ReALIS projekt: a szóképzés általánosítása a számítógépes fordításban

Alberti Gábor[1], Kleiber Judit[1], Ohnmacht Magdolna[2],
Szilágyi Éva[1], Anne Tamm[3], Viszket Anita[1]

[1] Pécsi Tudományegyetem, Bölcsészettudományi Kar, Nyelvtudományi Tanszék,
7624 Pécs, Ifjúság útja 6.
[2] Szegedi Tudományegyetem, Bölcsészettudományi Kar, Nyelvtudományi Doktori Iskola,
6722 Szeged, Egyetem u. 2.
[3] Università degli Studi di Firenze, Dipartimento di Filologia moderna,
Via Santa Reparata 93-95, 50129 Firenze
gelexi@btk.pte.hu

Kivonat: A ReALIS projekt célja egy olyan nyelvelemző program megalkotása, amely minden eddiginél alaposabb szemantikai reprezentációt társít szövegekhez, és azt mint nyelvfüggetlen közvetítőnyelvet alkalmazva gépi fordító rendszerként is hasznosítható. Az elemző elméleti hátteréül a totálisan lexikalista GASG szolgál [3], a szemantikai reprezentációt pedig mint a ReALIS [1] implementációját képzeljük el. A totálisan lexikalista morfológia kiterjesztéseként az inflexiós morfémák mellett a produktív *derivációs morfémákhoz* is közvetlenül lexikai egységeket rendelünk, a szóalaktanilag „láthatatlan" *konverziós* [17] eseteket is figyelembe véve. A nyelvenkénti sokszínűség a nyelvfüggetlen szemantikai reprezentációs szinten eltűnik, így nem igényel bonyolultabb mechanizmust a nagyon különböző nyelvek közötti fordítás sem. A lexikalista elméletekre épülő elemzők és gépi fordító rendszerek egyre nagyobb térhódítása igazolni látszik, hogy szükség van a kifinomult nyelvelméleti eszköztárra a számítógépes nyelvészet olyan igényes területein, mint például a gépi fordítás. Projektünk alapvető célja a végsőkig vitt lexikalizmus kipróbálása ezen a területen.

1 Bevezetés

Mint a szerzők névsora mutatja, a *ReALIS* projekt kutatócsoportja a *GeLexi* (*Ge*nerative *Lexi*con) elméleti és számítógépes nyelvészeti kutatócsoport [3] és a *LiLe* (*Li*nguistic *Le*xicon) adatbázis-készítő team [6] egyesülése révén jött létre, további elméleti nyelvészek bevonásával. Alapvető célunk változatlanul a *kifinomult nyelvelméleti eszköztár* hasznosságának igazolása a számítógépes nyelvészet olyan igényes területein, mint amilyen például a *gépi fordítás* [4].

Azon az úton haladunk tovább, melyet a *totális lexikalizmus* elvének a morfémák szintjén való érvényesítése és egy korszerű DRT alapú szemantikához [16] való közvetlen (azaz nem egy hagyományos generatív szintaxison keresztül történő) hozzáférés fémjelez. A ReALIS (*Reciprocal And Lifelong Interpretation System*) [1] mint

háttér a DRT szemantikának egy olyan pragmatikai kiterjesztésére utal, amelynek alapján „interpretálói tudásbázist / információállapotot” építünk fel a parser támogatására.

Egy konkrétumot szeretnénk ebben a tanulmányban kiemelni. Az inflexiós morfémák mellett a produktív *derivációs morfémákhoz* is lexikai egységeket rendelünk, a szóalaktanilag „láthatatlan” *konverziós* [17] eseteket is figyelembe véve. A cikk második felében a lexikalista alapú gépi fordítás jelenlegi állásának áttekintése után projektünk megközelítését ismertetjük.

2 Elméleti háttér

A nyelvelméletben az utóbbi évtizedekben lexikalista fordulat következett be: a kezdeti szintaxis-központú elméleteket egyre inkább felváltják a szótári komponenst előtérbe helyező megközelítések. Ennek hatására egyre több lexikalista elmélet és számítógépes alkalmazás születik. Kutatócsoportunk alapvető célja megvizsgálni, hogy a végsőkig vitt (totális) lexikalizmus mint elméleti keret mennyire sikeres elméletileg, illetve intelligens nyelvtechnológiai alkalmazások fejlesztése céljára.

2.1 Totális lexikalizmus

A totális lexikalizmus azt jelenti, hogy a mondat összeépüléséhez szükséges minden információt a lexikonban tárolunk, így nincs szükség külön szintaktikai szabályrendszerre (frázisstruktúra-szabályokra). A nyelvtan tehát nem más, mint egy nagy adatbázis, amelyben a lexikai egységeket és azok tulajdonságait tároljuk, illetve egyetlen művelet – az *unifikáció* – mint a mondatok összeépülésének motorja.

Az a hipotézisünk, hogy egy ilyen homogén lexikalista nyelvtan mind elméletben, mind gyakorlatban jól működő, hatékony tud lenni; célunk ennek igazolása. Első lépésként kidolgoztuk a nyelvtant a magyar nyelv egy fragmentumára, majd azt implementáltuk Prolog programnyelven [3]. A parser bemenete egy magyar mondat, kimenete annak morfológiai és szintaktikai elemzése (szintaxison a függőségi viszonyokat értve), illetve a mondathoz rendelhető diskurzus-szemantikai reprezentáció.

