Stockolms universitet
Institutionen för lingvistik
Magisteruppsats i datorlingvistik, ht 2002

Prepositioner
och
automatisk textgranskning för andraspråksinlärare

Jens Eeg-Olofsson

Sammanfattning

Uppsatsen presenterar ett förslag till komplettering av ett befintligt system – Granska – för datorstödd språkgranskning av svensk text. Kompletteringen avser granskning av prepositionsanvändning. Syftet är att anpassa systemet för användare med annat modersmål än svenska. Uppsatsen redovisar också den analys av skrivfel i autentisk text som ligger till grund för kompletteringsförslaget. Prepositionsrelaterade skrivfel har extraherats ur en textkorpus som innehåller uppsatser i svenska, författade av andraspråksinlärare vid olika stadier i inlärningsprocessen. Klassificering av iakttagna konstruktioner har resulterat i en uppsättning nya så kallade granskningsregler som identifierar problem med prepositionsanvändning och som är tänkta att implementeras i en ny version av textgranskningsverktyget. I uppsatsen utvärderas och kommenteras slutligen de nya reglernas detektion, diagnos och korrektion av prepositionsrelaterade skrivfel hos andraspråksinlärare.

Handledare: Ola Knutsson, Nada, KTH
Examinator: Lars Borin, Institutionen för lingvistik, Stockholms universitet

Innehåll

1 Inledning och forskningsbakgrund

1.1 Granska och CrossCheck

På Institutionen för numerisk analys och datalogi vid Kungliga tekniska högskolan i Stockholm, arbetar man med ett system för automatisk stavnings- och grammatikkontroll. Detta system har sedan 1998 utvecklats under namnet Granska. På förekommen anledning startade man i januari 2002 ett nytt forskningsprojekt med syftet att utveckla Granska så att verktyget bättre kunde möta behoven hos användare med annat modersmål än svenska.

På Nada drivs för närvarande två parallella forskningsprojekt med fokus på andraspråksinlärare. Det ena projektet ”Språkliga datorstöd och andraspråksinlärning” har sin utgångspunkt i människa-datorinteraktion och pedagogiska aspekter på datorstött skrivande. Projektet finansieras av Vetenskapsrådet. Mitt uppsatsarbete hör hemma inom det överlappande projektet ”CrossCheck – svensk grammatikkontroll för andraspråksskribenter”, som finanseras av Vinnova inom Språkteknologiprogrammet och som är ett samarbete mellan Nada och Institutionen för lingvistik vid Stockholms universitet.

CrossCheck är egentligen arbetsnamnet på den version av verktyget Granska som är målet för forskningsprojektet, nämligen ett verktyg för automatisk grammatikkontroll av svensk text, särskilt utformat för skribenter som lär sig svenska som andraspråk. CrossCheck-projektet är alltså särskilt inriktat på att utveckla själva verktyget och dess funktionalitet, och är utpräglat språkteknologiskt. Det nya verktyget ska vara beväpnat med en statistisk granskningsmodul som är tänkt att arbeta parallellt med de lingvistiskt formulerade reglerna som ligger till grund för feldetektionen (avsnitt 1.3). Dessutom avser man bygga upp ett textkorpusmaterial sorterat i två huvudgrupper – andraspråksinlärarproducerad respektive standardspråklig text. Korpusarna ska bland annat tjäna som underlag för utvärdering av verktyget och som träningsdata för den statistiska granskningen. Korpusen ska också erbjuda material för typologiskt-kontrastiva eller komparativa studier, därav projektnamnet.

1.2 Syfte

Uppsatsen ansluter till målsättningen i CrossCheck-projektet, nämligen att utveckla Granska mot bättre funktionalitet för andraspråksinlärare.

Granska saknar i princip granskningsregeler för prepositionsproblem. Mer specifikt är syftet med uppsatsen därför att komplettera Granskasystemet med regler för detektion och korrektion av prepositionsrelaterade problem i andraspråksinlärartext.

Det finns ingen ambition att utreda huruvida prepositionsrelaterade skrivfel skulle vara relativt överrepresenterade hos andraspråksinlärare. Ingen ambition finns heller att kartlägga eventuella skillnader i feltypsmönster med avseende på modersmålsgrupp. Strukturingen av undersökningsdata lämnar emellertid öppet för en analys av det senare slaget.

1.3 Teknisk beskrivning av Granskasystemet

Min beskrivning av Granskasystemet har den huvudsakliga funktionen att ge nödvändig information för fortsatt läsning av uppsatsen. Uttömmande beskrivningar av systemet finns i första hand i Ola Knutssons licentiatavhandling (Knutsson 2001) samt i artiklar och forskningsrapporter från systemets konstruktörer och utvecklare (Domeij et al 1994; Domeij et al 1998; Kann et al 1998; Domeij & Knutsson 1999; Carlberger et al 2000; Knutsson 2002; Knutsson et al 2002; Bigert & Knutsson 2002). Mitton (1996) erbjuder en ytterst läsvärd presentation av system för automatisk textkontroll i allmänhet.

1.3.1 Ordklasser, frasklasser och felklasser

Granska gör först en *tokenisering* av texten som ska granskas – teckensekvenserna separeras för att man ska kunna skilja analysenheterna åt. Sedan ordklassanalyseras – *taggas* – texten. Granskas *tagger* är en statistisk ordklasstagger av markovmodell, en så kallad *Hidden Markov Model*, där varje ord tilldelas alla tänkbara analyser (ordklass*taggar*) genom matchning mot ett lexikon med all nödvändig information.[1] Granskas referenslexikon är konstruerat utifrån en ordklasstaggad standardspråkskorpus – the Stockholm-Umeå Corpus (SUC), som presenteras i avsnitt 2.2 – och har kompletterats med ordformer ur 1986 års version av Svenska Akademiens ordlista. I markovmodellen beräknas för varje sekvens om tre ord – ett *fönster* eller *ordtrigram* – den mest sannolika motsvarande taggsekvensen. För varje steg i denna procedur tas det närmast följande ordet med i beräkningen, samtidigt som fönstret lämnar bakom sig det ord som tidigare stod längst till vänster. För varje nytt steg har taggern alltså redan gjort sin statistiska analys för de två första orden i fönstret. Kalkylen *uppdateras* nu tillsammans med den nya sista enheten i fönstret. På det här viset tuggar sig taggern igenom hela texten. Varje ord, även ord som avviker i kontexten eller inte återfinns i lexikonet, tilldelas av markovtaggern alltid en och endast en tagg. Efter ordklassanalysen matchas den taggade texten mot villkorsregler som representerar typiska felsekvenser, regler som alltså fångar upp sådana sekvenser i indata. Dessa regler kallas för *granskningsregler* och presenteras mer utförligt nedan, men först några ord om vad som bestämmer vad granskningsreglerna ska innehålla.

Mängden av alla möjliga skrivfel är naturligtvis oändlig (liksom mängden av alla möjliga strängar i naturligt språk) och därför krävs till att börja med ett ganska omfattande arbete för att identifiera feltypsmönster för granskningsreglerna. Här gäller det att kombinera lingvistisk kunskap och intuition med emprisk examination av textkorpusar. Man försöker klassificera felkonstruktioner i oredigerad text, till exempel gymnasieuppsatser, samtidigt som distributioner för olika konstruktionstyper undersöks i standardspråkskorpusar. Om en mycket vanlig typ av skrivfel råkar sammanfalla med en mycket vanlig konstruktionstyp, blir feltypen särskilt intressant för regelskrivning. En sådan feltyp som det i Granskasystemet lagts ner extra mycket möda på att fånga upp, är kongruensfel i nominalfraser, som **en viktigt mening* eller **det arga tanten*, en annan är särskrivningar som **radio kanal* och **alkohol förtäring*.[2]

[1] Egentligen är det inte *ord* utan *tokens* som taggern arbetar med, men jag kommer av läsbarhetsskäl även fortsättningsvis att referera till analysenheterna som *ord*.

[2] I grammatiklitteraturen används vanligen framförställd asterisk för ogrammatiskhet och framförställt frågetecken för vacklande grammatiskhet. Båda används i uppsatsen, men aldrig på exempel försedda med index (som alltid är autentiska exempel).

De flesta program för automatisk grammatikkontroll av det här slaget kräver långtgående regelgeneralisering, alltså samma regel ska kunna detektera många olika (men ändå lika) fel, detta för att programmet ska kunna hitta fel och ge ändringsförslag inom rimliga tidsmarginaler. Granskas regelexekvering är dock osedvanligt snabb och inte särdeles känslig för stickspår och mycket specifika matchningar.

1.3.2 Regelspråk och regelstruktur

Granska är en forskningsprototyp som är framtagen på en institution med statligt utbildningsuppdrag. Därför är det naturligt att forskningsprojektet är föremål för kontinuerlig genomströmning av studenter som lämnar sina bidrag till utvecklingen av Granska genom examensarbeten och liknande. Detta ställer ganska höga krav på Granska vad gäller överskådlighet, flexibilitet, läsbarhet och effektivitet. För att möta behoven har utvecklarna konstruerat ett eget regelspråk för Granska. Reglerna skrivs huvudsakligen i *objektorienterad* notation. Detta lämpar sig särkilt väl för grammatisk analys, eftersom analysenheterna behandlas som objekt med särdragsattribut som *ordklass*, *numerus*, *species* och *lemmaform*. Med angivelse av motsvarande värden på attributen kan programmet identifiera en specifik ordform, till exempel *substantiv*, *pluralis*, *bestämd form* och ”*spel*”. Just denna beskrivning av objektet motsvarar ordet *spelen* men också *spelens*, eftersom jag utelämnat *kasus* i särdragsbeskrivningen.

Granskningsreglerna anger alltså *kontextvillkor* som idealt matchar vad som klassats som typiska felkonstruktioner, till exempel **Jag mötte ett björn* eller **Vi använder ett dator program som rättar stavfel*. Till sin hjälp har granskningsreglerna så kallade ***hjälpregler*** som tar hand om analysen av till exempel komplexa nominalfraser, så att detektionen av **Jag mötte ett respektingivande men snäll björn med blåbär på nosen* inte blir mer komplicerad än detektionen av **Jag mötte ett björn*. Hjälpregeln matchar alltså i typfallet korrekta sekvenser, och anropas i själva granskningsregeln. Det är bara de sekvenser som matchas i granskningsreglerna som blir föremål för en grammatisk analys utöver den som redan gjorts av taggern, och även denna omständighet bidrar till hög exekveringshastighet.

Granskningsreglerna består av ett vänster- och ett högerled. I vänsterledet anger man matchningsvillkoren, till exempel inkongruensvillkor för matchning av sekvensen **ett björn*. I högerledet anger man i olika *fält* vilka operationer regeln ska utföra. Här kan man manipulera de tokens som matchats i vänsterledet. Vänsterledets matchningsvariabler och deras särdragsbeskrivning, är objekt som de metoder och funktioner som används i de olika fälten opererar på. Tabell 1 visar en lätt förenklad form av granskningsregel som detekterar såväl **ett björn* som **ett respektingivande men snäll björn med blåbär på nosen*.

Tabell 1. Exempel på granskningsregel.

Granskningsregel	**Kommentarer**

`exempelregel@kongruensregler`
`{`
`X(wordcl=dt),`
`(NPall/Y)(gender!=X.gender)`
`-->`
`mark(X)`
`corr(X.form(gender:=Y.gender))`
`action(scrutinizing)`
`}`
	En granskningsregel måste ha ett `namn@` följt av *kategoritillhörighet*.
	
	I vänsterledet anges en sekvens av matchningsvariabler, samt värden på de grammatiska särdrag som ska bindas till var variabel. Till exempel så ska värdet på särdraget *genus* för variabel `Y` **inte** vara samma som för variabel `X` (utropstecknet är en negationsoperator). `NPall` är namnet på en hjälpregel för nominalfrasigenkänning.
	
	I högerledet markeras objektet som är bundet till variabeln `x` (nämligen ordet *ett*). Samma objekt korrigeras med avseende på genus och får ett nytt värde på detta särdrag så att `X` delar detta särdragsvärde med `Y` (alltså nominalfrasen).
	
	Det sista fältet är obligatoriskt och metoden `scrutinizing` utför granskningsuppdraget.

Tråkigt nog detekterar regeln också *Jag såg ett björnar emellan ovanligt beteende*, som korrigeras till **Jag såg en björnar emellan ovanligt beteende*, liksom **Vi använder ett dator program*, som korrigeras till **Vi använder en dator program*.

Konstruktioner med prepositionen (eller snarare *postpositionen*) *emellan* låter ofta lite högtravande men är inte så ovanliga, och det finns ett sätt att undvika falska alarm på mindre typiska men korrekta konstruktioner som matchas av granskningsreglerna. Man skriver en så kallad ***accepterande regel*** med samma struktur som en granskningsregel men med en annat obligatoriskt fält. Den accepterande regeln detekterar den korrekta sekvensen och blockerar matchning av sekvensen i kongruensreglerna, helt enkelt genom att lyfta det som matchats i vänsterledet förbi kategorin kongruensregler. De accepterande reglerna tillämpas därför alltid före granskningsreglerna.

Granska har naturligtvis en regel i kategorin `sär` som detekterar **dator program* och korrigerar sekvensen till *datorprogram*. Men faktum är att **ett dator program* också detekteras av en kongruensregel, liknande exempelregeln ovan, som korrigerar sekvensen till **en dator program*. Hela frasen får alltså två detektioner, vilket kanske inte är så dumt, eftersom det i det här fallet är förhållandevis lätt för användaren att själv bedöma vari felet bestod.

Det finns faktiskt också hjälpregler för detektion och automatisk korrektion av feltaggningar. Det står utom allt tvivel att taggern gör systematiska feltaggningar, den analyserar till exempel artikel som pronomen i vissa kontexter. Man har identifierat vilken typ av sekvens som orsakar

misstaget och skrivit en hjälpregel som detekterar sekvensen och modifierar taggningen, och granskningsreglerna får den korrekta taggningen som indata istället för den felaktiga.

1.3.3 Teknik och grammatik i förening

Rättstavningsmodulen i Granska har ytterligare några år på nacken. Den kallas Stava och är integrerad i granskningsreglerna. Stavas inre byggnad får illustrera Granskasystemets fascinerande sammansmältning av avancerad programmeringsteknik och lingvistisk kunskap. Stavas ordlista är kodad på ett sinnrikt sätt för att minska det elektroniska lagringsutrymmet, optimera programmets exekveringshastighet och säkra ordlistan mot kommersiellt utnyttjande. I princip går kodningen till så att varje alfabetiskt tecken tilldelas ett värde och för varje ord utförs på ordets bokstavsvärden en serie räkneoperationer. Kodningen resulterar i en motsvarande serie heltal som sorteras in i en så kallad *hashtabell*. Distributionen i hashtabellen blir i princip unik för varje ord och det är denna distribution som är koden, det vill säga kombinationen av de platser som ockuperats i tabellen. Man kan säga att ordet får en egen profil. Varje ord i den text som sedan ska stavningsgranskas, kodas på samma sätt och om resultatet matchar någon profil i hashtabellen så är ordet rättstavat. Det hela fungerar ungefär som ett kolvlås som passar till en viss uppsättning nycklar. Varje ord som ska granskas konverteras till en nyckel. Passar nyckeln accepteras ordet. Passar den inte är det ett stavfel och ordet skickas vidare för korrektion.

”Vanliga” rättstavningsprogram för svenska tvingas leva med det faktum att mängden av alla möjliga sammansättningar i svenskan är oändlig och sammansättningar ger därför väldigt lätt upphov till falska alarm. Den lingvistiska dimensionen av Stava inbegriper bland annat en modul för sammansättningar som gör att programmet utan vidare accepterar ord som *honungsmelonhalva* eller varför inte *andraspråksinlärarforskningskontext*. Programmet behöver alltså inte ha med sammansättningar i ordlistan. Om man sammansätter fel, till exempel glömmer ett foge-s, så reagerar programmet. Det betyder också att det finns lingvistisk kunskap inbyggd som släpper igenom *melonhalva* utan s-fog, men inte **melonshalva*. Dessutom är Stava utrustad med ett paket suffixregler som matchar avledningar och böjningar som kan bildas till de ord som finns i ordlistan. Alla dessa programpaket samverkar vid rättstavningskontrollen.

Korrektionsmodulen i Stava utnyttjar det faktum att mängden av alla möjliga sekvenser om fyra bokstäver i svenskan i allra högsta grad är ändlig, men fungerar annars i princip som många andra program för stavningskontroll (avsnitt 1.4).

1.4 Automatisk stavnings- och grammatikkontroll

Datorprogram för automatisk stavningskontroll började utvecklas för engelska redan på femtiotalet. Gemensamt för de flesta stavningskontrollprogram är förekomsten av en mer eller mindre omfattande ordlista mot vilken användarens ordsträngar matchas. Om användaren skrivit ett ord som inte återfinns i listan, till exempel strängen ***ordwt***, så reagerar programmet och pekar på misstaget på ett eller annat sätt. Det finns också rättstavningsprogram som inte alls använder sig av ordlistor, utan som istället delar upp texten i teckensekvenser om till exempel tre tecken (bokstäver), så kallade *trigram*, som programmet sedan gör en frekvenskalkyl på. I fallet med den missformade strängen ***ordwt***, så kommer programmet troligen inte att notera mer än en enda förekomst av sekvensen ***rdw***, och ännu mindre sannolikt hitta två trigram med utseendet ***dwt***. Strängen ***ordwt*** är högst osannolik och markeras som stavfel. Detta om man så vill helt olingvistiska angreppssätt har den fördelen att det är språkoberoende.

De flesta stavningsprogram erbjuder dessutom rättningsförslag. Generering och rankning av rättningsförslag är en betydligt mer komplicerad procedur. Vanligen bygger programmen även här på samma principer, nämligen *stränglikhet* och *ordfrekvens*.

Den missformade ordsträngen restaureras i många rättstavningsprogram i termer av *redigeringsavstånd* eller *levenshteinavstånd* (Kukich 1992; Mitton 1996; Prütz 2002). Programmet söker i princip efter det minst kostsamma sättet att modifiera det felstavade ordet så att det blir identiskt med ett ord i programmets ordlista. För strängen ***ordwt*** räcker det till exempel med att byta ut en bokstav för att skapa ordet ***ordet***. Kostnaden för operationen, eller redigeringsavståndet, är 1. Kostnaden för att komma till ordet ***orkat*** är större – två bokstäver måste bytas ut, och ***orkat*** får lägre prioritet som rättningsförslag. Ordet ***oddset*** kräver tre operationer (två utbyten och en insättning) och får ännu lägre prioritet. Men redigeringsavståndet till ordet ***ordat*** är också bara 1. Vilket ord ska programmet ge som rättningsförslag? Nu viktas för varje rättningsalternativ värdet på redigeringsavståndet med rättningsalternativets position i en frekvensordlista. Naturligtvis måste man sätta en tröskel för hur stor restaureringskostnaden ska få lov att vara (redan vid kostnaden 5 innehåller antalet rättningsförslag till strängen ***ordwt*** till att börja med alla fembokstavsord i listan). Vi kan utgå ifrån att programmet kommer att välja ***ordet*** som första rättningsförslag, kanske följt av ***orkat*** och ***ordat***.

Men om användaren nu istället skrev ***ordat*** när avsikten var att skriva ***ordet***, då upptäcks inte misstaget av stavningsprogrammet, eftersom det felstavade ordet redan finns i listan. Dessa så kallade *real word errors* är mycket vanliga, och väldokumenterade (Kukich 1992; Mitton 1996; Golding & Schabes 1996). Mittons undersökning av nära tusen engelska gymnasieuppsatser visade att hela 40 procent av alla stavfel resulterade i ett annat existerande ord. Jag använder hädanefter beteckningen *RWE* för *real word error*.

