



UPPSALA  
UNIVERSITET

Institutionen för lingvistik och filologi  
Språkteknologiprogrammet  
Examensarbete i datorlingvistik  
27 juni 2006

# Utveckling av ett svensk-engelskt lexikon inom tåg- och transportdomänen

Hans Axelsson, Oskar Blom

Handledare:  
Anna Sägval Hein, Uppsala Universitet  
Eva Petterson, Uppsala Universitet  
Helena Wretman, Explicon AB

## **Sammandrag**

This paper describes the process of building a machine translation lexicon for use in the train and transport domain with the machine translation system MATS. The lexicon will consist of a Swedish part, an English part and links between them and is derived from a Trados translation memory which is split into a training(90%) part and a testing(10%) part.

The task is carried out mainly by using existing word linking software and recycling previous machine translation lexicons from other domains. In order to do this, a method is developed where focus lies on automation by means of both existing and self developed software, in combination with manual interaction. The domain specific lexicon is then extended with a domain neutral core lexicon and a less domain neutral general lexicon.

The different lexicons are automatically and manually evaluated through machine translation on the test corpus. The automatic evaluation of the largest lexicon yielded a NEVA score of 0.255 and a BLEU score of 0.190. The manual evaluation saw 34% of the segments correctly translated, 37%, although not correct, perfectly understandable and 29% difficult to understand.

# Innehåll

|   |          |
|---|----------|
| <b>Sammandrag</b>                                     | <b>2</b> |
| <b>Innehåll</b>                                       | <b>3</b> |
| <b>Tabeller</b>                                       | <b>4</b> |
| <b>Förord och upplägg</b>                             | <b>6</b> |
| <b>1 Inledning</b>                                    | <b>7</b> |
| <b>2 Syfte</b>  | <b>8</b> |
| <b>3 Bakgrund</b>                                     | <b>9</b> |
| 3.1 Domänanpassning inom maskinöversättning . . . . . | 9        |
| 3.1.1 Vad är en domän? . . . . .                      | 9        |
| 3.1.2 Ämnesområde . . . . .                           | 9        |
| 3.1.3 Texttyp . . . . .                               | 9        |
| 3.1.4 Domänneutralitet . . . . .                      | 10       |
| 3.1.5 Varför domänanpassa? . . . . .                  | 10       |
| 3.2 Domänanpassningens delar . . . . .                | 11       |
| 3.2.1 Tvättning av materialet . . . . .               | 11       |
| 3.2.2 Tokenisering . . . . .                          | 11       |
| 3.2.3 Lemmatisering . . . . .                         | 11       |
| 3.2.4 Länkning . . . . .                              | 12       |
| 3.3 Maskinöversättningsystemet MATS . . . . .         | 13       |
| 3.3.1 UCP Light . . . . .                             | 13       |
| 3.3.2 MULTRA . . . . .                                | 14       |
| 3.3.3 MatsLex . . . . .                               | 15       |
| 3.4 Tidigare framtagna resurser inom MATS . . . . .   | 16       |
| 3.4.1 Kursdatabasprojektet . . . . .                  | 16       |
| 3.4.2 Scania . . . . .                                | 16       |
| 3.4.3 Scarrie . . . . .                               | 17       |
| 3.4.4 Jordbrukslexikon . . . . .                      | 17       |
| 3.5 Att utvärdera en testöversättning . . . . .       | 17       |
| 3.5.1 BLEU . . . . .                                  | 17       |
| 3.5.2 NEVA . . . . .                                  | 18       |
| 3.5.3 MT Quality Evaluation Toolbox . . . . .         | 19       |
| 3.5.4 Manuell utvärdering . . . . .                   | 20       |

|          |  |           |
|----------|--|-----------|
| <b>4</b> | <b>Metod</b>                                   | <b>22</b> |
| 4.1      | Förfärdning                                    | 22        |
| 4.2      | Den svenska delen av lexikonet                 | 22        |
| 4.3      | Översättningsrelationer och den engelska delen | 23        |
| 4.4      | Utvärdering                                    | 23        |
| <b>5</b> | <b>Utförande</b>                               | <b>25</b> |
| 5.1      | Hjälpmiddel för att ta fram lexikala resurser  | 25        |
| 5.1.1    | Materialet och dess förutsättningar            | 25        |
| 5.1.2    | Tvättning                                      | 25        |
| 5.2      | Tokenisering                                   | 26        |
| 5.3      | Lematisering                                   | 26        |
| 5.3.1    | Automatiska metoder                            | 26        |
| 5.3.2    | Manuella metoder                               | 28        |
| 5.4      | Ett svenskt lexikon som utgångspunkt           | 28        |
| 5.5      | Länkning                                       | 29        |
| 5.5.1    | Översättningsrelationer och den engelska sidan | 29        |
| 5.5.2    | Dataintegritet                                 | 31        |
| 5.6      | Sammanlagning av lexika                        | 32        |
| 5.6.1    | Extraktion av allmänlexikon                    | 32        |
| <b>6</b> | <b>Resultat</b>                                | <b>33</b> |
| 6.1      | Automatisk utvärdering                         | 33        |
| 6.2      | Manuell utvärdering                            | 35        |
| <b>7</b> | <b>Sammanfattning</b>                          | <b>37</b> |
| <b>8</b> | <b>Framtida utveckling</b>                     | <b>38</b> |
|          | <b>Litteraturlörteckning</b>                   | <b>39</b> |
| <b>9</b> | <b>Appendix</b>                                | <b>41</b> |

## Tabeller

|     |  |    |
|-----|--|----|
| 3.1 | Extrakt ur databasingång i MATS                    | 15 |
| 5.1 | Statistik över träningsmaterialet                  | 26 |
| 5.2 | Exempel på stavfel som förstör statistik           | 27 |
| 5.3 | De 20 mest frekventa löporden i träningsmaterialet | 30 |
| 5.4 | Tabell över ordklassdistribution i BombLex.        | 30 |
| 5.5 | Översikt över lexikonstorlek                       | 32 |

|     |   |    |
|-----|---|----|
| 6.1 | Maskinöversättning i Mats med Allmänlexikon. Neva- och Bleupoäng visas för 28 segment om 150 meningar var. . . . .    | 34 |
| 6.2 | Maskinöversättning i Mats med BombLex. Neva- och Bleupoäng visas för 28 segment om 150 meningar var. . . . .          | 34 |
| 6.3 | Maskinöversättning i Mats med BombLex Extended. Neva- och Bleupoäng visas för 28 segment om 150 meningar var. . . . . | 34 |
| 6.4 | Snittvärden för maskinöversättning med olika lexikon, poäng i Neva och Bleu . . . . .                                 | 35 |
| 6.5 | Övergripande resultat av manuell utvärdering . . . . .  | 35 |
| 6.6 | Feltyper i testdatan . . . . .  | 36 |

## Förord och upplägg

Denna uppsats beskriver bakgrunden till och arbetet med att ur TRADOS-minnen ta fram BombLex, en tvåspråkig baseline bestående av domänspecifika lexikala resurser på engelska och svenska, anpassade för maskinöversättningssystemet MATS (Methodology and Application of a Translation System) som är utvecklat vid Institutionen för Lingvistik och Filologi vid Uppsala Universitet. Uppsatsen kommer inledningsvis att ge en beskrivning av historik inom ämnesområdet och belysa metoder och resonemang som är tongivande i sammanhanget. Därefter beskrivs arbetsgången med att ur träningsdata skapa domänspecifika lexikala resurser från ax till limpa. BombLex utökas till den slutgiltiga BombLex Extended med en allmänlexikal del som består av ingångar som delas av ett flertal domäner. Slutligen genomför vi ett flertal körningar med olika lexikon på testdatan. Detta utvärderas sedan med hjälp av utvärderingsmåttene NEVA och BLEU. Avslutningsvis presenteras förslag på tänkbara fortsättningar inom utvecklingsarbetet med resurserna.

Arbetet är utfört under januari 2005 till och med april 2006. Arbetsinsatserna består av tre delområden: ett gemensamt och två personliga. Det gemensamma arbetsområdet har präglat hela arbetsgången och har bestått av planering och upplägg, insamling av referenser, den automatiska länkningen, kontinuerlig lexikonanalys och justering, manuell länkning och lemmatisering samt allmän bruksprogrammering. De personliga arbetsområdena innebär att vi fokuserat individuellt på vissa komponenter under arbetsgången. Det finns dock inga vattentäta skott mellan dessa, eftersom vi kontinuerligt har diskuterat och hjälpt varandra. Oskars individuella arbetsinsatser har bestått av programmering av databasskript för lemmatisering, den automatiska utvärderingen, testkörningar, sammansättningsanalys och skript för behandlig av länkingsresultat. Hans individuella arbetsinsatser har bestått av tvättning och tokenisering av materialet, den manuella utvärderingen, sammanställning av statistik, versionshantering och typsättning av uppsats.

Våra främsta tack går till Anna Sågvall Hein och Eva Pettersson för handledning, rådgivning och teknisk assistans. Tack även till Brita Norlén för omsorgsfull korrekturläsning och Eva Forsbom för hjälp med utvärderingsformalismerna. Vi vill även tacka Anna och Åsa tillsammans med våra goda vänner Örjan Berglund, Eva Ericsson, Anna Hedström, Jens Moberg, och Peter Strömbäck.

# 1 Inledning

Maskinöversättning är en disciplin som givits stort forskningsutrymme sedan de första datorerna introducerades. Tanken var att datorn skulle göra det möjligt att snabbt och effektivt kunna översätta stora mängder text med relativt små insatser. Det naturliga språket visade sig dock vara svårhanterligt ur flera synvinklar och bristen på effektiva angreppssätt har under lång tid lett till att kvaliteten på maskinella översättningar inte levt upp till förväntningarna. (Hutchins och Somers, 1992) De senaste årtiondena har forskningen inom detta ämnesområde utmynnat i en rad olika framgångsteorier, dvs. olika sätt att förbättra resultatet av maskinell översättning. Domänanpassning är en av de främsta framgångsteorierna och innebär att avgränsa översättningsmaterial till specifika områden, domäner, med avseende på ämnesområde och texttyp. Genom att dela upp maskinöversättningens resurser i olika områden, eller domäner, elimineras eller förminskas en del av de företeelser som tidigare legat till grund för dålig översättning. Det är viktigt att komma ihåg att maskinöversättning är en mycket svår uppgift av den enkla enledningen att översättning i sin helhet är svårt, även för människor.

## 2 Syfte

Syftet med examenarbetet är att göra en utvidgning av resursbanken tillhörande maskinöversättningsystemet MATS genom att utvinna lexikala resurser specifika för att hantera domänen tågtrafik. För detta har vi genom översättningsföretaget Explicon rekviderat ett omfattande parallellkorpusmaterial i form av översättningsminnen. Materialet har delats in i test- respektive träningsdata, där vi genom att extrahera kunskap ur träningsdatan skapar en uppsättning lexikala resurser, som sedan används av MATS för att maskinöversätta testdatan. Som gränsvärde för vilka ord som extrahe- ras ur träningsmaterialet sätts ett tröskelvärde på 3 lemmaförekomster.

Vi har valt att koncentrera domänanpassningen till vokabulär, och inte skapat do- mänspecifika syntaktiska regelverk. Detta dels pga. att en sådan uppgift skulle utöka omfattningen av vårt examensarbete alldeles för mycket, men även av anledningen att MATS-systemet saknar specifika grammatiska regler för respektive domäner. Sys- temet har tidigare tränats att hantera domänerna lastbilstillverkning, jordbruk och utbildningsinformation och det finns resurser som till viss del är lämpliga för åter- vinning i form av lexikondatabaser och transferregler. Resultatet kommer att utgöra en resursmässig baseline bestående av domänspecifika lexikala resurser på engelska och svenska med upprättade översättningsrelationer dem emellan. Till varje språksi- da kommer det även att finnas en uppsättning mönsterord för att hantera morfologi på respektive språk.

Vi kommer att utvärdera en testkörning och diskutera möjlig vidareutveckling och förbättring av metoder och resurser. Arbetet utförs i samarbete mellan Institutio- nen för Lingvistik, Bombardier Transportation och Explicon.

## 3 Bakgrund

### 3.1 Domänanpassning inom maskinöversättning

#### 3.1.1 Vad är en domän?

Begreppet språklig domän lanserades i slutet av 1960-talet av en forskargrupp vid New Yorks Universitet, ledd av Naomi Sager och har sitt ursprung i den amerikanske lingvisten Zellig Harris' matematiska teori om relationen mellan ett standardspråk och mängden av dess subspråk, sk. domäner (Somers, 2003). Bakgrunden är antagandet att ett standardspråk innefattar specialiserade domäner, dvs en mängd subspråk, som används av talare inom tydligt avgränsade områden med avseende på ämnesområde och texttyp. Dessa båda begrepp är nyckelord för att beskriva och kategorisera en domän och behandlar lexikala respektive syntaktiska egenskaper hos texter.

