Stockholms universitet
Institutionen för lingvistik
Påbyggnadskurs i datorlingvistik
C-uppsats HT 2001

Feltaxonomi

för automatisk språkgranskning av svensk text

Jens Eeg-Olofsson
Handledare: Ola Knutsson, Nada
Examinator: Lars Borin

Sammanfattning

Forskning kring automatisk språkgranskning av text har under det senaste decenniet bedrivits ganska intensivt på flera håll i Sverige. Metoder och infallsvinklar är delvis olika, men målet är i allmänhet gemensamt för de olika forskningsinstitutionerna, nämligen att ta fram ett effektivt korrekturläsningsprogram som klarar av att hjälpa skribenten att upptäcka och korrigera inte bara typiska stavfel, utan också grammatiska felkonstruktioner.

Ett idealt program för automatisk språkkontroll analyserar och förstår allt i en text, inklusive allt som skribenten avsett att förmedla när han eller hon uttryckt sig på ett olyckligt sätt. Detta program existerar naturligtvis inte. Vad man i första hand utvecklar idag är program för detektion, diagnos och korrektur av lokala syntaxbrott, en nog så svår utmaning. Ett sådant program måste kunna göra grammatiska analyser av text och det implicerar att det måste vara kontextkänsligt. Analysen ska alltså dessutom leda till upptäckt, diagnos och korrektur av felaktiga konstruktioner. Det betyder att man måste ge programmet regler för att ta hand om sådana konstruktioner. Vad som är rätt och fel, grammatiskt och ogrammatiskt, är sällan självklart och okontroversiellt. Därför måste man göra en lingvistisk (inte språkvårdande) analys av felkonstruktioner genom studier av autentiska texter. Identifierade fel måste också klassificeras för att kunna integreras i regeluppsättningen för korrekturläsningsprogrammet.

Denna uppsats beskriver arbetet med att klassificera fel i ett autentiskt material. Arbetet har bedrivits inom ramen för ett projekt för automatisk språkgranskning – ”Granska” – vid Institutionen för numerisk analys och datalogi (Nada) vid Kungliga Tekniska Högskolan i Stockholm.

Arbetet är samtidigt en utvärdering av det elektroniska verktyg för annotering av skrivfel som tagits fram inom Granska-projektet. Målet har varit att komma fram till en struktur för klassificering av fel, en feltypolgi, som ska bestämma utformningen av en ny version av annoteringsverktyget.

De texter som analyserats utgörs av gymnasieuppsatser i ämnet svenska. Granska-projektet har beviljats forskningsstöd för insamling av texter på svenska av andraspråksinlärare. Feltypologin ska kunna användas för fortsatt klassificering av fel i texter av andraspråksinlärare. I förlängningen ska en utökad feltypologi kunna ligga till grund för utveckling av regeluppsättningen i språkgranskningsprogrammet.

Feltypologin har växt fram i två annoteringsetapper. Den första etappen presenteras i uppsatsen som ”försöksstudie” och motsvar avsnitt 3. Försöksstudien beskriver annoteringsarbete och analys av fel med en preliminär feltypologi. Denna resulterar i en ny feltypologi som prövas på liknande sätt i etapp två i arbetet, som beskrivs i avsnitt 4. Resultatet är den tredje och (för tillfället) slutgiltiga typologin som ska styra utformningen av annoteringsverktyget.

Innehållsförteckning

1 Bakgrund och syfte

Korrekturläsningsprogram av typen stavningskontroll började utvecklas på sextiotalet (Kukich 1992). Det typiska stavningsprogrammet reagerar på ord – eller ”icke-ord” – som inte återfinns i en given ordlista, som de fetstilta orden i exempel 1 nedan.

1. *Det är för mig **beklämande** att på stan få höra **tjuåringar** svära åt sina kompisar*

Ett fenomen som genom åren gäckat stavningsprogrammen, dess konstruktörer och användare, är när ett felstavat ord just genom felstavningen blir ett annat existerande ord som man inte haft för avsikt att skriva och som stavningsprogrammet inte reagerar på eftersom ordet är representerat i ordlistan (exempel 2 och 3).

2. *Svenska svordomar är på ett eller annat **sett** kopplade till bibeln*

3. *det var ju den aggressiva tonen som barnet reagerade på inte ordens betydelse i **säg***

Mittons studie från 1987 (för engelska språket) av nära tusen gymnasieuppsatser, visade att hela 40 procent av alla stavfel resulterade i ett annat existerande ord (Kukich 1992). Denna typ av felstavningar kräver en analys av kontexten för att kunna upptäckas av ett rättstavningsprogram. Likadant är det med grammatiska felskrivningar som bryter mot regler för språkets syntax (exempel 4 och 5).

4. *Det **har gått börjar gå** till överdrift*

5. *Slangen anses som **fula** och opassande*

Forskning för att utveckla kontextkänsliga program för automatisk språkkontroll har för engelska bedrivits sedan början av åttiotalet. Sådana program har oftast haft kunskaps- eller regelbaserade system för grammatisk analys i botten. Med tiden har statistiska metoder för kontextuell analys vunnit mark (Kukich 1992).

Svenska automatiska språkgranskare började utvecklas först på nittiotalet. Dessa utgör på inget sätt en direktöversättning av engelska system. Förutsättningarna för automatisk språkgranskning av svensk text ser annorlunda ut, bland annat på grund av bristen på verktyg för fullständig satsanalys. Man är för den grammatiska analysen i princip låst vid ordklass- och frasidentifiering. Automatisk språkgranskning är dessutom i ganska hög utsträckning språkspecifik (Knutsson 2001).

Algoritmer för analys av språklig struktur är traditionellt utformade för att kunna analysera korrekta eller grammatiska konstruktioner och välja mellan olika tolkningsmöjligheter, under det att en regelvidrig konstruktion lämnas utan analys.

Unga män och kvinnor är välkomna	analys a	‹ ‹ *Unga män och kvinnor* › *är välkomna* ›
	analys b	‹‹‹ *Unga män* › *och kvinnor* › *är välkomna* ›
**Unga män och kvinnor är välkommen*	ingen analys	

Tabell 1. Tvetydig grammatisk (Nivre 1991) respektive ogrammatisk konstruktion[1]

[1] Traditionellt används i lingvistisk litteratur asterisk för att markera ogrammatiskhet. Denna konvention förbehålls här aktuell tabell, då anförda exempel annars nästan uteslutande representerar någon form av felkonstruktion.

Ett system för automatisk språkgranskning av text ska kunna analysera även ogrammatiska eller felaktiga strukturer, och därtill kunna leverera korrektur, och man har därför att implementera ytterligare regler i systemet, regler som matchar typiska felkonstruktioner. För att kunna bygga upp ett dylikt regelverk behöver man veta hur de typiska felkonstruktionerna ser ut. Man behöver en feltypologi. Det är arbetet med att utforma en sådan feltypologi som utgör ämnet för föreliggande uppsats.

Vid Institutionen för numerisk analys och datalogi (Nada) på KTH bygger man upp ett textgranskningsinstrument, Granska. Målet är att ta fram en automatisk korrekturläsare med användarvänligt gränssnitt. Granska är en hybrid mellan ett statistiskt och ett kunskaps- eller regelbaserat system (Knutsson 2001). I nuvarande form detekterar och korrigerar Granska i huvudsak dels särskrivningar, dels inkongruens i nominalfraser och predikativ. Som predikativ sorterar i det här sammanhanget det som i skolgrammatiken kallades ”predikatsfyllnad” och som i moderna grammatikor går under beteckningen ”bundet predikativ” (Knutsson 2001: 115 ff).

Särskrivning	*En privat **radio kanal***
Inkongruens i nominalfras	*En **viktigt** mening*
Inkongruens i predikativ	*Deras betydelse är **mångbottnat***

Tabell 2. Feltyper som detekteras av Granska

Man har härutöver utökat regeluppsättningen med regler för till exempel felaktiga sammansättningar, böjningsfel efter preposition och ordföljdsfel, i första hand för att visa att systemet har kapacitet för detektion, diagnos och korrektion av andra feltyper.

Granska har kontinuerligt utvärderats i arbeten av i första hand dess konstruktörer (Domeij, Knutsson, Larsson, Severinson Eklundh, Rex 1998; Carlberger, Domeij, Kann, Knutsson 2000; Knutsson 2001). (Öhrman 1998) och (Staerner 2001) har bidragit till utvecklingen av Granska med avseende på särskilda feltyper (särskrivningar respektive ordföljdsfel).

För svenskt vidkommande har utveckling av språkgranskningsverktyg, vid sidan om Granska, också skett vid institutionen för lingvistik på Göteborgs universitet samt vid institutionen för lingvistik på Uppsala universitet. Projektet i Uppsala, som går under namnet Scarrie, är ett samarbete med norska och danska institutioner. Alla dessa språkgranskningsverktyg är forskningsprototyper (Knutsson 2001: 19 ff). En kommersiell språkgranskare med bland annat en svensk version har släppts av finska Lingsoft (Lingsoft 2001). De danska och norska versionerna av Scarrie bygger på ett uppmärksammat holländskt system för automatisk språkgranskning, CORRie (Vosse 1994). Som vi ska se kommer felklassificeringen i CORRie att få inflytande över föreliggande arbete.

Dessutom erbjuder marknaden verktyg för stilkontroll. Dessa representerar emellertid en annan typ av system som bygger på frasmönstermatchning med ett enda stort fraslexikon – alltså samma struktur som ett klassiskt rättstavningsprogram – till skillnad från en grammatikgranskare som bygger på ett system av produktiva regler som kan generera både korrekta och inkorrekta konstruktioner (Vosse 1994: 41 f).

Det uppmärkningsarbete som ska definieras och beskrivas i denna uppsats bildar den empiriska grunden till en klassificering av fel, en feltypologi. Denna feltypologi ska kunna användas, dels för att förbättra utformningen av det annoteringsverktyg (se avsnitt 2) som

tagits fram på Nada, dels för utvärdering av Granska som språkgranskningsverktyg samt eventuellt för implementering av fler felregler i systemet.

Den resulterande strukturen är dessutom tänkt att fungera som avstamp för en vidareutveckling av feltypologin för texter skrivna av andraspråksinlärare. Granskagruppen har beviljats forskningsanslag för att ta fram ett andraspråksinlärarmaterial och planen för föreliggande uppsats inbegriper en del 2 (läs: D-uppsats) med bland annat en grundläggande hypotes att felskrivningar bör kunna klassificeras efter skribentens modersmål. I väntan på textmaterial från andraspråksinlärare finns ambitionen att redan i nuläget bereda mark för en utvidgad feltypologi, bland annat genom att i arbetet med feltyperna ta hänsyn till vad som beskrivits i studier av texter av andraspråksinlärare. (Öhrman 2000) och (Staerner 2001) har i sina studier av andraspråksinlärarmaterial haft för ögonen en utveckling av Granska-systemet.

Feluppmärknings- eller annoteringsarbetet ska vara lingvistiskt till sin karaktär, det vill säga analysen ska baseras på lingvistisk intuition och kunskap, snarare än på maskinell strukturanalys och mönstermatchning, trots att granskasystemet i sig som nämnts är ett hybridsystem i det här avseendet. Det är dessutom viktigt att leverera en så heltäckande lingvistisk feltypologi som möjligt. Detta betyder också att delar av felanalysen med nödvändighet kommer att gå utanför vad som är möjligt att i nuläget implementera i Granskas regelsystem. Med andra ord så kommer grammatiskt oklanderliga konstruktioner som exempel 6 att märkas upp för semantiskt regelbrott.

