BENGT SIGURD

Erfarenheter av Swetra—ett svenskt MT-experiment

Abstract

Swetra is primarily a research project, but its computer programs and reports may result in commercial products. The research group at Lund consists of a few persons only, which has advantages and disadvantages. The main languages studied are Russian, English and Swedish. The grammatical model is Referent Grammar (RG) and the project is a spin-off of the development of that grammar. Referent Grammar is inspired by generalized phrase structure grammar (GPSG) and written directly in Prolog (DCG), which is why the grammars can run directly on computers. Arity Prolog and PC's are used. The work within Swetra consists mainly of writing grammar modules and lexicons for different languages. These modules are connected in automatic translation. Referent grammars are bidirectional and can be used both for analysis and synthesis (generation). A certain number of transfer rules are also needed in translation, however, in order to make necessary changes of functional representations, add features such as definiteness when translating from Russian into English etc. The paper gives a survey of the experience from Swetra research so far.

1 Inledning

Swetra (Swedish Computer Translation Research) har pågått cirka två år, och det finns en del erfarenheter att förmedla—erfarenheter som andra som ger sig ut på maskinöversättningens gungfly kan ha glädje av att känna till. Swetra, som stöds av Svenska Forskningsrådet för Samhällvetenskap och Humaniora, är ett minimalt projekt ifråga om ekonomiska, personella och maskinella resurser. I projektet arbetar—förutom undertecknad—Mats Eeg-Olofsson (halvtid), Lars Gustafsson (kvartstid), Barbara Gawrońska-Werngren (halvtid). Programmeringen har främst skett på PC med användande av Arity Prolog.

Jag skall meddela mina erfarenheter av projektet under följande huvudrubriker: Organisation, Vägval (texttyp, språk, inställning till syntax, grammatikmodell), Vad Swetra kan, Publiktrycket.

Proceedings of NODALIDA 1989, pages 229-236

2 Organization

Jag tror att en liten forskargrupp lokaliserad på en plats har stora fördelar. Man vet vad de andra gör och behöver inte resa runt eller samlas till stora seminarier. Ledningen förenklas. Det räcker ofta med några ord till de andra för att de skall vara med på noterna. Uppenbarligen säger jag detta mer eller mindre omedvetet med tanke på Eurotra, där medarbetarna synes ha tvingats använda mycket tid på att resa för att informera varandra och komma överens om vilka lingvistiska teorier, programmeringsspråk, format och datorer man skall använda. Den lilla tätt sammansvetsade gruppen har säkert stora fördelar när det gäller att utveckla den grundläggande teorien och prototypen, men när det sedan gäller att implementera teorien och gå från prototyp till produkt, att skriva stora lexikon, att förbättra programmen och att testköra systemet då har större arbetsgrupper säkert fördelar.

Vi hade i själva verket inget val i Lund, när vi började pröva på maskinöversättning. Jag hade länge talat och skrivit om hur svårt maskinöversättning är utan att själv ha någon riktig datorerfarenhet av det—bara pappersspekulationer. Swetra är en biprodukt av grammatikforskning och uppstod sedan jag insett att det gick att använda referentgrammatik för översättning. Swetra är snarast ett försök att se hur långt man kan komma med små medel. Vi ägnar föga tid åt att visa vad man inte kan klara eller bevisa att maskinöversättning är omöjligt, mest tid åt att visa vad man kan klara.

Anpassningen till PC bestämde vi oss tidigt för. Den underlättar flyttning och demonstration av programmen. Vi tänker nu också använda snabba Macintoshdatorer som också kan läsa DOS-disketter och möjliggör flyttning mellan olika miljöer.

3 Vägval

3.1 Texttyp

Det är nödvändigt att göra ett antal vägval när man går in i maskinöversättningsbranschen. Flera av dessa val hör ihop. Man måste välja mellan att försöka sig på att översätta godtycklig text (fritext) eller text inom något specialområde som teknik (t.ex. bilar, datorer), medicin, affärskontrakt, väderleksrapporter, etc. Inledningsvis träffade vi inte något sådant val—man var glada att det gick att översätta några meningar överhuvudtaget. De meningar vi först översatte var typiska lingvistmeningar, typ "En flicka kom", "Pojken slog hunden som sprang", "Hunden, som pojken, som flickan kände slog sprang". Efter att ha satt upp vad vi tyckte var syntaktiska regler som borde klara grundläggande strukturer försökte vi pröva dem på empiriska texter, bl.a. en liten Greenpeace-text om kärnkraftverk i Sellafield som fortfarande släpper ut plutonium och sedan några notiser från Pravda. Mötet med empiriska texter var omtumlande; man inser att teoretiska lingvister lever i ett reservat (med ett svenskt uttryck: en skyddad verkstad). Varenda mening ledde till att vi fick revidera grammatiken, ändra eller lägga till regler och inte bara utöka lexikonet som man ju hoppas.

