

Pilotstudie om maskinöversättning
inom ramen för Projekt

Kursdatabas

- Utveckling av språkliga resurser för ett
vetenskapsområde samt utvärdering

Eva Pettersson
evapet@stp.ling.uu.se

Uppsala universitet - Institutionen för lingvistik och
filologi

31 januari 2005

Innehåll

1	Bakgrund och syfte	4
2	Arbetsgång	4
2.1	Framtagande av lexikala resurser	4
2.1.1	Framtagande av ett baslexikon för allmänspråket	5
2.1.2	Utbildningsdepartementets svensk-engelska ordbok för utbildnings- och forskningsområdet	5
2.1.3	Terminologin i de medicinsk-farmaceutiska kursplanerna	6
2.1.4	Manuell tilldelning av viss syntaktisk och semantisk information	12
2.1.5	Resultat	13
2.2	Grammatikanpassning	14
2.2.1	Samordning av nominalfraser	15
2.2.2	Samordning av verb	16
2.2.3	Övriga samordningar	16
2.2.4	Subjektsigenkänning	17
2.2.5	Dubbla konjunktioner	18
2.2.6	Relativa adverb som bisatsinledare	18
2.2.7	Partikelverb	19
2.2.8	Reflexiva verb	19
2.2.9	Partiklar vs adverb	19
2.2.10	Preposition följt av bisats/infinitivsats	19
2.2.11	Inledande meningsfragment med avslutande kolon	20
2.2.12	Infinitivsats som efterställt attribut i nominalfras	20
2.2.13	Adverbial i infinitivsats	21
2.2.14	Återstående arbete	21
2.3	Evaluering	22
2.3.1	Automatisk utvärdering	23
2.3.2	Manuell utvärdering	23
2.4	Utveckling av förenklat gränssnitt för körning av översättningssystemet	25
3	Sammanfattande diskussion	25

Figurer

1	Gränssnitt för manuell eftergranskning av lexikoningångar . . .	8
2	Gränssnitt för manuell eftergranskning av översättningsrelationer	10
3	Det förenklade gränssnittet: startsidan	26
4	Det förenklade gränssnittet: resultatsidan	26

Tabeller

1	Exempel på lexikoningångar	13
2	Exempel på översättningsrelationer i lexikonet	13
3	Resultat av den automatiska utvärderingen	23
4	Resultat av den manuella utvärderingen enligt skalan korrekt- begriplig-svårbegriplig	24
5	Resultat av den manuella utvärderingen med J2450-måttet .	24

1 Bakgrund och syfte

Syftet med Projekt Kursdatabas är att samla all utbildningsinformation, såsom kursplaner och utbildningsplaner, i en universitetsgemensam databas. I denna databas ska all information finnas tillgänglig både på svenska och på engelska.

Ett delmål inom Projekt Kursdatabas, har varit att undersöka möjligheten med automatisk översättning av utbildningsinformation från svenska till engelska, med hjälp av MATS-systemet[WHF⁺04], utvecklat vid Institutionen för lingvistik och filologi vid Uppsala universitet.

Pilotprojektet med automatisk översättning har löpt över tre personmånader. Inom ramen för projektet har vi i första hand inriktat oss på översättning av kursplaner inom det medicinska och farmaceutiska vetenskapsområdet. Denna domän lämpade sig väl, då det fanns en hel del redan översatt material att tillgå inom detta område.

I denna rapport redogörs för anpassningen och vidareutvecklingen av MATS-systemets lexikon och grammatik till att hantera den vokabulär och de grammatiska strukturer som används i de medicinska och farmaceutiska kursplaner som tillhandahållits.

2 Arbetsgång

Arbetet med vidareutvecklingen av MATS-systemet har skett i tre delsteg:

1. Framtagande av lexikala resurser
2. Grammatikanpassning
3. Evaluering

Parallellt med detta har även ett förenklat gränssnitt för körning av översättningssystemet utvecklats. Nedan beskrivs de olika stegen mer utförligt.

2.1 Framtagande av lexikala resurser

I ett första steg togs ett lexikon fram, som ska täcka den vokabulär som används inom domänen.

För varje ingång i lexikonet krävs morfo-syntaktisk information i form av lemma, lexem, teknisk stam, böjningsmönster och ordklass. Verben ska dessutom tilldelas uppgift om konstruktionsbenägenhet, s k valens.

I de fall där jag varit osäker på exakt vilka morfo-syntaktiska särdrag som bör tilldelas en viss lexikoningång, har jag tagit ledning av Svenska Akademiens Ordlista[Aka86], Svenska skrivregler[spr00] och Svenska Akademiens Grammatik[THA95].

Lexikonet utvecklades i tre etapper. I den inledande fasen sammanställdes ett baslexikon, som ska täcka de allmänspråkliga delarna av det som ska översättas. Därpå utvidgades lexikonet till att även innefatta utbildningsspecifika termer hämtade från Utbildningsdepartementets *Svensk-engelsk ordbok för utbildnings- och forskningsområdet*[Reg03]. I det avslutande skedet lades ämnesspecifik (medicinsk-farmaceutisk) terminologi in i lexikonet, extraherad ur en kursplanekorpus med översättningar som tillhandahölls av Studerandebyrån. Nedan följer en mer detaljerad beskrivning av lexikonarbetet.

2.1.1 Framtagande av ett baslexikon för allmänspråket

Det allmänspråkliga lexikonet skapades på basis av på institutionen tillgängliga lexikon. Svensk-engelska lexikon anpassade för MATS-systemet, har tidigare utvecklats för översättning av lastbilsmanualer, jordbrukstexter och SÄPO-texter. Dessutom finns ett större enspråkigt lexikon för svenska, SCARRIE-lexikonet, tillgängligt. Även SCARRIE-lexikonet innehåller information om lemma, teknisk stam, böjningsmönster och ordklass, enligt det format som MATS kräver.

I lastbilslexikonet finns uppmärkt vilka av ingångarna som kan anses som allmänspråkliga och vilka som är specifika just för lastbilsdomänen. De ingångar som märkts som allmänspråkliga, fördes in i det blivande baslexikonet, medan övriga ingångar i lexikonet ignorerades.

Från jordbrukslexikonet och SÄPO-lexikonet, extraherades endast de ingångar, vars källsprakssegment dessutom fanns upptagna i SCARRIE-lexikonet. På så vis rensades de (mest) domänspecifika orden bort.

Det resulterande allmänspråkliga lexikonet består av 7 718 svenska ingångar, 5 638 engelska ingångar och 6 952 svensk-engelska ingångar.

2.1.2 Utbildningsdepartementets svensk-engelska ordbok för utbildnings- och forskningsområdet

I lexikonutvecklingens andra etapp, utvidgades lexikonet till att även innefatta sådan terminologi som används i utbildningssammanhang. För detta arbete utnyttjades Utbildningsdepartementets *Svensk-engelsk ordbok för utbildnings- och forskningsområdet*, som innehåller cirka 1 360 svenska utbildningstermer med engelska översättningar.

I ordboken finns inga uppgifter om böjning, ordklass och annan morfosyntaktisk information. För att komma åt denna information, slogs de svenska ordformerna och deras engelska översättningsekvivalenter automatiskt upp i på institutionen befintliga lexikon. I de fall där ordformen fanns att tillgå i något av dessa lexikon, tilldelades lemma, stam, böjningsmönster och ordklass i enlighet därmed.

De svenska ordformer som inte återfanns i något lexikon, kördes genom en på institutionen befintlig sammansättningsanalysator. Denna sammansättningsanalysator har SCARRIE-lexikonet som grund, och ger utifrån detta ifrån sig ett antal analyser av sammansättningen. Däribland väljer vi, på automatisk väg, den mest tillförlitliga analysen med hjälp av ett program, utvecklat på institutionen [Åbe03]. I de fall där sammansättningsanalysatorn segmenterat upp ett ord i flera lemman, tilldelades morfo-syntaktisk information utifrån högerledet i sammansättningen. De ord som inte heller på detta vis fick någon analys, tilldelades morfo-syntaktiska särdrag manuellt.

I de fall där ett lemma återfanns både i det allmänspråkliga lexikon som skapats i steg 1 och i Utbildningsdepartementets ordbok, kontrollerades huruvida översättningen skilde sig åt mellan de båda lexikonerna. Om så var fallet, skrevs den allmänspråkliga översättningen över av den översättning som anges i ordboken.

