BARBARA GAWROŃSKA-WERNGREN

Identifiering av diskursrefenter vid maskinöversättning från ryska till svenska

Abstract

The problem of tracking discourse referents in the process of machine translation which will be discussed in this paper is a part of my work at SWETRA (Swedish Computer Translation Research) at the Dept of Linguistics, Lund. The SWETRA-programs, implemented in Prolog and based on a GPSG-inspired formalism called Referent Grammar, abbreviated RG (Sigurd 1987, 1988), allow translation of a large repertoire of syntactic constructions between Russian, Swedish and English. Currently, the research is concentrated on translation from Russian into Swedish. One of the difficulties arising in the process of translating from a Slavic language into a Germanic one is inserting correct definiteness values in the noun phrases, as e.g. Russian and Polish do not make use of definiteness as a regular grammatic category. Generating appropriate definiteness values and their morphological representations in the target language is a very complicated task, as the choice between definite and indefinite NPs depends not only on endophoric (textual), but also on many exophoric (external) factors. The most general rules for use of definite noun phrases in Swedish and in English are, nevertheless, accessible for implementation in Prolog.

Our approach to MT requires tracking discourse referents and the paper will also discuss some problems connected with this approach.

1 Begreppen diskursreferent och koreferens

Innan vi presenterar den (preliminära) algoritmen som används i SWETRA-program för att identifiera diskursreferenter och välja nominalfrasernas bestämdhetsvärde, vill vi kortfattat beskriva den diskursmodell, som proceduren bygger på och försöka definiera själva begreppet "diskursreferent". Informellt brukar diskursreferenter karakteriseras som "saker och fakta som man talar om", vilket implicerar en exoforisk (icke-textuell) relation — man refererar till objekt i den icke-språkliga verkligheten. Definitionen av begreppet diskursreferent måste dock samtidigt vara relaterad till texten (det gäller bl a att ägna uppmärksamhet

Proceedings of NODALIDA 1989, pages 170-182

åt sättet att introducera nya diskursreferenter i texten och till de lingvistiska faktorer som möjliggör associering av två fraser med samma diskursreferent). Den preliminära modellen för identifiering av diskursreferenter som kommer att presenteras nedan innehåller därför både endoforisk och exoforisk komponent; i den senare spelar begreppet "mental verklighet"(dvs en värld av kognitiva enheter och relationer mellan dem) en primär roll. Innan vi övergår till den referent-grammatiska diskursmodellen, vill vi kortfattad kommentera en klassisk definition av sättet att introducera diskursreferenter formulerad av Karttunen (Karttunen 1976):

> the appearance of an indefinite noun phrase establishes a discourse referent just in case it justifies the occurence of a coreferential pronoun or a definite noun phrase later in the text.

Ovanstående definition kan inte tillämpas i SWETRA:s översättningsprocedur, eftersom distinktionen mellan indefinita och indefinita nominalfraser är (vid översättning från ryska eller polska) omarkerad i inputtexten. Man måste dessutom ta hänsyn till det faktum, att en ny diskursreferent kan introduceras inte enbart av en NP, utan också av t ex ett verb eller en hel mening — som i (1):

(1a) Idag förföljde en oidentifierad ubåt en svensk trålare .
(1b) Jakten pågick i ungefär en timme.

eller av ett adjektiv — som i (2):

(2a) Köp inte gula blommor till henne.
(2b) Den färgen tycker hon inte om.

Den diskursmodell, som vi vill föreslå här, tar hänsyn till olika sätt att introducera nya diskursreferenter — bl a till dem som illustrerades i (1) och (2). Den preliminära modellen innehåller fyra nivåer (som i sin tur kan indelas i subnivåer; nedanstående beskrivning är i viss mån förenklad):

nivå 1: Texten.

nivå 2: Mentala koncept som uppstår på basis av lingvistisk kunskap (dvs kunskap om ordens intensioner och extensioner och förmågan att tolka syntaktiska strukturer); dessa koncept vill vi preliminärt indela i tre huvudgrupper — s k nominaliserare (nominalizers), egenskaper och relationer; skillnaden mellan dessa begrepp kan (förenklat) förklaras i predikatskalkylens termer: "relationer" och "egenskaper" kan jämföras med predikat, och "nominaliserare" med argument. Varje enhet på denna nivå kan associeras med flera ord eller ordsekvenser i texten.

