Utveckling av ett svensk-engelskt lexikon för maskinöversättning inom jordbruksdomänen

Ebba Gustavii & Eva Pettersson
{ebbag,evapet}@stp.ling.uu.se

Uppsala Universitet, Institutionen för lingvistik

17 december 2003

Innehåll

Tabeller

Figurer

1 Inledning

I denna rapport beskrivs metoder och resultat för utvecklandet av ett svensk-engelskt lexikon för maskinöversättning inom jordbruksdomänen. Lexikonet har utvecklats som en del i ett samarbete mellan institutionen för lingvistisk (UU) och EC Systran och ska utgöra den lexikala basen i en svensk-engelsk Systran-prototyp.

Lexikonet är korpusbaserat och bygger på två meningslänkade korpusar som ställts samman av EC Systran. Den ena korpusen bygger på EU-dokument från en mängd domäner (däribland jordbruksdomänen) och utgörs av närmare 800 000 löpord. Den andra korpusen är en delmängd av den allmänna EU-korpusen och bygger uteslutande på texter från jordbruksdomänen. Jordbrukskorpusen inbegriper drygt 100 000 löpord.

Det resulterande lexikonet består av tre delar: ett baslexikon, ett flerordslexikon och ett lexikon innehållande svenska lemman som inte kan ges någon översättning på egen hand, utan endast som en del av ett flerordsuttryck. Enligt överenskommelse med EC Systran inbegriper lexikonet en fullständig källspråksbeskrivning. Däremot har vi endast tagit de första stegen mot en morfologisk målspråksbeskrivning.

Upplägget av rapporten återspeglar den övergripande arbetsgången och vi kommer först att beskriva utvecklandet av lexikonet på källspråkssidan, och därefter fokuserar vi på framtagandet av översättningsrelationer.

2 Utveckling av ett svenskt lexikon

I projektets första fas utvecklades källspråkslexikonet, som i första hand täcker alla enkla ordformer i jordbrukskorpusen, och vidare, ett antal flerordsuttryck. Nedan beskriver vi först arbetet med de enkla ordformerna, och därefter arbetet med flerordsuttrycken. Därpå beskrivs lexikonets format och tilldelning av valensdefinitioner till verben.

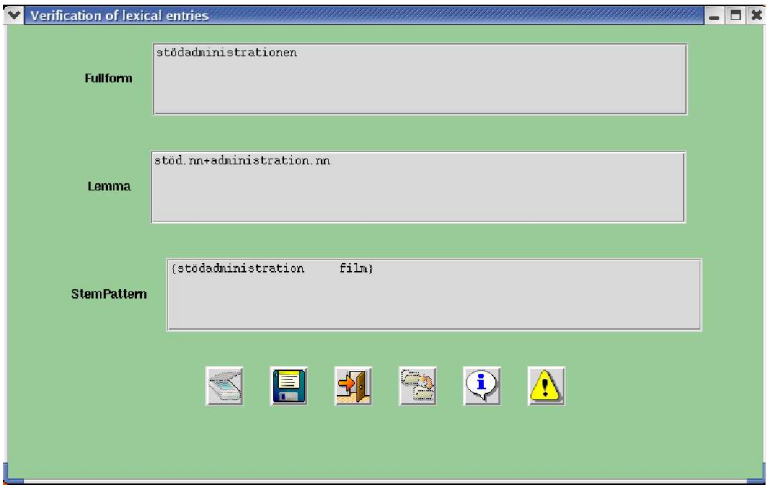
2.1 Enordingar

Ur jordbrukskorpusen har extraherats en lista över de ordformer som förekommer i korpusen. Dessa ordformer slog vi på automatisk väg upp i på institutionen befintliga lexika; den svenska MATS-databasen, Scarries fullformslexikon och Scarries stamlexikon. På så sätt fick vi fram 3 891 lemman med tillhörande tekniska stammar och böjningsmönster.

Resterande ordformer kördes genom UCP:s sammansättningsanalys i kombination med ett program utvecklat på institutionen, som automatiskt väljer den bästa av UCP:s alla sammansättningsanalyser [Åbe03]. De lemman som i och med detta gavs en analys, tilldelades morfosyntaktisk information i enlighet med huvudets morfologiska egenskaper.