Mittons resultat jämte andra liknande forskningsrapporter och utvärderingar av befintlig programvara har provocerat fram en ny generation program för stavningskontroll, som på olika sätt tar hänsyn till kontexten. Dessa program brukar kallas för *grammar checkers* eftersom de också ska kunna ta hand om problem med grammatisk struktur. Att ta hänsyn till kontexten innebär i alla händelser en analys av språkstruktur över ordnivå.

I botten på de flesta sådana grammatikkontrollprogram ligger en modul för automatisk ordklassanalys, inte sällan en markovmodell som i Granska. Redan ordklasstilldelningen är i praktiken en kontextanalys – det finns ingen ordklasstagger som i taggningsproceduren inte tar hänsyn till den närmaste kontexten vid taggning av ett ord. Den första stora grammatikkontrollen, *Critique*, som togs fram på IBM och som var föregångaren till Microsofts nuvarande grammatikkontroll för engelska, gjorde rentav en fullständig syntaktisk satsanalys (några sådana verktyg finns fortfarande inte för svenska, bland annat på grund av brist på ett tillräckligt stort elektronisk lexikon).

Låt oss nu säga att vårt RWE begåtts i följande kontext:

> ***Det talade*** *ordat* ***måste värnas!***

En ordklasstagger av den typ som används i Granskasystemet skulle kunna returnera följande analys

> *Det* <pronomen> *talade* <verb i preteritum> *ordat* <verb i supinum> *måste* <modalt verb i presens> *värnas* <infinit verb i passivum> *!* <interpunktionstecken>

Stavfelet lurar taggern på analysen av det ord som omedelbart föregår ***ordat***, och som är en particip, vilket i sin tur lurar taggern på analysen av artikeln som istället blir demonstrativt pronomen. En hypotetisk verbkedjeregel som matchar fyra verbtaggar i serie skulle visserligen korrekt flagga för fel men programmet kan fortfarande inte upptäcka stavfelet, även om användaren kanske gör det själv om hon/han blir uppmärksammad på kontexten genom programmets kommentar att ”det är mycket ovanligt med fyra verbformer i rad” eller något liknande.

RWE:s utgör alltjämt ett svårbemästrat problem och en spännande utmaning för forskning kring automatisk textgranskning.

En annan typ av fel som får likartade konsekvenser, men som inte är tangentbordsmissar utan snarare kunskapsfel, är när ett specifikt ord påtagligt ofta förväxlas med ett annat antingen semantiskt (*talträngd/tystlåten*), idiomatiskt (*till stor grad/till stor del*) eller fonologiskt (*ände/ende*) närliggande ord. Denna kategori skrivfel har kommit att få ökad uppmärksamhet, inte minst med analysen av texter producerade av andraspråksinlärare. Ett sätt att angripa denna typ av problem i automatisk textgranskning, är att laborera med så kallade *confusion sets* (Golding & Schabes 1996), eller *förväxlingsmängder* som jag kallar min tillämpning av metoden (avsnitt 3.4.4, 4.5).

1.5 Andraspråksinlärning och automatisk språkgranskning

I en datorlingvistisk uppsats som denna finns inte mycket utrymme för analys av själva fenomenet andraspråksinlärning. Forskning kring andraspråksinlärning har under flera decennier fokuserat *processen* vid inlärning av ett andraspråk, utifrån såväl lingvistiskt-grammatiska som neurologiska, psykologiska, sociologiska och inte minst pedagogiska aspekter. Flera konkurrerande och samtidigt mycket inflytelserika teorier har dessutom löpt parallellt genom forskningsområdet. Jag kommer inte att redogöra för denna omfattande teoribildning, utan hänvisar istället till lämpliga introduktioner och forskningsöversikter (Viberg 1987; Ellis 1997; Mitchell & Myles 1998; Berggren & Tenfjord 1999).

1.5.1 Grundläggande definitioner

Ett ***andraspråk*** är ett språk som lärs in efter det att modersmålet är befäst. Man brukar räkna med att huvuddragen i modersmålsstrukturen är etablerad vid tre års ålder. Inget språk som lärs in före tre års ålder är alltså ett andraspråk, och alla språk som lärs in efter modersmålet är andraspråk. ***Målspråket*** är det andraspråk som ***inläraren*** i en given situation strävar efter att tillgodogöra sig. I uppsatsen talar jag uteslutande om svenska som målspråk.

I Skandinavien har vi en tradition att göra skillnad mellan ***andraspråksinlärning***, som står för inlärning av andraspråk i målspråkets språkmiljö, och ***främmandespråksinlärning*** som innebär att andraspråket lärs in någon annanstans, där målspråket inte är standardspråket. Dessutom gör man ibland skillnad på ***formell inlärning***, som är undervisningsstyrd, och ***informell inlärning***, som så att säga sker utan skola.

Inlärning av ett andraspråk i målspråksmiljö har alltid en informell dimension, lärarledda studier av målspråket är då ett formellt komplement. Textmaterialet som ligger till grund för min studie har producerats i en sådan inlärningssituation. En person med annat modersmål än svenska, som lär sig svenska på någon utbildningsinstitution i Sverige, studerar ett ämne som etablerats med beteckningen ***Svenska som andraspråk***.

I språkvetenskapliga sammanhang används ofta förkortningarna **L1** för modersmål eller ***förstaspråk*** och **L2** för målspråk eller andraspråk.

1.5.2 Interferens

Ett mycket vanligt och väldokumenterat fenomen vid andraspråksinlärning är så kallad *interferens*. Detta betyder att språkliga strukturer, i typfallet från inlärarens modersmål, inverkar eller interfererar på målspråket. Inläraren klär så att säga målspråket i modersmålets dräkt. Ibland lyckas en sådan överföring och inläraren producerar ett yttrande enligt målspråkets standardnorm, därför att målspråket delade modersmålets struktur. Överföringen kan också resultera i ogrammatiskhet. Därför talar man ibland om *positiv* respektive *negativ överföring* eller *transfer*. Jag kommer genomgående att använda termen *interferens*, synonymt med *negativ överföring*.

Målspråket kan till exempel få låna ordföljden från modersmålet (1), en mycket vanlig interferenseffekt när målspråket är svenska, med sina ovanligt komplicerade ordföljdsregler i samband med underordnade satser och satsadverbial.

(1) *Då han kan inte stanna ensam.*

Det är också vanligt med interferens över ett annat andraspråk, som hos denna skribent, med arabiska som modersmål (2). Skribenten har enligt egen utsago goda kunskaper i engelska.

(2) *Nästa höst ska jag förtsetta att studera **in** T.H. Skolan **för** två år **mer** .*

En särskild typ av interferens som jag vill kalla ***fonologisk interferens*** kan medföra mer eller mindre kraftig deformering av ord (3).

(3) *i kväl en gammal man läser bocken som heter "Mordet på Cirkus" Han läggar sig i sängen. Det är **möligt** ute.*

Det är inte så lätt att omedelbart utläsa vad skribenten avsett att skriva, alltså ordet ***mörkt***. Skribentens modersmål är japanska. Östasiatiska språkbrukares svårighet med `/r/`-fonemet är allmänt bekant. Konsonantkluster av typen `[rkt]` förekommer inte heller i japanska. Exemplet är särskilt intressant därför att det så tydligt illustrerar hur fonologiska egenskaper kan få konsekvenser för den grafiska representationen av ett ord, och hur påtagligt fonologisk interferens kan förändra standardformen.

I termer av redigeringsavstånd krävs i det här fallet två ersättningar och en borttagning för att restaurera ***mörkt***. Detta ligger utom räckhåll för en stavningskontroll. Granskas stavningsmodul ger följande ersättningsförslag: [möjligt mjöligt mögligt]. Samtliga ligger bara en insättning ifrån det okända ordet.

Det finns inte heller några möjligheter att upptäcka och rätta till felet med hjälp av granskningsoch hjälpregler som tar hänsyn till kontexten, med mindre än att man tar till en kontextanalys på pragmatisk eller diskursiv nivå (”Vad handlar texten om? Den berättar om en person som på kvällen ligger och läser en bok i sängen”). Man skulle kunna tänka sig en uppsättning algoritmer för detektion och länkning av innehållsord, men en diskurskänslig *grammar checker* har mig veterligt ännu inte lanserats ens för engelska. Evidens för min analys av strängen ***möligt*** hittar man faktiskt i originalkorpusen, där det visar sig att aktuell uppsats skrivits som illustration till en bildserie där det alldeles uppenbart är mörka natten utanför mannens fönster.

1.5.3 Problem med analys av fel

Feldetektionen i språkgranskningssystem utgår på ett eller annat sätt från en redan ordklassanalyserad text och avviker råmaterialet tillräckligt mycket i fråga om grammatisk struktur blir ordklassigenkänningen inte längre meningsfull. Om en mycket stor del av orden resulterar i okända ord för programmet och en lika stor del har blivit andra existerande ord än de avsedda, då havererar kanske hela projektet att hjälpa användaren (Tabell 2).

Tabell 2. Svåranalyserad text ur SSM-korpusen. Konsekvent och oreflekterat har första alternativet i Granskas rättningsförslag valts vid korrekturen.

Ursprungstext	**...med korrigeringar efter kontroll med Granska**
Jag bor i Fetja. Jag åker från till Frescati halv timmer vi måste sitter i tunlbana och bussen. Jag måste vackna klockan sex, därför att det är lång väg och börjar min klass klockan 8, och Jag hoppa. Kan jag sultat kalas och börjar fakolititet	*Jag bor i Fetja. Jag åker från till Frescati halvtimmer vi måste ha suttit i tungbana och bussen. Jag måste vackra klockan sex, därför att det är lång väg och börjar min klass klockan 8, och Jag hoppa. Kan jag slutat kalas och börjar farolititet*

Korrigeringarna av uppsatstexten i tabell 2 har kanske inte allvarligt försvårat förståelsen av texten, men de har heller knappast hjälpt. En mänsklig bedömare med intresse för andraspråksinlärning skulle emellertid kunna peka på misstag som illustrerar *strategier i inlärningsprocessen*. Man kan förmoda att till exempel dubbelprepositionen *från till* är ett försök att beskriva en rörelse från A till B. Detta visar i sin tur att inläraren antagligen har tillgodogjort sig (eller håller på att tillgodogöra sig) de båda prepositionernas huvudsakliga användning i språket.

Anta att en automatisk grammatikkontroll utrustad med regler för att hitta prepositionsfel, detekterar dubbelprepositionen, ställer diagnosen ”Två prepositioner i följd efter finit verb verkar underligt” och ger två rättningsförslag – ”åker *till* Frescati” och ”åker *från* Frescati”. Med denna typ av återkoppling (feedback) riskerar man att undergräva skribentens kreativitet och i värsta fall blir hon eller han osäker på betydelsen av *till* och *från*.

1.5.4 Utmaningar och uppgifter för automatisk språkgranskning

Det går inte att komma ifrån att analys av skrivfel, med det entydiga syftet att korrigera gentemot en förutbestämd eller intuitivt närvarande standardnorm, är en i många stycken ofullkomlig analys. Detta problem utgör ytterligare en betydelsefull utmaning för utvecklarna av system för automatisk språkgranskning. Det innebär också att automatisk grammatikkontroll för närvarande kanske har sin viktigaste uppgift att fylla i en undervisningssituation, där programmen kan vara viktiga pedagogiska hjälpmedel i undervisningen (Knutsson et al 2002; Cochran Crocker 2002).

1.5.5 Granskning av Granskas granskning

Ett par examensarbeten över hur Granska fungerar på texter av andraspråksinlärare har redan gjorts (Öhrman 2000; Staerner2001). Öhrman bygger upp en generell feltypologi efter analys av en mindre, longitudinell andraspråkskorpus (ASU) och hon är inte optimistisk vad gäller verktygets utsikter att bättre kunna detektera felaktigt prepositionsval. Staerner har också använt ASU-korpusen och koncentrerar sig på ordföljdsfel. Hon lyfter dessutom fram den språkdidaktiska aspekten genom intervjuer med andraspråkslärare. ASU-korpusen är insamlad och arkiverad av Björn Hammarberg på Institutionen för Lingvistik, Stockholms universitet, precis som SSM-materialet som jag arbetat med.

2 Material

2.1 SSM

Akronymen *SSM* står för *Svenska som målspråk*, ett projekt för insamling och analys av svensk text av andraspråksinlärare som bedrevs vid Institutionen för lingvistik vid Stockholms universitet under åren 1972 – 1980 under ledning av Björn Hammarberg och Åke Viberg (Hammarberg 1977). Korpusen innehåller texter producerade vid preparandkurser i svenska för högskolebehörighet, vid dåvarande *Institute for English-Speaking Students* på Stockholms universitet. Detta betyder inte nödvändigtvis att studenterna var engelskkunniga men däremot att de hade motsvarande högskolebehörighet från hemlandet och att de hade något annat modersmål än svenska. Texterna består av skriftliga inträdesprov, klassrumsuppsatser och slutprov i uppsatsform. Slutprovet motsvaras idag av det så kallade *TISUS – test i svenska för universitets- och högskolestudier* – som administreras av Institutionen för nordiska språk.

Uppsatsskribenterna representerar tio modersmålsgrupper. Språkgrupperna har valts ut för att balansera korpusen språktypologiskt, men också för att reflektera aktuell distribution av olika invandrarspråk i Sverige. Förutom angivelse av modersmålstillhörighet och datum för excerpering, samt information om antal textbidrag och sammanlagd textstorlek på bidragen, innehåller korpusen för varje skribent uppgifter om geografisk härkomst, kön, födelseår, vistelselängd i Sverige, kursstadium i svenska och kunskaper i övriga språk, allt enligt skribentens egen utsago. Materialet har samlats in under åren 1973 och 1974. Originalen är handskrivna. Uppsatserna har typnormaliserats med elektrisk skrivmaskin.

I och med att CrossCheck-projektet drogs igång kunde man inleda arbetet med att ta fram en elektronisk version av SSM-korpusen. Detta går till så att den maskinskrivna versionen läses optiskt och lagras i textformat. Resultatet korrekturläses mot inte bara den maskinskrivna versionen utan också mot de fotostatkopierade originalen. Jag har fått till mitt förfogande den första delen av korpusen som blev färdig i elektronisk version, motsvarande modersmålsgrupperna arabiska (55 uppsatser), engelska (57 uppsatser) och japanska (49 uppsatser), sammanlagt 161 uppsatsfiler om drygt 30 000 ord. Dessa tre språkgrupper råkar kontrastera varandra språktypologiskt vad gäller grundläggande ordföljd. Jag gör här en kort språktypologisk presentation – ur ett prepositionsperspektiv – av vart och ett av språken. Jag har använt Andersson (1987) som faktareferens.

Arabiskan är ett utpräglat *flekterande* språk, med grammatiska böjningsmorfem som affixeras till en semantiskt motiverad ordrot. Prepositioner används precis som i svenskan. Arabiskan har genitivsuffix med böjningsmorfologi. Grundläggande ordföljd: VSO.

Engelskan står typologiskt nära svenskan och använder prepositioner på samma sätt. Engelskan har också grammatiskt genitivmorfem som ser ut som svenskans, men kan ibland använda prepositionskonstruktion istället. Grundläggande ordföljd: SVO.

Japanskan är ett *agglutinerande kasusspråk*, man staplar avledningsmorfem på (nominala) huvudord, utan bruk av prepositioner i vår mening. Semantiska relationer som markeras med

preposition i svenskan, markeras alltså på annat sätt i japanskan. Man kan emellertid identifiera motsvarigheter till prepositioner såsom postpositioner. Grundläggande ordföljd: SOV.

Jag har lyft ut drygt en tiondel av materialet för att använda som testtext vid utvärdering (avsnitt 3.2). I analysen är därmed de japanska uppsatsskribenterna representerade med 7760 ord, de arabiska med 10 550 och de engelsktalande med 10 600 ord. Siffrorna ligger i överkant, eftersom textfilerna innehåller filnamn och kommentarer från korrektur av den optiska scanningen.

2.2 SUC

Det har uppstått behov av att konsultera en standardspråkskorpus för information om distribution och frekvens för vissa konstruktioner eller kontexter. Därför har analysen av prepositionskonstruktioner kompletteras med sökningar i SUC-korpusen – *the Stockholm-Umeå Corpus*, version 2.0 1998 (Källgren 1998). SUC är en balanserad SGML-taggad standardspråkskorpus, en miljon ord stor. Samplingen har gjorts ur tio genrer: *reportagetexter*, *ledarspalter*, *recensioner*, *religion*, *förvärvsliv*, *populärvetenskap*, *biografier*, *blandat*, samt *vetenskap och skönlitteratur*. Korpusurval och sökmetoder presenteras i avsnitt 3.4.5.

3 Metod

3.1 Vad är en preposition?

Prepositionens status som grammatisk kategori har varit ganska konstant under de senaste hundra åren av svensk grammatikteori, även om den inte har samma pondus som verb eller substantiv. Däremot skiftar omfånget på kategorin ganska avsevärt beroende på vilken grammatiker man konsulterar. Det är särskilt prepositionens avgränsning gentemot partikeladverb och subjunktioner som är tänjbar, och jag ska i det följande försöka åskådliggöra denna problematik.

De flesta grammatikteoretiker verkar vara eniga om att prepositionen prototypiskt

1. binder ihop ett nominalt led med ett annat, och
2. är morfologiskt sluten, det vill säga oböjlig

Prepositionen står i allmänhet med sina nominala led, eller *referenter*, på var sida om sig (*smulorna **på** bordet* eller *smulorna **under** bordet* eller *bordet **med** smulorna* eller *bordet **utan** smulor **på***) Den sista frasen utgör exempel på en *elliptisk* konstruktion, där ***på*** har en underförstådd referent (bordet) på sin högra sida, fast platsen gapar tom.

De flesta grammatiker menar också att prepositionen i första hand hör samman med sin efterställda referent, som får sitt uttryck i det jag kallar för prepositionens *rektion*, i linje med bland annat Beckmans *Svensk språklära* (Beckman 1968) och *Svenska Akademiens grammatik*

(Teleman et al. 1999).[3] Rektionen är vanligen en nominalfras men kan också vara till exempel en infinitivsats med nominal funktion, eller ett adverb (*Jag kände behov av **att ta en paus**. Tack för **igår**.*) Prepositionen bildar tillsammans med sin rektion en *prepositionsfras*, som oftast har adverbial funktion i satsen (Thorell 1982:177f).

Den referent som beskrivs i rektionen kallas i för *B-referent*. På andra sidan prepositionen hittar vi *A-referenten*. Strukturen för A-referentens uttryck är generellt svårare att identifiera, men A-referenten är semantiskt ofta mycket inflytelserik i det som jag något vagt kallar för *prepositionskonstruktion* och som omfattar prepositionsfras och A-referent(er). Med följande exempel vill jag illustrera hur semantiska egenskaper hos prepositionen och dess referenter får konsekvenser för syntaktisk struktur.

> *Jag satte mej **vid** bordet **utan** smulor **på***
> De semantiska relationerna mellan *Jag satte mej*, *bordet* och *smulor* blockerar en läsning med *mej* som underförstådd rektion till *på*, och därmed är identiteten fastställd för rektionen till *vid = bordet utan smulor på*
>
> *Jag åt upp smulorna **på** bordet*
> Prepositionskonstruktionen börjar med *smulorna*
>
> *Jag lämnade smulorna **på** bordet*
> Prepositionskonstruktionen börjar med *Jag* (inte med *smulorna*)
>
> *Jag torkade **av** smulorna **på** bordet*
> Rektionen till *av = smulorna på bordet*
>
> *Jag torkade **av** smulorna **från** bordet*
> Rektionen till *av = smulorna* (inte *smulorna från bordet*)

Det är just förbindelser som *torkade **av*** som utgör vattendelare mellan olika definitioner av begreppet preposition. Många grammatiker räknar *torkade **av*** som ett sammansatt verb eller *partikelverb*, där ***av*** analyseras som *verbpartikel* eller *partikel* eller *partikeladverb* eller rätt och slätt *adverb*, i kontrast till prepositionen ***av*** i en sats som *Smulorna torkade **av** värmen från solen*. Både SAG och Olof Thorell (Thorell 1982) analyserar partikeln som preposition – SAG med motiveringen att den kan ”ta rektion” – men med *funktion* som partikeladverbial.

