



UPPSALA
UNIVERSITET

Institutionen för lingvistik och filologi
Språkteknologiprogrammet
Examensarbete i datorlingvistik

9 Juni 2006

Utveckling av ett gränssnitt för uppdatering av lexikondatabasen MatsLex

Örjan Berglund

Handledare:

Anna Sågvall Hein, Uppsala Universitet

Gustav Öquist, Uppsala Universitet

Eva Pettersson, Uppsala Universitet

Beata B. Megyesi, Uppsala Universitet

Sammandrag

This paper describes the creation of a prototype of an interface for updating the lexical database MatsLex, which is used by the machine translation system MATS. The prototype is mainly aimed at those users that to some extent lack linguistic knowledge. Thus, care is taken that the linguistic information used is the minimal amount needed to match the pattern words. The parts of speech that the prototype covers are nouns and proper nouns.

The prototype consists of the actual user interface and the rules that are used for matching the pattern words. The basis of the update mechanism is to let the user provide the linguistic information needed for the words that they want to add to the database. Then the pattern words that match the given information are used to allow the user to choose the morphological pattern to be assigned to the entered word. Therefore, it is desirable that the pattern words match as few other pattern words as possible, preferably only themselves. The abilities of the prototype of matching its own pattern words is evaluated by manually matching all of the pattern words that are used within the prototype. For Swedish and English noun pattern words, the percentage of exclusive matches is 75.96% and 76.19%, respectively. For both Swedish and English proper noun pattern words the percentage of exclusive matches is 100.00%.

Innehåll

Sammandrag	ii
Innehåll	iii
Figurer	iv
Förord	v
1 Introduktion	1
1.1 Syfte	1
1.2 Översikt	2
2 Området maskinöversättning	3
2.1 Maskinöversättning	3
2.2 MATS	4
2.3 MatsLex	5
2.4 Mönsterord för lexikonrepresentation	5
3 Ett gränssnitt för uppdatering av lexikon	9
3.1 Introduktion till designprinciper inom människa-datorinteraktion . .	9
3.2 Gränssnittets förutsättningar och beståndsdelar	11
3.3 Gränssnittets utseende	12
3.3.1 Gränssnittets första steg	12
3.3.2 Gränssnittets andra steg	13
3.3.3 Gränssnittets tredje steg	16
4 Utvärdering och resultat	18
4.1 Utvärdering av matchningen av mönsterord	18
5 Diskussion	20
6 Sammanfattning	21
Litteraturförteckning	22

Figurer

3.1	Gränssnittets första steg.	13
3.2	Gränssnittets första steg. Felmeddelande.	13
3.3	Gränssnittets andra steg. Före inmatning av substantivets ordformer. . .	14
3.4	Gränssnittets andra steg. Efter inmatning av substantivets ordformer. . .	14
3.5	Gränssnittets andra steg. Felmeddelande.	15
3.6	Gränssnittets andra steg. Egennamnens fält.	15
3.7	Gränssnittets andra steg. Egennamnens fält med alternativ för personnamn.	15
3.8	Gränssnittets tredje steg. Det inskrivna ordet är ett substantiv.	16
3.9	Gränssnittets tredje steg. Det inskrivna ordet är ett egennamn.	17
3.10	Gränssnittets tredje steg. Det inskrivna ordet är ett substantiv och de mor- fosyntaktiska koderna visas.	17
4.1	Matchning av svenska substantivmönsterord.	18
4.2	Matchning av engelska substantivmönsterord.	19
4.3	Matchning av svenska egennamnsmönsterord.	19
4.4	Matchning av engelska egennamnsmönsterord.	19

Förord

Författaren tackar Anna Sågvall Hein, Gustav Öquist, Beata Megyesi och Eva Petersson för handledning i arbetet samt teknisk hjälp och vägledning.

Ett varmt tack till Hans Axelsson, Oskar Blom, Eva Ericsson, Anna Hedström, Jens Moberg och Peter Strömbäck för vänskap, stöd och samkväm.

1 Introduktion

Med maskinöversättning menas översättning som görs med hjälp av datorbaserade system från ett språk till ett annat, med eller utan människors inblandning i översättningsprocessen (Hutchins och Somers, 1992). Föresatser till och teorier kring automatisk översättning från ett språk till ett annat har funnits inom vetenskapen sedan 1600-talet. Den teknik som vår tid lever med var självklart inte uppfunnen, och de teorier som lades fram handlade främst om att skapa ett universellt språk som kunde förstås av alla människor och som var baserat på logiska principer. Sedan dess har många sådana språk tagits fram; det mest kända torde vara esperanto. Försök att med teknikens hjälp automatisera översättning lät dock vänta på sig i flera hundra år; primitiva prototyper för maskinöversättning togs fram under 1930-talet. Dock var det under 1950-talet med dess framväxt av datortekniken som forskningen kring maskinöversättning kom igång på allvar.

Maskinöversättning är en forskningsgren som faller tillbaka på många andra; den använder sig av lingvistik, datavetenskap, artificiell intelligens med flera. Det hägrande målet inom området är system som producerar felfria översättningar från ett språk till ett annat utan någon som helst inblandning av människor. Än så länge är det dock långt kvar till det målet, särskilt för andra språk än engelska. Språk är komplexa och uppbyggda på olika sätt, vilket självklart försvårar översättningsprocessen. Det finns ännu inget översättningssystem, som inte opererar inom snäva och från början bestämda ramar, vilket producerar översättningar som står på egna ben utan en översättares för- eller efterbearbetning. Utanför triviala översättningssystem är lexikon för alla språk som är inblandade i översättningen en förutsättning, och om översättningen har högre krav på sig än att översätta varje ord för sig genom tvåspråkiga lexikon, utan hänsyn till grammatik eller ordval, krävs teknik som kan analysera språkens morfologiska och grammatiska struktur.

1.1 Syfte

Målsättningen för det föreliggande examensarbetet är att skapa en prototyp till ett gränssnitt för uppdatering av en lexikal databas, MatsLex, som används av maskinöversättningssystemet MATS (Hein m.fl., 2002). Gränssnittet skapas i form av en webbsida som skall kunna ta emot de ord som dess användare ger det och klassificera dem genom att ge dem ett mönsterord för att sedan ta fram viss information som behövs vid inläggningen av orden i databasen. De språk som förekommer i MatsLex är svenska och engelska. Av yppersta vikt är att gränssnittet skapas för att användas av personer som helt eller delvis saknar lingvistisk kompetens. Detta får till följd att del av arbetet vigs åt att komma fram till vilken språklig information som är den minimala för att kunna identifiera ett ords mönsterord. För att göra detta studeras en

uppsättning mönsterord och deras morfosyntaktiska egenskaper. För att skapa ett så bra gränssnitt som möjligt konsulteras en uppsättning designregler för gränssnitt, så kallade 'gyllene regler'. Den huvudsakliga begränsning av arbetets omfång är valet av de ordklasser som systemet bakom gränssnittet kan behandla. Ordklasserna som valts ut är substantiv samt egennamn. En utvärdering av gränssnittets matchningsförmåga av sina mönsterord kommer att göras, samt en diskussion om de möjligheter till vidareutveckling som finns för gränssnittet och dess bakomliggande rutiner.

