Komplex (magyar) szintaktikai elemző rendszer mint beágyazott rendszer

Kis Balázs, Naszódi Mátyás, Prószéky Gábor

MorphoLogic
{kis,naszodim,proszeky}@morphologic.hu

Az előadás az 1995 óta fejlődő HumorESK mondatelemző rendszer fejlődését, új lehetőségeit és alkalmazásait mutatja be. A fejlesztők a HumorESK architektúráját úgy alakították ki, hogy a rendszer a legkülönbözőbb alkalmazások beágyazott komponense lehessen. A program teljesen adatvezérelt, ami azt jelenti, hogy a feldolgozandó nyelvet, a nyelvtant és az elemzés mélységét illetően semmilyen előfeltételezéssel nem él: ennek minden paraméterét a felhasználó határozhatja meg. Az előadás során a szerzők vázlatosan ismertetik azokat az alkalmazásokat és projekteket, amelyekben eddig felhasználták a HumorESK mondatelemző rendszert.

1. A HumorESK program elméleti alapja

A HumorESK program az – ugyancsak a MorphoLogic által kifejlesztett – MetaMorpho (MMO) nyelvtani formalizmus első implementációja. Az előadás 1. része ezért többnyire általában is érvényes a MetaMorpho-formalizmusra, ugyanakkor a rendszer fejlődése során e formalizmus olyan elemekkel is kiegszül, illetve kiegészült, amelyek kimondottan a HumorESK-implementációra jellemzőek.

A HumorESK program nem valósít meg forradalmian új elemzési algoritmust. Alulról felfelé végzi a szöveg elemzését; az egyes szimbólumokhoz egyszerűsített struktúrájú jegyszerkezeteket (*feature structure*) kapcsol, és az elemzési erdő építése során az itt feltüntetett jegyek értékeit ellenőrzi, illetve örökölteti. A HumorESK működésének logikája leginkább a PATR-II formalizmusnak felel meg (Shieber, Uszkoreit, Pereira, Robinson, Tyson 1983).

A HumorESK megvalósításában a szabályillesztés módja és a szabályok megfogalmazása számít újdonságnak. A szabályokat véges mintahalmaz formájában írjuk le, így a rendszer a szabályok illesztéséhez a morfológiai elemzőkéhez hasonló lexikont kap. Az egyes minták alulspecifikált elemekből – szimbólumokból – épülnek fel: egyes szimbólumok esetén csak a szintaktikai szerepet jelző címkét ismerjük, mások esetében meg van adva a lemma vagy éppen a felszíni szóalak is.

A fentiek miatt a HumorESK-et nem lehet egyértelműen a szabályalapú vagy a szótáras – a gépi fordítástól kölcsönzött kifejezéssel: példaalapú – rendszernek nevezni: az elemzési adat, vagyis a nyelvtan elemi komponense egyfelől olyan szabály, amelynek egyes elemeit lexikailag megszorítjuk; ugyanez másfelől olyan minta (példa), amelynek egyes részei nincsenek teljesen, a felszíni jelsorozat szintjén specifi-

kálva. Ez feltételezésünk szerint lehetővé teszi, hogy a rendszer elméletben bármelyik meglevő nyelvtani formalizmussal ekvivalens legyen, különösen ha a rendszerbe bekerül a sorrendfüggetlen szabályillesztés mechanizmusa is.

A nyelvtanban lehetnek ugyanarra a nyelvi jelenségre általánosabb és specifikusabb minták is. Ezért a HumorESK által alkalmazott formalizmusnak fontos eleme a szabályok közötti felülbírálati mechanizmus; a specifikusabb szabály jellemzően felülbírálja az általánosabbat. Ezzel egyfelől csökken a rendszer túlgenerálása, másfelől pedig a kapott elemzésekben az egyes lexikai elemekre (terminális szimbólumokra, „szavakra") vetített elemzések jobban megfelelnek a szimbólumok környezetének. A HumorESK így tulajdonképpen sokszor anélkül is meghatározza az egyes szavak környezetnek megfelelő „jelentését", ha maga jelentés semmilyen formában nincs reprezentálva a nyelvtanban. Példa:

```
NP=ADJ+N:155261
HU.NP[ ... ] = ADJ(...) + N(...)

NE=ANY+NE(nev):16772
HU.NE[...] = ANY( casetype=UPPERINITIAL ) + N( prop = FIRSTNAME )
!155261
```

A fenti esetben mindkét minta illeszkedik a mondatkezdő „Fekete Péter" karaktersorozatra. A második azonban specifikusabb: ez abból látszik, hogy az ott leírt N szimbólumnak rendelkeznie kell a prop tulajdonsággal, annak pedig a FIRSTNAME értékkel. Ez azt jelent(het)i, hogy a jelzett helyen olyan főnévnek kell szerepelnie, amelyet egy korábbi minta vagy éppen a morfológiai elemző modul személynévként azonosított. A második minta kiegészül a `!155261` sorral, ami azt jelenti, hogy amennyiben a minta „elsül", illeszkedik egy bemeneti jelsorozatra, akkor ennek a mintának felül kell bírálnia a 155261 azonosítójú másik mintát, amennyiben az is illeszkedett ugyanarra a bemenetre.