Vanligast är emellertid att dessa olika *av* också får olika ordklassanalys. Lindberg (1980) hävdar att prepositionen, till skillnad från sin homonym i funktion som partikeladverbial, **måste** ta rektion, eller *styrning* som hon kallar det. Visserligen menar hon också att det i ett exempel som *ta för dig av tårtan* ”är till ingen nytta att grubbla över om *för* är preposition eller ej. Det ingår i en fast verbfras som lyder *ta för sig*, och ordets status av tryckstark verbpartikel är det enda som säkert kan konstateras.” (Lindberg 1980:114f). Likväl kallar hon den för adverb, och ytterligare analys av verbpartikeln lyser med sin frånvaro såväl under adverb- som verbparagraf, även om

[3] Hädanefter hänvisar jag till *Svenska akademiens grammatik* med akronymen *SAG*. Grammatiken är utgiven i fyra delar enligt analysnivå. Jag har i första hand använt mig av den morfologiska beskrivningen – del 2.

partikelverben avhandlas kort i avsnittet om satsstruktur (Lindberg 1980:177). I många grammatiska beskrivningar analyseras verbpartikeln inte alls.

Granskas tagger gör skillnad mellan partikel (<pl>) och preposition (<pp>). Eftersom taggern har gjort sina frekvensberäkningar på SUC-korpusen så är taggerns ordklasstilldelning en spegling av analysen i SUC (Granska har också nästan identiska taggbeteckningar – grundtagguppsättningen presenteras i sin helhet i Appendix). Verbpartiklarna tilldelas i princip taggen <pl> i SUC. Ordgruppen har emellertid fått byta identitet mer än en gång sedan första versionen av SUC – partiklarna har i olika omgångar blivit taggade både som adverb och som prepositioner. En viktig poäng i sammanhanget är att jag som regelkonstruktör självklart måste anpassa mig till systemets ordklassanalys när jag anger särdrag och värden för objekt som ska matchas i granskningsreglerna (avsnitt 4.1).

Prepositionen ***utan*** i *bordet utan smulor på* kan också byta grammatisk kategori (*Jag satte mig inte vid bordet **utan** stannade en bit därifrån*). I sådana fall väljer grammatikerna enhälligt att klassa prepositionens homonym som *konjunktion*. Värre är det med prepositioner som ***innan*** och ***sedan***. Den senare är kanske vanligast som adverb, men de båda får i den grammatiska beskrivningen en gemensam kluven identitet som preposition och *underordnande konjunktion* (=*subjunktion*), som i vissa fall är svår att motivera. Följande exempel har jag stulit från Ljunggren (1951).

*De har levt lyckligt **sedan** giftermålet* – preposition
*De har levt lyckligt **sedan** de gifte sig* – subjunktion
*De levde lyckligt **innan** giftermålet* – preposition
*De levde lyckligt **innan** de gifte sig* – subjunktion

Skälet till att skilja på ordklasstillhörigheten ska man söka i rektionen (subjunktioner tar rektion analogt med prepositioner i SAG:s analys). Prepositionen kan mycket väl ha en bisats som rektion, men aldrig utan hjälp av ett bisatsinledande element som *att*, *om*, *när* eller varför inte *innan* (*De har levt lyckligt **sedan innan** de gifte sig*). Vissa språkvårdare avvisar prepositionellt *innan* och föreskriver *före*. Granska taggar konsekvent *innan* som subjunktion.

En del grammatiker väljer istället att fokusera likheterna mellan preposition och subjunktion. Man kan säga att prepositioner besitter själva kardinalegenskapen hos subjunktionerna, nämligen att de underordnar ett element (B-referenten) i relation till ett annat (A-referenten). Jespersen (1968:87ff) underkänner rentav själva distinktionen och argumenterar för att avskaffa kategorierna adverb, konjunktion, preposition och interjektion i den grammatiska beskrivningen, och istället analysera alla dessa som *partiklar*!

Thorell (1982) ansluter sig lite försiktigt till denna analys då han presenterar en kategori *partiklar* med subkategorierna *adverb*, *fogeord* och *interjektioner*. Fogeorden får i sin tur underavdelningarna prepositioner, subjunktioner, konjunktioner och infinitivmärket (att). En partikel, till exempel *sedan*, kan då ha vad Thorell kallar en *huvudfunktion* (adverbfunktionen) och en eller flera *bifunktioner* (funktion som preposition eller som *bisatsinledare*=subjunktion).

Collinder (1970) inför också en kategori *partiklar och kasusord (adverb)*, till vilken han för adverb, sam- och underordnande konjunktioner, prepositioner och interjektioner. Lindberg (1980)

kallar prepositioner och konjunktioner för ”vårt språks förbindarord” och Beckman (1968) slutligen, använder en överordnad kategori *bindeord* för prepositioner och sam- och underordnande konjunktioner. I tabell 3 har jag sammanställt några olika analyser av prepositionsord.

Tabell 3. Alternativa ordklassanalyser av ord med prepositionsegenskaper.

Exempel	**Ordklassanalys**			
	SAG	**Thorell**	**Lindberg**	**Granska**
*Hon klev **av** tåget*	preposition	preposition	adverb/verbpartikel	preposition
*Hon klev **av** i Rimforsa*	preposition	preposition	adverb/verbpartikel	partikel
*Flickan gick **över** till andra sidan gatan*	preposition	preposition	adverb/verbpartikel	partikel
*De har levt lyckligt **sedan** giftermålet*	preposition	preposition	preposition	preposition
*De har levt lyckligt **sedan** de gifte sig*	subjunktion	subjunktion	subjunktion	subjunktion
*Prepositioner uppträder **som** underordnande konjunktioner*	subjunktion	preposition	pronomen?	konjunktion
*Lejonen kom **nära** oss*	preposition	?	?	preposition
*Lejonen kom **närmare** oss*	preposition?	?	?	adjektiv

Granskas tagger analyserar alltså *av* som preposition i *Hon klev **av** tåget* (liksom för övrigt i *Jag torkade **av** smulorna*) men som partikel i *Hon klev **av** i Rimforsa*. Problemet är säkerligen rent statistiskt – *av* är generellt vanligare i prepositionsfunktion (Megyesi 1996:42f) men knappast i en kontext där *av* omedelbart följs av en annan preposition. Taggerns val utreds mer i detalj i avsnitt 4.1. Exemplet är annars en utmärkt illustration till likheten mellan relationen verb – objekt och relationen preposition – rektion. Man kan argumentera för att rektionen lika gärna kan kallas för objekt (Jespersen 1968:87ff).

> *Hon klev av tåget* – rektionen är objekt till *av*, som är en transitiv preposition
> *Hon klev av i Rimforsa – av* saknar rektion och är intransitiv preposition

Prepositionsfrasen *i Rimforsa* är här med SAGs term ett fritt adverbial i satsen. *av* står i denna sats *absolut* (utan rektion), och jämförs i SAG med ”transitiva verb utan sitt objekt” (SAG2:626).

Ordet ***nära*** intar något av en särställning, då det uppträder i skepnad av såväl adverb (*Han står mig nära*), som adjektiv (*Hon är en nära vän*) och preposition (*Hon bor nära mig*). ***Nära*** hör dessutom till de prepositioner som inte riktigt anpassar sig till grundkriterium nummer 2 ovan, eftersom det håller sig med böjningsmorfologi.

Mycket ofta förekommer en mer eller mindre stark lexikal bindning mellan preposition och endera eller båda referenterna.

> *Jag brukar handla **på Kvantum** men du handlar **i kvartersbutiken**.*
> *Han är mycket **intresserad av** boken.*
> *Kalle är **van vid** att få som han vill.*
> *Jag **längtar efter dej**.*
> *Jag **längtar till** (eller **efter**) sommaren.*
> *Jag **längtar till efter** sommaren* (här är sista prepositionsfrasen själv rektion till *till*)

Dessa förbindelser tycks vila på mycket lös eller godtycklig semantisk grund och förfaller snarare vara ett uttryck för ett slags stelnad hävd i språkbruket. Somliga prepositionella uttryck är helt lexikaliserade eller stelnade, som *till exempel* och *istället för*, eller äldre genitivkonstruktioner som *till bords* och *till sjöss*, andra har rentav förlorat sin identitet som prepositionsfras, som *igår*. I själva verket skulle man kunna rada upp prepositionskonstruktionerna längs en kontinuerlig skala efter grad av lexikalisering, från ”tillfällig förbindelse” till ”idiom” (avsnitt 4.3 och 7.2).

3.2 Kontrollgrupp

Jag har valt ut tjugo uppsatsfiler (3 800 ord) och sparat som referens- och utvärderingsmaterial. Tack vare SSM-materialets ganska utförliga bakgrundsinformation om skribenterna, anser jag att jag har kunnat göra ett balanserat urval där hänsyn tagits till kön, ålder, övriga språkkunskaper, textmängd (en del skribenter är representerade med upp till sju uppsatser, många bidrar med en enstaka, längden på texterna varierar också avsevärt), vistelsetid i Sverige och kursstadium. Vistelsetid och kursstadium gäller för varje skribent tidpunkten för första excerperingen. Det betyder att i vissa fall ytterligare fyra månader kan ha förflutit fram till den sista excerperingen från en och samma skribent. Angivelsen av kursstadium rimmar ganska illa med språket i motsvarande texter. Jag kan knappast mer än kommentera dessa omständigheter i min undersökning och jag har inte strukit kursstadium eller vistelsetid som parametrar för balanseringen.

Ett alternativt tillvägagångssätt kunde varit att slumpa ut lika många filer, men materialet är inte tillräckligt stort för att kunna ge statistisk validitet åt ett slumpmässigt urval.

3.3 Identifiering av prepositionskonstruktioner

För att få hjälp att hitta prepositionskonstruktioner kan man ge Granska en regel för detektion av prepositionstagg. Men det är som sagt ingalunda självklart vad som är en preposition och därför inte heller vad som taggas som preposition eller vilka ord som får prepositionstagg. Först har man alltså att kontrollera taggerns analys av prepositioner, entydiga prepositioner kan ha blivit taggade som subjunktioner eller partiklar och är feltaggningen systematisk så behöver man utöka matchningsvillkoren för den regel som ska hitta prepositionskonstruktioner. För att arbetsbördan inte ska bli för tung är det nödvändigt att extrahera ett stickprov ur materialet. I det här fallet tyckte jag det var lämpligt att göra ett slumpmässigt urval. En slumpgenerator tog en lista av filer (alla 141 uppsatsfiler som var aktuella för analys) och skrev ut en ny fil med lika många rader som filer i listan, en slumpmässigt utvald rad från varje fil. På så sätt fick jag en nonsenstext på 141 rader.

Jag konkatenerade sedan uppsatsfilerna för att lättare kunna redigera nonsensfilen så att varje rad formulerades om till en komplett mening, enligt tabell 4. Sedan lät jag Granska ordklasstagga resultatet. Den ordklasstaggade texten examinerade jag för hand.

Tabell 4. Illustration till stickprovsmetod för granskning av Granskas prepositionstaggning. För varje rad har jag kompletterat första sekvensen fram till meningsslut, och strukit resten. Lägg märke till ”feltaggningen” av *i morgon Kaffé*. Den kommenteras i avsnitt 4.1.

Första två raderna i nonsens-filen	*leva ensamma. Men i Egypten de Kanske måste vara hemma hos sina föräldrar tills de slutar*
	1973 gick han som vanling till jobbet ock där travade han sin chef i morgon Kaffé. Jag vill
Första två meningarna i den redigerade filen	*När pojkarna och flickorna här I Sverige blir 16 år går de hemifrån ofta och försöka att leva ensamma. Den 15 Februari 1973 gick han som vanling till jobbet ock där travade han sin chef i morgon Kaffé.*
Taggad version av första två meningarna	*När* <ha> *pojkarna* <nn.utr.plu.def.nom> *och* <kn> *flickorna* <nn.utr.plu.def.nom> *här* <ab> *I* <pp> *Sverige* <pm.nom> *blir* <vb.prs.akt.kop> *16* <rg.utr/neu.plu.ind/def.nom> *år* <nn.neu.plu.ind.nom> *går* <vb.prs.akt> *de* <pn.utr/neu.plu.def.sub> *hemifrån* <ab> *ofta* <ab.pos> *och* <kn> *försöka* <vb.inf.akt.mod> *att* <ie> *leva* <vb.inf.akt> *ensamma* <jj.pos.utr/neu.plu.ind/def.nom> . <mad> *Den 15 Februari 1973* <rg.yea> *gick* <vb.prt.akt> *han* <pn.utr.sin.def.sub> *som* <kn> *vanling* <nn.utr.sin.ind.nom> *till* <pp> *jobbet* <nn.neu.sin.def.nom> *ock* <ab> *där* <ab> *travade* <vb.prt.akt> *han* <pn.utr.sin.def.sub> *sin* <ps.utr.sin.def> *chef* <nn.utr.sin.ind.nom> *i morgon* <ab> *Kaffé* <pm.nom> . <mad>

Det visade sig att prepositionstaggningen var tillfredsställande (avsnitt 4.1), varför jag inte behövde modifiera sökregeln för prepositioner. Jag lät sedan webb-versionen av Granska tagga hela textmaterialet samtidigt som sökregeln tillämpades. Webb-versionen är inte den mest uppdaterade versionen av Granska men den har den fördelen att matcshningar rödmarkeras.

Sökregeln matchade ingen kontext runt prepositionerna och returnerade hela texten till skärm, taggad och med endast prepositioner med prepositionstagg rödmarkerade.

3.4 Analysmetoder

3.4.1 Översikt och problemdefinition

Materialet har först blivit föremål för en typologisk analys, där jag kommenterat och klassificerat prepositionsrelaterade skrivfel (avsnitt 3.4.2). Jag har också kontrollerat och kommenterat Stavas detektion och Granskas taggning och diagnos där jag tyckt att det varit befogat, till exempel vid utpräglade stavfel.

Alla förekomster av problem i prepositionskonstruktionen som jag klassat som prepositionsfel (se nedan, och avsnitt 3.4.2) har jag lyft ut med tre tokens för- och tre tokens efterkontext, inklusive ordklasstaggar, för att sedan kunna examinera med hjälp av textverktyg under Unix (avsnitt 3.4.3). Denna analys bildar underlag för nya regler till systemet.

En kompletterande analysmetod har använts för ord eller ordformer som återkommer i materialet som missuppfattningar eller förväxlingar av de korrekta eller uppenbart avsedda ordformerna. Jag har grupperat dessa i så kallade *confusion sets* eller *förväxlingsmängder* (avsnitten 3.4.4 och 4.5).

Jag har genomfört min analys oberoende av om somliga fel upptäcks och korrigeras på ett tillfredsställande sätt i nuvarande version av Granska. Granska levererar som sagt (avsnitt 1.3.2) redan nu multipla analyser. Ett ord som detekteras av Stava kan dessutom samtidigt och av någon annan orsak matchas av en granskningsregel, man kan alltså som användare råka ut för att bli upptäckt av bägge och få två olika diagnoser på samma fel, som i denna granskning av en mening ur SSM-materialet (programmets rättningsförslag inom hakparenteser).

Hunden sovar gott.

sovar Okänt ord (stav1@stavning) Stava
[sover
svar
sova
solar]

sovar gott misstänkt särskrivning (sär_stava1A@sär)
[sovargott]

Tillämpningen av särskrivningsregeln ser ut som en typisk löjeväckande felrättning av datorn. Det är den väl också. Samtidigt avslöjar misstaget på ett intressant sätt hur Granska och Stava interagerar. Orden *gott* och *var*, men också *so*, accepteras var för sig av Stava (*so* är ett annat ord för *sugga*). Stavas sammansättningsregler accepterar också *vargott* och *sovargott* men alltså inte *sovar.* Granskas särskrivningsregel är en regel i gruppen ”särskrivna stavfel” och matchar ett felstavat ord som blir rättstavat vid sammanslagning med närmast följande ord. Den ovanliga sammanskrivningen av finit verb och adverb blockeras inte i matchningsvillkoren för regeln.

I materialet förekommer ofta avvikelser i rektionen (4). Tumregeln för dessa fall har varit att endast de fel där problem föreligger i själva prepositionen blivit föremål för vidare analys.

(4) *after **Kaffé** Ringa han till **hans syster** i Göteborg ock prata på **hans resar till där***

I exempel (4) hittar vi således fyra rektionsproblem men endast två prepositionsproblem (***after*** *Kaffé* och *prata **på** hans resar*). Mycket ofta är det emellertid en tolkningsfråga om felet ska betraktas som ett problem i rektionen eller som ett prepositionsproblem som i (5) och (6) från en och samma uppsats. Skribenten har japanska som modersmål, deklarationen om språket i hemmet har med uppsatsämnet att göra.

(5) ***I hemma** talar vi turkiska.*
(6) *Ibland hämtar de med mat **till hem**.*

Jag har analyserat (5) och (6) som fall av otillbörlig insättning av preposition. Valet bygger på en parallelltolkning av stilnivå i hela texten. Också den sista prepositionsfrasen i exempel (4) kan tolkas som en överflödig förekomst av preposition, men tar man bort ***till*** så står man ändå med ordsträngen ***där*** som i sin tur måste modifieras för att passa in – korrektion av felet blir alltför komplex om det analyseras som prepositionsfel.

Samma analysprincip gäller vid lexikonfel på A-referentens sida om prepositionen, det finns alltså inget prepositionsproblem i en konstruktion som ***länga*** (istället för *längta*) ***efter våren***.

Jag har också analyserat vissa felaktigt konstruerade idiom som prepositionsfel, även om felet inte ligger i prepositionsvalet, detta på grund av den fullständiga prepositionsfrasens starka identitet som idiom (7) och (8).

(7) *Jag tänkte att försöka läsa om oregelbund verben **för den sista gången***
(8) ***För mestsdels** , läser jag saker om politik .*

(9) är ett gränsfall (som lämnats utan analys)

(9) *atmosfären av romanen **i allmänheten** är seriös och tänkfull*

3.4.2 Feltypologi

Tabell 5. Feltypologin

Feltyp	**Beskrivning**	**Exempel**
insertion (insättning)	Olycklig insättning av preposition.	*Fängelseanstalten är en gammal byggnad från **i** mitten 1800 talet*
deletion (borttagning)	Olyckligt utelämnad preposition.	*Fängelseanstalten är en gammal byggnad från i mitten **[ ]** 1800 talet*
substitution (ersättning)	Olyckligt prepositionsval, ofta vid lexikala bindningskrav.	*Det är svårt **till** mig att prata om politik här isverige*
fel ordform	Missformad preposition, eller preposition som RWE.	*hon **i** fortfarande till att födda barn << Fabrik för barn >> och ta hand om dem **ettan** en reeal hjälp från manen.* *Han var dum och där **för** hade staten tagit honom till ett hem för **efter** blevna.* *Det är svårt till mig att prata om politik här **isverige***
fel i idiom	Missformat prepositionellt idiom, ibland med godtyckliga gränser gentemot substitution.	*Jag tycker att det är intressant att läser nyheter i USA **genom svenska ögen***
genitivfel	Genitivkonstruktion med ”av”.	***Landena av Trädje Världen** vacknar*
semantiskt fel	Semantiskt problematisk konstruktion, eller på annat sätt svårklassificerat eller svårformaliserat prepositionsfel.	*Romanen skriver om bankrånare.* *Han accepterar inte fredspriset i tiden som vietnam folk kämper för sin frihet.*

Fortsättningsvis kallar jag kategorierna *insertion*, *deletion* och *substitution* för *insättning*, *borttagning* och *ersättning* respektive.