6. *Vill Sverige vara ett land som talar som grottmänniskor?*

Nu finns det ändå gränser för vad den lingvistiska felanalysen kan ta hänsyn till. Metakontextuella analyser är i princip inte aktuella, och det ska visa sig att det också i allmänhet bär för långt att arbeta med analysenheter över meningsnivå. Som vi ska se är emellertid gränsdragningen i många fall långtifrån okomplicerad (se särskilt avsnitt 3.3.2).

2 Metod

2.1 Material

På Institutionen för nordiska språk i Uppsala arbetar man med att föra över gymnasieuppsatser från handskrivet till elektroniskt format. 77 uppsatser som ”korrekturlästs”, det vill säga den elektroniska versionen är konrollerad en extra gång för överensstämmelse med det handskrivna originalet, har levererats för det feltypologiska arbetet.

Uppsatserna är alla skrivna som nationellt prov i svenska, kurs B, år 1998 och 1999. Samtliga är så kallade argumenterande uppsatser, det vill säga uppgiften har varit att argumentera för en ståndpunkt eller åsikt. [2]

[2] Det är en fördel i sig att materialet är nytt för Granska - ju större indata desto bättre feldetektion får man hoppas. Den nya kontakten med Nordiska språk i Uppsala är förhoppningsvis också av betydelse inte bara inom ramen för föreliggande uppsats, utan också för forskningssamverkan mellan de olika inblandade institutionerna i ett större sammanhang.

2.2 Verktyg

På Nada har man arbetat fram ett annoteringsverktyg som producerar annoteringar i XML-format. Drivkraften bakom tillkomsten av annoteringsverktyget har varit ökade krav på effektiv utvärdering av Granska-systemet, i kombination med en önskan om ett mer generellt utdataformat.

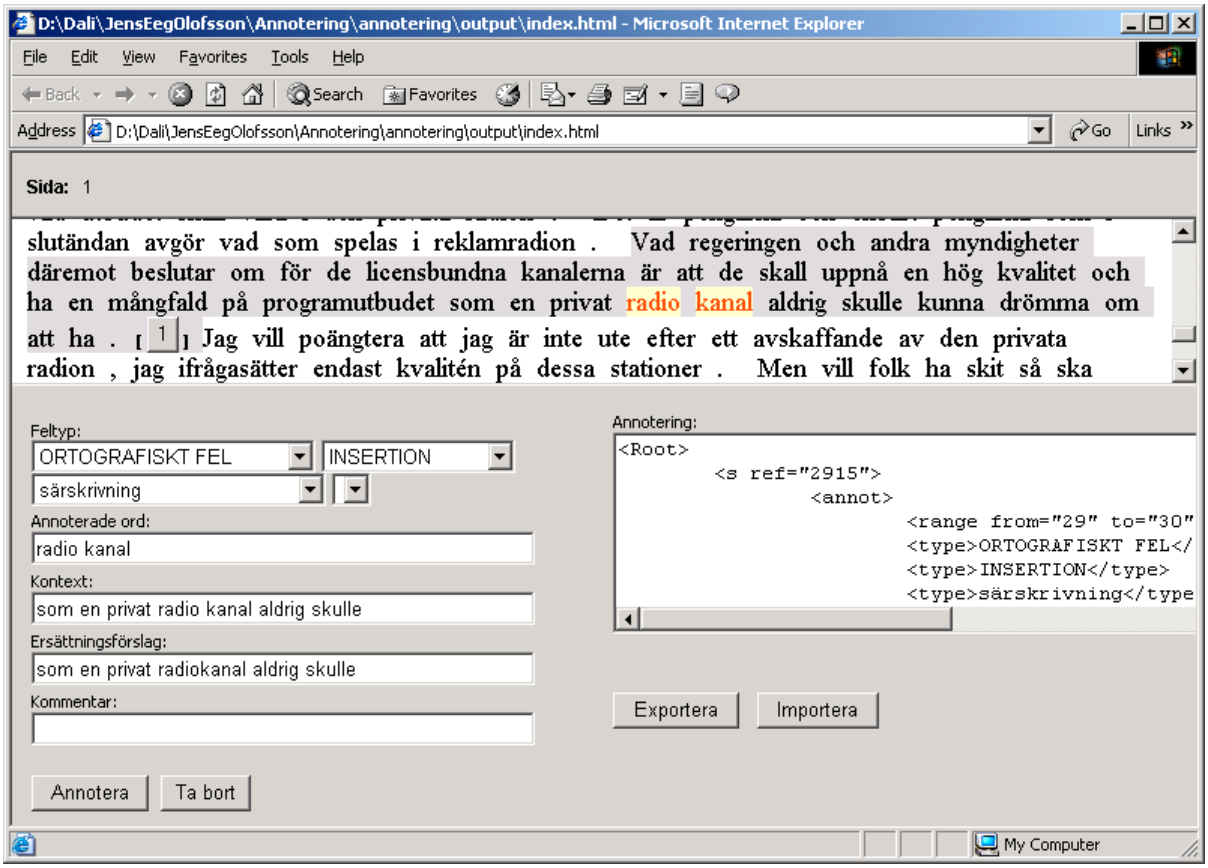
XML (Extended Markup Language) är ett uppmärkningsspråk som utvecklats med visionen att skapa ett universellt format för lagring och överföring av elektronisk information. XML kan också betraktas som en delmängd av SGML (Standardised Generalised Markup Language) – det filformat som under en tioårsperiod använts i språkvetenskapliga sammanhang för morfologisk och syntaktisk uppmärkning av text. SGML är i sin tur föregångare till HTML (Hypertext Markup Language) som blivit standardformatet för web-sidor, och det är inte minst i denna elektroniska miljö som XML på senare tid fått allt större betydelse, tack vare sin relativt enkla form och den därmed sammanhängande egenskapen att nästan alltid vara kompatibel med ett annat format.

Annoteringsarbetet sker också som vi ska se i ett HTML-dokument. Gymnasieuppsatserna är ursprungligen i Word-format och behöver formateras om på Nada och gå igenom Granska. Infilerna för annoteringsarbetet kommer därmed att utgöras av tokeniserade XML-versioner av uppsalafilerna.

2.2.1 Arbetsgång

Annotering sker först för hand på pappersutskrifter av uppsalamaterialet. Sedan annoteras en fil i taget i annoteringsverktyget. Ändringar i felkategorierna görs genom att programfilen som innehåller feltyperna i XML-kod öppnas i en texteditor där koden kan modifieras.

Annoteraren kan exportera annoteringarna till ett kodfönster där varje annotering (med felkategorier, kontext och ändringsförslag) representeras i XML-kod. Sedan kopieras koden och klistras in i en texteditor. Det är dessa kodfiler som sedan är tänkta att användas för vidareutveckling av Granskas regelsystem. XML-koden visar dessutom ytterligare information som finns inbyggd i systemet: varje mening i infilen har ett referensnummer. Dessutom har varje analysenhet ett positionsnummer i meningen. I kodfilerna med annoteringarna sorterar annoteringarna dels under referensnummer, dels under position i meningen.



I kodfönstret syns endast en del av koden för den aktuella annoteringen. Så här ser hela koden ut:

```
<Root>
	<s ref="2915">
		<annot>
			<range from="29" to="30"/>
			<type>ORTOGRAFISKT FEL</type>
			<type>INSERTION</type>
			<type>särskrivning</type>
			<text>som en privat radio kanal aldrig skulle </text>
			<comment></comment>
			<suggestion>som en privat radiokanal aldrig skulle </suggestion>
			<annotatedWords>radio kanal </annotatedWords>
		</annot>
	</s>
</Root>
```

Med utgångspunkt i andra arbeten inom feltaxonomi[3], dels för svenskt vidkommande i Granska-projektet (Öhrman 2000, Staerner 2001) och i Scarrie-projektet (Wedbjer Rambell 1998), dels internationellt (Vosse 1994; Bredenkamp, Crysman, Klein 1999), har ett slags prototyp-typologi konstruerats och implementerats i annoteringsverktyget. Prototypen (och i viss mån själva annoteringsverktyget) har sedan successivt reviderats under annotering av tio uppsatsfiler. En jämförelse mellan feltypologens och Granskas feldiagnoser har gjorts på en av filerna. Analysen har resulterat i en ny preliminär typologi. Prototyptypologin kallas i det följande **feltypologi A**. Den nya preliminära typologin kallas **feltypologi B**.

[3] Termen ”feltaxonomi” ska förstås ungefär som ”forskningsområdet ’feltypologi’”.

Feltypologi B har sedan prövats (och reviderats) under annotering av ytterligare tio uppsatsfiler, varefter resultatet har analyserats och utvärderats ånyo. Den feltypologi som växt fram ur denna andra utvärdering kallas för **feltypologi C**, och är identisk med den struktur som kommer att utgöra stommen för vidareutveckling av annoteringsverktyget samt för fortsatt feltypologiskt arbete utifrån texter av andraspråksinlärare.

Arbetet med prototyp-typologin avhandlas i det följande för sig och kallas för **försöksstudie**.

3 Försöksstudie

3.1 Feltypologi A

Feltypologi A fick till att börja med en struktur omfattande en överordnad kategorinivå och tre subnivåer.

(Vosse 1994) presenterar i sin avhandling om CORRie och grammatikbaserad stavningskontroll för holländska, fyra mycket generella klasser av fel och en femte kategori som särskilt beskriver särdragsproblem (Vosse, 1994). De fyra generella kategorierna, som också är kända som ”Levenshtein-operationer” från dynamisk programmering, introducerades i automatisk stavningskorrektion redan på sextiotalet (Kukich 1992: 394). I Vosses modell används kategorierna för att sortera misstag såväl på skrivteckennivå som på grammatisk nivå. På teckennivå tjänar de till att organisera olika typer av ”performansfel”, typiskt tangentbordsmissar. På grammatisk nivå återanvänds kategorierna för att beskriva syntaktiska fel, nu kompletterade med den femte kategorin **feature mismatch** (Vosse 1994: 104 f). Termen ”performansfel” är tänkt att användas komplementärt med en term ”kunskapsfel” (Vosse 1994: 35 ff). (Kukich 1992) talar om ”typografiska fel” vs ”kognitiva fel”.

Dessa generella felkategorier togs som utgångspunkt för feltypologi A och översta kategorinivån sammanföll i inledningsskedet med Vosses kategorier

Kategori	Beskrivning	Exempel
deletion	skribenten utelämnar en eller flera språkliga enheter	*Toleransen är olika hög hos människor vad de anser är fult språk.* *Det är bra att man **tabuläger** svordomar och könsord.*
insertion	skribenten lägger till en eller flera enheter	*De som är emot tycker att det inte behövs **inte** i språket.* *Men tyvärr använder man svordomar **allt för** ofta.* *Det stämmer **väll** ganska bra det.*
substitution	skribenten förväxlar en enhet med en annan	*Till exempel på en fotbollsmatch reagerar inte många **för** svordomarna.*
transposition	skribenten byter plats på enheter	*En sak **tycker jag** att vi alla kan hålla med om är att svordomar och könsord är onödigt i språket.*
feature mismatch	skribenten misstar sig på särdragsvärden, typiskt kongruensfel	*En sak tycker jag att vi alla kan hålla med om är att svordomar och könsord är **onödigt** i språket.* *Slangen anses som **fula** och opassande*

Tabell 3. Presentation av Vosses feltyper

Vosses kategorier visade sig vara ett mycket effektivt hjälpmedel i annoteringsarbetet, det vill säga felen lät sig i stort sett klassificeras på det här viset. En intressant egenskap hos dem är också att diagnosen på sätt och vis implicerar korrektionen. Ett fel i kategorin **insertion** korrigeras med en operation av typen **deletion** och vice versa.