I Utbildningsdepartementets ordbok, förekom även att en svensk term givits ett flertal engelska översättningsalternativ. Då MATS-systemet arbetar utifrån en defaultöversättning per ord, behövde man i dessa fall ta ställning till vilket av dessa översättningsalternativ som skulle väljas. Som regel har då det första översättningsalternativet valts, i enlighet med Utbildningsdepartementets egna rekommendationer.

I en del fall har de olika översättningsalternativen markerats som tillämpliga på universitetsnivå (univ), skolnivå (skol), inom vuxenutbildningen (vux) respektive inom folkhögskolan (fhs). För våra ändamål har då universitetsalternativet valts som översättning.

Vissa av ordbokens översättningsalternativ har även markerats som brittisk (Br) respektive amerikansk (Am) engelska, med avseende på ordval och stavning. Här har det brittiska alternativet valts, då det enligt Utbildningsdepartementet är brukligt att använda brittisk engelska då man vänder sig till läsare som inte har engelska som modersmål.

2.1.3 Terminologin i de medicinsk-farmaceutiska kursplanerna

Som tidigare nämnts, satsar vi i pilotstudien i första hand på översättning av kursplaner inom den medicinsk-farmaceutiska fakulteten. Till grund för extraherandet av lexikala termer inom denna domän, tillhandahöll Studerandebyrån en korpus bestående av 203 kursplaner översatta från svenska till engelska (cirka 345 740 löpord på den svenska sidan).

Svensk terminologi I ett första steg bearbetades den svenska delen av korpusen. Ur de svenska kursplanerna, extraherades alla unika ordformer, vilket gav ett resultat på cirka 5 000 typord (efter omvandling av meningsinledande versal).

Dessa typord tilldelades lemma, stam, mönsterord och ordklass, med hjälp av på institutionen befintliga lexikon i kombination med en sam-

mansättningsanalysator (se vidare 2.1.2) och det fritt tillgängliga programmet **Linguistica**, som utifrån en korpus automatiskt tilldelar de ingående orden morfologiska egenskaper. Linguistica använder sig av tekniker för inlärning utan förlaga, och ger bland annat ifrån sig en fil innehållande löporden segmenterade i tekniska stammar med tillhörande ändelser, enligt nedan:

```
laboration NULL.er.erna
```

Denna fil har körts genom ett program, utvecklat på institutionen, som jämför de suffix som av Linguistica associerats till vissa stammar, mot de suffix som är förknippade med speciella mönsterord i MATS-systemets morfologiska modell[Gus04]. Programmet använder sig dessutom av TnT-taggar[Bra00] för tilldelningen av morfologiska särdrag till ordformerna. På så vis kan exempelvis mönsterord som definierats för utrala substantiv, exkluderas från mängden av möjliga mönsterord att tilldela en ordform som taggats som neutrum.

Ordformer som inte kunde tilldelas morfo-syntaktisk information med hjälp av någon av ovanstående resurser, antogs vara substantiv och tilldelades det vanligast förekommande mönsterordet (FILM) och en stam identisk med grundformen.

Innan termerna lades in i lexikonet, gjordes en semi-manuell eftergranskning, med hjälp av en modifierad version av ett gränssnitt som utvecklats i samband med tidigare projekt på institutionen (se vidare [GP03]). Gränssnittet, som illustreras i figur 1, visar lemma, stam, mönsterord och ordklass och ger användaren möjlighet att acceptera, ändra eller kasta bort ingångar.

I arbetet med att utröna vilka ordformer i korpusen som var att betrakta som medicinska termer och vilka som helt enkelt utgjordes av felstavningar, utnyttjades den elektroniskt tillgängliga engelsk-svenska medicinska thesaurusen MeSH (Medical Subject Headings)[Ins04] och i viss mån även FASS[Läk04] och Svenska Akademiens ordlista. Utöver denna rensning, togs även de ingångar bort som redan fanns med i lexikonet i och med utvecklandet av baslexikonet och inlägget av utbildningsspecifika termer. De återstående 2 392 svenska termerna lades in i lexikonet.

Extraktion av översättningsekvivalenter Efter att de svenska termer som ska in i lexikonet tagits fram, påbörjades arbetet med att extrahera översättningsekvivalenter för dessa termer. Detta steg innefattar både att etablera översättningsrelationer mellan svenska och engelska termer, och att tilldela de engelska ingångarna morfo-syntaktisk information.

Extraktionen av översättningsekvivalenter grundar sig på en ordlänkning utförd med hjälp av Clue Aligner, ett länkningsprogram utvecklat på insti-



Figur 1: Gränssnitt för manuell eftergranskning av lexikoningångar

tutionen av Jörg Tiedemann[Tie03]. Länkingsprogrammet går automatiskt igenom parallellkorpuser och parar ihop enheter i källspråkstexten med motsvarande enheter i målspråkstexten, utifrån både statistiska beräkningar och lingvistiska jämförelser baserade på ordklassstaggning och en grov syntaktisk analys, s k chunkning. Programmet länkar först på meningsnivå, för att sedan gå in på ord- och frasnivå. Programmet förutsätter inte att det är ett ett-till-ett-förhållande mellan segmenten, dvs ett segment i källspråkstexten kan länkas till noll, ett eller flera segment i målspråkstexten.

Resultatet av länkningen är en fil med svenska ordformer sammankopplade med engelska ordformer och fraser, med en frekvens som anger hur många gånger den svenska ordformen har länkats till sin engelska motsvarighet. För de medicinsk-farmaceutiska kursplanerna resulterade länkningen i 14 487 länkar, på nedanstående format:

```

3X Omtentamen Re-examination
2X omtentamina re-examinations
1X omtentamina the re-examinations

```

Länkingsmaterialet efterbearbetades sedan på automatisk väg. Till att börja med gjordes ordformsinitiala versaler om till gemener. Därefter slogs de länkar, vars källspråkssegment var identiska, samman till en enda ingång med ett antal frekvensordnade översättningsekvivalenter. I de fall där samma översättningsekvivalent förekom flera gånger för samma källspråkssegment, räknades dess olika frekvenser samman och slogs ihop till ett enda översättningsalternativ.

Efterbearbetningen gav följande resultat för ordformerna *omtentamen* respektive *omtentamina* (givet ovanstående länkresultat):

```
omtentamen          {3X re-examination}  
omtentamina {2X re-examinations | 1X the re-examinations}
```

Då vi enbart var intresserade av de länkar, vars källspråkssegment finns med i den medicinsk-farmaceutiska terminologi som tagits fram i föregående etapp, rensades alla länkar bort som inte uppfyllde det villkoret.

Därefter lemmatiserades källspråkssidan. Lemmatiseringen utfördes med hjälp av det svenska stamlexikon som utvecklades i föregående steg (se vidare 2.1.3). Då detta lexikon innehåller information om teknisk stam och böjningsegenskaper, kan man för varje ingående lemma generera alla dess möjliga ordformer. De ordformer som förekom på källspråkssidan i länkningsmaterialet, jämfördes mot de ordformer som genererats utifrån lexikonet, varpå länkningsmaterialets ordform byttes ut mot det lemma som angavs för denna ordform i lexikonet. Om ordformen var homograf, och sålunda svarade mot flera olika lemman i lexikonet, skapades en ingång för varje lemma i länkningsmaterialet.