#### 3.1.2 Ämnesområde

Ett språk kan ha ett subspråk som utskiljer sig genom sitt ämnesområde, dvs. domänen bestäms av vad innehållet i texterna handlar om. En texts ämnesområde gestaltar sig tydligast i textens lexikala särdrag. Det innebär att klart avgränsade och tydliga domäner har en särpräglad vokabulär som är starkt knuten till ämnesområdet i fråga, t.ex talar man ofta om kemispråk, läkarspråk, juristspråk och militärspråk. I denna typ av fall finns det en gemensam, bakomliggande kunskap hos gruppen talare, som tar sig uttryck genom en karaktäristisk specialvokabulär. Domänen har skapats som ett resultat av att en grupp specialister kommunicerar med varandra. Den gemensamma kunskapen delas oftast inte med talare av standardspråket, och subspråket kan ofta uppvisa stor variation gentemot standardspråket. Gränsen mellan standardspråk och domän blir extra tydlig genom att ord kan ha vitt skilda betydelser i olika domäner såväl som i jämförelse med standardspråket, t.ex. det engelska ordet *process*, som har en skild betydelse inom datavetenskap resp. juridik, och ordet *car* som ska tolkas som antingen tågagn eller bil, helt beroende av kontexten.

#### 3.1.3 Texttyp

Som synes kan lexikala egenskaper karaktärisera ett subspråk, men det finns även syntaktiska konstruktioner som på ett tydligt sätt avgränsar domäner ifrån standardspråket. Detta visar sig genom att några konstruktioner är vanligare än andra i en viss domän, medan somliga konstruktioner helt kan saknas. Begreppet texttyp avser att kategorisera en text med utgångspunkt i dess syntax och dess syfte. Exempel på olika texttyper är instruktioner, beskrivningar och rapporter. De har alla olika syften. Syftet med en instruktion är ofta att vägleda vid ett genomförande, en beskrivning

syftar till att avbilda en produkt eller ett förfarande och en rapport till att beskriva en händelse eller ett läge. Olika texttyper kan uppvisa stora skillnader vad gäller syntax. Tydliga exempel på detta är recept eller andra typer av instruktioner, där det ofta återfinns många imperativformer samtidigt som det saknas fråge- och utropkonstruktioner. I texttypen nyhetsbrev kan däremot båda sistnämnda typer av konstruktioner vara vanligt förekommande, medan imperativer är mycket ovanliga. Nyhets- och väderrapporter tenderar ofta att innehålla futurumkonstruktioner medan annonstexter är rika på konstruktioner där både verb och hjälpverb utelämnats. Lagspråk däremot och framförallt avtalstexter är exempel på en domän som kännetecknas av en oerhört komplex och elaborerad syntax.

### 3.1.4 Domänneutralitet

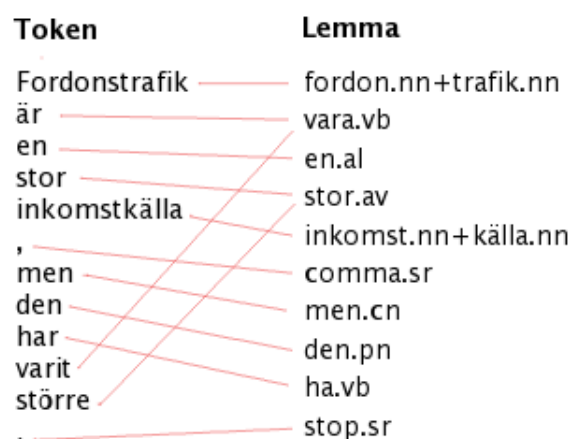
Tillsynes kännetecknas språkliga domäner av syntaktiska och lexikala särdrag som separerar dem ifrån varandra. Men även om domäner särskiljer sig genom sina olikheter, delar de även gemensamma drag. Dessa drag kan sägas vara allmänspråkliga. I motsats till domänspecifika, fackspråkliga uttryck, kan allmänspråkliga uttryck sägas vara oberoende av kontext. Detta innebär att det allmänspråkliga uttrycket inte har en begränsad användning i sig självt samt att det inte implicerar något konkret. T.ex verbet 'klippa' inte domänneutralt eftersom det implicerar 'sax'. (Moberg, 2005)

Enligt Moberg (2005) skiljer man på resurserna allmänlexikon respektive kärnlexikon. Urvalskriterierna för ett allmänspråkligt lexikon bygger på ett antagande om frekvens, där ett ord ska tillhöra de  $n$  vanligaste orden i  $n$  stycken korpusar. Genom att filtrera det allmänspråkliga lexikonet genom semantiska urvalskriterier erhålls ett kärnlexikon, vars ingångar blir ännu mer domänneutrala. Tanken med kärnlexikonet är att tillföra en rörlig lexikal basresurs i MATS, som delas av samtliga domäner.

### 3.1.5 Varför domänanpassa?

Naturligt språk har en oerhörd uttrycks kraft, men det innebär också att det samtidigt kan vara ambiguöst och flerbottnat. Tolkning och betydelse är starkt beroende av kontextuella parameterar, och ytformen av ett språk är relativt uttrycksfattig. Inom maskinöversättning är detta ett stort problem, eftersom det inte finns möjlighet att ta hänsyn till alla omständigheter kring ett yttrande. Det har visat sig att domänanpassning är ett av de framgångsrika angreppssätten för att förbättra resultatet av maskinell översättning mellan språk. (Somers, 2003). Genom att avgränsa översättningen till en specifik domän reduceras en stor del av de problem som ligger till grund för felaktiga maskinöversättningar. Inom domänen får varje ord en snävare tolkning, vilket leder till färre ambiguiteter och högre antal korrekta översättningsresultat. En vanlig tankegång är att vokabulären i specifika domäner tenderar att ha en begränsad storlek. I praktiken skulle detta innebära att det är möjligt att extrahera alla ord som kan tänkas dyka upp inom en viss domän. Detta är endast delvis sant, eftersom det med stor sannolikhet kommer att dyka upp nya namn och benämningar på komponenter och metoder(Somers, 2003).

Figur 3.1: Relationen token - lemma



## 3.2 Domänanpassningens delar

### 3.2.1 Tvättning av materialet

Tvättning innebär att förfina råmaterialet till en mer processvänlig form. I detta fall utgörs träningsmaterialet av översättningsminnen, framtagna av Bombardier Transportation med hjälp av översättningsverktyget Trados Translators Workbench. Detta är en kommersiell programvara som tillsammans med Microsoft Word används inom professionell översättning för att dela översättningar mellan medarbetare inom företag. Ett flertal översättningsföretag använder det för att snabbt bygga upp stora resurser genom att återanvända redan gjorda översättningar. Råmaterialet som levererats är i binärformat och innehåller mycket information som inte är av intresse för vår uppgift. Det är därför nödvändigt att tvätta texten för att i största möjliga mån bli av med onödig data.

### 3.2.2 Tokenisering

Tokenisering är processen att dela in en text i enheter. Den kan genomföras på ett flertal olika sätt, men vanligtvis låter man textuella avgränsare vara interpunktionstecken, radbrytning och blanksteg. Genom att göra detta kommer man åt att isolera mellanliggande strängar. Även interpunktionstecken och mellanslag ses i förlängningen som en grupp tokens. Tokenisering är det första steget mot att extrahera potentiella kandidater till lexikonet.

### 3.2.3 Lemmatisering

Inom lexikalisk teori omfattar begreppet lemma ett ords grundform eller uppslagsform och representerar klassen av samhöriga böjningsformer, stavningsformer och uttalsformer.(NE, 2005) Lemmatisering är en form av lingvistisk bearbetning som går ut på att från en tokeniserad syntaktisk enhet och dess böjningsformer extrahera tillhörande lemma, se förhållande i figur 3.1.

Lemmatisering är i vårt fall en automatisk såväl som manuell process, automatisk i den bemärkelsen att det görs automatiska slagningar i befintliga databaser där

man tillsammans med sammansättningsanalys tar vara på lemmainformation som kan vara till nytta. Det manuella arbetet består i att förse lexikoningångar med språklig information och är även delvis automatiserat.

### **Sammansättningsanalys**

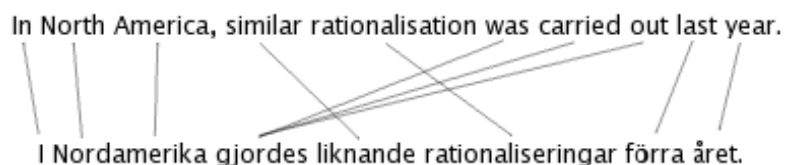
En viktig del i arbetet med att framställa ett lexikon är att under lemmatiseringen identifiera sammansättningar av ord. Detta är nödvändigt för att automatiskt kunna tilldela böjningsinformation till sammansättningarna. Sammansättningsanalysen genomförs helt automatiskt. För att välja ut de mest sannolika substantivsammansättningarna används UCP:s sammansättningsanalys i kombination med ett urvalsprogram utvecklat av Stina Åberg vid Institutionen för Lingvistik och Filologi (Åberg, 2003). Programmet väljer den bästa av flera möjliga analyser. Varje potentiell analys genererad av UCP:s sammansättningsanalysator behandlas av urvalsprogrammet. Om fler än en analys per lemma påträffas, tillämpas speciella urvalsregler på dessa och de mest sannolika sammansättningarna väljs ut. Urvalsreglerna består i huvudsak att jämföra dels antalet sammansättningsled och dels stränglängden hos dessa. I första fallet väljs det lemma som har minst antal led ut. Om antalet led är lika jämförs de första delarna i de tekniska stammarna och den längsta väljs. Om dessa är lika långa, jämförs den andra delen av lemmat. Eftersom UCP:s sammansättningsanalysator i nuläget bara stöder analyser av ord vars samtliga konstituenten är substantiv, kan bara sådana ord identifieras.

### **3.2.4 Länkning**

Länkning är ett kärnbegrepp inom maskinöversättning och andra lexikografiska applikationer. Dess syfte är att med statistiska och i vissa fall lingvistiska mått och metoder identifiera relationer mellan lexikala enheter i parallellkorpusmaterial för att på så sätt extrahera översättningsekvivalenter mellan två språk. Resultatet kan användas vid lexikonskapande, där det kan vara mycket tidsbesparande att ha tillgång till högkvalitativ ordlänkning. Manuell lemmatisering är nämligen både tidskrävande och tenderar att vara enormt repetitiv och därigenom mottaglig för felaktig input från användaren.

Det finns flera typer av ordlänkningssystem, vilka alla delar det gemensamma målet att identifiera kopplingar mellan lexikala enheter i parallellkorpusar. När det handlar om att skapa tvåspråkiga lexikon är det framförallt innehållsenheter, dvs fraser, termer och innehållsord som är intressanta. Länkar mellan funktionsord är irrelevanta för skapandet eller utvidgningen av tvåspråkiga lexikon. (Ahrenberg m.fl., 2000). Det räcker därför inte med att länka varje ord i källspråket, utan man måste ta hänsyn till parametrar såsom kontext, ordklass, funktion, frekvens osv. Ordlänkningssystem länkar ord till ord men det finns också ett behov av att hantera länkar mellan enstaka ord och flerordsenheter. Detta är en nödvändighet i fallen där lexikala enheter inte kan delas in i separata ord med korrekt översättning på källspråket, se figur 3.2.

I stort finns det två övergripande infallsvinklar för att genomföra maskinell ordlänkning. Den första grundar sig på idén med en statistisk översättningsmodell, där man modellerar länkning som dolda parametrar i en Markovmodell. De främsta företräddarna för detta är Franz Josef Och samt Hermann Ney. Den andra, lingvistiska, infallsvinkeln baserar sig på stränglikhetsmått språken emellan och associationstester



**Figur 3.2:** Länkar mellan ord och flerordsenheter.

som berör samförekomst och kontext mm. (Tiedemann, 2003) Som tidigare nämnts finns det en stor mängd mjukvara för att genomföra länkning och i detta arbete används länkmjukvaran Clue Aligner.

### Clue Aligner

Clue Aligner är en automatisk länkare utvecklad av Jörg Tiedemann på Institutionen för Lingvistik och Filologi, Uppsala Universitet. För att fastställa relationer mellan ord och fraser på käll- och målspråk används en kombination av så kallade "clues" - ledtrådar, som baseras på statistiska och lingvistiska mått och som består av associations-sannolikheter. Exempel på "clues" är viktade mått på samförekomst, stränglikhet och frastyp. Clue Aligner är konfigurerbart i stor omfattning och använder sig av ett flertal moduler, däribland Giza++, som är en del av verktygssamlingen EGYPT utvecklad vid Center for Language and Speech Processing vid Johns Hopkins University i Baltimore.

Ett problem med automatisk ordlänkning är tokenisering. Många översättningsrelationer innehåller komplexa enheter, såsom idiomatiska och sammansatta enheter. Eftersom tokenisering därmed ibland förutsätter länkning, kan token-enheter komma att ändras under själva länkningen. Därför kombinerar Clue Aligner en enklare förtokenisering med en dynamisk tokenisering under länkningen.

Ytterligare ett problem med traditionell ordlänkning är den totala avsaknaden av lingvistisk information. Clue Aligner, däremot, använder sig av ordklassstagning och syntaktisk (yt)parsning, som mynnar ut i en uppsättning "clues".

