nyertük ki. Ez utóbbiban megtalálhatóak klinikai pszichológiai populációkkal (depressziós, borderline, droghasználó, krízisben lévő betegekkel) készített mélyinterjúk, többgenerációs traumatizált családinterjúk, normál populációkkal (teljesítmény-, veszteség-, párkapcsolati interjúk) felvett féligstruktúrált interjúk, valamint nemzeti és etnikai vonatkozású szövegkorpuszok. Az általános korpuszokból a magyar nyelvben használatos gyakori szóalakok morfoszintaktikailag elemzett formái

kerültek be az általunk használt szótárba, amit a speciális pszichológiai szövegkorpusz gyakran előforduló szavaival egészítettünk ki.

A szoftver motorja véges állapotú technológián alapul, grafikus felülete lehetővé teszi a nyelvtanok gráfként való megjelenítését és szerkesztését. Ezáltal olyan környezetet biztosít, melyben egységesen kezelhetők a nyelvi elemzés különböző szintjei (inflexiós és derivációs morfológia, szintaktikai elemző és transzformációs szabályok). Az automatizált narratív pszichológiai elemzés megközelítésében a gráfokban megjelenő lokális nyelvtanok olyan algoritmusoknak tekinthetők, amelyek pszichológiailag releváns kifejezések beazonosítását végzik.

Ennek megfelelően első lépésben minden modul esetében megtörtént az adott modul tematikájába illeszkedő szavak szótári leválogatása a Magyar Nemzeti Szövegtár leggyakoribb 10 000 igéje, határozói és névutói alapján[1]. Bizonyos modulok esetében további jelentésteli dimenziók is bevezetésre kerültek, mint csoportosító változók: például az érzelmi állapotok „pszichológiai annotációja” [4] során, a valencia mellett, a primer és a társas érzelmek elkülönítése is csoportosító szempontként jelent meg.

A legtöbb modul esetében a gráfok két típusba sorolhatóak: szólistás és szintaktikai gráfok. Mivel az automatizált pszichológiai tartalomelemzés gyakorisági adatokkal dolgozik, a szólistás gráfok készítése értelmes technológiai eljárásnak bizonyul. Ilyenkor a gráfban csupán az adott pszichológiai jelentéskategóriába illeszkedő szavak listája kerül be, minimális szintaktikai megszorítással vagy anélkül (1. ábra).

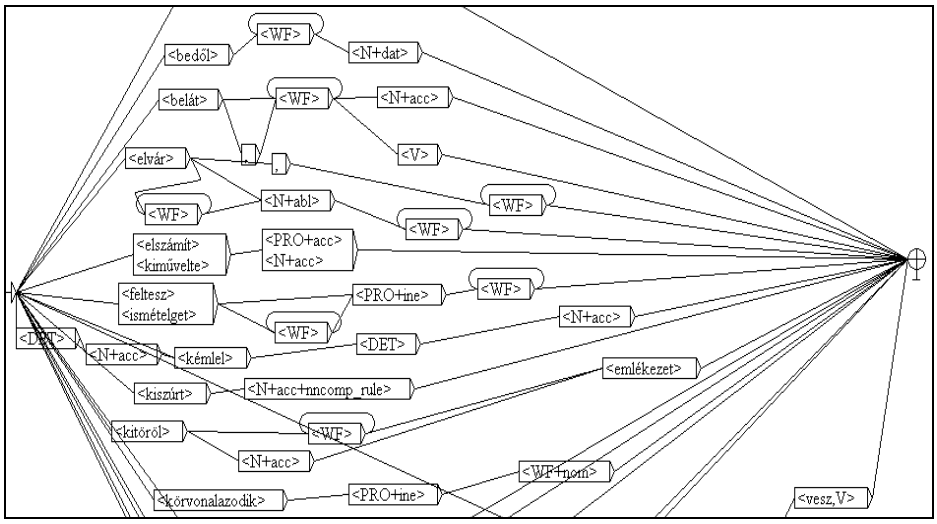


1. ábra. Érzelem modul algráfja.

