Részleges gépi fordítás a NooJ rendszerben

Váradi Tamás

MTA Nyelvtudományi Intézet
1068 Budapest Benczúr u. 33

Kivonat Az előadás ismerteti azokat az eredményeket, amelyeket a NooJ nyelvelemző fejlesztőrendszer[8][5] gépi fordításra való alkalmazása terén elértünk. Bemutatja a rendszer azon új képességeit, amelyek alkalmassá teszik a lokális grammatikákat kétnyelvű felhasználásra. A lokális grammatikák kiválóan alkalmasak az egyedi lexikai szabálytól a szóosztályokra érvényes általános szabályok megfogalmazására. A dolgozat fő tézise, hogy a rendkívül könnyen kezelhető, gyors és interaktív NooJ rendszer jól alkalmazható részleges gépi fordítást igénylő feladatokra.

Kulcsszavak: gépi fordítás, véges állapotú nyelvelemezés, lokális grammatika, NooJ

1. Bevezetés

A jelen dolgozat célja, hogy bemutassa azokat a lehetőségeket, amelyeket a NooJ keretfejlesztő rendszer gépi fordításhoz kínál. Egy gépifordító rendszer létrehozása természetesen rendkívül összetett folyamat, amely kihívást jelent a számítógépes nyelvészet egésze számára, és nagy szerep jut benne a szoftver technológiának is. Ebben a dolgozatban az alulról építkezés jegyében áttekintjük azokat az elveket, amelyek szerint a mondatok gépi fordítása legalább részlegesen elvégezhető, és bemutatjuk ezek megvalósítását a NooJ rendszerben. A 2. részben ismertetjük a részleges, minta alapú gépi fordítás elveit, a a 3. részben bemutatjuk a NooJ eszköztárát, amellyel lehetővé válik gépi fordítás megvalósítása, a 4. részben példákat mutatunk be a főnévi csoportok fordítására.

2. Elméleti háttér és motiváció

A nyelvi modell, amely a jelen munkálatot motiválja Morris Gross lokális grammatika fogalmára[3] épül. Ez az elmélet a végesállapotú technológiával jól kezelhető lokális grammatikai viszonyokra helyezi a hangsúlyt, sok tekintetben a generatív grammatika uralkodó ágával szemben, ahol a távoli függőségek vizsgálata a domináns. Gross a ma már igen elterjedt lexikális nyelvtanok, sőt mondhatni a konstrukciós grammatika előfutárának tekinthető. További fontos jellemzője munkásságának nemcsak a lexikon és szintaxis határának, de a szintaxis és szemantika közötti viszony újrafogalmazása. A nyelvi szerkezetek szemantikai alapú megközelítésével a szellemi elődének tekintett Harris[4] munkásságát követi.

2.1. Minta alapú fordítás

A gépi fordítás technológiáját tekintve eljárásunk alulról felfelé haladó (bottom up) *minta alapú fordításnak* tekinthető, amely az alábbi elvekben megegyezik a Metamorpho rendszer[6][7] megközelítésével. A minta fogalma a lokális grammatikában is folytonos átmenetet jelent az egyedi szavakat tartalmazó lexikon valamint a csak szóosztályokra vonatkozó szintaktikai szabályok között. Ez a rugalmasság az implementáció síkján is megvalósul, mert a NooJ lexikai komponense ugyanúgy véges állapotú transzdúszerekből áll, mint a szintaktikai rész, sőt a lexikai redundancia kezelésére a lexikonban is szabályokat alkalmazunk.

A lokális grammatikák végesállapotú transzdúszerek, amelyeket lépcsőzetesen alkalmazunk (*cascaded finite state transducers*) oly módon, hogy először a legszűkebb hatókörű (azaz a legtöbb egyedi lexikai elemet tartalmazó) szabályokat futtatjuk, majd az egyedi lexikai elemeket fokozatosan szóosztályok váltják fel, míg végül a legáltalánosabb, tehát csak szóosztályokra vagy szintaktikai csoportokra hivatkozó szabályok következnek. Ez az eljárás biztosítja azt, hogy a lehető leghamarabb megtaláljuk az adektvát fordítási megfelelőt. Az egész eljárás robosztusságát az adja, hogy a grammatikák lexikai és egyéb megkötéseit végsőkig feloldva eljutunk a szó szerinti megfeleltetéshez. Vagyis, ha semmi egyéb lokális grammatika nem volt illeszthető egy kifejezésre, végső soron megkapjuk az alkotóelemeknek a szótárban felsorolt célnyelvi megfelelőit. Ha több van belőlük, akkor mindegyiket.