Jag hade till att börja med också en kategori *transposition* för konstruktioner av typen **Jag såg omkring mig*. Men de flesta felplacerade prepositioner dök upp i konstruktioner som över huvud taget var syntaktiskt mycket svårrestaurerade (*Ante tittade inte omkring sig själv* eller *men nu vanade jag ganska vid här*), konstruktioner som istället hänförts till kategorin semantiska fel och alltså lämnats utanför analysen (avsnitt 4.2.4). En transposition kan dessutom skrivas om i termer av borttagning och insättning och bidrar inte till förenkling av regelkonstruktionen.

Många specifika innehållsord i omedelbar anslutning till preposition är överrepresenterade i materialet (*tåget, Stockholm, sängen, skolan, Sverige, svenska, vintern*). Detta beror delvis på att skribenterna läst samma kurser i svenska, och därmed fått samma uppsatsämnen.

3.4.3 Kontextfönster

Prepositionsfelen har blivit föremål för kvantitativ analys med avseende på såväl lexikal kontext som taggkontext. För varje identifierat prepositionsfel har jag extraherat den prepositionella enheten – det vill säga felaktig preposition, saknad preposition, eller ord som skulle varit preposition – med tre ords för- och tre ords efterkontext, med tillhörande ordklassanalys (taggar). Varje prepositionsfel blir då representerat som ett *7-gram* med den prepositionella enheten i mitten. Jag kallar fortsättningsvis dessa 7-gram för *kontextfönster* (Tabell 6).

Denna typ av representation tillåter statistisk analys med hjälp av textverktygen under Unix. Jag har kunnat snäva in kontexten, eller bara titta på taggsekvensen eller bara ordsekvensen, eller undersöka alla kombinationer av ord- och taggsekvens.

Tabell 6. Exempel på kontextfönster. Ord fram till punkt och ord efter punkt representeras med hakparenteser, med namnlös tagg. Vid RWE:s till följd av särskrivning hamnar den oavsedda prepositionen i mitten. Bortfall av preposition markeras med hakparenteser, men med namn på taggen.

Textavsnitt med prepositionsfel	*, sen förtsett till England. After sex månader comde Jag till båcka att förtset min studing*
Taggad version	*,* <mid> *sen* <ab> *förtset* <nn.neu.sin.def.nom> *till* <pp> *England* <pm.nom> . <mad> *After* <pm.nom> *sex* <rg.nom> *månader* <nn.utr.plu.ind.nom> *comde* <vb.prt.akt> *Jag* <pn.utr.sin.def.sub> *till* <pp> *båcka* <nn.utr.sin.ind.nom> *att* <sn> *förtset* <nn.neu.sin.def.nom> *min* <ps.utr.sin.def> *studing* <nn.utr.sin.ind.nom>
Kontextfönster	[ ] <[ ]> [ ] <[ ]> [ ] <[ ]> *After* <pm.nom> *sex* <rg.nom> *månader* <nn.utr.plu.ind.nom> *comde* <vb.prt.akt>
	månader <nn.utr.plu.ind.nom> *comde* <vb.prt.akt> *Jag* <pn.utr.sin.def.sub> *till* <pp> *båcka* <nn.utr.sin.ind.nom> *att* <sn> *förtset* <nn.neu.sin.def.nom>
	Jag <pn.utr.sin.def.sub> *till* <pp> *båcka* <nn.utr.sin.ind.nom> [ ] <[pp]> *att* <sn> *förtset* <nn.neu.sin.def.nom> *min* <ps.utr.sin.def>

Multipelt bortfall i preopositionskonstruktionen (*Satte jag* [mig i]? *stolen för att läsa, men kunde jag inte* eller *Jag tog promenad istället* [för att] *läsa*) får som konsekvens att den icke prepositionella enheten fortfarande saknar representation i kontextfönstret.

En multipel särskrivning som *försvuma* ***hem i från*** *över helgerna* resulterar i två kontextfönster.

Eftersom det inte är ord utan tokens som taggas, så hamnar till exempel det avgörande elementet ***säjer*** i *Alla säjer "hej" på mig* utanför kontextfönstret (ett citationstecken är ett token).

3.4.4 Förväxlingsmängder

När det gäller automatisk grammatikkontroll är felanalys med så kallade *confusion sets* en metod för att angripa problemet med RWE:s (Golding & Schabes 1996). Metoden innebär ganska enkelt att man först identifierar ord som brukar förväxlas med varandra, antingen som konsekvens av fonologiska stavfel (*än/en*) eller som konsekvens av att skribenten missuppfattat betydelsen

(*vederlägga/bevisa*) eller användningen (*stämning/atmosfär*) av ett ord. Även högfrekventa förväxlingar av grafiskt närliggande ord (*men/med*) är lämpliga kandidater. Dessa ordpar är vart och ett exempel på en *confusion set*.

En *confusion set* kan innehålla två eller fler element. Uppgiften för programmet blir att för varje matchning med ett element som tillhör en *confusion set*, avgöra vilket av orden i mängden som skribenten troligast avsett att skriva. Varje element i mängden utsätts nu för statistisk analys och det ord väljs som visar sig vara sannolikast i aktuell kontext. I princip fungerar proceduren likadant som när en statistisk trigram-tagger väljer analys för ett ord som *om*, med den skillnaden att elementen i förväxlingsmängden då är ordklasstaggar. Programmet får information om att *om* uppträder med någon av taggarna <sn>, <pp> och <pl> och gör en kontroll av sannolikheten för var och en av ordklassanalyserna i aktuell kontext. Den tagg vinner som får högst sannolikhetsvärde. Poängen med *confusion sets* är alltså att ge systemet möjlighet att utföra den här typen av operationer på lexikala enheter som man redan grupperat i en särskild modul.

För att återvända till ordet *om* så detekteras och korrigeras ett stavfel som **omm* utan problem av Granskas stavningsmodul, men däremot inte ett RWE som *öm.* Ordet kommer att taggas som adjektiv oavsett hur kontexten ser ut, därför att sannolikheten för att ordet uppträder med någon annan ordklasstillhörighet är noll. I värsta fall uppnår man en dominoeffekt så att programmet på det här stället i texten snuvas på kontextvillkor för en granskningsregel som annars hade tillämpats. Om *öm* nu istället hade varit ett element i en *confusion set* så hade man kunnat ge programmet möjlighet att göra om sin analys. Antingen matchas då förväxlingsmängden redan i taggningsmomentet, ordet taggas om och den motsägelsefulla kombinationen av *öm* och låt oss säga subjunktionstagg, kanske detekteras av en granskningsregel. Eller så kan man (som jag har gjort) matcha förväxlingsmängden i en granskningsregel direkt.

Mina förväxlingsmängder är inte *confusion sets* i egentlig mening. De uppträder i skepnad av granskningsregler och anger standardformen av en preposition (eller i vissa fall ett idiom) som mål för den operation som ska utföras på något av elementen i förväxlingsmängden. Även förväxlingar som inte är RWE:s får tillhöra förväxlingsmängden. Analysmetoden har i första hand använts för formfel. Dessa behandlas i avsnitt 4.5.

3.4.5 Sökning i standardspråkskorpus

SUC är i sig en balanserad korpus för jag skulle vilja påstå vårdat standardspråk. Det har inte varit nödvändigt eller ens särskilt lämpligt att använda hela SUC (en miljon ord) som referensmaterial. Korpusens filer är mycket samarbetsvilligt namngivna, så att det är lätt att säkra representation från alla genrer genom att till exempel plocka ut alla filer med sifferkombinationen 01 i filnamnet. Så har jag också gjort. Den relativa storleken för varje genre motsvarar då också hela korpusens, vad gäller antalet filer. Filerna kan ju vara lite olika stora, men så långt har jag inte kontrollerat representativiteten. Mitt referensmaterial blev på det här sättet ganska precis en tiondel så stort som hela korpusen, eller ungefär etthundratusen ord.

3.5 Regelskrivning

Granskningsreglerna har jag skrivit och redigerat i en texteditor och kompilerat och testat i DOS-versionen av Granska. För detta har jag använt ett paket med befintliga hjälpregler men utan Granskas granskningsregler. På så vis har jag kunnat koncentrera mig på resultatet av granskning med mina regler.

3.6 Utvärderingsmetoder

3.6.1 Endast prepositionsregler

Den färdiga regeluppsättningen har jag utvärderat på den textmassa jag kallar kontrollgrupp (avsnitt 3.2). Resultatet har analyserats manuellt med avseende på *täckning* och *precision* (se nedan). Jag har inte kompletterat aktuell version av Granska med mina regler och jämfört resultatet med granskning utan mina regler – om prepositionsreglerna ger resultat självständigt ökar de täckningen också generellt. Beräkning av precisionen för mina regler är dessutom helt oberoende av interaktion med övriga granskningregler, eftersom någon sådan interaktion inte finns. Prepositionsreglerna fungerar oberoende av andra regelpaket, ungefär som Stava-modulen i Granska arbetar oberoende av granskningsreglerna. Som systemets regelspråk är konstruerat är det överhuvudtaget mycket ovanligt att en regel skuggar indata för en annan regel.[4]

3.6.2 Täckning och precision

Begreppen *täckning* (*recall*) och *precision* är fundamentala i språkteknologisk utvärderingsmetod. För ett textgranskningssystems vidkommande är täckning lika med kvoten av antalet korrekt detekterade fel i en text och det absoluta antalet fel i texten. Detta ”absoluta” belopp (facit om man så vill), motsvaras ofta av resultatet av mänsklig korrektur. Värdet för precisionen är lika med kvoten av de korrekt detekterade felen och samtliga feldetektioner.

t = täckning
p = precision
n = absolut eller verklig felmängd
d = upptäckta (detekterade) fel som verkligen var fel – d är en delmängd av n
f = upptäckta fel som inte var fel (falska alarm)

$t = d / n$
$p = d / (d + f)$

Missade (oupptäckta) fel definieras som n – d

För automatiska språkgranskare gäller typiskt att hög täckning betyder låg precision, och omvänt att hög precision betyder låg täckning. Jag ska förtydliga med ett exempel.

[4] En kalibrering av prepositionsregler gentemot övriga regler, för att undvika till exempel trippeldetektioner, får utarbetas vid ett senare tillfälle, i samband med städning i regelfilerna.

Vi föreställer oss en granskningsregel som detekterar konstruktionen **Kalle **läste på** tidningen att Örgryte vunnit derbyt* och föreslår *läste **i** tidningen*. Regeln är konstruerad så att den detekterar alla förekomster av *läste* med efterföljande preposition *på* och därpå följande nominalfras. Det betyder att regeln också upptäcker ?*läste på boken* och **läste på veckan*, men också *läste på mjölkpaketet* och *läste på distans* med rättningsförslag ?*läste **i** mjölkpaketet* och **läste **i** distans*. De så kallade *falska alarmen* kommer kanske rentav att bli fler än de ”äkta”. Regeln har uppnått hundraprocentig täckning för just inkorrekt förekomst av *på* efter *läste*. Men hur är det med precisionen? Om vi tänker oss att en mänsklig utvärderare fann att de äkta alarmen var 4 och de falska uppgick till 6, då blev värdet för precisionen 4/(4+6), eller 40 procent.

Ett system med hög täckningsprestanda kan vara önskvärt för en språk- eller skrivsäker användare, som lätt kan förkasta diagnosen vid falska alarm och ignorera rättningsförslagen. Jag har istället försökt begränsa antalet falska alarm till ett minimum (avsnitten 4.2.2 och 6), för att undvika att den osäkre användaren gör kontraproduktiva korrektioner i sin text. Därmed tappar systemet samtidigt i täckning, det vill säga programmet går miste om många fel.

Knutsson (2001) påpekar att användning av denna utvärderingsmetod exklusivt, är otillfredsställande när det gäller att visa hur ett system för automatisk grammatikkontroll fungerar i praktiken, det vill säga för riktiga användare. Min egen utvärdering illustrerar detta problem, även om jag inte har varit i närheten av en utvärdering med användare (avsnitt 6).

3.6.3 Statistiska förutsättningar

Ett annat mer generellt fundament i kvantitativ metod handlar om storleken på undersökningsdata. Ju större mängd undersökningsdata, desto trovärdigare resultat. Mina undersökningsdata är inte omfångsrika. Större svenska andraspråksinlärarkorpusar i elektronisk format fanns helt enkelt inte när uppsatsarbetet initierades. Den elektroniska versionen av SSM-korpusen har växt under arbetets gång och det är ju till glädje för den som vill pröva min metod på ett större material. Om jag tycker att skribentens modersmålstillhörighet är ovidkommande för skrivfel i andraspråksinlärartext så skulle jag därför också kunna utvärdera mitt regelpaket på ett stort nytt textmaterial. Jag har emellertid valt att begränsa mig till den lilla kontrollgruppen, som är tänkt att spegla undersökningsmaterialet, och inte ta in fler delar av SSM-korpusen i utvärderingen.

4 Analys

4.1 Taggerns prepositionsanalys

Först som sist kan jag konstatera att Granskas tagger är stabil och inte ställer till med några stora problem vid regelskrivandet (avsnitt 3.3). Vid genomgång av den taggade hundrafyrtioen meningar långa testfilen upptäckte jag ingen otvetydigt feltaggad preposition. De enda prepositioner som ”feltaggats” var felstavade varianter av *i* (*J*), *på* (*pa*) och *mellan* (*medan*), en ihopskrivning (*efternyckeln*) samt konstruktionen *i morgon Kaffé* istället för *i*(eller *på*) *ett morgonkafé*. *I morgon* flerordstokeniseras och taggas som adverb. Fem partiklar hade taggats

som preposition – tre representationer av verbförbindelsen *tycka om* och en vardera av *känna till* och *hitta på*. Till sist fanns tre särskrivningar som isolerade prepositionen i sammansättningen (*under kläder*, *komma i håg* och *till baka*).

Jag gjorde alltså ingen sökning på vare sig subjunktions- eller partikeltaggar utan bara på prepositionstaggar. Den enkla sökregeln för identifiering av prepositioner returnerade hela textmängden, med preposition och prepositionstagg rödmarkerade i texten. Detta betydde att jag i praktiken läste all text i taggad version och hittade därför också en och annan feltaggning, oftast orsakad av återkommande ovanliga eller felaktiga konstruktioner i indata. När det kommer till valet mellan partikel- och prepositionstagg har jag tagit hänsyn till potentiella feltaggningar vid regelkonstruktionen. Jag kommenterar i det följande några typfall och ett specialfall av taggningsproblem.

4.1.1 Prepositioner och partiklar

I de flesta fall där det föreligger en tvetydighet vad gäller partikel eller preposition föredras prepositionstaggen, som i *Sonja tycker mycket **om*** <pp> *Göteborg* (men också *Sonja tycker **om*** <pp> *Göteborg*, liksom *Sonja känner **till*** <pp> *Göteborg*). Att den statistiska analysen väljer prepositionstagg på ***för*** i exempel (10) är inte så konstigt med tanke på dubbeldistributionen partikel/preposition. Megyesis (1996) sammanställning av partiklars och prepositioners distribution visar, inte oväntat, att *för* dominerar som preposition och *emot* dominerar som partikel.

(10) *Det finns bara två saker i mitt hemland, de som är **för** <*pp> *Amerika och de som är **emot** <*pl> *det, och alla som är **emot** <*pl> *är kommunister.*

Om två prepositionsord uppträder i serie tyckte jag att taggern konsekvent verkade välja partikel- framför prepositionstagg för första ordet, som i *hon stannade **till*** <pl> ***på*** <pp> *torsdag*. Jag skrev därför en sökregel för identifiering av taggsekvensen <pl> <pp> eller <pp> <pp> . Min teori kom på skam, det fanns sammanlagt 113 sådana sekvenser, varav fem utgjordes av multipla matchningar av samma sekvens (17), men jag hittade bara en enda korrekt prepositionskonstruktion som felaktigt partikeltaggats (11).

(11) *Att bo i kärnfamilj, storhushåll eller ensam är en fråga jag har tänkt på* <pl> *i* <pp> *flera veckor, månader och kanske år.*

Felaktiga konstruktioner har däremot många gånger lurat taggern

(12) *Jag längtade efter* <pl> *till* <pp> *dig*
(13) *Måste en läkare prata* <vb.inf.akt> *om* <pl> *för* <pp> *en patient om* <pp> *denne skulle snart dö, eller skulle det vara hemligt?*
(14) *En pojke som berättar om* <pl> *i* <pp> *fängelse.*
(15) *Han läser en bok som köpte på* <pl> *efter* <pp> *middag på bokhandlen.*
(16) *Om man kan läser bara samma tidning varje dag så blir man samma tanka på* <pl> *om* <pp> *tidning.*
(17) *Ofta mamma och jag tuvngna att försvuma hem* <pl> *i* <pp> *från* <pp> *över* <pp> *helgerna.*
(18) *Ja min man tillbaka, måste vi disktera med* <pl> *i* <pp> *lugna igen.*

4.1.2 Ett specialfall

I åtminstone ett fall har taggerns analys till synes gått över styr. ***åt*** i exempel (19) och (20) taggas som preposition.

(19) *Han **åt** lunch på tåget*

(20) *Jag åkte T-bana kl. kvart över åtta sedan jag hade druckit kaffe och **åt** en bulle på ett kafé.*

åt lunch på tåget med aktuell taggning, igenkänns alltså av Granska som prepositionsfras. Även i (21) blir ***åt*** taggat som preposition, men här är det syntaktiskt rimligt.

(21) *som vanlig tvättade han sig **åt** en hestig frukost och åkte bil till jobbet*

4.1.3 Potentiella flerordstaggar

Eftersom taggern inte gör någon syntaktisk analys avslöjas inte något samband mellan komponenterna i så kallade *flerordsprepositioner* (SAG) som ***i** stället **för*** och ***för** länge **sedan***. En extremt fast flerordsförbindelse som *till exempel* sammanförs av Granskas tokeniserare till ett enda token, som får adverbtagg. Flerordsprepositionerna är svårare att tokenisera som en enhet, dels på grund av att mittenenheten i många fall har varierande form, dels för att sekvensen inte alltid behöver vara en flerordspreposition (*Ställ cykeln i stället för cyklar! Begrunda inte skissen för länge sedan du en gång blivit färdig!*).

Ännu vanskligare är det med en förbindelse som ***för att*** i konstruktioner som *Sonja ville och flytta till Göteborg **för att** hon tycker om rör och trafiken*, som många grammatiker gärna analyserar ytsyntaktiskt i sin helhet som subjunktion. En flerordstokenisering av *för att* skulle emellertid resultera i många feltaggningar.

4.2 Generella iakttagelser och ställningstaganden

4.2.1 Hur svåra är prepositionskonstruktionerna?

Andraspråkslärare pekar entydigt ut ordföljd, verbböjning, kongruens, utelämnade ord och prepositionsbruk som sina studenters huvudsakliga grammatiska problemområden (Staerner 2001; Knutsson et al 2002). Svårigheter med verbtempus har ännu inte angripits i Granskasystemet, ordföljdsproblem (Staerner 2001) har utretts under ett par års tid och utgör ett svårbemästrat problem i ett system för svenska. Kärnan i Granskasystemet är analys av nominalfraser och kongruensfel. SSM-materialet visar på mycket utbredda kongruenssvårigheter vad gäller anaforisk referens (*Petterson och sin flick Anitta stannade där flera dager* eller *Det är inte svårt att anpassa mig till livet här i Sverige*) och systemet har förutsättningar att utvecklas mot en ganska träffsäker korrektion av dessa fel.