A szintaktikai gráfok készítése azoknál a kifejezéseknél fordul elő, ahol szintaktikai függőségi viszonyok figyelembevétele szükséges az adott pszichológiai jelentés megragadásához. Például a „bedől” ige csak abban az esetben minősül kognitív kife-

[1] A szótárak fejlesztését a PTE Pszichológia Intézet és az MTA Pszichológiai Kutatóintézet, a szótárak morfológiai annotációját az Szegedi Tudományegyetm és az MTA Nyelvtudományi Intézete végezte.

jezésnek, ha részeshatározós esetű főnév követi <N+dat>. Míg a „belát” igénél az azt követő főnév tárgyas vonzata a feltétel <N+acc> (2. ábra)



2. ábra. Szintaktikai szekvenciára épülő elemző algoritmus.

A lokális nyelvtanok találatait a program konkordanciába rendezi, az adott találat kívánt számú karakterkörnyezetével együtt. Mivel a lokális nyelvtanok lezárása egy szemantikai indexszel történik, a program arra is lehetőséget biztosít, hogy a találatokat szemantikai indexükkel együtt a teljes szövegkörnyezetben lássuk (3. ábra).

```
A nemzet és a császár kibékítésében nagy része volt Erzsébet császárnénak is.
Az ifjú fejedelemasszony <V POZITÍVÉRZELEM">együtt érzett az</V> elnyomott
magyar nemzettel. <V NEGATÍVÉRZELEM"">Fájt</V> neki a nemzet <V
NEGATÍVÉRZELEM"><N>szenvedése</N></V> és <V POZITÍVÉRZELEM""><N>öröme</N> telt
benne, ha javíthatott a sorsán. Állandóan <V "KOGNITIV"><V>érdeklődött</V>
a</V> viszonyai iránt, <V "KOGNITIV">megtanulta</V> nyelvét, történelmét és
szószólója volt császári férje oldalán. A császár <V „KOGNITIV">ismerte e</V>
<N>vonzalmát</N> és felesége kedvéért sokszor gyakorolt kegyelmet, hogy <V
POZITÍVÉRZELEM""><N>örömet szerezzen</N> a császárnénak. De a nemzet is <V
„KOGNITIV">tudta, hogy Erzsébet császárné a pártfogója s azért nagy <V
POZITÍVÉRZELEM""><N>örömmel</N> fogadta a hírt, hogy az országgyűlésre ő is
lejön császári urával.
```

3. ábra. Szemantikai indexek megjelenítése a szövegben.

2 Technikai fejlesztések

A modulok technikai fejlesztését több tényező is lehetővé tette. A Szegedi Tudományegyetemnek köszönhetően az elemzések alapjául szolgáló szótár szemantikai adatbázis információval bővült. Az MTA Nyelvtudományi Intézetben elkészült a nyelvtani, valamint a tematikus szerepek beazonosítására szolgáló lokális nyelvtan, amihez a szövegeinket a MorphoLogic Moose szintaktikai elemzőprogramja [11] készíti elő.

2.1 A szótár szemantikai bővítése

Az alapszótárban a főnevek pszichológiailag releváns szemantikai jegyekkel bővültek. A Szegedi Tudományegyetem által elkészített főnévi adatbázis 20788 főnévi lemmához társít szemantikai információt, melyek különböző szociális kapcsolatokat (rokon, egyéb társadalmi kapcsolat, szűk családi kapcsolat), csoportok jellegét (etnikai, vallási) és egyéb, a tartalomelemzés szempontjai szerint releváns jellemzőket kódolnak (1. táblázat).

1. táblázat: Szemantikai jegyek példája.

szó	Ember	nem	foglalkozás	kapcsolat	csoport	etnikai
betörő	X	xy				
házasságtörő	x	xy	x	x		
jégtörő						
szentségtörő	x	xy				
kitörő						

2.2 Tematikus szerepek beazonosítása

Bármilyen jellegű pszichológiai szövegelemzésben elengedhetetlenül fontos a nyelvtani és a tematikus szerepek beazonosítása. Mivel erre egyenlőre a NooJ szoftver nem képes, egy segédprogram beiktatása vált szükségessé.