De så kallade performansfelen är emellertid mycket svåra att diagnosticera. Det går mycket sällan att vara säker på om ett stavfel som *väll* beror på om skribenten råkat lägga till en enhet (performansfel) eller om skribenten permanent stavar ordet på det här viset (kunskapsfel). Anförda exempel antyder också att för lingvistisk intuition väldigt olika typer av fel kommer att sortera under samma kategori.

Låt oss titta i tabellen på kategori **insertion**. Ett satsadverbial för mycket, får samma ingång i analysen som ett stavfel, eller som mellanslaget i en särskrivning. Om feltypologin ska vara av lingvistisk karaktär förefaller utgångspunkten i Vosses feltyper mer eller mindre kontraintuitiv. Bör man inte ha en kategori som på en överordnad analysnivå skiljer ett stavfel som *väll* från ett klart syntaxbrott som att stoppa in ett satsadverbial för mycket? Stavfelet har i det här fallet begåtts på en nivå som inte rör betydelsebärande enheter. Man kan säga att det är ett fel på teckennivå, och det får heller inga konsekvenser över denna nivå. Satsadverbialet är emellertid en betydelsebärande enhet. Felet kan sägas ha begåtts på symbolnivå. För att undvika att denna för lingvistisk intuition grundläggande information hamnade för långt ner i analysen flyttades Vosses kategorier ner en nivå för att istället ge plats åt följande överordnade kategorier:

Kategori	Kommentar	Exempel
ortografiskt fel	Teckennivå. Till exempel stavfel.	*Men tyvärr använder man svordomar **allt för** ofta.* *Det stämmer **väll** ganska bra det*
syntaktiskt fel	Symbolnivå Detta ska förstås som syntaxfel på såväl sats- och frasnivå som på ordnivå, man kan tänka sig en subkategorisering i syntax och morfosyntax .	*De som är emot tycker att det inte behövs **inte** i språket.* *En sak **tycker jag** att vi alla kan hålla med om är att svordomar och könsord är **onödigt** i språket.*
semantiskt fel	Symbolnivå	*Vill Sverige vara ett land som talar som grottmänniskor?*

Tabell 4. Överordnade felkategorier

De första två nivåerna i feltypologi A, jämte Granskas paradfeltyper på nivå 3, ser då ut såhär

Nivå 1
ORTOGRAFISKT FEL
SYNTAKTISKT FEL
SEMANTISKT FEL

Nivå 2
DELETION
INSERTION
SUBSTITUTION
TRANSPOSITION
FEATURE MISMATCH

Nivå 3
särskrivning
inkongruens i nominalfras
inkongruens i predikativ

För att förtydliga feltypologin och hur den används, presenteras den nedan med avseende på ortografiska och syntaktiska fel, i tabellform med exempel.

Kategori nivå 1	**Kategori nivå 2**	**Kategori nivå 3**	**Exempel**
ortografiskt fel	DELETION		*likriktade **radokanaler*** *När **användsdetta** uttryck?*
	INSERTION	särskrivning	*En privat radio kanal* *Det stämmer **väll** ganska bra det* *En annan **artickel***
	SUBSTITUTION		*man ser nästan ned **vå** människan som talar* *på ett eller annat **sett***
	TRANSPOSITION		*radions **portibalitet*** ***(**”Att få välja sina ord – vilken lyx” Svd 22.2,1997)*
syntaktiskt fel	DELETION		*Jag håller med Lars-Gunnar Andersson att slang är fult*
	INSERTION		*De som är emot tycker att det inte behövs **inte** i språket.* *Man kan tala om vad man är för människa och till vilken grupp man **till**hör*
	SUBSTITUTION		*Till exempel på en fotbollsmatch reagerar inte många **för** svordomarna.*
	TRANSPOSITION		*En sak **tycker jag** att vi alla kan hålla med om är att svordomar och könsord är onödigt i språket.*
	FEATURE MISMATCH	inkongruens i nominalfras inkongruens i predikativ	*en **viktigt** mening* *Slangen anses som **fula** och opassande*

Tabell 5. Feltypologi A

3.2 Analys av resultat i försöksstudien

Efter annotering av tio uppsatsfiler kan man konstatera att huvudparten av felkategorierna på ett eller annat sätt finns representerade redan i den första uppsatsen. Nästan alla exempeltexter i detta avsnitt är hämtade ur den först annoterade uppsatsfilen. Nya kategorier har successivt tillkommit på nivå 3 och 4 som en naturlig följd av att nya typer av fel dykt upp. Men tillkommande felkonstruktioner och alternativa versioner av redan diagnosticerade fel har

också satt nytt ljus på redan gjorda annoteringar, och avslöjat brister eller inkonsekvens i analysen.

En omedelbar reflektion är vidare att kategorin **ortografiskt fel** har fått stor plats. Man kan också konstatera att preprocessningen av materialet i kombination med utformningen av annoteringsverktyget gett upphov till problem vad gäller konsekvensen i analysen, inte minst i just ortografisk felkontext.

I försökssstudien blir det också uppenbart, dels i hur förbluffande hög grad felanalysen är avhängig bedömningen hos den enskilde annoteraren, dels att analys- eller abstraktionsnivåerna på inget sätt är entydiga och okontroversiella. Sålunda sorterar till exempel **interpunktionsfel** i studien genomgående under **ortografiskt fel**. Men dessa ”ortografiska misstag” resulterar i inte så få fall i syntaktisk ambiguetet och hela uttrycket kan få en oavsedd tolkning, en ny betydelse helt enkelt. Det är alltså ingalunda självklart att vi ska kategorisera ett uteblivet kommatecken som ortografiskt fel, däremot har det visat sig lämpligast att göra så, av konsekvens- och läsbarhetsskäl (se exempel 11 nedan).

För normvägledning och grammatisk konsultation har i huvudsak dels Svenska skrivregler (Svenska språknämnden 2000), dels Svenska Akademiens grammatik (Teleman, Hellberg, Andersson 1999), använts[4]. En mer omfattande diskussion kring normproblematiken sparas till avsnitt 5.2.

Iakttagelserna kommer i det följande att exemplifieras och diskuteras, i första hand med avseende på ortografiska och sematiska fel. Syntaktiska fel avhandlas i jämförelsen med Granska.

3.2.1 Effekter av preprocessningen på annoteringsarbetet

I Uppsalafilerna har man markerat avstavning medelst lodstreck. I den tokeniserade XML-versionen representeras då ett avstavat ord som två ord med ett lodstreck emellan. Detta utgör egentligen sällan ett problem, vill man annotera ett ord som *lik | som* så är det ingenting som hindrar att man markerar tre analysenheter i infilen. Men någon annotering för felaktig avstavning har inte funnits anledning att göra och det är tveksamt om denna feltyp har någon större relevans för studien. Automatiska språkgranskningssystem är huvudsakligen utformade för elektroniska medier och därmed är feltypen utmanövrerad.[5] Vad som i själva verket är mest egenartat med Uppsalamaterialet är att originaltexterna är handskrivna, och vi ska se att denna egenskap får ytterligare konsekvenser för feltypologin.

Annoteringsverktyget är som tidigare beskrivits utformat på det sättet att annoteringar automatiskt sorteras efter placering i texten (referensnummer för meningsbörjan) och plats i meningen (positionsnummer för analysenhet). Detta innebär att annotering för substitution av kommatecken med punkt och därpå följande gemen med versal måste annoteras i två annoteringsenheter (med två olika platsreferenser). Punkt annoteras då i första meningen och versal istället för gemen annoteras i den efterföljande meningen och kommenterats som ”följdfel”, som i exempel 7:

[4] I det följande kommer (Teleman, Hellberg, Andersson 1999) att refereras till som SAG.
[5] Skrivnormen har i det här avseendet också förändrats snabbt, och avstavningsregler har tappat i betydelse, säkert i första hand som en konsekvens av att skrivverktyget för särskilt formell text i så hög utsträckning blivit elektroniskt.

7. *Eftersom vi människor har olika smak och åsikter om vad som är fint och vad som är fult. Kan man inte få ett enkelt och objektivt svar på frågan.*

```
<s ref="4039">
      <annot>
            <position pos="18" />
            <type>ORTOGRAFISKT FEL</type>
            <type>SUBSTITUTION</type>
            <type>interpunktionsfel</type>
            <text>som är fult . </text>
            <comment></comment>
            <suggestion>som är fult ,</suggestion>
            <annotatedWords>. </annotatedWords>
      </annot>
</s>
<s ref="4128">
      <annot>
            <position pos="0" />
            <type>ORTOGRAFISKT FEL</type>
            <type>SUBSTITUTION</type>
            <type>interpunktionsfel</type>
            <text>Kan man inte </text>
            <comment>följdfel</comment>
            <suggestion>kan man inte</suggestion>
            <annotatedWords>Kan </annotatedWords>
      </annot>
</s>
```

I till exempel de fall där punkt saknas kan man annotera i samma annoteringsenhet. Då blir det istället tveksamt vilket element man ska välja att markera för annotering – sista ordet i meningen som saknar punkt, eller kanske både detta ord och första ordet i påföljande mening. I försöksstudien har felmarkering gjorts enligt det senare alternativet (exempel 8). Det finns ingen möjlighet att markera tomrum (mellanslag).

8. *Men vad tycker jag om svordomar Jag tycker att det är onödigt och fult.*

```
<s ref="1036">
      <annot>
            <range from="5" to="6"/>
            <type>ORTOGRAFISKT FEL</type>
            <type>DELETION</type>
            <type>interpunktionsfel</type>
            <text>jag om svordomar Jag tycker </text>
            <comment></comment>
            <suggestion>jag om svordomar? Jag tycker </suggestion>
            <annotatedWords>svordomar Jag </annotatedWords>
      </annot>
</s>
```

Exempel 9 visar ett lite mer intrikat specialfall av syntax- och interpunktionsbrott.

9. *Man kan ge dem alterna|tiv på andra ord som man kan använda. Förklara för barnen att man blir mer respekt|erad som vuxen att använda smarta och välformulerande ord.*

Här har först kontexten ‹ . *Förklara* › markerats och annoterats som **insertion** och **interpunktionsfel**. ‹ *Förklara* › har sedan markerats och annoterats som **syntaktiskt fel / deletion / saknas satsfog** (konjunktion för att markera satsförkortning i det här fallet). Sedan har samma ord markerats igen och annoterats för **interpunktionsfel** (versal istället för gemen) och kommenterats som följdfel. Det går alltså bra att annotera för multipla fel i samma analysenhet, men man tvingas i princip göra en enhetsintern felanalys och bestämma sig för i vilken ordning felen bör annoteras. Dessutom kan man få problem med konsekvens i ändringsförslag (se avsnitt 3.3.1).