För prepositionsrelaterade fel har det som sagt inte funnits någon beredskap i Granskasystemet. Felaktigt prepositionsval har ganska hög frekvens i feltypologiska undersökningar på andraspråksinlärartext (Pitkänen-Koli 1990; Öhrman 2000). Min undersökning ger istället vissa indikationer om att prepositionskonstruktioner kanske inte vållar så stora problem i svenskan. Jag

har bland annat konstaterat en förkrossande övervikt för korrekt prepositionsbruk i mitt material när jag undersökt enskilda prepositioner och deras distribution och användning (avsnitt 4.3.2). Prepositioner som *på* och *i* hör till språkets vanligaste ord. Möjligen har detta bidragit till hög frekvens för prepositionsrelaterade problem i Pitkänen-Kolis och Öhrmans undersökningsresultat. I Pitkänen-Kolis fall är det dessutom (unga) finsktalande skribenters texter som analyserats. De semantiska relationer som i svenskan (och i engelskan och arabiskan) uttrycks med hjälp av prepositioner, uttrycks på ett annorlunda sätt i finskan (liksom i viss mån i japanskan), vilket möjligen skulle kunna vara en ytterligare förklaring till de många prepositionsproblemen.

4.2.2 Feldetektion, diagnos och korrektion

Skrivfelsmönstren i materialet avslöjar att de feltyper som Granska är särskilt utformat för att ta hand om – inkongruens i nominalfras och i predikativ samt särskrivningar – verkligen förefaller att dominera över andra typer av fel, vid sidan om inkongruens i anaforiska element. Vid granskning med Granska av enskilda uppsatser är det emellertid påfallande många fel som ändå inte detekteras. Mängden falska alarm är samtidigt ytterst liten. Iakttagelsen stöds av Öhrmans (2000) utvärdering av Granska på ASU-korpusen (Knutsson 2001:139). Regelkonstruktörerna har valt att sätta en hög tröskel för falska alarm, till priset av att missa detektionen av många fel som så att säga avviker från normen. Precision och täckning står i omvänt proportionellt förhållande till varandra (avsnitt 3.6.2).

Ett vanligt problem är att programmet upptäcker en felaktig konstruktion men ställer fel diagnos och ger fel rättningsförslag (**Jag läste på bankrånet när jag slog upp morgontidningen* kanske detekteras och korrigeras till **Jag läste i bankrånet när jag slog upp morgontidningen*). Ett sätt lösa problemet är att leverera flera rättningsförslag, där regeln returnerar den troligaste korrektionen överst i en alternativlista och den minst troliga underst. Man överlåter alltså det sista steget i revisionsprocessen till användaren, som i det anförda exemplet förhoppningsvis väljer programmets andra alternativ, som råkar vara *om*. Rangordningen ges genom till exempel sökning på verbkontexten i en standardspråkskorpus. Det är så här som rättstavningsprogram oftast fungerar, inklusive Stava, fast stavningsmodulen maskingenererar rättningsförslagen och deras inbördes ordning.

För en användare med svenska som andraspråk kan en sådan återkoppling med alternativa rättningsförslag vara särskilt konstruktiv i en undervisningssituation med tillgång till mänsklig handledning. Konkreta exempel på denna lösning är exempelregeln i avsnitt 4.3.2 och regeln `substpp3` i avsnitt 5.2.2.

4.2.3 Multipla fel i prepositionskonstruktion

Materialet är mycket rikt på multipla fel. En utesluten preposition kan till exempel ha sällskap av missformade ord och särskrivningar. Det betyder att om regeln som fångar den uteslutna prepositionen ska kunna detektera felet, så måste man ge den en uppsättning typiskt missformade kontexter, förutom den korrekta, som indata. Jag uppfattar detta som ett centralt problem för CrossCheck och en viktig faktor att hålla i huvudet vid regelkonstruktionen. Begrunda följande exempel (22) jämte den taggade versionen (23).

(22) *After sex månader* ***comde Jag till båcka att förtset*** *min studing*
(23) *After* <pm.nom> *sex* <rg.nom> *månader* <nn.utr.plu.ind.nom> *comde* <vb.prt.akt> *Jag* <pn.utr.sin.def.sub> *till* <pp> *båcka* <nn.utr.sin.ind.nom> *att* <sn> *förtset* <nn.neu.sin.def.nom> *min* <ps.utr.sin.def> *studing* <nn.utr.sin.ind.nom>

En regel som detekterar utesluten preposition med vänsterkontext bestående av verbfras följd av vissa partikel- eller adverbtaggade element (hem, tillbaka...), och med infinitivfras i högerkontext, är naturligtvis chanslös på exempel (22). Jag har konstruerat regler som detekterar den här typen av multipla fel, som tar en hel uppsättning av specifika ordformsfel i kontexten som indata. Men det blir orimligt betungande för regelkonstruktören att arbeta med kontexten på det här viset. Dessutom kan det bli svårt för användaren att förstå vari felet låg, när hennes/hans text i rättningsförslaget är korrigerat till oigenkännlighet! Därför förefaller det bättre att konstruera reglerna så att regeln för den uteslutna prepositionen i exempel (22) träder i kraft först sedan användaren (idealt) med hjälp av programmet korrigerat särskrivningen, men också *båcka*, samt det missformade infinita verbet. [5]

Stavfel i särskrivningar måste ändå i allmänhet tas om hand simultant i regeln för särskrivning. Den särskrivna formen av *tillbaka* råkar kvalificera för analys i min studie, eftersom den producerar en preposition. Granskas befintliga regeluppsättning innehåller en specialregel för just *till baka*, men regeln tar ingen hänsyn till felstavat för- eller efterled. RWE:s som *bocka*, *boka* och *backa* passerar odetekterade och *båcka* korrigeras av Stava till just *bocka* eller *backa*. *Tillbaka* är ett högfrekvent ord och därför anser jag att det är motiverat att anpassa sig till variation i stavning vid regelkonstruktionen.

Sett ur ett annorlunda perspektiv, som analyserar en dynamisk inlärningsprocess (Hammarberg 1977; Knutsson et al 2002:3), är sekvensen *till båcka* i själva verket en indikation på att skribenten tillgodogjort sig en viktig fonologisk distinktion i målspråket, som handlar om vokalkvalitet, och där stavningen *båcka* är en lösning för att representera skillnaden mellan de båda vokalljuden.

4.2.4 Prepositionsfel utanför analysen

En mycket avvikande syntaktisk struktur i indata kan bereda stora problem för ett system med automatisk ordklassigenkänning, men också för en mänsklig läsare. Missformade och svårtolkade satser sorterar i min analys under semantiska fel (24) – (27). De utgör konstruktioner som enligt min mening varit omöjliga att vidare systematisera inom ramen för min undersökning, än mindre att ange matchningsvillkor för i någon granskningsregel .

(24) *Men nu vanade jag ganska vid här*
Men nu är jag ganska van vid att vara här (tolkningsförslag)

(25) *Om man kan läser bara samma tidning varje dag så blir man samma tanka på om tidning*

[5] Problemet med multipla fel aktualiserar egentligen frågor om programmets funktionalitet – ska granskningen vara en statisk engångsrevision eller ska regelexekveringen börja om från början efter varje ändring som användaren gör enligt programmets rekommendationer? En utförligare diskussion kring interaktion med användaren ryms emellertid inte inom ramen för uppsatsen.

(26) *Under jag var bussen tänkte jag annan sak utan svenska*
Under tiden jag var på bussen tänkte jag på allt utom på svenska (tolkningsförslag)

(27) *Då kom om idags skrivning till mitt huvd*

Kategorin semantiska fel i prepositionskonstruktion är störst för de japanska skribenterna – 38 stycken, mot 28 för arabisktalande och 23 för engelsktalande skribenter – och denna kategori är som sagt inte representerad i kontextfönstren.

Många felkonstruktioner med prepositionen *av* som centralt element, utgörs av genitivkonstruktioner. Av sammanlagt 19 kontextfönster med *av* eller felstavningen *ov* i mitten, är sju stycken alldeles klart genitivkonstruktioner av interferenskaraktär. Jag bedömer genitivkonstruktion med *av* som alltför svår att fånga kontextuellt och kommer inte att behandla feltypen i det följande.

4.2.5 Prepositionsfelmängden

Felsökningen har resulterat i 400 kontextfönster, 116 från japanska skribenter, 158 från de arabiska och 126 problem i prepositionskonstruktion från engelsktalande skribenter. Från japanska och arabiska skribenter har jag extraherat precis lika stor felmängd relativt hela textmängden för motsvarande modersmålsgrupp. Engelsktalande skribenter bidrar med något mindre relativ prepositionsfelmängd. Den relativa storleksrelationen mellan de tre felmängderna är 1 : 1 : 4/5. Hela detta material om 400 observerade problem i prepositionskonstruktion, kommer jag fortsättningsvis att referera till som *felmängden* eller *kontextfönstren*.

4.3 Ersättning

4.3.1 Fasta och mindre fasta förbindelser

Prepositionsvalet är mycket ofta beroende av referenternas identitet och inbördes förhållande, och ibland kan det det tvärtom vara referenterna som påverkas av prepositionsvalet. Ofta resulterar felaktigt prepositionsval i en ganska subtil betydelseglidning (28), ibland i mer akut ogrammatiskhet (29).

(28) *Han föll rakt ner **på** ett träd **i** gården och fick två sår **i** västra armen*
(29) *Jag tycker att man bara kan förstår Vad hända i hela värld, när man **läser på tidningen***

Arbetar man på tidningen är ***på*** korrekt prepositionsval.

Prepositionsvalen i exempel (28) och (29) är beroende av kontext inom prepositionskonstruktionen. Men ibland är prepositionsvalet beroende snarare av diskursiv än syntaktisk eller semantisk kontext (30).

(30) ***I söndags gick jag med min vän som heter Per tillsammans in till centrum stan, och tittade på fönstren***.

Ett fel som ***på*** *fönstren* i exempel (30) kan omöjligt upptäckas i ett system för automatisk grammatikkontroll, med mindre än att man inför moduler för pragmatisk analys, och även sedan man till äventyrs lyckats med detta kan man inte avskriva en läsning av (30) som tar i beaktande att skribenten och hans vän har ett specialintresse för fönster.

Jag ska försöka visa hur godtyckliga de lexikala bindningskraven kan vara (Tabell 7), och hur svårt det kan vara att generalisera felaktigt val av preposition, till och med om prepositionen rentav fått fel form och blivit partikel.

Tabell 7. Exempel på svårigheter med att ta hjälp av kontexten för att åtgärda felaktigt prepositionsval. Felkonstruktionerna är autentiska exempel från mitt SSM-material.

	Prepositionskonstruktion		
Ogrammatisk	*läser* <vb>	*på* <pp>	*tidningen* <nn >
Grammatisk	*läser* <vb>	*på* <pp>	*mjölkpaketet* <nn >
Grammatisk	*Arbetar* <vb>	*på* <pp>	*tidningen* <nn >
Ogrammatisk	*studerar* <vb> *Jag* <pn>	*i* <pp>	*Universtet* <nn>
Ogrammatisk	*gick* <vb> *jag* <pn>	*vid* <pp>	*univestit* <nn>
Grammatisk	*studerade* <vb> *jag* <pn>	*i* <pp>	*sängen* <nn>
Grammatisk	*gick* <vb> *jag* <pn>	*i* <pp>	*skolan* <nn>
Grammatisk	*studerade* <vb> *jag* <pn>	*vid* <pp>	*universitetet* <nn>
Grammatisk	*pluggade* <vb> *jag* <pn>	*på* <pp>	*universitetet* <nn>
Grammatisk	*skolkade* <vb> *jag* <pn>	*från* <pl>	*skolan* <nn>
Ogrammatisk	*studera* <vb>	*in* <pl>	*T.H.* <pm> *Skolan* <nn>
Grammatisk	*studera* <vb>	*in* <pl>	*materialet* <nn>
Grammatisk	*skola* <vb>	*in* <pl>	*barnet* <nn>

Bindningskraven för de olika prepositionsförbindelserna i tabell 7 är inte särskilt starka. Regler för felaktigt prepositionsval kan bara komma ifråga för misstag i tillräckligt fasta (lexikaliserade) förbindelser och idiom, och måste alltså med nödvändighet bli mer eller mindre specifika.

4.3.2 Man får se sig omkring – kontextanalys

Vid sökning i hela uppsatsmaterialet på högfrekventa prepositioner är det i själva verket förbluffande hur sällan det har blivit fel i prepositionskonstruktionen. Om man istället koncentrerar sig på kontexten kring de misstag som trots allt begåtts vad gäller själva prepositionen så framträder ibland mönster, till exempel i omedelbar vänsterkontext. Fortsatt sökning efter detta mönster har i vissa fall visat sig effektiv. Det verkar med andra ord vara mer konstruktivt att examinera **referenterna** kring felaktigt prepositionsval.

Låt oss titta på prepositionen *om* och dess omgivningar i textmaterialet. Ordets trippelfunktion som preposition, subjunktion och partikel tycks inte bereda några större svårigheter för uppsatsskribenterna. Felkonstruktioner kring ordet *om* visar sig nästan uteslutande höra ihop med antingen verbet *tänka* i prepositionsförbindelse (*tänka på*), eller något av verben *diskutera* och *debattera*. De senare behandlas i avsnitt 4.4 nedan. Verbet *tänka* är en typisk representant för en grupp högfrekventa lexikala enheter som representerar mycket grundläggande abstrakta begrepp, och som kan uppträda i tusen och en skepnader, och man kunde mycket väl föreställa sig att regelskrivande för sekvenser som **Pelle tänkte om Lisa hela natten* antingen kostar för många falska alarm eller också kostar för mycket programkod, i förhållande till resultatet. Men sökning på *tänk* i felmängden (och för säkerhets skull också på *tank*, men utan träff) visar ett ganska entydigt mönster av det aktuella verbet, följt av prepositionen *om*, följt av ett substantiv.

```
X1(lemma="tänka"),
X2(text!="på"),
X3(wordcl=nn | wordcl=pn | wordcl=pm | wordcl=dt)
```

Sökning på samma sträng i hela uppsatsmaterialet komplicerar analysen något. Bland annat finner man *Tänk om alla i Sverige talade franska!* Vid förekomst av till exempel satsadverbial i satsen förändras mönstret också, på grund av ändrad ordföljd. Vid regelskrivningen bör man också ta hänsyn till partikelverbet *tänka om* (tryckstarkt *om*). Nu blir Granskas prepositionstaggning intressant. För att undvika att regeln detekterar partikelförbindelsen, duger det inte att bara ange prepositionstagg som matchningsvariabel, eftersom taggern gärna väljer prepositionstaggen också för partiklar, och *om* är vanligast som preposition (Megyesi 1996). Därför gör man klokt i att begränsa högerkontexten till nominalfraser och nominala infinitivfraser, så att regeln inte detekterar *Jag fick tänka om och skriva en ny uppsats*, men däremot **Pelle tänkte bara om att komma hem*.

Jag inbjuder läsaren att begrunda de olika kontexter i vilka uppsatsskribenterna gjort fel med verbet *tänka* i prepositionsförbindelse (inkorrekt prepositionsval i fetstil):

*Sedan tänkte han **om** sin hund*
*jag har tänkt mycket **med** skrivningen*
*Om man tänker **om** halsa och trivsel, är det bättre för koppen om man går eller cyklar*
*Sedan tänkade hon **om** hur kunnde hon åka till biografen.*
*De tänker **om** nagon mjörlig sätt (metod) för att öppna dörren*
*kvinnor tänker **om** problemet på ett särskilt sätt*
han hade mycket tid i det kallt lilla hus för att tänker varför han var där
*jag hade tänkt **om** snön och klimatiserat mig vid klimet.*
*Du tänker **om** centern i Stockholm.*
Då tänker de ett sätt för att lösa sitt problem.
*Han tänkte **om** kläder*
*Han tänkte **om** hans syster som bor i Göteborg*
Under jag var bussen tänkte jag annan sak utan svenska.
De tänker vad de kan göra
*Han tänker alltid **om** electric apparater.*

Av 11 felaktiga prepositionsval är det faktiskt 7 som matchar variabelsekvensen ovan. För att detektera modifierad ordföljd måste man alltså ange optionell nominalfras eller adverbkedja mellan verbet och prepositionen. Prepositionen *med* är i själva verket mycket olämplig att matcha, den är en säker fälla för falska alarm i det här fallet, så *med* lyfts ur variabelsekvensen. Däremot får *med* plats bland rättningsförslagen. Men hur troligt är det att skribenten ville skriva *tänka med*? Det bästa man kan göra är en frekvensanalys i normalspråkskorpus. En sökning i SUC visade att *på* var sex gånger vanligare direkt efter någon form av verbet *tänka*. Någon annan preposition efter verbet förekom inte. Nedanstående regel tjänar som exempel och är inte tänkt att läsas som någon färdig regel, därför har den inte heller vare sig namn eller kategoritillhörighet.

```
{
X1(lemma="tänka" & vbf!=imp),
(ADV)()?,
(NPall)()?,
X4(wordcl=pp & text="om"),
(NPall/X5)()
-->
mark(X1 X4)
corr(X4.replace("med"))
corr(X4.replace("på"))
info("Du har nog valt fel preposition till verbet"
italics(X1.real_text))
action(scrutinizing)
}
```

Programmet returnerar rättningsförslagen i omvänd ordning relativt `corr`-raderna i regelkod, så korrektion med *med* som rättningsförslag anges först i regeln. För att fånga möjlig alternativ ordföljd anropas en hjälpregel för adverbkedjor, och en för nominalfraser optionellt (frågetecknet), och så återstår bara att blockera detektion av imperativformen *tänk*. `info`-fältet motsvarar diagnosen. Jag har i mitt regelskrivande strävat efter att leverera så hövliga, begripliga och pedagogiska diagnoser som möjligt.

I princip kan man fånga även utebliven preposition genom att sätta frågetecken efter matchningsuttrycket för variabel `X4`, men då måste man markera andra tokens, programmet kan givetvis inte markera något som eventuellt inte finns. Dessutom måste man skriva om `corr`-fältet, så att metoden i händelse av utebliven preposition har ett token att utföra operationen på. Man kan till exempel välja att sätta in prepositionen framför nominalfrasen, med metoden `insert`. Om man matchar även utebliven preposition ökar risken för falska alarm avsevärt. Slutsatsen måste bli att man gör klokt i att ta hand om bortfall av preposition i en annan regel (avsnitt 4.4).

Följande konstruktioner är exempel på förbindelser i samband med felaktigt prepositionsval, som också blivit föremål för regelkonstruktion.

> *Jag vet ingenting om jag ska ble* ***interessare för*** svensk språk
> *Han* ***var framme till Göteborg*** *klockan 14,30*
> *Han* ***kom tillbaka i Stockholm*** *klockan halv tre.*
> *Vi kan simma* ***i vinteren***
> *Hans fru och tre barn kom och vinkade av honom* ***i Central stationen****.*
> *Människor driker vodka för att kanske undvika att tänka på andra saker och*
> [forts…] *politiksystemet låter det för att* ***sitta bättre i makten****.*

4.3.3 Förväxlingar och missuppfattningar av idiomatiska uttryck

Vid användning av utpräglat idiomatiska uttryck förekommer inte så sällan sammanblandningar med andra liknande idiom (*för mestsdels, till en stor grad*). Särskilt gäller detta idiom som man gärna ser flerordstaggade som adverb (precis som *till exempel*). Dessa omständigheter gör att feltypen lämpar sig särskilt väl för analys med hjälp av *förväxlingsmängder*, i likhet med stavfel och särskrivningar, och därför utreds feltypen i avsnitt 4.5.