A Moose szintaktikai elemzőprogram a nyers szöveg nyelvi elemzése során a szöveget bekezdésekre, mondatokra és tokenekre bontja, elvégzi a szavak morfológiai elemzését, valamint beazonosítja a főnévi (NP) és igei (VP) csoportokat. Az igei csoportok beazonosításánál a program a vonzatkeret-adatbázis segítségével az igéhez sorolható vonzat és szabad határozó NP-ket is beazonosítja.

A tematikus szerepek kiosztásához a MetaMorpho rendszer vonzatkeret-leíró formalizmusát kibővítettük egy új jeggyel (*theta*). A theta jegy a vonzathoz rendelt meghatározott tematikus szerep. Lévén, hogy a pszichológia tartalomelemzésben a tematikus szerepek azonosítása különösen fontos az értelmezés szempontjából, ezért minden modul esetében kiválogattuk a vonzatos igéket és egyszerű példamondatokon keresztül 2640 vonzatkeret-leírást készítettünk, amelyekkel végül kibővült a MetaMorpho rendszer vonzatkeret-leíró formalizmusa. Az automatikus ellenőrzés és

a felmerült hibák javítását tartalmazó validációs ciklus után összesen jelenleg 2322 tematikus szereppel annotált vonzatkeret áll rendelkezésre a rendszerben (2. táblázat).

2. táblázat: Annotált vonzatkeretek tematikus szerepeloszlásai.

Összes vonzatkeret:	2322
Th-jeggyel annotált vonzat összesen:	3174
AG (ágens) jeggyel annotált vonzat:	1447
PAT (páciens) jeggyel annotált vonzat:	749
EXP (experiens) jeggyel annotált vonzat:	646
STI (stimulus) jeggyel annotált vonzat:	270
BEN (beneficiens) jeggyel annotált vonzat:	55
REC (recipiens) jeggyel annotált vonzat:	5
SRC (forrás) jeggyel annotált vonzat:	1
INS (instrumentum) jeggyel annotált vonzat:	1
GOAL (cél) jeggyel annotált vonzat:	0

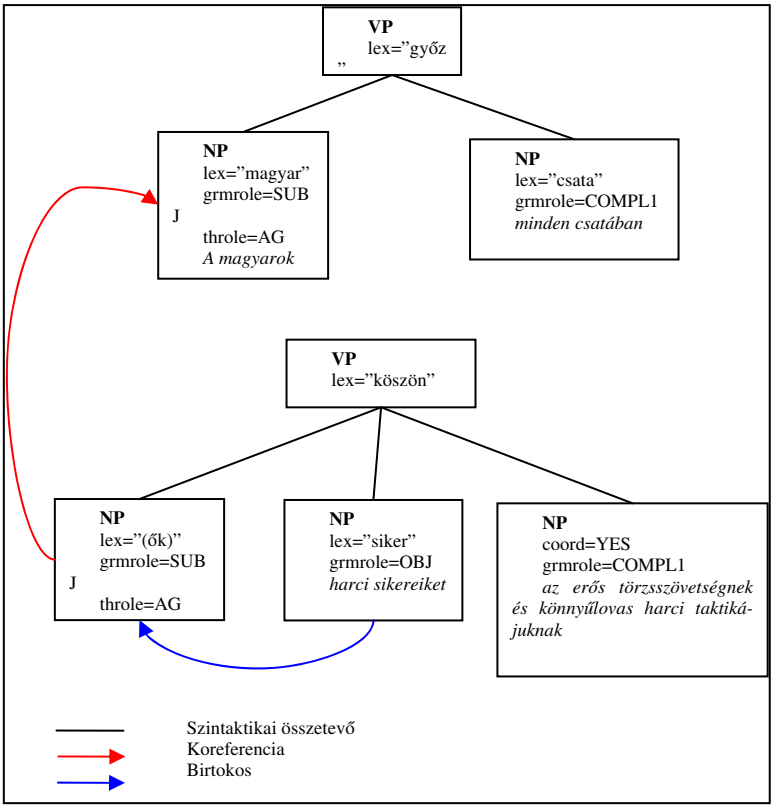
2.3 Szövegbeli utalások feloldása

A szövegekben előforduló utalások természetes jelenségek, ami nem okoz különösebb nehézséget az olvasó számára a szöveg követésében. A tartalomelemzés során az NP-k közötti utalás, azaz amikor a főnévi csoportok egy része nem közvetlenül utal a való világ entitásaira, hanem a szövegben korábban bevezetett ilyen kifejezésre hivatkozik, nem elhanyagolható mennyiségű találati hibát okoz.