Naturligtvis är detta inte den enda möjliga analysen av exempel 9. Man skulle till exempel kunna tolka skribenten så att andra meningen i själva verket är en imperativ sats, en uppmaning till envar att förklara för barnen. Kvar som interpunktionsfel blir isåfall en substitution för utropstecken med punkt, allra sist. Dependensrelationen mellan interpunktion och syntax blir mycket tydlig.

Ett annat intressant fall av multipelt fel återfinns i tabellexemplet på **deletion** ovan.

10. *Det är bra att man **tabuläger** svordomar och könsord.*

Här markeras ‹ *tabuläger* › och annoteras för stavfel. Enheten kan emellertid markeras en gång till, beroende av i vilken utsträckning man anser att *tabulägger* saknar prefixet be- som i *tabubelägger*. Felparet kan tolkas som ett fall av performansfel och ett fall av kunskapsfel i enlighet med diskussionen ovan. Men återigen blir det hartnär omöjligt att avgöra om frånvaron av ett extra ‹ *g* › har att göra med skrivslarv eller om vi har att göra med en kunskapslucka. Man kan tänka sig att denna typ av stavfel (tillsammans med till exempel särskrivningar) i ett handskrivet material i större utsträckning utgörs av genuina kunskapsfel, jämfört med motsvarande fel i ett tangentbordsproducerat material.

Preprocessningen och utformningen av annoteringsverktyget aktualiserar också frågan om vilka analysenheter som ska få bestämma annoteringsspannet. Det förefaller ganska naturligt att infilen till annoteringsverktyget bär information om grafisk struktur – meningar och ord (eller skiljetecken). För lingvistisk intuition är det naturligare att utgå ifrån grammatisk struktur i analysen. Det ska visa sig att en för annoteraren rimlig tumregel är att i princip felanalysera inom fullständig sats (se avsnitt 3.3.2 samt avsnitt 4.1.4 och 6.1).

3.2.2 Ortografiska fel

Kategorin **ortografiskt fel** är som nämnts välrepresenterad. 36 av 72 annoteringar i första uppsatsfilen är diagnosticerade som ortografiska fel. Många av de fel som här subkategoriserats som interpunktionsfel, försvårar avsevärt läsningen av texten och felannotering förefaller i hög grad motiverad. Ett fel som upprepas i de flesta av uppsatserna är underlåtenhet att markera anföring och citat, typiskt i exempellistor där man refererar till ett språkbruk (exempel 11).

11. *De småord som retar är **va, ju, väl, då, så, hmm, liksom**.*

Denna typ av fel har blivit subkategoriserad som **deletion** och **interpunktionsfel**, med en ytterligare specifikation på nivå 4 – **citation eller anföring saknas**. Feltypen ser påtagligt ortografisk ut men feldiagnosen är starkt semantiskt färgad. Man kan också säga att konsekvensen av misstaget har en avsevärd semantisk tyngd. Men det är viktigt hålla i minnet att skribenten inte rimligen kan ha skrivit detta utan att mena att presentera en refererande

lista. En annorlunda läsning blir alltför bisarr och strandar syntaktiskt. Samtidigt är en dylik ”naiv” läsning naturligtvis helt okontroversiell för en maskin, och för regelkonstruktören till en automatisk korrekturläsare kan en sådan läsning vara konstruktiv. Men för den feltypologi vi utarbetar här vill vi ha en mer intuitiv läsning som utgångspunkt. Dessutom vill vi vid annoteringen fokusera misstag snarare än konsekvens och då är felet närmast ett brott mot skrivnormen och sorterar rimligtvis under ortografiska fel. Konsekvensen av misstaget är försämrad läsbarhet, men själva misstaget ligger i utelämnandet av vederbörligt tecken.

Förkortningar uppträder i olika skepnader.

> 12. *Som när man **tex** träffar sina kompisar så sätter man sig direkt framför TV:n istället för att **exv** lyssna på musik och ”snacka” lite allmänt.*

I exempel 12 har ‹ *tex* › annoterats som **ortografiskt fel / deletion / saknas mellanslag**, ‹ *exv* › har annoterats som **ortografiskt fel / deletion / stilfel** med ändringsförslag ‹ *exempelvis* ›.

3.2.3 Semantiska fel

Typiskt semantiska felanalyser hamnar som tidigare antytts ofta utanför Granska-systemets räckvidd. Men somliga fel som i förstone kan förefalla otvivelaktigt ortografiska (exempel 13), har också kommit att annoteras som **semantiskt fel:**

> 13. *mellan 6 – 12 års ålder*

med ändringsförslag ‹ *mellan 6 och 12 års ålder* ›. Substitution av konjunktion med bindestreck på det här sättet återkommer ofta i materialet och förekommer dessutom ofta i tal (om man översätter bindestrecket med ordet *till*). Den implicita relationen mellan bindestrecket och prepositionen *till*, samt det förhållande att ordet *mellan* styr frasinnehållet på ett sätt som analyserats som i första hand semantiskt, har föranlett valet av kategori på nivå 1.

Feltypen har vidare kategoriserats som **substitution** på nivå 2 och som **semantisk inkongruens** på nivå 3. Kategorin ”semantisk inkongruens” är tveksam eftersom termen ”semantisk kongruens” i SAG används i snävare bemärkelse. Termen syftar då på en konstruktion, typiskt en predikativ, där vad som framstår som semantisk kongruens, i språkbruket får företräde framför grammatisk kongruens (SAG 2: 226) som i *Köttbullar är gott*. (Kukich 1992: 416) använder däremot termen ”semantisk inkongruens” på samma sätt som i försöksstudien.

Exempel 14 och 15 illustrerar ett annat återkommande fel som trots allt ganska okontroversiellt kategoriserats som semantiskt fel, är felskrivning eller förvrängning av idiomatiska uttryck.

> 14. *Jag tycker det är fruktansvärt när barn **i den lägre åldern** skricker hora och andra mindre trevliga ord.*

> 15. *Det är **närmare** en statistisk fråga.*

‹ *i den lägre åldern* › respektive ‹ *närmare* › är annoterat som **semantiskt fel / substitution / fel i idiom** med ändringsförslag ‹ *i de lägre åldrarna* › respektive ‹ *närmast* ›. I den mån Granska ska kunna detektera den här typen av fel krävs ett idiom-lexikon. Ett litet sådant är

egentligen redan inbyggt i systemet – Granska signalerar för fel i fasta uttryck enligt Svensk handordbok (Knutsson 2001: 22).

Exempel 16 och 17 visar en konstruktion med kongruensproblematik och som i försöksstudien också annoterats som sådan.

16. *Jag kan förstå att **dessa orden** inte passar i mer formella sammanhang*

17. *Det fyller en funktion i **detta sammanhanget***

Vi har snarare att göra med en lätt otillbörlig dubblering än en disharmoniering av särdragsvärden, i det här fallet bestämdhet. Feltypen skulle kunna kategoriseras som sådan, alternativt som stilfel med kommentar för dialekt eller talspråk (konstruktionen är regel i till exempel göteborgska). Beskrivning av specieskongruens i svenska nominalfraser är en komplicerad avdelning och i SAG tvekar man att tala om kongruens i nominalfraser med avseende på bestämdhet (Knutsson 2001: 96 f).

3.2.4 Jämförelse med Granska

I den först annoterade uppsatsfilen detekterar Granska 20 fel (jämfört med feltypologens 72), varav hälften är antingen falska alarm eller feldiagnosticerade fel. Resultatet stämmer överens med tidigare utvärderingar av Granska för texter av gymnasie- och högskolestudenter (Knutsson 2001: 178). Dessa feldiagnosticerade fel visar sig i vissa fall innehålla viktig specifik information om hur systemet arbetar, men dessutom generell information om tolkningsalternativ vid syntaktiska tveksamheter (exempel 18 – 21).

18. *Det beror på, om man menar att ett **rätt ord** ska finnas med i Svenska Akademiens ordlista så är ööhh fel. Men om man menar att ett **rätt ord är när det** talas av personer med svenska som modersmål, så är ööhh ett rätt ord.*

Granska detekterar här båda instanserna av ‹ *rätt ord* › och diagnosticerar konstruktionen som särskrivning, i likhet med konstruktioner som *rätt stavning*. I försöksstudien har *rätt* accepterats som adjektiv och har bara annoterats i andra instansen, och då som **syntaktiskt fel / transposition / predikativ i fel position**. Man kan utgå ifrån att Granska inte detekterar några fel i detta avsnitt om man ersätter *rätt* med *riktigt*. Eftersom Granska har problem med *rätt* som adjektiv och eftersom det ur normsynpunkt förefaller rimligt att inte acceptera *ett rätt ord* eller för den delen *ett fel ord*, så bör dessa konstruktioner fortsättningsvis annoteras som till exempel **syntaktiskt fel / substitution / lexikonfel**. Andra instansen av konstruktionen i exempeltexten blir lämpligen två annoteringar, ‹ *rätt* › markeras och annoteras som lexikonfel och ‹ *är när* › markeras och annoteras för **syntaktiskt fel / deletion / saknas predikativ**.

SAG anför sex okontroversiella exempel på adjektivistisk användning av *rätt* eller *fel*, såsom *rätt lösning* och *rätt tro*. Analysen är beroende av semantisk kontext.

Ett liknande resonemang kan föras utifrån Granskas reaktion på texten i exempel 19

19. *Speciellt könsord för de är inte vackra.*

Här har Granska reagerat på ‹ *de* › och föreslår objektsformen. Det kan verka förvånande, men det är konjunktionen *för* som av Granska tolkas som preposition och föranleder feldetektionen. Detta betyder inte att Granska aldrig accepterar *för* som konjunktion, bara att

Granskas statistiska ordklasstaggare väljer den tolkning av *för* som är sannolikast givet den närmaste kontexten. I det här fallet är det emellertid inte lyckat att felannotera ‹ *för* ›. Med ett ersättningsförslag som ‹ *ty* › gör sig i själva verket annoteraren skyldig till ett stilbrott.

Ett tredje fall där det kan finnas anledning att så att säga kalibrera Granskas och feltypologens feldetektion (exempel 20), rör infinitivmärket *att*

20. *Jag tycker att vi ska fortsätta hitta på nya ord och uttryck.*

Granska reagerar på att skribenten uteslutit *att* före infinit verb. Feltypologen har låtit konstruktionen passera. På samma sätt har efter visst övervägande följande (vanliga) konstruktionstyp (exempel 21) lämnats oannoterad.

21. *Jag tycker bara det är roligt när man gör om ord och hittar på nya.*

Exempel 21 har således inte annoterats för utebliven satsfog (*att*). Första uppsatsen har fyra instanser av denna feltyp, och ingen detekteras heller av Granska. Man kunde hävda att ett inkonsekvent bruk av satsfogen skulle föranleda felannotering, men i så fall är man alldeles klart och otillbörligen uppe på ett metakontextuellt plan i felanalysen.

Det finnas anledning att ta hänsyn till Granskas felregler och fortsättningsvis annotera för utelämnat infinitivmärke, men låta konstruktionstypen i exempel 21 passera även fortsättningsvis utan annotering. Infinitivsatser är emellertid ett mycket komplicerat diagnoskapitel (SAG 4) och man kan knappast kräva en konsekvent analys av infinitivsatser inom ramen för detta arbete. I det här fallet bör man tillämpa intuition och stanna vid den.

Granska har regler för att ta hand om vad som kallas för kontamination, av typen *han tillhör en av de främsta trumpetarna i landet*, där korrektionen föreslår: antingen *...tillhör de främsta...* eller *Han är en av de främsta....* Granska reagerar emellertid inte på följande kontexter.