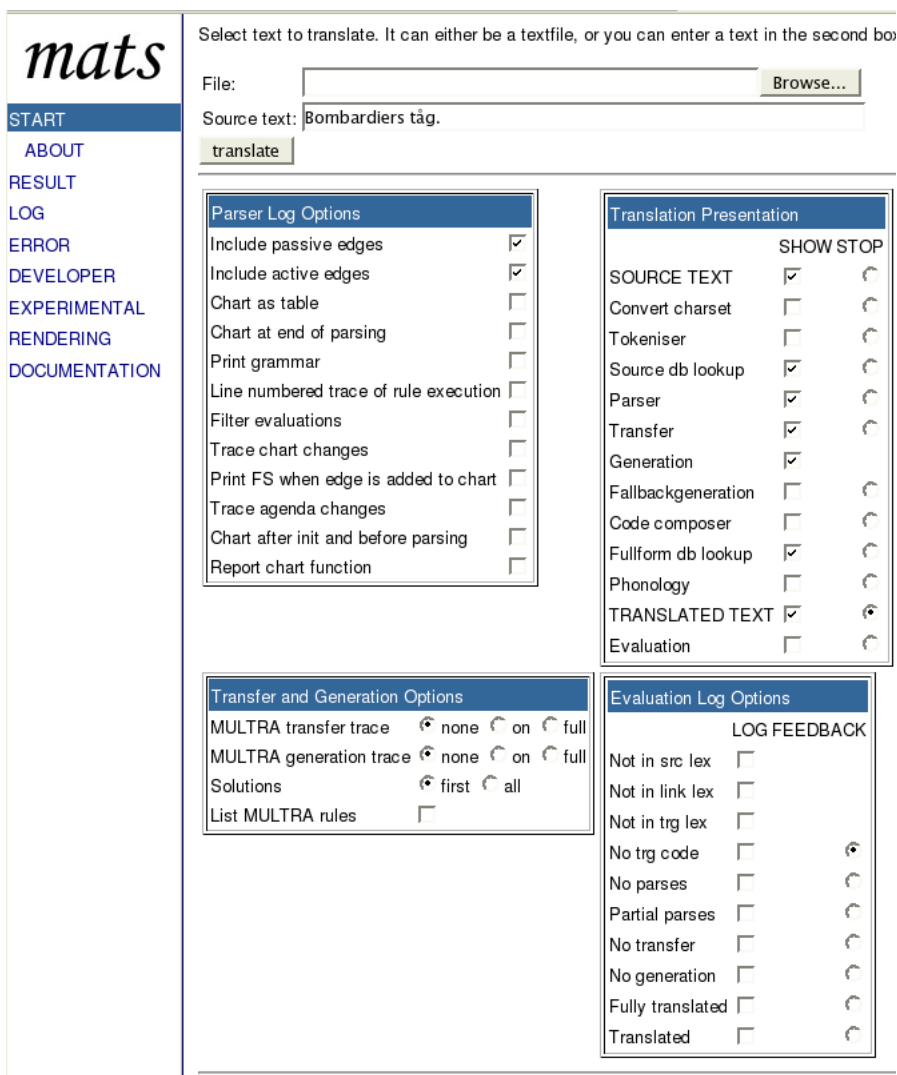
För att sammanställa länkarna ställs en 2-dimensionell "clue"-matris upp, där raderna och kolumnerna utgörs av mål- respektive källspråksenheter från samma meningssegment. Värdena i matrisen utgörs av en kombinerad sannolikhet för varje möjligt ordpar. Utifrån länkkandidaterna väljs sedan de mest sannolika ut.

## 3.3 Maskinöversättningssystemet MATS

MATS är ett fullt automatiserat maskinöversättningssystem utvecklat i samarbete mellan Institutionen för Lingvistik och Filologi vid Uppsala Universitet, Scania CV AB och översättningsföretaget Explicon AB (Sågvall Hein m.fl., 2003). Systemet är helt modulärt och består av tre huvudkomponenter: Uppsala Chart Parser Light, MULTRA och MatsLex.

### 3.3.1 UCP Light

UCP Light (Uppsala Chart Parser Light) är en chartparser utvecklad av Per Weijnitz vid Institutionen för Lingvistik och Filologi vid Uppsala Universitet. Den är baserad



Figur 3.3: Användargränssnitt i maskinöversättningssystemet MATS

på UCP2, en tidigare LISP-implementation av en chartparser introducerad av Sågvall Hein 1980. UCP Light började utvecklas inom projektet Scarrie, eftersom det till programmet ScarCheck fanns behov av en snabb och fristående parser-modul. UCP Light är implementerad i C och cirka 21 gånger snabbare än sin föregångare UCP2.(Weijnitz, 2002)

### 3.3.2 MULTRA

MULTRA (Multilingual Support for Translation and Writing) är en unifieringsbaserad forskningsprototyp för maskinöversättning som utgör grunden i maskinöversättningssystemet MATS. Multra är baserat på transfer-modellen och översättningar görs inom väl avgränsade domäner, där varje domän har sin egen uppsättning lexikala resurser och där de olika domänerna delar en gemensam uppsättning syntaktiska regelverk. Eftersom MULTRA är ett regelbaserat system i det traditionella transferparadigmet består det av fyra större moduler: analys, selektion, transfer och gene-

**Tabell 3.1:** Extrakt ur databasingång i MATS

| lemma  | teknisk stam | mönsterord | affixlista               | realiseringsregel |
|--------|--------------|------------|--------------------------|-------------------|
| bil.nn | bil          | STOL       | en, ens, ar, arna, arnas | \$                |

rering. Varje moduls input och output är textbaserad, och modulerna kommunicerar sinsemellan via ett protokoll. Detta möjliggör enkel spårning och god överskådning, i och med att alla delresultat i en översättningprocess är åtkomliga och läsbara (Sågvall Hein m.fl., 2003)

### 3.3.3 MatsLex

MatsLex är en flerspråkig lexikal databas som används inom maskinöversättnings-systemet MATS. Vid en översättning inom en specifik domän måste s.k. 'run-time'-lexikon utifrån domänens lexikala databas kompileras. Ur de kompilerade lexikonerna kan sedan relevant lexikal information extraheras av respektive översättningsmodul.

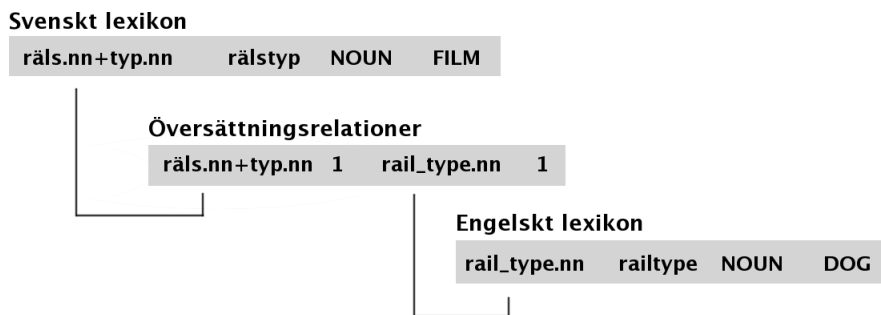
Fördelarna med det här tillvägagångssättet är att varje modul i översättningsprocessen får tillgång till ett lexikon på ett format som passar just den modulen. Samtidigt innehåller inget 'run-time'-lexikon någon överflödigt data utan bara för ändamålet relevant information. Detta ger upphov till snabb dataåtkomst, vilket i sin tur minskar tidsåtgången för hela översättningsprocessen. Ytterligare en fördel är möjligheten att kunna administrera/modifiera lexikondatabasen utan att påverka en pågående översättning. Eftersom kompileringen av lexikon sker vid en viss tidpunkt märks inte eventuella ändringar i databasen vid en översättning förrän vid nästa kompilering. Dessutom underlättar lexikonet vid jämförelse av resultat, pga att flera olika lexikon som är kompilerade vid olika tidpunkter och som därför utgör olika instanser av databasen, kan användas (Tiedemann, 2002).

#### Lexikondatabasens struktur

Lexikonets interna struktur är uppbyggd med en relationell databas som modell och är implementerad i database management-systemet MySQL. Varje domänspecifikt maskinöversättningslexikon innehåller ett flertal separata enspråkiga lexikon. Databasens enspråkiga delar består av diverse tabeller, som i sin tur innehåller morfologisk, syntaktisk och semantisk information. För att minska redundansen i lexikonet och för att underlätta inlägg innehåller databasen inga explicita fullformer, utan använder sig av mönsterord för att beskriva tex ett substantivs deklination eller ett verbs konjugation. Detta innebär att för varje lemma i databasen finns en koppling till en eller flera tekniska stammar samt till ett specifikt mönsterord. Mönsterordet är i sin tur kopplat till en uppsättning av affix med tillhörande morfologiska genereringsregler, som består av reguljära uttryck för just det mönsterordet. När det gäller språkkombinationen svenska och engelska handlar det i många fall om att konkatenera en teknisk stam med ett suffix. På så vis kan korrekta ordformer av ett specifikt lemma genereras utan att samtliga former av ett lemma behöver lagras individuellt, se 3.1.

Ett potentiellt problem med den här typen av lexikonstruktur är att den inte är lämplig för alla språk, i värsta fall kan varje ord vara sitt eget mönsterord och därmed ha ett eget böjningsparadigm. (Tiedemann, 2002).

För varje enskilt enspråkslexikon i maskinöversättningslexikonet föregås tabellerna tillhörande det specifika språket av ett språkunikt prefix. T ex svStem, enStem.



**Figur 3.4:** Kopplingen mellan svenskt lexikon, översättningsrelationer och engelskt lexikon.

Denna struktur ger upphov till möjligheten att lätt kunna lägga till ett extra enspråkigt lexikon för en viss domän. Ytterligare en fördel är att översättningslänkar lätt kan skapas genom att man parar ihop ett lexem från ett språk med ett lexem från ett annat. Detta görs i en separat tabell som innehåller själva länken mellan de olika lexemen, se figur 3.4, samt optionell information som ursprung och kommentar.

### 3.4 Tidigare framtagna resurser inom MATS

Det har tidigare genomförts fyra större projekt inom MATS vid Institutionen för Lingvistik och Filologi. Dessa har resulterat i lexikon och grammatiska regelverk som båda kommer att återanvändas i detta arbete.

#### 3.4.1 Kursdatabasprojektet

Syftet med Projekt Kursdatabas är att samla utbildningsinformation såsom kursplaner och utbildningsplaner i en universitetsgemensam databas. I denna databas ska all information finnas tillgänglig både på svenska och på engelska. (Pettersson, 2005). En del av projektet är inriktat på att med hjälp av MATS maskinellt översätta kursplaner mellan svenska och engelska, och har mynnat ut i en pilotstudie inriktad på att ta fram resurser för maskinöversättning inom inom det medicinska och farmaceutiska vetenskapsområdet. Projektet löpte under perioden maj 2004 - januari 2005 på Institutionen för lingvistik och filologi i vid Uppsala Universitet och resulterade i bla. ett lexikon innehållande 10 292 svenska termer, 8610 engelska termer och 10 248 översättningsrelationer. Ett särdrag hos detta lexikon är att allmänspråkliga ingångar, dvs delar som delas av ett brett antal domäner, är uppmärkta, något som användes i ett senare skede av arbetet. Detta lexikon kallas fortsättningsvis i uppsatsen för 'KursLex'.

#### 3.4.2 Scania

Scania-lexikonet har utvecklats i ett samarbete mellan Institutionen för Lingvistik och Filologi och Scania CV AB. Lexikonet grundar sig på teknisk skrift i form av lastbilsmanualer och var det första domänspecifika lexikonet inom ramen för MATS-projektet. Lexikonets svenska sida består av ca. 20 000 lemman. Detta lexikon kallas fortsättningsvis i uppsatsen för 'ScaniaLex'.

### 3.4.3 Scarrie

Scarrie är ett program för stavningskontroll och viss grammatikkontroll. Det har utvecklats i ett EU-samarbete med Danmark, Norge och Sverige. Den svenska delen har utvecklats vid Institutionen för Lingvistik vid Uppsala Universitet med hjälp tidningsmaterial från de båda svenska tidningarna Svenska Dagbladet och Upsala Nya Tidning. Projektet har koordinerats av WordFinder Software AB i Växjö. Scarrie innehåller ca. 150 000 svenska lemman men saknar översättningsrelationer och engelska lemman. Detta lexikon kallas fortsättningsvis i uppsatsen för 'LingLex'.

### 3.4.4 Jordbrukslexikon

Jordbrukslexikonet är en produkt av ett samarbete mellan Institutionen för Lingvistisk vid Uppsala Universitet och EC Systran och utgör den lexikala basen i en svensk-engelsk Systran-prototyp. Lexikonet bygger på två meningslänkade korpusar som ställts samman av EC systran om ca 900 000 ord (Gustavii och Pettersson, 2003). och innehåller 6304 svenska lemman, 4991 engelska, och 5923 översättningsrelationer. Detta lexikon kallas fortsättningsvis i uppsatsen för 'JordbruksLex'.

## 3.5 Att utvärdera en testöversättning

Ett viktigt delområde inom såväl manuell som maskinell översättning är naturligtvis att genomföra utvärderingar och fastställa kvaliteten på översättningar. Det är i det här sammanhanget eftersträvansvärt att genomföra denna typ av utvärderingar på automatiserad väg, eftersom manuell utvärdering tenderar att vara både kostsam och tidsödande. (Forsbom, 2003). Detta är ett allvarligt problem för den moderna maskinöversättningsindustrin, eftersom det ofta finns ett behov av daglig utvärdering för att skilja bra idéer från dåliga i utvecklingsarbetet (Papineni m.fl., 2001). I detta arbete används två framträdande ramverk för att automatiskt utvärdera maskinell översättning, BLEU och NEVA.