2.2. Részleges fordítás

A harmadik rokon vonás a Metamorpho rendszerrel, hogy a lokális grammatikák kimenetét a célnyelvi megfelelők adják, azaz a mintaillesztéssel egyben előáll a hozzá tartozó célnyelvi megfelelő. Naivitás azzal áltatnunk magunkat, hogy egyenes út vezet a lépcsőzetes eljárás első szintjein előállt célnyelvi megfelelőktől a koherens célnyelvi mondat megformálásáig. Természetesen további kutatások szükségesek annak az eljárásnak a kidolgozására, amelynek során a sok-sok egyedi megfelelés egyetlen mondattá áll össze.

Bár Gross a lokális grammatikát végső soron az egész mondat struktúrájának leírására alkalmasnak tartotta, mi szerényebb célt követünk. Először a felszini szerkezetből szótár és minta-alapú eljárások segítségével megragadható mondatrészeket igyekszünk előállítani. Ez az eljárás elveiben hasonlít a részleges felszíni elemzéshez (chunking)[1][2], de nem áll meg az elemi szerkezeteknél, hanem a végesállapotú eljárást a maximális kiterjesztésű főnévi csoportokig viszi. A dolgozat egyik központi tétele, hogy a mondatban a maximális kiterjesztésű főnévi csoportok gépi fordítása elérhető célt jelent, amely kétnyelvi gyakorlati alkalmazásokban (pl. információkinyerés) önmagában is gyakorlati értékkel bír. A főnévi csoportok sikeres gépi fordításának elvi esélyét egy szintaktikai és egy szemantikai tényező is növeli.

Szintaktikailag kedvező az a tény, hogy a főnévi csoporton belüli szórend még a magyar esetében is kötött, ami azt jelenti, hogy a főnévi csoport végesállapotú

technológiával jól kezelhető. Ennek a szempontnak az érvényesülését csorbítják magyarban az igeneves szerkezetek, amelyek szinte teljesen nyitottá teszik a módosító szerkezetet. Angolban a prepozíciós szerkezetek főnévi csoporthoz sorolásának [*PP attachement* jól ismert problémája okoz gondokat. Mindezekkel együtt döntően lokális függőségekkel kell számolnunk.

A szemantikai érvet talán helyesebb inkább hipotézisként megfogalmaznunk. E szerint a maximális kiterjesztésű főnévi csoportok olyan szematikailag funkcionális (referáló) egységek, amelyek „megőrződnek" a fordítás során. Jogos az elvárás, hogy az alábbi mondatban

> [A biztonsági szolgálat emberei] [pontban hatkor] megnyitották [a Magyarországra települt legújabb áruházlánc első üzletének a kapujait].

a szögletes zárójelek között szereplő részek pontos, tényszerű fordítása tartalmazza ezek önálló célnyelvi megfelelőit a célnyelvi mondatban. Azt nyilvánvalóan nem várhatjuk el, hogy ezek belső szerkezete is azonos legyen (pl. az igeneves módosító szerkezetnek valószínűleg vonatkozó mellékmondat felelne meg az angolban fordításban) de azt igen, hogy az egész egységnek legyen hasonló szintaktikai státuszú megfelelője. Jegyezzük meg, hogy itt most nem a „megnyitotta kapuját" kifejezésről van szó, amely többszavas lexikai egységként funkcionál, és a kifejezés *egészének* van fordítási megfelelője, ami viszont nem feltétlenül tartalmazza a „kapu" szó megfelelőjét (pl. lehet „launched its operations, set up shop, started trading" stb.)

