Argumentumstruktúrák gépi azonosítása (Szemantikai modul a Hunpars elemzőhöz)

Babarczy Anna[1], Gábor Bálint[1], Hamp Gábor[2], és Rung András[3]

[1] Kognitív Tudományi Tanszék, BME, 1111 Budapest, Stoczek u. 2.
{babarczy, bgabor}@cogsci.bme.hu

[2] Szociológia és Kommunikáció Tanszék, BME, 1111 Budapest, Stoczek u. 2.
hampg@eik.bme.hu

[3] Nyelvtudományi Intézet, MTA-ELTE, 1068 Budapest, Benczúr Gy. u. 33.
runga@artitude.hu

Kivonat: A Hunpars projekt folytatásaként a III. MSzNy Konferencián bemutatott mondattani elemző alkalmazásunkat szemantikai modullal egészítjük ki A fejlesztés elsődleges célja tagmondatszintű szemantikai tudások beemelésére, a frázisstruktúra tematikai címkézése. Erre a célra egy strukturált vonzatkerettárat fejlesztünk. Bár első lépésben az igék argumentumszerkezetére helyezzük a hangsúlyt, a fejlesztés olyan általános elméleti alapokon nyugszik, melyekkel bármely predikátumfunkciót betöltő nyelvi elem kezelhető. Az argumentumszerkezetek leírásában a Role and Reference Grammar fogalomrendszerét ötvözzük a FrameNet projekt módszereivel és a Konstrukciós nyelvtan egyszintű, lexikalista filozófiájával.

1 Bevezetés

A Hunpars-projekt folytatásaként a III. MSzNy Konferencián bemutatott mondattani elemző alkalmazásunkat [1] szemantikai modullal egészítettük ki. A Hunpars jelenleg implementált moduljai automatikusan végzik bármilyen értelmezhető magyar mondat szintaktikai elemzését. Az elemző a frázisstruktúra nyelvtanok alapelveinek felhasználásával a mondat szavait hierarchikus szerkezetekbe, frázisokba szervezi, és szintaktikai jegyekkel felcímkézett, zárójelezett mondatszerkezetet ad kimenetként.

A projekt második szakaszában célunk elsősorban tagmondatszintű szemantikai tudások beemelése, a frázisstruktúra tematikai címkézése. Tematikai címkézés alatt egy olyan rendszert értünk, melyben a mondatban szereplő predikátumokhoz tartozó főnévi, illetve határozói frázisokat tagmondatbeli szerepüket jelölő címkékkel látunk el. A fejlesztés jelenlegi szakaszában a hangsúly nem annyira az elemzés technikai részletein, mint inkább az erőforrásként használt tematikai nyelvtan kidolgozásán van. A projekt olyan általános elméleti alapokon nyugszik, melyekkel bármely predikátumfunkciót betöltő nyelvi elem kezelhető, bár első lépésben az igék keretrendszerének kidolgozása a cél. A generatív nyelvelméletben elterjedt, a tematikus szerepek fogalmi keretére épülő megközelítést a klasszikus funkcionalista igetipológiával [2,

3], a konstrukciós nyelvtan lexikalista elveivel [4] és ezek korpusznyelvészeti alkalmazásaival [5] társítjuk.

2 A tematikai elemző-modul szerkezete

Az elemző elsődleges erőforrása a vonzatkerettár. A vonzatkerettár szerkezeti felépítése az 1. ábrán bemutatott példán látható. A tár alapeleme a Frame, melyet egy meghatározott vonzatkeret definiál. A vonzatkeret meghatározásában a morfoszintaktikai és a tematikai jegyek azonos súllyal szerepelnek. Egy-egy Frame egy vagy több, a vonzatkeretébe illeszthető nyelvi elemet, vagy konstrukciót, foglal magába – ezeket lexikális tételnek nevezzük. Egy lexikális tétel állhat egyetlen szóból, de lehet szónál kisebb elem (például igekötő, képző) vagy többszavas kifejezés is (idióma, kollokáció). A tematikai elemzés a Frame és a zárójelezett, morfoszintaktikailag annotált mondat illesztéséből áll.