Következő lépésként kidolgoztuk a nyelvtant az angol nyelv egy kis szeletére, és – kihasználva azt, hogy amit a Prolog elemezni tud, azt generálni is – kipróbáltuk a totálisan lexikalista megközelítést a gépi fordítás területén is [4]. Elképzelésünk az, hogy az univerzálisnak tartott szemantikai reprezentáción keresztül elemzőnk kétirányú használatával bármely két nyelv között megvalósítható a fordítás. Nem kell megírni minden nyelvpárra külön a fordító mechanizmust, csak rendelkeznünk kell az adott nyelvek nyelvtanának implementációjával.

Működő programunk bizonyította, hogy érdemes a totális lexikalizmus eszméjét a nyelvtechnológia területén alkalmazni. A következő lépés annak vizsgálata, hogy nagyobb adatbázison is működik-e a mechanizmus, illetve hogy bármely nyelvi jelenségről számot tud-e adni. Célunk egy pártízezer lexikai egységet tartalmazó relációs adatbázis és egy azt működtető program megalkotása, totálisan lexikalista alapokon. Szeretnénk továbbá minden eddiginél alaposabb gépi fordítást megvalósítani, amely nemcsak mondatok, hanem összefüggő szövegek fordítására is alkalmas, és számot tud adni többek között a retorikai viszonyról, a diskurzusfunkciókról (topik,

fókusz), a hangsúlyról és az aspektusról is. Ehhez egy minden eddiginél alaposabb szemantikai reprezentációt nyújtó elméletre van szükség.

2.2 ReALIS

A ReALIS (Reciprocal and Lifelong Interpretation System) egy reprezentacionalista, dinamikus szemantikai rendszer. Az interpretáció folyamatának az eddigieknél „reálisabb" szemléletét nyújtja. Négy komponensből áll: tartalmaz egy modellt a „külső" világról, a létező entitásokkal és a köztük lévő relációkkal, egy parciális függvényt, amely folyamatosan változó („felfrissülő") információállapotokat társít minden egyes interpretálóhoz, valamint egy-egy függvényt a dinamikus információállapot-változás és a statikus igazságértékelés formális kifejezésére.

Elsődlegesen magának az interpretálónak az információállapotát hivatott ábrázolni, és csak közvetve a feldolgozott diskurzusokét – ezért mondjuk a ReALIS interpretációs megközelítését *élethossziglaninak* ('lifelong'). A rendszer karakterisztikus tulajdonsága a kölcsönös ('reciprok') információábrázolási technika: az interpretáló egy interpretálási folyamat során nemcsak az adott témáról meglévő saját tudására építhet, hanem a másokról feltételezhető tudásra is. Ez olyan konkrét nyelvi tények magyarázatában is szerepet játszhat, mint a névelők, névmások és más anaforikus elemek használata adott mondatbeli helyzetekben.

Előnye más szemantikai rendszerekhez képest, hogy pragmatikai tényezőket is figyelembe vesz, illetve hogy nagyon alapos szemantikai reprezentációt társít nem csupán egy-egy mondathoz, hanem szövegekhez is. Ezek a tulajdonságai teszik alkalmassá intelligens nyelvtechnológiai alkalmazásokra, mint például a gépi fordítás.

3 Szóképzés

A totálisan lexikalista morfológia kiterjesztéseként az inflexiós morfémák mellett a produktív *derivációs morfémákhoz* is közvetlenül lexikai egységeket rendelünk – Alberti (2006) [2] szellemében, a szóalaktanilag „láthatatlan" *konverziós* [17] eseteket is figyelembe véve. Mi több, az alábbi fordítási problémák a „szóképzés" fogalmának még merészebb általánosítását látszanak szükségessé tenni, amennyiben szeretnénk megőrizni egy olyan kézenfekvő fordítási megfeleltetést (legalábbis kiindulópontként), mely szerint a forrásnyelv t_f relatív tövéből létrehozott $K_f(t_f)$ képzett alak célnyelvi megfelelője egy $K_c(t_c)$ képzett alak lesz, ahol a t_c tő a t_f tő megfeleltetettje, a K_c képzési „függvény" pedig a K_f képzésé.

Nézzünk néhány példát a nyelvekben található változatos eszközökre, ahogy a különféle „képzéseket" megvalósítják!

Míg az angol a *progresszív* alakot hagyományos értelemben vett igenévképzéssel hozza létre (ld. (1a): t_f: *set up* ⇒ $K_f(t_f)$: ***sett**ing up*), addig a magyarban úgy foghatjuk fel, hogy a képzett igealak máshova rendeli az igekötőjét ((1b): t_c: ***fel**állítottuk* ⇒ $K_c(t_c)$: *állítottuk **fel***). Megint másként jár el az észt nyelv [23]: a tárgy esetét módosítja ((1c): a $t_{c'}$ argumentumszerkezetében: *Accusativus* ⇒ a $K_{c'}(t_{c'})$ argumentumszerkezetében: *Partitivus*).