Végezetül felmerül a kérdés a vállalkozás motivációjáról: nem eleve nagyfokú naivitás egy ilyen komplex műveletet mint amilyen a gépi fordítás, egyetlen szoftver eszközzel megkísérelni? Nem hamis illúziókat keltünk, amikor a NooJ-t mint gépi fordítórendszert próbáljuk beállítani? Úgy vélem, hogy bár ezek a kérdések sok tekintetben jogosak, nyomós érvek szolgálnak a NooJ gépi fordításra való alkalmazása mellett. Először is, amint a 3. részben látni fogjuk, a NooJ igen komplex rendszer, ami ugyanakkor ingyenesen elérhető és sokoldalúan fejleszthető: valójában tetszőleges nyelvre akár teljesen az alapokról a kívánt nyelvi kódokkal felépíthető egy működő nyelvelemző rendszer. A NooJ tehát készen szállítja a nyelvészeknek a szükséges szoftvertechnológiai infrastruktúrát. Amint a fentiekben rámutattunk, a vele kifejlesztett nyelvelemző alkalmazások, esetünkben a részleges gépi fordítás, akár rapid alkalmazásfejlesztő munkaeszközként, akár önálló alkalmazásként hasznos tud lenni.

3. A NooJ eszköztára gépi fordításra

Ebben a részben összefoglaljuk a NooJ rendszer azon új funkcióit, amelyek lehetővé teszik a gépi fordításra történő alkalmazását. Mint ismeretes, a NooJ elődje, az INTEX rendszer, teljesen egynyelvű alkalmazás volt, egyszerre csak egy nyelvvel lehetett dolgozni, amelyet a munkamenet elején meg kellett adni. A NooJ kezdettől fogva törekedett a többnyelvűség támogatására. Már eddig is nemcsak számos file formátumot, de nyelvet is kezelt a rendszer. Mégis valójában csak a legutóbbi időkben valósult meg a nyelvek közötti átjárhatóság. Ennek alapja a kétnyelvű lexikon.

3.1. A lexikon

Minden fordítórendszer legfontosabb eleme a kétnyelvű szótár. A NooJ valójában a létező szótárakat kapcsolja egybe. A forrásnyelvi szótárban az idegennyelvi megfelelőket a szemantikai jegyekhez hasonló módon adhatjuk meg. Az 1. ábra részletet mutat be egy magyar-angol szótárból: a címszó és a szófaj mellett a ragozási útmutató (`+FLX` jegy) valamint az angol megfelelő (`+en` jegy) található. Figyeljük meg, hogy a többjelentésű szavak minden külön jelentése önálló bejegyzést kap a szótárban. Ennek megfelelően a forrásnyelvi szöveg kezdetben több értelmezést kap. Ez a notáció azt is lehetővé teszi, hogy egyéb célnyelveken is megadjuk a megfelelőket, íly módon tetszőleges n-nyelvű szótárat készíthetünk.

Nagyon fontos elv, hogy egy szótári elem több szóból is állhat. Minden olyan többszavas kifejezést, amelynek minden eleme és azok sorrendje rögzített, a szótárban a legcélszerűbb szerepeltetni. A leghosszabb megfelelő (*longest match*) elve gondoskodik arról, hogy ilyen esetben a többszavas kifejezés illeszkedik először a szövegre. A fordító rendszer hatékonyságát és robosztusságát döntő mértékben megszabja a szótárban található többszavas kifejezések mennyisége. A többszavas kifejezések mellett a NooJ rendszer természetesen megtalálja az azokat alkotó egyedi szavakat is, így tehát a `házi feladat` kifejezés mellett a rendszer egyenként is elemzi a `házi` és a `feladat` szavakat is, ami adott esetben feleslegesen növeli a szöveg többértelműségét. Ha le akarjuk tiltani az ilyen típusú többes elemzéseket, használhatjuk a `+UNAMB` jegyet, ami letiltja ugyanennek a szónak vagy szókapcsolatnak egyéb értelmezéseit.

```
ernyő,N+FLX=3AD1+en=umbrella
autó,N+FLX=2A4+en=car
iskola,N+FLX=1A3+en=school
kormány,N+FLX=1A5b+en=government
kormány,N+FLX=1A5b+en=steering wheel
házi feladat,N+FLX=1A7b+en=homework+UNAMB
```