nivå 3: Diskursreferenter, dvs kognitiva enheter som konstitueras på basis av de slutsatser som kan dras med utgångspunkt från de egenskaper hos nominaliserare och de relationer mellan nominaliserare vilka representeras på nivå 2. Slutsatserna kan vara baserade antingen på lingvistiska eller icke-lingvistiska kunskaper. En diskursreferent kan associeras med flera enheter på nivå 2 (koreferens).

nivå 4: Objekt och fakta i den "verkliga världen"; denna nivå behöver inte vara representerad i alla diskurstyper, eftersom objektets existens respektive icke-existens i den fysiska verkligheten inte påverkar möjligheten att konstituera diskursreferenter som i ett av exemplen i Frarud (Frarud 1986):

(3a) We don't have a dog.
(3b) He would only fight with the cat.

I enlighet med den ovan skisserade modellen uppstår koreferens om två (eller fler) mentala koncept framkallade av ord eller ordsekvenser kan associeras med samma kognitiva enhet på en högre nivå — en föreställning av t ex ett objekt, ett faktum, en egenskap etc. Förutsättningar för koreferens uppstår således på nivå 2 (begrepp baserade på rent lingvistisk kunskap); i fortsättningen kommer vi dock att — för att förenkla procedurbeskrivningen — ibland använda formuleringar som "koreferenta ord", "koreferenta substantiv" etc — i stället för den mer exakta, men alltför långa frasen: "ord associerade med koreferenta begrepp".

Ett av villkoren för potentiell koreferens (som stora delar av vår komputationella modell bygger på) kan formuleras på följande sätt: två koncept (C1 och C2) framkallade av ord eller ordsekvenser i texten (T1 och T2) kan i regel associeras med samma referent R, om det senare introducerade konceptet (C2) är inte ojämförbar med C1 och om C2 inte är mera specifikt än C1. Med "mera specifikt" menar vi följande relation: C1 är mera specifikt än C2, om föreställningen "att vara C1" implicerar (direkt eller indirekt) "att också vara C2". Resonemanget kan illustreras med exempel (4) och (5):

(4a) Jag träffade min granne.
(4b) Mannen var berusad.

Nominalfraserna *min granne* och *mannen* (eller, mera exakt, de "nominaliserare" som associeras med dem — låt oss kalla dem för N1 och N2) uppfattas som koreferenta eftersom det stereotypa konceptet av "granne" kan representeras som "en människa som bor nära talaren/den tidigare nämnda personen" (vi bortser för eventuell metaforisk användning av ordet i fråga). Att vara någons granne implicerar normallt att vara en manniska, vilket i sin tur implicerar att vara antingen en man eller en kvinna — N2 är alltså mindre specifik än N1. Däremot i en text som:

(5a) Jag träffade en man.
(5b) Min granne var berusad.

upplevs nominalfraser *en man* och *min granne* som icke-koreferenta, eftersom konceptet "min granne" är mera specifikt än "en man".

Distinktionen mer/mindre specifik är naturligtvis inte alltid lika klar; exempel (6) — ett fragment av SWETRA-översättning av en rysk tidningstext — visar ett mera problematiskt fall:

(6a) Israeliska flygplan utförde idag tre bombattacker över libanesiskt territorium.
(6b) Femton personer dödades som resultat av luftpiraternas barbariska aktion.

Det faktum, att begreppen "israeliska flygplan" och "luftpiraterna" interpreteras som koreferenta, kan möjligan förklaras på följande sätt: C1 – som kan representeras som `israeli(N1)`, där N1 motsvarar begreppet "flygplan" (begreppet inkluderar inte bara flygplan som maskiner men också deras "animata delar" – piloter) implicerar — i enlighet med de värderingar som är aktuella i sändarens kultur vissa negativa egenskaper (att vara en israelisk pilot kan innebära — med en hög sannolikhetsgrad — att vara också en luftpirat, en bandit osv). Vid en sådan interpretation skulle det ovan formulerade villkoret för möjlig koreferens vara uppfyllt. Ett försök att med hjälp av en komputationell procedur identifiera koreferens i liknande fall kommer att presenteras i avsnitt 2.1.