1.2 Översikt

Först ges en bakgrundsbeskrivning av det språkteknologiska området maskinöversättning samt MATS och MatsLex. Sedan följer en beskrivning av mönsterord och hur sådana använts inom svensk språkteknologisk forskning. Metodkapitlet ger förutom en bakgrund till skapandet av gränssnittet en introduktion till området människa-datorinteraktion; på detta beskrivs gränssnittets beståndsdelar. Därefter görs en utförlig genomgång av gränssnittets samtliga steg. En utvärdering av matchningen av mönsterord görs. Till sist ges en diskussion om det arbete som utförts och vad som hade kunnat göras annorlunda, samt en sammanfattning av uppsatsen.

2 Området maskinöversättning

I detta kapitel ges först en översikt över området maskinöversättning som sedan går vidare till en introduktion till maskinöversättningssystemet MATS samt dess lexikala databas, MatsLex. Sedan ges en introduktion till företeelsen mönsterord samt en mer ingående beskrivning av den uppsättning mönsterord som inkorporerats i föreliggande systems mönsterordsmatchning.

2.1 Maskinöversättning

Att översätta en text från ett språk till ett annat är svårt och kräver mycket av översättaren förutom grundliga kunskaper i de aktuella språken. Det språkteknologiska intresseområdet maskinöversättning kan sägas utgöra försök att genom varierande metoder med datorers hjälp automatisera så mycket som möjligt av översättningsprocessen (Jurafsky och Martin, 2000). Ofta är detta för svårt för att kunna göras på ett tillfredsställande sätt. Dock finns det enklare översättningsproblem som kan lösas genom befintliga maskinöversättningstekniker. Dessa är först och främst sådana problem där en ungefärlig översättning ses som tillräcklig, eller översättningar som senare bearbetas av en mänsklig översättare. Det rör sig också om sådana översättningar som ligger inom en mindre språkdomän, vilket gör att en fullgod automatisk översättning är en möjlighet. Ett exempel på en sådan språkdomän är översättning av väderleksrapporter. Flera sätt att utföra maskinöversättning har skapats. I denna sektion ges en introduktion till dem.

Oaktat hur ett maskinöversättningssystem går tillväga när det utför sina översättningsrutiner, kräver de språkliga resurser för att kunna utföra sina uppgifter (Hutchins och Somers, 1992). Dessa resurser kan i sin tur delas in i kategorierna lexikala data och grammatiska data. Med 'grammatiska data' menas den information som innefattas av och impliceras av systemets rutiner och regler för språklig analys och generering. 'Lexikala data' avser all information som finns lagrad i systemet och associerad till individuella lexikonenheter. Dessa data skiljer sig betydligt från de lexikon eller ordböcker som används av människor. Dessa kan innehålla en definition av orden i form av deras ordklass och innebörd, i förekommande fall översättningsekvivalenter, etymologi och anvisningar om uttal. Dessa fakta, som de presenteras i lexikon avsedda för människor, är ointressanta för ett lexikon som skall användas för maskinöversättning. Sådana lexikon måste innehålla all information som systemet kräver, det vill säga den information som krävs för syntaktisk och semantisk bearbetning. Detta kan inkludera förutom ordklass en subkategorisering, vilken visar på de egenskaper ordet har inom sin ordklass, till exempel huruvida ett verb är transitivt eller intransitivt. Semantisk information inkluderas också. Sådan information kan till exempel vara ett substantivs klassning som animat eller inanimat, eller om ett verb är tvunget

att ta ett mänskligt subjekt.

Ett maskinöversättningssystem kan vara tvåspråkigt eller flerspråkigt. Ett tvåspråkigt system kan vara skapat för att översätta från ett språk till ett annat eller mellan två språk; ett flerspråkigt för att översätta mellan fler än två språk. Dylika system kan också delas in i kategorierna direkta och indirekta. Kategorierna syftar på den metodik som används av systemen för att kunna producera en översättning; sker någon bearbetning av källspråkstexten, eller målspråkstexten, innan den slutgiltiga översättningen presenteras? Av de strategier som här behandlas kan endast direktöversättning hänföras till den direkta kategorin; de övriga tillhör den indirekta.

Den tidigaste och mest primitiva strategin för maskinöversättning som uppfanns kallas direktöversättning. System som använder sig av direktöversättning är i allmänhet skapade för ett specifikt språkpar, det vill säga att de är tvåspråkiga. Namnet kommer av att den saknar mellanled i sin översättningsprocess; den enda bearbetning som görs av källspråks- och målspråkstexten är den bearbetning som krävs för att kunna presentera en översatt målspråkstext. I korthet görs en morfologisk analys av källspråkstexten. Denna analys finner ordslut och böjda former och översätter genom att matcha informationen mot en tvåspråkig ordlista. Sedan följer eventuellt en viss omplacering av orden i texten för att få fram en mer acceptabel översättning, och målspråkstexten genereras.

Den första indirekta strategin för maskinöversättning kallas Interlingua. Den översätter mellan språk genom att analysera texten på målspråket och representera dess innehåll i en språkoberoende representationsform. Representationen av källspråkstexten ligger sedan till grund för genereringen av översättningen till målspråket. Detta möjliggör också översättningar från och sedan tillbaka till målspråket.

En annan strategi för indirekt maskinöversättning kallas för transfer-strategin. Namnet transfer syftar på dess indelning av översättningsprocessen i tre faser: analys, transfer och generering. Analysen görs på källspråkstexten och skapar en syntaktisk representation av den. Nästa steg, transfer, överför representationen av källspråkstexten till en representation av målspråkstexten. Genereringen, det sista steget, skapar en målspråkstext med utgångspunkt i den av transferen skapade representationen. Det bör påpekas att de båda representationerna är språkberoende, till skillnad från Interlinguametodens.

2.2 MATS

MATS (Hein m.fl., 2002) är ett regelbaserat maskinöversättningssystem utvecklat vid Uppsala Universitet genom ett samarbete mellan Institutionen för Lingvistik och Filologi, Explicon AB och Scania CV AB. Det är tvåspråkigt och översättningen sker i en riktning, från svenska till engelska. Översättningen sker inom flera olika domäner, bland andra fordonsmanualer och jordbruk.

MATS utgår från det transferbaserade översättningssystemet MULTRA (Weijnitz m.fl., 2004). Översättningsprocessen i MATS är strikt modulärt uppbyggd i självständiga steg. Varje steg är fristående och har möjlighet att inspektera resultaten av varje tidigare steg. Huvuddragen i översättningsprocessen ges nedan.

Först görs en extraktion av meningsenheter från källdokumentet. Dessa meningsenheter tokeniseras sedan. Information, morfosyntaktisk och semantisk, hämtas för de tokeniserade enheterna och en förstahandsöversättning skapas; källspråksmeningen analyseras också grammatiskt. Sedan överförs källspråkets strukturer till målspråk-

kets strukturer. Här kan, om nödvändigt, transferregler lägga till, ta bort eller omforma data för att utföra uppgiften på bästa sätt. När detta är gjort görs en första översättning till målspråket, som sedan kompletteras genom hämtning av fulla ordformer och fraser. När översättningen är klar görs diverse justeringar och förfiningar av slutprodukten (till exempel att omvandla engelskans 'a' till 'an' där så krävs). Slutligen återskapas det ursprungliga dokumentet med de översatta partierna inplacerade på sin rätta plats.