1. ábra. Az idegennyelvi megfelelő megadása a NooJ szótárban

Két szótár összekapcsolása természetesen legfeljebb egy szótárprogram számára elégséges. Egy fordítóprogram számára minimális követelmény, hogy mindkét szótár segítségével szóalakok morfológiai elemzése és generálása is lehetséges legyen. A morfológiai elemzés mindig is része volt az INTEX/NooJ rendszernek. Több módon is megoldható volt, kezdetben a szóalakok elemzésükkel együtt szerepeltek a szótárban, amelyet egy szótőlista és paradigmák segítségével maga a rendszer állított elő. Ez az út nem adott lehetőséget az alakok generálására. A NooJ újabb változatai inflexiós vagy derivációs file-okat használnak, amelyek reguláris kifejezésekkel definiált transzdúszerek. Ez a megoldás nemcsak elemezni, hanem generálni is képes alakokat. Az 5. ábrán láthatjuk azt is, hogyan képes a célnyelvi megfelelők helyes alakjait előállítani.

3.2. Lokális grammatika

A NooJ rendszer egyik vonzó jegye a véges állapotú transzdúszerek könnyű kezelhetősége, amelyet egy jól kezelhető grafikai felület biztosít. E rövid áttekintésben a lokális grammatikák NooJ-beli implementációjának olyan új jegyeit említem, amelyek különösen hasznosak a gépi fordítás számára.

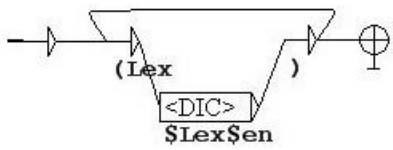
- változók használata
- hivatkozás lexikai jegyekre
- jegyek hozzárendelése az annotált egységekhez
- lexikai jegyek örökítése
- lexikai megkötések
- komplex változók használata lexikai megkötésekben

A változók használata elengedhetetlen ahhoz, hogy a lokális grammatika ne csak teljes egészében a bennük meghatározott egyedi szavakra legyen érvényes. A lexikai jegyekre való hivatkozás teremti meg a lehetőségét, hogy a transzdúszer kimenetében a szótári egység jegyeként számontartott célnyelvi megfelelőt alkalmazzuk, ráadásul a megfelelő toldalékolt alakban. Mivel a fordítási megfelelések tipikusan nem szavak hanem szintagmák között állnak fel, fontos, hogy a szintagmát is jegyekkel láthassuk el valamint, hogy a fej jegyeit is be tudjuk emelni a szintagma jegyei közé.

Terjedelmi korlátok miatt nem térhetünk ki az egyes funkciók részletesebb ismertetésére, a 4. részben bemutatott gráfok illusztrálják használatukat.

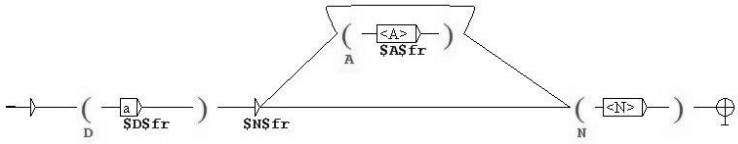
4. Előzetes eredmények

Miután áttekintettük az eszköztárat, tekintsük az alapesetet, hogy miként lehet a NooJ rendszerben egy direkt fordítórendszert megvalósítani?