### 3.5.1 BLEU

BLEU (Bilingual Evaluation Understudy) är ett N-grambaserat mått som ofta används inom maskinöversättning. Det är baserat på medelvärdet av antalet n-grammatchningar mellan en maskinellt föreslagen översättningskandidat och ett antal referensöversättningar. BLEU stämmer väl överrens med mänsklig uppfattning vad gäller graden av korrekthet och flyt i översättningar (Papineni m.fl., 2001). Måttet BLEU räknar antalet n-gram hos kandidatöversättningen som förekommer hos referensöversättningen och ger dem samma vikt. Om kandidaten är kortare än referensen ges en straffpoäng(BP), se formel 1.

$$\text{BLEU} = \text{BP} \cdot \exp \left( \sum_{n=1}^N w_n \log p_n \right) \quad (1)$$

där

$$\text{BP} = \begin{cases} 1 & \text{om } c > r \\ e^{(1-\frac{r}{c})} & \text{om } c \leq r \end{cases}$$

$r$  = referensmeningens längd  
 $c$  = kandidatmeningens längd

$$N = 4$$

$$w = \frac{1}{N}$$

För att beräkna n-gramprecision jämförs referensmeningens mot kandidatmeningen:

$$p = \frac{\sum_{C \in \{Cand\}} \sum_{n\text{-gram} \in \{C\}} \text{Count}_{clip}(n)}{\sum_{C \in \{Cand\}} \sum_{n\text{-gram} \in \{C\}} \text{Count}(n)}$$

Det finns dock svagheter med BLEU när man genomför utvärdering på segmentnivå:

1. Måttet hanterar inte meningsssegment som är kortare än det längsta definierade n-grammet (vanligtvis fyra ord) på ett rättvist sätt.
2. BLEU använder det geometriska medelvärdet, vilket resulterar i 0 poäng om det inte finns någon matchning i någon av n-gramklasserna (t.ex. trigram), även om det finns matchningar i andra klasser (t.ex. i bigram).

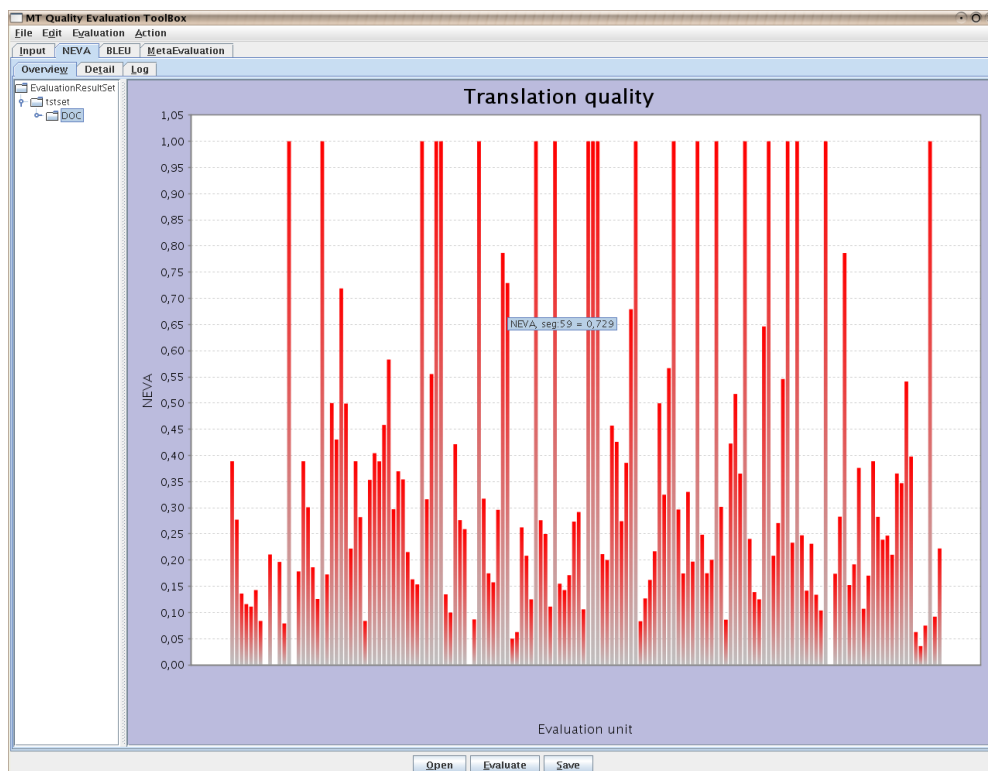
Dessa svagheter leder till utvärderingsproblem med vissa typer av översättningsmaterial, framförallt tekniska manualer som kännetecknas av korta meningsssegment t.ex. listor och tabeller (Forsbom, 2003).

### 3.5.2 NEVA

Måttet NEVA är en omdefiniering av BLEU, genomförd av Eva Forsbom vid Institutionen för Lingvistik och Filologi vid Uppsala Universitet. Syftet med NEVA är att ta fram ett mätverktyg som inte begränsas av ovan nämnda problem. Genom att använda ett aritmetisk medelvärde istället för ett logaritmiskt i BLEU (dvs, genom att funktionerna EXP och LOG ur BLEU utesluts) kan man få en mer rättvisande utvärdering på meningstyper som tidigare gav missvisande värden. Liknande modifieringar av BLEU har även genomförts av Stephan Oepen och Erik Velldal inom utvärdering av maskinöversättningsystemet LOGON (Velldal och Oepen, 2005):

$$\text{NEVA} = \text{BP} \cdot \sum_{n=1}^N w_n p_n \quad (2)$$

där



Figur 3.5: Översikt över utvärderade meningssegment i MT Quality Evaluation Toolbox.

$$N = \begin{cases} N_{max} & \text{om } c \geq N_{max} \\ c & \text{om } c < N_{max} \end{cases}$$

Båda utvärderingsmetoderna ger ett resultat mellan 0 och 1, där 1 ges för översättningar som är identiska med referensmeningen. Det är viktigt att komma ihåg att automatisk utvärdering av översättning är mycket strikt och dogmatiskt. I de fall där maskinöversättningssystemet producerat en målspråksmening som inte delar lexikala enheter med någon av referensmeningarna ges samma fel som vid en felöversättning. Graden av synonymi som täcks in är alltså helt avhängig antalet referensmeningar, som i detta fall endast är en per källspråksmening. Av den anledningen kompletteras resultaten med en manuell utvärdering av en del av översättningarna.

### 3.5.3 MT Quality Evaluation Toolbox

MT Quality Evaluation Toolbox är ett program för automatisk evaluering av maskinöversättning utvecklat av Eva Forsbom på Institutionen för Lingvistik och Filologi, Uppsala Universitet. För utvärdering krävs en SGML-fil med tre komponenter: en med källspråkssegment, en med referensöversättning samt en med maskinöversättning. Utvärderingen görs genom en jämförelse mellan referensöversättningen och maskinöversättningen och tillåter en uppsättning användarbestämda statistiska mått. Både Neva- och Bleuresultatet presenteras i form av grafiska diagramstaplar där utvärderingen av varje segment kan beskådas enskilt 3.5.3.

### 3.5.4 Manuell utvärdering

Den manuella utvärderingen syftar till att ge en mer nyanserad bild av översättningskvaliteten än vad automatisk utvärdering kan göra. Neva- och Bleu kan som tidigare nämnts ge låga poäng när en översättning innehåller lexikala enheter som inte förekommer i referensmeningen, trots att den språkmässigt sett kan vara helt korrekt. Nedan ges tre exempel på meningar som i en manuell testutvärdering skulle givits graden *korrekt*, men som i en automatisk utvärdering skulle få relativt låg poäng. Benämningen *Källmening* är det svenska segment som ligger till grund för översättningen till engelska. *Kandidatmening* är resultatet av den maskinella översättningen av källmeningen och *referensmening* är den engelska översättningen gjord av en professionell översättare. Det är alltså den referensmeningen som jämförs med kandidatmeningen vid poängtilldelning i Neva och Bleu.

- (3) Källmening:  
*"Jag ser fram emot att samarbeta med er alla i min nya roll."*  
Referensmening:  
*"I am looking forward to working with all of you in my new role."*  
Kandidatmening:  
*"I look forward to a continued cooperation with you all in my new role."*  
Poäng vid automatisk utvärdering: Neva:0.372, Bleu: 0.317

I 3 ges ett exempel på en mening som har Neva- och Bleupoäng på 0.372 respektive 0.317. Referensmeningen har en progressiv form av verbet "look" som den automatiskt översatta kandidatmeningen saknar. Vidare har "att samarbeta" översatts olika i referens- och kandidatmeningen. Trots syntaktiska skillnader mellan kandidat- och referensmening svarar de båda felfritt mot källmeningen, något som de båda automatiska utvärderingsmetoderna inte har möjlighet att ta hänsyn till.

- (4) Källmening:  
*"Komponentöversyn innebär besiktning och översyn av reparerbara komponenter enligt standardspecifikation."*  
Referensmening:  
*"Component Overhaul provides the inspection and overhaul of repairable parts to a standard specification."*  
Kandidatmening:  
*"Component overhaul means examination and overhaul of repairable components according to standard specification."*  
Poäng vid automatisk utvärdering: Neva:0.379, Bleu: 0.274

Exempel 4 visar en mening som fått Neva- och Bleupoäng på 0.379 respektive 0.274. Konstruktionen "provides of" som används i referensmeningen för att översätta "innebär" återkommer inte i kandidatmeningen. Istället har "innebär" i kandidatmeningen översatts med "means". Trots att meningens innebörd bevarats och att kandidatmeningen har korrekt syntaktisk struktur, ger automatiska utvärderingsmetoder inte poäng som står i relation till kvaliteten på översättningen.

- (5) Källmening:  
*"Effektivisera uppläggning av nya konton"*  
Referensmening:  
*"Streamline Account Enrollment Processes"*  
Kandidatmening:  
*"Streamline arrangement of new accounts"*  
Poäng vid automatisk utvärdering: Neva:0.05, Bleu: 0

I 5 ses ett exempel som kan fungera som rubrik, alternativt imperativsats. Syntaktiskt utgörs källmeningen av ett verb tillsammans med en nominalfras bestående av huvudord och prepositionsfras. Kandidatmeningen däremot har samma semantiska innehåll, men utgörs av ett verb tillsammans med en nominalfras bestående av ett sammansatt substantiv. Den automatiska utvärderingen ger i detta segment väldigt låga Neva- och Bleupoäng, (0.05, respektive 0). Anledningen till det låga Bleuvärdet är att referensmeningen endast innehåller fyra ord, något som diskuterats i tidigare kapitel.

Dessa tre utdrag är bara några av de exempel som tydligt visar varför det är viktigt att komplettera automatisk utvärdering av maskinöversatt text med manuell. Vad gäller de olika utvärderingssätten kan sägas att den automatiska utvärderingen är tidssnål och konsekvent. Men ingen hänsyn tas till synonymi och alternativa grammatiska konstruktioner om som i fallet ovan bara en referensöversättning nyttjas. Ett manuellt evalueringstillvägagångssätt ger däremot en mer nyanserad bild. Problemet här är istället det som är fördelarna med den automatiska utvärderingen, nämligen tidsåtgången och subjektiviteten.

## 4 Metod

Metoderna som avses att användas för att framställa lexikonet är implementerade i programkomponenter som till stor del är inriktade på att utnyttja befintliga resurser i form av lexikondatabaser och programvara. Slutprodukten kommer att bestå av en relationsdatabas i MySQL-format, men innan den har nått den formen lagras informationen i separata textfiler. Den övergripande processen består av fem delar: Förbearbetning, sammanställning av det svenska lexikonet, skapande av översättningsrelationer, sammanställning av det engelska lexikonet och slutligen utvärdering.

### 4.1 Förbearbetning

Inledningsvis språkseparatoras korpusmaterialet i en svensk och en engelsk del, därefter används tvättningsscript skrivna i Perl för att rensa materialet från uppmärknings- och stiltaggar. Dessa script är specifikt anpassade för Bombardiermaterialet och tar bort information som inte ska användas i det fortsatta arbetet. Originalmaterialet sparas även det, eftersom det fortsatta arbetet grundar sig på dels tvåspråkiga, tvättade respektive otvättade korpusvarianter, dels på språkseparatorade, tvättade och otvättade varianter.

### 4.2 Den svenska delen av lexikonet

Först extraheras typorden ur den svenska delen av testdatan i översättningsminnet. För att få reda på vilka löpord som redan finns representerade i någon av de befintliga lexikaliska resurserna genererar man de fullständiga ordformerna ur lexikonerna med hjälp av databasens inbyggda morfologi. Om ett löpord i översättningsminnet matchas mot en ordform i någon av de tillgängliga lexikonerna extraheras all nödvändig information för ett lexikoninlägg ur databasen. Här gäller något olika attribut beroende på ordklass. Substantiv har i regel en kod som återspeglar ett lemmas semantiska särdrag. Verb har istället ett valensattribut som beskriver hur många och vilka typ av objekt verbet i fråga tar. Gemensamt för alla ordklasser är information om ordklass, mönsterord, lemma och teknisk stam.

De typord som på automatisk väg inte tilldelats någon lexikoninformation körs igenom UCP:s sammansättningsanalys. De ord som analyseras och får en eller flera analyser, dvs är potentiella substantivsammansättningar, filtreras genom ett script Åberg (2003) som väljer den bästa av flera möjliga analyser.

De typord som varken återfinns i någon av lexikondatabaserna eller analyserats som sammansättningar, och som därför inte automatiskt kunnat tilldelats någon lexikoninformation, måste manuellt förses med lexikoninformation. Till de löpord som

är substantiv används programmet PAT Starbäck och Tiedemann (1997) som med hjälp av användarinformation härleder deklination och därmed mönsterord.