2 En preliminär modell för identifiering av diskursreferenter

SWETRA-program för översättning mellan vissa slaviska och germanska språk är — som det har påpekats tidigare — baserade på en GPSG-inspirerad formalism — referentgrammatik (Referent Grammar, RG; Sigurd 1987, 1988). Möjligheter att använda RG för parsing och översättning av vissa polska och ryska syntaktiska konstruktioner har beskrivits i Gawrońska-Werngren (1988) och i Sigurd & Gawrońska-Werngren (1988).

Proceduren som för närvarande används vid översättning från ryska till svenska kan indelas i följande tre huvudstadier:

1. parsing av input-meningen och formulering av en sorts interlinguarepresentation, s k funktionell representation (f-representation), som innehållller information om meningens syntaktiska struktur (uttryckt med hjälp av sådana traditionella termer som subjekt, objekt, predikat, adverbial och satsadverbial), ordens betydelsekoder (formulerade i "maskinengelska") samt koder för vissa grammatiska drag, som numerus, genus, värdet +/- animat, tempus osv. T ex en enkel rysk mening som

mal'čik	bežal	domoj
pojke	sprang	hem

   får efter analysen följande f-representation (förenklat):

   ```
   s(subj(np(r(_,m(boy,sg),D,sg,ani,ma,_,_),H,Rel)),
     pred(run,past),
     sadvl([]),sadvl([]),advl(home),advl([]),advl([])).
   ```

   Symbolen [] (tomma mängden) avspeglar det faktum att meningen inte innehåller några satsadverbial och inga fler adverbial än *domoj* (home). Prolog-atomen med funktorn `r` brukades i några tidigare skrifter om RG kallas för "referentbeskrivning" (referent deskription), vilket naturligtvis måste betraktas som en approximation; denna enhet innehåller snarare

en beskrivning av nominalfrasens huvudord, som under översättningsprocessens nästa etapp möjliggör identifiering av diskursreferenter. I fortsättningen kommer den delen av nominalfrasens representation (ofta betecknad med symbolen `R`) kallas för "referent nucleus". Variabeln `D` — bestämdhet — förblir oinstantierad under den första fasen av översättningsproceduren. Variablerna `H` och `Rel` används för att lagra information om eventuella attribut: `Rel` kan innehålla en funktionell representation av en relativsats, medan koder för övriga pre- och postnominala attribut placeras i enheten `H`.

2. betydelsekoder för syntaktiska konstituenter lagras i listor; därefter börjar sökning efter eventuella koreferenta fraser och analys av den tidigare textuella informationen; procedurens mål är att instantiera variabeln `D` som "def" (+definit) i de fall, då textuella faktorer implicerar bestämdhet. Meningens funktionella representation lagras också i databasen.

3. generering av inputmeningens ekvivalent i målspråket. På stadiet B sker en omformulering av meningens f-representation till en PROLOG-lista — en enkel transitiv sats får då följande form:

```
[subj(X),pred(Y),obj(Z),sadvl(S1),sadvl(S2),advl(A1),
                    advl(A2),advl(A3)]
```

Därefter jämförs varje nominal konstituent med tidigare översatta substantiv, verb och attribut. De tidigare översatta ordens betydelsekoder är samlade i två separata listor — listan som innehåller substantivens och verbens betydelsekoder kommer i fortsättningen att kallas för R-listan, listan i vilken attributiva bestämningar placeras — för A-listan. Skälet för den sortens indelning är icke-lingvistiskt - sökningsproceduren verkar helt enkelt fungera mera effektivt, om huvudordens koder samlas i en separat lista. Sökning efter koreferenta konstituenter och eventuella bestämdhetsindicer genomförs med hjälp av en rekursiv procedur, som terminerar när den funktionella representationen inte innehåller flera nominella konstituenter. Det ryska lexikonet som inkluderar en viss (mycket förenklad) information om ordens semantiska kategoritillhörighet är tillgängligt under denna del av proceduren.