Man kan tänka sig att rita diagram som visar ökningen av de grammatiska reglerna och lexikon för varje ny mening som skall översättas. I början måste mötet med en ny mening nödvändigtvis leda till stor ökning både av grammatiken och lexikonet, men kurvan bör så småningom plana ut, när nästan alla meningar klaras av utan att någon förbättring behöver göras. Ett färdigt system bör klara allt och inte kräva mera utbyggnad. Om det är ett interaktivt system bör det inte fråga operatören alltför ofta. Jag kan inte säga att vi kommit så långt i Swetra att vi känner att kurvan håller på att plana ut; varenda liten text vi försöker oss på kräver flera komplettringar—och det är inte alltid bara fråga om att sätta in fler ord i lexikon.

Det val i fråga om texter vi avser att träffa innebär att vi säger att Swetra inte kan översätta godtycklig text, bara speciell text, men vi vill specificera vilken speciell text Swetra kan översätta på ett lingvistiskt sätt. Vi vill inte specificera dessa texter genom hänvisning till en genre som t.ex. tekniska manualer, väderleksrapporter, utan ge specifikationen mera lingvistiskt och säga att meningarna inte får innehålla samordnade relativsatser, inte mer än två satsadverbial och tre andra adverbial, inte inbäddade genitiver som "pojkens hunds svans" etc. Lexikalt kan specifikationen göras genom att man t.ex. säger att orden måste tillhöra den marina sfären och röra sig om olika typer av båtar som rör sig i Östersjön.

I själva verket borde nog många MT-system ha specificerat sina begränsningar på detta sätt. Så vitt jag förstått av de demonstrationer Eurotra företagit i Brüssel har Eurotra egentligen också gått på denna linje: Deltagarna fick bara föreslå meningar som hade högst 7 ord, bara ett adjektiv i nominalfrasen, endast subjektiva relativsatser, inga samordningar etc. Det subspråk som Swetra vill kunna översätta definieras alltså i första hand genom vissa syntaktiska begränsningar. Sedan är det en självklarhet att orden måste finnas i lexikonet för att översättningen skall fungera. Jag skall sedan visa typiska Swetra texter som vi brukar köra i vår Demo (se också Sigurd & Gawrońska-Werngren 1988) .

3.2 Språk

Självklart måste man välja vilka språk man vill översätta mellan, men märkligt nog har vi tvekat länge på denna punkt i Swetra. Det beror naturligtvis på att vi åtagit oss att forska i maskinöversättning, inte att leverera ett färdigt program. Det är i själva verket intressant att pröva hur bra den grammatiska modellen referentgrammatik fungerar på olika språk. Vi har erfarenhet av översättning till och från franska, polska, georgiska, samoanska, svenska, engelska, ryska. På senare tid har vi alltmer koncentrerat oss på ryska, engelska, svenska och främst på översättning från ryska till svenska—sedan Barbara Gawrońska-Werngren ställde sina språkkunskaper till förfogande. De olika referentgrammatiska modulerna som programmeras direkt i Prolog (DCG=Definite Clause Grammar) är i princip bidirektionella och kan köras både i analys och syntes (generering). Men vill man göra mera avancerad översättning, använda specifika transferregler, anknyta diskurssemantiska procedurer m.m. och göra ett effektivt program är det bäst att bestämma sig för en riktning.

Vi har bedömt översättning mellan ryska och engelska som intressant därför att detta par länge intresserat MT-forskarna. Ett framgångrikt system som kan översätta mellan ryska och engelska har uppenbara praktiska tillämpningar. Språkens typologiska skillnader gör översättning också lingvistiskt intressant. Eftersom många personer inte kan ryska blir demonstrationer dessutom mera imponerande. Åskådarna blir mera förbryllade av att se hur en obegriplig text översättes till en begriplig än att se t.ex. engelska, som de flesta kan, översättas till svenska. (Å andra sidan blir åskådarnas möjligheter att bedöma översättningens kvalitet mera begränsade).