Efter att ha lemmatiserat källspråkssidan av översättningsminnet återstår etablerandet av översättningsrelationer och målspråkslemmatisering.

### 4.3 Översättningsrelationer och den engelska delen

För att fastställa översättningsrelationer ordlänkar man översättningsminnets källspråkssegment och målspråkssegment med hjälp av Clue Aligner. Detta genererar ordlänkar på typordsnivå med frekvensinformation. Dessa använder man sedan som underlag för att bestämma översättningsrelationer målspråkslemman och källspråkslemman emellan. Genom att sedan automatiskt expandera ordformer ur lemmarna i den svenska sidan av lexikonet är det möjligt att matcha dessa mot typord i resultatet av länkningsresultatet. De olika länkningsfrekvenserna för de svenska typord som tillhör samma lemma sammanställs, sorteras och blir till en lista som utgör grunden för valet av den bästa länken.

När ett källspråkstypord i länkningsresultatet matchas mot en ordform tillhörande ett lemma i det svenska lexikonet, eftersöks det målspråkstypord som det är länkat till i de tillgängliga lexikon databaserna. I det fall målspråkssidan av länken består av fler än ett ord, och därmed med största sannolikhet är en sammansättning, eftersöks enbart det sista ordet.

Om målspråkstypordet hittas, läggs dess lemma till i en lista över länkkandidater på lemmanivå. Vidare inkrementeras en frekvensräknare med frekvensen för typordslänken. För de länkar där målspråksord inte hittas, måste översättningsrelation och målspråkslexikon ingång upprättas manuellt.

Alltså görs typordslänkarna producerade av Clue Aligner om till en lista med lemmalänkar där varje svenskt lemma knyts ihop med en lista på potentiella engelska lemmakandidater med tillhörande frekvens. Vidare har den större delen av den engelska sidan av lexikonet automatiskt tagits fram. Listan går igenom manuellt för att man ska kunna verifiera översättningsrelationerna och engelska lexikon ingångar samt för att fylla i information som inte fastställts på automatisk väg.

#### (6) Skiss över bearbetningsförlopp

1. Utgångspunkten är ett lemma i det svenska lexikonet, 'man l.nn+timme.nn'
2. Genom att generera alla fullformer skapas bla. fullformen 'mantimmar'
3. Ordformen 'mantimmar' hittas i länkfilen och är länkat till 'man hours'
4. Eftersom 'man hours' är en sammansättning, är huvudordet sist. Med information om det går det att skapa ett nytt lemma.
5. man\_hours.nn skapas genom att huvudet 'man' konkateras med det befintliga lemmat 'hour.nn'. All morfologisk information ifrån 'hour.nn' förs vidare till lemmat 'man\_hour.nn'
6. En ingång i översättningsrelationerna skapas: man l.nn+timme.nn - man\_hour.nn

### 4.4 Utvärdering

I skapandet av lexikonet bestäms att 90% av korpusmaterialet ska utgöra testdata, och de resterande 10% testdata. För att man ska kunna utvärdera resultaten av över-

sättningarna används MT evaluation Toolbox, ett javaprogram för att automatiskt utvärdera maskinöversatt text genom att mäta NEVA och BLEUpoäng. Dessa resultat kompletteras med manuell utvärdering, där kvaliteten för 300 individuella meningsegment kategoriseras i tre grupper med avseende på korrekthet och feltyp.

## 5 Utförande

### 5.1 Hjälpmedel för att ta fram lexikala resurser

Den baseline som arbetet avser att skapa består av tre lexikala komponenter: ett svenskt lexikon, ett engelskt lexikon och översättningsrelationer dem emellan. För att genomföra detta använder vi oss av en till stor del egenutvecklad mjukvara. Till sammans med resurser i form av MySQL-databaser och mjukvara under GNU public license utgör dessa samlade komponenter ett kraftfullt och mångsidigt paket för att göra lexikonutvinningar ur stora textmängder.

#### 5.1.1 Materialet och dess förutsättningar

Råmaterialet erhålls från Bombardier Transportation via vår kontaktperson Helena Wretman och utgörs av översättningsminnen i MS-word format. Den interna strukturen på materialet har mycket stor likhet med konventionella parallellkorpusar, dvs hela satser är strukturerade parvis i käll- och målspråk med intern metatagging och intern representation av specialtecken, se 5.1. Materialets ämnesområde kretsar till största delen kring affärsverksamhet inom tågtrafik såsom nyhetsbrev, säljrapporter och affärsplanering. Det finns även många inslag av teknisk texttyp som instruktioner och dokumentation till system. Även avtalstexter och överrenskommelser omfattas av materialet. Till synes är texttyperna i allra högsta grad varierade, materialet innehåller både meningssegment med enkel syntax och segment med mer komplicerad och utvecklad satsbyggnad.

#### 5.1.2 Tvättning

Korpusmaterialet innehåller i rå form stora mängder redundant data, i form av RTF-taggar blandat med uppmärkning och datumstämplar genererade av Trados Translators workbench. Innan materialet kan börja bearbetas måste det därför tvättas. En undersökning av materialet visar att informationsredundansen kan delas in i tre separata delar:

- (7) Metainformation om start och slut av segment samt om skapare och tidpunkt för skapandet, tex. <Seg>, <TrU> och <Crd>
- (8) Typografiska stilinstruktioner, t ex {ul}, {br}, {super} och {strong}.
- (9) Representation av specialtecken som t ex \quote, \endash

För att utvinna lexikala resurser ur träningsdata krävs naturligtvis att innehållet är skilt från uppmärkningen och det första steget är därför att tvätta materialet. Tvättningsprocessen består av en serie komponenter som är egenutvecklade.

```

<TrU>
<CrD>14042003, 14:09:56
<CrU>HHN
<ChD>15042003, 09:51:52
<ChU>HHN
<Seg L=EN-US>Expertise in the Rail Industry.
<Seg L=SV-SE>Tekniskt \backslash$endash kunnande i järnvägsindustrin
</TrU>

```

**Figur 5.1:** Exempel på TRADOS översättningsminne.

**Tabell 5.1:** Statistik över träningsmaterialet

| Otvättat material   |        | Tvättat material    |        |
|---------------------|--------|---------------------|--------|
| Svenska löpord      | 252393 | Svenska löpord      | 252393 |
| Engelska löpord     | 254257 | Engelska löpord     | 254257 |
| Uppmärkningsord     | 362975 | Totalt antal löpord | 506650 |
| Totalt antal löpord | 906580 | Svenska typord      | 24784  |
|                     |        | Engelska typord     | 14575  |

Eftersom materialet levererades i ett binärformat, konverteras det först till ett bearbetningsvänligt textformat. Materialet går därefter igenom en tvättprocess, som via Perl- och Pythonskript avtaggar och genomför teckensubstitution för den interna uppmärkningen och teckenrepresentationen. Materialet sparas i tvåspråkig, tvättad form, men det separeras även i två delar, en svensk och en engelsk, som ska användas i länkingsprocessen. Tio procent av materialet läggs åt sidan för att senare användas som testkorpus. De återstående 90 procenten är träningsdata som därefter tokeniseras för att man ska få fram alla unika ordformer.

## 5.2 Tokenisering

Tokenisering genomförs efter att så mycket redundant data som möjligt tagits bort ut materialet. Syftet är att isolera samtliga förekommande typord för att sedan med automatiska såväl som manuella metoder extrahera deras motsvarande lemma.

## 5.3 Lemmatisering

För att ta fram en uppsättning domänspecifika lexikoninlägg krävs användning av båda automatiska och manuella metoder. De automatiska utgör den största delen och kompletteras med manuella.

### 5.3.1 Automatiska metoder

De automatiska metoder som används för att lemmatisera är egenutvecklade med grund i databasens inbyggda morfologi. Metoderna syftar till att i högsta grad åter-

**Tabell 5.2:** Exempel på stavfel som förstör statistik

|                                |   |
|--------------------------------|---|
| bonusiiivå - 'bonusnivå'       | Borgarrd - 'Borgard'                              |
| bovenkmap - 'Bovenkamp'        | dataöverföring används - 'dataöverföring används' |
| därflr - 'därför'              | eftesom - 'eftersom'                              |
| fakturormått - 'fakturor mått' | finnsa - 'finnas'                                 |
| framfö - 'framför'             | funktionewr - 'funktioner'                        |

använda befintliga databasresurser och kan utifrån ordform i de allra flesta fall härleda lemma. Eftersom varje lexikoninlägg kräver information om åtminstone lemma, stam, mönsterord och ordklass är som resurserna dessa kan utvinnas ur klart begränsade. Det första steget är att gå igenom lexikaliska databasresurser knutna till andra MATS-domäner. På så sätt extraheras, när det är möjligt, nödvändig information ur MatsLex. Det andra steget är sedan att köra de ord som inte kunnat tilldelas någon information igenom UCP:s sammansättningsanalys. Det sista steget består av att manuellt lemmatisera kvarvarande ord.

Eftersom översättningsminnet är inmatat manuellt förekommer i materialet många stavfel. Detta är ett problem eftersom felstavade ord kommer att behandlas som fristående ordformer och inte räknas tillsammans med lemmat de avser att beteckna. Detta kan i vissa fall leda till att somliga ordformer inte når upp i det tröskelvärde på tre förekomster som satts som lägsta gräns.

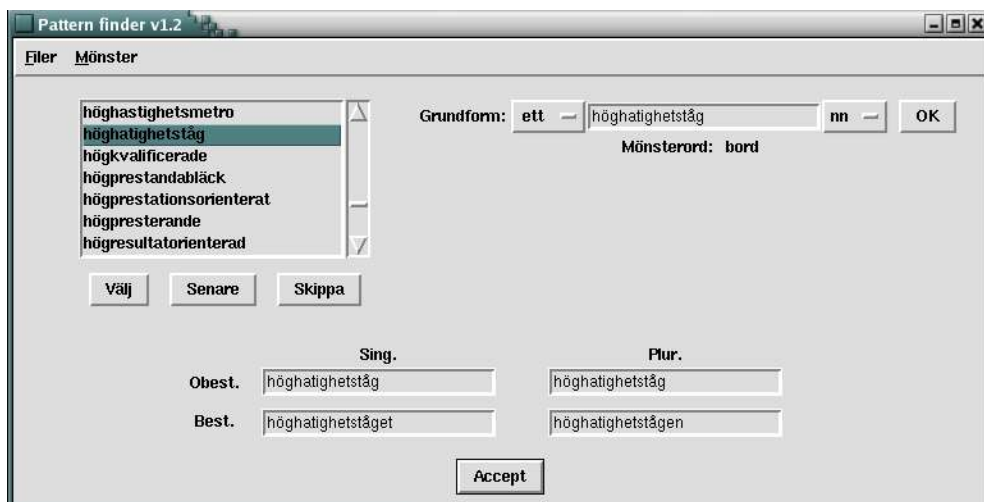
## Databasuppslag

Efter tvättningen och upprensningen av översättningsminnena genereras en lista av unika löpord, dvs. typord. Dessa uppgår till 25921. Utifrån stammarna i de tillgängliga, befintliga databaslexikonerna 'LingLex', 'ScaniaLex', 'KursLex' och 'JordbruksLex', genererades fullformer ur ingående lemmat som sedan användes för att matcha typorden i översättningsminnet. Detta görs i en flerstegsprocess på automatisk väg med programvara skriven i programmeringsspråket python. För att genom python kommunicera med databaserna används API-modulen MySQLdb. När ett typord matchar en fullform från någon av lexikondatabaserna extraheras den information som krävs för ett databasinlägg, dvs lemma, stam, mönsterord, ordklass och annan ordklassspecifik data.

Eftersom ett lemma på den svenska sidan av lexikonet kan innehålla flera tekniska stammar (t ex: bonde, bönder) och därmed utgör två lexikoninlägg, extraheras även samtliga tekniska stammar knutna till ett typord, även om inget löpord med just den stammen förkommer i översättningsminnet. Tillvägagångssätt täcker in 15079 av typorden. Dessutom hittas 626 extra stammar.

## Sammansättningsanalys

De resterande 10544 ordtyperna körs igenom UCP:s sammansättningsanalys. Ut-datan från sammansättningsanalysen går sedan vidare till ett filtreringsprogram utvecklat på Institutionen för Lingvistik och Filologi vid Uppsala Universitet. Totalt väljs 5120 ord ut som troliga sammansättningar. Eftersom UCP:s sammansättningsanalys utgår ifrån LingLex finns också information nödvändig för ett lexikoninlägg i scarie-databasen. För detta ändamål skrivs ett python-skript som extraherar nödvändig lexikalisk information ur scarie-databasen givet en sammansättningsfinala



Figur 5.2: PATtern Finder

led. Detta eftersom en svensk sammansättning i regel böjs efter sitt finala led, t ex *slut1.nn+montering.nn+aktivitet.nn*. Denna information rensades därefter manuellt och korrekta sammansättningar lades till lexikonet.