2.3 MatsLex

MatsLex är den databas som MATS använder för att hämta lexikal information (Tiedemann, 2002). Den uppsättning lexikala enheter av vilka MatsLex består är tillskrivna morfologiska, syntaktiska och semantiska egenskaper, samt länkar mellan enheterna. Enheternas egenskaper beskrivs med olika koder. En av databasens viktigaste egenskaper är att den inte lagrar några kompletta ordformer; istället används mönsterord med associerade tekniska stammar och ordformssuffix för att ta fram korrekt information om lemmat ifråga. En stor fördel med detta tillvägagångssätt är att uppdateringen av databasen blir enklare. Ett lemmas samtliga former finns tillgängliga när lemmat läggs in med sitt mönsterord i databasen. Detta är dock ett tillvägagångssätt som för det första inte passar till alla språk och för det andra är helt beroende av korrekt definierade mönsterord och förmågan att kunna länka rätt lemma till rätt mönsterord. I det absolut värsta (men tämligen osannolika) fallet skulle varje lemma få ett eget mönsterord.

Databasen MatsLex är uppbyggd för att kunna innehålla lexikala data för flera språk, och för att med hjälp av dessa skapa länkade översättningslexikon för två språk. MATS-systemet använder databasen för att hämta den lexikala information det behöver för en viss översättning och kompilerar denna information till temporära lexikon, vilka det senare använder sig av. Detta möjliggör anpassning av data till översättningsprocessen, så att varje steg får sin indata i det format som det använder. Dessutom kan varje stegs indata begränsas till precis den mängd som steget behöver. Olika versioner av de temporära lexikonerna kan också jämföras med varandra. För att säkra översättningsprocessens integritet kan uppdateringar endast göras till den centrala databasen, aldrig till något temporärt lexikon. Dessa är helt oberoende av databasen MatsLex. För att uppdateringar av MatsLex skall få effekt i de temporära lexikonerna måste dessa kompileras om.

2.4 Mönsterord för lexikonrepresentation

Ett mönsterord kan sägas symbolisera en grupp av ord inom en ordklass som delar morfologiska och grammatiska egenskaper. Om ordet 'tidning' ges mönsterordet 'stol' betyder det att 'tidning' har sitt böjningsmönster gemensamt med 'stol', samt eventuella andra språkliga egenskaper. Som visas nedan delar de två orden böjningsmönster och kan därför ges samma mönsterord.

```
stol -> tidning
stol|s -> tidning|s
stol|en -> tidning|en
```

stol|ens -> tidning|ens
stol|ar -> tidning|ar
stol|ars -> tidning|ars
stol|arna -> tidning|arna
stol|arnas -> tidning|arnas

Den huvudsakliga nyttan av att använda sig av mönsterord är att samtliga former för ett ord i ett lexikon kan genereras utan att de anges explicit för varje ord, vilket sparar lexikonresurser.

Staffan Hellberg har vid Göteborgs Universitet utvecklat en textanalysator för svensk text som består av en ordlista och en uppsättning mönsterordsdefinitioner (Hellberg, 1978). Mönsterordsdefinitionerna är uppdelade i paradigmer, där varje paradigm ges ett nummer och innehåller ett specifikt mönsterord tillsammans med detta mönsterords ordklassstillhörighet samt i förekommande fall nödvändig morfosyntaktisk information för mönsterordet. Förutom detta anges hänvisningar till subrutiner som kontrollerar eventuella ambiguiteter hos de analyserade orden.

Som en del i projektet En Lexikonorienterad parser för Svenska (LPS) togs ett stort stamlexikon för svenska fram (Hein och Sjögreen, 1990). Stamlexikonet har sin utgångspunkt i samma lexikala databas som ligger till grund för Svensk Ordbok (Allén, 1986). Detta stamlexikon är en delkomponent i LPS-analysatorn, en böjningsmorfologisk analysator för svenska. LPS-analysatorn opererar på enskilda ordformer och klassificerar dem morfologiskt, under förutsättning att stammarna står att finna i lexikonet och att grammatiken accepterar deras morfologiska struktur. För varje lexikoningång anges dess böjningstyp genom ett mönsterord. Varje mönsterord hänvisar till en böjningsmorfologisk regel som formellt definierar de egenskaper som mönsterordet medför.

Genom ett samarbete mellan Uppsala Universitet och EC Systran utvecklades ett svensk-engelskt lexikon med utgångspunkt i jordbruksdomänen (Gustavii och Petersson, 2003). Till detta lexikon skapades en uppsättning mönsterordsdefinitioner. Dessa baseras huvudsakligen på de mönsterord som används i MATS-databasen, vilka har sin utgångspunkt i Hellbergs definitioner. Definitionerna har gjorts för svenska och engelska. För varje mönsterordsdefinition ges först mönsterordet. Sedan ges alla morfosyntaktiska koder som angår just detta mönsterord och ett eventuellt reguljärt uttryck som är till för att visa huruvida ett ord böjs oregelbundet, tappar bokstäver vid böjning eller liknande. Sist visas de ändelseaffix som ordens alla former uppvisar. Nedan ges först en kort förklaring till de morfosyntaktiska koderna, sedan en introduktion till de olika mönsterordsdefinitionerna.

En morfosyntaktisk kod är en bokstavskombination som visar upp ett ords olika egenskaper. De egenskaper som är relevanta för de koder som ges de ordklasser som förekommer i detta arbete, substantiv och egennamn, är ordklass, genus, numerus, bestämdhet samt kasus. Ordklass betecknas antingen substantiv ('NN') eller egennamn ('PN'). Genus är specifikt för de svenska mönsterorden och har varianterna utrum ('U') samt neutrum ('N'). Numerus innefattar singular ('S') samt plural ('P'). Egenskapen bestämdhet, vilken enbart förekommer för svenska substantiv, har alternativen indefinit ('I') samt definit ('D'). Kasus har två former, nämligen grundform ('B') samt genitiv ('G'). Dessutom kan vissa egenskaper underspecificeras, vilket görs genom ett 'X'. Underspecificering innebär att den aktuella egenskapen inte påverkar ordformen.

Den morfosyntaktiska koden för de svenska substantiven har följande syntax:

ordklass | genus | numerus | bestämdhet | kasus

FLICKA	NNUSIB	\$	a
FLICKA	NNUSIG	\$	as
FLICKA	NNUSDB	\$	an
FLICKA	NNUSDG	\$	ans
FLICKA	NNUIB	\$	or
FLICKA	NNUIG	\$	ors
FLICKA	NNUPDB	\$	orna
FLICKA	NNUPDG	\$	ornas

Egenskaperna genus, numerus, bestämdhet och kasus kan underspecificeras för svenska substantiv. Den första delen av de svenska substantivmönsterorden är indelad efter mönsterordens morfologiska egenskaper i deklinationer. Det finns sju deklinationer, vilka följer nedan.

