

Utveckling av lexikala resurser för ett språkgranskningssystem för svenska

Leif-Jöran Olsson
ljo@stp.ling.uu.se

Examensarbete i datorlingvistik
Språkteknologiprogrammet
Uppsala universitet · Institutionen för lingvistik och filologi

15 januari 2004

Handledare:
Anna Sågvall Hein, Uppsala universitet
Biträdande handledare:
Mats Dahllöf, Uppsala universitet

Sammandrag

Denna uppsats beskriver utveckling, anpassning och utvärdering av tre lexikala komponenter i ett språkgranskningssystem för svenska, Scarrie-piloten. De tre komponenterna är: ett ordformslexikon, en sammansättningsgrammatik, och en grafem-till-fonem-överföringsgrammatik.

Ordformslexikonet har sammanställts från en tidningstextkorpus och innehåller godkända ord och fraser, samt icke-godkända ord och fraser. Lexikonet lagras i en lexikal databas, ScarrieLex, för att underlätta underhåll och uppdateringar. Lexikoningångarna/posterna konverteras automatiskt vid export från databasen till ett trie-baserat uppslagningsformat som används vid körningen av piloten.

Den produktiva ordbildningsprocessen med konkatenerade sammansättningar i svenskan gör det nödvändigt att känna igen icke-lexikaliserade sammansättningar. Sammansättningsanalysatorn interagerar med ordformslexikonet och är implementerad som en deterministisk finit automat (DFA). Sammansättningsreglerna, som uttrycker kategoriberoenden, representeras som reguljära uttryck.

Grafem-till-fonem-överföringen har baserats på en sammanställning av en specifikation av fonemrepresentation och en grammatik för generering av fonetiska representationer. Grafem-till-fonem-överföringsgrammatiken bygger på en kontextberoende formalism, där regler opererar i tre steg, med utgångspunkt i ordformssträngen. Överföringsgrammatiken används som en del i en kombinerad ljud- och stavningsbaserad ordkorrektur för svenska.

Systemevalueringen och komponentfunktionsvalideringen följer attributbaserade riktlinjer.

Nyckelord: Adaption, integration, dokumentation, standardisering, lexikala resurser, utvärdering och validering.

Förord

Arbetet som ligger till grund för denna uppsats har utförts vid Institutionen för lingvistik, Uppsala universitet, och har finansierats via EU-projektet Scarrie (LE3-4239).

Först vill jag tacka huvudhandledare och projektkoordinator Anna Sågvall Hein och biträdande handledare Mats Dahllöf för deras råd och stöd.

Jag vill också tacka följande personer involverade i Scarrie-projektet eller helt enkelt för att ha varit arbetskamrater under tiden arbetet med uppsatsen förflutit: Bengt Dahlqvist, Ebba Gustavii, för hennes återkoppling på sammansättningsanalysen, Olga Wedbjer Rambell, för hennes skarpsinne och entusiastiska diskussioner, Per Starbäck, för hans filosofiska inställning, Jörg Tiedemann, för hans arbete med den första versionen av ScarrieLex-databasen och samarbetet därefter, Erik Tjong Kim Sang, dina tofflor är i tryggt förvar :), Per Weijnitz, för uppfriskande implementationsdiskussioner, Eva Forsbom, för stöd i lexikonrensningsfrågor samt tämjandet av slutsatserna, Lars Borin, för mönstermodellen över vägen mot vishetens källa, och Jenny Wiksten Folkeryd för hennes uppskattade hjälp med uppställningen av den fonemiska representationen.

Jag vill tacka Theo Vosse för hans snabba och hjälpsamma svar på frågor angående den ursprungliga Corrie-prototypen i början av projektet.

Dessutom vill jag tacka Richard M Stallman¹ för de ovärderliga verktyg som framkommit genom hans förmåga att inspirera till att utveckla och sprida datormjukvara enligt de fyra friheterna: 0. Frihet att köra programmet för vilket ändamål som helst. 1. Frihet att analysera och ändra programmet efter behov. 2. Frihet att sprida kopior av programmet. 3. Frihet att förbättra och vidare distribuera programmet med dessa förbättringar för att gagna allmänheten.

Slutligen ett stort tack till Ylva Berglund, Per Weijnitz, och Karin Erlandsson för deras hjälp med korrekturläsning av uppsatsen i någon av dess olika faser.

¹Grundare av Gnu-projektet och Free Software Foundation.

Innehåll

Förord	
Innehållsförteckning	
Tabeller	iv
Figurer	v
1 Inledning	1
1.1 Syfte	1
1.1.1 Lexikonet — Scarrie master dictionary	2
1.1.2 Sammansättningsanalys	2
1.1.3 Uttalsbaserad korrektion	3
1.1.4 Överväganden och förberedelser inför adaptionen	4
1.2 Uppsatsens upplägg	4
2 Bakgrund	5
2.1 Korrektion — generering av ersättningsförslag	5
2.2 Standardisering	6
3 Lexikonet	8
3.1 Källor för det svenska ordformslexikonet	8
3.2 Urvalsförfarandet	9
3.3 ScarrieLex — en lexikal databas	9
3.3.1 Struktur och innehåll i ScarrieLex	10
3.3.2 Finjustering	13

3.3.3	ScarrieLex och det intermediära lexikonformatet	13
3.4	Några ord om framtiden	14
4	Sammansättningsanalys	15
4.1	Sammansättningsgrammatiken	15
4.2	Sammansättningsregler	17
4.2.1	Sammansättningsbarhetsattribut	17
4.2.2	Kommenterade sammansättningsregler	18
4.2.3	Sammansättningsregelutvecklingsprocessen	18
4.2.4	Finjusteringar	20
4.3	Sammansättningsutvärdering	21
4.4	Morfologisk avledning	21
4.5	Några ord om framtiden	21
5	Grafem-till-fonem-överföring	23
5.1	Grafem-till-fonem-överföringsregler	25
5.1.1	Grafonematiska relationer i svenskan	25
5.1.2	Steg-ett-regler	25
5.1.3	Steg-två-regler	29
5.1.4	Steg-tre-regler	29
5.2	Utvärdering av uttalsreglernas effekt på förslagskvaliteten	30
5.3	Några ord om framtiden	30
6	Systemevaluering och komponentvalidering	31
6.1	Evalueringsmetoder	31
6.2	Validering av funktionsattribut	32
6.2.1	Resultat av körning på testsviterna	32
6.2.2	Resultat för körning på testtext	33
6.3	Systemevaluering	33