I början av översättningsprocessen är både R- och A-listan tomma. Proceduren som används för att lagra den information, som senare kan utnyttjas för identifiering av diskursreferenter, kan beskrivas på följande sätt:

1 Är subjektsplatsen i den funktionella representationen tom? (subjektet representeras som en tom mängd t ex vid analys av ryska opersonliga konstruktioner av typen:

ubito *pjat' čelovek* — 'fem människor dödades'.
döda fem människor
+unpers

Om ja, gå över till f-representationens nästa konstituent (predikatet), lagra dess betydelsekod i R-listan och fortsätt till nästa konstituent

— tills en konstituent som innehåller en NP påträffas. Efter att ha hittat den första NP, gå över till 2. Om subjektsplatsen inte är tom, gäller det också att gå över till 2.

2 Kontrollera först, om enheten H innehåller några attributkoder. Om alla platser i H är tomma, placera huvudordets betydelsekod i R-listan och förse den med numret N1 = N + 1, där N = antalet element i R-listan. Gå över till 5. Om några attribut förekommer, gå över till 3.

3 Kontrollera om platsen avsedd för räkneord innehåller en konstant, t ex 2. Om inte, gå över till 4. Om ja, ändra huvudordets betydelsekod från `m(X,pl)` till `m(X,2)` och placera den i R-listan (med lämpligt indexnummer). Denna del av proceduren används för att möjliggöra generering av bestämd form i texter av typen: *Två pojkar sprang. Den ene var liten. Den andre* ... Gå över till 4.

4 Innehåller den sista platsen i enheten `H` en kod av typen `m(X,prop)`, dvs ett egennamn? (appositionsfall). Om ja — placera huvudordets betydelsekod och egennamnets representation i R-listan och förse bägge med samma nummer (för att hantera fall som t ex *professor Andersson*, där samma referent kan i fortsättningen åberopas antingen med hjälp av enbart egennamnet eller enbart titeln). Instantiera variabeln `D` (bestämdhet) som `def` (+bestämd). Om nominalfrasen innehåller andra attributkoder, placera dem i listan A. Gå till 5. Om enheten `H` inte innehåller något egennamn i apposition, men andra attribut, placera också deras koder i A-listan, placera huvudordets kod i R-listan och gå över till 5.

5 Om den funktionella representationen innehåller fler konstituenter, sök efter nästa NP och — efter att ha funnit den — upprepa proceduren fr o m punkt 2. Om den funktionella representationer inte innehåller flera nominalfraser, omformulera den funktionella representationen till dess ursprungliga form (en PROLOG-atom med funktorn s) och gå över till etapp C (generering av meningen i målspråket).

Den ovan beskrivna proceduren innehåller i praktiken flera stadier: bl a en delprocedur som möjliggör insättning av svenska possessiva pronomen framför beteckningar för släktskap och liknande relationer (typ *morfar*, *granne* etc) och en del andra substantiv som kräver förekomsten av ett possessivt attribut i svenskan, men inte i ryskan och polskan (en inputmening av typen *spotkalem sąsiada* skulle alltså översättas till svenska som *jag träffade min granne*, även om den polska nominalfrasen *sąsiada* (*granne*) inte innehåller något possessivt pronomen).

Delar av koreferenssökningsproceduren kommer att illustreras med några test-exempel (demos).

2.1 Exempel på koreferensidentifiering i SWETRA

Vi antar, att åtminstone en mening har blivit analyserad och översatt. R-listan (substantiv- och verbkoder) och möjligen också A-listan (attributkoder) innehåller således några element. Den första frågan som ställs när en NP påträffas i den aktuella f-representrationen är: kan något av de tidigare översatta orden associeras med samma referent som den aktuella nominalfrasen? Första steget i svarsproceduren är att undersöka, om den lexikala enheten som motsvarar den aktuella NP:s huvudord är subklassificerat som "relation" eller "egenskap". Om den lexikala enheten (rlex) innehåller konstanten "property" med eventuell vidare subklassificering (exempelvis "colour"), består procedurens nästa steg i att undersöka, om A-listan innehåller ett adjektiv som tillhör samma kategori; om så är fallet, kan konstanten "def" placeras i det aktuella substantivets karakteristik. Detta är naturligtvis en approximation; algoritmen borde berikas med en del restriktioner, men den ger i regel korrekta resultat vid översättning av enkla textfragment. En analog princip tillämpas om det aktuella substantivet är i lexikonet subklassificerat som "relation" eller "aktion" — i det här fallet söker programmet i första hand igenom R-listan och letar efter ett verb vars semantiska beskrivning i lexikonet skulle möjliggöra koreferens med den aktuella nominalfrasen. Den delen av proceduren möjliggör korrekt översättning av sekvenser som (1). Den ryska versionen av exempel (1) är följande:

(1a)	Segodnja	neopoznannaja	podvodnaja	lodka	presledovala	švedskij	trauler
	idag	oidentifierad	ubåt		följde	svensk	trålare
(1b)	Presledovanie	prodolžalos'	okolo	časa			
	jakt	pågick	ungefär	timme			

De ryska lexikonenheter som motsvarar det transitiva verbet och det verbala substantivet i texten har följande form (förenklat):

```
rlex(presledovat',m(chase,_),v,vt,inf,_,_,
                  [follow,chase,hunt],_,_,_,_).
rlex(presledovanie,m(hunt,sg),n,sg,ina,ne,nom,
                  [follow,chase,hunt],rel2,_,_,_).
v    --- verb
vt   --- transitive verb
inf  --- infinitive
n    --- noun
sg   --- singular
ina  --- inanimate
ne   --- neutrum
nom  --- nominative
rel2 --- 2-argument-relation
```

Om programmet inte hittar något koreferent adjektiv eller verb, börjar det söka efter en tidigare översatt koreferent NP. Det enklast fallet är naturligtvis koreferens mellan nominalfraser med identiska betydelsekoder. Om R-listan innehåller en kod som inte skiljer sig från den aktuella, återstår bara att undersöka

eventuella motindicier (programmet kontrollerar t ex om det aktuella ordet inte föregås av ett attribut av typen *annan*) och — ifall inga sådana finns — instantiera variabeln D som def. Den delen av algoritmen möjliggör generering av bestämd/obestämd form i exempel som (7) och (8):

Input:

(7a)	Južnee	Sajdy	pojavilsja	izrail'skij	samolet.
	söder om	saida	dök upp	israelisk	flygplan
(7b)	Samolet	prodvigaetsja	na	zapad.	
	flygplan	förflyttar sig	västerut		

Svensk output:

(7c) Ett israeliskt flygplan dök upp söder om Saida.
(7d) Flygplanet förflyttar sig västerut.

Input:

(8a)	Kakoj-to	samolet	pojavilsja	južnee Sajdy.
	något	flygplan	dök upp	söder om saida
(8b)	Potom	pojavilsja	drugoj	samolet.
	senare	dök upp	annan+ma	flygplan

Svensk output:

(8c) Något flygplan dök upp söder om Saida.
(8d) Senare dök det upp ett annat flygplan.

Om R-listan inte innehåller någon kod som är identisk med det aktuella ordets betydelsekod, fortsätter sökningsproceduren; det gäller bl a att ge svar på följande frågor:

- om den aktuella nominalfrasen har värdet +plural: innehåller R-listan åtminstone två enheter med samma betydelsesymbol (den första delen av meningskoden), men med värdet "sg" ? Eller har man tidigare påträffat åtminstone två enheter som bildar en mängd som kan vara koreferent med den aktuella frasen?

- innehåller R-listan en kod, som kan associeras med ett koncept ("nominalizer") som inte är ojämförbart och inte mer specifikt än det aktuella ordets "nominalizer"? Dessa delar av proceduren identifierar koreferens i sekvenser av typen:

Input:

(9a)	Mal'čik	vstretil	devočku.
pojke	träffade	flicka+ack	
(9b)	Rebjata	pobežali	domoj.
	barn	sprang	hem

Output:

(9c) En pojke träffade en flicka.
(9d) Barnen sprang hem.

Substantiven *mal'čik* (pojke) och *devočka* (flicka) är i lexikonet specifierade som subkategorier av "barn"; det PROLOG-predikatet som identifierar koreferens i fall som (9) är formulerat som:

```
coref(m(A,pl),Rlist):- hyponyms(m(A,sg),[H|T],Rlist),
      T/=[].
```

Ovanstående regel kan läsas på följande sätt: ett substantiv med betydelsekod `m(A,pl)` kan associeras med en mängd bestående av åtminstone två tidigare introducerade referenter, om åtminstone två tidigare nämnda substantiv kan tolkas som hyponymer till det aktuella substantivets singularform (symboliserad som `m(A,sg)`). Detta är en av de enklaste varianterna av coref-predikatet (i praktiken används en del ytterligare restriktioner; deras utformning kräver fortsatt arbete).