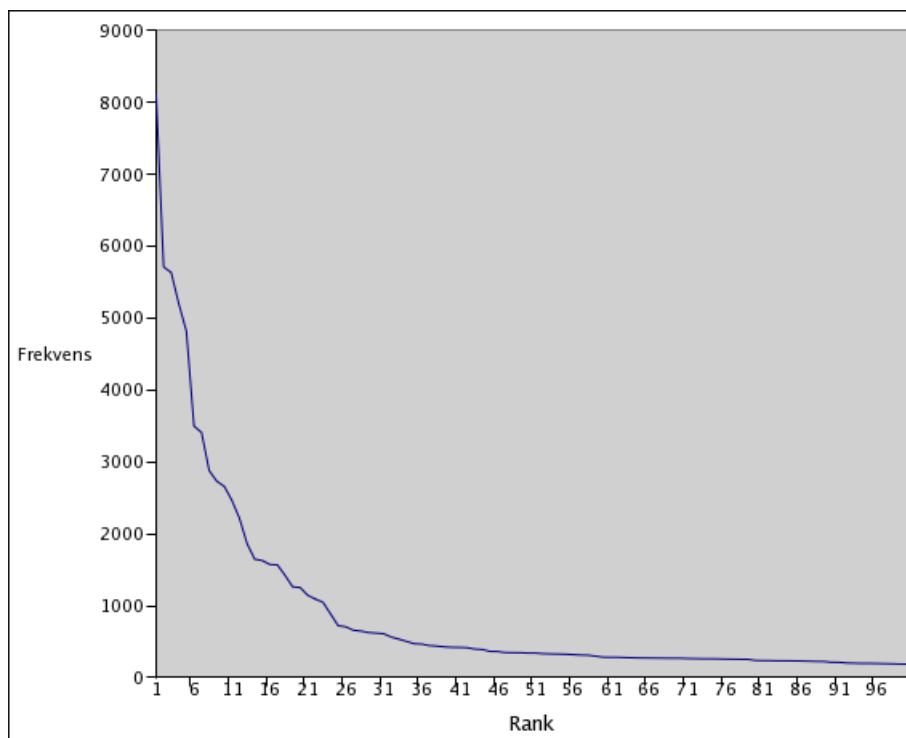
### 5.3.2 Manuella metoder

De ordformer som inte kan slås upp i andra MATS-lexikon kräver manuell lemmatisering. Efter att ha lemmatiserat automatiskt kvarstår ca 5000 ord för manuell lemmatisering. Eftersom lexikonet grundar sig på mönsterord för generering av fullformer består lemmatiseringen till största del av att identifiera varje lemmas ordklass, mönsterord samt tekniska stam. Det är även nödvändigt att se till att det finns mönsterord i databasen som täcker alla ords morfologi. För att lemmatisera substantiven används programmet PAT, (PATtern Finder), skrivet av Per Starbäck vid Institutionen för Lingvistik och Filogi vid Uppsala Universitet. Detta program försöker med hjälp av information från användaren sluta sig till aktuellt mönsterord givet genus och en eller flera former av det svenska substantiv som användaren anger, se 5.2.

## 5.4 Ett svenskt lexikon som utgångspunkt

Efter att ha lemmatiserat den svenska sidan av lexikonet har ca 14000 svenska lemmor extraherats med automatiska och manuella metoder. För att minimera det manuella arbetet med att skapa engelska lexikoningångar samt översättningsrelationer upprättas ett slags tröskelvärde.

Zipfs lag säger att det finns ett samband mellan ett ords frekvens och dess ranking i en korpus. Det finns en konstant,  $k$ , för varje typord i en korpus sådan att  $k = f * r$  där  $f$  är typordets frekvens och  $r$  är typordets ranking i korpusen. Tilläggas bör att Zipfs lag inte bör betraktas som en lag utan snarare som en relativt träffsäker generalisering baserad på empiriska fakta (Manning och Schütze, 1999). Det finns en liten uppsättning av mycket frekventa ord, en mellanstor uppsättning av mellan-



**Figur 5.3:** Zipfs lag applicerad på träningsmaterialet

frekventa ord och en stor uppsättning av lågfrekventa ord. Texten utgör typord med frekvens 1 c:a 50% av typorden i den aktuella korpussen.

Sett ur detta perspektiv nås till slut en gräns där arbetet med att utvinna och lägga till lexikoningångar samt översättningsrelationer inte motsvarar den resultatmässiga vinsten. Gränsen sätts till att varje lemma måste vara representerat åtminstone 3 gånger i träningsmaterialet. Förhoppningen är även att skapa ett bättre underlag till den automatiska genereringen av översättningsrelationer, eftersom högre frekvens innebär tillförlitligare ordlänkar. Detta resulterar i ett lexikon om ca 8000 svenska lemmor.

## 5.5 Länkning

Den tvättade och språkseparatorade träningsdatan ska nu länkas, och till det används Clue Aligner med default-värden. Inledningsvis konverteras de båda filerna till tokeniserad XML. Därefter görs en meningslänkning, där meningssegment paras ihop och representeras enligt 5.4. Det meningslänkade materialet går sedan vidare till ordlänkning där varje enskilt svenskt ord kopplas till motsvarande engelskt, se 5.5

### 5.5.1 Översättningsrelationer och den engelska sidan

Det engelska lexikonet skapas med utgångspunkt ur det svenska. Genom att generera typord av de lemmor som ingår i den svenska sidan av lexikonet lokaliserar man deras representation i parallellkorpussen. Detta innebär att genereringen av det engelska lexikonet skedde i samma steg som skapandet av översättningsrelationer. När en

**Tabell 5.3:** De 20 mest frekventa löporden i träningsmaterialet

| Typord     | Frekvens | Rank | <i>k</i> |
|------------|----------|------|----------|
| och        | 8121     | 1    | 8121     |
| i          | 5711     | 2    | 11422    |
| att        | 5632     | 3    | 16896    |
| för        | 5197     | 4    | 20788    |
| av         | 4817     | 5    | 24085    |
| som        | 3498     | 6    | 20988    |
| på         | 3408     | 7    | 23856    |
| till       | 2879     | 8    | 23032    |
| är         | 2731     | 9    | 24579    |
| med        | 2658     | 10   | 26580    |
| en         | 2464     | 11   | 27104    |
| det        | 2214     | 12   | 26568    |
| den        | 1863     | 13   | 24219    |
| de         | 1647     | 14   | 23058    |
| ett        | 1628     | 15   | 24420    |
| har        | 1573     | 16   | 25168    |
| bombardier | 1567     | 17   | 26639    |
| om         | 1422     | 18   | 25596    |
| eller      | 1263     | 19   | 23997    |
| vi         | 1252     | 20   | 25040    |

**Tabell 5.4:** Tabell över ordklassdistribution i BombLex.

| Frekvens | Ordklass |
|----------|----------|
| 4939     | NOUN     |
| 2454     | VERB     |
| 1005     | ADV      |
| 978      | ADJ      |
| 559      | PNOUN    |
| 217      | PREP     |
| 75       | PRON     |
| 74       | CONJ     |
| 35       | NUM      |
| 18       | SEP      |
| 9        | ART      |
| 3        | INT      |
| 1        | MW       |
| 1        | IE       |

```
<?xml version="1.0" encoding="utf-8"?>
<!DOCTYPE cesAlign PUBLIC "-//CES//DTD XML cesAlign//EN" "">

<cesAlign toDoc="1988en.xml" version="1.0" fromDoc="1988sv.xml">
<linkGrp targType="s" toDoc="1988en.xml" fromDoc="1988sv.xml">
<link certainty="1295" xtargets="s1.1;s1.1" id="SL0.1" />
<link certainty="3" xtargets="s2.1;s2.1" id="SL1.1" />
<link certainty="6" xtargets="s3.1;s3.1" id="SL2.1" />
<link certainty="3" xtargets="s3.2;s3.2" id="SL2.2" />
<link certainty="34" xtargets="s3.3;s3.3" id="SL2.3" />
<link certainty="7" xtargets="s3.4;s3.4" id="SL2.4" />
<link certainty="-1390" xtargets="s4.1;s4.1 s4.2" id="SL3.1" />

```

**Figur 5.4:** Meningslänk

**Figur 5.5:** Länkning på ordnivå

```
<?xml version="1.0" encoding="utf-8"?>
<!DOCTYPE cesAlign PUBLIC "-//CES//DTD XML cesAlign//EN" "">

<cesAlign version="1.0"> <linkGrp targType="s" toDoc="1988en.xml" fromDoc="1988sv.xml">
<link certainty="1295" xtargets="s1.1;s1.1" id="SL0.1">
<wordLink certainty="0.0163524885533091" lexPair="REGERINGSFÅ<96>RKLARING;Statement of Government
xtargets="w1.1.1;w1.1.1+w1.1.2+w1.1.3+w1.1.4" />
</link>
<link certainty="3" xtargets="s2.1;s2.1" id="SL1.1">
<wordLink certainty="0.072701606262909" lexPair="av;of the" xtargets="w2.1.12;w2.1.12+w2.1.13" />
<wordLink certainty="0.250323740369646" lexPair="Eders;Your" xtargets="w2.1.4;w2.1.4" />
<wordLink certainty="0.210867095125943" lexPair="herr;Mr" xtargets="w2.1.8;w2.1.8" />
<wordLink certainty="0.065697227648027" lexPair="Sveriges;Swedish" xtargets="w2.1.13;w2.1.14" />
<wordLink certainty="0.210867095125943" lexPair="talman;Speaker" xtargets="w2.1.9;w2.1.9" />
<wordLink certainty="0.238486746796535" lexPair="Majestät;ter;Majesties" xtargets="w2.1.2;w2.1.2" />
<wordLink certainty="0.210867095125943" lexPair="Kungliga;Royal" xtargets="w2.1.5;w2.1.5" />
...

```

```
städning.nn      Most likely first:
['cleaning.nn', 2]
['clean.vb', 2]
```

**Figur 5.6:** potentiella länkkandidater

```
man1.nn+timme.nn['1', 'mantimmar', 'EN:', 'man', 'hours'] hour.nn ['NOUN', 'DOG', 'hour']
english lexicon entry: man_hour.nn      NOUN      DOG      man hour
translation relation: man1.nn+timme.nn      1      man_hour.nn      1
```

**Figur 5.7:** Samlad lexikon- och översättningsinformation

engelsk översättningsekvivalent fastställts med hjälp av länkningsresultatet söks de engelska sidorna av databaserna igenom efter möjlig information. Om den engelska översättningsekvivalenten innehåller fler än ett ord eftersöks endast det finala ledet, eftersom det är det enda som påverkas av böjning. Om ett svenskt lemma är automatiskt länkat till flera engelska lemman genereras en lista sorterad efter frekvens. Om flera engelska lemman har samma frekvens hamnar det med samma ordklass som det svenska lemmat överst (figur 5.6). När all information extraherats ur databaserna skapas en fil som innehåller lexikoninformation och potentiell översättningsrelation, se figur 5.7. Denna information är grunden till den engelska sidan av lexikonet och översättningsrelationerna, som därefter bearbetas manuellt och rensas upp. Tack vare att informationen strukturerats enhetligt och märkts upp på varje rad underlättas detta efterarbete. Svenska lemman som inte automatiskt fått någon engelsk länkkandidat, korrigeras genom att skapa den engelska länken manuellt och samtidigt utöka det engelska lexikonet.

Ett alternativt tillvägagångssätt hade varit att för varje svenskt lemma slå upp de redan befintliga översättningsrelationerna. Detta hade gett upphov till ett komplett engelskt lexikon, som dessutom hade kunnat genereras på nolltid. Men då hade kärnan i lexikonet, själva översättningsrelationerna inte blivit särskilt tillförlitliga, eftersom det är här den domänspecifika kopplingen görs. Eventuella översättningar hade troligtvis haft semantiska felaktigheter, även om de varit syntaktiskt riktiga. Å andra sidan är detta ett lämpligt tillvägagångssätt för de mera domänneutrala, finita ordklasserna, som t ex prepositioner eller pronomen, som har samma översättningslänkar obereonde av domän.

## 5.5.2 Dataintegritet

Som sagt har stor del av lexikonsammanställningen skett på automatisk väg genom att man slår upp fullformer i befintliga databaser och därigenom extraherar nödvändig information om parameterar som böjningsmönster, teknisk stam, semantiska data etc. Detta har sparat mycket tid eftersom det reducerat mängden data som behövs lemmatiseras manuellt. Det har dock visat sig finnas viss diskrepans lexikonerna emellan. Gemensamma ingångar uppvisar inte identiska metadata, och det är både svårt och tidsödande att säga vilket lexikon som har den mest rättvisande informationen.

**Tabell 5.5:** Översikt över lexikonstorlek

|                         | Basic | Bomblex | Bomblex Extended |
|-------------------------|-------|---------|------------------|
| Svenska lemman          | 7127  | 8752    | 13157            |
| Engelska lemman         | 5813  | 7127    | 11293            |
| Översättningsrelationer | 7127  | 8752    | 13157            |

Efter att manuellt ha undersökt och uppskattat databaserna och deras innehåll väljs att för varje lemma genomgå lexikonerna i ordningen *KursLex*, *JordbruksLex*, *ScaniaLex*, *LingLex*.

## 5.6 Sammanslagning av lexika

Det domänanpassade lexikonet är nu extraherat ur träningsmaterialet. Två uppsättningar lemman, en svenskspråkig och en engelskspråkig del, har tagits fram. Dessa svarar mot de svenska lemman som förekommer tre eller fler gånger i någon ordform i träningsmaterialet. Med både automatiska och manuella metoder har morfologiska och semantiska data för alla ingående lemman tagits fram. Med hjälp av till största delen automatisk länkning, men även manuell har domänspecifika relationer orden emellan etablerats. Detta lexikon innehåller nu 7432 ingångar. Därefter utökas lexikonet med en kärnvokabulär om 2158 lexikonringångar (Moberg, 2005) Innehållet i denna resurs överlappar till stor del det egna materialet. Det sammanslagna lexikonet bestående av våra egna och kärnlexikonets resurser kallas i fortsättningen *Bomblex*. Lexikonet består nu av 8752 ingångar. Kärnlexikonet består av domänneutrala funktions- och innehållsord och utgör en baskomponent i MATS som har som syfte att täcka in den mest grundläggande vokabulären.