Efter att dessa steg gåtts igenom, återstod cirka 1 000 ordformer som inte tilldelats någon morfosyntaktisk information. Dessa ordformer analyserades manuellt. För att underlätta den manuella tilldelningen, utvecklade vi ett gränssnitt. I detta

gränssnitt visas ordformen för användaren och det finns rutor för att skriva in motsvarande lemma, stam och mönsterord. Från gränssnittet kan man sedan välja att spara ingången i baslexikonet eller slänga bort ingången (om det exempelvis rör sig om ett stavfel). Man kan också slå upp mönsterord och se dess olika böjningar. Gränssnittet illustreras i figur 1 nedan.



Figur 1: Gränssnitt för manuell tilldelning av morfosyntaktisk information

2.2 Flerordingar

I arbetet med att inkorporera flerordsuttryck i lexikonet, utgick vi ifrån de automatiskt framtagna flerordsuttryck som finns listade i *Multiword Expressions for Swedish* [Wed98]. De av dessa flerordsuttryck som fanns med i jordbrukskorpusen, lades in i lexikonet. Detta resulterade i 176 flerordsuttryck.

I [Wed98] återfinns endast kontinuerliga flerordsuttryck och därmed inga lexikala enheter bestående av verb och partikel eller verb och reflexivt pronomen. För att få fatt i korpusens partikelverb, gjorde vi en sökning på verb följt av en ordform listad som adverb eller preposition i enordslexikonet. De potentiella partikelverben extraherades och granskades manuellt. Eftersom en partikel inte alltid följer direkt på verbet kan vi därmed ha missat något partikelverb, men förhoppningsvis lokaliserades de mest frekventa. För att underlätta det manuella arbetet sorterades sökresultatet på lemman, vilka listades tillsammans med de fullformer som attesterats. I de fall det var oklart huruvida det verkligen var fråga om ett partikelverb granskade vi ordformerna i korpuskontext. Den manuella genomgången resulterade i en lista om 127 partikelverb. Under arbetet med översättningslexikonet lokaliserades ytterligare 35, varmed det totala antalet partikelverb uppgår till 162. Under den manuella granskningen lokaliserades dessutom 77 konstruktioner med verb och preposition, t.ex. *bestå av*. Även denna lista kom senare att utökas och uppgår totalt till 100.

Motsvarande metod användes för att lokalisera reflexiva verb, varmed en lista över 38 attesterade reflexiva verb ställdes samman.

Kontinuerliga flerordsuttryck har lagts in i baslexikonet, medan diskontinuerliga dito har sparats i ett separat fraslexikon.

2.3 Format

En ingång i baslexikonet består av lemma, teknisk stam, böjningsmönster och optionella extra argument inom hakparenteser:

- kvinna.nn {kvinn flicka}
- man.nn {man man män män}
- angå.vb {angick gick angå gå} [no_PCP no_VBP]

En ingång i flerordslexikonet består av lemma samt extra argument. För flerordsuttrycken anges inte teknisk stam eller böjningsmönster eftersom dessa kan fås genom att de ingående orden slås upp i något av enordslexikonen. Exempel:

- betala.vb_in.ab [attach_PCP attach_PCA attachable head=betala.vb]

2.3.1 Lemma

Varje lemma i lexikonet har någon av följande extensioner:
nn substantiv
av adjektiv
vb verb
pn pronomen
pm egennamn
ab adverb
pp preposition
cn konjunktion
sn subjunktion
nl numeral
al artikel
in interjektion
ie infinitivmärke

Modifierarledet och huvudet åtskiljs av ett '+' för lemman som analyserats som sammansättningar:

- justering.nn+man.nn {justeringsman man justeringsmän män}

För diskontinuerliga flerordsuttryck gäller att ingående lemman åtskiljs av ett understreck:

- arbeta.vb_fram.ab [attach_PCP no_PCA attachable head=arbeta.vb]

Kontinuerliga flerordsuttryck behandlas som ett enda lemma och har därmed en gemensam lemmaextension:

- Arla_Ost_AB.pm {Arla_Ost_AB abf}

2.3.2 Teknisk stam

Den tekniska stammen är definierad som den del av lemmat som är gemensam för alla lemmats böjningsformer. Lemman som genomgår omljud eller andra stora förändringar, har tilldelats mer än en stam:

- man.nn {man man män män}

2.3.3 Böjningsmönster

För varje stam finns ett mönsterord, som anger hur lemmat ska böjas. Dessa mönsterord utgör i huvudsak en kombination av de mönsterord som används i MATS-databasen respektive Scarrie-lexikonet och som utvecklats på institutionen med ledning av *The Morphology of Present-Day Swedish* [Hel78]. En del mönsterord har dock modifierats en aning, medan andra inte används alls. I de fall där vi stött på lemman med böjningsmönster som inte täcks av något av de befintliga mönsterorden, har vi också definierat nya mönsterord med utgångspunkt i *Svenska Akademiens Grammatik* [THA95].