I ex (9) identifieras alltså substantiven *mal'čik* och *devočka* som hyponymer till den lexikala enheten med betydelsekod `m(child,sg)` och koreferensen upptäckts med hjälp av följande predikat, som rekursivt letar efter möjliga hyponymer till det aktuella ordets (här:*rebjata* — *barn*) singularform:

```
hyponyms(m(A,N),[m(B,N) | Rest],[r(_,m(B,N)) | Rest1]):-
            more_restricted(m(B,N),m(A,N)),
            hyponyms(m(A,N),Rest,Rest1).

hyponyms(m(A,N),Hyponymlist,[H|T]):-
hyponyms(m(A,N),Hyponymlist,T).

hyponyms(m(A,N),[],[]):-!.
```

Variabeln `Hyponymlist` (symboliserad i huvudpredikatet "coref" som `[H|T]`) betecknar naturligtvis en lista, i vilken eventuella hyponymer lagras under sökningsproceduren. Predikatet `more_restricted` är — i dess enklaste version — formulerat som:

```
more_restricted(A,B):- rlex(_,A,_,_,_,_,_,Features1,
              Features2,_,_,_),
              rlex(_,B,_,_,_,_,_,Features2,_,_,_,_).
```

Regeln säger, att begreppet associerat med en lexikal enhet med betydelsekoden A är mera specifik än begreppet associerat med B, om A har de semantiska drag (Features2) som anses vara mest karakteristiska för B, och dessutom åtminstone ett drag, som är mera specifikt. I ex (9) innehåller de lexikala enheterna för "pojke" och "flicka" specifika drag "male" resp "female" och dessutom alla drag som tillåter användningen av enheten "barn" och, följaktigen, tolkas de som hyponymer till samma ord. Formatet för lexikala enheter är i det här fallet följande (förenklad notation):

```
rlex(mal'čik,m(boy,sg),n,sg,ani,ma,nom,[male],
                          [child],_,_,_).
rlex(devočka,m(girl,sg),n,sg,ani,fe,nom,[female],
                          [child],_,_,_).
rlex(rebenok,m(child,sg),n,sg,ani,ma,nom,[child],
                          [human],_,_,_).
```

Det ovan visade fallet av koreferensidentifiering är naturligtvis mycket enkelt; att bygga upp en hierarki av semantiska drag som skulle fungera effektivt vid sökning efter hyponymer i mera komplicerade texter är ingen enkel uppgift; utformning av lexikala enheter i SWETRA är i detta avseende än så länge inte fullständig. Vid översättning av korta textfragment är det dock möjligt att identifiera en gemensam diskursreferent i fall som är mindre "självklara" än ex (9) — som t ex den tidigare diskuterade koreferensen mellan "israeliska flygplan" och "luftpirater". I den sortens fall används följande PROLOG-predikat:

```
possible_coref(A,B):- rlex(_,A,_,_,_,_,_,F1,_,_,_,_),
                      rlex(_,B,_,_,_,_,_,F2,F3,_,_,_),
                           evaluation_in(F2),
                           co_elt(F1,F2).
```

Predikatet `evaluation_in` innebär, att listan F2 (semantisk karakteristik) innehåller ett drag som är klassificerat som "värdering"; det enklaste sättet att i en komputationell modell identifiera koreferens mellan ett substantiv som är markerat i fråga om värderingskomponenten och ett omarkerat sådant är att betrakta värderingskomponenten enbart som ett uttryck för sändarens attityd, och inte som en faktor som påverkar referensrelationen. Eftersom de lexikala enheterna i vårt exempel innehåller gemensamma drag (i listorna F1 och F2) och programmet inte hittar några indicer mot koreferens, associeras både "israeliska flygplan" och "luftpirater" med samma referent. Nedanför visas — i en förenklad notation — de lexikala enheter som i det diskuterade fallet möjliggör identifiering av koreferensrelationen:

```
rlex(samolety,m(airplane,pl),n,pl,ina,ma,nom,
     [airplane,pilot],[machine,human],_,_,_).
rlex([vozdu$nyje,piraty],m(air_pirate,pl),n,pl,ani,
     ma,nom,[neg,pilot],[human],_,_,_).
```

Frågan, vilka och hur många gemensamma drag som behövs för att två "nominalizers" ska uppfattas som koreferenta, är naturligtvis komplicerad, och den aktuella versionen av programmet utesluter inte vissa fel och övergeneraliseringar.