FRAME	**Ösztönöz**	
	LEXICAL ENTRIES	ösztönöz bátorít buzdít biztat unszol sarkall
	ACTOR	ösztönző <CAS<NOM>>
	UNDERGOER	ösztönzött <CAS<ACC>>
	NON-MACROROLE	goal <CAS<SBL>>; <CLAUSE<SUBJ-IMP>> manner <CASE<INS>>

[*A vállalatvezetők*]$_{\text{ACTOR}}$ [a magasabb profit érdekében] [minden évben] [*prémiummal*]$_{\text{MANNER}}$ [***ösztönzik***] [*jobb munkára*]$_{\text{GOAL}}$ [*a dolgozókat*]$_{\text{UNDERGOER}}$

PERIPHERY	location <CAS<INE;…>> source <postpp/ÉRDEKÉBEN;…> (…)

1. **ábra** Frame és periféria a vonzatkerettárban

2.1 A tematikus szerepek

A Frame-et definiáló vonzatkeret leírása a Van Valin nevéhez fűződő Role and Reference Grammar (RRG) [3] fogalmaira épít. A RRG megkülönböztet két kitüntetett szerepű argumentumot, azaz két makroszerepet: az ACTOR-t és az UNDERGOER-t. A két tematikai funkciót a vonzatkeret formális logikai szerkezete alapján határozzuk meg: informálisan fogalmazva, az actor az esemény aktív szereplője, míg az undergoer a viszonylag passzív résztvevő. Bár egy tipikus tranzitív szerkezetben az actor az alany, az undergoer pedig a tárgy szintaktikai funkcióhoz rendelhető, ettől a mintától eltérő vonzatkeretek is előfordulhatnak (pl. az állapotot vagy visszaható eseményt kifejező egy-argumentumú igék alanya UNDERGOER funkciót tölt be, míg a személytelen igék dativus argumentuma bizonyos esetekben ACTOR szerepet kap).

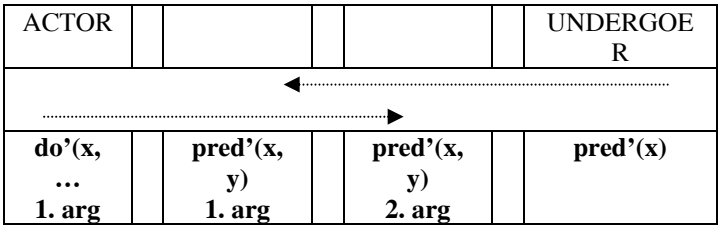
A predikátum logikai szerkezetét az Aktionsart típusa határozza meg. A Hunpars szempontjából lényeges Aktionsart kategóriák és a hozzájuk rendelt intranzitív és tranzitív logikai struktúrák az (1) táblázatban láthatók. A táblázatot követő példák a táblázat sorait illusztrálják sorrendben.

1. Táblázat: Logikai szerkezetek Aktionsart típus szerint

Aktionsart	Logikai szerkezet
állapot	**pred'**(x) vagy (x, y)
atelikus cselekvés	**do'**(x, [**pred'**(x) vagy (x, y)]
állapot változás	BECOME **pred'**(x) vagy (x, y)
telikus cselekvés	**do'**(x, [**pred'**(x, (y))]) & BECOME **pred'**(z, x) vagy (y)