A címkék diverzitása, illetve a jegyszerkezetek kihasználásának mértéke a nyelvtanban tetszés szerint választható meg. A HumorESK architektúrájától nem idegen a mintáknak az elemek felszíni sorrendjétől független illesztése sem – ez a funkció jelenleg fejlesztés alatt áll –, ennek megvalósítása esetén a szimbólumok nemprojektív módon is származtathatók.

Fontos lehetőség a HumorESK-ben, illetve a MetaMorpho-formalizmusban, hogy a szabályokat leképező mintákhoz transzformációk rendelhetők. Így a minták alkalmas megfogalmazása esetén a rendszer már elemzési időben logikai struktúrává „fordíthatja" a szöveget.

2. A HumorESK megvalósításának lényeges vonásai

A fejlesztők a HumorESK architektúráját úgy alakították ki, hogy a rendszer a legkülönbözőbb alkalmazások beágyazott komponense lehessen. A program teljesen adatvezérelt, ami azt jelenti, hogy a feldolgozandó nyelvet, a nyelvtant és az elemzés mélységét illetően semmilyen előfeltételezéssel nem él: ennek minden paraméterét a felhasználó határozhatja meg.

Így lehetőség van teljes mondatok vagy mondatfeletti struktúrák nagy mélységű elemzésére is; ezzel például részletes tartalomelemző alkalmazások működtethetők – a 3. részben ilyen alkalmazást is bemutatunk. Ugyanakkor olyan nyelvtant is készíthetünk, amellyel a HumorESK csak egyes részstruktúrák kis mélységű elemzését végzi el, így alkalmas például NP-kivonatolásra (*NP chunking*), illetve általános kollokációkeresésre is.

Az elemzés mélysége és a bemeneti szegmensek lefedése egyazon nyelvtannal is lehet különböző: a HumorESK-ben a nyelvtan szintekre bontható, a konfigurációban pedig előírható, hogy az elemzés mely szintig történjen meg.

A HumorESK valós idejű rendszer; az elemzési idő korlátozható, s akkor is kiolvashatók hasznos elemzési eredmények, ha az algoritmus logikája szerint még nem fejeződött be a szegmens eredménye.

A HumorESK implementációja statikus programkönyvtárként, C- és C++-illesztőfelülettel áll rendelkezésre. Jelenleg a 32-bites Windows alatti megvalósítás érhető el; a Unix/Linux-rendszerekben használható változat fejlesztés alatt áll.

A mondatelemző program a cikk írása idején az alkalmazásokra jellemző kétféle mélységű magyar nyelvtannal működik. A mélyebb elemzést előíró nyelvtan kb. 20 000 mintát tartalmaz. Ezzel a nyelvtannal egy szegmens (mondatjelölt) elemzése, hibakereső üzemmódban, átlagos PC-n 10-300 ms időt vesz igénybe, az átlagos elemzési idő 50 ms alatt van olyan szövegekben, ahol egy mondatjelölt jellemzően 20 szónál hosszabb.

3. A HumorESK alkalmazásai

3.1. Üzleti rövidhírek tartalomelemzése

2003 közepén zárult le egy NKFP-projekt, amelynek célja üzleti rövidhírek tartalomelemzése volt. Ez olyan alkalmazás – a *NewsPro* – készítését jelentette, amelynek elemeznie kell a rövidhírek mondatait, s a mondatelemzés eredményeire olyan szemantikai kereteket kell illesztenie, amelyek lehetővé teszik az egyes mondatok által leírt események, illetve az események szereplőinek azonosítását. Ha például egy bank megnöveli tulajdonrészét egy cégben, akkor ezt – a tulajdonrész meglétét és növekedését – a rendszernek megfelelően azonosítania kell mint eseményt, és fel kell ismernie, hogy az esemény szereplői között megjelenik a bank mint vevő (és tulajdonos), a cég mint az adásvétel (és a tulajdonlás) tárgya, az eredeti tulajdonrész (ha meg volt adva), és a növekedés mértéke.