22. *Man kan ge dem **alterna\|tiv på andra ord** som man kan använda.*

23. *Man kan tala om vad man är för människa och **till vilken grupp man tillhör** genom modeuttrycken.*

Exempel 22 och 23 har i försöksstudien annoterats som **syntaktiskt fel / insertion / dubbelreferens**. Det är en betydligt mer generell analys som omfattar fler felkonstruktioner än tvättäkta kontaminationer. Feltypen diskuteras vidare i avsnitt 4.1.3.

24. (10.) *Det är bra att man **tabuläger** svordomar och könsord.*

Granska misstänker i exempel 24 faktiskt särskrivning och föreslår ‹ *tabulägersvordomar* ›, stavfelet har här på känt manér resulterat i ett annat ord och Granska tolkar *tabuläger*, som ett sammansatt substantiv. Kontexten med dubbla substantiv matchar felreglerna för särskrivning.

Intressant är att Granska bara genom att fokusera denna felkontext, uppmärksammade feltypologen på det multipla stavfelet i *tabuläga*. Granska detekterade dessutom två stavfel i uppsatsfilen som undgått den mänsklige annoteraren, nämligen *faoul* (en basket-term)

respektive *sistnämda*, och då måste man ändå säga att texten verkligen nagelfarits vid feluppmärkningen. En slutsats man kan dra är att det säkerligen skulle förbättra täckningen hos den mänskliga felanalysen att låta varje fil i materialet också gå igenom Granska. En utförligare jämförelse med Granska har av olika skäl inte fått plats plats inom ramen för föreliggande arbete.

3.3 Slutsatser av försöksstudien

3.3.1 Revision av feltypologi A

På översta nivån får kategorin **syntaktiskt fel** en explicit syskonkategori **morfosyntaktiskt fel**. Vi har egentligen med en subkategori att göra men då annoteringsverktyget i sin nuvarande utformning inte har en hierarkisk struktur utöver just nivåhierarkin hamnar den nya kategorin ändå på nivå 1. Ett komparationsfel som ***mer religiösare*** och ett annat komparationsfel som emellertid snarare är ett böjningfel, ***storaste***, får nu olika ingångar i analysen (syntaktiskt respektive morfosyntaktiskt fel), vilket svarar väl mot lingvistisk intuition. En felaktig konjugationsform som ***ätade*** blir klassificerad som morfosyntaktiskt fel och svävar på det viset inte i skymningslandet mellan ortografiskt (stav-) fel och syntaktiskt fel. Kategorin **morfosyntaktiskt fel** antas dessutom ha hög relevans för en utveckling av feltypologin för andraspråksinlärarmaterial.

Kvar som **stavfel** blir då antingen fel av fonologisk karaktär, exempelvis ***sjuta***, eller uppenbara så kallade performansfel, som ***med*** istället för ***men***, men eftersom tangentbordsfel inte kan förekomma i materialet och distinktionen performans- och kunskapsfel så sällan går att göra, kommer ingen subkategorisering att göras för **stavfel**.

Vosses kategorier är effektiva och sväljer i princip alla skrivfel under solen. Detta har att göra med graden av generell kapacitet hos feltyperna. I själva verket bär Vosses feltyper mycket lite språkspecifik information. Med en mer datalogisk infallsvinkel skulle man troligen skriva om tre av dem som antingen **insertion** eller **deletion** eller en kombination av dessa båda. De passar allra bäst för att ta hand om performansfel av typen tangentbordsfel. Hos Tono (Tono, 1999) sorterar motsvarande kategorier under ”ytstrukturell taxonomi” i motsats till ”lingvistisk kategoriklassificering”.

Trots att man alltså lätt kan argumentera emot en fortsatt närvaro av Vosse-kategorierna i feltypologin kommer de att finnas kvar. De utgör på sätt och vis en transitnivå där den lingvistiska analysen ”vilar” för sortering i generella klasser, och detta underlättar för fortsatt lingvistisk klassificering. Faktiskt så kommer den enda av kategorierna som i någon mån bär lingvistisk information, **feature mismatch**, att avlägsnas eftersom **feature mismatch** är ett specialfall av **substitution** (Vosse 1992). Tillräcklig information om särdragsinkongruens lämnas på nivå 3. Alla fel som annoterats för **feature mismatch** har också annoterats för särdragsinkongruens på nivå 3.

Transposition har i försöksstudien använts vid flyttningar över flera lexikala enheter. Denna analys är intuitivt överlägsen en analys i två annoteringar – en **deletion** och en **insertion**. Kategorin **transposition** gör kategorin **ordföljdsfel** överflödig. I exempel 25 markeras ‹ *kvar* › och annoteras som **predikativ i fel position**.

25. *En hel del människor värnar om sitt språk och att renheten **kvar blir** i det.*

Vad gäller de annoteringsproblem som hänger samman med utformningen av annoteringsverktyget så kommer uppmärkningen att fortsätta på inslagen linje. För konsekvens i ersättningsförslagen gäller dessutom: Ersättningsförslag anges successivt, det vill säga att redan gjorda diagnoser och korrektioner kommer att ”följa med” kontexten för ersättningsförslaget.

Förkortningar av typen ***exv*** kommer fortsättningsvis att annoteras för **substitution**, i övrigt behålles analysen för dessa feltyper.

Feltyperna i exempel 13 – 15 kommer även fortsättningsvis att behandlas som semantiska fel. Kategorin **semantisk inkongruens** utgår. ”Dialekt eller talspråk” blir kommentar till **stilfel**. ”Fel preposition” och ”fel adverbial” blir kommentarer till **lexikonfel**.

Dubblering av bestämdhet (exempel 16 och 17) felannoteras inte.

Felkategorin **saknas infinitivmarkör** avlägsnas, och i de fall feltypologen bedömer avsaknad av infinitivmärke som kandidat för feluppmärkning, annoteras felet som **saknas satsfog**.

Underlåtenhet att markera referens kommer att redan på nivå 3 kategoriseras som just sådan, utan att passera över **interpunktionsfe**l. Kategorin byter namn från ”citation eller anföring saknas” till **saknas markör för referens**. Detta var i princip den enda kategorin som kunde motiveras på nivå 4 och nivåerna har därför skurits ner till tre.

3.3.2 Kontextuell begränsning för felannotering

Det antyddes inledningsvis att gränsdragningen för vilken nivå man kan tillåta sig att lägga felanalysen på, är problematisk. Exempel 26 får illustrera.

> 26. *Ta till exempel enkla uttryck som att man är hungrig eller trött. För att förstärka dessa ord använder man ofta **fan** och **jävlar** framför. Varför gör man det? Kan man inte istället använda ord som hemskt eller fruktansvärt?*.

En möjlig felanalys utgår ifrån en tolkning av skribentens intentioner på så vis att ***fan*** och ***jävlar*** här är tänkta som kandidater till attribut till predikativen ***hungrig*** och ***trött***, på samma sätt som ***hemskt*** och ***fruktansvärt***. I så fall saknas dels markör för referens, alltså citationstecken, dels kongruerar inte kandidaterna med predikativen, dels upptäcker man att kandidaterna morfosyntaktiskt står mycket långt från en attributiv form, ***fan*** är rentav omöjlig att transformera för ändamålet. Problemet är att vi har att göra med en refererande användning av orden, de behöver egentligen inte ha attributiv form om man begränsar felkontexten till satsnivå. Man kunde mycket väl tänka sig att endast annotera ***fan*** och ***jävlar*** som **ortografiskt fel/ deletion / saknas markör för referens**, eftersom det är svårt att motivera någon annan, ska vi säga lokalsyntaktisk, felannotering. Ett tredje alternativ är att annotera för semantiskt fel, men frågan är på vilken nivå man i så fall diagnosticerat felet. Det är egentligen inte svårt att förstå vad skribenten menar och en eventuell semantisk feldiagnos blir i det här fallet starkt syntaktiskt färgad.

Exemplet har i försöksstudien annoterats för **inkongruens i predikativ** med en kategori **meningsextern inkongruens** på nivå 4. Denna senare kategori komplicerar ytterligare läsningen av kodfilerna. En lämpligare annotering av kontexten är att, som föreslagits ovan, markera ‹ *fan* › och ‹ *jävlar* › och annotera för underlåtenhet att markera referens, med ändringsförslag ‹ *”fan”* › och ‹ *”jävlar”* ›.

Med en konsekvent begränsning till lokalsyntaktisk felannotering skulle emellertid exempel 27 annoteras endast för substitution av frågetecken med punkt:

> 27. *Eftersom vi människor har olika smak och åsikter om vad som är fint och vad som är fult. Kan man inte få ett enkelt och objektivt svar på frågan.*

Det för utan tvekan för långt från lingvistisk intuition med ett så rigoröst lokaltextuellt spann för annoteringen. Ibland blir det också helt kontraintuitivt att inte ta hänsyn till exempelvis satsextern inkongruens som vid annotering av inkongruens i anafor. En rimlig slutsats är att fortsatt annotering görs inom fullständiga satser, utom i de fall vi har att göra med anaforisk referens, som alltså kan ha stort spann, men oftast inte har det. Anafor ska i det här sammanhanget förstås som en pronominell enhet som refererar bakåt mot en nominalfras, det så kallade korrelatet. [6]

> 28. *Är svordomar en krydda i språket eller anses **det** som stötande?*

I exempel 28 har ‹ *det* › markerats och annoterats för **syntaktiskt fel / feature mismatch / inkongruent anafor**, och kommer alltså fortsättningsvis att annoteras för **substitution** på nivå 2.

Ett par exempel till från försöksstudien får illustrera annoteringar med för stort spann.

> 29. *Om man ex tar min mamma som är kristen. Varenda gång som hon hör en svordom så blir hon ledsen och tycker att man gör gud "ledsen".*

I försöksstudien har exempel 29 annoterats över två annoteringsenheter som **transposition** och **adverbial** (‹ *varenda gång som hon hör en svordom* ›) **i fel position**, med ändringsförslag (efter tilläggskorrigeringar) ‹ *Om man exempelvis tar min mamma som är kristen så blir hon ledsen varenda gång hon hör en svordom och tycker att man gör Gud ”ledsen”.* ›. Annoteringsspannet är alldeles för stort.

> 30. *Det finns en anledning varför man exempelvis inte skulle svära om man träffade kungen och hans familj. För att **de** är fula ord som kränker en person.*

Här har ‹ *de* › markerats och annoterats som **semantisk inkongruens** med den lätt bisarra kommentaren ”anafor istället för korrelat” (ersätter man med ett ”korrelat” som *svordomar* blir konstruktionen överhuvud taget inte anaforisk). Också här har diagnosen för stort spann, och implicerar dessutom för mycket tolkning av intention hos skribenten. Hittar man inget korrelat som passar så får det styra diagnosen. Alltså **syntaktiskt fel / inkongruent anafor** med kommentar ”korrelat saknas”.

Härutöver kommer som redan antytts ytterligare kategorier i feltypologi A att försvinna eller revideras. Ändringar i den nya typologin – feltypologi B – får under det fortsatta annoteringsarbetet bara göras på nivå 3.

---

[6] I Chomskys GB-teori kallas anaforer med satsextern referens inte längre för anaforer utan helt enkelt för pronomen, eller ”referentiella uttryck” om explicit referens saknas.