6.4	Några ord om framtiden	36
7	Slutsatser	37
	Referenser	39
	Bihang	42
A	ScarrieLex	43
A.1	Struktur	43
A.1.1	Lemma — svlemma	43
A.1.2	Lexem — svlexeme	43
A.1.3	Stam — svstem	43
A.1.4	Affix — svaffix	44
A.1.5	Morfomvandlingsmönster — svpattern	44
A.1.6	Ordklass — svpos	44
A.1.7	Böjningsmönster — svinflexion	45
A.1.8	Morfosyntaktisk kod — svmorphcode	45
A.1.9	Frekvens på ordformsnivå — svfreq	45
A.1.10	Stilinformation på ordformsnivå — svstyle	45
A.1.11	Sammansättningsegenskaper på ordformsnivå — svderiv	46
A.1.12	Ersättningsalternativ på ordformsnivå — svreplacement	46
A.1.13	Stilinformation på lemmanivå — svlemstyle	48
A.1.14	Sammansättningsegenskaper på lemmanivå — svlemderiv	48
A.1.15	Ersättningsalternativ på lemmanivå — svlemreplacement	48
A.1.16	lexikon — lex	49
A.2	Användning av ScarrieLex	49
B	Användning av den svenska Scarrie-demonstratorn	53
B.1	Input	54

B.2	Output	54
B.2.1	Diagnos/ersättningsförslag	55
C	Testord och exempelkorrektioner använda under utvecklingen av grafem-till-fonem- överföringsreglerna	56
C.1	Testord	56
C.2	Exempelkorrektioner	56

Tabeller

4.1	Syntax för reguljära uttryck i sammansättningsgrammatiken.	16
4.2	Specialtecken i reguljära uttryck i sammansättningsgrammatiken.	16
4.3	Attribut som begränsar sammansättningsbarheten i grammatiken.	17
4.4	Igenkända sammansättningar, del ett.	19
4.5	Igenkända sammansättningar, del två.	20
5.1	Fonemrepresentationer av grafem.	24
5.2	Effekt av uttalsreglerna på förslagskvaliteten.	30
6.1	Sammanställning av resultaten för testsvitsvalideringen.	32
6.2	Täckning för ordfel och förslagskvalitet för testsviterna (från [Sågvall Hein et al 1999]).	33
6.3	Generell funktionssammanställning för systemevalueringen.	34
6.4	Kategoriserad funktionssammanställning för systemevalueringen.	35
7.1	Evalueringsjämförelse med MS Word 97.	38
C.1	Testord använda under utvecklingen av grafem-till-fonem-reglerna.	56

Figurer

2

1.2	Finit automat för sammansättningsexempel.	3
3.1	Användargränssnittet för arbete med ScarrieLex - vy lexikon.	10
3.2	Användargränssnittet för arbete med ScarrieLex - vy lemma.	11
3.3	Användargränssnittet för arbete med ScarrieLex - vy stam.	11
3.4	Användargränssnittet för arbete med ScarrieLex - vy ordform.	12
3.5	Användargränssnittet för arbete med ScarrieLex - vy böjningsmönster.	12
A.1	Exempel på information i fältet svstem.	44
A.2	Stilinformationsexempel ett.	46
A.3	Stilinformationsexempel två.	47
A.4	Ett frasersättningsexempel.	47
A.5	Exempel på lemmastilinformation och ersättning av alla former för hela lemmat.	48
A.6	Begränsning av sökning med hjälp av operatorn LIKE och jokertecken %.	49
A.7	Begränsning av sökning med hjälp av operatorn LIKE och jokertecken _.	50
A.8	Begränsning av sökning med hjälp av operatorn LIKE (storleksberoende).	50
A.9	Begränsning av sökning med hjälp av operatorn = (identitet, storleksberoende).	51
A.10	Begränsning av sökning med hjälp av skilt från-operatorn <> (not).	51
A.11	Begränsning av sökning med hjälp av operatorn REGEXP.	52
B.1	Webbgränssnittet för den svenska Scarrie-demonstratorn.	53
B.2	Granskningsrapport från webbgränssnittet.	54

Kapitel 1

Inledning

Det EU-finansierade projektet Scarrie (LE3-3249), Scandinavian Proof-reading Tools, har haft som mål att utveckla ett högkvalitativt korrekturläsningshjälpmedel, hädanefter kallat Scarrie-piloten eller endast piloten, för den skandinaviska publiceringsindustrin, och att minska skillnaderna i tillgång till språkgranskningsresurser mellan de skandinaviska språken och de större europeiska språken. För bakgrundsinformation, se Scarrie Technical Annex, [Scarrie TA 1996]. Slutanvändare från publiceringsindustrin, har deltagit både som partner och underleverantörer i projektet för att försäkra att Scarrie-piloten är av högsta kvalitet för den tänkta användargruppen.

Scarrie-piloten är baserad på en holländsk programprototyp för icke-interaktiv stavningskorrektion, Corrie [Vosse 1994]. Denna prototyp fanns uppfylla kraven för att vara utgångspunkten för det nya korrekturläsningssverktyget [Tjong Kim Sang 1997]. Corrie-prototypen anpassades till de tre skandinaviska språken danska, norska, och svenska genom en process av adaptation och till någon del internationalisering och lokalisering, se avsnitt 2.2 för definitioner.