Ebben a rendszerben a HumorESK mondatelemző végzi a rövidhírek mondatainak elemzését. Ehhez meglehetősen bonyolult szerkezeteket lefedő, viszonylag nagy mélységű elemzést adó magyar mondatnyelvtant kellett készíteni. Ez a mondatnyelvtan három lényeges komponenst tartalmaz, amelyek külön-külön is jelentős fejlesztést igényeltek, és általában is nagy mértékben járultak hozzá a magyar számítógépes szintaxis fejlődéséhez:

(1) *Tulajdonnév-felismerés.* Az üzleti rövidhírek nagy mennyiségben tartalmaznak személy-, cég-, intézmény-, helyneveket, dátum- és időmeghatározásokat, illetve

pénzösszegeket. Ezek felismerésére kiterjedt résznyelvtan készült, amelyet az MTA Nyelvtudományi Intézetének Korpusznyelvészeti Osztályán készítettek elő a Clark programmal, s a MorphoLogic munkatársai ezt követően adaptáltak a HumorESK számára. Ez a tulajdonnév-felismerő rendszer – annak ellenére, hogy a NewsPro-projekt tesztkorpusza alapján készült – általánosan is használható, s más projektekben, sőt részben más nyelvű nyelvtanokban is megjelenhet tulajdonnév-felismerő modulként.

(2) *Főnévi csoportok felismerése.* Az üzleti hírek szövegeinek megfelelő bonyolult főnévicsoport-nyelvtant kellett készíteni, amely más jellegű szövegekben nem kívánt módon túlgenerálhat, ezért általános szövegekhez még adaptálni kell. Ezzel kapcsolatban viszont jelentős eredmény, hogy a NewsPro-rendszerhez készített nyelvtan a magyar főnévi csoportokban megjelenő legtöbb jelenséget lefedi (beágyazott melléknévi igenévi szerkezetek, különféle birtokos szerkezetek, értelmező jelzők stb.), s mint ilyen, a magyar főnévi csoportok eddigi legteljesebb számítógépes leírása. Elméleti szempontból viszont nem egységes, mivel pragmatikus szempontok szerint, egy igen koncentrált korpusz által reprezentált nyelvváltozat leírására szolgál.

(3) *Igevonzatok felismerése.* A NewsPro-rendszerhez készített HumorESK-nyelvtan az MTA Nyelvtudományi Intézetének Korpusznyelvészeti Osztálya által készített igevonzat-szótár adaptált változatát alkalmazza. Több mint 8000 igevonzatot, illetve egyes esetekben ezek általánosított változatát tartalmazza. Az eddigi magyar számítógépes szintaxisok közül e tekintetben is a legteljesebbnek számít.

3.2. Narratív pszichológiai tartalomelemzés

Egy másik NKFP-projektum keretében a MorphoLogic az MTA Pszichológiai Intézetével közösen pszichológiai narratívumok – interjúk során rögzített, az alanyok által elmondott történetek – elemzését végzi. A nyelvi elemzés feladata itt viszonylag korlátozott: úgynevezett parciális vagy lokális nyelvtanok segítségével meghatározott nyelvi markereket (pszichológiai szempontból jelentős, a narratívumban megjelenő nyelvi jelenségeket) kell felismerni. Itt nincs szükség teljes mondatok mély elemzésére: elegendő meghatározott nyelvi jelenségek jelenlétét észlelni, és a jelenségeket a szövegekben megjelölni. A megjelölt markerek alapján pszichológusok statisztikát készítenek, s ezeket használják fel további kutatási célokra (lásd László et al. 2003, elhangzik ugyanezen a konferencián).

Az elemzés során a HumorESK-nek a következő markertípusokat kell felismernie:

(1) idő és idővel kapcsolatos megnyilvánulások

(2) a közelítés és távolítás kifejezései,

(3) a narratív perspektíva kifejezései.

A megfelelő nyelvtanok kifejlesztése során különösen nagy problémát jelentett az időt kifejező határozók és határozói szerkezetek felismerése; erről ugyanezen a konferencián külön előadás szól (Naszódi 2003).

Ebben a projektben a HumorESK mondatelemző modult egy LinTag nevű programba ágyaztuk, amely az elemzési eredményeket olyan formába alakítja, amely az Atlas.ti nevű statisztikai programcsomagban használható fel.