2. ábra. Szó szerinti fordítás angolra

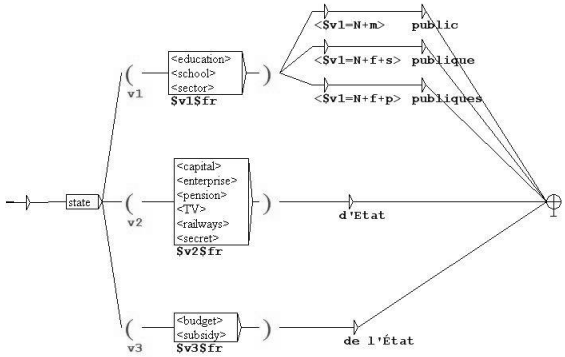
A 3. ábrán látható transzdúszer a Lex változóban tárolt tetszőleges szótári egység (`<DIC>`) angol célnyelvi megfelelőjét adja kimenetként, melyre a `$LEX$en` komplex változó alakjában hivatkozunk (feltételezve az 1. ábrán található szótári kódolást). Erre a megoldásra természetesen csak végső soron, minden egyéb lehetőség kimerítése után hagyatkozunk. Jellemzőbb azonban a szintaktikai szerkezet függvényében meghatározni a fordítási megfelelőt. Noha a 3 ábrán taláható minta akár pontos angol megfelelőjét tudná nyújtani *„a magyar állami oktatás”*



3. ábra. Egy egyszerű magyar főnévi csoport fordítása franciára

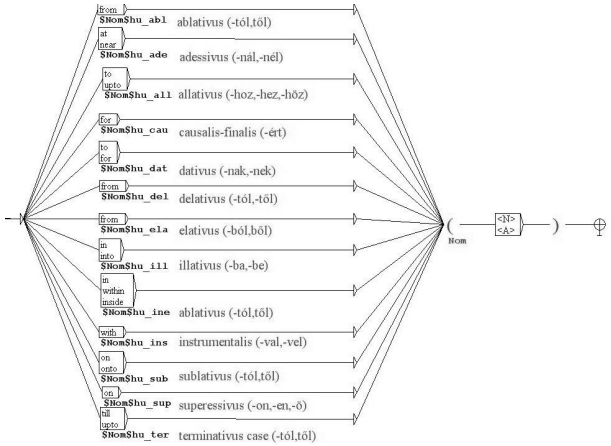
kifejezésnek, ugyanez franciára az eltérő sorrend miatt már csak a 3 ábrán látható lokális grammatikával leheetséges.

Ez a megfeleltetés is hamar túl elnagyoltnak bizonyul több okból. Nem tartalmaz például semmilyen finomabb disztribúciós megkötést a jelzők sorrendjére vonatkozóan. Ha ez mindkét nyelvben azonos, akkor természetesen nem is kell vele foglalkoznunk, hiszen egy fordítórendszer nem önmagáért való célként elemzi a forrásnyelvet, hanem csak olyan szintig, ameddig az releváns a két nyelv eltérései szempontjából. A 3 azért is fogyatékos volt, mert nem tartalmazta a franciában létező nembeli egyeztetést sem. Ennek megvalósítására a 4 ábra jobb fölső sarkában találunk példát. Az ábra nagyobb része a *„state”* angol szó francia megfelelőit adja három listával definiált kontextus függvényében. Ebben az eggyszerű esetben a szótárban is listába vehettük volna a kéttagú kifejezéseket. Könnyen elképzelhetünk azonban olyan gráfot, ahol általánosabb, szintaktikai `<N>` vagy szemantikai `<+bodypart>` jegyekkel határozzuk meg az odaillő lexikai elemeket.



4. ábra. Lexikai és morfológiai szelekciós megszorítások

Magyar-angol, magyar francia vonatkozásában gyakori eset, hogy egy szótári egységnek, tipikusan prepoziciónak, egy kötött morféma, azaz esetrag felel meg. Ilyenkor a szótári megfeleltetés nehézkes, és nehéz általánosan érvényes grammatikát találni. Az 5 ábrán található durva megoldást is csak mindegy egyéb lehetőség kimerítése után alkalmazzuk.



5. ábra. Angol prepozició magyar esetrag megfelelés alapesetei

Az eddigi ábrákon bemutatott esetek túl általánosak és elnagyoltak voltak, inkább csak a NooJ lehetőségeit illusztráló példaként szolgáltak. A lokális grammatikák alkalmazásának legfontosabb terepe ott van, ahol a minták részben lexikalizáltak, részben nyitott lexikai osztályokat tartalmaznak, és ezért szótári listázásuk vagy nem célszerű vagy lehetetlen is. Ilyen lehet a névkifejezések zöme (már ahol egyáltalán szükség van bármilyen fordításra) különösen például a dátumok, földrajzi helymeghatározások, időpont kifejezések.