Arbetet med den svenska Scarrie-piloten utfördes vid Institutionen för lingvistik, Uppsala universitet. Projektkoordinatör var professor Anna Sägval Hein.

Uppsatsen beskriver utvecklingen av de språkspecifika resurserna och valideringen av den språkliga funktionaliteten. Dessutom tas adaptationen av den svenska versionen av piloten upp.

1.1 Syfte

Uppsatsens syfte är att beskriva utvecklingen av de tre lexikala resurser som jag utvecklat för språkgranskning av svenska med Scarrie-piloten:

1. Lexikonet — Scarrie master dictionary
Lagras och underhålls i en lexikal databas, ScarrieLex. Kan användas tillsammans med domän- och användarspecifika lexikon.

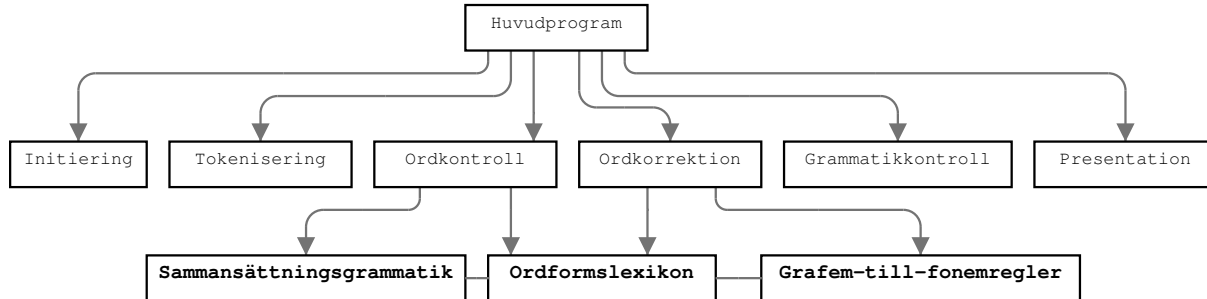
2. Sammansättningsanalys

Ordformskonkatenering med prefixavledning. Suffixavledning utförs med enkel strängbaserad analys.

3. Uttalsbaserad stavningskorrektur

En uttalslikhetsjämförelse används tillsammans med editeringsavståndsbaserad information i stavningskontrollen för att korrigera kompetensfel.

De tre lexikala resurserna är markerade i arkitekturöversikten i figur 1.1.



Figur 1.1: Arkitekturöversikt¹

1.1.1 Lexikonet — Scarrie master dictionary

Lexikon utgör en mycket viktig del av ett korrekturläsningssystem. Scarrie-systemet är inget undantag. Lexikonmaterialet är korpusbaserat och sammanställt från tidningstext (se kapitel 3), eftersom systemet just är tänkt att stödja korrekturläsning på svenska tidningsredaktioner. I Scarrie-piloten kan man använda två typer av lexikon. Kompilerade och icke-kompilerade. Båda typerna presenteras nedan.

Kompilerade lexikon är tänkta att svara för huvuddelen av den kvantitativa lexikala täckningen och därför vara mer långsiktiga och förändras långsamt. Kompilerade lexikon är optimerade för snabb inläsning till internminnet och för snabb uppslagning vid körning av Scarrie-piloten. Den lexikala informationen exporteras från den lexikala databasen ScarrieLex för användning i Scarrie-piloten. Se avsnitt 3.3.1 för en utförligare genomgång av informationen som finns i Scarrie master dictionary.

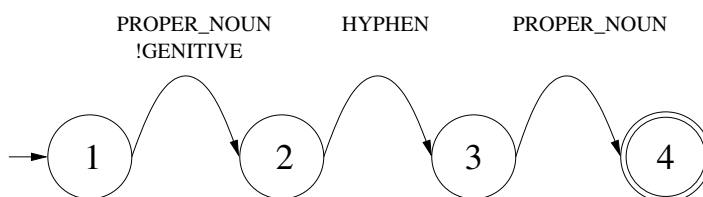
De icke-kompilerade lexikonerna innehåller inte lika mycket information som de kompilerade. De icke-kompilerade lexikonerna används endast tillsammans med minst ett kompilerat lexikon, som supplement för individuella behov inom specifika domäner. I icke-kompilerade lexikon kan man lägga både enskilda ord och fraser. Inläsningen av icke-kompilerade lexikon sker efter inläsningen av kompilerade lexikon. Eftersom icke-kompilerade lexikon är rena textfiler, läses de in mycket långsammare, och ingångarna lagras tillsammans i det interna lexikonet under körningen, är det en fördel om antalet ingångar i icke-kompilerade lexikon inte är allt för många.

1.1.2 Sammansättningsanalys

Stavningskontroll är i grunden lexikonbaserad på ett eller annat sätt, vanligtvis används explicita lexikon; fullformslexikon eller stamlexikon med morfologisk analys eller implicita lexikon baserade på

statistisk metoder. Ibland förekommer även kombinationer av explicita lexikon och statistiska metoder. [Kukich 1992]:380–385. Lexikonet sätter gränserna för hur bra ett stavningskontrollprogram är. Ordbildning i form av sammansättningar är en produktiv process i svenskan, därför är de kombinatoriska möjligheterna i teorin obegränsade. Detta gör det omöjligt att skapa lexikon med komplett täckning och därför krävs robusta mekanismer för igenkänning av icke-lexikaliserade sammansättningar. Med ett lexikon som innehåller morfosyntaktisk information kan detta göras genom att kombinera ord och deras kategorier från lexikonet till godkända sammansättningar.

Sammansättningsgrammatiken är implementerad som en deterministisk finit automat (DFA). Sammansättningsgrammatiken uttrycker kategoriberodenden med hjälp av reguljära uttryck. Se avsnitt 4.1 för en beskrivning av grammatikkomponenterna, och avsnitt 4.2 för språkspecifika detaljer.



Figur 1.2: Finit automat för sammansättningsexempel.