4 Analys

4.1 Feltypologi B

I tabell 6 presenteras feltypologi B, dels som resultat av revisionen av feltypologi A (vänsterkolumn), dels som den såg ut efter annotering av ytterligare tio uppsatsfiler.

Nivå 1 SYNTAKTISKT FEL MORFOSYNTAKTISKT FEL ORTOGRAFISKT FEL SEMANTISKT FEL Nivå 2 INSERTION DELETION SUBSTITUTION TRANSPOSITION Nivå 3 stavfel särskrivning interpunktionsfel inkongruent anafor inkongruens nominalfras inkongruens ipredikativ böjningsfel saknas markör för referens saknas mellanslag saknas satsfog saknas subjekt saknas objekt saknas verb saknas preposition saknas partikel saknas optionsmarkör dubblering av satsfog dubblering av satsadverbial dubblering av verb dubblering av objekt dubbelreferens predikativ i fel position adverbial i fel position ofullständig sammansättning lexikonfel stilfel fel i idiom	Nivå 1 SYNTAKTISKT FEL MORFOSYNTAKTISKT FEL ORTOGRAFISKT FEL SEMANTISKT FEL Nivå 2 INSERTION DELETION SUBSTITUTION TRANSPOSITION Nivå 3 stavfel särskrivning interpunktionsfel inkongruent anafor inkongruens nominalfras inkongruens ipredikativ komparationsfel böjningsfel saknas markör för referens saknas mellanslag saknas satsfog saknas subjekt saknas objekt saknas verb saknas adjektiv saknas adverbial saknas preposition saknas partikel saknas optionsmarkör saknas infinitivmärke dubblering av satsfog dubblering av adverbial dubblering av verbfras dubbl av prepositionsfras dubblering av objekt dubbelsyftning predikativ i fel position adverbial i fel position hjälpverb i fel position ofullständig sammansättning felaktig satsförkortning lexikonfel stilfel fel i idiom

Tabell 6. Feltypologi B

Nivå 3 har inte oväntat fått fler kategorier. Vad som är smärtsamt tydligt i feltypologi B är att det är mycket svårt att se något sammanhang i kategorisalladen på nivå 3. Bland annat kan man konstatera att semantiska, syntaktiska, lexikala och ortografiska kategorier blandas på ett till synes godtyckligt sätt. En oistringent och inkonsekvent terminologi gör det hela än mer förvirrande.

Vad som troligen har skett är att egenskaper hos kategorierna på nivå 2 tillåtits fortplanta sig till nivå 3, så att det viktigaste i en feltyp som ”saknas objekt” fortsatt att vara att någonting saknas. Vilken typ av enhet som utelämnats har därmed fotfarande i princip inte ägnats någon analys utöver klassificeringen på nivå 1, och därmed är inte mycket vunnet. Klassiska satsfunktionstermer som ”subjekt” och ”adverbial” blandas till synes godtyckligt med ordklasstermer som ”verb” och ”adjektiv”. En feltypologi bör ha en grammatisk struktur som analyserar funktionsklass och lexikal klass på olika nivåer. Man behöver också se ett tydligare flödessamband vad gäller subkategorisering.

Dessutom används ogenerat kategorier som ”lexikonfel” och ”böjningsfel” utan vidare beskrivning (här utnyttjades i regel kommentar-möjligheten för att specificera). Dessa feltyper är alltför generella för att inte subkategoriseras och möjligen har vi att göra med ytterligare en sidoeffekt av egenskaperna hos de generella kategorierna på nivå 2. Det är också önskvärt att analysen där så är möjligt förs ner till lexikal nivå, med tanke på att ordklassanalysen är fundamental för Granska-systemet, men kanske också med tanke på att misstag, diagnos och korrektion för användaren nästan alltid ser ut att försiggå på ordnivå. Ett förslag till ny struktur för feltypologin (och annoteringsverktyget), nämligen den tidigare omtalade fetypologi C, kommer att presenteras sist i detta avsnittt.

4.1.1 Morfologiska fel

Kategorin ”böjningsfel” representerar en feltyp som givet definitionen i försöksstudien sorterar under morfosyntaktiska fel. Det gör den därför att misstaget och den syntaktiska konsekvensen på något sätt stannar inom ordet. Analysen implicerar existensen av en ordintern syntax. Detta är en egentligen onödigt okonventionell definition av begreppet morfosyntax. Vanligare är att betrakta *Det **har gått börjar gå** till överdrift* som syntaktiskt fel, och *Slangen anses som **fula** och opassande* som exempel på morfosyntaktiskt regelbrott.

(Bredenkamp, Crysmann, Klein 1999) har i sin tyska feltypologi en böjningsfelskategori ”morfologiskt fel”. Termen ”morfologiskt fel” svarar bättre mot vad som eftersträvas i försöksstudien med kategorin ”morfosyntaktiskt fel”. En feltyp ”morfologiskt fel” gör kategorin ”böjningsfel” överflödig. Alla böjningsfel, alltså kunskapsfel av typen ***storaste***, ***ätade***, som inte påverkar syntaxen på satsnivå, liksom avledningsfel, kommer automatiskt att sortera under ”morfologiskt fel”. Denna kategori kommer härutöver att prövas för att skilja ut de svåranalyserade stavfelen av kognitiv karaktär, till exempel ***aggrivisitet***.

Formen ***storaste*** är exempel på en så kallad övergeneralisering, alltså man har tillämpat en generell regel på ett undantagsfall. Man kan räkna med att denna feltyp har hög relevans för andraspråksinlärarmaterial. Men hur väljer man att klassificera ett fel som ***storaste*** på nivå 2? Är det en enkel substitution? Eller är det en komplex kombination av två instanser av substitution och en insertion? Hur man än använder Vosses fetyper för att klassificera böjningsfelet så blir analysen missvisande. Vosses kategorier passar bäst där de hörde hemma

till att börja med, nämligen under **ortografiskt fel** och under **syntaktiskt fel**. De lyfts ur feltypologin som subkategorier till morfologiskt fel och semantiskt fel.

Felet i *Blåvalen är det storaste däggdjuret* skulle i feltypologi B klassificeras som **morfosyntaktiskt fel / substitution / böjningsfel**. Feltypologi C kommer att erbjuda **morfologiskt fel / komparationsfel / övergeneralisering**.

Exempel 31 visar hur feltypologen arbetat med Vosse-kategorierna i feltypologi B för annotering av det ur typografisk synpunkt trippel-felstavade ordet ***aggrivisitet***

31.

```
<annot>
      <position pos="6" />
      <type>ORTOGRAFISKT FEL</type>
      <type>SUBSTITUTION</type>
      <type>stavfel</type>
      <text>bättre att all aggrivisitet kommer ut </text>
      <comment></comment>
      <suggestion>bättre att all aggrevisitet kommer ut </suggestion>
      <annotatedWords>aggrivisitet </annotatedWords>
</annot>
<annot>
      <position pos="6" />
      <type>ORTOGRAFISKT FEL</type>
      <type>DELETION</type>
      <type>stavfel</type>
      <text>bättre att all aggrivisitet kommer ut </text>
      <comment></comment>
      <suggestion>bättre att all aggrevissitet kommer ut</suggestion>
      <annotatedWords>aggrivisitet </annotatedWords>
</annot>
<annot>
      <position pos="6" />
      <type>ORTOGRAFISKT FEL</type>
      <type>TRANSPOSITION</type>
      <type>stavfel</type>
      <text>bättre att all aggrivisitet kommer ut </text>
      <comment></comment>
      <suggestion>bättre att all aggressivitet kommer ut</suggestion>
      <annotatedWords>aggrivisitet </annotatedWords>
</annot>
```

Vad beskriver annoteringen? Kanske hur korrektionen går till, på ett maskinellt, typografiskt plan. Men här finns ingen information om typ av misstag. Inte ens klassen ”stavfel” säger något specifikt om feltypen. Felet indikerar en oklar representation av ordet *aggressivitet* hos skribenten själv. Det hela rör sig antagligen om så mycket mer än ett stavfel, och den omständighet att ***aggrivisitet*** förekommer två gånger i samma text talar ytterligare för detta. Men det kan man fortfarande som feltypolog inte vara säker på. Och hur klassificerar man något man uppfattar som en ”oklar mental representation” av standardformen? Det är svårt att se hur man kan komma så mycket längre i beskrivningen av felet, eller ens av feltypen, än att klassificera det hela som morfologiskt fel och stavfel.

Ordet ***stadiumen*** har med feltypologi B felannoterats som morfosyntaktiskt fel. Lättar man på normkravet kan man hävda att ordet inte ska felannoteras alls, men exempel 32 visar tydligt

skillnaden i beskrivningskraft hos feltypologi B jämfört med feltypologi C. Exempel 32 visar annoteringen med feltypologi B

32.

```
<type>MORFOSYNTAKTISKT FEL</type>
<type>SUBSTITUTION</type>
<type>lexikonfel</type>
<text>på de lägre stadiumen . </text>
<comment></comment>
<suggestion>på de lägre stadierna . </suggestion>
<annotatedWords>stadiumen </annotatedWords
```

Här har ”lexikonfelet” som synes inte ens på kommentarraden fått någon vidare beskrivning. I feltypologi C kan man tänka sig olika subkategoriseringar, antingen med explicit ordklassangivelse som hos (Öhrman 2000)[7]

morfologiskt fel
felböjt substantiv
fel pluralform

eller utan

morfologiskt fel
fel i speciesböjning

Den senare varianten är tvetydig eftersom den kan betyda att skribenten har gjort en typ av kongruensfel och till exempel använt obestämd istället för bestämd form. En lösning är att subkategorisera för stilfel

morfologiskt fel
speciesfel
stilfel

En lämpligare lösning kunde vara att annotera även denna typ för **övergeneralisering**.

4.1.2 Fel på teckennivå

Kategorin **interpunktionsfel** är fortsatt högfrekvent i annoteringarna med feltypologi B, och fungerar i stort sett bra under förutsättning att man accepterar att kategorin saknar subkategorier som till exempel ”kommateringsfel” och ”versalfel”. Det skulle emellertid bära alltför långt i det här sammanhanget att ägna interpunktionsfelen samma ingående analys som syntaktiska och morfologiska fel, även om interpunktionsfelen som tidigare diskuterats, ofta nog kunde klassas som syntaktiska fel. Distinktionen mellan fel på teckennivå och fel på symbolnivå har emellertid visat sig fungera tillräckligt bra, och kvarstår.

Möjligheten att i analysen skilja ut olika typer av stavfel, kommer att prövas i feltypologi C. Vi har redan sett att stavfel av kunskapskaraktär kommer att sortera under morfologiska fel. Ortografiska fel resulterar som sagt ofta i syntaxbrott på satsnivå. Så är det också med de stavfel som blir ett nytt ord. Dessa stavfel får en egen subkategori. Subkategoriseras gör också

[7] Öhrman har dock ”ortografiska fel” som överordnad kategori för böjningsfelen

fonetiska stavfel av typen ***sjuta***. Stavfel av den typ vi kallat performansfel, som ***med*** istället för ***men***, får ingen subkategori.

Vosse-kategorierna är särskilt välmotiverade för sortering av fel på teckennivå då de utgör en del av beskrivningen av dessa fel. Ett återstående problem med Vosse-kategorierna är emellertid deras namn. Det verkar omotiverat att ha en engelskspråkig nivå i feltypologin, när resten är svenskspråkig. Vosse-kategorierna får alltså namn på svenska.