### 5.6.1 Extraktion av allmänlexikon

För att ytterligare utöka omfattningen av lexikonet slås det sedan ihop med ett allmänlexikon om 7127 ord och resultatet därav kallas *Bomblex Extended*. Detta allmänlexikon är en del av *KursLex* och är extraherat ur ett flertal andra databaser (*ScaniaLex*, *LingLex*, och *JordbruksLex*), där man filtererat bort domänspecifika enheter och på så sätt samlat ord som är av allmänspråklig karaktär. När dessa slagits samman är storleken på lexikonet 13157 svenska lemman, 11293 engelska lemman och 13227 översättningsrelationer.

Även då lexikonet slås samman med kärn- och allmänlexikonet uppstår problem med dataintegriteten. Eftersom både kärn- och allmänlexikonet till viss grad överlappar vårt eget lexikon finns också här gemensamma ingångar som skiljer sig åt med avseende på metadata. Konsekvensen av detta vid en översättning är att lexikonuppslagningen vid analys får två kandidater med olika metadata. I vissa fall kan detta leda till felaktig översättning, som manuellt får redigeras efter utvärderingen. Detta berör framförallt de fall där felaktiga semantiska eller morfosyntaktiska data angivits, och där det kan bli översättningsfel med avseende på numerus, artikel o dyl. För att kringgå dessa potentiella problem görs en sammanställning av samförekomst mellan lexikonerna, där ingångarna i *Bomblex* ges företräde.

## 6 Resultat

### 6.1 Automatisk utvärdering

När lexikonresurserna lästs in och kompilerats körs den 35220 ord stora testdatan igenom MATS för översättning. För att få bättre översikt över resultaten delades materialet in i 18 smådelar om 150 segment var. Alla segment körs sedan i sekvens och respektive resultat utvärderas med hjälp av MT Evaluation Toolbox. Körningarna separeras även lexikonmässigt där tre fristående körningar görs med de tre olika lexikonerna i ordningen:

1. Endast BombLex
2. Endast Allmänlexikon
3. BombLex Extended (BombLex och Allmänlexikon)

Resultaten som följer 6.1 till 6.3 visar Neva- och Bleupoäng för varje segment med olika lexikon. Det är tydligt att det i alla tre körningarna finns en viss poängmässig spridning segmenten emellan. Detta beror på att testdatan är mycket blandad med avseende på texttyp och ämnesområde, och detta resulterar i att översättningskvaliteten varierar kraftigt inom segmenten. Som visas i 6.4 ger en översättning av testdatan med allmänlexikonet en Nevapoäng på 0.203 och en Bleupoäng på 0.127. En körning med BombLex ger en Nevapoäng på 0.234 och en Bleupoäng på 0.190. Sammanslagningen av de båda ger det bästa resultatet, en Nevapoäng på 0.255 och en Bleupoäng på 0.190.

En maskinöversättning med lexikaliska resurser baserade på domänanpassning av Bombardier-materialet, som levererats av Explicon, förbättrar Nevapoängen med ca 23%, eller ca 5 procentenheter i jämförelse med körning med endast allmänlexikaliska resurser. Motsvarande förbättring för Bleupoängen är ca 49% eller ca 6 procentenheter.

Resultaten beror givetvis på flera saker. Korpusen, som lexikonet är uppbyggt av är inte någon sluten domän. Den innehåller, som tidigare nämnts, en mängd ämnesområden, som avtalstexter, tekniska instruktioner och säljrapporter inom tåg- och transportområdet. Lexikonet bör därför betraktas som en vokabulär som beskriver ett mellanting mellan allmänt språk och subspråk.

Vidare har det i evalueringstadiet endast funnits en referensöversättning att tillgå. För att ett maskinöversatt segment ska få poängen 1 måste det i detta fall vara identiskt med referensöversättningen. I praktiken ger detta upphov till en klart snäv automatisk poängbedömning, helt utan alternativa översättningssätt. För att ge ett pålitligare resultat bör den automatiska utvärderingen tillföras ett flertal alternativa referensöversättningar.

**Tabell 6.1:** Maskinöversättning i Mats med Allmänlexikon. Neva- och Bleu-poäng visas för 28 segment om 150 meningar var.

| Segment  | NEVA  | BLEU  |
|----------|-------|-------|
| 150.xml  | 0.147 | 0.075 |
| 300.xml  | 0.205 | 0.142 |
| 450.xml  | 0.176 | 0.080 |
| 600.xml  | 0.163 | 0.082 |
| 750.xml  | 0.200 | 0.130 |
| 900.xml  | 0.221 | 0.148 |
| 1050.xml | 0.233 | 0.166 |
| 1200.xml | 0.223 | 0.144 |
| 1350.xml | 0.238 | 0.150 |
| 1500.xml | 0.218 | 0.140 |
| 1650.xml | 0.224 | 0.138 |
| 1800.xml | 0.211 | 0.141 |
| 1950.xml | 0.200 | 0.134 |
| 2100.xml | 0.239 | 0.148 |
| 2250.xml | 0.201 | 0.129 |
| 2400.xml | 0.172 | 0.093 |
| 2550.xml | 0.196 | 0.121 |
| 2700.xml | 0.197 | 0.127 |

**Tabell 6.2:** Maskinöversättning i Mats med BombLex. Neva- och Bleu-poäng visas för 28 segment om 150 meningar var.

| Segment  | NEVA  | BLEU  |
|----------|-------|-------|
| 150.xml  | 0.173 | 0.106 |
| 300.xml  | 0.233 | 0.178 |
| 450.xml  | 0.206 | 0.143 |
| 600.xml  | 0.199 | 0.129 |
| 750.xml  | 0.236 | 0.173 |
| 900.xml  | 0.264 | 0.208 |
| 1050.xml | 0.228 | 0.167 |
| 1200.xml | 0.218 | 0.144 |
| 1350.xml | 0.266 | 0.183 |
| 1500.xml | 0.261 | 0.191 |
| 1650.xml | 0.277 | 0.199 |
| 1800.xml | 0.276 | 0.227 |
| 1950.xml | 0.243 | 0.188 |
| 2100.xml | 0.272 | 0.176 |
| 2250.xml | 0.221 | 0.158 |
| 2400.xml | 0.196 | 0.132 |
| 2550.xml | 0.228 | 0.170 |
| 2700.xml | 0.218 | 0.163 |

**Tabell 6.3:** Maskinöversättning i Mats med BombLex Extended. Neva- och Bleu-poäng visas för 28 segment om 150 meningar var.

| Segment  | NEVA  | BLEU  |
|----------|-------|-------|
| 150.xml  | 0.184 | 0.115 |
| 300.xml  | 0.242 | 0.189 |
| 450.xml  | 0.217 | 0.156 |
| 600.xml  | 0.217 | 0.149 |
| 750.xml  | 0.260 | 0.197 |
| 900.xml  | 0.288 | 0.232 |
| 1050.xml | 0.277 | 0.224 |
| 1200.xml | 0.246 | 0.184 |
| 1350.xml | 0.282 | 0.198 |
| 1500.xml | 0.281 | 0.210 |
| 1650.xml | 0.292 | 0.211 |
| 1800.xml | 0.286 | 0.238 |
| 1950.xml | 0.261 | 0.205 |
| 2100.xml | 0.291 | 0.192 |
| 2250.xml | 0.247 | 0.187 |
| 2400.xml | 0.227 | 0.166 |
| 2550.xml | 0.246 | 0.189 |
| 2700.xml | 0.244 | 0.189 |

**Tabell 6.4:** Snittvärden för maskinöversättning med olika lexikon, poäng i Neva och Bleu

|                  | Neva  | Bleu  |
|------------------|-------|-------|
| Allmänlexikon    | 0.203 | 0.127 |
| BombLex          | 0.234 | 0.168 |
| BombLex Extended | 0.255 | 0.190 |

**Tabell 6.5:** Övergripande resultat av manuell utvärdering

| Korrekt  | Förståelig | Svårförståelig |
|----------|------------|----------------|
| 99 (33%) | 113 (38%)  | 88(29%)        |

## 6.2 Manuell utvärdering

Den manuella utvärderingen genomförs på 300 slumpvis valda delar av testdatan som maskinöversatts med BombLex Extended. Utvärderingen delas enligt vår subjektiva uppfattning in i tre graderingar med avseende på översättningskvalitet: *korrekt, förståelig och svårförståelig*.

- *Korrekt* innefattar meningssegment som har fått korrekt översättning. Denna typ av meningar kan skilja något sig ifrån referensmeningen lexikalt sett, men ändå ha samma semantiska innebörd. Det är här ingen tvekan om vad meningens betyder.
- *Förståelig* innefattar meningssegment som har fått en förståelig översättning. I dessa fall kan det saknas översättning av enstaka antal lexikala enheter. Det går att förstå meningen utan större ansträngning.
- *Svårförståelig* innefattar meningssegment som har fått en svårförståelig översättning. Denna typ av meningar kan dels vara fragmentariskt parsade, och därigenom översatta ord för ord. Transferregler för verb- genitiv eller prepositions konstruktion kan ha förstört syntaxen i meningen. I dessa fall kan man behöva slå i lexikon eller undersöka transferregler för att reda ut innebörden av meningen.

Som visas i tabell 6.5 utgörs 34% av det översatta testmaterialet av översättningar som fått etiketten "korrekt". 37% utgörs av meningar som tilldelats "förståelig" och 29 % består av meningar som klassats som svårförståeliga. Den manuella utvärderingen visade att meningar som fått etiketterna "förståelig" eller "svårförståelig" uppvisade typer av fel som var återkommande i hela utvärderingsmaterialet. Av den anledningen gavs också varje "förståelig" och "svårförståelig" mening en grov kategorisering som talar vad inom vilken kategori felet ligger. Kategorierna för att beskriva detta är "Lexikoningång(ar) saknas", "Felaktigt ordval" och "Felaktig grammatisk konstruktion". Det kan vara relativt svårt att analysera och märka upp feltyper, men det tre grupperna avser att beteckna fel enligt nedan:

- *Lexikoningång(ar) saknas* innebär att den översatta meningen innehåller ord som inte finns med i lexikonet. Resultatet blir att ordet inte översätts alls.
- *Felaktigt ordval* betyder att ett visst ord har valts framför ett annat vid översättningen, och att detta har lett till översättningsfel. Denna typ av fel är relaterad

**Tabell 6.6:** Feltyper i testdatan

|     |                                  |
|-----|----------------------------------|
| 109 | Felaktig grammatisk konstruktion |
| 49  | Felaktigt ordval                 |
| 122 | Lexikoningång(ar) saknas         |

till homografi, t.ex *inför* (presens av verbet 'införa' kontra prepositionen 'inför') Felaktigt ordval är ofta förekommande i meningar som inte fått fullständig analys, och uppkommer ofta när systemet inte kunna analysera källspråksmeningen och faller tillbaka på ord-för-ordgenerering. Ett annat fel i denna grupp är när det finns engelska ord i källspråket. Den engelska preposition 'for' analyseras som preteritumformen av det svenska verbet 'fara'

- *Felaktig grammatisk konstruktion* beskriver feltyper som kan bero på ett flertal faktorer, med den gemensamma nämnaren att de leder till grammatiska fel i målspråksmeningen. Detta manifesterar sig t ex i numerus- och personkongruensfel hos substantiv respektive verb, felaktig distinktion mellan adjektiv och adverb samt felaktig negationsplacering. En annan vanlig feltyp är felaktig ordföljd, som kan bero på att det saknas transferregler för en viss konstruktion och att systemets försök i att generera en översättning leder till syntaktiska fel. Den sistnämnda feltypen innefattar många idiomatiska uttryck.

Anledningen till att det finns osäkerhet kring orsaken bakom en felöversatt mening är att flera typer av problem och fel sammanflätade kan vara sammanflätade. Dessa kan vara både svåra och tidsödande att diagnosticera. En gruppering och sammanfattning av feltyperna visas tillsammans med antal förekomster i tabell 6.6. Som synes utgörs saknade lexikoningångar den allra största delen av feltyper, tätt följd av grammatiska felkonstruktioner. De fall där fel ord har valts i genereringen är en endast hälften så många som övriga två feltyper.

## 7 Sammanfattning

Denna uppsats beskriver hur ett maskinöversättningslexikon för tåg- och transportområdet tagits fram genom parallellkorpusbearbetning. Befintlig och egenutvecklad mjukvara i kombination med lexikonresurser har minimerat det manuella arbetet och därmed påskyndat lexikonframställandet och gjort det mindre felbenäget. Som parallellkorpus har ett professionellt sammanställt översättningminne använts.

Inledningsvis har den svenska sidan av lexikonet skapats. Denna har sedan använts som utgångspunkt för skapandet av den engelska sidan samt översättningsrelationer dem emellan. Vidare har lexikonet utökats med ett kärnlexikon om c:a 2000 lemmar samt ett allmänlexikon om c:a 6000 lemmar.

Lexikonet har utvärderats både automatiskt och manuellt. Den automatiska utvärderingen resulterade i en genomsnittlig Nevapoäng på 0,255 och en Bleupoäng på 0.190. Som jämförelse har även allmänlexikonet utvärderats på testdatan, Detta visar en genomsnittlig Neva- och Bleupoäng på 0.203 respektive 0.127.