1.1.3 Uttalsbaserad korrektion

Som mål för den uttalsbaserade korrektionsuppgiften har stått att ge bättre ersättningsförslag tillsammans med ortografisk editeringsavståndsberäkning. Dessa två komponenter ska användas tillsammans i en kombinerad ljud- och stavningsbaserad ordnivåkorrektion för svenska. För att uppnå detta skapades en överföringsgrammatik och ett fonetiskt lexikon, som skall användas för att jämföra uttalslikhet. Först ställdes en specifikation av fonemisk representation upp som grund för överföringsgrammatiken, som i sin tur användes för att generera det fonetiska lexikonet för svenska. Överföringsgrammatiken används dessutom under körning av Scarrie-piloten för att generera fonetiska representationer för de okända ord som upptäckts vid körningstillfället, vilka jämförs med representationerna i lexikonet för att hitta det mest troliga ersättningsförslaget. Resultatet av jämförelsen viktas sedan ihop med editeringsavståndsmåttet för ett kombinerat mått.

Grafem-till-fonem-överföringen sker i tre steg. I dessa är det olika kontextberoende regler som opererar på indatasträngen, med början i den ursprungliga ordformen:

1. ytliga grafem-till-fonem-överföringsregler
2. metaregler
3. efterbearbetningsregler

Från den ursprungliga ordformssträngen ger det första stegets ytliga regler en sekvens av fonem. Detta sker deterministiskt. Metareglerna opererar på fonemsekvensen till dess att det inte kan göras några fler regelappliceringar. Efterbearbetningsreglerna opererar strikt vänster-till-höger. Se kapitel 5 för exempel.

1.1.4 Överväganden och förberedelser inför adaptationen

Eftersom adaptationen skulle genomföras för tre olika språkversioner av Scarrie-piloten, som dessutom skiljer sig något i vad de ska klara av, behövdes en hel del planering av hur adaptationsprocessen med de språkspecifika egenskaperna skulle genomföras. Språkliga egenskaper och utvärdering av språkteknologiprojekter har varit i fokus för Eagles (European Language Engineering Standards) standardiseringsprojekt. Ett par exempel är riktlinjer för utveckling av språkteknologiprogram [Véronis & Ide 1996] och utvärdering av språkteknologiprogram [EAGLES 1996].

1.2 Uppsatsens upplägg

Detta inledande kapitel presenterar uppgift och syfte, det vill säga utvecklande av språkspecifika språkliga resurser och validering av den språkliga funktionaliteten, samt anpassning av den svenska versionen av korrekturläsningspiloten. Ett bakgrundskapitel följer. Därefter viks ett kapitel var åt de tre lexikala resurserna: lexikonet, sammansättningsanalysen, och den uttalsbaserade stavningskorrektionen. I kapitel 6 presenteras funktionsvalideringen av de lexikala resurserna samt systemevalueringen. Slutligen, i kapitel 7, presenteras slutsatser av komponentutvecklingen, anpassningsprocessen, evalueringen och komponentvalideringen.

Kapitel 2

Bakgrund

Dagens språkgranskningsprogram utför oftast en kombination av ord-, stil- och grammatik kontroll. När *textkontroll* från början dök upp i samband med datoriseringen, var det som *stavningskontroll*. Då användes helt enkelt listor där alla godkända ord var listade rakt upp och ned. Fanns ordet i listan, så var det OK. Annars betraktades det som felaktigt. När man senare började med *stilkontroll*, så var det icke-önskvärda ord som listades. Fanns ordet inte i listan, så var det OK och uppfyllde den stilnivå som avsågs. De här två metoderna är exempel på hur positiv respektive negativ textkontroll utförs. För att upptäcka felstavningar som gett upphov till ett annat riktigt ord (real-word errors), krävs grammatikregler eller samförekomststatistik (co-occurrence statistics) på meningsnivå. Under 1980-talet började *grammatikkontroll* därför användas. Dessa tre metoder tillsammans kallas gemensamt för språkgranskning. I början på 1990-talet förfinades teknikerna och gjordes mer applicerbara på olika språkgranskningsuppgifter [Kukich 1992]. Generellt befinner sig ett språkgranskningsprogram någonstans i rummet som spänns upp av följande enhetsvektorer, några kontinuerliga andra diskreta:

- ordformslexikon / bokstavssamförekomster (positivt/negativt)
- feligenkänning (ger inga förslag) / korrektion (ger förslag)
- grammatikregler / ord- eller ordklassamförekomster (positivt/negativt)

Hur man bygger upp lexikonet beror på vilken tillämpning det ska användas till. Att hantera regelbundenheter är samförekomststatistikens styrka. Samtidigt är det även dess svaghet. Har man skrivit ett felaktigt ord som följer något mönster kommer det inte att upptäckas, ett exempel skulle kunna vara felaktig preteritumböjning, *ge*, **gedde*; *ha*, **hadde*.

2.1 Korrektion — generering av ersättningsförslag

För att kunna utföra korrektion krävs att man ställer upp hypoteser om vad som har blivit fel. Fel brukar ofta delas upp i två kategorier; performansfel (felskrivningar) eller kompetensfel (felstavningar). När det gäller performansfel har man funnit att fyra enkla operationer kan appliceras för att korrigera nästan alla förekomster av dessa fel [Damerau 1964]:

- Insättning (insertion) — en bokstav extra har satts in, **insertion*
- Borttappning (deletion) — en bokstav har tappats, **deltion*
- Omkastning (transposition) — två bokstäver är omkastade, **tranpsosition*
- Ersättning (substitution) — en bokstav har ersatts med en annan, **substition*

Många fall kan klassas till flera olika kategorier. Kontextberoendet gör behovet av en tvåvägskoppling till ett grammatikkontrollprogram tydligt. För att korrigera kompetensfel krävs ofta även uttalsbaserad information.