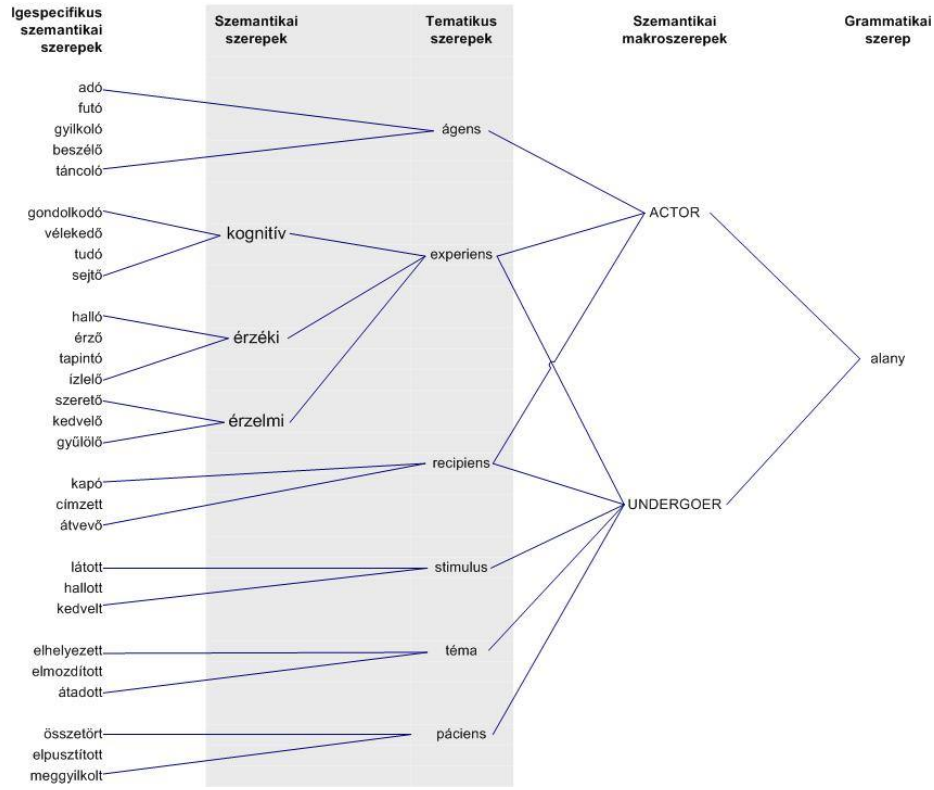
1. Vilmos álmos. Vilmos ágyban van.
2. Edit szánkózik. Feri almát hámoz.
3. Vilmos felébredt. Edit a lejtő aljára ért.
4. Edit visszaliftezett a pálya tetejére. Feri kenyeret süt.

A két makroszerep kiosztása az actor-undergoer hierarchia szerint valósul meg, amint a 2. ábrán látható:



2. ábra. Az ACTOR és az UNDERGOER makroszerepek kiosztását vezérlő hierarchia. Forrás: [9], 177. o.

A logikai szerkezetben előforduló egyéb, nem-makroszerepű argumentumokat tematikai címkékkel jellemezzük.. Döntő kérdésnek tekintettük, hogy hol kell meghúzni a tematikai kategóriák határait. A Hunpars célja a kategóriák számának minimalizálása: új címkét csak akkor vezetünk be, ha ez szükséges ahhoz, hogy kifejezhető legyen egy-egy predikátum argumentumai közötti kontraszt. Ezzel a döntéssel a nehezen definiálható, sokszor „megérzéseken" alapuló tematikai osztályok alkalmazásából adódó kategorizációs problémákat igyekeztünk elkerülni, illetve azokat minimálisra csökkenteni. (Lásd 3. ábra)



3. ábra. A szemantikai szerepek hierachiáját mutatja a [8] (31. o.) alapján készült ábra, amelyen a szürkével jelölt tartomány nem jelenik meg Hunpars-mondatelemzésben.

Annak, hogy az elemzés csak kevés szerepcímkét használ, van néhány következménye. Például a *vár* ige tárgyesetű és -rA ragos bővítménye azonos szerepcímkét kap, mert mindkettő nem fordulhat elő egy argumentum szerkezetben. A -tÓl ragos argumentumot ezzel szemben megkülönböztetjük, mivel szerepelhet a tárggyal együtt:

5. Mit$_{[\text{UNDERGOER}]}$ vármy?
6. Mire$_{[\text{UNDERGOER}]}$ várs?
7. *Mire mit várs?
8. Mit$_{[\text{UNDERGOER}]}$ várs tőle$_{[\text{SOURCE}]}$?

Bár a *vár* ige esetében a két UNDERGOER címkével jelölt argumentum kétségtelenül hasonló tematikai funkciót tölt be a mondatokban, a tárgyesetű és valamilyen oblique esetű alternatívák azonos címkézése nem szemantikai, hanem disztribúciós alapon történik. A *tud* ige tárgy és oblique argumentumai éppen ezért két különböző tematikai címkét kapnak:

9. Mit$_{[\text{UNDERGOER}]}$ tudsz?
10. Miről$_{[\text{THEME}]}$ tudsz?
11. Miről$_{[\text{THEME}]}$ mit$_{[\text{UNDERGOER}]}$ tudsz?