Efter ovanstående efterbearbetning återstod 1 477 länkar, på nedanstående format:

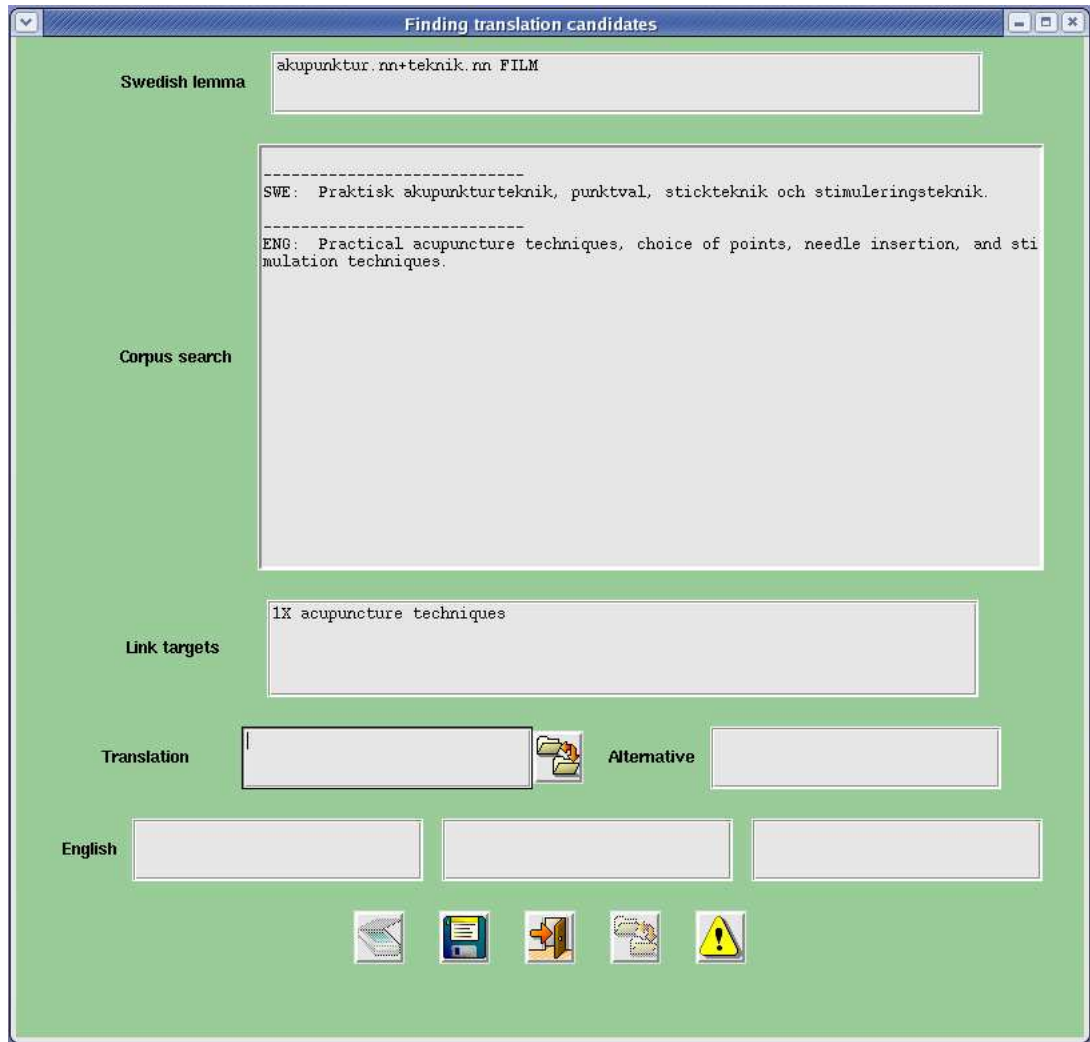
```
omtentamen.nn {3X re-examination | 2X re-examinations | 1X the re-examinations}
```

I det avslutande steget genomfördes en semi-manuell eftergranskning av resultatet och tilldelning av morfo-syntaktiska särdrag på målspråkssidan. För detta ändamål modifierade jag ett gränssnitt som tidigare utvecklats på institutionen (se vidare [GP03]). Det modifierade gränssnittet illustreras i figur 2.

Gränssnittet innehåller fem fält. I det översta fältet visas det källspråkslemma som ska tilldelas en översättning. I fältet därunder visas alla de meningar i kursplanekorpusen, där detta lemma förekommer i någon av dess former. I samma ruta visas även de engelska meningar som källspråkssegmenten länkats till. På så vis får man tillgång till kontexten, och kan avgöra huruvida ett visst lemma har flera olika översättningar som används i olika kontexter.

I det tredje fältet presenteras (det automatiskt efterredigerade) resultatet av ordlänkningen, med de olika översättningsalternativen i fallande frekvensordning.

I det fjärde fältet finns två parallella rutor. I den vänstra rutan skriver man in det engelska lemma man vill ha som översättning i lexikonet. I detta skede kollar programmet igenom de svensk-engelska lexikon som redan finns på institutionen. Om det svenska lemmat förekommer i något av dessa lexikon, presenteras i denna ruta den översättning som förekommer i detta lexikon. Användaren kan dock välja att byta översättning, med ledning av länkningsresultatet och den domänspecifika kontexten. I den högra



Figur 2: Gränssnitt för manuell eftergranskning av översättningsrelationer

rutan kan man ange alternativ till källspråkslemmat. Detta kan exempelvis komma till pass i samband med adjektiv som i sin singulara neutrumform är homografa med adverb, t ex *munlig-muntligt*. I vissa fall har då endast adjektivtolkningen tagits med i det svenska lexikonet, men korpusen visar att även adverbtolkningen bör finnas med.

Bredvid den vänstra rutan i det fjärde fältet finns en knapp, som man trycker på när man skrivit in sitt översättningsförslag. Då aktiveras det nedersta fältet, som innehåller tre parallella rutor. Här skrivs det engelska lemmats stam, mönsterord respektive ordklass in. Programmet ger alltid ett förslag till stam, mönsterord och ordklass, som användaren kan välja att spara eller ändra. De värden som visas för användaren, bestäms utifrån

det lemma användaren skrivit in som översättningsekvivalent. Om lemmat förekommer i något av de på institutionen befintliga lexikonerna, visas de värden lemmat har i detta lexikon. I övriga fall bestäms värdena utifrån heuristiska principer. Stammen skapas då utifrån lemmat, genom att lemmaextensionen kapas av och eventuella underscore-tecken omvandlas till mellanslag. Ordklass väljs utifrån lemmaextensionen. Även vid bestämningen av vilket mönsterord som ska visas för användaren, tas lemmaextensionen i beaktande. Här väljs helt enkelt det vanligast förekommande mönsterordet inom ordklassen.

Problem och resultat Ett problem i samband med länkningen och extraheringen av översättningsekvivalenter, har varit att de parallellställda texterna inte alltid innehållit samma mängd information. I vissa fall har information som delgivits i den svenska texten, utelämnats helt eller skrivits om på den engelska sidan, som i exemplet nedan:

Källspråkssegment: *ALLMÄNNA UPPLYSNINGAR: Kursen ges på A-nivå dels som stödämne och dels inom huvudämnet biomedicinsk laborietvetenskap.*

Motsvarande målspråkssegment: *GENERAL INFORMATION:*

I dessa fall kan det vara svårt att veta hur vissa termer ska översättas. Dessutom ger det upphov till en ”skevhet” i det övriga länkingsresultatet, då systemet förväntar sig att översättningen är fullständig, och om någon mening helt saknar översättning, kan det leda till en felaktig länkning, som fortplantar sig vidare i texten.

Denna skevhet uppstår även i det motsatta fallet, dvs när information som saknas på den svenska sidan, lagts till på den engelska sidan, som i nedanstående exempel:

Källspråkssegment: *ALLMÄNNA UPPLYSNINGAR:*

Motsvarande målspråkssegment: *GENERAL INFORMATION: This is a compulsory course of the fourth year of the 160 point study programme in Biomedicine.*

Problemet med icke överensstämmande översättningar torde vara en av anledningarna till att endast 1 478 av totalt 2 349 svenska lemmorna fanns med i länkingsmaterialet.

Det har också varit svårt att veta vilken översättningsekvivalent som ska väljas, i de fall där det i korpusen förekommer flera olika översättningsalternativ. I första hand har jag då arbetat utifrån att det översättningsalternativ som har högst frekvens är det mest önskvärda. Om de olika alternativen har lika hög frekvens, eller om översättningen verkar underlig (och man kan

misstänka att det rör sig om ett stavfel alternativt felaktig användning av terminologi), samt i de fall där en term helt saknar översättning i korpusen, har jag använt mig av olika källor för att förvissa mig om att få fram en så bra översättningsekvivalent som möjligt. Den främsta källan för kontroll av detta slag har varit den engelsk-svenska medicinska thesaurusen MeSH[Ins04]. För översättningar av mer allmänspråklig karaktär, har även *Stora Engelska Ordboken*[Nor88] varit till nytta. I de fall där jag varit osäker på stavningen av den översättning som anges i korpusen, har jag dessutom, utöver MeSH, utnyttjat *Oxford English Dictionary*[Pre04].