Kompetensfel är ett fel som beror på en brist i skribentens stavnings- eller ordbildningskunskap. Ofta gör detta att det felstavade ordet stavas uttalslikt¹. Dock kan felet i de flesta fall av vokallängd som påverkar konsonantdubbling, betoning eller tonalitet, korrigeras genom att applicera felskrivningsoperationerna; insättning (**lamma*), borttappning (**pankaka*) eller ersättning (**värklig*). Men om felet kräver mer än en felskrivningsoperation för att blir korrekt är uttalsbaserad korrektion fördelaktigare.

- Felaktig ordkunskap — **värdens grej*, **missommar*
- Övergenerering från regler — **gedde*
- Kongruensfel — **en hus*
- Orddelning — **jätte stor*²
- Registerfel — **en jävligt fin middag*

För att hitta användbara uttalsbaserade korrektioner får den fonematiska transkriptionen varken vara för grov eller för fin. Grammatiken måste kunna klara fonemets alla realiseringar i tabell 5.1 sidan 24. Likaljudande sekvenser av fonem ska dessutom överföras till samma representation. Exempel på använda regler finns i avsnitt 5.1.1.

Feltypologier kan ställas upp där fel från texter klassificeras, till exempel [Wedbjer Rambell 1998a]. Då får man underlag för att ranka ersättningsförslag i den ordning man finner mest trolig.

Användning av felregler eller felsamförekomster är vanligare än användning av positiva regler eftersom det är mycket svårare att göra en komplett språkbeskrivning än en beskrivning över ett fragment som inte anses tillhöra språket.

2.2 Standardisering

Fastän mycket arbete har lagts ner på engelska språkgranskningsprodukter har utvecklingen av icke-engelska produkter varit eftersatt på grund av avsaknaden av standardiserade språkliga resurser. Dessutom saknas öppna standarder för interaktion med program på operativsystemsnivå, det vill säga en

¹Böjnings- och avledningsfel måste identifieras på meningsnivå.

²Korrekt om frasgräns.

infrastruktur för språkgranskning i form av ett öppet programmeringsgränssnitt, application program interface (API), där de språkliga resurserna kan kopplas in på lika villkor. Dessa två begränsningar har lett till en tveksamhet bland programföretag att satsa på nya produkter då de riskerar att låsa sig till en viss leverantörs språkliga resurs. Av ekonomiska skäl har de små marknaderna för mindre språk också gjort utvecklingen lågprioriterad för kommersiella företag. Tidigare forskningsprojekt fokuserade ofta på engelska fastän de genomfördes i icke-engelsktalande länder. En orsak till det är att engelska har en någorlunda enkel morfologi och ordföljd som ger snabb återkoppling för små mängder indata (morfologiska regler eller grammatikor).

Lokalisering, internationalisering och adaption

Globaliseringsriktlinjer håller på att utvecklas, vilket säkerställer att internationaliserings- och lokaliseringsansträngningar följer standarder och använder metoder som resulterar i högkvalitativa program som fungerar för språk världen över/jorden runt. Följande stycken ger kort definitionerna för globaliseringskomponenterna:

Lokalisering (L10N)³. Med lokalisering menas “Översättning och anpassning till den lokala marknaden följt av kompilering och provkörning.” [Esselink 1998]. Lokalisering har varit en mycket liten del av arbetet. Felmeddelanden och övriga textsträngar har översatts och programmet har anpassats för detta.

Internationalisering (i18N)⁴. Internationalisering specificerar utvecklings- och designfrågorna gällande utveckling av flerspråksprodukter. Det vill säga på den lägsta nivån; att undvika explicita textsträngar i koden. Textsträngar ska placeras i externa resursfiler för enkelt underhåll och ökad flyttbarhet. Språk-specifika kodavsnitt ska ej blandas med generell kod. För att undvika programfel och oväntade resultat bör internationaliseringssäkrade textsträngsfunktioner användas för till exempel teckenlikhet och -ekvivalens, lexikografisk ordning — sortering, teckenklasser och konverteringar. Eftersom ordföljd och morfotax kan skilja sig från källspråkets bör textsträngskonkatenering undvikas. Internationalisering har varit en något större del av arbetet än lokalisering.

Adaption är tillägg av egenskaper eller ändringar på kodnivå, i ett annars internationaliserat program, för att kunna utföra samma uppgifter som med det ursprungliga programmet. Adaption har utgjort en större del av arbetet när det gäller framtagandet av Scarrie-piloten.

³L10N är en förkortning av ordet localisation — l+ocalisation(=10 bokstäver)+n

⁴i18N är en förkortning i analogi med den för lokalisering — i+nternationalisatio(=18 bokstäver)+n

Kapitel 3

Lexikonet

Detta kapitel beskriver kort framtagandet av det svenska ordformslexikonet och dess lexikala täckning. I anknytning till detta kapitel, presenteras i bihang A strukturen på den lexikala databas, ScarrieLex, som lexikonets data är lagrade i, samt en användarhandledning för handhavandet av ScarrieLex.

Lexikonet innehåller ord och fraser av två typer:

godkända Ord eller fraser i angiven form accepteras antingen i alla definierade stilar eller i explicit angiven stil. Former som endast ska accepteras i sammansättningar kan också anges.

icke-godkända Ord eller fraser som inte ska accepteras alls eller ersättas antingen i alla definierade stilar eller i explicit angiven stil.

Detta möjliggör specifikation av både en positiv och en negativ vokabulär. Den positiva används som grund för den lexikala täckningen och även dess utökning i sammansättningsanalysen för att känna igen icke-lexikaliserade sammansättningar. Den negativa vokabulären höjer i första hand precisionen genom att man antingen kan säga att något är fel eller ger en explicit ersättning för den icke-godkända enheten. Alla ordenheter, tagna från korpusen (se avsnitt 3.2 nedan), i lexikonet har getts information om frekvens, grammatisk kategori, morfosyntaktiska särdrag, och stil. Frasenheter har getts information om grammatisk kategori, och morfosyntaktiska särdrag, samt positiv eller negativ stil. Icke-godkända enheter, ord och fraser som markerats för ersättning, har därutöver information om korrektionsersättningar (se avsnitt 3.3.1 för mer information).