4.1.3 Kalibrering med Granska

En feltyp som i feltypologi A fick heta ”dubbelreferens”, har i feltypologi B döpts om till ”dubbelsyftning”, för att undvika förväxling med den typ av referens som avses i feltypen **saknas markör för referens**. **Dubbelsyftning** beskriver samma typ av fel som (Knutsson 2001) kallar ”kontamination”. Beteckningen i feltypologi B är oklar och svår att motivera. Nedan följer några exempel (exempel 33 – 38) på konstruktioner som med feltypologi B annoterats som **dubbelsyftning**, ibland som **semantiskt fel** men oftast som **syntaktiskt fel** på nivå 1, och alltid som **insertion** på nivå 2

33. *men det är fel om det överdoseras i mängder*

34. *Så ni får absolut inte missförstå mig alls*

35. *ett år senare använder ett par antal tusen om inte miljoner det som vardagsord*

36. *göra gott intryck ifrån sig*

37. *ge bra intryck ifrån sig*

38. *Vart kommer den att leda till?*

Samtliga anförda konstruktioner går utmärkt att beskriva som kontaminationer, det vill säga sammanblandningar av två uttryckssätt som orsakar mer eller mindre problem i syntaxen. Feltypologi C annammar termen **kontamination**. Feltypen får analysen **syntaktiskt fel / insertion / kontamination**.

Katagorin **saknas infinitivmärke** återinförs men används fortfarande med intuition som ledmärke.

Granskas lyckosamma analys av sammansatta verbkonstruktioner som ”verbkedjor” (Knutsson 2001) avspeglas i nya feltypologin.

4.1.4 Semantiska fel

Kategorin **semantisk inkongruens** återinförs. Dels finns stöd för en bredare användning av termen i (Kukich 1992), dels kan kategorin behövas för att beskriva brott mot vad som betecknas som semantisk kongruens i SAG.

39.
Psykisk misshandel finns det skäl att tala om men det var ju den aggressiva tonen som barnet reagerade på inte ordens betydelse i säg. Man måste ju skilja på att använda svordomar och på att kränka ***något oralt*** .

Feltypologen har annoterat exempel 39 så här med fetypologi B

```
<annot>
      <position pos="12" />
      <type>SEMANTISKT FEL</type>
      <type>SUBSTITUTION</type>
      <type>böjningsfel</type>
      <text>på att kränka något oralt . </text>
      <comment>kontexten kräver semantisk kongruens</comment>
      <suggestion>på att kränka någon oralt . </suggestion>
      <annotatedWords>något </annotatedWords>
</annot>
<annot>
      <position pos="13" />
      <type>SEMANTISKT FEL</type>
      <type>SUBSTITUTION</type>
      <type>lexikonfel</type>
      <text>att kränka något oralt . </text>
      <comment></comment>
      <suggestion>att kränka någon verbalt . </suggestion>
      <annotatedWords>oralt </annotatedWords>
</annot>
```

Den lokalsyntaktiskt mest rimliga läsningen av ***något oralt*** är som nominalfras och objekt till *kränka*, med *något* som determinerare och *oralt* som huvudord (”det är någonting som är av oral art som man kränker”). Vad skribenten alldeles uppenbart menar är emellertid att objektet, pronomenet *något* är en anafor som pekar tillbaka, kanske på *barnet*, och *oralt* ska fungera som adverbialt komplement (”det är någonting som man kränker på ett oralt sätt”). Problemet är att pronomenet *något* bara fungerar som ”fri anafor” – det kan inte vara bundet av ett explicit korrelat. Men det ska referera implicit till någonting, rimligen en människa (låt vara vem som helst) med namn och identitet, alltså en ”någon”. Vi har alldeles klart med ett semantiskt kongruensbrott att göra (notera att ett semantiskt fel i det här fallet alltså reulterat i syntaktisk tvetydighet). Som semantiskt fel klassas också det tvivelaktiga valet av ordet *oralt* istället för (det avsedda?) *verbalt*. I feltypologi C klassas då ena felet som **semantiskt fel / semantisk inkongruens** (**/ fel pronomenform**). Andra felet kategoriseras som **semantiskt fel / lexikonfel / fel adverbial**.

Före presentationen av feltypologi C några exempel på anaforisk referens (exempel 40 och 41) med stort avstånd mellan anafor och korrelat. Anafor, och vad som tolkats som korrelat, kursiveras:

40.
*Jag tvivlar på att ungdomen nu för tiden använder fler **slangord och uttryck** än ungdomen gjorde för femtio år sedan.*
*Att börja skylla på föräldrarna tycker jag är ett stort misstag. För den största delen av det fula språket lär sig nog ungdomen i skolan. Om de avhåller sig ifrån att säga **dem** hemma eller i sina föräldrars närvaro så är det inte så mycket de kan göra åt det.*

Annotering med feltypologi B:

```
<type>SEMANTISKT FEL</type>
<type>SUBSTITUTION</type>
<type>lexikonfel</type>
<text>ifrån att säga dem hemma eller </text>
<comment>långväga korrelat</comment>
```

```
<suggestion>ifrån att säga de fula orden hemma eller </suggestion>
<annotatedWords>dem </annotatedWords>
```

Denna typ av konstruktioner skulle enligt utvärderingen av försöksstudien motivera undantaget från regeln om satsinternt annoteringsspann. Men man måste nog säga att avståndet mellan korrelat och anafor är orimligt långt för att omfattas av analysen. Den analys som iallafall lämnats är dessutom samarbetsvillig i överkant, det ser nästan ut som om annoteraren trätt in för att försöka ”rädda” syntaxen i texten, ett tillvägagångssätt som i och för sig utmärker ett tidigt och mycket uppmärksammat program för automatisk språkgranskning, Critique (Kukich 1992; Knutsson 2001).

Exempel 41 (exempel 30 i avsnitt 3.3.2) visar på ett liknande problem

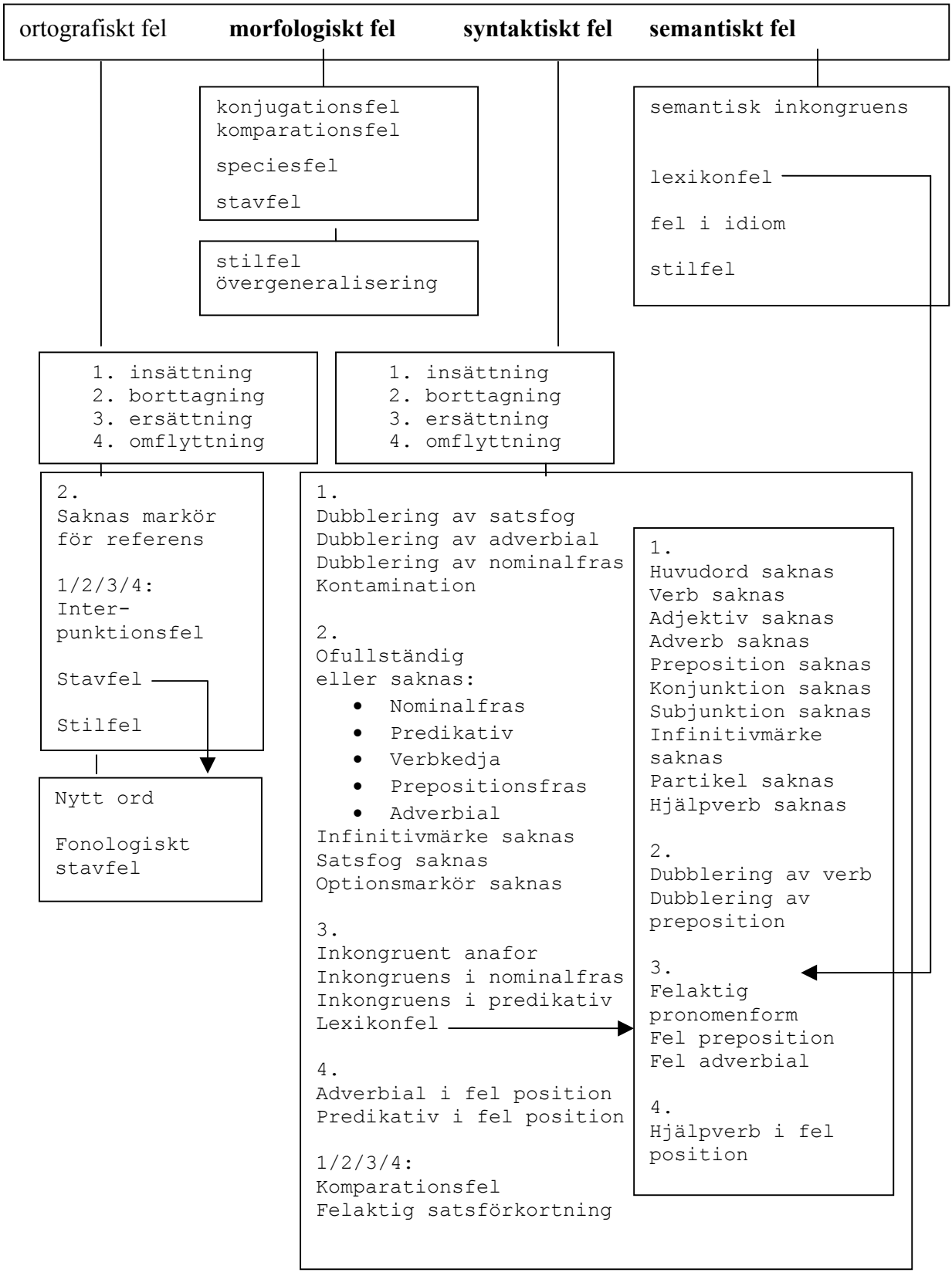
> 41.(30.)
> *Det finns en anledning varför man exempelvis inte skulle svära om man träffade kungen och hans familj. För att **de** är fula ord som kränker en person.*

Exemplen antyder att tumregeln, som säger att annotering bör göras inom huvudsats, snarare ska tillämpas undantagslöst. Exempel 40 och 41 blir då inte annoterade för vare sig syntaktisk eller semantisk inkongruens.

Det kan också vara värt att påpeka att fenomenet ”nytt stycke” som syns i exempel 40, inte syns i input-filerna till annoteringsverktyget, alltså de tokeniserade XML-versionerna av uppsatstexterna. Denna omständighet definierar på sätt och vis hur annoteraren bör förhålla sig till texterna. Redigeringsenheten ”nytt stycke” ligger klart utanför (eller över) den kontext som är aktuell för felklassificeringen.

5 Resultat Feltypologi C

Med nedanstående schema presenteras den resulterande feltypologin.



6 Diskussion

6.1 Vad visar annoteringarna?

Många gånger i föreliggande uppsats återkommer frågeställningen om vad som egentligen föranlett en annotering, och om annoteringen blir missvisande med en annorlunda analys av texten, eller kanske överflödig. Än uttrycks farhågor för att annoteringen är en analys av konsekvens snarare än misstag, än för att den snarare är en beskrivning av korrektionen.