A technikai fejlesztések során kétféle, főnévi csoportok közötti utalástípussal foglalkoztunk: a) koreferencia, b) elvált birtokos. Ezek feloldására a Moose szintaktikai elemzőprogram olyan szabályalapú algoritmusokat alkalmaz, amelyek behelyettesítik a hivatkozott kifejezések szótári alakját az utaló kifejezésekbe, ezáltal a NooJ alkalmazásban egyszerű lexikális alakok keresésére nyílik lehetőség.

A Moose szintaktikai elemzőprogram hat különböző NP-koreferencia feloldását végzi el: egyszerű ismétlés, tulajdonnév-variánsok, szinonimák, hipernima, névmási és zérónévmási anafora. Továbbá beazonosítja az összetartozó birtokosoknak és birtokoknak megfelelő kifejezések közötti viszonyokat a szövegben, különös tekintettel azokra az esetekre, ahol a birtokosnak és a birtoknak megfelelő NP-k nem közvetlenül követik egymást.

A nyelvi elemzés során tehát, amit a Moose szintaktikai elemzőprogram végez, megtörténik a nyelvtani és a tematikus szerepek beazonosítása, valamint a hivatkozások feloldása (4. ábra).

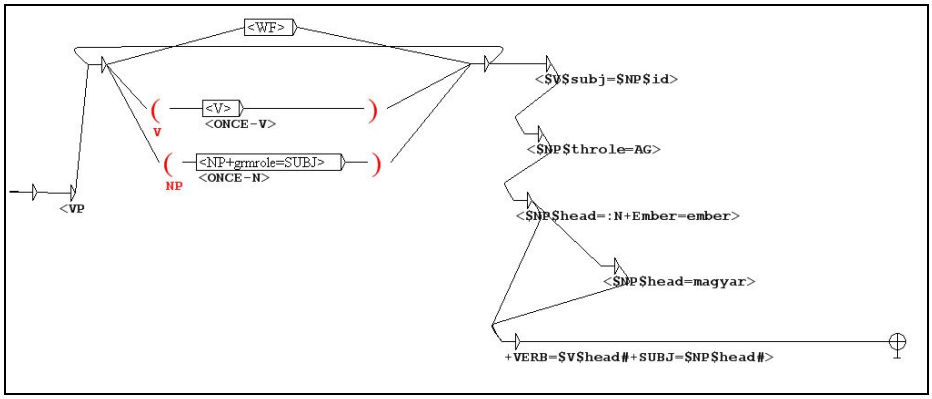


4. ábra. A nyelvi elemzés folyamata.

2.4 Protézisnyelvtan a NooJban

A Moose szintaktikai elemzőprogram által biztosított nyelvtani elemzés a nyers szöveget olyan XML struktúrában jeleníti meg, amiben a dependenciaviszonyokat a szövegszavakhoz társított attribútumok értékei kódolják. Az így előállt szöveg képezi a NooJ bemenetét, ahol a pszichológiai mintázatok beazonosítása történik. Ahhoz, hogy az egyes pszichológiai modulokhoz tartozó korábban kidolgozott lokális nyelvtanok az elemzett mondat szóelemeinek teljes dependenciaviszonyát lefedjék, szükség volt egy ún. *protézisnyelvtan* kidolgozására [1] (5. ábra). A protézisnyelvtan jelentősége, hogy szabad szórendű nyelvekben az összetevők közötti függőségi viszonyok és egyeztetési jelenségek kezelését, illetve a lexikai és a függőségi tulajdonságok szerinti lekérdezést teszi lehetővé. A NooJ-ban ennek technikai hátterét a szoftver új funkciói (a felismert elemek változókban való tárolása, lexikai megszorítások) valósítják meg, melyek így a NooJ-t a véges automatákénál nagyobb leíró kapacitással ruházzák fel.