Det svenska Scarrie-lexikonet har sammanställts med korrekturläsning av tidningstext som sitt huvudsakliga användningsområde.

3.1 Källor för det svenska ordformslexikonet

Källorna som den huvudsakliga lexikala täckningen är baserad på, är de följande:

1. Det generella lexikonet till Uppsala Chart Processor (UCP) 1997, [General Dictionary of UCP 1997]

2. 60 395 artiklar från Upsala Nya Tidning (UNT) 1995–1996
3. 159 691 artiklar från Svenska Dagbladet (SvD) 1995–1996

3.2 Urvalsförfarandet

En principiell idé som ligger till grund för den lexikala täckningen är att urvalet ska vara korpusbaserat. Endast de ordformer som förekommer i SvD/UNT-korpusen (källorna 2 och 3 ovan) inkluderades i en första omgång. Dessutom har de lemmor som finns representerade i urvalet expanderats med alla sina former.

SvD/UNT-korpusen är sammanställd från 220 086 artiklar, vilka innehåller cirka 70 miljoner ordförekomster (1 672 993 ordtyper). Frekvensinformation hämtades från korpusen och lagrades i en fil, där varje enhet innehåller information om totalfrekvens, samt frekvens med initialversal respektive initialgemmen. Ytterligare bearbetningar utfördes på dessa frekvensdata [Dahlqvist 1998]. Alla ordtyper som endast förekom en gång i korpusen togs bort, vilket lämnade 618 099 ordtyper i filen. Ordformerna kategoriserades utifrån vilka teckentyper de innehöll. Denna kategoriseringen var utgångspunkten för en grovprocessning för att avgöra om de skulle bearbetas vidare eller ej. Till exempel har alla poster bestående enbart av sifferuttryck uteslutits. Egennamn klassificerades ytterligare i följande kategorier; personnamn, företagsnamn, namn på geografiska platser, akronymer och flerordsenheter. Detta arbete utfördes till stor del av några sommarjobbare.

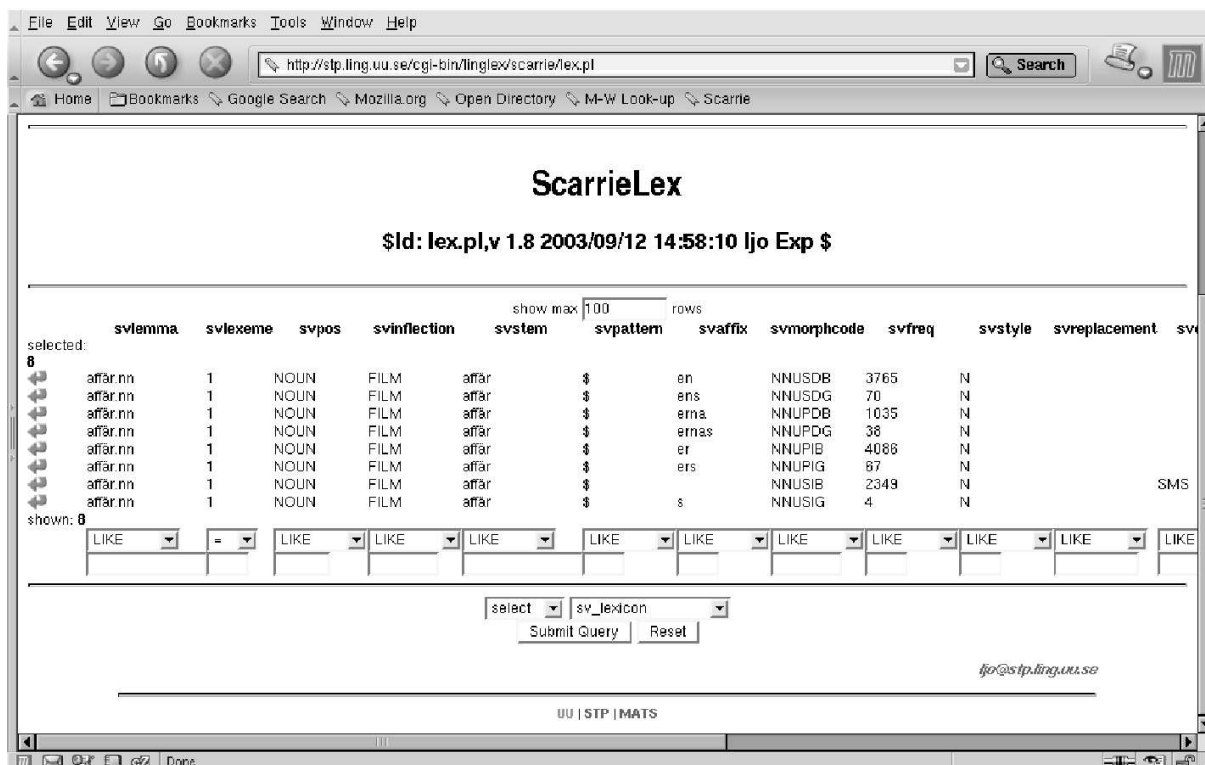
De resterande 350 000 ordtyperna, med frekvens två och högre, inkluderades för morfologisk analys i UCP-systemet, ett chartparsningssystem, med lexikonuppslagning, morfologisk och syntaktisk analys, utvecklat vid Institutionen för lingvistik, Uppsala universitet [UCP]. Bearbetningen omfattade sammansättningsanalys i UCP. En grovgallring av semantiskt felaktiga segmenteringar av lexikaliserade sammansättningar gjordes av ett par sommarjobbare. Resultatet av analysen efter grovgallringen var 252 180 ordformer vilka representerar 88 325 lemmor i det generella lexikonet i UCP. Analysresultaten konverterades till ett transportformat, kallat Master dictionary-formatet, och lagrades i den lexikala databasen ScarrieLex, utvecklad för detta syfte (se avsnitt 3.3). För mer information om urvalet av flerordsenheter, se [Wedbjer Rambell 1998c]. Efter detta har jag manuellt åtgärdat de felaktigheter eller problem som upptäckts.