2.3.4 Extra argument

För diskontinuerliga flerordsuttryck anges det syntaktiska huvudet som ett extra argument inom hakparenteser:

- bunden.av_till.pp [head=bunden.av]

Vidare har en del verbingångar tilldelats extra argument av två typer. Den ena typen begränsar antalet tillämpliga former i böjningsparadigmet. För lemmat *uppgå* har vi t ex blockerat perfekt particip och de passiva formerna:

- uppgå.vb {uppgick gick uppgå gå} [no_PCP no_VBP]

Blockeringsmekanismen har använts för perfekt particip (no_PCP), presens particip (no_PCA) samt för passiva former (no_VBP). Vi har valt att arbeta med dessa extra argument för att undvika en explosion av antalet mönsterord.

Den andra typen av extra argument har använts för partikelverb och definierar i vilken grad verbet och partikeln kan skrivas ihop. Argumenten attach_PCA och attach_PCP anger att partikeln konkateneras till verbet i presens particip respektive perfekt particip, t ex: *driva in -> indrivande, indriven, indrivet, indrivna*. Det mer generella argumentet attachable anger att det finns parallella, ihopskrivna former med (i stort sett) bibevarad betydelse för de övriga formerna i paradigmet, t ex:

- jämna.vb_ut.ab [attach_PCP attach_PCA attachable head=jämna.vb]

För verbet jämna.vb_ut.ab kan således följande former genereras: *jämna ut, jämnar ut, jämnade ut, jämnat ut, jämnas ut, jämnats ut, jämnades ut, utjämna, utjämnar, utjämnade, utjämnat, utjämnas, utjämnades, utjämnats, utjämnande* samt *utjämnad*.

Tilldelningen av argument har gjorts manuellt, på intuitiv basis.

2.4 Valenstilldelning

I samband med att lexikonet senare har modifierats för att passa formatet i MATS-databasen, tilldelades baslexikonets verb valensbeteckningar. Tilldelningen gjordes manuellt, med ledning av korpusexempel. För ändamålet utvecklades ett gränssnitt för att enkelt få tillgång till samtliga korpusträffar.

Valensdefinitioner hämtades från MATS-grammatiken vari de betecknas med ord typiska för en viss valensram, såsom va.plundra för verb som tar direkta objekt. Antalet möjliga valenstyper utvidgades något men är alltjämt relativt litet och snarast ett verktyg för grovindelning. Det bör vidare framhållas att vi, på grund av tidsbrist, inte haft möjlighet att efterkontrollera tilldelningarna och att det kan finnas behov av att fingranska resultatet.

3 Utveckling av ett svensk-engelskt lexikon

Det svensk-engelska lexikonet grundar sig på det svenska lexikon som beskrivits ovan. Lämpliga engelska översättningar till de svenska ingångarna har tagits fram semi-automatiskt, genom automatisk ordlänkning och manuell efterbearbetning.

Den automatiska ordlänkningen har genomförts på den allmänna EU-korpus som tillhandahållits av Systran, då jordbrukskorpusens ringa storlek inte skulle givit tillförlitligt länkningsresultat. Under den manuella efterbearbetningen har varje översättningsrelation granskats och den meningslänkade jordbrukskorpusen har använts för att ge ledning om det domänspecifika språkbruket. Vidare har de svenska lexikoningångarna automatiskt slagits upp i den allmänspråkliga delen av MATS-databasen.

3.1 Länkning

Den allmänna EU-korpusen länkades på ordnivå av Jörg Tiedemann med den ordlänkare, Clue Aligner, som han beskriver i [Tie03]. Ordlänkaren kan utnyttja flera typer av källor. Förutom rent statistisk information, involverade länkningsprocessen, i vårt fall, ordklasstaggning (av både den svenska och den engelska korpusen) samt chunkning av den engelska korpusen.