Dekl.1 : '-or' som pluralsuffix
Dekl.2 : '-ar' som pluralsuffix
Dekl.3 : '-er' som pluralsuffix
Dekl.4 : '-r' som pluralsuffix
Dekl.5 : '-n' som pluralsuffix
Dekl.6 : inget pluralsuffix
Dekl.7 : '-s' som pluralsuffix

Den senare delen består av substantiv som har en avvikande böjning, eller saknar singular- eller pluralformer.

Den morfosyntaktiska koden för svenskans egennamn har följande syntax:

ordklass | numerus | kasus

PER	PMUBM	\$	
PER	PMUGM	\$	s
GUNILLA	PMUBF	\$	
GUNILLA	PMUGF	\$	s

För de svenska egennamnen kan egenskapen kasus underspecificeras. I de fall mönsterorden för de svenska egennamnen är specifikt maskulina eller feminina markeras detta genom ett 'M' eller ett 'F' sist i den morfosyntaktiska koden.

Den morfosyntaktiska koden för engelskans substantiv har följande syntax:

ordklass | numerus | kasus

BUSH	NNPB	\$	es
BUSH	NNPG	\$	es\'
BUSH	NNSB	\$	
BUSH	NNSG	\$	\'s

Både numerus och kasus kan underspecificeras för de engelska mönsterorden.
Den morfosyntaktiska koden för engelskans egennamn har följande syntax:

ordklass | numerus | kasus

SCANIA PMSB \$

SCANIA PMSG \$ \ ' s

För de engelska egennamnen kan både numerus och kasus underspecificeras.

3 Ett gränssnitt för uppdatering av lexikon

I detta kapitel ges först en introduktion till människa-datorinteraktion samt en kort förklaring till vad en uppsättning 'gyllene regler' för design är tänkt att åstadkomma. Därefter redogörs för de delar som gränssnittet består av och förutsättningar för det samma. Sedan följer en översikt över själva gränssnittet, där dess alla steg redovisas utförligt.

3.1 Introduktion till designprinciper inom människa-datorinteraktion

Forskningsområdet människa-datorinteraktion (MDI) är ett förhållandevis nytt sådant. Sedan 1940-talet har forskning skett inom interaktionen mellan människor och maskiner, men det är under de senaste decennierna som datorn kommit i fokus som maskinen ifråga (Dix m.fl., 2004). Termen 'människa-datorinteraktion' kom till under 1980-talet.

Disciplinen MDI faller tillbaka på ett flertal vetenskapliga discipliner, men det är inom datavetenskapen och systemdesign som dess huvudanvändning ligger. Dock är det inte nödvändigtvis så att interaktionen ifråga iscensätts av en ensam datoranvändare vid en stationär dator. Termen 'användare' kan betyda en ensam användare, en grupp av samtida användare av systemet, en sekvens av individuella användare som har olika uppgifter att utföra med hjälp av systemet.

Med 'dator' innefattas all datorbaserad teknologi från en vanlig dator till ett storskaligt datorsystem. Också system som kontrollerar processer och andra teknologier kan komma ifråga. Som 'interaktion' ses all kommunikation mellan en användare och en dator, antingen direkt eller indirekt. Med direkt interaktion menas en dialog mellan användaren och systemet som innefattar återkoppling och att användaren har kontroll över den pågående processen. En indirekt kommunikation kan vara att systemet bearbetar data från användaren. Dessa tre huvudtermer förklaras mer ingående nedan.

Som användare avses en eller flera människor. Människor har vissa egenskaper och begränsningar, vilka måste tas i beaktande vid skapandet av ett interaktivt system. En människa tar in, lagrar, bearbetar, reagerar på och använder information genom sina sinnen. Vid användande av datorer används i första hand känsel, syn och beröring för upptagande av information. Denna information lagras i människans minne, antingen för stunden i det sensoriska minnet eller i arbetsminnet, eller permanent i långtidsminnet. Förståelse och kompenserande för bristerna i människans informationsbearbetning och informationslagring kan hjälpa en designer skapa en bättre interaktion.

Inte bara människor har egenskaper och begränsningar. Datorer och datorsystem har dem också, om än skilda från människors. När datorer lagrar information använder de ett lagringsmedium, företrädesvis en hårddisk eller ett RAM-minne, vars utrymme inte är obegränsat. Vid interaktion med en dator använder sig människor av hjälpmedel för att överhuvudtaget kunna interagera med datorn. Dessa hjälpmedel kan till exempel vara en skärm, ett pekdon, ett tangentbord eller något annat hjälpmedel för textinmatning, röststyrning med flera andra. Skapare av interaktiva system måste vara väl medvetna om egenskaperna hos den apparatur som deras system skall användas i eller med. Dessa egenskaper måste i lika hög grad tas hänsyn till som de hos människorna.

Människan och datorn möts, vilket kanske anats, inom interaktionen. Interaktion kan ses som en dialog mellan användaren och ett system. Den kan ske på en mängd olika sätt och är beroende av många olika faktorer. Dessa kan vara det gränssnitt som systemet använder, användarens erfarenhet av systemet, den sociala och organisatoriska kontext i vilken interaktionen sker och så vidare. Det finns flera olika metoder för att modellera den interaktion som äger rum mellan ett system och en användare. Den kanske mest kända metoden är skapad av Donald A. Norman och kallas Norman's Model (Norman, 1988). Den utgår helt från användarens sida av interaktionen och går i korthet ut på att användaren formulerar en handlingsplan, vilken sedan utförs med hjälp av systemets gränssnitt. När planen helt eller delvis utförts kontrollerar och utvärderar användaren resultatet av den via gränssnittet. Användaren formulerar sedan en ny handlingsplan med utgångspunkt i resultatet av den förra.

Genom att studera abstrakta principer för design och sedan skapa regler som förkroppsligar dessa har flera forskare inom MDI skapat regeluppsättningar för design av interaktiva system, 'gylle regler' med ett annat uttryck (Dix m.fl., 2004). Dessa regeluppsättningar är i regel generella och kan ibland inte användas fullt ut vid skapandet av vissa system. Dock är de en god hjälp i designprocessen och om de används blir slutresultatet ofta bättre än om de inte använts.

En sådan regeluppsättning har skapats av Ben Schneiderman (Schneiderman, 1998). Det är denna uppsättning som använts som vägledning i skapandet av det gränssnitt som är detta arbetes mål. Reglerna ges nedan; de står i kursiverad stil och undertecknads förklaringar står i normalt typsnitt.

1. *Sträva efter enhetlighet i händelsesekvenser.* Upplägg, terminologi, kommandon och andra komponenter bör visa upp en så stor konsekvens som möjligt.

2. *Möjliggör för vana användare att använda genvägar.* Exempel på genvägar kan vara förkortningar och makron, för att erfarna användare skall kunna utföra välkända och ofta förekommande moment snabbare.

3. *Ge informativ återkoppling.* Alla handlingar som användaren kan utföra inom systemet bör följas av en bekräftelse som passar handlingens vikt inom systemet.

4. *Utforma dialoger så att användaren vet när de slutar.* Det underlättar enormt för användaren om han eller hon vet att en uppgift är slutförd.

5. *Förhindra fel och använd felhantering.* Användarna bör så långt det är möjligt hindras från att göra misstag och om de ändå gör fel, ge dem klara instruktioner om hur de kan rätta till felet.