3.3 Inställning til syntax

Ett annat vägval rör inställningen till syntax. Många—i synnerhet tidigare—system betraktar ordöversättning som den primära operationen vid översättning. Man försöker sedan sekundärt se till att ordens kongruensböjning stämmer och att ordföljden är rätt genom att försöka identifiera vilka ord som hör ihop i fraser och vilka ord (fraser) som är subjekt, predikat, objekt och adverbial—något som ofta är ett minimikrav för att klara böjning och ordföljd. Swetra har inte ord som primära enheter, utan satser och fraser i olika funktioner (såsom subjekt, objekt, predikat, satsadverbial, andra adverbial, attribut). Swetra är på detta sätt en typisk produkt av sin tid—den tid i lingvistiken då syntax skriven såsom generativ grammatik står i fokus. Jag skall förklara detta närmare senare när jag beskriver den grammatiska modell (Referentgrammatik) som Swetra hela tiden arbetat med.

Ett system som prioriterar ordöversättning kan ofta översätta många sorters texter snabbt givet ett stort lexikon, men kvaliteten blir lidande om man inte håller reda på satsstrukturen i detalj. Förvånande många tvetydiga ord blir entydiga om man tar ut satsdelarna—standardexemplet är: Var var det var? Böjning av orden i målspråket ger sig liksom ordföljden om man har detaljerade upplysningar om ordens funktioner i fraser och satsdelar. Å andra sidan är satslösning (parsning) tidsödande, och alla problem kan inte lösas med syntax—ökända är prepositionsfraserna vars anknytning till verb eller nominalfras ofta endast kan avgöras på semantisk väg. När Swetra går grammatikvägen, innebär det att man endast accepterar och översätter meningar vars grammatiska struktur systemet har regler för att identifiera. Ett lexikonorienterat system kan alltid föreslå en översättning även om systemet inte har gjort någon detaljerad satslösning.

3.4 Grammatisk modell

En modern lingvist har till synes flera väl genomdiskuterade formella grammatiska modeller att välja på, t.ex. Transformationell generativ grammatik med frasstruktur och transformationer (EST, Government and Binding), Lexical-Functional Grammar (LFG), Generaliserad frastrukturgrammatik (GPSG), Dependensgrammatik, Funktionell Grammatik av Hallidays typ. Det är påfallande att dessa grammatiker nästan bara har prövats på typiska lingvistmeningar

av ovannämnda typ ("En flicka sprang", "Hunden, som, pojken, som flickan kände slog sprang"). Man ser t.ex. aldrig en GB-grammatiker visa hur alla meningarna på en viss sida text skall analyseras, vilka träddiagram man bör rita och vilka problem man möter. Gängse teoretiska grammatikmodeller visar främst hur vissa formella idéer som t.ex. frasstruktur, transformationer, dependens, subjacency, c-command fungerar. De har inte det primära syftet att visa hur vanliga meningar skall analyseras så att beskrivningen förklarar varför orden har den form de har och står i den ordning de gör. Allra minst förklarar de varför den föreliggande meningen betyder vad den gör.

Såsom framgått var det inte så att medarbetarna i Swetra bestämde att de skulle översätta med dator mellan två bestämda språk och sedan började leta efter en lämplig grammatik. I stället var det så att jag höll på med de pilregler som DCG erbjuder och insåg att man som resultat av analysen (satslösningen, parsningen) kunde få en sorts universell semantisk representation: en predikatslogisk formel eller en funktionell representation, en representation som talar om vad som är subjekt, predikat, objekt, adverbial, etc. och vad orden betyder. (Chomskys klassiska generativa grammatikregler ger endast ett S som resultat av analysen, vilket bara säger att satsen var grammatisk). Sedan jag skrivit en sådan grammatik för svenska och en för engelska insåg jag att man måste kunna översätta via den gemensamma funktionella representationen om man standardiserade den. Det är i princip det vi hållit på med sedan i Swetra. Den funktionella representationen fungerar som ett mellanspråk, *interlingua*.