(1) a. We set up the tent. / We were setting up the tent.
b. Felállítottuk a sátrat. / Állítottuk (éppen) fel a sátrat (mikor...).
c. Panime telgi püsti / Panime telki püsti.
put.past.1pl tent.acc up / put.past.1pl tent.part up

Az utóbbi két esetben tehát maga az igei szótest nem módosul (*konverzió* [17]), *általánosított képzésként* foghatjuk azonban fel a vonzatszerkezet-módosulást [2]. Az (1) példasorban bemutatott szófajváltást megvalósító K_c (igenév-) képzésnek tehát a két tekintett célnyelvben a finit igei alakot megőrző operáció feleltethető meg. Az észtbeli $K_{c'}$ operációt olyan argumentumszerkezet-módosító igeképzésnek tekinthetjük, mint a magyarban az *átúszik a folyón → átússza a folyót* példa által szemléltethető változást: más lexikai egység szerepel a kimenetben, mint a bemenetben, hiszen már esetkeret jelölődik ki. A magyarbeli K_c operációt abban az értelemben tekinthetjük képzésnek, hogy a bemeneti alakban prefixummá inkorporálódó igekötői argumentum a kimeneti alakban szóvá önállósul [2]; a különbséget tulajdoníthatjuk eltérő lexikai leírásoknak (ahogy az *eszik egy almát → almát eszik*, *alakult egy kórus → kórus alakult* példapárok esetében is eltérő lexikai tétel megadásával szeretnénk számot adni a bemenet és a kimenet különbségéről, ami az egyik argumentum inkorporált helyzetét illeti).

Ugyanígy, nyelvenként különbözően valósulhat meg a passzív progresszív alakok előállítása (lásd a (2) példasorban: K_f igenévképzés, K_c új igetétel létrehozása egy argumentum inkorporált helyzetének megszüntetése révén, míg $K_{c'}$ új igetétel létrehozása egy argumentum esetének módosítása révén [23]), illetve az aktív passzív képzés (a (3) pontban):

(2) a. The owls were checked. / The owls were being checked.
b. Átvizsgálták a baglyokat. / Vizsgálták (éppen) át a baglyokat (mikor...).
c. Vaadati öökullid üle. / Vaadati öökulle üle.
look.impers owl.nompl over / look.impers owl.partpl over
(3) a. Harry kissed his mother. / Harry was kissed by his mother.
b. Harry megcsókolta az anyját. / Harry-t megcsókolta az anyja.
c. Harri suudles oma ema. / Harrit suudles ta ema.
H.nom kiss.3sg.past his-own mother.part H.part kiss.3sgpast his mother.nom

Ez utóbbi esetében (3) az angol ismét igenévképzéshez folyamodik, míg a magyar és az észt megőrzi az időjeles igeformát, sőt az esetkeretet is, továbbá argumentuminkorporációval kapcsolatos változás sem történik. Ami változik, az az argumentumok szórendi helyzete. Amennyiben úgy tekintjük, hogy a magyarban és az észtben nem *alulspecifikált* az igei régens ellenőrzése alatt tartott argumentumok sorrendje, hanem különféle sorrendek különféle lexikai tételekhez tartoznak, akkor még a (3b-c) pontokban bemutatott „topikalizációt" is tekinthetjük általánosított képzésnek.

Visszatérve az angol *–ing* igenévképzésre, annak egyes esetekben akár a magyarban is igenévképzés feleltethető meg (pl. *runn**ing*** ~ *fut**ó***), máskor azonban (igekötő híján) formai szempontból „identikus függvénynek" kell tekintenünk a képzést ((4): [*ran ~ futott*] [*was runn**ing*** *~ futott*]), vagy egy mátrixigének kell megfeleltetnünk ((5): [*dim ~ ostoba*] [*be**ing*** *dim ~ ostobán* ***viselkedett***].

(4) a. He ran. / He was running.
b. Futott. / Futott (éppen, mikor...).
c. Ta jooksis. / Ta jooksis.
s/he run.3sgpast / s/he run.3sgpast

(5) a. Harry thought Ogden was (being) extremely dim.
b. Harry úgy gondolta, hogy Ogden rendkívül ostoba (... ostobán viselkedik).
c. Harri arvas, et Ogden [on erakordselt tobe]. / [käitub erakordselt tobedalt].
H.nom think.3sgpast that O.nom [be.3sg extremely stupid] / [behave.3sg exteremely stupidly]

Ez utóbbi (5) a szerint a gyakori fordítási megfelelés szerint működik, amit az angol és a magyar viszonylatában a másik irányban tapasztalunk gyakrabban: amikor is magyarbeli szuffixálással megvalósuló képzésnek (pl. K_f: *-(t)At*, *-hAt*, *-Ul*, *-ít, -(V)ll*), mely igei, illetve melléknévi töveken működik, egy mátrixigei régens bevonása feleltethető meg, mely önálló nem finit igealakot, illetve melléknevet szelektál (rendre: *dolgoztat ~ make sy work*, *dolgozhat ~ can/may work*, *barnul ~ become/grow brown*, *barnít ~ make sg brown*, *drágáll ~ consider sg expensive*).