Den manuella utvärderingen ger en något mer nyanserad bild och påvisar även brister med den automatiska utvärderingen. I den manuella utvärderingen klassificerades 33% av testdatan som korrekt, 38% som förståelig och 29% som svårförståelig.

## 8 Framtida utveckling

Under arbetet med detta lexikon har rutiner och processer ständigt förbättrats och kontinuerligt strömlinjeformats. Metoder som varit effektiva och rationella under vissa förhållanden har i nya miljöer och tillämpningar visat sig vara otillräckliga och därför behövs omformas och utvidgas för att tillfredsställa nya behov. Detsamma gäller den slutgiltiga produkten, själva lexikonet. Efter att ha genomgått en serie av tester och finjusteringar har lexikonet nått en baselinenivå, i vilken funktionaliteten och graden av korrekthet är tillfredställande. Det finns dock vidare utvecklings- och förbättringsmöjligheter både av processer och metoder för att framställa lexikonet och av själva lexikonet som produkt.

Vad gäller utveckling av metoder finns först och främst viktiga förbearbetningssteg som skulle förbättra resultatet. Översättningsminnet visade sig vara blandat, både med avseende på texttyp och på ämnesområde. Sett ur detta perspektiv borde översättningsminnen av den här typen egentligen betraktas som flera mindre domäner, snarare än en stor. I praktiken hade med stor säkerhet någon form av klustring av översättningsminnet varit lönsamt. På så vis skulle man kunna tänka sig att bygga upp flera mindre lexikon och därmed utföra översättningar på ännu mer avgränsade subdomäner. Det hade även gett möjligheten att sortera ut de segment av texttyp som inte alls lämpar sig att maskinöversätta.

Som den manuella utvärdering visar, beror ungefär 40% av felen på saknade lexikon-ingångar. Ett spår inom vidareutveckling av lexikonet kunde därför vara att utöka storleken genom att lägga till fler ingångar. Ytterligare en möjlighet till förbättring är att i korpusen ta till vara domänspecifika kollokationer, dvs ofta förekommande ordkombinationer, som inte låter sig översättas med grammatiska eller semantiska regler. Uttryck som "I alla fall" sparas med fördel som en lexikal enhet i lexikonet istället för att genereras utifrån grammatiska regler.

Gemensamt för alla förbättringar är att de förutsätter en djupgående analys av omfattande målspråksmaterial där alla typer av problem klassificeras med en uttömmande feltypologi.

# Litteraturförteckning

Nationalencyklopedin, 2005.

Ahrenberg, Lars, Merkel, Magnus, Sågvall Hein, Anna, och Tiedemann, Jörg. *Evaluation of Word Alignment Systems*. 2000. In *Proceedings of LREC 2000, Athens/Greece*. .

Forsbom, Eva. *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, held in conjunction with MT SUMMIT IX, pp. 29-36. New Orleans, Louisiana, USA, September 27., 2003*. URL <http://stp.lingfil.uu.se/~evafo/Papers/mtsummitixeval.pdf>.

Fürst, Kalle och Aaeng, Eivind. *Brødrene Dal og professor Drøvels hemmelighet*. 1979.

Gustavii, Ebba och Pettersson, Eva. *Utveckling av ett svensk-engelskt lexikon för maskinöversättning inom jordbruksdomänen*. Inst. för lingvistik och filologi, 2003. <http://stp.lingfil.uu.se/evapet/agri.pdf>.

Hutchins, W. J. och Somers, H. L., redaktörer. *An Introduction to Machine Translation*. Academic Press, London, 1992.

Manning, Christopher D. och Schütze, Hinrich. *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, Massachusetts, 1999. URL [citeseer.ist.psu.edu/635422.html](http://citeseer.ist.psu.edu/635422.html).

Moberg, Jens. Språkliga basresurser i maskinöversättningssystemet mats. Examensarbete, 2005.

Papineni, K., Roukos, S., Ward, T., och Zhu, W. Bleu: a method for automatic evaluation of machine translation, 2001. URL <http://www1.cs.columbia.edu/nlp/sgd/bleu.pdf>.

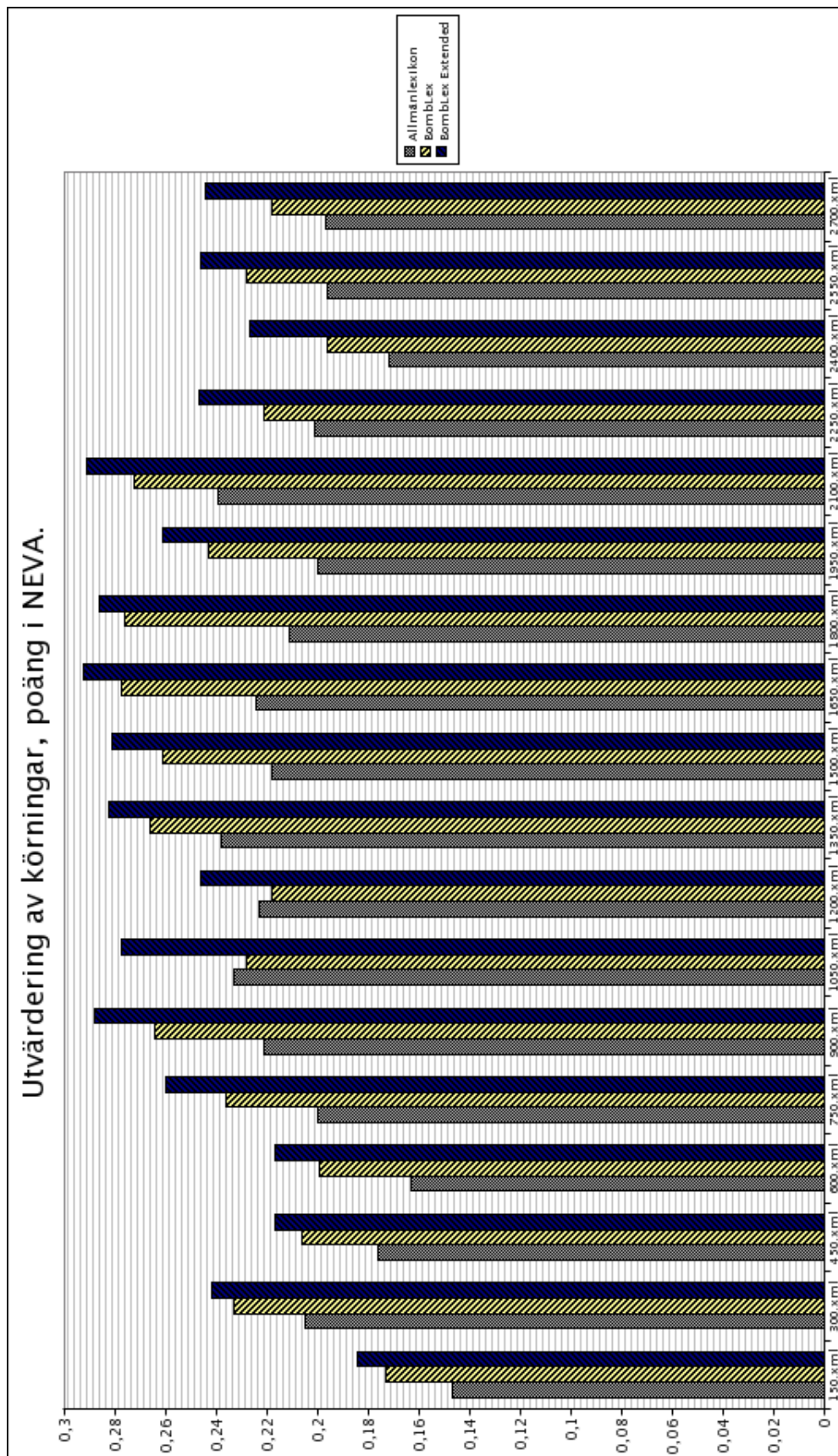
Pettersson, Eva. *Pilotstudie om maskinöversättning inom ramen för projekt kursdatabas*. Inst. för lingvistik och filologi, 2005. <http://stp.lingfil.uu.se/evapet/kurspilot.pdf>.

Somers, Harold. *Computers and Translation, a Handbook (pre-final draft)*. Unknown, 2003.

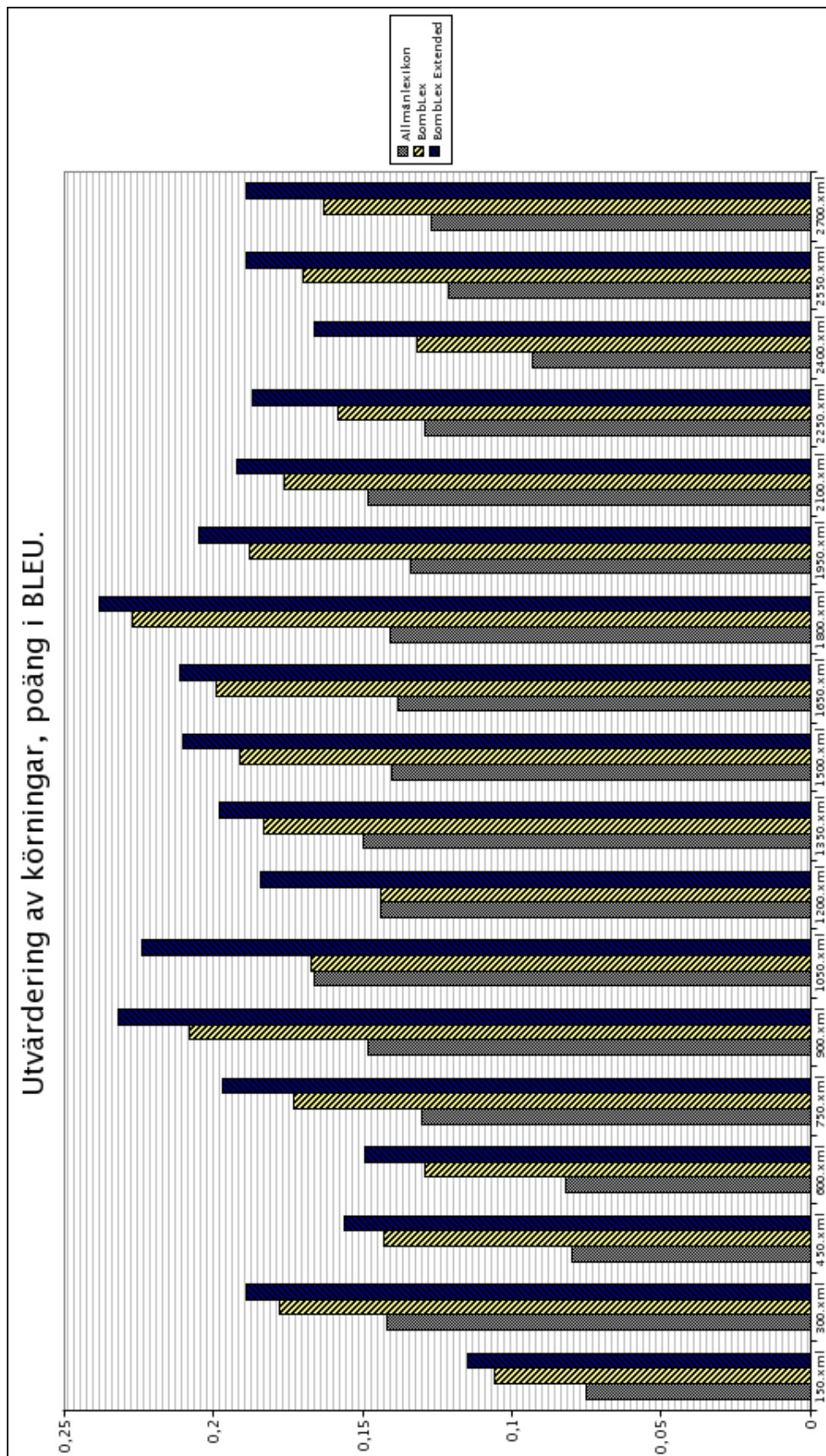
Starbäck, Per och Tiedemann, Jörg. *Användarhandledning till ScanCheck*. Uppsala universitet. Institutionen för lingvistik, 1997.

- Sågvall Hein, Anna, Forsbom, Eva, Weijnitz, Per, Gustavii, Ebba, och Tiedemann, Jörg. Mats - a glass box machine translation system, 2003.
- Tiedemann, Jörg. *MatsLex - a Multilingual Lexical Database for Machine Translation*. Department of Linguistics, Uppsala University, 2002. <http://www.let.rug.nl/tiedeman/blog/paper/coling04.pdf>.
- Tiedemann, Jörg. *Word to Word Alignment strategies*. Department of Linguistics, Uppsala University, 2003. Available at <http://www.let.rug.nl/tiedeman/blog/paper/coling04.pdf>.
- Velldal, Erik och Oepen, Stephan. Maximum entropy models for realization ranking. ss 109–116, Phuket, Thailand, September 2005.
- Weijnitz, Per. Uppsala chart parser light - improving efficiency in a chart parser. Examensarbete, August 2002.
- Åberg, Stina. Datoriserad analys av sammansättningar i teknisk text. Examensarbete, December 2003.

## 9 Appendix



Figur 9.1: Nevapoäng för 28 segment



Figur 9.2: Bleupoäng för 28 segment