6. *Tillåt användaren att enkelt gå tillbaka till ett tidigare steg.* Detta minskar eventuell oro och stress samt uppmuntrar utforskning av gränssnittet, då användaren vet att han eller hon alltid kan gå tillbaka till det förra steget.

7. *Stöd ett internt kontrollokus.* Användaren kontrollerar systemet, vilket svarar på hans eller hennes handlingar.

8. *Minska belastningen på användarens korttidsminne.* Genom att göra gränssnittet lättfattligt, använda flera sidor inom gränssnittet och ge tid för inlärningsperioder minskar belastningen på användarens korttidsminne, vilket minskar risken för att fel görs.

3.2 Gränssnittets förutsättningar och beståndsdelar

Det gränssnitt som skapats består först och främst av en huvudfil som innehåller själva gränssnittet och vissa funktioner som behövs för behandlingen av orden i det samma. Huvudfilen tar även hjälp av separata filer innehållandes språkregler, vilka hjälper den att komma fram till rätt mönsterord, samt en fil som innehåller funktioner. Dessa funktioner används först och främst av språkregelfilerna.

Grundförutsättningen för uppdateringsfunktionen i gränssnittet är att användaren ger det information om det ord som skall läggas in i databasen. Systemet tar hand om informationen och använder den för att ge det inskrivna ordet ett eller flera alternativa mönsterord, vilka användaren sedan använder för att välja vilken morfologisk information som skall läggas till ordet i databasen. För att kunna komma fram till vilket mönsterord ett ord har skapades regelfiler för detta ändamål. Dessa regelfiler är fyra till antalet, en fil för varje språk och ordklass. De tar fram mönsterord med hjälp av matchning av de ordformer som användaren angett gentemot de mönsterordsdefinitioner som angetts av Eva Pettersson och Ebba Gustavii (Gustavii och Pettersson, 2003). Dessa regelfiler är uppställda efter de i mönsterordsdefinitionerna angivna deklinationerna. Vilka krav som ställs för att matcha ett mönsterord beror i någon mån på vilket mönsterord det handlar om, men i normalfallet kontrolleras all information som efterfrågas i gränssnittet. För svenska substantiv innebär detta ordets genus, singular- och pluraländelser samt ordets bestämda form; vad gäller de engelska dito efterfrågas ordets singular- och pluraländelser. Studier av mönsterordsdefinitionerna visade på att dessa tre respektive två former var de som i mycket stor utsträckning skilde de olika mönsterorden åt; dock finns svenska mönsterord som endast kan skiljas åt genom kontroll av den bestämda pluralformen (till exempel 'kilo' och 'garage'). Detta skulle dock innebära att en ny ordform måste efterfrågas av användaren, vilket sågs som överflödigt med tanke på hur få av de svenska mönsterorden som har dessa egenskaper. I de fall ett mönsterord har alternativa former för en eller flera av sina ordformer, läggs båda in i regeln om ordformerna efterfrågas av matchningsfunktionen. Egennamn kontrolleras genom ordets grundform och genitivform, vilket gäller för bägge språken.

Ett mindre antal mönsterord har uteslutits från matchningsreglerna. De mönsterord som uteslutits som är gemensamma för svenskan och engelskan är de mönsterord som beskriver kvantitativa fraser, det vill säga mönsterorden 'en grupp' och 'a handful'. Förutom dessa uteslöts bland de svenska egennamnen tre mönsterord. Dessa är 'af', 'jordbruksavd' och 'jordbruksv'. Det första uteslöts på grund av att det är en del av ett efternamn, vilka redan finns representerade bland egennamnsmönsterorden. De två andra har det gemensamt att de är förkortningar, och uteslöts därför att de sågs som för lika redan existerande mönsterord. Vad gäller de svenska substantiven uteslöts där två mönsterord som betecknar månader, 'januari' och 'mars'. Förutom dessa uteslöts fyra mönsterord för sådana substantiv som enbart står i pluralform. Skälet till detta är systemets utformning. Eftersom det strävar efter att efterfråga så lite grammatisk information som möjligt - för de svenska substantivmönsterorden

formerna obestämd singular, bestämd singular, obestämd plural - kan det inte skilja mönsterord som står i enbart plural från varandra, då det får endast en ordform att arbeta med. Mot denna bakgrund togs dessa mönsterord - 'kläder', 'vägnar', 'kr' samt 'glasögon' - bort.

Funktionsfilen innehåller funktioner som skrivits för att kunna skriva mer precisa regler för bestämning av ett inskrivet ords mönsterord. Till exempel finns här funktioner för att ta fram ett ords stam, för att kontrollera om en bokstav är en konsonant eller en vokal samt för att kontrollera huruvida ett svenskt ord böjs med omljud i pluralformen, och i så fall med vilka vokaler.

3.3 Gränssnittets utseende

De designprinciper som föreliggande gränssnitt stödjer sig mot är Schneidermans åtta regler, vilka tidigare redovisats. Innan varje steg i gränssnittet redogörs för i detalj, märk de regler som där är allerstädes närvarande. Främst gäller detta den åttonde regeln, vilken påbjuder att så långt det är möjligt minska belastningen på användarens korttidsminne och inte tvinga honom eller henne att fylla i all information i ett enda steg. Detta är ett av flera skäl till att gränssnittet är uppdelat i tre delar istället för, till exempel, en enda eller kanske två. Uppdelningen i tre steg minskar inte bara belastningen på användarens korttidsminne, utan gör det också lättare för användaren att veta var i inläggningsprocessen han eller hon befinner sig jämfört med möjligheten att all information skall skrivas in i ett och samma steg. Den tredje regeln följs i den meningen att felmeddelanden ges om användaren gjort fel och att gränssnittet flyttar fram ett steg om användaren gjort allt rätt; den fjärde regeln efterlevs genom ett popup-fönster som talar om för användaren att ordet lagts till databasen, för att sedan transportera användaren tillbaka till gränssnittets första sida. Den andra regeln efterlevs egentligen inte i detta gränssnitt eftersom användaren måste fylla i alla former enligt gränssnittets ramar, vilket nämns i diskussionen. Enligt den femte regeln ges felmeddelanden i de fall användaren glömt fylla i någon av de nödvändiga formerna eller om ett mönsterord inte hittas för något av orden. Schneidermans sjätte regel säger att användarna alltid skall ha möjlighet att göra handlingar ogjorda, vilket uppmuntrar utforskande och minskar oro för att göra fel. I enlighet med denna regel har användarna alltid möjligheten att gå tillbaka ett eller flera steg i gränssnittet om de märker att de gjort något fel. Kanske har användaren stavat någon av ordformererna fel och vill ändra den. Det finns knappar för detta ändamål som låter användaren gå ett steg tillbaka från alla steg utom det första. Angående den sjunde regeln gör systemet ingenting utan att användaren har instruerat det att göra något, vilket gör att användaren har full kontroll över systemet. Till sist har konsekvens i gränssnittet gällande dess utseende särskilt eftersträvat, efter Schneidermans första regel. Allt som specifikt gäller något av språken har placerats ihop i en kolumn i gränssnittet, vilket underlättar för användaren att hålla de olika språken skilda från varandra.