Totálisan lexikalista morfoszintaktikai megközelítésünk kiegészítve a képzés felfogásának fenti erőteljes általánosításával lehetővé teszi tehát, hogy a fordítás azon eseteit is képesek legyünk *szabályalapúan* (és kompozicionális jelentéskalkulációval társítva) kezelni, ahol elvész a szófaji megfelelés, sőt még a szó~szó megfelelés is két nyelv között a lexikai egységek mondattá szerveződésének folyamatában.

4 Gépi fordítás

A gépi fordító rendszereket két nagy csoportra oszthatjuk: szabályalapú [15] és mintaalapú [22] megközelítésekre. A kezdetekre szabályalapú rendszerek fejlesztése volt a jellemző, míg az utóbbi években inkább hatalmas korpuszok és fordítómemóriák segítségével próbálják létrehozni a minél kisebb emberi beavatkozást igénylő gépi fordító programokat.

A szabályalapú rendszereken belül többféle megközelítést alkalmazhatnak. A fordítás történhet (1) direkt módon, kis elemzéssel, (2) közvetítőnyelven keresztül, illetve (3) transzferrel, amikor a forrásnyelvi mondatból egy absztrakt forrásnyelv-közeli reprezentáción, majd egy célnyelv-közeli reprezentáción keresztül állítják elő a mondat célnyelvi megfelelőjét.

A fordításhoz két lépésben jutnak el. Az első lépés az analízis (elemzés), amely a forrásnyelvi szöveg szintaktikai elemzésén túl tartalmazhat morfológiai, és valamilyen mértékű szemantikai elemzést is – ez utóbbira a többértelműségek kezeléséhez van (minimálisan) szükség. A második lépés pedig a szintézis (generálás), amikor a mondat célnyelvi megfelelőjét állítják elő lehetőleg ugyanazzal a mechanizmussal. Így a modularitás és a megfordíthatóság fontos tulajdonságaik ezeknek a rendszereknek.

Miután a szabályalapú megközelítés nem bizonyult kellően hatékonynak, még statisztikai módszerek bevonásával sem, továbbá rengeteg befektetett munkát igényelt a nyelvtaníró részéről, a 90-es évektől kezdtek tért hódítani a mintaalapú fordítórendszerek. Hatalmas korpuszok születtek, amelyeket minél alaposabban annotáltak, annál jobb minőségű elemzőket lehetett rájuk építeni, viszont a ráfordítási idő is jelentősen megnőtt. Közülük leghatékonyabbnak a treebankok (mondatszerkezettel is annotált korpuszok) bizonyultak, hiszen azok tartalmazzák a lehető legtöbb információt; viszont előállításuk nem egyszerű feladat, sok munkaórát igényel. Éppen ezért napjainkban egyre több olyan rendszert fejlesztenek, amely annotálatlan korpuszból képes „megtanulni” a nyelv elemeit és azok tulajdonságait különféle módszerek (analógia, disztribúciós módszer) segítségével, így kevés ráfordítással készíthetnek haté-

kony elemzőket. Nagyon elterjedt továbbá gépi fordításkor a párhuzamos korpuszok, illetve különféle fordítómemóriák használata, melyeket például az interneten fellelhető több nyelven elérhető anyagok szinkronizálásával állítanak elő.

A korpuszalapú megközelítések sem hoztak teljes sikert. Nem minden jól formált kifejezés található meg bennük, különösen a gazdag morfológiájú nyelvek esetében, mint a magyar, és nagyon ritka az igazán alaposan és jól annotált anyag, vagy megfelelően illesztett párhuzamos korpusz. Emiatt születnek különféle hibrid megoldások: alapvetően korpuszra támaszkodó, de valamilyen szintű elemzést is végző, vagy szabályalapú, de a többértelműségek kezelésében korpuszt is használó rendszerek.

Napjainkban egyre többen kezdenek visszatérni a szabályalapú megközelítéshez, de nem sekélyelemzést használnak, ami csak részleges elemzést ad, és hiába robusztus, ha nem annyira precíz [12], így komplexebb célokra (pl. igazán jó minőségű fordítás) nem alkalmas. A mélyelemzést végző rendszerek sokkal alaposabbak és pontosabbak, és az utóbbi időben lefedettségben is felveszik a versenyt a sekélyelemző rendszerekkel. Az ilyen nyelvészeti alapú, kézzel írt nyelvelemzők készítéséhez szükséges kezdeti nagyobb energiabefektetés pedig megtérül később, például amikor új nyelvekre dolgozzák ki [10]. Az igazán jól működő és hatékony mélyelemző rendszerek unifikációs mechanizmusokat használnak, így előnyeik között szerepel, hogy egyszerűbb szabályokat lehetséges megfogalmazni, gyorsabb, egyszintű, lexikalista és megfordítható [15]. A többértelműség kezelésére ezek a rendszerek is használhatnak korpuszt, mivel sokszor szabályba nem foglalható tényezők szükségesek az egyértelműsítéshez (pl. kontextus), illetve optimalitás-jelölő rangokat, hogy a ritka alakzatokat is elfogadja az elemző, de alapesetben a gyakoribb elemzés mellett döntsön [10].

A legtöbb gépi fordító rendszert egy adott nyelvpárra dolgozzák ki, sőt, akár csak az egyik irányban működnek, de léteznek többnyelvű rendszerek is, ahol a mechanizmus univerzalitására törekszenek.