Bearbetningen av den medicinsk-farmaceutiska kursplanekorpusen resulterade slutligen i 2 327 svenska ingångar, 1 987 engelska ingångar och 2 269 översättningsrelationer.

2.1.4 Manuell tilldelning av viss syntaktisk och semantisk information

Som tidigare nämnts bör det för verben, utöver information om lemma, stam, mönsterord och ordklass, även anges en valens. Om ett verb redan tidigare tilldelats en valens i något av de på institutionen befintliga lexikonerna, kopierades denna valens helt enkelt över till detta lexikon. Övrig valens-tilldelning har skett manuellt, på basis av min lingvistiska intuition och faktisk användning i korpusen.

För att möjliggöra mer generella grammatikregler, har dessutom substantiven tilldelats semantisk information. Detta arbete utfördes, vid institutionen, av Jan Hellström. Den semantiska klassificeringen delar in substantiven enligt nedanstående kriterier¹:

- konkreta vs abstrakta substantiv (concr)
- dividuativa vs individuativa substantiv (indiv)
- animata vs inanimata substantiv (anim)
- substantiv som enbart förekommer i pluralis (plt)
- substantiv som enbart förekommer i singularis (slt)
- kvantitetsbetecknande substantiv (quant)
- tidsangivande substantiv (temp)
- substantiv som (tillsammans med en preposition) kan uttrycka rumsförhållanden (loc)

¹Se vidare [Hel01] för en utförligare beskrivning av hur de semantiska särdragen tilldelats. Hellströms rapport beskriver tilldelningen av semantiska särdrag inom ramen för ett tidigare projekt, men principen är densamma även i detta projekt.

- substantiv som (tillsammans med en preposition) kan uttrycka sätt
- måttsbetecknande substantiv (quant)
- substantiv som betecknar verbhandlingar (verb)

Lemmat *ambulans.nn* har, enligt ovanstående, tilldelats följande egenskaper:

+concr +indiv -anim -plt -slt -quant -temp +loc -manner -measure -verb

2.1.5 Resultat

Det resulterande lexikonet kan sägas bestå av tre delar: ett svenskt lexikon, ett engelskt lexikon och ett svensk-engelskt lexikon. I praktiken är dock all information lagrad i en och samma databas.

I de enspråkiga delarna av lexikonet finns information om lemma, lexem, stam, böjningsmönster, ordklass och domän. På den svenska sidan finns dessutom viss semantisk information för substantiven och valensinformation för verben. Se tabell 1 för exempel.

Lemma	Lexem	Stam	Mönsterord	Ordklass	Domän	Semantik/ Valens
kurs.nn	1	kurs	SATS	NOUN	kd	-concr +indiv -anim -plt -slt -quant -temp -loc -manner -measure -verb
blöda.vb	1	blö	TYDA	VERB	gd	va.accelerera

Tabell 1: Exempel på lexikoningångar

I den svensk-engelska delen av databasen, anges översättningsrelationer på lemma- och lexemnivå. Se tabell 2 för exempel.

Källspråks- lemma	Källspråks- lexem	Målspråks- lemma	Målspråks- lexem
kurs.nn	1	course.nn	1
blöda.vb	1	bleed.vb	1

Tabell 2: Exempel på översättningsrelationer i lexikonet

Efter den semi-manuella eftergranskningen av lexikoningångarna, bestod lexikonet av 10 195 svenska ingångar, 8 471 engelska ingångar och 10 156 svensk-engelska ingångar.

De lexikoningångar som togs fram i det inledande steget, dvs de ingångar som utgör baslexikonet, har märkts med domänbeteckningen *gd* (general

dictionary), medan de domänspecifika ingångarna märkts med *kd* (kursdatabas). Sammanlagt har 6 951 svenska ingångar och 5 638 engelska ingångar märkts med *gd*.

2.2 Grammatikanpassning

MATS-systemet översätter mening för mening, och innehåller tre olika grammatikkomponenter:

1. *analysgrammatiken* som gör en grammatisk analys av källspråksmeningen
2. *transfergrammatiken* som omvandlar källspråksanalysen till motsvarande grammatiska struktur på målspråkssidan
3. *genereringsgrammatiken* som utifrån målspråksanalysen genererar en målspråksmening

I analyssteget genereras alla möjliga grammatiska analyser av indatasträngen. I de fall där MATS-systemet misslyckas med att analysera hela meningen som ett sammanhängande segment, till exempel på grund av stavfel eller grammatiska fel i indatan, eller på grund av att meningen innehåller strukturer som grammatiken ännu inte hanterar, ger analysgrammatiken ifrån sig partiella analyser som skickas vidare till transfern var för sig.

I transfergrammatiken hanteras både strukturella och lexikala skillnader mellan språken. De lexikala transferreglerna utgör ett komplement till lexikonet, då dessa definierar kontextuellt beroende översättningar, som inte går att komma åt i lexikonet.

En vanligt förekommande typ av lexikal transferregel anger val av preposition i samband med vissa verb, substantiv eller adjektiv. Ett exempel är frasen *på svenska*. I lexikonet anges översättningen *on* för prepositionen *på*, men i detta sammanhang är *in Swedish* en bättre översättning än *on Swedish*, vilket hanteras i transfergrammatiken.

Exempel på andra kontextuellt beroende översättningar som hanteras med hjälp av transferregler är idiomatiska uttryck såsom *äga rum*, som därmed översätts till *take place* istället för det mer bokstavstroga *own rooms*.

I genereringsmodulen ordnas målspråksstrukturens ingående enheter i enlighet med målspråkssyntaxen. Denna modul är helt skild från analyssteget, vilket innebär att den genererade målspråksordföljden normalt sett är helt oberoende av källspråkets ordföljd. I de fall där genereringsregler saknas för en viss struktur, dvs när systemet inte vet hur det ska ordna de ingående enheterna, kommer dock målspråksenheter ut i samma ordning som de motsvarande källspråksenheter.

Arbetet med att anpassa de tre grammatikerna till att hantera de strukturer som förekommer i de medicinsk-farmaceutiska kursplanerna, baseras på de 203 kursplaner från det medicinsk-farmaceutiska vetenskapsområdet, som Studerandebyrån tillhandahållit. Materialet delades upp i en träningsdel och en utvärderingsdel. Var 20:e kursplan lades till träningskorpussen, medan övriga kursplaner fördes till utvärderingskorpussen.

Nedan beskrivs hur jag har hanterat några av de grammatiska problem som jag stött på i arbetet med grammatikanpassningen.

2.2.1 Samordning av nominalfraser

Ett återkommande problem vid körningen av träningskorpussen genom MATS-systemet, bestod i analys och generering av samordnade nominalfraser eller nominalgrupper. I materialet är det mycket vanligt förekommande med samordningar av ett stort antal nominalfraser, i och med att ett antal kursmoment eller förkunskapskrav räknas upp inne i meningen, till exempel:

Kursen behandlar viral och cellulär tillväxt, spridningsvägar och förekomst av mikroorganismer t.ex. hos sjuka och smittobärare, viktiga patogena mikroorganismer, metoder att påvisa och kontrollera infektionssjukdomar, immunsystemets uppbyggnad och funktion, virulensegenskaper hos viktiga patogener, olika typer av virus och deras förökningssätt, odlingstekniker lämpade för mikroorganismer och virus, viktiga antimikrobiella läkemedels verkningsmekanismer och användningsområden, läkemedelsresistensens olika verkningsätt, hur den kan undersökas, mätas och motverkas samt kunskap om resistensens spridningssätt.

I analyssteget gav nominalfrassamordningar med många led upphov till problem, på grund av ett snabbt ökande antal analyser. Detta eftersom grammatiken syftar till att täcka alla möjliga tolkningar av samordningen. Ett alltför stort antal kombinationer av analyser leder till att programmet ”ger upp” och stannar, utan att någon översättning har valts, dvs indatasträngen kommer ut oförändrad.