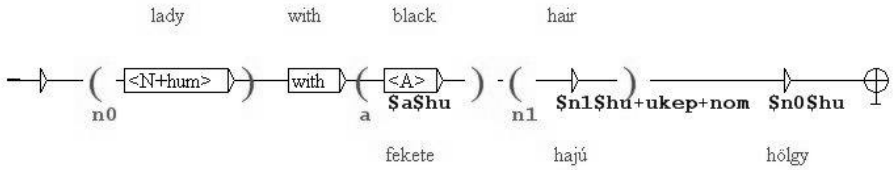
Az 1 táblázat csak utalásszerűen tartalmaz további olyan megfeleléseket magyar és angol főnévi szerkezetek között, amelyekben tipikusan lexikai szelekciós szabályok uralkodnak és a NooJ rendszer fent ismertetett eszközeivel jól kezelhetők.

`N with A N`	*girl with black hair*	`A N-Ú N`	*fekete hajú lány*
`A-speaking N`	*Spanish-speaking students*	`A nyelvű N`	*spanyol nyelvű diákok*
`N of N`	*freedom of assembly*	`A N`	*gyülekezési szabadság*
`N (Adv) Adv`	*house immediately opposite*	`A N`	*közvetlen szemközti ház*
`N P N`	*people at the reception*	`A N`	*a fogadáson lévő emberek*

1. táblázat. lokális szerkezeti megfelelések magyar-angol főnévi csoportokban

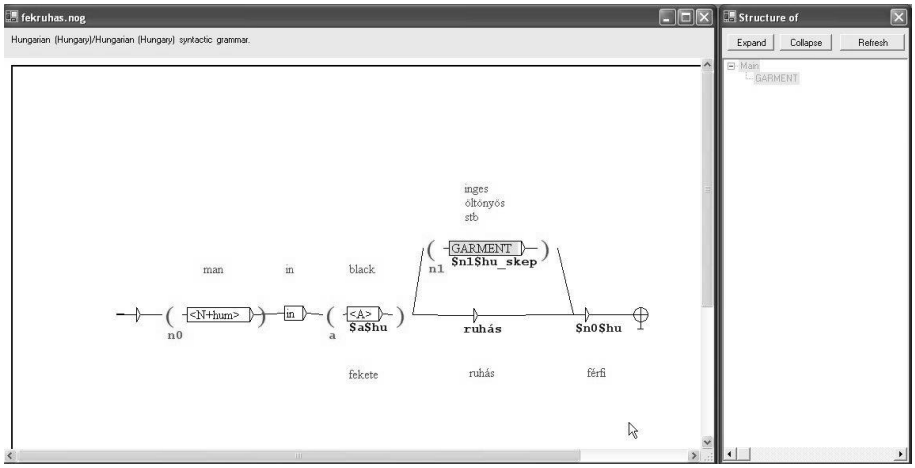
Példaként tekintsük a táblázat első sorában látható megfelelést. A 6. ábrán található az a vázlatos lokális grammatika, amely megvalósítja az angol kifejezés magyarra fordítását. Könnyedén meg tudjuk oldani a változókban tárolt angol lexikai egységek magyar megfelelőinek, beleértve az *-Ú* képzős alaknak az előállítását és helyes sorrendbe rendezését. Figyeljük meg, hogy az egyszerűség kedvéért csak azzal az egy szemantikai megkötéssel éltünk, hogy a *n0* elem `+hum` jeggyel rendelkezzen, az *n1* elemre nézve nem írtunk elő semmilyen megkötést. Márpedig ez helytelen alakhoz vezet akkor, ha *n1* nem testrészt vagy egyéb elidegeníthetetlen birtokot illetve inherens tulajdonságot jelent. Ez esetben ugyanis

az *-Ú* képző helyett a *-Vs* képzőt kell alkalmaznunk (*man with a black umbrella*:*fekete esernyős férfi*).