4.1 Létező rendszerek

Egyetértés mutatkozik abban, hogy ha a cél egy igazán intelligens gépi fordító rendszer létrehozása, nem elégségesek a sekélyelemzést végző programok. A mintaalapú és statisztikai megközelítések ellen pedig az szól, hogy a legtöbb nyelvre nem létezik igazán jól használható párhuzamos korpusz [7]. További érv a szabályalapú gépi fordítás mellett az újrahasznosíthatóság: hogy a gépi fordítás fejlesztésének eredményei más nyelvtechnológiai alkalmazásokban is használhatók, illetve más területek eredményeit a gépi fordítás is hasznosíthatja. Ezért ha olyan gépi fordító rendszer fejlesztése a cél, ami bármely két nyelv esetében[15] jó minőségű és pontos fordítást ad, akkor mélyelemzést végző szabályalapú rendszer alkalmazása tűnik a legjobbnak.

Napjainkban erre a célra az unifikációs mechanizmusokat használó lexikalista elméletek látszanak a legmegfelelőbbnek, amik nemcsak a konfigurációs nyelveket (mint az angol) kezelik hatékonyan, hanem a magyarhoz hasonló szabadabb szórendű nyelveket is. Továbbá a jó minőségű fordításhoz szükségesnek látszó valamiféle

[15] Rokon, vagy nagyon hasonló nyelvek közötti fordításkor elég lehet a sekélyelemzés, ha nagyfokú a morfológiai és szintaktikai hasonlóság. Például csehről szlovákra lehet szinte szóról szóra fordítani [14].

szemantikai reprezentáció hozzárendelésében is jobb eredményeket érnek el, mint a frázisstruktúra-nyelvtanok.

A lexikalista nyelvtanok közül a legeredményesebbek LFG vagy HPSG formalizmust használnak. Számos nyelvre léteznek már nagy lefedettségű elemzőik, amik nagyon jó minőségű és alapos elemzést adnak, továbbá szemantikai reprezentációt is társítanak a mondatokhoz. Különböző nyelvekre alkalmazzák ugyanazt az elméleti keretet, így tudják elérni, hogy a különböző nyelvű elemzések szinte teljesen párhuzamosak legyenek, ami jelentősen megkönnyíti az elemzőkre épülő gépi fordító rendszerek fejlesztését.

Ilyen nyelvtan például az ERG (English Resource Grammar), ami a legnagyobb HPSG-alapú nyelvtan angol nyelvre, implementálva az LKB (Linguistic Knowledge Building) rendszerben [9]. HPSG-re kifejlesztettek egy keretrendszert is (Grammar Matrix [5]), ami nem egy nyelvtan, hanem nyelvtanok feletti általánosítások gyűjteménye. Négy függetlenül fejlesztett HPSG-alapú nyelvtanból indult ki: angol, japán, német és norvég. A cél egy egységes nyelvtan megalkotása az egyes nyelvtanok feletti általánosítások alapján, hogy további megszorítások bevezetésével még egységesebb elemzések születhessenek, és hogy újabb nyelvek nyelvtanainak megírását gyorsabban lehessen elkezdeni. A Grammar Mátrixszal több kompatibilis nyelvtant is fejlesztettek. Ilyen például a JACY [21], ami egy nagy lefedettségű, nagy pontosságú japán nyelvtan. HPSG-alapú, MRS-t (Minimal Recursion Semantics[16]) használ a szemantikai reprezentáció hozzárendeléséhez. Eredetileg beszéltnyelvi dialógusok gépi fordítására fejlesztették, később automatikus email-megválaszolásra és egyéb (többnyelvű) nyelvtechnológiai feladatokra is használták.

Az LFG-t mint elméleti keretet használó elemzők közül érdemes megemlíteni a Parallel Grammar (ParGram [8]) projektet. Lényege, hogy a hasonló szerkezetek elemzései a különféle nyelvekben amennyire csak lehet, párhuzamosak. Így hasonló számítógépes alkalmazásokban használhatók, és a gépi fordítás is egyszerűsíthető. Eredeti célja az LFG-formalizmust tesztelni: univerzalitását, lefedettségének határait, és hogy mennyire tartható a párhuzamosság a különféle nyelvek között. A párhuzamosságot az f-struktúra (funkcionális szerkezet) szintjén érik el[17]. Az eredmények bíztatóak, nagyfokú párhuzamosság érhető el, így új nyelvtanok építése is könnyebben, gyorsabban lehetséges. Az elemzőt szándékosan nagyon különböző nyelvekre dolgozták ki először: angolra, németre, japánra, urdura és norvégra. Némelyik nagy lefedettségű, ipari alkalmazás, némelyiknél az s-struktúrát (szemantikát) is kidolgozták. Az elméletet az XLE (Xerox Linguistic Environment) platformon implementálták.

Jónéhány a különféle lexikalista alapú elemzők közül tehát elérte már azt a lefedettséget, amit korábban csak nagy korpuszok vagy statisztikai módszerek alkalmazásával lehetett megvalósítani. Ezek az elemzők mélyelemzést végeznek, kimenetükben szemantikai (vagy szemantika-közeli) reprezentáció is szerepel, továbbá nagyfokú párhuzamosságot képesek megvalósítani akár nagyon különböző nyelvek között is. Mindezek lehetővé teszik, hogy jó minőségű gépi fordító rendszerek épüljenek

[16] Lapos szemantikai reprezentációt rendel szavakhoz és frázisokhoz. Mélyebb szintű, mint csupán a predikátum-argumentum viszonyok. Rendezetlen struktúra, hatókör tekintetében alulspecifikált.