Gränssnittet är skapat med hjälp av scriptspråket PHP samt databashanteraren MySQL. Det återfinns här: <http://stp.ling.uu.se/~xyu/1/>

3.3.1 Gränssnittets första steg

Det användaren tänks göra i detta steg är att skriva in singularformerna / grundformerna av det ordpar han eller hon vill lägga in i sin databas. Orden skrivs in med sin

svenska och sin engelska motsvarighet. Redan här etableras en konsekvens i gränssnittet, i enlighet med Schneiders första regel, genom att placera textrutorna för inmatning av de svenska och engelska orden bredvid varandra i två kolumner.

Steg 1

Skriv in det ord du vill lägga in i lexikonet på svenska och engelska.

Svenska:	Engelska:
<input type="text" value="öga"/>	<input type="text" value="eye"/>
<input type="button" value="Fortsätt"/>	

Figur 3.1: Gränssnittets första steg.

Detta sker konsekvent under de steg där någon form av inmatning sker, det vill säga de två första. Förutom detta är textrutorna samt den för textrutorna utmärkande texten ('Svenska' samt 'Engelska') ovanför dem placerade i rak vinkel mot varandra.

I det fall användaren försöker gå vidare till nästa steg och inte har fyllt i något ord för antingen engelska eller svenska hindras han eller hon att gå vidare och blir visad ett felmeddelande, i enlighet med Schneiders femte regel.

Svenska:	Engelska:
<input type="text" value="öga"/>	<input type="text"/>
<input type="button" value="Fortsätt"/>	

Ett ord på båda språken måste anges.

Figur 3.2: Gränssnittets första steg. Felmeddelande.

De två ord som här skrivs in skickas vidare till nästa steg i gränssnittet.

3.3.2 Gränssnittets andra steg

I detta steg måste användaren först ange huruvida det inskrivna ordet är ett substantiv eller ett egennamn. Substantiv ses som normalfallet och visas automatiskt vid stegets början.

I det fall ordet är ett substantiv anger användaren det svenska ordets genus, pluralform och bestämda singularform. För att underlätta för användaren anges substantivets genus framför dess bestämda form genom att ange 'Den' eller 'Det' baserat på det val av genus som gjorts med hjälp av rullisten framför den obestämda singularformen. Användaren skriver också in det engelska ordets pluralform i den högra kolumnen.

Steg 2

Välj en ordklass för det svenska ordet *öga* och fyll i korrekta värden för genus samt alla angivna ordformer.

Ordklass:

Svenska:		Engelska:	
<input type="text" value="En"/>	<input type="text" value="öga"/>	One	<input type="text" value="eye"/>
Flera	<input type="text" value="öga"/>	Several	<input type="text" value="eye"/>
Den	<input type="text" value="öga"/>		

Figur 3.3: Gränssnittets andra steg. Före inmatning av substantivets ordformer.

Ordklass:

Svenska:		Engelska:	
<input type="text" value="Ett"/>	<input type="text" value="öga"/>	One	<input type="text" value="eye"/>
Flera	<input type="text" value="ögon"/>	Several	<input type="text" value="eyes"/>
Det	<input type="text" value="ögat"/>		

Figur 3.4: Gränssnittets andra steg. Efter inmatning av substantivets ordformer.

Om användaren lämnar något av de nödvändiga fälten tomt ges ett felmeddelande. Se Figur 3.5.

För svenska egennamn anges förutom genitivformen även huruvida egennamnet är ett personnamn, icke ett personnamn men inte heller en förkortning, eller en förkortning. Detta görs genom att klicka i en av tre rutor under egennamnets grundform. Se Figur 3.6.

Om det rör sig om ett personnamn, anger användaren också vilken sorts personnamn det handlar om; mansnamn, kvinnonamn eller ett efternamn. I de fall egennamnet varken är ett personnamn eller en förkortning anges dess genus genom att kryssa för antingen 'Den' eller 'Det'. De engelska egennamnen lämnas utan åtgärd från användaren. Se Figur 3.7.

Textfälten för de svenska och de engelska ordformerna är även i detta steg placerade i två separata tabeller, vilka är lagda bredvid varandra. Också här är den utmär-

Svenska:	Engelska:
Ett <input type="text" value="öga"/>	One <input type="text" value="eye"/>
Flera <input type="text" value="ögon"/>	Several <input type="text" value="eyes"/>
Det <input type="text"/>	

Du måste fylla i alla nödvändiga fält.

Figur 3.5: Gränssnittets andra steg. Felmeddelande.

Ordklass:

Svenska:	Engelska:
Detta är <input type="text" value="Gunnar"/> <input type="checkbox"/> Ordet är ett personnamn. <input type="checkbox"/> Ordet är inte ett personnamn. <input type="checkbox"/> Ordet är en förkortning.	This is <input type="text" value="Gunnar"/>

Figur 3.6: Gränssnittets andra steg. Egennamnens fält.

Svenska:	Engelska:
Detta är <input type="text" value="Gunnar"/> <input checked="" type="checkbox"/> Ordet är ett personnamn. <input type="checkbox"/> Ordet är inte ett personnamn. <input type="checkbox"/> Ordet är en förkortning.	This is <input type="text" value="Gunnar"/>
<input type="radio"/> <i>Man</i> <input type="radio"/> <i>Kvinna</i> <input type="radio"/> <i>Efternamn</i>	

Figur 3.7: Gränssnittets andra steg. Egennamnens fält med alternativ för personnamn.

kande texten och textrutorna placerade i rak vinkel mot varandra. Knappraden under textrutorna är utökad jämfört med den som fanns i det första steget, vilket endast hade en knapp, 'Fortsätt'. I det andra och det tredje steget består knappraden av tre knappar. I det andra steget är dessa knappar 'Fortsätt', 'Tillbaka' och 'Avbryt'. Den första av de tre låter användaren fortsätta till nästa steg i gränssnittet under förutsättning att alla villkor för det steg han eller hon står i är uppfyllda, alltså att inga fält är tomma. 'Tillbaka' låter användaren förflytta sig ett steg tillbaka i gränssnittet, ifall någonting behöver rättas till. Knappen 'Avbryt' transporterar användaren till gränssnittets första sida oberoende av vilket steg han eller hon befinner sig på.

3.3.3 Gränssnittets tredje steg

Det tredje steget visar användaren två tabeller, en för det svenska ordet och en för det engelska. Tabellen för det svenska ordet består av två kolumner, vilket tabellen för det engelska ordet också gör. I den vänstra kolumnerna anges alla grammatiska former för ordet enligt det mönsterord som ordet blivit tilldelat. I det fall ordet har matchat mer än ett mönsterord kan användaren välja bland dem i en rullist som finns placerad snett ovanför respektive tabell. Denna rullist finns alltid synlig, men om antalet mönsterord är färre än två kan den enbart visa det gällande mönsterordet. Högerkolumnen i tabellen för det svenska ordet visar alla grammatiska former för det ord användaren tidigare angett i en grammatisk kontext. Tanken bakom detta är att användaren skall jämföra de grammatiska former som mönsterordet uppvisar och de som finns representerade genom det ord som tidigare angetts. Se Figur 3.8. Steget är utformat på precis samma sätt för egennamnen. Se Figur 3.9.