För att förhindra detta, har jag arbetat med att begränsa antalet analyser i fallet med nominala samordningar. Detta har bland annat inneburit att jag tagit bort nominalgruppstolkningen för strukturer som även kan tolkas som nominalfraser, dvs plurala och/eller definitiva konstruktioner. Sålunda ges de ingående segmenten i följande struktur nu endast beteckningen ’nominalfras’: *föreläsningar, seminarier, laborationer, videodemonstrationer och datorövningar.*

Antalet analyser växer ytterligare då ett eller flera efterställda attribut förekommer, till exempel i frasen *godkänd teori och deltagande på samtliga obligatoriska kursmoment.* Här finns två olika tolkningar:

1. det efterställda attributet hör till hela samordningen: [(godkänd teori och deltagande) (på samtliga obligatoriska kursmoment)]
2. det efterställda attributet hör till det senaste ledet i samordningen: [(godkänd teori) och (deltagande på samtliga obligatoriska kursmoment)]

Med ett stort antal ambiguiteter av detta slag i en och samma samordning, exploderar antalet analyser. För att undvika detta, angavs i grammatiken att endast den senare tolkningen ska gälla, dvs att det efterställda attributet ska analyseras som tillhörande det senaste ledet i samordningen.

Medan analysgrammatikens regler för nominalfrassamordningar är rekursiva, tillåtande i princip hur många samordningsled som helst, är detta inte fallet med genereringsgrammatiken. Hittills hade endast samordningar med upp till tre led förekommit, men för denna domän krävdes en utökning av reglerna till att täcka strukturer med upp till sex samordningsled.

2.2.2 Samordning av verb

Även samordningar av verb förekommer relativt frekvent i kursplanematerialet, t ex i meningen *Genom kursen skall den studerande förvärva träning i att **planera, genomföra och redovisa** en undersökning.* I denna typ av elliptiska konstruktion, ”delar” flera verb på samma objekt. För att få en så tillförlitlig analys som möjligt, och därmed undvika övergenerering, krävs att alla de ingående verbens valenser kontrolleras innan strukturen sparas som giltig.

För att hantera denna struktur, infördes fraskategorin `cong` för verb-samordningar. Denna verbfras innehåller information om alla de ingående verbens valensramar.

När en verbfras av detta slag har lästs, vandrar parsern framåt i analysen, och kontrollerar att de argument som det första verbet i samordningen kräver, följer. Därpå kontrolleras för vart och ett av de övriga verben i samordningen, huruvida dess obligatoriska argument är närvarande, samt att inga argument utanför valensramen förekommer.

Eftersom *planera* såväl som *genomföra* och *redovisa*, har givits en valens som tillåter ett efterföljande direkt objekt, godkänner sålunda grammatiken konstruktionen ***planera, genomföra och redovisa** en undersökning.* Däremot godkänns inte *planera, genomföra och **svimma** en undersökning,* då *svimma* enligt sin valensram inte kan ta något direkt objekt.

2.2.3 Övriga samordningar

Övriga samordningar som förekom i materialet och som tidigare inte behandlats av grammatiken, är samordningar av prepositionsgrupper (tidigare fungerade endast samordningar av prepositionsfraser) och samordningar

av olika typer av satser. Jag har således definierat regler för samordning av prepositionsgrupper samt för följande satstyper:

1. infinitivsatser², ex: *att självständigt söka information och lösa problem*
2. deklarativa huvudsatser, ex: *Undervisningen pågår i 8 veckor och genomförs parallellt med kursen i Oto-Rhino-Laryngologi*
3. bisatser, ex: *att den studerande har deltagit i samtliga obligatoriska kursmoment samt visar sig nöjaktigt ha inhämtat både teori- och laborationskurser*

Samordning av infinitivsatser är relativt triviale, då de ingående samordningsleden inte ”delar” på något argument.

När det gäller samordning av bisatser eller deklarativa huvudsatser, förekommer dock elliptiska konstruktioner där subjektet är underförstått i de senare leden, som i fall 2 och 3 ovan. Genereringsmodulen förväntar sig dock ett subjekt närvarande, då verbkongruensen är beroende av subjektets numerus och persontillhörighet.

Om man i analyssteget helt enkelt lägger in subjektet i de senare leden också, kan detta emellertid leda till att subjektet realiserar flera gånger på målspråkssidan. Jag har därför valt att, i de underspecificerade leden, ärvä in subjektets särdrag i en speciell variabel, som slängs bort i genereringssteget. På så vis realiserar subjektet endast det antal gånger det faktiskt förekommer i källsprakssegmentet.

2.2.4 Subjektsigenkänning

För att korrekt böjning av verbet ska kunna garanteras, krävs att systemet har lyckats hitta satsens subjekt, vilket inte är helt triviale. I många fall återfinns subjektet i den deklarativa huvudsatsens fundament. MATS-systemet hade ändå i vissa fall svårt att känna igen segmentet som ett subjekt. Detta gällde främst i de fall där subjektet var komplext, exempelvis innehållande inflettade bisatser eller förtydliganden inom parentes eller mellan kommatecken, som i *Skriftlig examination med frågor av MEQ-typ (”bläddertenta”) genomföres vid kursens slut*. Grammatiken utökades sålunda till att hantera ett antal konstruktioner av denna typ.

Subjektsigenkänningen försvårades även i de fall där fundamentet utgjordes av andra funktioner än subjektet. Oftast var det då adverbial i fundamentet som inte analyserades korrekt. I grammatiken anges vilka konstruktioner som får förekomma som satsinitiala adverbial. Denna information var

²I traditionell grammatik talar man snarare om infinitivfraser än infinitivsatser. I MATS-grammatiken hanteras denna typ av konstruktion dock som en sats, varför begreppet *infinitivsats* används i detta sammanhang

inte fullständig, och utökades följaktligen till att täcka de nytillkomna konstruktionerna.

Varje mening i träningskorpusen, där verbkongruensen misslyckats, inspekterades manuellt, och i de fall där det bedömdes som motiverat utökades grammatiken till att hantera dessa fall.

2.2.5 Dubbla konjunktioner

I grammatiken saknades adekvata regler för hanterandet av dubbla konjunktioner av typen *varken... eller...* Jag har således utökat grammatiken till att hantera följande konstruktioner:

- *dels* + nominalfras/nominalgrupp + *dels* + nominalfras/nominalgrupp/explikativ bisats
- *varken/vare sig* + nominalfras/nominalgrupp + *eller* + nominalfras/nominalgrupp

Exempel: *För antagning till fristående kurs gäller **dels** grundläggande behörighet enligt Högskoleförordningen, **dels** att den studerande har minst 15 poäng i farmaceutisk bioteknik eller motsvarande naturvetenskaplig utbildning.*

2.2.6 Relativa adverb som bisatsinledare

Relativa bisatser inleds av subjunktionen *som* (som ibland kan utelämnas) eller av en relativ satsbas (relativt pronomen eller adverb).[THA95]

Tidigare har grammatiken endast hanterat relativa bisatser inledda med *som*. I kursplanekorpusen förekommer emellertid även relativa bisatser inledda med relativt adverb, t ex *Därefter vidtar CBB2 (8 veckor) **då** cellens organeller (...) genomgås i detalj.*

De relativa adverbena är *när*, *då*, *där* och *dit*. [THA95] Jag klassificerade dessa som relativa även i vårt lexikon.

Den grammatikregel som fanns för relativa bisatser, behövde också modifieras. Den ursprungliga regeln tillät att den övergripande satsens korrelat fungerade som subjekt eller objekt i den underordnade satsen. Jag lade in restriktionen att detta endast gäller när bisatsinledaren inte utgörs av ett relativt adverb. Jämför nedanstående exempel:

- laboratorium där verksamheten anknyter till utbildningens inriktning
- *laboratorium där anknyter till utbildningens inriktning
- laboratorium som anknyter till utbildningens inriktning

2.2.7 Partikelverb

Partikelverben utgör en typ av konstruktion som lämpligen behandlas med transferregler. För att på ett systematiskt sätt komma åt så många partikelverb som möjligt för denna domän, taggades hela kursplane-korpusen med hjälp av TnT-taggararen[Bra00]. Ur det taggade materialet togs alla segment fram som taggats som partiklar, tillsammans med det närmast förekommande verbet i vänsterkontexten. På så vis fick jag en lista över sexton potentiella partikelverb, av vilka tre sållades bort som feltaggade. För de kvarvarande tretton partikelverben skrevs transferregler och definierades valensramar (med ledning av användningen i korpusen). Exempel på vanligt förekommande partikelverb i korpusen är *bygga upp*, *ha rätt till* och *ta upp*.

2.2.8 Reflexiva verb

Även för de reflexiva verben utnyttjades den TnT-taggade korpusen, vilket resulterade i en lista på åtta olika reflexiva verb, såsom *tillgodogöra sig*, *lära sig* och *visa sig*. På samma sätt som för partikelverben, tilldelades också de reflexiva verben valensinformation och transferregler definierades för att komma åt översättningen.