3.3 ScarrieLex — en lexikal databas

En lexikal databas, ScarrieLex, har utvecklats för att stödja underhåll av lexikala resurser, både för den lexikonsammansställning som beskrivs här, och senare för underhåll av slutanvändarna på plats. Strukturellt och erfarenhetsmässigt bygger ScarrieLex på två tidigare databaser, en tidigare projektspecifik databas för Scarrie-piloten [Tiedemann 1999], samt MatsLex, som är en flerspråkig databas avsedd för maskinöversättning [Tiedemann 2002]. I [Olsson 2003] beskrivs processen för att lägga till ytterligare information i databasen för att göra lexikonet mer generellt och lämpat för andra användningsområden. Ett steg mot generalisering var att morfologiskt blåsa upp lexikonet så att det innehåller alla ytformer av de lemmor som fanns i den projektspecifika databasen, för detta krävs till exempel teknisk stam och i vissa fall mer specifika böjningsmönster. Se [Olsson 2003] för mer information.

Databasen innehåller alla lexikala resurser; godkända såväl som icke-godkända former av både enkla ord och fraser, samt böjningsinformation för morfologisk generering. Databasen kan nås via ett webbgränssnitt som kan presentera data i flera olika vyer (se figur 3.1– 3.5). För att exportera information från ScarrieLex till önskat användningsformat används olika exportskript.

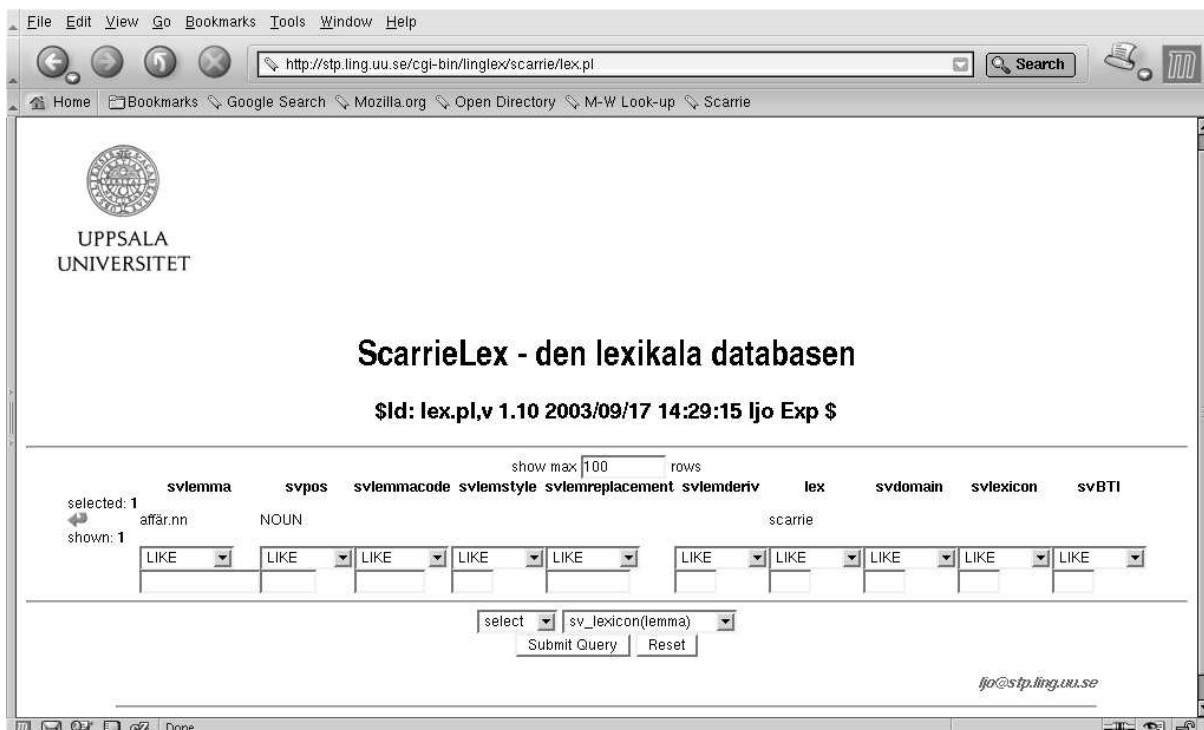
Databasen bygger på frekvensutvalda ordformsdata bestående av 265 552 poster för enkla ordformer. Detta inkluderar 67 792 egennamn. 265 447 av dessa poster är godkända former och 105 icke-godkända. Dessutom baseras databasen på enskilda former för 4 899 godkända fraser och 46 flerordsenheter. De lemman som finns representerade, cirka 150 000, i ovanstående ordformsdata blir morfologiskt expanderade cirka 1,8 miljoner poster.