[17] Néha szándékosan nem párhuzamosak az f-struktúrák, pl. a névszói állítmány esetében (az *It is red* mondat esetében az angolban a kopula a fej, a japánban a melléknév), vagy amikor egy nyelvben megvan egy bizonyos jegy (pl. a németben a nem), egy másikban pedig nincs.

rájuk. Legnagyobb részük még kísérleti fázisban tart, de eredményeik nagyon ígéretesek. Több nyelvre is kidolgoztak már működő gépi fordító rendszereket, amelyek ezeket az elemzőket (és generálókat) használják. Legtöbbjük transzfer-alapú, így minden nyelvpárra és mindkét fordítási irányra külön transzfer-komponenst kell kidolgozni. A magyarra ilyen rendszer eddig még nem készült[18].

LFG formalizmust használ például az XTE (Xerox Translation Environment), ami a ParGram projekt fordítórendszere [11]. A nyelvfüggetlennek tartott f-struktúra teszi alkalmassá többnyelvű nyelvtechnológiai alkalmazásokra, mint a gépi fordítás. A hatékony elemzőt és generálót tartalmazó XLE platformot kiterjesztették egy transzfer-komponenssel. A transzfer az f-struktúrán keresztül történik, amiről elismerik, hogy néha nem elég univerzális (pl. a fejváltó fordítás esetében), de a nyelvek közötti különbségek legnagyobb része már ezen a szinten eltűnik. A legjobb az s-struktúrán keresztüli fordítás lenne, viszont ekkor ki kellene dolgozni a szintaxis-szemantika interfészt mind az elemzés, mind a generálás oldaláról (azóta ez irányban folytak a kutatások). A többértelműséget a döntés végsőkig való halasztásával kezeli. HPSG-formalizmust használ például a DELPH-IN (Deep Linguistic Processing with HPSG Initiative [7]), ami egy nyílt forráskódú, szemantikai transzfer-alapú gépi fordító rendszer. Létező forrásokat alkalmaz: elemzőket, generálókat, kétirányú nyelvtanokat és transzfer-motort. Kezdetképpen japánról angolra fejlesztették ki. A cél a mechanizmus működésének megmutatása, ezért egyelőre száznál kevesebb transzfer-szabályt, és csupán párezres transzfer-lexikont tartalmaz. A rendszer a forrásnyelvet (japán) elemzi egy szabályalapú forrásnyelvi nyelvtannal (JACY), a legjobb elemzést valószínűségi rangok alapján kiválasztja; kimenete egy precíz, alulspecifikált (nyelvspecifikus) forrásnyelvi szemantikai reprezentáció (MRS_S). Ebből a transzfer-motor előállítja az alulspecifikált (nyelvspecifikus) célnyelvi (angol) szemantikai reprezentációt (MRS_T) újraíró szabályokkal. Ha több megoldás van a szabályalkalmazásnál, több fordítás lesz. Végül a célnyelvi generátor (ERG) angol nyelvű szöveget készít belőle. További céljaik között a lefedettség növelése szerepel, valamint választ kapni néhány elméleti kérdésre: mennyire lehet egységes a szemantikai reprezentáció, lehet-e (kvázi)automatikusan növelni a nyelvtanokat és lexikonokat, mi lehet a lexikális szemantika szerepe, stb. A fejlesztés egyelőre kísérleti stádiumban van, azonban nagyon ígéretes.

Létezik olyan rendszer is, amely LFG és HPSG alapú elemzőket is használ, a norvég-angol gépi fordítást megvalósító szemantikai transzfer-alapú LOGON [18]. Olyan célokra kezdték fejleszteni, ahol a fordítás minősége fontosabb a nagy lefedettségnél. A fordítás három lépésben történik: első a norvég mondat LFG-alapú grammatikai és szemantikai mélyelemzése, amihez az XLE platformon fejlesztett NorGram-ot használják (amit a ParGram projekten belül fejlesztettek). Kimenete egy nyelvspecifikus logikai szemantikai reprezentáció (MRS). Majd ezeknek a reprezentációknak a transzfere történik nyelvspecifikus angol reprezentációkba (MRS). Végül a szemantikai reprezentációból angol nyelvű mondatot generálnak HPSG alapú elemző (generáló) segítségével (célnyelvtan: ERG, generátor: LKB). A projekt hosszú távú, bár célja egyelőre csak egy demonstráció fejlesztése, amely megmutatja, hogy a mechanizmus működik. A többértelműség kezelésére valószínűségi rangokat használ.

[18] A magyarra működő gépi fordító rendszerek vagy mintaalapúak (Hunlish [13]), vagy alapvetően szabályalapúak (MetaMorpho [19]), de nem lexikalisták. Nem céljuk továbbá olyan minőségű fordítás, mint a cikkben tárgyalt egyéb alkalmazásoknak.

Fontos jellemzője a modularitás, így mindig a legfrissebb elemzőt rakhatják a rendszer mögé.