Den grammatiska modellen för Swetra var alltså given, men samtidigt som vi arbetat med översättningsprogrammen har vi utvecklat referentgrammatik och tagit beslut som standardiserat den och framförallt dess funktionella representation och lexikonformat. Referentgrammatik (RG, se referenser i litteraturlistan) är en sorts generaliserad frasstrukturgrammatik (GPSG), men i motsats till GPSG arbetar RG i de generativa reglerna med två representationer: den ytliga kategorirepresentationen betecknad o-representationen, och den funktionella representationen betecknad f-representationen. Dessa beteckningar påminner om dem som användes i lexical-functional grammar (LFG), en grammatik som nog också kan användas vid automatisk översättning.

Referentgrammatik är bidirektionell, dvs kan användas både i analys och syntes (generering). Det betyder att den ställer samma hårda krav på sitt input och sitt output. En korrekt skriven RG-modul genererar bara korrekta meningar och accepterar bara korrekta meningar. För att en grammatik skall vara användbar för MT krävs att den kan specificera korrekthet på alla nivåer, genererar korrekta former av artiklar, räkneord, pronomen, adjektiv, substantiv, verb, sätta orden i rätt ordning, etc. RG gör det—den genererar t.ex. också korrekta former av relativpronomen i ryska och polska en uppgift som kräver att flera faktorer tas hänsyn till. RG är dessutom lätt att skriva för en lingvist och vi har inte funnit något som den inte kan hantera. Den klarar också s.k. *unbounded dependencies*, dvs flyttningar som kan vara hur långa som helst.

RG använder liksom GPSG defekta syntaktiska kategorier, men de formaliseras inte som "slash-kategorier" i RG utan benämns `sdsent` = subjektsdefekt sats, `odpp` = objektdefekt prepositionsfras etc. I programmet flyttas den saknade

konstituenten till rätt position, vilket gör att man kan säga att RG har dolda transformationer, precis som GPSG.

För att en grammatikmodell skall kunna användas i maskinöversättning måste den också ha tagit ställning till hur ordbetydelser skall beskrivas och vilket format de skall ha i lexikonet. Många teoretiska modeller lämnar den frågan öppen, men det duger inte i ett MT-system där alla komponenter måste finnas och kunna samverka. RG har ett bestämt lexikonformat (se artiklar i litteraturlistan) och lexikonet är en särskild fil på vilken man kan låta diverse operationer arbeta. Ordbetydelsen beskrivs f.n. i en engelskliknande form (*machinese*), men lexikonet har utrymme för många grammatiska och semantiska uppgifter om orden. RG har också utvecklat vissa procedurer för morfologi (ordböjning och ordbildning) och tillämpat s.k. implikationell morfologi, en modell där den morfologiska kunskapen ses som kunskap om vissa ordformer kompletterad med kunskap om hur existensen av en viss form implicerar existensen av en viss annan form. Om man kan en form i språket och vet vad den betyder, kan man dra slutsatser om hur många andra former bör se ut och vad de betyder.

Den grammatiska diskussionen de senaste decennierna har p.g.a. Chomskys starka inflytande mycket berört universella syntaktiska principer för naturliga språk, t.ex. begränsningar på flyttning av konstituenter. Flyttningar och hoppande på complementizers (i engelska) har t.ex. diskuterats åtskilligt, och relativsatser har t.ex. beskrivits som en flyttning av den relativiserade konstituenten till komplementizern = relativmarkören efterlämnande ett osynligt spår. Men det finns mycket annat som är viktigt att beskriva om man skall kunna översätta relativsatser. Man finner emellertid inom EST/GB och andra teoretiska grammatikmodeller inte t.ex. några konkreta regler för val av rätt form av relativpronomen, något som är nödvändigt om man vill implementera MT. Det talas ofta vid presentationen av dessa moderna grammatikmodeller som om böjning vore en trivial uppgift — men det är den inte ens i engelskan. En grammatik som skall användas för MT måste ha regler som genererar "who", respektive "whom" (för att inte tala om "whose"). I ett språk som georgiska är ordböjning huvuduppgiften: om man lyckas få verbformen rätt är det mesta av satsen avklarat. I både den ryska och den svenska grammatikmodulen upptas en ansenlig mängd av reglerna med att böja orden rätt.