Förutom de nämnda predikaten innehåller programmet också en preliminär procedur för identifiering diskursreferenter även när nominalfrasens huvudord är elliderat. Proceduren kräver vidare elaborering, men dess nuvarande utformning ger möjlighet att generera rätt artikel och kontrollera kongruens i sekvenser av typen:

Input:				
(10a)	Dva	samoleta	pojavilis'	južnee Sajdy.
	två	flygplan	dök upp	söder om saida
(10b)	Odin	prodvigaetsja	na zapad	
	en+ma	förflyttar sig	mot väst	
Output:				
(10c)	Två flygplan dök upp söder om Saida			
(10d)	Det ena förflyttar sig mot väster.			

Efter referentidentifiering följer översättningsprocessens sista etapp — generering av outputmeningen. Om variabeln `D` fortfarande är oinstantierad (som t ex vid översättning av textens första nominalfras eller i de fall, då ingen koreferent fras har upptäckts), genereras den svenska NP:s morfologiska form i enlighet med språkspecifika regler i programmets svenska modul. Den obestämda artikeln genereras alltså inte automatiskt när variabeln `D` inte har fått värdet def under etapp B — programmet undersöker först, om nominalfrasens huvudord är klassificerad som "unikt föremål" (som *solen*, *jorden* etc) eller som ämnesnamn, om det föregås av indefinita, deiktiska eller possessiva pronomen osv. Ogrammatiska konstruktioner som **en min bil* eller **en ingen bil* kan således inte genereras. Om variabeln `D` har blivit instantierad, är genereringsproceduren betydligt kortare — man behöver bl a inte undersöka, om substantivet i fråga är en benämning på ett unikt föremål, eller om det föregås av indefinita pronomen.

Enheten `R` ("referent nucleus") används i genereringsproceduren för att kontrollera kongruens inom nominalfrasen. Efter översättningsprocessen första fas (etapp A) innehåller enheten R följande konstanter:

- betydelsekod — t ex `m(boy,sg)`
- numerusvärde — här, sg
- det genusvärde, som substantivet har i källspråket
- värdet +/- animat

Återstående platser i enheten `R` kan senare fyllas med symboler för de grammatiska drag, som är nödvändiga för att åstadkomma kongruens i målspråket. Under genereringsprocessen söker programmet efter en svensk lexikal enhet med samma betydelsekod och — om variabeln `D` (bestämdhet) har fått ett konstant värde under etapp B — med samma bestämdhetskod som den aktuella nominalfrasen i f-representationen. Om t ex en nominalfras med huvudordets betydelsekod `m(boy,sg)` har fått värdet def efter referentidentifieringsproceduren, väljs följande lexikala enhet (om frasen inte innehåller attribut som utesluter användning av efterställd artikel):

```
lex(pojken,m(boy,sg),n,def,sg,re,ma,_,_,_,_,_).
re --- realis
```

och målspråkets genus- och sexusvärden överförs till "referent nucleus", som får följande form:

```
r(_,m(boy,sg),def,sg,ani,ma,re,ma).
```

Om frasen innehåller ett eller flera adjektivattribut, väljs den bestämda artikelns rätta numerus- och genusform enligt följande regel:

```
art(R,D) --> [X],R=r(_,_,_,D,N,_,_,G,_),
                  lex(X,D,art,N,G,_,_,_,_,_,_,_).
```

Regeln innebär, att kategorin "artikel" (art) med bestämdhetsvärde D kan realiseras som en ordform X om X har samma bestämdhets-, numerus- (N) och genusvärden (G) som nominalfrasens huvudord. Analoga regler används för att välja de övriga kongruerande attributens morfologiska former. Kongruenskontroll med hjälp av "referent nucleus" fungerar mycket effektivt, och några fel har i detta avseende inte observerats.