Ordlänkningen resulterade i närmare 118 000 länkar. Dessa efterbearbetades automatiskt. Länkningsingångar som var identiska bortsett från inledande versal eller gemen slogs ihop:

beslut {1X:Decisions 46X:decision 24X:decisions},
Beslut {6X:Decision 2X:Decisions 1X:decision}
=>
beslut {3X:Decisions 6X:Decision 47X:decision 24X:decisions}

Därefter sorterades de länkar ut vars svenska ordform kunde analyseras med det svenska enordslexikonet. Därmed fick vi fatt i de länkar som kunde vara av intresse för utvecklandet av det svensk-engelska lexikonet för jordbruksdomänen. Dessa länkar lemmatiserades på den svenska sidan och i de fall en ordform var tvetydig skapades en länkningsuppsättning för varje lemmatolkning, t ex:

bra {4X:good 1X:well} => bra.av {4X:good 1X:well}, bra.ab {4X:good 1X:well}

Avslutningsvis sorterades länkningsalternativen efter frekvens.

3.2 Upprättande av översättningsrelationer

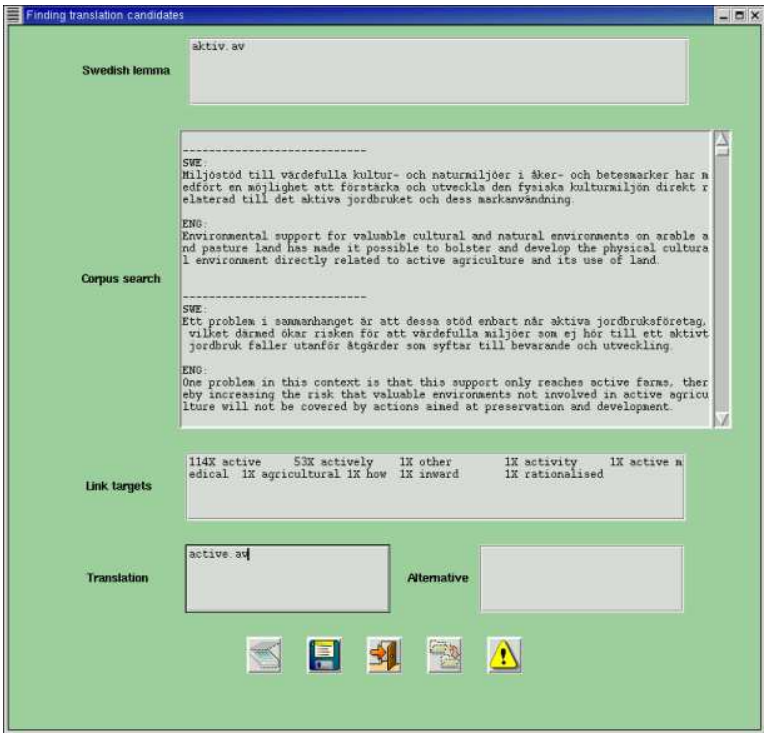
De slutgiltiga översättningsrelationerna definierades manuellt, med ledning av de automatiskt framtagna resurserna samt de översättningar som fanns tillgängliga i den allmänspråkliga delen av MATS-databasen. För att underlätta arbetet utvecklade vi ett gränssnitt i vilket all tillgänglig information presenteras. Detta gränssnitt illustreras i figur 2 nedan.

I det översta fältet presenteras källspråkslemmat som skall tilldelas en översättning. I fältet därunder visas alla de meningar, i jordbrukskorpusen, där lemmat förekommer. För varje källspråksmening presenteras även den länkade målspråksmeningen. Sökningen görs på alla böjningsmönster som kan genereras utifrån lemmat och går via en indexerad fil. I det tredje fältet presenteras resultatet av ordlänkningen med länkningsalternativen i fallande frekvensordning. Därefter presenteras två fält. I den vänstra förväntas användaren fylla i den mest lämpliga översättningen (i lemmaformat). Om det finns en översättning av lemmat i MATS-databasen visas denna i rutan, men kan ändras av användaren. I den högra rutan kan användaren skriva in alternativa översättningar.

Vi har strävat efter att återspegla språkbruket i jordbrukskorpusen och har endast i undantagsfall definierat översättningsrelationer som skiljer sig från de som finns attesterade däri. Därmed har vi inte arbetat med att eliminera inkonsekvenser i språkbruket, såsom inkonsekvent översättning av huvud i sammansättningar, t ex:

- *företagsstöd -> business support*
- *småföretagarstöd -> small-business aid.*

På EC Systrans begäran, har vi vidare strävat efter att inte tilldela ett svenskt lemma mer än en översättning. Ofta har vi dock stött på översättningar som varit beroende av den närliggande kontexten, och om möjligt, har vi i sådana fall definierat