6. ábra. Angol-magyar módosító szerkezetek

Szemantikai megszorításokat többféle módon tehetünk: a lexikai elemek felsorolásával, beágyazott gráfok illetve szemantikai jegyek segítségével. A 7. ábrán azt mutatjuk be, hogyan lehet beágyazott gráfokkal szemantikai alosztályokat képezni. A GARMENT nevű algráf jelen esetben nem más, mint a ruhadarabok diszjunktív halmaza, amelyet ebben a szélsősegesen egyszerű esetben természetesen a főgráfban is megadhattunk volna. De még ebben az esetben is célszerű volt így eljárnunk, mivel az algráfok újrafelhasználása útján nagyobb modularitást érhetünk el, az algráf nevének alkalmas megválasztásával pedig a főgráf áttekinthetőségét és a rendszer karbantarthatóságát növelhetjük. Nem is beszélve arról az esetről, amikor a beágyazott gráf szerkezete lényegesen bonyolultabb, melyet érdemes is rejtve tartani a magasabb szintek elől.



7. ábra. Megfelelések szemantikai osztályokkal kifejezett megkötésekkel

5. Összegzés és további feladatok

A NooJ nyelvi fejlesztő rendszer mára kialakult eszköztára komoly szoftertechnológiai segítséget ad komplex nyelvi elemző rendszerek kifejlesztéséhez. Ezek között reális vállalkozásnak tűnik a részleges gépi fordítás megvalósítása. A maximális kiterjesztésű főnévi csoport több szempontból alkalmas célpont a fordításhoz. Az igeneves illetve prepozíciós módosító szerkezetektől eltekintve belső szerkezetében döntően lokális függőségi viszonyok dominálnak, amelyek véges állapotú grammatikával jól kezelhetőek. Nem idiomatikus, nem metaforikus stb. használatban szemantikailag is olyan önálló egységet képviselnek, amelyek belső szerkezetük mégoly eltérő volta ellenére is várhatóan megfelelésben állnak egymással. Kétnyelvi alkalmazásokban elengedhetetlenül fontos például a névkifejezések, dátumok, időkifejezések stb. fordítása. Egyéb gyakorlati alkalmazásként megemlíthetjük például a nyelvoktatást, ahol bizonyos nyelvi jelenségek gyakoroltatásához szintén hasznos lehet egy ilyen részlegesen fordító rendszer.

A részleges fordítás célként tételezése nem feltétenül jelenti azt, hogy a vázolt eljárást inherensen alkalmatlannak tartanánk a mondat teljes szerkezetének a megragadására. Ezt a célt inkább egy kutatási program első állomásának tekintük. A további feladatok a mondat szintaktikai vázának megragadására irányulnak, melyekben az ige és vonzatkeretének az integrálása kapja a központi szerepet.

Hivatkozások

1. Steven Abney: Partial parsing via finite-state cascades. 2. évf. (1996) 4. sz., *Journal of Natural Language Engineering*, 337–344. p.
2. Steven P. Adney: Parsing by chunks. In Carol Tenny (szerk.): *The MIT Parsing Volume, 1988-89*. http://www.vinartus.net/spa/89d.pdf, 1989, MIT Press.
3. Maurice Gross: The construction of local grammars. In E. Roche – Y. Schabes (szerk.): *Fininte State Language Processing*. Cambridge, Mass., 1997, The MIT Press, 329–352. p.
4. Zellig S. Harris: *Papers on Syntax*. Synthese Language Library sorozat, 14. köt. Dordrecht:Holland, 1981, D. Reidel Publishing Co.
5. Svetla Koeva – Denis Maurel – Max Silberztein (szerk.). *Nooj pour la Linguistique et le Traitement Automatique des Langues* (konferenciaanyag). Presses Universitaires de Franche-Comté, 2006.
6. Gábor Prószéky: Machine translation and the rule-to-rule hypothesis. In Krisztina Károly – Ágota Fóris (szerk.): *New Trends in Translation Studies. In Honour of Kinga Klaudy*. Budapest, 2005, Akadémiai Kiadó.
7. Gábor Prószéky – László Tihanyi: Metamorpho: A pattern-based machine translation project. In *24th Translating and the Computer Conference, 19-24* (konferenciaanyag). London, 2002, 19–24. p.
8. Max Silberztein: *NooJ Manual*. http://www.nooj4nlp.net/NooJ