3 Sammanfattning

Identifiering av diskursreferenter i SWETRA:s program bygger på en fyra-nivå diskursmodell (1:ord och ordsekvenser, 2:begrepp baserade på rent lingvistisk kunskap, 3:diskursreferenter — kognitiva enheter som kan associeras med flera begrepp på nivå 2 på basis av analytiska slutsatser och kunskap om den icke-lingvistiska verkligheten, 4: objekt och fakta i den verkliga världen). Den aktuella algoritmen för identifiering av diskursrefenter med hjälp av den textuella informationen befinner sig på ett experimentellt stadium och bygger dels på den textuella informationen (betydelsekoder och meningarnas funktionella representationer lagrade i olika PROLOG-listor), dels på preliminära försök att representera stereotypa begrepp i lexikonet (listor över semantiska drag, semantisk subkategorisering). Programmet möjliggör för närvarande identifiering av koreferens i korta textfragment (5–6 meningar) vid relationer av typen: koreferens mellan identiska begrepp, koreferens mellan ett mera generellt begrepp och dess mera specifika antecedent (typ *granne — människan*), koreferens mellan referentmängder (*pojke* och *flicka — barnen*) samt mellan deras element (*två flygplan — det ena flygplanet*). Referentidentifiering är dessutom möjlig i vissa ellipsfall (*två flygplan — det ena*), vid koreferens mellan ett emotionellt laddat begrepp och dess neutrala motsvarighet (*israeliska flygplan — luftpiraterna*), samt i de fall, då koreferenta begrepp instantieras i texten med hjälp av kategoriellt olika fraser (t ex om diskursreferenten introduceras av ett adjektiv eller en hel mening). Generering av bestämdhetsvärde i målspråket sker primärt på basis av koreferensrelationer, därefter tillämpas språkspecifika regler (nominalfrasens form väljs med hänsyn till attributtyp och substantivens subkategorisering i det svenska lexikonet). Proceduren kräver vidare elaborering — de semantiska representationerna i lexikonet måste utvidgas, dessutom gäller det att åstadkomma ett mera generellt system för utformning av listor med semantiska drag. Databasen måste dessutom berikas med en förenklad representation av stereotypa kunskaper om den icke-lingvistiska världen (detta är en uppgift som kan förverkligas enbart i begränsad utsträckning och leder ur lexikonet in i encyklopedien). Det finns också ett behov av vidare studier kring de faktorer som implicerar valet av bestämd form i svenskan i de fall då ingen koreferensrelation föreligger (i många empiriska svenska texter har t ex den första nominalfrasen bestämd form). Den komputationella modellen som presenterats här har alltså en rad begränsningar, men vid översättning av korta texter med ett begränsat antal diskursreferenter fungerar den tämligen effektivt. Procedurens utveckling är föremål för fortsatt arbete inom SWETRA.

Litteratur

Frarud, Kari. 1986. The introduction och maintenance of discourse referents. *Papers from the Ninth Scandinavian Conference of Linguistics.* Stockholm.

Gawrońska-Werngren, Barbara. 1988. A Referent Grammatical Analysis of Relative Clauses in Polish. *Studia linguistica* 42(1):18-48.

Karttunen, Lauri. 1976. Discourse Referents. *Syntax and Semantics*, vol. 7:383–386. New York: Academic Press.

Sidner, Candace L. 1983. Focusing in the Comprehension of Definite Anaphora. Brady, M. & Berwick, R. C. [eds.] *Computational Model of Discourse*:267–330. MIT Press, Cambridge, Mass.

Sigurd, Bengt. 1987. Referent Grammar (RG). A generalized phrase structure grammar with built-in referents. *Studia linguistica* 41(2):115–135.

Sigurd, Bengt. 1988. Using Referent Grammar (RG) in computer analysis, generation and translation of sentences. *Nordic Journal of Linguistics* 11:129–150.

Sigurd, Bengt, & Gawrońska-Werngren, Barbara. 1988. The potentials of SWETRA, a multilanguage MT-system. *Computers and Translation* 3:238–250.

Lund University
Dept of Linguistics and Phonetics
Helgonabacken 12
Lund, Sweden