Figur 2: Gränssnitt för manuell tilldelning av översättningsrelationer

flerordsuttryck. Vid ett fåtal tillfällen har översättningsvariationen berott på homografi som inte kan tas om hand i termer av flerordsuttryck. Då betydelserna har varit tydligt åtskilda har vi definierat flera översättningsalternativ, t ex:

- art.nn {art film} => species.nn/nature.nn

En ytterligare strävan har varit att definiera översättningar med samma syntaktiska funktion som källspråkslemmat. Vid ett tjugotal tillfällen har vi dock inte funnit någon tillfredsställande översättning inom samma ordklass och tvingats tillåta ett byte av ordklass. Detta har framför allt gällt adjektiv som översatts till substantiv, t ex:

- procentuell.av {procentuell polig} => percentage.nn

I de fall inget tillfredsställande översättningsförslag har kunnat extraheras ur korpusmaterialet, på grund av alltför kreativa eller kontextberoende översättningar, har vi förlitat oss på vår egen lingvistiska kompetens samt *Stora Engelska Ordboken* [Nor88].

3.3 Format

De engelska lemmana har samma format som de svenska, dvs de listas med grundformen följt av en lemmaextension. Uppsättningen möjliga lemmaextensioner är

densamma med undantag för v-ing, som har lagts till för att beteckna verb i progressiv form. Vi har valt att lägga till denna kategori, trots att det traditionellt sett snarare är fråga om en böjnings- än en ordklasskategori, eftersom den progressiva formens funktion ofta är adjektivisk eller substantivisk snarare än verbal.

Källspråks- och målspråksinformationen åtskiljs med '=>'. Om det finns flera översättningsalternativ separeras dessa med '/', t ex:

- affär.nn {affär film} => shop.nn/business.nn

Om översättningen utgörs av ett flerordsuttryck med ett syntaktiskt huvud, anges detta inom hakparenteser:

- ambition.nn+nivå.nn {ambitionsnivå nivå} => level_of_ambition.nn [head=level.nn]
- lägga.vb_ned.ab [attach_PCP no_PCA head=lägga.vb] => close.vb_down.ab [head=close.vb]/spend.vb

Precis som på den svenska sidan behandlas de kontinuerliga flerordsuttrycken som ett enda lemma och ges en gemensam lemmaextension, medan samtliga ingående lemman i diskontinuerliga uttryck tilldelas separata extensioner.

3.4 Tilldelning av morfosyntaktisk information på den engelska sidan

I samband med att lexikonet konverterades till formatet i MATS-databasen, tilldelades ingångarna i baslexikonet morfosyntaktisk information även på den engelska sidan. De mönsterordsdefinitioner som har använts är desamma som i MATS-databasen, med en del smärre modifieringar.

Tilldelningen av mönsterord har skett automatiskt. I första hand hämtades böjningsinformation från MATS-databasen. De lemman som inte återfanns där tilldelades böjningsmönster enligt enkla heuristiska strategier. Resultatet har inte utvärderats och är att betrakta som ett första automatiskt steg som bör följas upp med en manuell bearbetning. För vissa ordklasser, såsom pronomen, formulerades ingen heuristik, och lemman tillhörande dessa klasser har sparats undan i en särskild fil och bör tas om hand manuellt.

4 Ord på -ande/-ende

Svenska ord med avledningssuffixen -ande eller -ende, kan i regel fungera både som substantiv och som adjektiv. För att få fram statistik över vilken som är den vanligaste tolkningen för specifika ord av denna typ, taggade vi den allmänna korpusen med hjälp av TNT-taggaren. (Se [Bra00] för en utförligare beskrivning av TNT-taggaren)

Ur den taggade korpusen, extraherade vi sedan de ord som slutade på -ande eller -ende. För dessa ord förenklades taggarna på automatisk väg till att endast ange ordklass. Identiska ord slogs samman och frekvensberäkningar för ordklasstillhörighet

utfördes. Resultatet blev en lista över ordformer med de olika ordklassfrekvenserna inom krullparenteser:

- utnyttjande {30 NN 2 ADJ}

5 Resultat

Det slutliga lexikonet består av tre delar; ett baslexikon, ett flerordslexikon och ett lexikon innehållande svenska lemman som inte kan ges någon översättning när de förekommer självständigt, utan bara som del av ett flerordsuttryck.

I baslexikonet utgörs större andelen lemman av substantiv, följt av verb och adjektiv, enligt tabell 1 nedan.