2.2.9 Partiklar vs adverb

Enligt *Svenska Akademiens ordlista*[Aka86], kan ord som *på* och *av* fungera både som prepositioner och adverb. Denna indelning återfanns även i vårt lexikon. Detta resulterade dock i en hel del oönskade ambiguiteter, där adverbtolkningen valdes framför prepositionstolkningen. Sålunda kunde segment som *av nya behandlingsmetoder* felaktigt översattas till *off new treatment methods* istället för *of new treatment methods*.

Då adverbtolkningen endast förekommer i de fall där ordet ingår i en partikelverbskonstruktion, valde jag därför att ta bort adverbtolkningen ur lexikonet, och endast behålla prepositionstolkningen. Dessa prepositioner tilldelades istället ett särskilt särdrag för att ange att de kan fungera som partiklar. De prepositioner detta gällde är *av*, *om*, *på* och *till*.

2.2.10 Preposition följt av bisats/infinitivsats

För konstruktioner där en preposition följs av en bisats med inledande subjunktion, bör man undvika att realisera prepositionen i den engelska översättningen. Som exempel kan nämnas den svenska frasen *information om att läkemedel absorberas i människokroppen*. Här är *information that drugs are absorbed in the human body* att anse som en bättre översättning än *information on that drugs are absorbed in the human body*. En transferregel skrevs därför som tar bort prepositionen på målspråkssidan, då den

följs av en bisats med inledande subjunktion.

På samma sätt tas prepositionen bort på den engelska sidan i de fall där en preposition följs av en infinitivsats med inledande infinitivmärke, t ex *angelägen om att*, som hellre bör översättas till *anxious to* än *anxious about to*.

2.2.11 Inledande meningsfragment med avslutande kolon

I det material som tillhandahållits, förekommer upprepade gånger en rubrik som avslutas med kolon och följs av en sats (eller en fras), t ex

MÅL: Genom kursen skall den studerande förvärva träning i att planera, genomföra och redovisa en undersökning.

Då segmenteraren inte hanterar kolon som meningsavskiljande segment, krävs att denna konstruktion hanteras av grammatiken, för att analysen av segmentet ska bli fullständig. Regeln för deklarativa huvudsatser anropas normalt endast i meningsbörjan, för att minska risken för övergenerering. Således skulle huvudsatsen som följer på kolon i ovanstående exempel, inte kunna analyseras utan att denna struktur hanteras.

För att komma tillrätta med detta, lät jag grammatiken analysera meningsinitiala nominalfraser och nominalgrupper med efterföljande kolon som ett meningsfragment, som kan följas av exempelvis en deklarativ huvudsats.

En vanligt förekommande rubrik i materialet är även *INGÅR I PROGRAM*, som i

INGÅR I PROGRAM: Läkarprogrammet, 220 p

Här utgörs rubriken inte av en nominalfras, utan av ett verb med tillhörande argument. Dessutom saknas subjektet i denna typ av konstruktion, vilket skapar problem i samband med verbböjningen. Jag har valt att spara även meningsinitiala verb med tillhörande argument följt av kolon, som ett meningsfragment.

I just fallet med *INGÅR I PROGRAM*, är detta en fast rubrik som återkommer i alla kursplaner med översättningen *STUDY PROGRAMME*. Jag har därför definierat en transferregel som gör om översättningen av *INGÅR I PROGRAM* till *STUDY PROGRAMME*, i de fall där segmentet förekommer som en rubrik, dvs på meningsinitial position och följt av ett kolon. Övriga fall med meningsinitiala verb följt av kolon, tilldelas tredje person singularis för verbböjningen.

2.2.12 Infinitivsats som efterställt attribut i nominalfras

I korpusen förekommer en del konstruktioner med infinitivsats som efterställt attribut i en nominalfras, som i

möjlighet att komplettera icke godkänd laboratoriekurs och förmåga att självständigt söka information.

Detta verkar dock vara begränsat till att gälla ett fåtal substantiv. För att undvika övergenerering, har jag därför endast applicerat denna egenskap på de attesterade substantiven *möjlighet, förmåga, kunskap, modell* och *ambition*.

2.2.13 Adverbial i infinitivsats

I grammatiken tilläts satsadverbial i infinitivsatsen på positionen direkt efter infinitivmärket, medan övriga adverbial endast tilläts sist i satsen, enligt nedan:

Placering av satsadverbial i infinitivsatsen: *att inte kunna bidra till en säker och rationell läkemedelsanvändning*

Placering av övriga adverbial i infinitivsatsen: *att kunna bidra till en säker och rationell läkemedelsanvändning på olika sätt*

Adverbkonstruktioner som normalt sett inte klassificeras som satsadverbial, förekommer dock relativt frekvent på satsadverbialspositionen. I vissa fall är detta även att föredra, som i nedanstående exempel:

att på olika sätt kunna bidra till en säker och rationell läkemedelsanvändning

Jag lade därför in i grammatiken att också övriga adverbial får förekomma på positionen direkt efter infinitivmärket i en infinitivsats.

2.2.14 Återstående arbete

Trots ansträngningarna att begränsa antalet analyser av indatan, framförallt i samband med samordningar, händer det fortfarande att vissa konstruktioner ger upphov till ett alltför stort antal analyser, vilket leder till utdata på svenska. Detta kan förhindras om användaren undviker alltför långa samordningar, och istället delar upp innehållet på flera meningar. Dock borde även bättre strategier utvecklas i MATS-systemet för att förhindra uppkomsten av detta problem. Arbete med att lösa detta problem med hjälp av taggning av källtexten, är på gång.

Ett annat problem utgörs av homografer. Många homografer disambigueras i analyssteget. Exempelvis kan adverbtolkningen av ordet *för* uteslutas i segment av typen *för betyget godkänd*, då adverbet *för* inte kan modifiera en nominalfras. Däremot räcker inte den grammatiska analysen till för att utesluta adverbtolkningen i segmentet *för godkänt betyg*, eftersom adverbet

för kan modifiera adjektiv. Alltså är både adverbtolkningen och prepositionstolkningen tillämpliga i det senare fallet, även om adverbtolkningen är att betrakta som osannolik ur ett rent semantiskt perspektiv.

Även homografproblemet kan i många fall lösas genom taggning av källtexten. Dessutom utgör Jan Hellströms arbete med att semantiskt klassificera substantiven, en god grund för semantisk disambiguering i vissa kontexter. Vilka dessa kontexter är, återstår att utforska.

En tredje fråga rör bruket av s k ”nakna” nominalfraser, dvs indefinita nominalfraser i singular som saknar kvantitetsattribut. Här skiljer sig svenskan och engelskan ofta åt. En naken nominalfras på svenska sidan kan motsvaras av naken nominalfras även på den engelska sidan, eller av en indefinit icke-naken nominalfras alternativt en definit nominalfras. Se nedanstående exempel:

svensk naken nominalfras - engelsk naken nominalfras: *ge kunskap*
- provide knowledge

svensk naken nominalfras - engelsk icke-naken nominalfras: *på externt laboratorium* - *at an external laboratory*

svensk naken nominalfras - engelsk definit nominalfras: *ha möjlighet att* - *have the opportunity to*

Det tycks inte finnas några klara regler för hur nakna nominalfraser ska hanteras, men heuristiska principer i grammatiken skulle kunna förbättra resultatet.

Sammansättningar av typen **substantiv + bindestreck + och + substantiv**, såsom *infektions- och tumörläkemedel*, skulle behöva hanteras bättre. I grammatiken finns regler för att hantera sammansättningar av denna typ, men dessa regler förutsätter att det första ledet i sammansättningen utgörs av substantivets singulara indefinita form, som i *barn- och ungdomsverksamhet*. I exemplet med *infektions- och tumörläkemedel* avslutas det första ledet istället med ett foge-s, som inte lexikon och grammatik känner igen (annat än som en genitivform). Här kan tänkas att man lägger in fogeformen i lexikonet (vilket tidigare gjorts i bl a SCARRIE-lexikonet). Det kräver dock en hel del arbete att utföra detta, då över 60% av ingångarna i lexikonet utgörs av substantiv.