4 Vad Swetra kan

Swetra kan analysera, syntetisera (generera) och översätta typiska lingvistmeningar med intransitiva, transitiva och dubbelt transitiva verb och upp till två satsadverbial och tre andra adverbial. Som adverbial kan även förekomma prepositionsfraser och underordnade konjunktionsbisatser. Swetra accepterar kopulativa satser (i ryska utan kopula) och satser med passivum, hjälpverb ("ha", modala hjälpverb som "kan" och aspektuella verb som "börja"). Swetra har omfattande komplex av regler för att hantera nominalfraser med bestämningar före och efter huvudet och komplicerad kongruens. Särskild uppmärksamhet har ägnats åt relativsatser; det referentbegrepp som givit RG dess namn växte i

själva verket fram ur analysen av relativsatser (se uppsatser om relativsatser i litteraturlistan).

Stora likvärdiga grammatikmoduler finns för engelska, ryska och svenska och tillhörande lexikon är för närvarande på några tusen ord. Demonstrationsprogram visar analys av satsen i källspråket, och generering av motsvarande sats i målspråket. Översättningen tar ofta bara några sekunder. Diverse uppsnabbningsknep användes (som vi här inte tänker avslöja). Grovöversättning görs via gemensam funktionell representation, men vissa transferregler har dessutom utarbetats för översättning i en viss riktning (särskilt ryska till engelska).

Typiska empiriska prov är deskriptiva texter, typ nyhetsnotiser. Swetra har bl.a. använt notiser ur Pravda, och det är troligt att Swetra specialiserar sig på notis- eller bulletintexter (se Sigurd & Gawrońska-Werngren 1988). Sådana texter liknar de texter som det textgenererande datorprogrammet Commentator kunde generera och samverkan mellan Commentator och Swetra för att generera och översätta sådana texter till några språk har diskuterats som ett framtida projekt. Ett exempel på en sådan text är följande: "Ett okänt flygplan närmade sig Öland från öster igår. Det girade söderut innan det kom in på svenskt område och försvann sedan söderut. Inget svenskt flygplan fanns då inom området. Flygplanet observerades av radar." Ytterligare exempel på texter som Swetra riktar in sig på finns i artiklar omnämnda i referenslistan.

5 Publiktrykket

När man säger att man håller på med automatisk översättning får man ofelbart frågan: "Hur många år dröjer det innan det går, tror Du?" Svaret: "Aldrig" accepteras inte och inte heller svar av typen: "Vi kan översätta texter med 80% kvalitet om de ligger inom en viss domän, t.ex. det marina området och grammatiken är så begränsad att den inte tillåter samordnade relativsatser och inte mer än två prepositionsfraser i en nominalfras etc."

Det finns ett publiktryck som de som sysslar med maskinöversättning alltid måste ha känt. Det är det tryck som lockar en att säga: "Om 5 år, eller om 10 år" och kanske mumla något ohörbart om textuella begränsningar. Det är det tryck som fått många forskare att lova för mycket och därigenom tidvis förstöra marknaden för forskning inom automatisk översättning. I ansökningarna för Swetra har vi aldrig lovat att leverera ett fungerande översättningssystem, bara att forska inom automatisk översättning.

Ofta sägs också: "Men poesi och skönlitteratur kommer väl aldrig att kunna översättas?" Och några tokroliga vandringexempel ges. Det finns ett tryck att då säga: "Nej, det kommer aldrig att gå." Men här känner jag mig ofta lockad att också gå på tvären och säga att: "Jo, poesi går lika bra, fast det blir kanske inte riktigt samma poesi utan ofta djärvare vändningar och fräschare metaforer." Men åhörarna vill inte heller höra på det örat. MT är ett svårt område—men det är jag inte den förste som konstaterat.

Litteratur

Gawrońska-Werngren, B. 1988. A referent grammatical analysis of Polish relative clauses. *Studia Linguistica* 42(1):18–48

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

— 1988. Using Referent Grammar (RG) in computer analysis, generation and translation of sentences. *Nordic Journal of Linguistics* 11(1–2):129–150.

—1989. A referent grammatical analysis of relative clauses. *Acta Linguistica Hafniensia* 21(2):95–115

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

Inst för Lingvistik
Helgonabacken 12
Lunds Universitet
Sverige