Övriga idiomfel i materialet är antingen diskursbaserade (*plus åtskilliga utländska underrättelsetjänster som jobbar i Sverige **på fri fot***), eller på annat sätt mycket svåra att kontextuellt separera från korrekta förekomster av identiska sekvenser, eller sekvenser som blir korrekta under restaureringsprocessen mot korrekt idiomform. Det var med andra ord kanske inte ett idiom man hade att göra med till att börja med (Tabell 8).

Tabell 8. Exempel på lyckad respektive misslyckad tillämpning av idiomregel.

	Idiomatisk sträng	**Icke idiomatisk sträng**
Ursprungssekvens	**Dålig utbilding **ligger på grunden för** de flesta.*	?*Lastfartyget **ligger på grunden för** styrman var full.*
Restaurering, steg 1	**Dålig utbilding **ligger på grund för** de flesta.*	*Lastfartyget **ligger på grund för** styrman var full.*
Restaurering, steg 2	*Dålig utbilding **ligger till grund för** de flesta.*	***Lastfartyget **ligger till grund för** styrman var full.*

4.4 Borttagning och insättning

4.4.1 Den tomma mängdens feltyp

Kategorin *borttagning* är stor, med 63 förekomster är det den näst vanligaste ”prepositionen” på position 4 (mittenpositionen) i mina kontextfönster. [ ] och *i* toppar statistiken på position 4 med sammanlagt 135 förekomster. Det betyder att i drygt en tredjedel av alla kontextfönster ockuperas prepositionsplatsen av någon av dessa båda. Vid analys av var modersmålsgrupp för sig upptäckte jag att över en tredjedel av prepositionsfelen hos de japansktalande skribenterna utgjordes av utebliven preposition. Man kunde mycket väl tänka sig att fenomenet hade sin grund i att japanskan som helhet står typologiskt långt från svenskan, i jämförelse med såväl engelskan som arabiskan. Men då borde jag också ha hittat generellt fler prepositionsfel hos japansktalande skribenter, och det har jag inte gjort (däremot är kategorin *semantiska fel i prepositionskonstruktion* större för de japanska skribenterna (avsnitt 4.2.4)).

Klassen *borttagning* har nu den egenartade egenskapen att den inte upptäcks av sökregeln för prepositionsidentifiering, helt enkelt därför att felet är lika med avsaknad av just det element sökregeln ska hitta. Alla förekomster av feltypen var därför borttagningar som jag så att säga snubblat över vid genomgång av materialet. Jag insåg därmed också att ”mörkertalet” kunde vara mer eller mindre stort. Jag prövade att extrahera prepositionskontexter i SUC för att om möjligt få ett sökmönster för utebliven preposition. Det överlägset vanligaste taggtrigrammet med preposition i mitten var <NN> <PP> <NN>, det vill säga prepositionen flankeras av substantiv (taggen anges med versaler i SUC). Jag sökte följaktligen i hela mitt SSM-material på bigrammet <nn> <nn>, men hittade mest genitivkonstruktioner och särskrivningar, och ingen utelämnad preposition. Därför gjorde jag istället en kontextanalys för de 53 utelämnade prepositioner jag dittills hittat. Resultatet visade inte något entydigt kontextmönster för feltypen, däremot var avvikelsen från SUC-korpusens trigram påfallande (Tabell 9).

Tabell 9. Närkontext för utelämnad preposition i undersökningsmaterialet och närkontext för preposition i SUC – de tolv vanligaste sekvenserna.

Närmaste taggar kring utelämnad preposition i felmängden (53 trigram)				**Närmaste taggar kring preposition i SUC (100 700 ord, 37 500 trigram)**			
Absolut frekvens	**Trigram**			**Absolut frekvens**	**Trigram**		
4	vb	<[pp]>	ps	2290	NN	PP	NN
4	vb	<[pp]>	nn	879	VB	PP	NN
3	vb	<[pp]>	ie	786	NN	PP	DT
3	pn	<[pp]>	ie	613	NN	PP	PM
3	pl	<[pp]>	pm	498	NN	PP	JJ
3	nn	<[pp]>	ie	435	AB	PP	NN
2	vb	<[pp]>	pn	354	VB	PP	DT
2	vb	<[pp]>	dt	281	MAD	PP	NN
2	pl	<[pp]>	nn	246	PN	PP	NN
2	nn	<[pp]>	nn	182	VB	PP	JJ
2	ab	<[pp]>	nn	168	MAD	PP	DT
1	vb	<[pp]>	vb	155	VB	PP	PM

Det vanligaste trigrammet för utelämnad preposition i felmängden kom först på 30:e plats bland SUC-trigrammen med 82 förekomster (vilket motsvarar en relativ frekvens på drygt 0,2 procent). Jag sökte på sekvensen verb-possessiv i hela mitt material och ”hittade” ytterligare fyra utelämnade prepositioner hos japansktalande skribenter och en hos engelsktalande skribenter.

Jag sökte också på ordkontext. Vanligaste ordet efter utelämnad preposition var *att*, vilket också avspeglas i tabellen (taggen <ie> är infinitivmärket). Jag hittade på det viset ytterligare två borttagningar hos engelsktalande skribenter. Dessutom hittade jag tre kontextfönster där jag själv utelämnat utelämnad preposition (jag hade glömt representera enheten med hakparenteser). Tillsammans med dessa restaurerade kontextfönster hittade jag ytterligare tio fall av utelämnad preposition. Sökningsförfarandet var mycket tidskrävande och kontexten för feltypen visade som sagt inget entydigt mönster, så jag bestämde mig för att nöja mig med vad jag funnit.

4.4.2 Infinitivmärke efter prepositionsenheten

Analys av hela högerkontexten avslöjade en del intressanta mönster. Det visade sig att samtliga förekomster av ordet *att* efter utelämnad preposition var infinitivmärken omedelbart följda av ett infinit verb. Tre av dessa *att* var feltaggade på grund av felaktig verbform (Tabell 10). Ordet *att* är över huvud taget det vanligaste ordet i kontextfönstren på position 5 (omedelbart efter prepositionen).

Tabell 10. *att* efter borttagning och *att* efter annat prepositionsfel, med högerkontext.

Utelämnad preposition			**Närvarande men felaktig (eller felaktigt närvarande) preposition**		
Bort-tagning	***att*+tagg**	**Högerkontext**	**Pre-position**	***att*+tagg**	**Högerkontext**
[ ]	*att* <sn>	*koma tillbaka*	*på*	*att* <sn>	*alla skulle*
[ ]	*att* <sn>	*förtset min*	*om*	*att* <sn>	*det inte*
[ ]	*att* <sn>	*födda barn*	för	att <sn>	*min föraldrar*
[ ]	*att* <ie>	*vilja få*	för	*att* <sn>	*jag kom*
[ ]	*att* <ie>	*titta på*	för	*att* <sn>	*jag har*
[ ]	*att* <ie>	*tala om*	*för*	*at* <pm.nom>	*läser sin*
[ ]	*att* <ie>	*tala om*	*för*	*att* <ie>	*studera .*
[ ]	*att* <ie>	*ta lunch*	*för*	*att* <ie>	*leva och*
[ ]	*att* <ie>	*skriva några*	*för*	*att* <ie>	*läsa och*
[ ]	*att* <ie>	*se "*	*för*	*att* <ie>	*läsa i*
[ ]	*att* <ie>	*samla informationen*	*för*	*att* <ie>	*komma in*
[ ]	*att* <ie>	*köpa biljeterna*	*för*	*att* <ie>	*hitta mina*
[ ]	*att* <ie>	*häsa på*			
[ ]	*att* <ie>	*göra så*			
[ ]	*att* <ie>	*gå på*			

Som vi ser i tabell 10 så är infinitivfraser vanliga också efter närvarande preposition. Tre *för* utan påföljande infinitivsats har bildats av särskrivningar av *därför* (avsnitt 4.5). Alla *för* i tabellen följs därmed av infinitivsats. Även det finita *läser* är nämligen tänkt som infinitiv (31). Samtliga *för* är dessutom motsatsen till borttagningar, alltså insättningar.

(31) *Han känner sjalv mycket battre och kan nu försatte för at läser sin boken.*

Hur är det med vänsterkontexten? Kontextfönstren förslår inte riktigt för att se A-referentens beskrivning. När det gäller *för* är vänsterkontexten mycket svåranalyserad med komplexa verbkonstruktioner som i exempel (31). Vid utesluten preposition rör det sig ofta om en stympning av subjunktionsförbindelsen *för att* efter vissa verbfraser, eller så fungerar infinitivfrasen ogrammatiskt som direkt objekt till intransitivt verb.

After sex månader comde Jag till båcka [ ] *att förtset min studing,*
Ante bestämmde sig [ ] *att koma tillbaka.*
Det var fridag kväll och jag hade bestämt mig [ ] *att gå på bio.*
Han bestämde sig [ ] *att ta lunch på tåget.*
Sen gick han [ ] *att se "Hamlet"*
19.30 gick han på stadsteatern [ ] *att titta på "Hamlet"*
De som väntar [ ] *att köpa biljeterna undrar vad hon gör*

4.4.3 Transitivitetsproblem

Om vi ser till feltypen borttagning som helhet så består taggarna närmast till vänster om prepositionen av finita verb (23 stycken), partiklar eller adverb (12 stycken), substantiv (10 stycken) och pronomen (7 stycken). Felaktig transitivitet är här ett mycket påtagligt problem, dels vid partikelverb (*tillbaka* dominerar som partikel), reflexiva verb och verb som ibland kallas för *prepositionsverb* (*kalla på, vänta på, prata med, ropa på, bidra med, vissla på, känna för, börja med*) [6], dels vid ”vanliga” intransitiva verb (*sova, ligga, skolka, bo, tänka, komma, sitta*) samt vid deponensverbet *trivas*.

Omständigheterna inbjuder till formalisering, men skenet bedrar. För det första är det mycket vanligt med rent idiosynkratiska fel, som i fallet *trivas* – tre gånger förekommer transitiv användning av *trivas.* Dessa härrör från en och samma skribent. Verbet används härutöver tolv gånger korrekt. Sekvensen *Prata varandra* hittar man också bara hos en skribent. Hos samma skribent finner vi även sekvensen *prata med varandra*. Verbet *prata* med böjningsformer förekommer ytterligare 55 gånger i materialet, med två unika problem i prepositionsförbindelse.

För det andra är transitivitet en mycket svår egenskap att laborera med i kontextvillkoren, åtminstone så länge en syntaktisk analys på satsnivå inte är implementerad. Det är mycket få verb som uppträder någorlunda entydigt i det här avseendet och svenskans ordföljdsregler är svåra att ta sig förbi. I mina kontextfönster förekommer finita former av verbet *vänta* 6 gånger omedelbart framför utesluten preposition. I svenskan används *vänta* på många sätt, även transitivt i många sammanhang och ibland fungerar det bara med direkt objekt (*vänta barn*). En regel som upptäcker transitivt *vänta* och blockerar indata av specifika ord som *barn*, kräver en ganska komplex accepterande regel som släpper igenom konstruktioner med omvänd ordföljd. Min granskningsregel detekterar även korrekta sekvenser som *Jag **väntade** hela natten **på** dig* eller *Hon **väntade** varje kväll **på** att han skulle höra av sig* (just ordet *varje* har redan en egen accepterande regel i Granska, men den regeln ingår inte i mitt regelpaket). Min accepterande regel tar hand om till exempel spetsställd prepositionsfras och frågekonstruktioner som i *Utanför hotellet **väntade** bussen **på** delegaterna* eller *Hur länge **väntade** jag **på** bussen?*

Partikel- och reflexivverben är trots allt något tacksammare. Jag har skrivit en regel som matchar taggsekvensen verb – reflexivpronomen – infinitivfras (regeln `delpp4` i avsnitt 5.2.2) och som alltså detekterar sekvenser som ***Jag bestämde mig att gå hem**. För reflexivverben har subjektet en tendens att vid omvänd ordföljd smyga in mellan verbet och reflexivpronomenet, och därför detekteras inte en korrekt konstruktion som *När **bestämde du dig** för att sluta röka?* (jämför *Hur länge **väntade du** på bussen?*). Visserligen detekteras då inte heller **När **bestämde du dig att sluta röka?***, men missarna blir förhoppningsvis få. Det finns en rad hjälpregler i Granska för att identifiera olika typer av verbkonstruktioner, bland annat just reflexivförbindelser av typen *bestämma sig*. Jag har vid regelskonstruktionen saxat ur bland annat denna hjälpregel (om jag anropade hela hjälpregeln skulle även utpräglat transitiva reflexivverb komma att tillhöra villkorsmängden). Jag har också kompletterat med verbförbindelser efter sökning i SUC på sekvensen verbtagg – pronomentagg.

[6] Termen *prepositionsverb* antyder en stark förbindelse mellan prepositionen och verbet och frågan är om inte prepositionen i vissa fall är hårdare knuten till verbet till vänster än till sin rektion – prepositionen hör hemma i både verbfras och prepositionsfras. Jag diskuterar detta fenomen i avsnitt 7.2.

Vid otillbörlig användning av preposition (*insättning*) råder motsatt förhållande, det vill säga verbet används intransitivt när kontexten blockerar en sådan användning, men mönstret är inte lika tydligt. Gemensamt för alla fel som klassats som insättning är att skribenten har använt preposition där det inte skulle vara något alls. Korrektion av ett sådant fel innebär ingenting annat än att en borttagningsoperation utförs på prepositionsenheten. Denna omständighet är en lättnad för regelskrivaren, som plötsligt kan mata in en ganska stor mängd felkonstruktioner i vänsterledet som får samma diagnos och korrektionsoperation (borttagning) i högerledet. Om en skribent gjort ett transistivt verb intransistivt genom att använda preposition, så spelar det ingen roll för korrektionen vilken preposition som använts och inte heller om prepositionen fått fel form (32) – (42).

(32) *Efter en halv timme ringde han till företaget och frågade vilken tid han skulle* ***träffa med ingenjör Karlsson***
(33) *De kunde inte försätta kriget i Vietnam då måste de* ***hindra med det****.*
(34) *Det är ganska nytt att man* ***debatterar om köns-rollerna****.*
(35) *Kissinger åkte många gånger till nord Vietnam för att* ***disktera om slut*** *av kriget.*
(36) *Jag* ***träffade i svenska Ampasador*** *i London för min visa.*
(37) *försök att* ***hjälpa till dem*** *som behöver*
(38) *Kremlmuren och andra särskilda plater* ***gav till Jersild*** *en ideé om auktoritet*
(39) *Jag* ***takade till henne*** *och gav blommor.*
(40) *Kan du* ***bjuda på mig****?*
(41) *Hon blir orolig därför att hon inte kan* ***hitta på nycklen****.*
(42) *Han* ***åt till lunch*** *på tåget klockan halv ett.*

Givet att Granskas moduler i samarbete med användaren lyckas restaurera felstavade verb, kan alla dessa konstruktioner i princip detekteras och korrigeras av en granskningsregel som reagerar på ospecificerad preposition efter transitivt verb, följd av nominalfras. Men redan vid en snabb reflektion inser man att också mängder av korrekta konstruktioner skulle detekteras av regeln, särkilt om förbindelserna i exempel (40) till och med (42) integrerades i matchningsvillkoren. Många verb uppvisar som vi redan sett en mycket dynamisk distribution vad gäller transitivitet. För korrektion av (40) förslår inte syntaktiska kontextmatchningar ens på satsnivå.

Även verbet *diskutera* har en kluven distribution. Jag har i SUC hittat verbet i flera prepositionsförbindelser och nästan en tredjedel av verbets förekomster är framför preposition. Verbet *debattera* borde uppvisa ett liknande distributionsmönster men det råkar förekomma noll gånger i mitt utsnitt av SUC-korpusen.

Jag avslutar denna genomgång med ett textutsnitt från mitt material, som visar hur svårtämjd verkligheten kan vara även när det gäller transitivitet.

Hunden söver.
– – –
Han är glöma söva. läsar bok.

4.4.4 Insättning i tid och rum

För övrigt kan man identifiera två insättningsfel som återkommer i materialet, dels framför rumsadverb

> ***I hemma*** *talar vi turkiska.*
> *Ibland hämtar de med mat* ***till hem****.*
> *Han tittar* ***på ute****.*
> *Sedan tog jag en buss* ***till hit****.*

dels i samband med tidsangivelser

> *Lasses gick till jobbet* ***i tisdag*** *den 15 Februari kl 8 50*
> *J tisdags den 15 februari 1973,* ***i klockan 8.50****, drack Lasse kaffe på jobbet.*
> *och sen till Sverige med flygplan* ***i 27 August***
> *Jag gick i skolan* ***i år 1972***
> ***I tisdag*** *den 15 februari tänkte Lasse att resa till Götebog*
> *Det fanns* ***i 1970*** *i Stockholm en stor demonstration emot privatbilism...*
> *en gammal byggnad* ***från i mitten 1800*** *talet*
> *Fängelset var nybyggt* ***i 1847***

4.5 Ordformsfel

4.5.1 Fördelar med klassificering i förväxlingsmängder

Mina *förväxlingsmängder* är som sagt (3.4.4) ingalunda några äkta *confusion sets*, men min tillämpning av metoden har ickedestomindre varit till stor hjälp vid klassificering av feltyper och vid regelskrivning, i synnerhet för formfel, vars förväxlingsmängder i praktiken listas i granskningsregelns vänsterled, under det att den avsedda standardformen återfinns i högerledet som rättningsförslag.

Genom att handgripligen rita förväxlingmängder efter identifiering av olyckliga prepositionsformer i kontextfönstren, blev formfelens relation till standardformen tydligare. Inte minst blev det tydligt hur ofta problemet bakom stavfelet tycks vara fonologiskt betingat, som vid förväxling av *av* med *ov* eller *mot* med *måt* eller *utan* med *ettan*, eller den egenartade förväxlingen av *är* med *i*. Jag tror att man skulle vinna mycket på en noggrannare analys av fonologiska stavfel, och inte bara i en applikation speciellt för andraspråksinlärare (Eeg-Olofsson 2002:31).

I flera fall har jag genom att generalisera ur de förväxlingar jag faktiskt hittat, lagt förväxlingsformer till den autentiska förväxlingsmängden (Tabell 11, 12 och 14). Tack vare en sådan komplettering detekterade formreglerna till exempel en icke observerad felstavning av *tillbaka*. Ett annat exempel på hur förväxlingsmängdsmetoden hjälpt mig är formfelsklassificeringen av idiom med prepositionellt innehåll som uppträder som adverb (Tabell 11). Det har huvudsakligen rört sig om ganska klassiska kontaminationer av typen *till en stor grad*. Regelstrukturen passar bra för att ta hand om den här typen av flerordsenheter. Man kan välja mellan att ge ett eller två korrektionsförslag (*i hög grad* och *till stor del*).

Tabell 11. Missuppfattningar av idiom med prepositionellt innehåll och ordklassidentitet som adverb.