2.3 Evaluering

När lexikon- och grammatikanpassningen slutförts, genomfördes en automatisk utvärdering både på träningskorpusen och på testkorpusen. Dessutom genomfördes en manuell utvärdering av en av kursplanerna i testkorpusen.

I nedanstående kapitel beskrivs tillvägagångssätt och resultat för de båda utvärderingsmetoderna.

2.3.1 Automatisk utvärdering

Den automatiska utvärderingen genomfördes med hjälp av ett utvärderingssystem utvecklat vid institutionen av Eva Forsbom[For04]. Programmet inkluderar de två utvärderingsmått BLEU[PRWZ01] och NEVA[For03]. Båda dessa mått bygger på en n-gramsjämförelse mellan den automatiskt översatta texten och en eller flera referensöversättningar. Resultatet är ett tal mellan 0 och 1, där 1 ges för översättningar som är identiska med referensöversättningen.

I vårt fall har den engelska delen av kursplanekorpusen använts som referensöversättning. Resultatet av utvärderingen redovisas i tabell 3 nedan.

	Träningskorporus	Testkorporus
BLEU	0,312	0,218
NEVA	0,339	0,255

Tabell 3: Resultat av den automatiska utvärderingen

2.3.2 Manuell utvärdering

Den manuella utvärderingen utfördes på de hundra första meningarna i det otränade materialet. Utvärderingen genomfördes i två etapper.

I den första utvärderingen, graderades meningarna enligt skalan **korrekt - begriplig - svårbegriplig**. Benämningen **korrekt** har de meningar fått som är grammatiskt korrekta, och som förmedlar samma innehåll som källspråkstexten. Meningar som innehåller mindre grammatiska eller lexikala fel (som inte påverkar förståelsen av innehållet), har definierats som **begripliga**, medan meningar klassificerats som **svårbegripliga** i de fall där de innehåller grammatiska och/eller lexikala fel som försvårar förståelsen av översättningen.

I den andra utvärderingsetappen, utvärderades meningarna enligt J2450-måttet[SAE01]. J2450 anger riktlinjer för uppmärkning av översättningsfel med avseende på feltyp. De feltyper som inkluderas i måttet är nedanstående:

- **Felaktig terminologi** (Wrong Term, WT)
- **Syntaktiskt fel** (Syntactic Error, SE)
- **Utelämnande** (Omission, OM)
- **Kongruensfel** (Word Structure or Agreement Error, SA)
- **Felstavning** (Misspelling, SP)
- **Interpunktionsfel** (Punctuation Error, PE)
- **Övriga fel** (Miscellaneous Error, ME)

Resultat Resultatet av utvärderingen enligt skalan korrekt - begriplig - svårbegriplig redovisas i tabell 4.

	Korrekt	Begriplig	Svårbegriplig
Antal meningar	62	16	15

Tabell 4: Resultat av den manuella utvärderingen enligt skalan korrekt-begriplig-svårbegriplig

Av de hundra meningarna, sållades sju stycken bort från utvärderingen, då de innehöll stavfel i källtexten. Dessa sju meningar har inte tagits med i tabellen över resultatet.

Som framgår av tabellen, är 78 av de 93 meningarna att betrakta som fullt begripliga. 62 stycken är helt grammatiskt korrekta, medan endast 15 meningar innehåller sådana fel som försvårar förståelsen av meningen. Dessa problem kan åtgärdas genom fortsatt utveckling av lexikon och grammatik.

I tabell 5 illustreras resultatet av utvärderingen enligt J2450-måttet.

	Antal meningar	Antal fel
WT	9	10
SE	10	10
OM	5	5
SA	9	11
SP	-	-
PE	-	-
ME	-	-

Tabell 5: Resultat av den manuella utvärderingen med J2450-måttet

Av denna tabell kan utläsas att det vanligast förekommande felet utgörs av kongruensfel. Detta handlar främst om substantiv som är identiska i singular och plural form på den svenska sidan, och där fel numerus sålunda har valts i översättningen. T ex har *diskussion av patientfall* översatts till det singulara *discussion of patient case* istället för det plurala *discussion of patient cases*.

De syntaktiska fel som förekommer involverar i de flesta fall prepositionsfraser, som tilldelats fel funktion i satsen, och sålunda placerats på fel ställe i målspråksmeningen. Som exempel kan nämnas meningen *Vid kursens början ges orientering om risker och skydd vid kemiskt experimentarbete*. I denna mening torde den sista prepositionsfrasen (*vid kemiskt experimentarbete*) tolkas som ett efterställt attribut till nominalfrasen *risker och skydd*. Systemet har dock istället tolkat prepositionsfrasen som ett adverbial, och placerat det sist i satsen, vilket ger följande översättning: *At the beginning of the course, orientation about risks and protection is given at chemical*

experimental work, istället för det mer passande *At the beginning of the course, orientation about risks and protection at chemical experimental work is given.*

Felklassen **Wrong Term** inkluderar fel val av preposition, som i **on laboratories** istället för **in laboratories** som översättning av **på laboratorier**.

Omission innefattar framför allt nominalfraser där svenskan saknar indefinit artikel, men denna borde ha inkluderats på den engelska sidan. Detta är exempelvis fallet i frasen *som individuellt skriftligt prov*, som av systemet översatts som *as individual written examination*, medan *as an individual written examination* hade varit att föredra.

2.4 Utveckling av förenklat gränssnitt för körning av översättningssystemet

Per Weijnitz har, inom ramen för pilotprojektet, anpassat det befintliga MATS-gränssnittet, till ett mer användarvänligt gränssnitt. Detta gränssnitt saknar viss funktionalitet, som är till nytta för utvecklare av systemet, men som inte fyller någon funktion för den vanlige användaren. Exempel på sådan funktionalitet är spåringsmöjligheter och illustration av den grammatiska analysen.

Gränssnittet är webb-baserat, och hanterar XML-filer, HTML-filer, SGML-filer och vanliga textfiler. Det finns även möjlighet att skriva in enstaka ord, fraser eller meningar i en textruta.

På resultatsidan visas översättningen, mening för mening. Man kan även välja att klicka på **DETAILS**, för att få se översättningen med färgmarkeringar. Färgmarkeringar uppkommer i de fall där MATS-systemet misslyckats i något översättningssteg, och översättningen sålunda är mindre tillförlitlig. Exempelvis markeras ett ord med röd färg om det saknas i lexikon, medan en grågrön färg antyder att indatan inte kunnat analyseras som ett sammanhängande segment, varvid partiella analyser ligger till grund för översättningen.

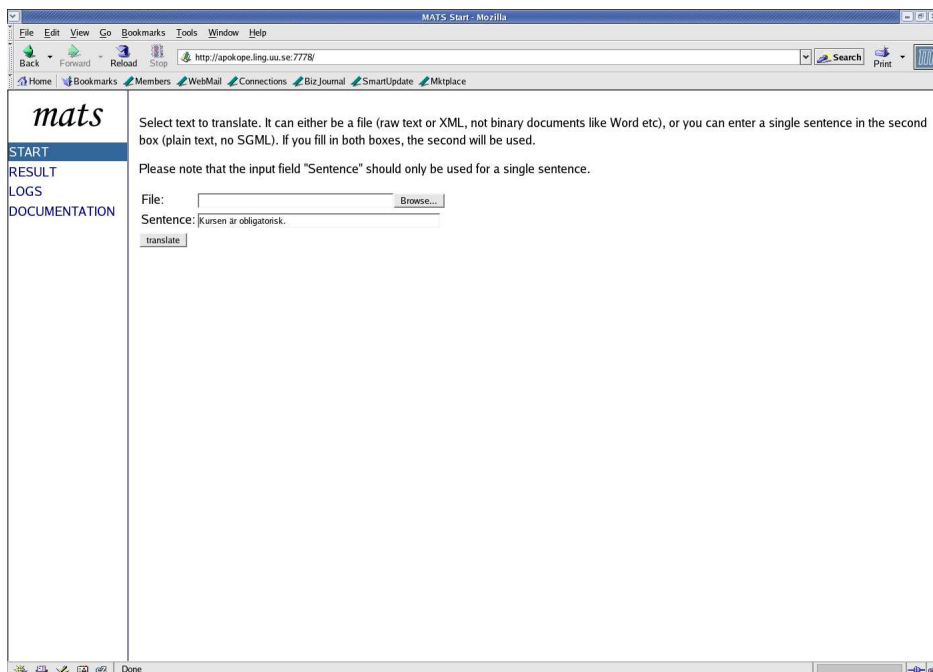
På resultatsidan finns också en länk till en fil där översättningen ges i samma format som den kom in, dvs om man skickat in en HTML-fil, återfinns här samma HTML-fil, fast med texten översatt.