Steg 3

Välj det mönsterord som passar det ord du skrivit in.

Svenska		Engelska	
<input type="text" value="öga"/>	<input type="text" value=""/>	<input type="text" value="dog"/>	<input type="text" value=""/>
Form	Exempel	Form	Exempel
öga	Ett öga	eye	One eye
ögas	Ett ögas	eye's	The eye's
ögat	Det ögat	eyes	Several eyes
ögats	Det ögats	eyes'	The eyes'
ögon	Flera ögon		
ögons	Flera ögons		
ögonen	De ögonen		
ögonens	De ögonens		

Visa morfkod

Figur 3.8: Gränssnittets tredje steg. Det inskrivna ordet är ett substantiv.

Under tabellen för det svenska ordet återfinns en kryssruta med beteckningen 'Visa morfkod'. Då denna klickas i visas en tidigare dold mittenkolumn som visar varje ordforms morfosyntaktiska kod. För att underlätta för användaren har funktioner implementerats som skriver ut tabellerna ordnade efter grammatiska egenskaper, i fallande ordning: obestämd singular, bestämd singular, obestämd plural, bestämd plural. Eventuella alternativa ordformer placeras tillsammans med sin egen form.

Svenska		Engelska	
<input type="text" value="per"/>		<input type="text" value="scania"/>	
Form	Exempel	Form	Exempel
Gunnar	Han heter Gunnar	Gunnar	It is called Gunnar
Gunnars	Den är Gunnars	Gunnar's	It is Gunnar's

Figur 3.9: Gränssnittets tredje steg. Det inskrivna ordet är ett egennamn.

Detta gjordes då det upptäcktes att ordningen annars kunde vara något omkastad, vilket skulle kunna vara till men för grammatiskt obehövande användare. Om alla former i tabellen visar sig vara korrekta har också det korrekta mönsterordet hittats. Se Figur 3.10.

Svenska			Engelska		
<input type="text" value="öga"/>			<input type="text" value="dog"/>		
Form	Morfkod	Exempel	Form	Morfkod	Exempel
öga	NNNSIB	Ett öga	eye	NNSB	One eye
ögas	NNNSIG	Ett ögas	eye's	NNSG	The eye's
ögat	NNNSDB	Det ögat	eyes	NNPB	Several eyes
ögats	NNNSDG	Det ögats	eyes'	NNPG	The eyes'
ögon	NNNPIB	Flera ögon			
ögons	NNNPIG	Flera ögons			
ögonen	NNNPDB	De ögonen			
ögonens	NNNPDG	De ögonens			

Figur 3.10: Gränssnittets tredje steg. Det inskrivna ordet är ett substantiv och de morfosyntaktiska koderna visas.

Om det skulle vara så att inget mönsterord kan tilldelas något av de inskrivna orden, eller båda, visas ett felmeddelande som uppmanar användaren att kontrollera de under steg 2 inskrivna ordformerna. I det tredje steget har knappraden längst ned ändrats något; knapparna är nu 'Lägg in', 'Tillbaka' och 'Avbryt'. Det som skiljer knappraden i det tredje steget från det andra är knappen 'Lägg in'. Denna knapp lägger in de två orden i en provisorisk databas tillsammans med nödvändiga uppgifter. När detta gjorts kommer användaren också transporteras tillbaka till det första steget för att lägga in ytterligare ett ordpar.

4 Utvärdering och resultat

Här redogörs för utvärderingen av systemet samt resultaten av densamma.

4.1 Utvärdering av matchningen av mönsterord

De huvudsakliga sätt som finns att tillgå för att utvärdera föreliggande systems matchningsförmåga är att matcha dess egna mönsterord, eller att använda en samling ord av lämplig längd, till exempel från ett allmänlexikon. Den metod som valdes var att testa matchningen av systemets egna mönsterord, då det är av vikt att dessa matchar så få mönsterord som möjligt.

För att kontrollera systemets matchningsförmåga av sina egna mönsterord testades alla i systemet förekommande mönsterord manuellt. De svenska mönsterordens alternativa former har testats i de fall de förekommer, det vill säga att 'liter' testats två gånger, först med nollpluralformen och sedan med sin alternativa pluralform (liter-litrar).

Det är viktigt att poängtera att inget mönsterord saknar matchning av sig självt. Detta innebär att begreppet 'enkelmatchning' syftar på sådana mönsterord som i systemet enbart matchas av sig själva. Begreppen 'dubbelmatchning' samt 'trippelmatchning' betecknar de mönsterord som matchar ett respektive två mönsterord förutom sig själva. Dessa flermatchningar kan delas in i två grupper:

1) De mönsterord som matchar fler mönsterord än sig själva på grund av för stora likheter mönsterorden emellan; detta sker på grund av att ett eller flera av mönsterorden har en alternativ form.

2) De mönsterord som matchar fler mönsterord än sig själva på grund av att skillnaderna dem emellan inte täcks av systemet utan att något av mönsterorden har någon alternativ form.

Nedan ges resultaten från utvärderingen.

	Antal (procent)
Antal mönsterord	104 (100.00%)
Antal enkelmatchningar	79 (75.96%)
Antal dubbelmatchningar	13 (12.50%)
Antal trippelmatchningar	12 (11.54%)

Figur 4.1: Matchning av svenska substantivmönsterord.

För merparten av de svenska mönsterord som matchar fler mönsterord än sig själva gäller att de tillhör den första kategorin, alltså att de uppvisar så stora likheter att systemet inte kan skilja dem åt på grund av att något av dem, eller fler än ett, har en

alternativ form som gör det omöjligt för systemet att särskilja dem. Till exempel matchar 'seger' alltid 'liter' då 'liter' har två olika pluralformer (en liter, flera liter samt en liter, flera liter) vilket gör att systemet inte kan särskilja dem. Som en blandning mellan de båda klasserna kan mönsterordet 'vin' nämnas. Det matchar, förutom sig självt, de båda mönsterorden 'parti' och 'intervall'. Detta sker på grund av att 'parti' har två alternativa former för bestämd form singular; 'partiet' och 'partit', vilket gör att det matchar 'vin'. Vad gäller 'intervall' överensstämmer dess ordformer helt och fullt med de för 'vin', med tillägget att det har två alternativa pluralformer, vilket gör att det står som ett eget mönsterord. En mer renodlad representant för den andra gruppen flermatchningar är mönsterordsparet 'kilo' och 'garage' vilka också har helt gemensamma former förutom den i bestämd form plural (garagen - kilona).

	Antal (procent)
Antal mönsterord	21 (100.00%)
Antal enkelmatchningar	16 (76.19%)
Antal dubbelmatchningar	5 (23.81%)

Figur 4.2: Matchning av engelska substantivmönsterord.

De engelska mönsterordens dubbelmatchningar uppkommer uteslutande av det faktum att de enbart kontrolleras genom substantivets singular- och pluralformer, det vill säga att de tillhör den andra gruppen dubbelmatchningar. De två mönsterord som saknar pluralsuffix, 'aircraft' samt 'series' kan därför inte särskiljas från 'meter', det mönsterord som enbart förekommer i bestämd form, oavsett om det står i singular- eller pluralform. På samma sätt kan inte 'music' skiljas från 'toe-in', vilka har liknande förutsättningar med skillnaden att dessa mönsterord enbart förekommer i singularformen.