Ezek alapján az ACTOR és az UNDERGOER mellett hatféle tematikus szerepet különböztetünk meg: CO-ACTOR, MANNER, SOURCE, LOCATION, GOAL, THEME, és ezek mindegyikének egy absztrakt és egy konkrét típusát, ami szükség illetve lehetőség szerint elválasztja például a téri elhelyezkedést (konkrét) az időbeli elhelyezkedéstől (absztrakt), vagy az eszközhasználatot (konkrét) a módtól (absztrakt):

12. Lassan$_{[\text{MANNER/ABSTRACT}]}$, nagy gonddal$_{[\text{MANNER/ABSTRACT}]}$ írt.
13. Lassan$_{[\text{MANNER/ABSTRACT}]}$ írt a tollal$_{[\text{MANNER/CONCRETE}]}$.
14. Lassan$_{[\text{MANNER/ABSTRACT}]}$ írt nagy gonddal$_{[*\text{MANNER/CONCRETE}]}$.
15. Nagy gonddal$_{[\text{MANNER}]}$ írt.

A fenti példákban a vonzatkeretnek megfelelően a módhatározót absztrakt módként címkézzük. Ha ez mellérendelő viszonyban áll egy –vAl esetű NP-vel, az utóbbi is absztrakt címkét kap, hiszen nagyvalószínűséggel azonos szerepű argumentumokat koordinálunk (4). Ha a két argumentum nincs mellérendelő viszonyban, a –vAl ragos főnevet eszköznek tekinthetjük, bár ez ritka esetben téves elemzéshez vezethet (6). Magában álló –vAl esetű NP semleges MANNER címkét kap (7), itt csak a szavak vagy a szövegkörnyezet részletes szemantikai elemzése dönthetne a két értelmezés között.

A fenti nyolc tematikus szerepet kiegészíti egy alacsonyszintű argumentum-leírás. Az *ösztönöz*-keretben az actor talán triviálisan az *ösztönző* leírást kapja, az undergoer pedig az *ösztönzött* lesz attól függetlenül, hogy a Frame melyik igéje szerepel a mondatban. Ennek jelentőségét a későbbiekben tárgyaljuk.

Ennyiben az elemző lexikális alapú. A magyar NooJ tematikai moduljához [6] hasonlóan, a lexikális megkötéseket egy default mapping rendszer egészíti ki, amely a magyar nyelv sajátosságainak megfelelően elsősorban esetragokra, névutókra és határozó típusokra épül, de bármilyen más formai feltételrendszer kifejezésére is alkalmas. A rendszer beépül a tematikai modul Frame szerkezetébe: a default mintát egy predikátum-független supra-lexikális Frame-nek tekintjük. A Frame típusok magyarázatát ld. alább. A default címkék szerepe kettős: egyrészt a vonzatkerettárban nem szereplő predikátumokat tartalmazó mondatok elemzését teszik lehetővé, másrészt a szabad határozók definiálásának kritériumát adják és ezek annotálását végzik. Szabad határozónak tekintünk minden olyan NP-t és határozót, amely az esemény logikai szerkezetén kívül esik, azaz amelynek a morfológiai vagy szintaktikai jegyei egy olyan default tematikai szerepet határoznak meg, amely bármilyen mondatba funkcióváltozás nélkül beilleszthető. Ezek a szereplők/körülmények a perifériába tartoznak, és nem játszanak szerepet a Frame definiálásában.A szabad határozók és

argumentumok megkülönböztetésében James Pustejovsky [7] definícióját követjük: eszerint valódi szabad határozónak csak olyan mondatrészek tekinthetők, amelyek szabadon előfordulhatnak bármilyen mondatszerkezetben. Ide tartoznak tehát az eseményt időben és térben elhelyező határozók és az esemény okát megnevező határozók. Bármilyen más bővítmény argumentumnak tekintendő, attól függetlenül, hogy a megnevezése kötelező-e, hiszen csak a mondat adott predikátuma engedheti az előfordulását, vagyis beépül a predikátum logikai szerkezetébe.