Gränssnittet illustreras i figur 3 och 4.

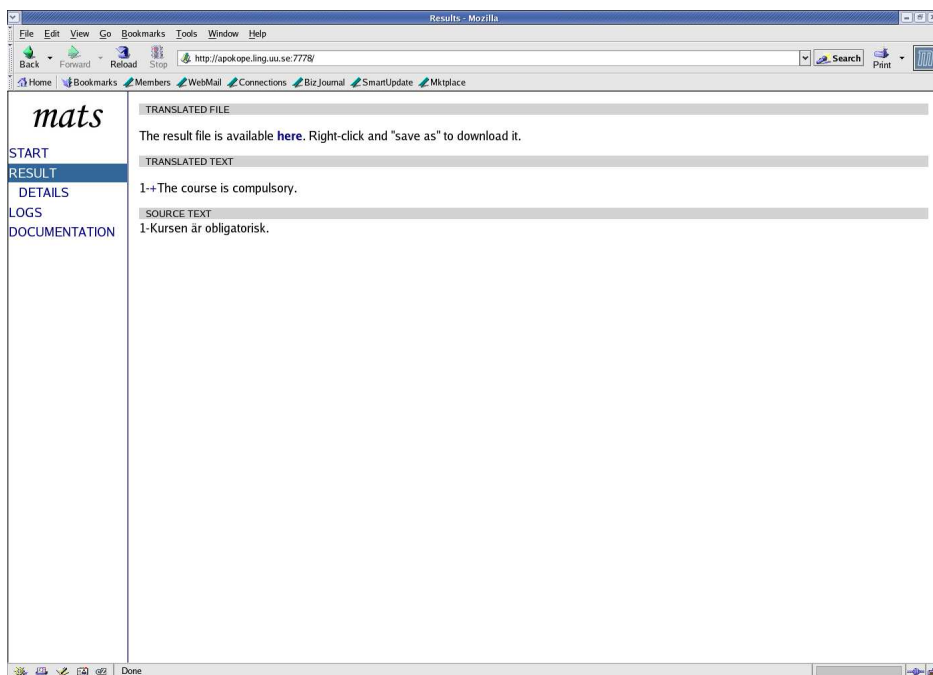
3 Sammanfattande diskussion

Inom ramen för pilotprojektet med maskinöversättning av medicinska och farmaceutiska kursplaner, utvecklades ett lexikon innehållande 10 292 svenska termer, 8 610 engelska termer och 10 248 översättningsrelationer³.

³I samband med grammatikanpassningsarbetet upptäcktes en del brister i lexikonet, och en del nya ingångar lades in, medan andra togs bort. Därför stämmer inte de slutgiltiga lexikonsiffrorna riktigt överens med de siffror som anges i 2.1.5



Figur 3: Det förenklade gränssnittet: startsidan



Figur 4: Det förenklade gränssnittet: resultatsidan

Stora delar av lexikonet kan återanvändas vid en eventuell utvidgning av projektet till att även innefatta kursplaner inom andra vetenskapsområden samt annan utbildningsinformation, såsom utbildningsplaner o dyl. Detta gäller givetvis de termer som markerats som allmänspråkliga i lexikonet, men även en stor del av den domänspecifika terminologin, då här ingår termer rörande utbildning och undervisning i allmänhet, och inte bara inom det medicinsk-farmaceutiska vetenskapsområdet.

Även den grammatiska anpassningen kommer naturligtvis till nytta även i ett utvidgat projekt, då en stor del av grammatikanpassningen varit av generell art.

Utvärderingsresultatet ser lovande ut. 78 av 93 meningar var fullständigt korrekta eller innehöll endast smärre fel. Bara 15 meningar innehöll sådana fel som försvårar förståelsen av meningen. Dessa fel kan åtgärdas genom vidareutveckling av lexikon och grammatik.

För de automatiska utvärderingsmått kan sägas att referensöversättningarnas antal och kvalitet, är av avgörande betydelse för resultatet. Då det finns så många olika sätt att översätta en mening på, ger dessa mått mest tillförlitliga resultat om de tillämpas tillsammans med ett flertal referensöversättningar.

I fallet med kursplanematerialet finns endast en referensöversättning att tillgå. Referensöversättningen är dessutom skev, då här både lagts till, tagits bort och skrivits om. Dessutom innehåller både källtexten och referensöversättningen en hel del stavfel, vilket påverkar utvärderingssiffrorna i negativ riktning.

En korrekt källtext och en bättre referensöversättning skulle sålunda ha givit högre siffror, som bättre speglar den faktiska översättningskvaliteten. De automatiska måtten bör i evalueringssammanhang kompletteras med manuella bedömningar, men ger ändå en värdefull indikation på översättningskvaliteten och är till stor nytta för att avspegla framsteg i utvecklingen av systemet.

Referenser

- [Aka86] Svenska Akademien. *Svenska Akademiens ordlista över svenska språket*. Norstedts Förlag, 1986.
- [Bra00] Thorsten Brants. Tnt - a statistical part-of-speech tagger. Technical report, Saarland University, Computational Linguistics, 2000.
- [For03] Eva Forsbom. Training a super model look-alike: Featuring edit distance, n-gram occurrence, and one reference translation. In *Proceedings of the Workshop on Machine Translation Evaluation: Towards Systemizing MT Evaluation, in conjunction with MT SUMMIT IX*, pages 29–36, 2003.

- [For04] Eva Forsbom. Mt quality evaluation toolbox. Tillgänglig på <http://stp.ling.uu.se/evafo/software/MTQualEvalTool/>, 2004.
- [GP03] Ebba Gustavii and Eva Pettersson. Utveckling av ett svensk-engelskt lexikon för maskinöversättning inom jordbruksdomänen. Tillgänglig på <http://stp.ling.uu.se/~evapet/systran.pdf/>, 2003.
- [Gus04] Ebba Gustavii. Combining resources for automatic extraction of lexical noun entries. Tillgänglig på <http://stp.ling.uu.se/ebbag/GSLT/NLP/Words/Report.pdf>, 2004.
- [Hel01] Jan Hellström. Semantiska särdrag för substantiv i multa. Tillgänglig på http://stp.ling.uu.se/janhell/sem_sardrag.pdf, 2001.
- [Ins04] Karolinska Institutet. Mesh. Tillgänglig på <http://mesh.kib.ki.se/swemesh/swemesh.cfm>, 2004.
- [Läk04] Läkemedelsindustriföreningen. Fass. Tillgänglig på <http://www.fass.se/LIF/home/index.jsp>, 2004.
- [Nor88] Norstedts. *Stora Engelska Ordboken*. Norstedts, 1988.
- [Pre04] Oxford University Press. Oxford english dictionary. Tillgänglig på <http://www.oed.com/>, 2004.
- [PRWZ01] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. Bleu: a method for automatic evaluation of machine translation. Technical Report RC22176, IBM Research Division, T. J. Watson Research Center, 2001.
- [Reg03] Regeringskansliet. Svensk-engelsk ordbok för utbildnings- och forskningsområdet. Technical report, Utbildningsdepartementet, dec 2003.
- [SAE01] SAE. Surface vehicle recommended practice: SAE J2450. Tillgänglig på <http://www.lisa.org/useful/2001/J2450Practice.pdf>, 2001.
- [spr00] Svenska språknämnden. *Svenska skrivregler*. Liber AB, 2000.
- [THA95] Ulf Teleman, Staffan Hellberg, and Erik Andersson. *Svenska Akademiens Grammatik*. Norstedts Ordbok, 1995.
- [Tie03] Jörg Tiedemann. Combining clues for word alignment. In *Proceedings of the 10th Conference of the EACL*, 2003.

- [WHF⁺04] Per Weijnitz, Anna Sågvall Hein, Eva Forsbom, Ebba Gustavii, Eva Pettersson, and Jörg Tiedemann. The machine translation system MATS - past, present & future, 2004.
- [Åbe03] Stina Åberg. Datoriserad analys av sammansättningar i teknisk text. Master's thesis, Uppsala universitet, 2003.