Vad gäller de svenska egennamnen, och de engelska, matchas samtliga mönsterord enbart av sig själva. I och med att de är så få - engelskan har ett enda mönsterord för egennamn - är det föga förvånande att matchningen blir hundra procentig.

	Antal (procent)
Antal mönsterord	8 (100.00%)
Antal enkelmatchningar	8 (100.00%)

Figur 4.3: Matchning av svenska egennamsmönsterord.

	Antal (procent)
Antal mönsterord	1 (100.00%)
Antal enkelmatchningar	1 (100.00%)

Figur 4.4: Matchning av engelska egennamsmönsterord.

5 Diskussion

Resultaten från utvärderingen av mönsterordsmatchningen visade på en icke obetydlig överensstämmelse mellan antalet enkelmatchningar av svenskans och engelskans substantivmönsterord, trots att deras antal var nästan fem gånger större för svenskan än engelskan. Dessutom var antalet enkelmatchningar av egennamnsmönsterorden hundraprocentigt i bägge fallen, vilket kanske förvånar mindre på grund av deras mindre antal (åtta respektive ett) samt att de är mindre komplexa, morfologiskt sett, jämfört med substantiven. De dubbelmatchningar och trippelmatchningar som förekommer bland substantivmönsterorden är i viss mån oundvikliga om de mönsterord som används skall kunna täcka någon betydande del av sitt språk, särskilt för den morfologiskt mer komplexa svenskan. I viss mån kan överlappande matchningar hindras på bekostnad av den lingvistiska variationsrikedomen inom språken; som exempel är det en möjlighet att eliminera alternativa former för de mönsterord som har sådana och att slå ihop mönsterord vars ordformer överlappar i hög grad - som exempel kan tas 'garage' och 'kilo', vilka enbart skiljs åt genom sin bestämda pluralform ('garagen' respektive 'kilona'). Den väg som väljs - variation eller förenkling - beror naturligtvis mycket på den förestående uppgiften.

Under arbetet med prototypen har i första hand reglerna för mönsterordsmatchning men även också i viss mån själva gränssnittet ständigt byggts på, omprövats och utvecklats. Matchningsrutiner och andra funktioner som används av matchningsrutinerna fungerade vid en viss punkt i utvecklingen men var tvungna att bytas ut eller ändras när nya mönsterord skulle matchas. Gränssnittet har under arbetets gång getts nya uppgifter och funktioner men dess utseende har inte ändrats i nämnvärd grad. Dock finns det flera möjligheter till utveckling och förbättring av prototypen, både vad gäller dess rutiner och själva gränssnittet.

Den mest grundläggande av förändringarna påverkar genom sin natur både gränssnittet och rutinerna, nämligen att antalet ordklasser som gränssnittet behandlar skulle kunna utökas. Detta skulle självklart göra det mer lämpat för sin uppgift, att uppdatera en lexikal databas. En mer utförlig utvärdering av mönsterordens matchning och täckningsgrad skulle också vara lämplig, till exempel att använda en större samling testdata än enbart systemets egna mönsterord, samt att undersöka huruvida det finns ett behov av fler eller färre mönsterord i systemet. En utvärdering av själva gränssnittet med avseende på dess användarvänlighet har heller inte gjorts inom detta examensarbete, vilket skulle visa på eventuella förbättringsmöjligheter för gränssnittet.

6 Sammanfattning

Denna uppsats har beskrivit utvecklingen av en prototyp till ett gränssnitt för uppdatering av lexikondatabasen MatsLex. Prototypen är riktad mot sådana användare som har begränsade kunskaper inom lingvistik. Inledningsvis gjordes en bedömning av vilken språklig information som är den minsta möjliga för att identifiera ett mönsterord. Denna information efterfrågas sedan av användaren av prototypen via ett inom arbetet skapat gränssnitt. Det har tagits fram med hjälp av en uppsättning mönsterord samt en samling designregler för gränssnitt.

Prototypens regler för matchningen av mönsterord har utvärderats manuellt genom att alla i reglerna uppställda mönsterord har körts igenom systemet för att kontrollera om de hittas av reglerna, och i så fall hur många andra mönsterord de matchar. Inget mönsterord saknade matchning av sig självt. Av de svenska substantivens mönsterord matchade 75.96% enbart sig själva och 12.50% matchade ett mönsterord förutom sig själva. Andelen som matchade två mönsterord samt sig själv var 11.54%. De svenska egennamnsmönsterorden matchade alla sig själva. Av de engelska substantivens mönsterord matchade 76.19% enbart sig själva och 23.81% matchade ett mönsterord förutom sig självt. Det engelska egennamnsmönsterordet matchade sig självt.

Litteraturförteckning

- Allén, Sture, redaktör. *Svensk Ordbok*. Esselte Studium AB, 1986.
- Dix, Alan, Finlay, Janet, Abowd, Gregory D., och Beale, Russell. *Human-Computer Interaction*. Pearson Education Limited, 3 utgåvan, 2004.
- Gustavii, Ebba och Pettersson, Eva. *Utveckling av ett Svenskt-Engelskt Lexikon Inom Jordbruksdomänen*. Institutionen för Lingvistik och Filologi, Uppsala Universitet, 2003.
- Hein, Anna Sågvall, Forsbom, Eva, Tiedemann, Jörg, Weijnitz, Per, Almqvist, Inger, Olsson, Leif-Jöran, och Thaning, Sten. *Scaling Up an MT Prototype for Industrial Use - Databases and Data Flow*. Institutionen för Lingvistik och Filologi, Uppsala Universitet, 2002.
- Hein, Anna Sågvall och Sjögren, Christian. *Ett Svenskt Stamlexikon För Datamaskinell Morfologisk Analys*. Svenskans beskrivning. Lund University Press, 1990.
- Hellberg, Staffan. *The Morphology of Present-Day Swedish*. Data Linguistica. Almqvist & Wiksell, 1978.
- Hutchins, W. John och Somers, Harold L. *An Introduction to Machine Translation*. Academic Press, 1992.
- Jurafsky, Daniel och Martin, James H. *Speech and Language Processing*. Prentice Hall Series in Artificial Intelligence. Prentice Hall, 2000.
- Norman, Donald A. *The Psychology of Everyday Things*. Basic Books, 1988.
- Schneiderman, Ben. *Designing the User Interface. Strategies for Effective Human-Computer Interaction*. Addison-Wesley, 3 utgåvan, 1998.
- Tiedemann, Jörg. MatsLex - a Multilingual Lexical Database for Machine Translation. I: *Proc. of the 3rd International Conference on Linguistic Resources and Evaluation (LREC'02)*, band VI, ss 1909–1912, 2002.
- Weijnitz, Per, Hein, Anna Sågvall, Forsbom, Eva, Gustavii, Ebba, Pettersson, Eva, och Tiedemann, Jörg. The Machine Translation system MATS - Past, Present and Future. I: *RASMAT'04 (Recent Advances in Scandinavian Machine Translation)*. Institutionen för Lingvistik och filologi, Uppsala Universitet, 2004.