Syftet med det feltaxonomiska arbete som beskrivits var i korta drag att komma fram till en ny feltypologi för Granskas annoteringsverktyg för att sedan kunna bygga vidare på feltaxonomin utifrån ett andraspråksinlärarmaterial, ett till synes mycket konkret ändamål. Men beskrivningen av annoteringsarbetet, och särskilt slutsatser som dragits ur problematiseringar och reflektioner, utgör i princip ett slags manual för hur annoteringsverktyget bör användas, hur olika feltyper bör klassificeras och så vidare. Beskrivningen ska kunna fungera som rättesnöre inte bara för uppsatsförfattaren i det fortsatta arbetet inom Granska-projektet utan även för andra användare av i första hand Granskas annoteringsverktyg. I valet av förfaringssätt för annoteringsarbetet ligger också med nödvändighet ett val av förhållningssätt till automatisk språkgranskning, och till språkkorrektion över huvud taget.

En återkommande reflektion rör oro för att fastna i ”tolkning av intention” hos skribenten. Men vad är ett automatiskt språkgranskningsprogram avsett för? Är det inte tänkt som ett hjälpmedel att göra sig förstådd? Kan det då inte vara lämpligt att i arbetet med att annotera autentisk text också anstränga sig för att förstå vad skribenten avser att förmedla?

Det kan vara lämpligt, men i princip fortfarande inte utöver satsnivå, och denna princip är viktig. Feltypologens uppgift är på sätt och vis att lägga sig någonstans mellan människa och maskin i analysen. Tanken är att feltypologen med den ”kallhamrade” lingvistiska analysen av felkonstruktioner lämnar det mest effektiva bidraget till utveckling av språkgranskningssystemet. Denna tanke bygger på premissen att språkgranskningsprogram behöver regler för grammatisk analys för att kunna fungera. Man kan visst hävda att det skulle vara önskvärt med språkgranskare som var känsliga för språkinnehåll på semantiskt-pragmatisk nivå, men vi saknar verktyg för sådant. Kärnan i språkgranskningssystemen är morfologisk och syntaktisk analys, och antagandet att text som produceras vid tangentborden har morfologisk och syntaktisk struktur. Utvärderingar får sedan visa om systemen blir bättre. Automatiska språkgranskningsprogram har fördelen att de kan utvärderas med hjälp av ”riktiga” användare.

Oron för att fastna i olika tolkningar av vad skribenten avsett att förmedla är i hög grad motiverad, särskilt som den feltypologiska uppgiften i det här fallet inbegriper identifiering av semantiska fel. Man kan också räkna med att problemet inte blir mindre vid annotering av andraspråksinlärartexter. Det kommer troligen att bli än mer angeläget att hålla sig till tumregeln om maximalt annoteringsspann över fullständig sats, och på denna nivå bör man styras av tolkning av vad skribenten vill förmedla. Det svåra är att undvika en sådan läsning på högre kontextuell nivå.

Det framgår vidare på flera ställen att ambitionen är att med annoteringarna beskriva själva misstaget och inte konsekvensen av misstaget. Det finns inget som säger att det ena är bättre

än det andra, poängen är att man inte bör blanda ihop de alternativa analyserna. Därför får ett misstag som i exempel 2

*Svenska svordomar är på ett eller annat **sett** kopplade till bibeln*

en uppmärkning för ortografiskt fel, under det att felet i exempel 3

*Deras betydelse är **mångbottnat***

blir annoterat som syntaktiskt fel, trots att båda resulterar i syntaxbrott.

Exempel 31 (‹ *aggrivisitet* › stället för ‹ *aggressivitet* ›) får representera ett tredje mycket intressant problem i den feltypologiska analysen. Felet har med feltypologi B annoterats mycket ”tekniskt” som en serie segmentella transformationer som beskrivs med hjälp av Vosses kategorier. Annoteraren uttrycker missnöje med analysen och flaggar för ”beskrivning av korrektion”. Men är det möjligen så att den metodiska och ”olingvistiska” annoteringen faktiskt bär information om vad som hänt? En verkligen rimlig diagnos är att betrakta felet som fonologiskt, därför att *aggressivitet* har samma tungvrickningsegenskaper som *CSN* som ofta blir *CNS* eller *CNN* i dagligt tal, eller *institution* som blir *inst**uti**tion*. Då skulle felet vara en helt konsekvent stavning av en fonologisk omsegmentering av den besvärliga standardformen. Exemplet belyser också de komplicerade relationerna mellan olika representationer av språket, förhållandet mellan fonologi och grafologi, och rent konkret nödvändigheten att tala om stavfel också i termer av fonetiska eller fonologiska fel.

Vad visar då annoteringarna? Sammanfattningsvis kan man säga att de visar en lingvistisk, lokalsyntaktisk analys av språkliga misstag, som utgår från en samarbetsvillig läsning av texten på samma lokala nivå.

6.2 Vad är rätt och vad är fel?

Analysen har en dimension av godtycke som är ofrånkomlig i dessa sammanhang och som har att göra med att felklassificeringen är beroende av den mänskliga annoterarens uppfattning om vad som är rätt och fel. Beskrivningen innehåller flera exempel på gränsfall, där den enskilde annoteraren blir tungan på vågen. I själva verket är det nog så att hela annoteringsarbetet är starkt färgat av annoterarens relation till fenomenet språknorm. (Knutsson 2001: 4) skriver om normen att ”det sverigesvenska skriftspråket är enligt Svenska Akademiens grammatik (Teleman et al, band 1, 1999) mycket enhetligt. I skriftspråket försöker man ofta hålla sig till den språkliga normen för att undvika oklarheter.”

Normen tycks verkligen vara ganska enhetlig på ett ortografiskt plan. ”Svenska skrivregler” (Svenska språknämnden 2000) avhandlar definitionsmässigt bara skrift och är utpräglat normativ. Det har genomgående rått samstämmighet mellan å ena sidan feltypologens intuition och å andra sidan anvisningar och rekommendationer i ”Svenska skrivregler”, en omständighet som förstås besparat feltypologen en del ansträngande avvägningar eller frustrerande eftergifter.

Ett material bestående av refererande gymnasieuppsatser visar emellertid med all önskvärd tydlighet att norm och stil hänger intimt samman och att konstaterandet om skriftspråkets enhetlighet verkar underligt när man studerar språket i ett innehållsligt perspektiv.

Diskussionerna omkring *rätt* som adjektiv, infinitivsatser eller enskilda ord som *tabulägga* eller *stadiumen* visar att normen också så att säga är kontextberoende. SAG är heller inte, och ska inte vara, normativ. Ofta beskrivs alternativa konstruktioner som optionella, inte bara i talspråk, och i den mån man väljer att lyfta fram ett alternativ framför ett annat så gör man det med hänvisning till att ”språkvårdare rekommenderar” den ena eller den andra konstruktionen.

Under arbetet med feltypologi B valde feltypologen till slut att inte annotera *svärord* för stilbrott eller talspråk eftersom ändringsförslaget i vissa fall resulterade i ett omvänt stilbrott, analogt med resonemanget kring *ty* som ersättningsförslag till *för* i exempel 19 (avsnitt 3.2.4). Bedömningen är naturligtvis subjektiv.

Det är alltså svårt att tänka sig en generell norm för skrivet nationalspråk. (Knutsson 2001) rapporterar mycket olika resultat av utvärdering av Granska på professionella texter jämfört med till exempel studentuppsatser, och kommer till slutsatsen att det inte är rimligt att använda samma version av verktyget för granskning av så olika genrer och att en utveckling av en möjlighet för användaren att välja granskningsalternativ (regeluppsättning) är önskvärd. Detta betyder inte att olika versioner skulle förhålla sig olika ortodoxt till en gemensam norm. Normen är helt enkelt olika för olika genrer.

Det är viktigt att vara medveten om att en feltypologi, och kanske i ännu högre grad själva språkgranskningsprogrammet, är färskvaror som ständigt behöver uppdateras. Det är visserligen oklokt att föregripa en sannolik utveckling och till exempel låta bli att annotera *tabulägga* för uteblivet prefix, men en sak kan man vara alldeles säker på, och det är att norm ingalunda är ett statiskt fenomen.

Det råder ingen tvekan om att särskrivningar som *radio kanal* ofta försvårar läsningen av en text och att uppmaningar till offentligheten av typen *rökning och alkohol förtäring förbjuden* är vanliga, och irriterande för många. De allra flesta människor tycks emellertid inte uppmärksamma ”särskrivningseländet” och denna omständighet (men även andra omständigheter, till exempel att särskrivningar verkar ha florerat också vid sekelskiftet) indikerar att sammansättningen inte behöver vara så fundamental för svenskan, och framför allt att den inte behöver vara evig. Därför, men också av andra skäl, kunde man motivera särskrivning av sammansättning av substantiv. Man skulle då undvika tvetydigheten i sammansättningar som *torparvinge*.

Utläggningen om särskrivningar tjänar till att illustrera hur viktigt det är att feltypologen med avseende på språknorm förhåller sig så förutsättningslöst som möjligt till det studerade materialet. På så vis kan typologen hoppas på att lägga sig på en analysnivå mellan människa och maskin. Ibland bryter emellertid ambitionen samman, vilket får illustreras av följande lätt desperata annotering.

```
<range from="4" to="8"/>
<type>SEMANTISKT FEL</type>
<text>Istället låter de censurens dova och intetsägande pip skrika .
</text>
<comment>semantiskt motsägelsefullt och absurt,
okorrigerbart</comment>
<suggestion></suggestion>
<annotatedWords>dova och intetsägande pip skrika </annotatedWords>
```

Referenser

Litteratur

Becker, Bredenkamp, Crysmann, Klein (DFKI Saarbrücken (1999)) "Annotation of error types for german news corpus", chapter 1. I *Proceedings of ATALA Workshop*, Paris.

Carlberger, Domeij, Kann, Knutsson (2000) *A Swedish Grammar Checker,* skickad till Computational Linguistics feb 2001.

Carlberger, Domeij, Kann, Knutsson (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.

Domeij, Knutsson, Larsson, Severinson Eklundh, Rex (1998) *Granskaprojektet 1996-1997*. Technical Report, IPLab-146, Nada, KTH, Stockholms Universitet.

Haegeman, L (1991) *Introduction to Government & Binding Theory* Basil Blackwell.

Knutsson, O (2001) *Automatisk språkgranskning av svensk text* Licentiatavhandling. Institutionen för numerisk analys och datalogi, KTH, Stockholm.

Kukich, K (1992) "Tecniques for Autimatically correcting Words in Text". *ACM Computing Surveys* 24:4 ss 337-439.

Nivre, J (1990) *Elementai i generativ grammatik*. Göteborgs universitet. Institutionen för lingvistik.

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

Svenska språknämnden (2000) *Svenska skrivregler* Liber.

Teleman, Hellberg, Andersson (1999) *Svenska Akademiens grammatik* Band 1 – 4, Nordstedts Ordbok, Stockholm.

Vosse, T (1994) *The Word Connection. Grammar-Based Spelling Error Correction in Dutch*. Doctoral Dissertation. Enschede: Neslia Paniculata.

Wedbjer Rambell, O (2000) *Error Typology for Automatic Proof-reading Purposes* Examensarbete, Språkteknologiprogrammet, Institutionen för lingvistik, Uppsala universitet.

Öhrman, L (1998) Felaktigt särskrivna sammansättningar. C-uppsats i datorlingvistik, Institutionen för lingvistik, Stockholms universitet.

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

Elektroniska publikationer

Lingsoft (2001)	http://www.lingsoft.fi/grammatifix/features/index.html
Tono, Y (1999) *Error Taxonomy*	http://www.lancs.ac.uk/postgrad/tono/taxonomy.html
XML Sweden (2001)	http://www.xml.se