Idiom	**Autentisk förväxlingsmängd**	**Reviderad Förväxlingsmängd**
för det mesta	*för mestsdels* *det mesta*	*för mestsdels* *för mestadels*
tills vidare	*till vidare*	*till vidare* *til vidare*
i hög grad	*till en stor grad*	*till en stor grad*
till stor del	*till en stor grad*	*till en stor grad* *till en stor del*
för sista gången	*för den sista gången*	*för den sista gången* *för den sista gång*
till exempel	*till exampel*	*till exampel* *till example* *til exampel* *til example*
på så sätt (?)	*i sådat satt*	

4.5.2 Real word errors

Förväxling av *från* med *fran* eller *på* med *pa* kan ha sin grund i rent ”slarv” med diakritiska tecken, men kan lika gärna ha fonologiska orsaker. Felstavningen *pa* är mycket vanlig – 12 förekomster på position 4 i kontextfönstren – och detekteras och korrigeras naturligtvis korrekt av Stava. Vid förväxlingar som verkar vara interferenser (*over* istället för *över* eller ännu hellre förväxling av *från* med *from*) kan det vara ännu svårare att avgöra om förväxlingen beror på semantisk eller kanske ortografisk eller fonologisk likhet. I fallet *from* kompliceras också *korrektionen* av att förväxlingen resulterar i ett RWE.

Mina förväxlingsmängder innehåller både okända ord och RWE:s. Korrektion av okända ord är okomplicerad och kräver inga kontextvillkor. I de fall vänsterledet alltså innehåller något RWE, till exempel *from* eller *for*, blir det nödvändigt att ange kontextvillkor i vänsterledet, för att inte regeln ska detektera *Han var en from själ* och *Där for hästskjutsen med min älskade*. I fallet *from* är det som vi ser ganska svårt, särskilt om taggern har svårt att skilja mellan artikel och promomen (**Han var en from Själ* är felaktig om *en* är pronomen. **Han var en from själ och hjärta snäll person* verkar vara ett felaktigt prepositionsval). Betydligt värre är det ändå med elementet *fram*, som tillhör samma förväxlingsmängd. I det här fallet var det så svårt att eliminera mängden falska alarm, att *fram* lyftes ur formregeln helt, och programmet upptäcker därför inte detta RWE. Tabell 12 visar alla felstavade prepositioner.

Tabell 12. Prepositioner som fått felaktig form.

Preposition	**Autentisk Förväxlingsmängd**	**Reviderad förväxlingsmängd**	**RWE**
på	*pa*	*pa*	
av	*ov*	*ov åv*	
efter	*after*	*after*	
från	*fran fram from fråm*	*fran from fråm*	*from*
för	*for fär*	*for fär*	*for*
genom	*inom*		*inom*
i	*in*	*in*	*in*
inom	*inon*	*inon*	
mellan	*mellon medan*	*mellon medan melen melon*	*medan melon*
mot	*måt*	*måt mut*	
nära	*nar när*	*nar när*	*nar när*
över	*over*	*over*	
till	*til tili*	*til tili*	
utan	*ettan*	*ettan uttan*	*ettan*
utanför	*utenför*	*utenför utenfor utanfor*	

I tabell 13 har jag sammanställt prepositioner som råkat ut för hopslagning med sin rektion, samt ord som blivit prepositioner när de skulle varit något annat. Den fonologiskt intressanta förväxlingen av *är* med *i* uppträder fem gånger, hos samma skribent.

Tabell 13. Felaktiga sammanskrivningar samt förväxling som gett upphov till preposition

Preposition	**Sammanskrivning**	**RWE**
i	*isverige*	
i	*istan*	
Avsett ord	**Preposition**	**RWE**
men	*med*	*med*
år	*ur*	*ur*
är	*i*	*i*

4.5.3 Regler för särskrivningar

Vad gäller särskrivningar (Tabell 14) så är de adverbiala ?*i bland*, *i går* och så vidare, ganska okomplicerade när det kommer till regelskrivning. De detekteras även av Stava. Problemet är bara att det är tveksamt om det över huvud taget är något fel med dem! Andra särskrivningar har inte kvalificerat för regelskrivning, antingen på grund av sin alltför tillfälliga karaktär (*efter bliven*) eller också för att de bereder alltför stora svårigheter att eliminera falska alarm (*Jag hittade bilnyckeln **under kläder** och serietidningar i barnens rum*).

Andra åter är mycket goda kandidater för regelskrivning, såsom förväxling av *därför* med olika kombinationer av *där*, *dar*, *for*, och *för*. I det här fallet tyckte jag det kunde vara mödan värt att försöka matcha både särskrivning och vokalförväxling i samma regel, eftersom felkomplexet är

såpass vanligt. I själva verket blev regeln också komplex. Särskrivning av *därför* med olika kombinationer av *där*, *dar*, *för*, och *for* går att fånga upp i de fall elementen är felstavade *där for* eller *dar for*, **eller** där sekvensen omedelbart följs av ett verb eller föregås av ett interpunktionstecken.

Tabell 14. Särskrivningar med prepositionella element.

Samman-sättning	**Särskrivning**	**Autentisk förväxlings-mängd**	**Reviderad Förväxlingsmängd**	**RWE**
hemifrån	*hem*			*hem*
	i			*i*
	från			*från*
tillbaka	*till*	*til*	*til*	*till*
	baka	*båcka*	*båcka backa bocka båka boka*	*backa bocka boka*
därför	*där*	*dar*	*dar*	*där dar*
	för	*for*	*for*	*for*
tillsammans	*till*		*til*	*till*
	sammans			
utanför	*utan*		*uten*	*utan*
	för		*for*	*för for*
Mellanöstern	*Mellan*			*Mellan*
	Östern	*Östern*	*Östern*	(*Östern*)
Centralstationen	*Central*			*Central*
	stationen			(*stationen*)
eftermiddag	*efter*			*efter*
	middag			*middag*
efterblivna	*efter*			*efter*
	blevna	*blevna*	*blevna*	
eftersom	*efter*	*after*	*after*	*efter*
	som			*som*
underkläder	*under*			*under*
	kläder			*kläder*
iväg	*i*			*i*
	väg			*väg*
ihåg	*i*			*i*
	håg			*håg*
ifrån	*i*			*i*
	från			
ibland	*i*			*i*
	bland			*bland*

5 Resultat

5.1 Regeluppsättningens struktur

Analys och formalisering har resulterat i två regelkategorier och sex subgrupper (Tabell 15).

Tabell 15. Feltyperna med motsvarande regeltyper.

Feltyp		**Exempel**	**Regeltyp**	
			Subgrupp	**Kategori**
Formfel	stavfel	*mellon, fär, from*	spellpp@	formpprules
	särskrivning	*till båcka, där för*	särpp@	
	idiomfel	*till en stor grad*	idiompp@	
Kontext-betingat fel	ersättning	*Han tänkte **om** kläder*	substpp@	lexpprules
	borttagning	*Sonja trivdes sitt jobb*	delpp@	
	insättning	*att träffa **med** vänner*	inspp@	

Regeluppsättningen består av 31 granskningsregler för detektion, diagnos och korrektion av stavfel, särskrivningsregler och idiom, 9 regler för utelämnad preposition (borttagning), 8 regler för felaktigt prepositionsval (ersättning), samt 5 regler för insättning.. Härutöver har jag skrivit en sammansatt accepterande regel för alternativ ordföljd och 3 hjälpregler som återanvänder befintliga hjälpregler, och som anropas i granskningsreglerna. Två av formfelsreglerna har jag kommenterat ut för revidering, eftersom de gett upphov till alltför många falska alarm.

5.2 Granskningsregler

Jag har valt ut några regler som får representera regelpaketet. Hjälpreglerna som anropas i reglerna är oerhört mycket mer omfångsrika med avseende på programkod, och får inte plats här.

5.2.1 Regler för formfel

Regeln **`spellpp3`** är en typisk ”förväxlingsmängdsregel”. Förväxlingarna matchas i vänsterledet. Den avsedda prepositionen returneras i utdata. Regeln detekterar **Jag är from Atlanta* och inte *Han var from som ett lamm* eller *Fransiskus var en from man*, däremot får man falskt alarm på *Fransiskus var from och vis*.

Indata:	**Jag är from Atlanta*	*Fransiskus var from och vis*
Utdata:		
detektion	*Jag är* [*from*] *Atlanta*	*Fransiskus var* [*from*] *och vis*
diagnos	Du menar antagligen *från*	Du menar antagligen *från*
korrektion	[*från*] *Atlanta*	[*från*] *och vis*

```
spellpp3@formpprules
{
X1(wordcl!=dt),
X2(text="fran" | text="fron" | text="fråm" | text="from"),
X3(text!="som" & wordcl!=pp)
-->
mark(X2)
corr(X2.replace("från"))
info("Du menar antagligen" italics("från"))
action(scrutinizing)
}
```

Regeln **`idiompp5`** tar hand om kontamination av ett par idiomatiska uttryck.

Indata:	**till en stor grad*
Utdata:	
detektion	[*till en stor grad*]
diagnos	Du menar antagligen *till stor del* eller *i hög grad*
korrektion	[*till stor del*] [*i hög grad*]

```
idiompp5@formpprules
{
X1(text="till"),
X2(text="en")?,
X3(text="stor"),
X4(text="grad" | text="del")
-->
mark(X1 X2 X3 X4)
corr(X1.replace("i") X2.delete() X3.replace("hög") X4.replace("grad"))
corr(X2.delete() X4.replace("del"))
info("Du menar antagligen" italics("till stor del") "eller" italics("i
hög grad"))
jump(endlabel, 4)
action(scrutinizing)
}
```

5.2.2 Regler för kontextbetingade fel

Regeln **`delpp4`** detekterar frånvaro av preposition efter reflexivt verb och före infinitivmärke+infinit verb. Typfallet är bortfall av *för* i syftesadverbial (eller i subjunktion med final funktion, beroende på hur man väljer att se det).

Indata: **Lisa gifte sig att få pengar*

Utdata:
detektion *Lisa gifte* [*sig att få*] *pengar*
diagnos Du har nog glömt preposition före *att få*
korrektion *sig* [*för*] *att få*

```
delpp4@lexpprules
{
X1(lemma="ansluta" | lemma="avreagera" | lemma="bekanta" |
lemma="bestämma" | lemma="betala" | lemma="bilda" | lemma="bosätta" |
lemma="dra" | lemma="förkyla" | lemma="förälska" | lemma="ge" |
lemma="gifta" | lemma="gömma" | lemma="huka" | lemma="hämta" |
lemma="hänga" | lemma="hävda" | lemma="infinna" | lemma="lata" |
lemma="lägga" | lemma="missta" | lemma="resa" | lemma="sjåpa" |
lemma="skingra" | lemma="skilja" | lemma="skrapa" | lemma="sköta" |
lemma="slå" | lemma="specialisera" | lemma="sprida" | lemma="staka" |
lemma="stegra" | lemma="stå" | lemma="sätta" | lemma="tvätta" |
lemma="uppenbara" | lemma="uppföra" | lemma="utmärka" | lemma="vidga" |
lemma="vila" | lemma="ångra" | lemma="återhämta"),
(NPall/X2)()?,
X3(wordcl=pn & pnf=obj & pnf!=sub),
X4(wordcl=ab)*,
(INFP1/X5)()
-->
mark(X3 X4 X5)
corr(X5.insert("för"))
info("Du har nog glömt preposition före" italics(X5.real_text))
action(scrutinizing)
}
```

Regeln **substpp3** upptäcker olyckligt prepositionsval efter *tillbaka* som partikel, förutsatt att det föregående (partikel)verbet underförstår *riktning*. Regeln detekterar till exempel **Jag kom tillbaka i Stockholm*. CHWDORDER är en hjälpregel som hjälper granskningsregeln att också matcha till exempel **Därför kom jag tillbaka i Stockholm*. Den alternativa korrektionen med *genom* är beredskap för konstruktioner som ?*Pojkarna gick tillbaka i tunneln*. Regeln ger falskt alarm för till exempel *Hon åkte tillbaka i en Alfa Romeo*.

Indata:	**Jag kom tillbaka i Stockholm*	*Hon åkte tillbaka i en Alfa Romeo*
Utdata:		
detektion	*Jag kom* [*tillbaka i*] *Stockholm*	*Hon åkte* [*tillbaka i*] *en Alfa Romeo*
diagnos	Du har nog skrivit fel preposition efter *tillbaka*	Du har nog skrivit fel preposition efter *tillbaka*
korrektion	*kom tillbaka* [*till*] *Stockholm*	*åkte tillbaka* [*till*] *en Alfa Romeo*

```
substpp3@lexpprules
{
X1(wordcl=vb & vbt!=kop & lemma="komma" | lemma="hitta" | lemma
="längta" | lemma="gå" | lemma="åka"),
(CHWDORDER/X2)()?,
X3(text="tillbaka"),
X4(text="i" | text="in"),
(NPall/X5)()
-->
mark(X3 X4)
corr(X4.replace("genom"))
corr(X4.replace("till"))
info("Du har nog skrivit fel preposition efter" italics(X3.real_text))
jump(endlabel, X2.no_of_tokens+2+X5.no_of_tokens)
action(scrutinizing)
}
```

I konstruktionen ?*Jag var tillbaka till Stockholm* hade det varit korrekt med *i* (det kopulativa verbet underförstår *tillstånd* inte riktning). Jag har därför skrivit en motsvarande regel som bara matchar kopulativt verb före *tillbaka* och föreslår *tillbaka i*, men också *tillbaka på* (*ön, jobbet*).

Regeln **`inspp2`** detekterar otillbörlig insättning av preposition före rums- och riktningsadverb. Det finns inga subtaggar för olika typer av adverb så därför har jag i vänsterledet listat så många jag kommit på som uppfyller de villkor jag vill matcha. Här har jag haft hjälp av adverbtabeller i Lindholm (1997).

Indata: **Sedan tog jag en buss till hit*

Utdata:
detektion *Sedan tog jag en buss* [*till hit*]
diagnos Det ska nog inte vara preposition före *hit*
korrektion [*hit*]

```
inspp2@lexpprules
{
X1(text="i" | text="till"),
X2(text="där" | text="dit" | text="hem" | text="hemma" | text="tillbaka"
| text="tillbaks" | text="bakåt" | text="framåt" | text="därifrån" |
text="härifrån" | text="ut" | text="in" | text="ute" | text="inne" |
text="hemifrån" | text="utomlands" | text="då" | text="bak" |
text="baktill" | text="norrut" | text="söderut" | text="österut" |
text="västerut")
-->
mark(X1 X2)
corr(X1.delete())
info("Det ska nog inte vara preposition före" italics(X2.real_text))
jump(endlabel, 2)
action(scrutinizing)
}
```

6 Utvärdering

Jag började med att manuellt annotera kontrollgruppen för felaktiga prepositionskonstruktioner, på samma sätt som jag gjort med hela materialet. Detta betyder att vissa mycket svåranalyserade sekvenser, som exemplen nedan, lämnades utanför prepositionsfelmängden

Klockan halv nio gick till gatan för buss station.
Men japanerna kunde inte jobba också även kunde inte ansöka sig förtillståndet.
De gudarna tyckte att det kan bli bra att han ska klä på kvinlig
Detta fenomen framträder tydligt i kyrkobesökare.

I textmängden om 3 800 ord hittade jag därmed 40 prepositionsfel. Detta var alltså programmets ”facit”. Därefter granskade jag kontrollgruppen med hjälp av mitt regelpaket. Programmet gjorde 11 detektioner. 7 av detektionerna var tillämpningar av regeln för förväxling av *på* med *pa* eller regeln för förväxling av *för* med *for*. Underligt nog missar programmet en förekomst av *pa* samt en förekomst av *får* istället för *för*. Dessa missar kan jag inte förklara. Kontrollgruppstexten innehöll dessutom ytterligare en variant av *från*, nämligen *från*.

De övriga fyra detektionerna var tillämpningar av regel 5 och 6 för felaktigt prepositionsval och regel 3 för insättning. Två av dessa, (43) och (44), utgjordes av vad man kunde kalla *oäkta falska alarm*. Denna besynnerliga term betyder att programmet reagerar på ett ställe i texten där ett fel verkligen begåtts, men inte det fel som regeln tillämpas för, ungefär som i exemplet med *ordat* i avsnitt 1.4.

(43) *I tisdag 15 Februari 1973, gick Lasse till arbetet.*
får diagnosen ”Du har nog skrivit fel preposition före *tisdag*” med korrektionsförslaget [*På*] *tisdag*

(44) *Vi träffade i Minneapolis.*
får diagnosen ”Du ska nog inte ha preposition efter verbet *träffade*” med korrektionsförslaget [*träffade*]

Den andra ersättningsregeln tillämpas på ett okontroversiellt sätt (45), under det att insättningsregeln får rycka in i ytterligare en problematisk situation (46).

(45) *Jag kunde inte sov i nätterna*
får diagnosen ”Du har nog skrivit fel preposition före nätterna” med korrektionsförslaget [på]nätterna

(46) *Klockan kvart i fyra träffade med ingenjör Karlsson, och pratade med honom om deras nya TV-apparater.*
får diagnosen ” Du ska nog inte ha preposition efter verbet *träffade*” med korrektionsförslaget [*träffade*] *ingenjör Karlsson*

(43) och (44) är korrekt detekterade men inkorrekt diagnosticerade och korrigerade. (46) är korrekt detekterad och korrekt diagnosticerad får man nog säga. Men en revision enligt programmets korrektionsförslag kan inte ersätta det förlorade subjektet.

Med definitionen av täckning och precision som gavs i avsnitt 3.6.2 (det vill säga man räknar bara på *detektioner* utan att titta på diagnos och korrektion) så motsvarar resultatet en täckning på 25 procent (kvoten av 11 korrekt detekterade prepositionsfel och 40 förekomster av prepositionsfel) och en precision på 100 procent (kvoten av 11 korrekta detektioner och summan av dessa och 0 falska alarm). Möjligen speglar dessa värden min ambition att begränsa antalet falska alarm på bekostnad av täckningen (avsnitt 3.6.2). Men det lysande värdet för precisionen avslöjar också hur trubbig utvärderingsmetoden är i dessa sammanhang, och samtidigt hur nödvändigt det är med en utvärdering med användare för att få en uppfattning om hur programmet fungerar. De två felaktigt diagnosticerade detektionerna visar också hur viktigt det kan vara med diagnos, och i synnerhet en hänsynsfull diagnos. Ett program som returnerar detektion av exempel (43) och (44) utan diagnos, är inte ett bra verktyg. Då är det bättre att utelämna korrektionen också.

Att utvärdera hur användare bedömer programmets återkoppling och vilken effekt granskningen har för hur användaren verkligen redigerar sin text, är en allt annat än trivial uppgift. Utvärdering med användare på Nada (Knutsson et al 2002) har i alla händelser tydligt visat att återkoppling (det vill säga diagnos och korrektion) har ett avgörande inflytande på revisionsprocessen.

7 Diskussion

7.1 Variabler utanför analysen – interferens och modersmål

I avsnitt 1.5.2 tog jag upp begreppet interferens. Termen lyser sedan snarast med sin frånvaro genom uppsatsen. I avsnitt 1.2 deklarerar jag att jag inte avser undersöka modersmålsvariabeln närmare. Varför? Borde det inte gå att identifiera skrivfel som är typiska för en speciell modersmålsgrupp, precis som det ofta går att identifiera modersmål hos talare av svenska som andraspråk?

Att det är glest med interferensanalyser beror inte på att materialet saknar ytterligare belägg för interferensfenomenet. Tvärtom har många felkonstruktioner av interferenskaraktär, som ersättning av *från* med *from* eller *tänka på* med *tänka om* (*think about*), gett upphov till granskningsregler. Det är däremot inte den troliga interferensen som har varit avgörande för formaliseringen. Regeln för formfel hos prepositionen *från* matchar också förväxlingarna *fråm* och *fran* varav särskilt den senare inte är någon stark interferenskandidat. Exemplet (och diskussionen i avsnitt 4.5.2) visar också hur komplicerad interferensanalysen är. Har interferensen i fallet *from* semantiska, ortografiska eller fonologiska orsaker, eller är det fråga om en kombination? Är det verkligen interferens med ett annat språk som ligger bakom svårigheterna med svenska ordföljdsregler (*Igår jag var på bio*) eller har problemen snarare att göra med att svenskan helt enkelt är strukturellt ganska avvikande härvidlag?