4.2 Javaslatunk

A ReALIS projekt a végsőkig viszi a lexikalista megközelítést. Azt próbálja igazolni, hogy nincs szükség frázis-struktúrára: a mondatok összeépüléséhez elégséges a lexikai egységek gazdag jegystruktúrája és egyetlen művelet, az unifikáció.

Schneider [21] is amellett érvel, hogy nem feltétlenül kell konfigurációs felszíni reprezentációs szint a nyelv szintaktikai elemzéséhez. Az LFG például főleg azért használ c-struktúrát (összetevős szerkezetet), mert az környezetfüggetlen. Ha csak a függőségi viszonyokat mutató f-struktúrát használná, az nem lenne hatékony, mivel környezetfüggő. Akárcsak a Függőségi Nyelvtan, ahol (eredetileg) szintén csak funkcionális szint van, nem használ összetevős szerkezetet. A szórendnek nincs elsődleges szerepe (néha segít az egyértelműsítésben), de valójában nincsenek megszorítások arra nézve, hogy egy fej hol keresse a bővítményeit. Schneider [21] megmutatja, hogyan lehet az f-struktúrát környezetfüggetlenül előállítani, így nem lenne szükség a teljes c-struktúrára, csak a fontosabb elemeire.

A totálisan lexikalista megközelítés a végsőkig viszi ezt a javaslatot: kipróbálja, lehetséges-e semmivé redukálni a c-struktúrát. Készíthető-e hatékony elemző (és gépi fordító) rendszer akkor, ha nem támaszkodunk a környezetfüggetlen c-struktúrára, vagy egyéb konfigurációs felszíni reprezentációs szintre. Véleményünk szerint a válasz igen: a megközelítés hatékony, mert a függőségi nyelvtanokkal szemben nálunk vannak megszorítások a szórendre, az ún. rangparaméterek formájában [3]. Ezekkel a paraméterekkel elegánsan megragadható az is, hogy egy nyelven belül milyen szórendi variánsok lehetségesek, és számot ad a nyelvek közötti szórendi különbségekről is.

Célunk tehát egy olyan (szabályalapú) elemző és gépi fordító rendszer fejlesztése, amely totálisan lexikalista, nyelvfüggetlen, és minden eddiginél alaposabb szemantikai reprezentációt társít nem csupán mondatokhoz, hanem szövegekhez. Programunk a 3. pontban bemutatott nyelvenkénti sokszínűséget ezt a (ReALIS alapú) szemantikai reprezentációt mint közvetítőnyelvet alkalmazva hidalja át, hiszen ez nyelvfüggetlen közös alapot jelent, amihez képest aztán a felszíni megjelenítés az egynyelvű komponensek feladata. Így a fordításhoz alaposan kidolgozott elemzőkre van csupán szükség, melyek legfontosabb része a szemantikai komponens; hiszen programunk ugyanazt a mechanizmust használná bármely két nyelv esetében, és a fordítási iránytól függetlenül.

Morfémaszintű totális lexikalizmusunk pedig oly módon fejleszti tovább a grammatika *egyszintűségének* a gondolatát, hogy a kategoriális többértelműség alább szemléltetett potenciális túlburjánzása kikerülhetővé válik, hiszen az adott példában egy moduláris rendszer szintaxisának 72 megemlített lehetséges bemenete helyett pusztán azt a három elemzési utat kell bejárnunk, amit az időjeles igeként elemezhető három egység indít el.

(6)

Dobom	az	ír	szánom.
N+poss1sg+Nom/N+poss1sg+Acc/V+1sg	Pron/D	V/A/N_1/N_2	N+poss1sg+Nom/N+poss1sg+Acc/V+1sg
3	·	2 · 4	· 3 = 72

A magyar nyelvre nem készült még olyan alapos elemző és gépi fordító program, mint amilyen a ReALIS projekt céljai között szerepel. Olyan jelenségekről is számot kívánunk adni, mint a retorikai viszonyok (hogyan kapcsolódnak a mondatok egymáshoz), a különféle diskurzus-funkciók (mint a topik és a fókusz), vagy a hangsúly és az aspektus. Mindehhez egy minden eddiginél alaposabban kidolgozott formális szemantikai eszköztár áll rendelkezésünkre.

5 Összegzés

Az elmúlt években a GeLexi projekt egy magyar és angol mondatokat elemző Prolog programot fejlesztett, hogy igazolja a totális lexikalizmus eszméjének a gyakorlatban való alkalmazhatóságát [3]. Megmutattuk, hogy az elemzőnk kétirányú használatával a gépi fordítás is lehetséges, a diskurzus-szemantikai reprezentációt mint közvetítőnyelvet alkalmazva [4]. Eközben a LiLe projekt egy relációs adatbázis létrehozásába fogott, egyéb célok mellett azért, hogy nagyobb adatbázison vizsgálhassuk a totálisan lexikalista megközelítés hatékonyságát [6].