3.3. Korpuszstatisztikai eszköztár

A HumorESK felhasználásával olyan korpuszstatisztikai eszköztár is készült, amely különösen alkalmas típusos kollokációk keresésére. A korpuszstatisztikai eszköztár oly módon von ki adatokat akár annotálatlan korpuszból is, hogy az az NSP nevű statisztikai programcsomaggal legyen feldolgozható (Pedersen 2003). A korpuszstatisztikai eszköztárt ugyancsak bemutatjuk ezen a konferencián (Kis-Ugray 2003).

A HumorESK modult ezúttal egy parancssori eszközbe (mlc_dataset) építettük, amely a korpuszbeli mondatként szegmentált szakaszok elemzését végzi el, s az eredményekből úgynevezett kivonatolási metaszabályok segítségével kikeresi a releváns származtatott szimbólumokat.

A korpuszstatisztikai eszköztárral végzett kísérleteinkhez a 3.1. alatt említett NewsPro-projekthez készített hét logikai szintből álló magyar mondatnyelvtan alsó három szintjét használtuk fel (az egyszerű főnévi csoportokkal bezárólag). Kihasználtuk a HumorESK azon lehetőségét, hogy az elemzés maximális szintjét futásidőben meg lehet határozni: így, bár a kivonatoló futtatásához eredendően komplex nyelvtant használtunk fel, a felső négy szint kikapcsolása révén a rendszer nem használt fel a szükségesnél több erőforrást – sem processzoridőt, sem tárolóhelyet.

4. Összefoglalás

Ez az előadás a HumorESK mondatelemző modul legújabb alkalmazásait mutatta be, bizonyítva, hogy az eredetileg 1995-ben felvázolt mondatelemzési modell és annak implementációja alkalmas a széles körű felhasználásra.

Köszönetnyilvánítás

A HumorESK alkalmazásainak és a nyelvtanok elkészítéséért köszönet illeti a következő kollégákat: Tihanyi László, MorphoLogic (MetaMorpho-formalizmus), Váradi Tamás és kollégái, MTA NyTI (tulajdonnév-felismerés, névszói csoportok, igevonzatok struktúrájának meghatározása), Benkő Borbála Katalin és Katona Tamás, BME HIT (a NewsPro-mondatnyelvtan szerkesztői és implementálói), Gyimóthy Tibor, Alexin Zoltán és kollégáik, SZTE (a NewsPro szemantikai kereteinek kidolgozása), László János, Ehmann Bea, Pólya Tibor, Pohárnok Melinda, MTA PI (a narratív pszichológiai tartalomelemzés elvi kidolgozása és az elemzési eredmények továbbfeldolgozása), Gosse Bouma és Begoña Villada Moirón, Humanities Computing, Rijksuniversiteit Groningen (az NSP statisztikai programcsomag adaptálása, a korpuszstatisztikai eszköztár kimenetének értékelése).

Irodalomjegyzék

KIS Balázs (1997): Mi van a szavakon túl? Nyelvtani szerkezetek felismerése számítógéppel. *Előadás a VII. Országos Alkalmazott Nyelvészeti Konferencián.* Külkereskedelmi Főiskola, Budapest, 1997

LÁSZLÓ János–EHMANN Bea (2004): Narratív pszichológia és narratív pszichológiai tartalomelemzés (kéziratban). In (várható): Magyar Pszichológiai Szemle, 2004/2., Budapest.

NASZÓDI Mátyás: Nyelvhelyesség-ellenőrzés számítógéppel (parciális szintaxis). Elhangzott a *VII. Országos Alkalmazott Nyelvészeti Konferencián* (Külkereskedelmi Főiskola, Budapest, 1997)

PEDERSEN–BANERJEE (2003): The Design, Implementation and Use of the Ngram Statistics Package. In: *Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics* (Mexico City).

PRÓSZÉKY Gábor–KIS Balázs (1999): *Számítógéppel - emberi nyelven.* SZAK Kiadó, Bicske.

PRÓSZÉKY, Gábor (1996): Syntax As Meta-morphology. *Proceedings of COLING-96*, Vol.2, 1123-1126. Copenhagen, Denmark.

PRÓSZÉKY, Gábor (1999): Lexical Information and Decisions in Parsing. In: Cristea, Dan, Dan Tufiş, Amalia Todiraşcu, Valentin Tablan & Cătălina Barbu (eds.) *4th Eurolan Summer School on Human Language Technology, Technical Report 99-02*, ISSN 1224-9327, Iaşi, Romania.

SHIEBER, S. M., H. Uszkoreit, F. C. Pereira, J. Robinson, and M. Tyson (1983). The formalism and implementation of PATR-II. In *J. Bresnan, editor, 23 Research on Interactive Acquisition and Use of Knowledge.* SRI International, Artificial Intelligence Center, Menlo Park, Cal.