Av mina reflektioner framgår att interferensanalysen kräver en ganska ingående undersökning av orsakssammanhangen och dessutom stora språktypologiska kunskaper. Min analys av problem i prepositionssammanhang har snarare ett konsekvensperspektiv. En sådan analys blir mindre flertydig och är att föredra när målet är att utveckla ett datorprogram för automatisk grammatikkontroll. Konsekvensanalysen blir med nödvändighet ganska ointuitiv i vissa situationer, till exempel när en särskrivning som *där för* analyseras som ett fall av otillbörlig insättning av preposition. För diskussion om en feltypologisk analys med orsaksperspektiv hänvisar jag till Eeg-Olofsson (2002:30f).

I avsnitt 1.5.2 utreder jag begreppet fonologisk interferens, som jag exemplifierar med ordet *mörkt* som en japansktalande skribent formar om till *möligt*. Fonologisk interferens i tal kan ibland erbjuda mycket säkra ledtrådar till talarens modersmål, såsom arabisktalande inlärares svårighet med tonlös labial (`[parkera]` blir `[barkera]`). Om en del av syftet med grammatikkontrollen var att identifiera användarens modersmål så skulle det vara mycket intressant att försöka formalisera typiska fonologiska interferenser. Men nu är syftet att försöka hjälpa användaren att producera svensk text, och då är det inte säkert att det är så relevant att känna till skribentens modersmål.

Orsakerna till skrivfelen tycks generellt vara multifaktoriella eller mycket svåra att identifiera. I min analys är feltypen borttagning överrespresenterad hos japansktalande skribenter (avsnitt 4.4.1), men jag kan inte därmed utan vidare dra slutsatsen att japansktalande skribenter typiskt tenderar att utelämna prepositioner. Dessutom är feltypen insättning också den överrepresenterad hos de japansktalande skribenterna (med mer än tre gånger så många förekomster som hos engelsktalande och arabisktalande respektive). Den sammanlagda felmängden är inte större hos japansktalande än hos arabisktalande men något större än hos engelsktalande skribenter (avsnitt 4.2.5). Mängden särskilt svårtolkade prepositionskonstruktioner är något större för japansktalande skribenter. Det finns emellertid ingenting som säger att dessa omständigheter avspeglar just prepositionsproblematiken, av den enkla anledningen att jag i princip inte undersökt skrivfel av andra sorter och alltså inte har något att jämföra prepositionsfelen med.

Identifiering av modersmålsspecifika skrivfel med interferensanalys är ett osäkert företag, bland annat för att interferenskandidater inte alls behöver vara modersmålsbetingade. Exempel (47) härrör från en japansktalande skribent.

(47) *Klockan fempton i fyra, träffade han med ingenjör Karlsson.*

Är förekomsten av prepositionen *med* en interferens med engelska eller kan den ha helt andra orsaker? Det enda jag kan konstatera är att det är ett fall av insättning.

Hypotesen att skribentens modersmål är en viktig variabel i en feltypologisk analys är naturligtvis intressant. Därför har jag också gjort en alternativ sortering av material och felmängd efter modersmålsgrupp för att det ska vara möjligt att utföra analyser av mitt material med modersmål som variabel.

7.2 Handlar det i själva verket om verb?

7.2.1 Villkor för prepositionsval

Jag har vid upprepade tillfällen (avsnitt 3.1; avsnitt 4.3; avsnitt 4.4) varit inne på det mått av godtycke som tycks kännetecknа prepositionsanvändning och -distribution i svenskan. Den som lär sig svenska som andraspråk är inför prepositionsvalet mycket ofta hänvisad till kunskap som inte har så mycket med språkets grammatiska struktur som med dess *lexikon* att göra. Det heter *Tänk på barnen!* och inte *Tänk om barnen!*, det heter *vänja sig vid kylan* och inte *vänja sig till kylan*. Det betyder naturligtvis inte att engelskans prepositionsval i motsvarande konstruktioner, nämligen *Think about the children!* och *get used to the chill*, skulle vara lättare eller rimligare. Semantiskt godtycke och tendensen till lexikalisering i prepositionskonstruktioner är ingenting speciellt för svenskan.

I avsnitt 4.3.2 kunde jag konstatera att analys av kontexten gav utförligare information om eventuella svårigheter med prepositionskonstruktionerna. Särskilt tacksam var metoden för att lokalisera svårigheter med prepositionsverb som *tänka (på)*. Detta säger oss någonting om var vi bör leta efter de enheter som styr vilken preposition som kan komma ifråga.

Låt oss betrakta prepositionen med hjälp av SAGs termer, som prototypiskt flankerad av en A- och en B-referent. Verbet på A-referentens sida verkar alltså ha mycket stor betydelse för prepositionsvalet. Det finns naturligtvis exempel på fullständigt lexikaliserade förbindelser med rektionen , som *till exempel*, *ibland* och *med mera*, men rektionen kan oftast uppträda med en mängd olika prepositioner. Nominalfrasen *tiden* till exempel, kan stå som rektion till åtminstone *i*, *ur*, *under*, *över*, *på*, *med*, *för*, *av*, *vid* och *till*.

Med verb som binder prepositionen till sig har man emellertid sällan så många alternativ, ändå tycks dessa verbförbindelser bereda större svårigheter för andraspråksinlärarna i mitt material (**man läser på tidningen*, **kvinnor tänker om problemet*, **När hon kom in i kassa tittar de på henne och skrattar på henne*). Ibland ”skänker” skribenten istället en preposition till ett verb som kräver direkt objekt (**då måste de hindra med det*, **Jag takade till henne*). Vanligast är kanske att den obligatoriska prepositionen utelämnas helt (**Jag väntade tåget*, **Då kallade Lasse en taxi*, **jag kände att skriva några rader*, **han vislar hans hund*).

Verbet behöver för all del inte vara ett entydigt verb, men har nästan alltid semantisk funktion som predikat i satsen. Man skulle kunna placera ut dessa predikat längs en skala enligt fallande verbidentitet

tänka *på*
vänja sig *vid*
irriterad *på*
van *vid*
allergisk *mot*

Man kan också tänka sig en uppställning enligt fallande *bindningsstyrka* mellan det predikativa elementet och prepositionen. Om vi väljer att betrakta prepositionshomonyma verbpartiklar som prepositioner, har vi den starkaste bindningen mellan partikelverbet och dess partikel.

lägga av
allergisk mot
skratta åt/med/…
lyssna på/till/med/…
arbeta på/till/för/i/med/under/…

Jag misstänker alltså att prepositionen ofta har sina starkaste band till A-referentens sida, närmare bestämt till den enhet som har huvudrollen i beskrivningen av A-referentens tillstånd, alltså verbledet eller predikatet (ibland predikativen) till vänster om prepositionen. Därför bör också prepositionen kunna integreras i beskrivningen av verb, participer och adjektiv. Upptäckten är knappast ny, vare sig inom den grammatiska beskrivningen eller i sammanhang som rör andraspråksinlärning.

7.2.2 Valensbeskrivningar

Den typ av information om orden som handlar om vad dessa så att säga ställer för krav på sin omgivning, kan man få i så kallade *valensbeskrivningar*, en term som lånats från *dependensgrammatiken* (Megyesi 1996:5ff) och som har vissa likheter med analys av predikat- och argumentstrukturer i språket. Valensbeskrivningar är integrerade i nyare ordböcker, bland annat i Nationalencyklopediens ordbok och i Skolverkets stora ordboksprojekt *Lexin* (Skolverket 1996; Skolverket 1992), som särskilt riktar sig till andraspråksinlärare. Det finns till och med en variant av Lexin, där uppslagsorden sorterar under valensbeskrivningarna (Skolverket 1992). Lexin levererar följande valensbeskrivning av verbet *tänka* i huvudbetydelsen *medvetet bearbeta i hjärnan, fundera*.

<A tänker (på x/att+SATS)>

Parentesuttrycket är den optionella bestämningen till verbet, som måste bestå av prepositionen *på*, följd av antingen en inanimat referent eller också en bisats inledd av *att*. I min demonstrationsregel i avsnitt 4.3.2, som jag nedan presenterar en gång till, motsvaras uttrycket av den sista matchningsvariabeln `NPall` – en hjälpregel. Demonstrationsregeln har i jämförelse med Lexin en vidare valensbeskrivning och tillåter dessutom prepositionen *med* tillsammans med verbet.

```
{
X1(lemma=”tänka” & vbf!=imp),
(ADV)()?,
(NPall)()?,
X4(wordcl=pp & text=”om”),
(NPall/X5)()
-->
mark(X1 X4)
corr(X4.replace(”med”))
corr(X4.replace(”på”))
info(“Du har nog valt fel preposition till verbet”
italics(X1.real_text))
action(scrutinizing)
}
```

Valensinformation ur Lexin i kombination med kontextsökning i standardspråkskorpusar skulle säkerligen ge mycket värdefull information för det fortsatta arbetet med anpassning av Granska för andraspråksinlärare. Den starkaste typen av vänsterförbindelse för prepositionerna är den som knyter prepositionen som partikel till verbet. Tillräckligt vanliga partikelverb är i Lexin egna komplexa uppslagsord som i och för sig sorterar under verbhuvudet som specialfall (*tänka* representeras till exempel med tre partikelvarianter *tänka om*, *tänka ut* och *tänka över*). Partikelverben måste integreras i en valensinriktad analys av prepositionskonstruktioner och då blir partikelns ordklasstagg kanske mindre intressant. Med fokus på valens blir regelskrivningen i första hand en fråga om att matcha listor av ord eller lemman, ungefär som jag gjort med reflexiva verb (regeln `delpp4` i avsnitt 5.2.2). Mycket vanliga missuppfattningar av valensstruktur kan naturligtvis också klassificeras som förväxlingsmängder.

Ingenting är emellertid så enkelt som det verkar. Precis som alla andra försök att kategorisera naturligt språk så stöter valensbeskrivningen på patrull i mötet med verkligheten, och eventuella valensregler kommer att behöva kompletteras med en försvarlig uppsättning accepterande regler för att kunna matcha alternativa ordföljder av typen *Därför **skrev** inte hyresgästen **på** avtalet*.

7.3 Slutord

Det mest påtagliga resultatet av min undersökning är regelsamlingen för upptäckt och korrektion av prepositionsrelaterade skrivfel som presenteras i avsnitt 5. Avsikten med denna är i första hand att min utformning av granskningsreglerna ska kunna bilda underlag för konstruktion av nya regler. Man skulle kunna kan säga att jag genom min undersökning levererat ett förslag till startpaket.

Det är generellt svårt att formalisera användningen av prepositioner. Prepositionen står som vi sett i många fall i tillfälliga förbindelser med en eller flera av sina referenter och deras beskrivningar. Mitt arbete har emellertid visat att det ändå kan vara mödan värt att konstruera granskningregler för prepositionsrelaterade fel, just med hänsyn tagen till ett mer eller mindre lexikaliserat distributionsmönster. Mer allmänt bör det finnas förutsättningar för att ta fram en version av Granska som erbjuder konstruktiv hjälp för skribenter med svenska som andraspråk, särskilt om grammatikkontrollen tillämpas som pedagogiskt komplement i en undervisningssituation.

Referenser

L-G Andersson (1987) *Språktypologi och språksläktskap* Skriptor

N Beckman (1968) *Svensk språklära för den högre elementarundervisningen* 2:a svenska upplagan 1916, Albert Bonniers förlag, Stockholm

H Bergreen & K Tenfjord (1999) *Andrespråksinlæering* Ad Notam Gyldendal, Oslo

J Bigert & O Knutsson (2002) "Robust Error Detection: A Hybrid Approach Combining Unsupervised Error Detection and Linguistic Knowledge" i *the Proceedings of Romand02*, 2nd Workshop on RObust Methods in Analysis of Natural language Data, Frascati, Italy.

J Carlberger & R Domeij & V Kann & OKnutsson (2000) "Granska - an efficient hybrid system for swedish grammar checking" i *proc 12th Nordic Conference in Computational Linguistics, Nodalida-99*. Department of Linguistics, Norwegian University of Science and Technology, Trondheim, ss 49-56.

T Cerratto & L Borin (2002) *The Use of Language Tools for Writers in the Context of Learning Swedish as a Second Language* Rapport, Nada KTH Stockholm

T Cerratto & L Borin (2002) *Overview of the Research Area* Rapport, Nada KTH Stockholm

A Cochran Crocker (2002) *The Grammar Issue: an Annotated Bibliography* Texas Tech University. URL: http://www.english.ttu.edu/carter/5369/AB/AmandaAB.doc

B Collinder (1971) *Svenska – vårt språks byggnad* P. A. Norstedts & Söners förlag, Stockholm

R Domeij & O Knutsson (1999) *Specifikation av grammatiska feltyper i Granska* Arbetsrapport, Nada KTH Stockholm

R Domeij & O Knutsson & S Larsson & K Severinson Eklundh & Å Rex (1998) *Granskaprojektet 1996-1997*. Rapport, IPLab-146, Nada KTH, och Stockholms Universitet

R Domeij & J Hollman & V Kann (1994) "Detection of Spelling Errors in Swedish Not Using a Word List En Clair" i *Journal of Quantitative Lingustics* 1994, Vol 1, nr 3 ss 195-201

J Eeg-Olofsson (2002) *Feltaxonomi för automatisk språkgranskning av svensk text* C-uppsats Institutionen för lingvistik Stockholms unviversitet

R Golding & Y Schabes (1996) *Combining Trigram-based and Feature-based Methods for Context-Sensitive Spelling Correction* i A Joshi & M Palmer edts. *Procedings of the Thirty-Forth Annual Meeting of the Assiciation for Computational Lingustics* (ss 71-78), San Francisco. Morgan Kaufmann Publishers

B Hammarberg (1977) *Svenskan i ljuset av invandrares språkfel* Nysvenska studier årg 57, Lund

B Hammarberg (1990) ”Aktuella teman i forskningen om andraspråksuttal” i G Tingbjörn red. *Andra symposiet om svenska som andraspråk* Skriptor, Stockholm

A Hansén-Eriksson & O Knutsson (2000) *En explorativ studie av språkgranskningsverktyg* Rapport, Nada KTH Stockholm

K Hyltestam & K Wassén (1984*) Svenska som andraspråk – en introduktion* Studentlitteratur, Lund

O Jespersen (1977) *The Philosophy of Grammar* (utgivningsår1924) George Allen & Unwin, London

V Kann & R Domeij & J Hollman & M Tillenius (1998) *Implementation aspects and Implications of a Spelling Correction Algorithm* Rapport, TRITA-NA-9813, 1998. Nada KTH Stockholm

O Knutsson (2001) *Automatisk språkgranskning av svensk text* Licentiatavhandling, TRITA-NA-01-5, Nada KTH Stockholm

O Knutsson (2002) ”Datorn som språkgranskare” i *Språkvård* 1:2002 ss 26-33 Svenska språknämnden

O Knutsson & T Cerratto Pargman & K Severinson Eklundh (2002) “Computer support for second language learners' free text production - Initial studies” för publicering i Proceedings of ICL2002, 5th International Workshop on Interactive Computer Aided Learning, Villach, Austria.

K Kukich (1992) “Techniques for Automatically correcting Words in Text” i *ACM Computing Surveys* 24:4 ss 337-439

G Källgren (1998) *Documentation of the Stockholm – Umeå Corpus* Institutionen för Lingvistik Stockholms universitet

E Lindberg (1980) *Beskrivande svensk grammatik* 2:a upplagan Almqvist & Wiksell, Stockholm

H Lindholm (1997) *Svensk grammatik – svenska som främmande språk* 5:e upplagan 1974 Kursverksamhetens förlag, Lund

KG Ljunggren (1951) “Towards a Definition of the Concept of Preposition” i *Studia Linguistica* 1951:5

B Megyesi (1996) *Implementering av partikelverb för projektet ‘Datorstödd inlärning av grammatik och språkteori’* C-uppsats i datorlingvistik, Institutionen för lingvistik, Stockholms universitet. URL: http://www.speech.kth.se/~bea/cuppsats.pdf

J Michel et al. (2001) ”Reference Manual” i *The Berkeley Utilities – Unix Commands for MS-DOS*. URL: http://www.openetwork.com/berk.html

R Mitchell & F Myles (1998) *Second Language Learning Theories* Arnold, New York

R Mitton (1996) *English Spelling and the Computer* Longman, New York

R Mitton (1996) “Spellchecking by the Computer” i *Journal of the Simplified Spelling Society* Vol 20 nr 1 1996 ss 4 –11. URL: http://www.dcs.bbk.ac.uk/~roger/spellchecking.html

C Prütz (2001) *Dynamic Time Warping* Kurs i talteknologi. Institutionen för lingvistik, Uppsala universitet. URL: inte längre tillgänglig elektroniskt

Skolverket (1996) *Lexin* URL: http://www-lexikon.nada.kth.se/skolverket/forord.shtml
URL: http://www-lexikon.nada.kth.se/skolverket/information-sv.shtml

Skolverket (1992) *Svenska ord – Lexin* 2:a upplagan
URL: http://scrooge.spraakdata.gu.se/lb/lexin/

A Staerner (2001) *Datorstödd språkgranskning som ett verktyg för andraspråksinlärning*. Examensarbete, Institutionen för lingvistik, Uppsala universitet

U Teleman (1974) *Manual för grammatisk beskrivning av talad och skriven svenska* Studentlitteratur, Lund

U Teleman & S Hellberg & E Andersson red. (1999) *Svenska Akademiens grammatik* Band 1 - 4, Norstedts Ordbok, Stockholm

O Thorell (1982) *Svensk grammatik* 2:a upplagan 1977, Esselte studium AB, Stockholm

T Pitkänen-Koli (1990) ”Fel i svenska uppsatser gjorda av finska grundskole- och gymnasieelever samt universitetsstuderande” i *Andra symposiet om svenska som andraspråk* G Tingbjörn (red.) Skriptor, Stockholm

Å Viberg (1987) *Vägen till ett nytt språk – andraspråksinlärning i ett utvecklingsperspektiv* Natur och Kultur, Stockholm

L Öhrman (2000) *Datorstödd språkgranskning och andraspråksinlärare* D-uppsats i datorlingvistik, institutionen för lingvistik, Stockholms universitet

Appendix

Granskas tagguppsättning – grundtaggarna

Tagg-beteck-ning[7]	Taggbeskrivning	Exempel
ab	adverb	ganska, inte, alls
dt	artikel, pronominell artikel	ett, den, alla, samma ,något
ha	frågeadverb, även som subjunktion	hur, när, där, då
hd	frågepronomen	vilken
hp	relativpronomen	som
hs	relativt eller frågande possessivpronomen	vars, vems
ie	infinitivmärke	att
in	interjektion	hej
jj	adjektiv	olika, rikare
kn	konjunktion	och, eller, men, än
mad	Interpunktion – meningsavskiljare	. ? !
mid	annan avskiljande interpunktion	, ; : –
nn	substantiv	skolor, hemland, åsikt
pad	omslutande interpunktion	’ ” (
pc	particip	oberoende, betydande
pl	verbpartikel	in, upp, om, till, tillbaka
pm	egennamn	Egypten, Erik
pn	pronomen	du, båda, många, dig
pp	preposition	i, om, på, av
ps	possessivpronomen	hennes, våra, sin
rg	räkneord – grundtal, datum	17, tre, 3 januari
ro	räkneord – ordningstal	tredje
sn	subjunktion	om, att, sedan
vb	verb	är, såg, försöka, måste

[7] Taggbeteckningarna motsvarar värden på särdraget `wordcl` i Granskas regelspråk.