Ordklass	**Antal ingångar**
Substantiv	3 450
Enkla adjektiv	703
Verb	857
Enkla pronomen	54
Egennamn	458
Enkla adverb	333
Enkla prepositioner	51
Konjunktioner	12
Enkla subjunktioner	17
Numeraler	7
Artiklar	2
Interjektioner	2
Infinitivmärket	1
Flerordsprepositioner	83
Flerordsadverb	178
Flerordspronomen	8
Flerordssubjunktioner	28
Flerordsadjektiv	7
Andra kontinuerliga flerordsuttryck	5
Kopulativa sammansättningar	2
Total	**6 258**

Tabell 1: Ordklassfördelning i baslexikonet

Ingångarna i flerordslexikonet är distribuerade enligt tabell 2.

I lexikonet innehållande självständigt oöversättbara svenska lemman, utgörs majoriteten av ingångarna av adverb (som fungerar som verbpartiklar). Fördelningen av ingångarna illustreras i tabell 3.

Konstituent	**Exempel**	**Antal ingångar**
Partikelverb	*arbeta fram*	162
Verb med preposition	*anmärka på*	100
Reflexiva verb	*etablera sig*	38
Komplexa verbkonstruktioner	*dra inför domstol*	147
Substantiv med prepostion	*förteckning över*	19
Adjektiv med preposition	*bunden till*	4
Modifierade substantiv	*sluten omröstning*	31
Modifierade adjektiv	*ytterst ansvarig*	3
Diskontinuerliga konjunktioner	*både och*	4
Totalt		**508**

Tabell 2: Fördelning av konstituenter i flerordslexikonet

Ordklass	**Antal ingångar**
Substantiv	5
Verb	6
Adverb	31
Konjunktioner	1
Totalt	**43**

Tabell 3: Ordklassfördelning i lexikonet med oöversättbara svenska lemman

6 Resurser

De resurser som framtagits under arbetet med lexikonutvecklingen, finns samlade i katalogen /home/stp98/evapet/Systran/Resurser/. Filerna listas nedan.

6.1 Korpora

Meningslänkad svensk-engelsk jordbrukskorpus:	jordbruksKorpus_sven.txt
Meningslänkad svensk-engelsk allmänkorpus:	allmänKorpus_sven.txt
Svensk allmänkorpus:	allmänKorpus_sv.txt
Taggad version av den svenska allmänkorpusen:	taggadAllmänKorpus_sv.txt

6.2 Lexika

Svensk-engelskt baslexikon:	basLexikon.txt
Svensk-engelskt flerordslexikon:	flerordsLexikon.txt
Lexikon innehållande oöversättbara svenska lemman:	oöversättbaraLemman.txt
Svenska fullformer genererade ur baslexikonet:	fullformsLexikon.txt
Engelskt lexikon med mönsterordsdefinition:	engLexikon.txt

6.3 Länkningsresultat

Länkningsresultat för svenska/engelska:	länkningsresultat_orig_sven.txt
Länkningsresultat för danska/engelska:	länkningsresultat_orig_daen.txt
Efterbearbetat länkningsresultat för svenska/engelska:	länkningsresultat_bearbetat_sven.txt
Efterbearbetat länkningsresultat för danska/engelska:	länkningsresultat_bearbetat_daen.txt

6.4 Övrigt

Förteckning över ord på -ande/-ende:	listaÖverAndeEnde.txt
Mönsterordsfil:	mönsterordsDefinitioner.txt
Förteckning över verb med valensinformation:	listaÖverValenser.txt
Valensdefinitionsfil:	valensDefinitioner.txt

Referenser

[Åbe03] Stina Åberg. Datoriserad analys av sammansättningar i teknisk text. Master's thesis, Uppsala universitet, 2003.

[Bra00] Thorsten Brants. Tnt - a statistical part-of-speech tagger. Technical report, Saarland University, Computational Linguistics, 2000.

[Hel78] Staffan Hellberg. *The Morphology of Present-Day Swedish*. Almqvist & Wiksell, 1978.

[Nor88] Norstedts, editor. *Stora Engelska Ordboken*. Norstedts, 1988.

[THA95] Ulf Teleman, Staffan Hellberg, and Erik Andersson. *Svenska Akademiens Grammatik*. NorstedtsOrdbok, 1995.

[Tie03] Jörg Tiedemann. Combining clues for word alignment. In *Proceedings of the 10th Conference of the EACL*, 2003.

[Wed98] Olga Wedbjer Rambell. Multiword expressions for swedish. In Anna Sågvall Hein, editor, *Working Papers in Computational Linguistics & Language Engineering*, volume 8. Uppsala Universitet, Institutionen för Lingvistik, 1998.