2.2 A Frame és lexikális tételek

Ez természetesen nem kell hogy azt jelentse, hogy minden cselekvést kifejező ige vonzatkeretébe külön-külön vesszük fel például a módhatározó argumentumot. Egy-egy Frame leírása, a konstrukciós nyelvtan filozófiáját követve, bármilyen kötöttségi szinten megvalósítható. Lehet a Frame viszonylag szűken definiált, mint például az 1. ábrán látható *ösztönöz*-keret, amelybe csak néhány (de legalább egy) lexikális tétel illeszthető, de alkothatunk tágan definiált, alulspecifikált ige-osztályokra épülő kereteket is, és sublexikális morfémákat jellemző kereteket is. Az előbbire példa a cselekvést jelentő vagy a mozgást jelentő ige-keret, ahol a vonzatkeret meghatározása csak olyan jegyeket tartalmaz, amelyek minden cselekvést illetve mozgást jelentő igét jellemeznek. Az utóbbi típusba az argumentumstruktúra szempontjából meghatározó igekötők és képzők keretei tartoznak. Mivel egy keretbe olyan lexikális tételek illeszthetők, amelyek közös tematikai szerepekkel és közös morfo-szintaktikai szerkezettel rendelkeznek, és a tematikai szerepek maximális hatáskörűek, a keret-tagság lehetséges létszámát a morfoszintaktikai megkötések szigorítása vagy lazítása vezérli. Ennek megfelelően a morfoszintaktikai feltétel rendszer kötetlen: a pontos szóalaktól kezdve a legtágabb szófaji kategóriáig minden szinten specifikálhatjuk a Frame vonzatainak formai követelményeit.

Egy lexikális tétel lehet egy vagy több Frame instanciája is. Ha egynél több kerethez tartozik, a tematikai elemzést a keretek összevonása adja, a következő szabályok szerint. (1) A Frame-eknek három típusa van: lehetnek alapszintűek, bővítők vagy módosítók. (2) Alapkeretnek tekintjük a legszűkebben definiált lexikális Frame-eket, vagyis azokat, melyek lexikális tételei megengednek azonos alacsonyszintű argumentum-leírásokat, azaz szinonimáknak tekinthetők. Egy lexikális tétel legfeljebb egy alapkeret instanciája lehet. (3) Az alapszintnek meg nem felelő Frame-ek operátor keretek, amelyek bővítik és/vagy módosítják az alapkeret adott vonzatstruktúráit. Az operátorkeretek nem határoznak meg önálló alacsonyszintű argumentum-leírásokat. (4) Az igeosztályokat leíró Frame-ek és a szupralexikális default Frame bővítő operátor keretek. Ha egy lexikális tétel egy alapkeret és egy vagy több bővítő keret instanciája is, a tematikai elemzést a keretek unifikációja határozza meg, azaz a bővítő keretben talált új szerepű argumentumokkal kiegészítjük az alapkeret által meghatározott bővítményeket. Az alábbi (leegyszerűsített) példa az *ír* ige keretének az összetételét mutatja: az első sorban meghatározott végső Frame az ige alapkeretének (16a), egy transzfer igeosztály bővítőkeretének (16b) és a cselekvés igeosztály bővítőkeretének (16c) unifikációjából áll össze.

16. ír{ír} [ACTOR/író<NOM>, UNDERGOER/szöveg<ACC>, GOAL<DAT>, MANNER<INS>]
  a. ír{ír} [ACTOR/író<NOM>, UNDERGOER/szöveg<ACC>]
  b. benefit{olvas, ír, …} [ACTOR/<NOM>, UNDERGOER<ACC>, GOAL<DAT>]

c. cselekvés{eszik, ír, …} [ACTOR/<NOM>, UNDERGOER<ACC>, MANNER<INS >]

A vonzatkerettár felépítése lehetővé teszi a gyors fejlesztést: besorolhatunk lexikális tételeket igeosztály Frame-ekbe, anélkül, hogy alapkereteket határoznánk meg. Ilyen esetben a bővítő keret(ek)ből jön létre a predikátum Frame-je. A viszonylag széles eseménykörben előforduló vonzatok egyszerű feltérképezése és a vonzatkerettár gyors fejlesztése mellett a bővítő operátor keretek a ritka vagy szokatlan vonzatszerkezetek azonosításában is szerepet játszanak. A *megkérdez* ige alapszintű Frame-je például a <CAS<DEL>> (*-rÓl*) morfoszintaktikai jegyet rendeli az esemény THEME argumentumához, és nem illeszkedik az alábbi nem-kanonikus mondatra:

17. A mellékhatások tekintetében kérdezze meg orvosát, gyógyszerészét.

Az ige azonban lehet instanciája a tudás-transfer igeosztály bővítő keretének, ahol a THEME morfoszintaktikai specifikációi megengedőbbek, tartalmazzák a <tekintetében>, <vonatkozólag>, stb. névutós szerkezeteket. A többszintű Frame-alkotás egyrészt gazdaságossági okokra vezethető vissza, másrészt azt a pszicholingvisztikai folyamatot próbálja implementálni, miszerint egy lexikális konstrukció rögzített szerkezete bizonyos körülmények között közeledhet a lexikális tétel szemantikai szomszédainak szerkezeti felépítéséhez, illetve analógiás úton kölcsönözheti ezek egyes elemeit.

(5) Az operátor keretek másik típusa egyben bővítő és módosító operátor. Az itt leírt argumentumok nemcsak új bővítményekkel egészíthetik ki az eredeti keretben megadott bővítményeket, hanem módosíthatják is azokat. Ezek a keretek alkalmasak az igekötők és a képzők vonzatkeretre gyakorolt hatásának leírására: egy ilyen sublexikális keret instanciája lehet egy vagy több sublexikális morféma (képző vagy igekötő). Ebben az esetben önmagukban nem alkotnak végleges Frame-t, hanem az elemzés alatt álló mondat predikátumának töve által előhívott alapkerettel és/vagy osztálykeret(ek)kel együttesen határozzák meg az elemzés vonzatstruktúráját. A sublexikális operátor keretben meghatározott argumentumok felülírják a lexikális keret azonos tematikai szerepű argumentumainak morfoszintaktikai specifikációit:

18. fürdet{fürdet, mosdat, ...}[ACTOR<NOM>, UNDERGOER/fürdő<ACC>]
   a. fürdik{fürdik, mosdik, …}[UNDERGOER /fürdő<NOM>]
   b. tat{VERB<CAUS>}[ACTOR<NOM>, UNDERGOER<ACC>]

(6) Amennyiben konfliktus merül fel az alkalmazható Frame-típusok között, vagyis több illesztési lehetőség is van, a „nyertes” Frame-kombináció az lesz, amelyik az elemzés alatt álló mondat legnagyobb számú argumentum frázisára illeszthető.

3 Összefoglalva

A Hunpars-elemzőben alkalmazott tematikai címkézés rugalmasnak tekinthető. Érveltünk amellett, hogy ez a rugalmasság rendszerelméleti szinten is józan próbálkozásnak látszik. Emellett azonban praktikus megfontolássokkal is lehet indokolni a Hunparsban alkalmzott megoldást. Míg az automatikus szemantikai elemzés precizi-

tásának növelése munka- és időigényes szűken definiált kereteket kíván, a használható lefedettség eléréséhez észszerű egy gyorsabb fejlesztési folyamat lehetőségét is megteremteni. Ennek érdekében terveink között szerepel az igék automatikus keretbesorolása morfoszintaktikailag elemzett korpuszból kinyert statisztikai minták alapján.

Bibliográfia

1. Babarczy, A., Gábor, B., Hamp, G., Kárpáti, A., Rung, A., Szakadát, I.: Hunpars: mondattani elemző alkalmazás. In: III. Magyar Számítógépes Nyelvészeti Konferencia. SZTE, Szeged (2005) 20–28.
6. Gábor Kata, Héja Enikő: Vonzatok és szabad határozók szabályalapú kezelése. In: III. Magyar Számítógépes Nyelvészeti Konferencia. SZTE, Szeged (2005) 245-256.
4. Goldberg, Adele**:** Constructions. A Construction Grammar approach to argument structure. Chicago: University of Chicago Press (1995)
7. Pustejovsky, James: The generative lexicon. Cambridge, MA: MIT Press (1996)
8. Van Valin, Robert: An introduction to syntax. Cambridge: CUP (2001)
9. Van Valin, Rober & LaPolla, Randy: Syntax. Structure, meaning and function. Cambridge: CUP (1997)
2. Vendler, Zeno: Linguistics in philosophy. Ithaca: Cornell Univ. Press (1967)
3. http://linguistics.buffalo.edu/research/rrg.html
5. http://framenet.icsi.berkeley.edu