A ReALIS projekt keretében a két kutatócsoport egyesült, és további nyelvészek bevonásával azt a célt tűzte maga elé, hogy minden eddiginél pontosabb gépi fordítást valósítson meg, lényegesen nagyobb adatbázison. Rendszerünk szabályalapú, mélynyelvészeti elemzést végez, a totálisan lexikalista GASG (Generatív Argumentumstruktúra Nyelvtan) alapján, amely nem rendel összetevős struktúrát a mondatokhoz, csupán a lexikai egységek tulajdonságaira támaszkodik az elemzés során, a szórendről pedig rangparaméterek segítségével ad számot. A mondatokhoz rendelt diskurzus-szemantikai reprezentáció, amely a ReALIS [1] implementációján alapul, az eddigi (LDRT-alapú) reprezentációnál jóval részletesebb, olyan jelenségekről is számot tud adni, mint a retorikai viszonyok, a diskurzusfunkciók vagy az aspektus. A reprezentáció nyelvfüggetlen, így a gépi fordítás során nem okoz problémát a nyelvek sokfélesége: bizonyos lexikai egységek az egyik nyelvben szavak, a másikban szuffixumok, a topikalizációt az egyik nyelv pusztán a szórenddel, a másik képzők hozzáadásával fejezi ki, vagy a progresszív aspektust az egyik nyelvben segédige jelzi, a másikban pedig egy vonzat esetének módosítása.

Az utóbbi években a mélynyelvészeti, lexikalista, unifikációra épülő elemzők és gépi fordító rendszerek egyre nagyobb teret hódítanak. Már nem csupán pontosabbak, mint a mintaalapú vagy sekélyelemzést végző rendszerek, hanem lefedettségben is kezdik felvenni a versenyt velük. Ezeknek a rendszereknek a sikere még inkább arra ösztönöz minket, hogy folytassuk a totálisan lexikalista megközelítés nyelvtechnológiai alkalmazhatóságának vizsgálatát, és létrehozzunk egy minden eddiginél pontosabb gépi fordító rendszert.

Bibliográfia

1. Alberti, G.: ReAL Interpretation System. L. Hunyadi – Gy. Rákosi – E. Tóth eds.: Preliminary Papers of the Eighth Symposium on Logic and Language, University of Debrecen (2004) 1–12
2. Alberti, G.: Changes in Argument Structure in the course of Derivation in Hungarian. Acta Linguistica Hungarica (2006)

3. Alberti, G., Kleiber, J.,Viszket, A.: GeLexi project: Sentence Parsing Based on a GEnerative LEXIcon. Acta Cybernetica 16 (2004) 587–600
4. Alberti G., Kleiber J., Viszket A.: GeLexi projekt: Fordítás totálisan lexikalista alapokon. In: Alexin Z. – Csendes D. (szerk.): II. Magyar Számítógépes Nyelvészeti Konferencia, Juhász Nyomda, Szeged (2004) 73-80
5. Bender, E. M., Flickinger, D., Oepen, S.: The Grammar Matrix: An Open-Source Starter-Kit for the Rapid Development of Cross-Linguistically Consistent Broad-Coverage Precision Grammars. In Proc. COLING 2002 Workshop on Grammar Engineering and Evaluation. Taipei (2002)
6. Bódis Z., Kleiber J., Szilágyi É., Viszket A.: LiLe projekt: Adatbázis mint „dinamikus korpusz". In: Alexin Z. – Csendes D. (szerk.): II. Magyar Számítógépes Nyelvészeti Konferencia, Juhász Nyomda, Szeged (2004) 11–18
7. Bond, F., Oepen, S., Siegel, M., Copestake, A., Flickinger, D.: Open source machine translation with DELPH-IN. In Proceedings of the Open-Source Machine Translation Workshop at the 10th Machine Translation Summit, Phuket, Thailand (2005) 15-22
8. Butt, M., Dyvik, H., King, T. H., Masuichi, H., Rohrer, C.: The Parallel Grammar project. In Proceedings of COLING-2002 Workshop on Grammar Engineering and Evaluation, Taipei, Taiwan (2002)
9. Copestake, A.: Implementing Typed Feature Structure Grammars. Stanford, CA: CSLI Publications (2002)
10. Forst, M., Kuhn, J., Rohrer, C.: Corpus-Based Learning of OT Constraint Rankings for Large-Scale LFG Grammars. In: Butt, M. – King, T. H. (eds.): Proceedings of the LFG'05 Conference, University of Bergen, CSLI Publications (2005) 154-165
11. Frank, A.: From parallel grammar development towards machine translation. In Proceeding of MT Summit VII (1999) 134–142
12. Frank, A.: Projecting LFG F-structures from Chunks --- or (Non-)Configurationality from a different Viewpoint. In: Butt, M. – King, T. H. (eds.): The Proceedings of the LFG'03 Conference, University at Albany, State University of New York, CSLI Publications (2003) 217-237
13. Halácsy P., Kornai A., Németh L., Sass B., Varga D., Váradi T., Vonyó A.: A hunglish korpusz és szótár. In: Alexin Z. – Csendes D. (szerk.): III. Magyar Számítógépes Nyelvészeti Konferencia, Juhász Nyomda, Szeged (2005) 134-143
14. Homola, P., Kuboň, V.: A Translation Model for Languages of Acceding Countries. In: John Hutchins (ed.): Broadening Horizons of Machine Translation and its Applications. Proceedings of the Ninth EAMT workshop. Foundation for International Studies, University of Malta, Valletta (2004) 90-97