A protézisnyelvtan lényege, hogy először rekurzívan begyűjti és változókban tárolja a mondat állítmányát és a névszói csoportokat, majd ún. lexikai[2] megkötések segítségével ellenőrzi, hogy ezek rendelkeznek-e bizonyos tulajdonságokkal. A pszichológiai elemzések általános céljával összhangban itt az ige és vonzatai közti szintaktikai és szemantikai viszony beazonosítása történik, azaz a vonzatok grammatikai és tematikus szerepe szerint szűrjük a találatokat.



5. ábra. Protézisnyelvtan.

Az elemzés során a gráf kigyűjti a szöveg mondataiból azokat a találatokat, melyekben az ige alanyi szerepű vonzata ágens tematikus szereppel rendelkezik (5. ábra alapján). Mivel a keresett elemek, vagyis az ige és bővítményei tetszőleges sorrendben követhetik egymást, valamint egyéb elemek is közéjük ékelődhetnek, ezért felismerésükhöz olyan gráfot kell készítenünk, mely egy rekurzív 'hurokban' tartalmazza mind az igét (<V>), mind jelen példában az alanyt (<NP+grmrole=SUBJ>, alanyi szerepű NP), melyek tetszőleges sorrendben követik egymást, és közéjük ékelődve tetszőleges egyéb elemeket (<WF>, word form: tetszőleges szóalak) is megenged. A gráf bal oldali része ezt a hurkot ábrázolja. A tetszőleges szóalakokon (<WF>) kívül a többi felismert elemet piros zárójelekkel jelölt $NP és $V változókban tároljuk, ez teszi lehetővé, hogy a gráf jobb oldalán a lexikai megszorításokban hivatkozhassunk rájuk.

A lexikai megszorítások szerkezete és a rendelkezésre álló jegykészlet
A grammatikai funkció szerinti szűréshez az alábbi jegykészlet használható:

NP+grmrole= COMPL (vonzat), MOD (szabad határozó), OBJ (tárgy), SUBJ (alany), UNKNOWN (egyéb, fel nem ismert)

Nem elég azonban a főnév funkcióját ellenőrizni, külön megszorítással kell megbizonyosodnunk arról is, hogy az adott grammatikai szerepet az adott ige bővítmé-

[2] A 'lexikai' ebben a kontextusban úgy értendő, hogy nem a szövegben, hanem a hozzá tartozó annotációs szerkezetben kódolt információról van szó, ám ez lehet szintaktikai természetű információ is.

nyeként tölti be (vagyis az összetett mondatokban sem keverednek össze a különböző igék bővítménykeretei). Ehhez az XML struktúrában szereplő azonosító (id) attribútumok értéket kell összehasonlítani:

<Vsubj=NPid>	*alany*
<Vobj=NPid>	*tárgy*
<Vcompl1=NPid>	*egyéb bővítmény*

A tematikus szerepek szerinti kereséshez az alábbi jegykészlet áll rendelkezésre:

NP+throle=AG (ágens), PAT (páciens), REC (recipiens), STI (stimulus), EXP (experiens), SRC (forrás), GOAL (cél), INS (eszköz), BEN (beneficiens), UNKNOWN (egyéb, fel nem ismert)

A tematikus szerep annotációját szintén a Moose szintaktikai elemző helyezi el a szövegben, ami az alábbinak megfelelő lekérdezést tesz lehetővé:

<NPthrole=AG>

A találatok tovább szűrhetők lexikai megszorítások hozzáadásával, illetve a pszichológiai modulok kombinálásával. Így például a cselekvő alanyú igék közül kiszűrhetjük azokat, melyeknek alanya egy etnikai csoportot jelölő főnév. Ezeket tovább csoportosíthatjuk az etnikumok szűrésével (pl. magyar cselekvők vs. egyéb népcsoportok). Ennek megfelelően a névszói bővítmény (fejének) szemantikai és/vagy lexikális tulajdonságaira vonatkozó megszorításokat a protézisnyelvtan alábbi csomópontjaiban adhatjuk meg:

szemantikus tulajdonságok:
<NPhead=:N+Ember=ember>
<NPhead=:N+Nem=Y>
<NPhead=:N+etnikai=N>

lexikális tulajdonságok:
<NPhead=magyar>
<NPhead=fejedelem>

2.5 A nyelvtechnológiai változtatások bevezetése a pszichológiai modulokba

Az újonnan alkalmazott Moose szintaktikai elemzőprogram, valamint az erre illeszkedő NooJban kifejlesztett protézisnyelvtan valamennyi, már kifejlesztett pszichológiai modult érintett: szükségesség tette az eddig használt lokális nyelvtanok egy részének átírását. Azokban az esetekben, ahol a pszichológiai modulok lokális nyelvtanai a szólistás algoritmust követik, a protézisnyelvtanban az NP és VP csoportok egyszerű konkretizálással szűkíthetők a pszichológiailag releváns NP és VP csoportokra. Azonban a szintaktikai algoritmust követő lokális nyelvtanokat, amelyek nem

szószintű, hanem szó feletti találatot adnak, nem lehet egy az egyben illeszteni a protézisnyelvtan VP/NP csoportjával. A probléma megoldása különösen lényeges a pszichológiai jelentés megragadása szempontjából, hiszen a találatok nem elhanyagolható részét képezik az ilyen, szintaktikai szekvenciákra épülő jelentések.

Hivatkozások

1. Váradi T, Gábor K.: A magyar Intex fejlesztéséről. In III. Magyar Számítógépes Nyelvészeti Konferencia, Szeged (2004) 3-10
2. Prószéky G., László T., Ugray, G.: Moose: a robust high-performance parser and generator. Proceedings of the 9th Workshop of the European Association for Machine Translation, Foundation for International Studies, La Valletta, Malta (2004) 138-142
3. Szalai K., László J.: Az aktivitás-passzivitás modul kidolgozása NooJ tartalomelemző programmal. IV. Magyar Számítógépes Nyelvészeti Konferencia, Szeged (2006)
4. Fülöp É., és László J.: Az elbeszélések érzelmi aspektusának vizsgálata tartalomelemző programmal. IV. Magyar Számítógépes Nyelvészeti Konferencia, Szeged (2006)
5. Vincze O. és László J.: A mentális igék szótára, valamint alkalmazása az automatikus tartalomelemzésben. IV. Magyar Számítógépes Nyelvészeti Konferencia, Szeged (2006)
6. Bigazzi S., Csertő I., Nencini, A.: A személy- és csoportközi értekelés pszicholingvisztikája. IV. Magyar Számítógépes Nyelvészeti Konferencia, Szeged (2006)
7. Ferenczhalmy R., László J.: Az intencionalitás modul kidolgozása NooJ tartalomelemző programmal. IV. Magyar Számítógépes Nyelvészeti Konferencia, Szeged (2006)
8. Ehmann B., Garami V., Szabó J.,: NooJ fejlesztések a szubjektív időélmény tartalomelemzéses vizsgálatára. IV. Magyar Számítógépes Nyelvészeti Konferencia, Szeged (2006)
9. Pólya, T., Ferenczhalmy R., Fülöp É., Vincze O.: A pszichológiai perspektíva előfordulása történelem tankönyvi szövegekben V. Magyar Számítógépes Nyelvészeti Konferencia, Szeged (2007)
10. Silberstein, M.: NooJ manual. Paris:Université de Franche-Comté (2005)
11. Váradi, T.: The Hungarian National Corpus. In Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC-2002), Las Palmas de Gran Canaria, (2002) 385-389
12. Csendes D., Alexin Z., Csirik J., Kocsor A.: A Szeged Korpusz és Treebank verzióinak története. IV. Magyar Számítógépes Nyelvészeti Konferencia, Szeged (2005), 409-412