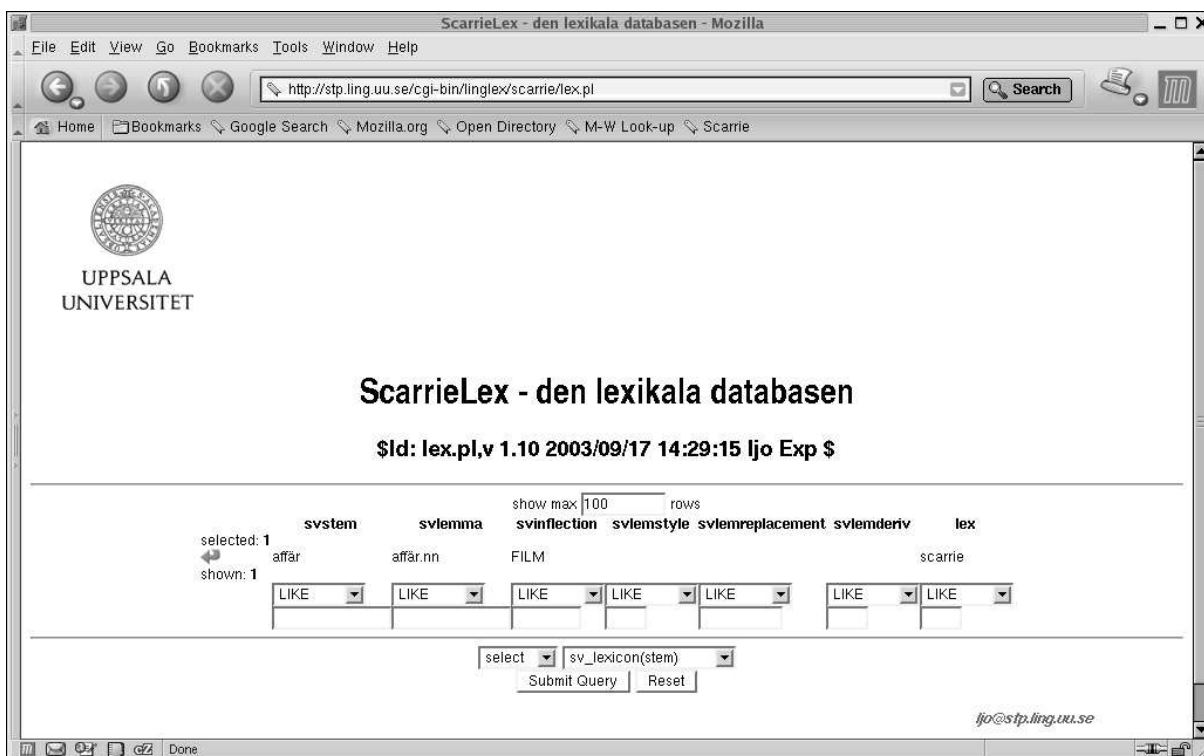
Figur 3.1: Användargränssnittet för arbete med ScarrieLex - vy lexikon.

3.3.1 Struktur och innehåll i ScarrieLex

En typisk ord- eller fraspost i ScarrieLex innehåller information om lemma, stam, böjningsmönster, ordklass och morfosyntaktiska egenskaper. De morfosyntaktiska egenskaperna och några semantiska aspekter är ihopslagna till en morfosyntaktisk kod. De olika fältens information kopplas antingen till lemmat (kan ses i lemmavyn figur 3.2) eller till enskilda böjningsformer (se figur 3.4). I vyn för böjningsinformation ses fälten böjningsmönster, affix, morfosyntaktisk kod och morfomvandlingsmönster (se figur 3.5). Information om sammansättningsegenskaper lagras också i databasen (se kapitel 4 för mer information om detta). En beskrivning av alla, för användaren, tillgängliga fält för poster i ScarrieLex (lexikonvyn, se figur 3.1) ges i bilag A.



Figur 3.2: Användargränssnittet för arbete med ScarrieLex - vy lemma.



Figur 3.3: Användargränssnittet för arbete med ScarrieLex - vy stem.

ScarrieLex - den lexikala databasen

\$Id: lex.pl,v 1.10 2003/09/17 14:29:15 ljo Exp \$

show max 100 rows

selected:	system	svaffix	svpattern	svlemma	svlexeme	svpos	svinflection	svmorphcode	svfreq	svstyle	svreplacement	sv
8	affär		\$	affär.nn	1	NOUN	FILM	NNUSIB	2349	N		SMS
	affär	en	\$	affär.nn	1	NOUN	FILM	NNUSDB	3765	N		
	affär	ens	\$	affär.nn	1	NOUN	FILM	NNUSDG	70	N		
	affär	er	\$	affär.nn	1	NOUN	FILM	NNUPIB	4086	N		
	affär	erna	\$	affär.nn	1	NOUN	FILM	NNUPDB	1035	N		
	affär	ernas	\$	affär.nn	1	NOUN	FILM	NNUPDG	38	N		
	affär	ers	\$	affär.nn	1	NOUN	FILM	NNUPIG	67	N		
	affär	s	\$	affär.nn	1	NOUN	FILM	NNUSIG	4	N		

shown: 8

LIKE LIKE LIKE LIKE = LIKE LIKE LIKE LIKE LIKE LIKE LIKE LIKE

select sv_lexicon(wordform)

Submit Query Reset

ljo@stp.ling.uu.se

Figur 3.4: Användargränssnittet för arbete med ScarrieLex - vy ordform.

UPPSALA UNIVERSITET

ScarrieLex - den lexikala databasen

\$Id: lex.pl,v 1.10 2003/09/17 14:29:15 ljo Exp \$

show max 100 rows

selected:	svinflection	svpattern	svaffix	svmorphcode
8	FILM	\$		NNUSIB
	FILM	\$	en	NNUSDB
	FILM	\$	ens	NNUSDG
	FILM	\$	er	NNUPIB
	FILM	\$	erna	NNUPDB
	FILM	\$	ernas	NNUPDG
	FILM	\$	ers	NNUPIG
	FILM	\$	s	NNUSIG

shown: 8

LIKE LIKE LIKE LIKE

select sv_inflections

Submit Query Reset

Figur 3.5: Användargränssnittet för arbete med ScarrieLex - vy böjningsmönster.

3.3.2 Finjustering

För att optimera lexikonet och grammatiken, med avseende i första hand på snabbhet, men även precision genom att göra viss information explicit för grammatiken, så reducerades i mitten av projektet systematiskt ordformshomonymer av ett par typer till en enda post, där följande morfosyntaktiska koder användes:

- Nomen och adjektiv: NANXIB, NAUSIB, NANSIB, till exempel *lik*,
- Verb och adjektiv: PCPXSDBT, PCPNSIBS, PCPXSIBS, PCPXSDBGP, exempelvis *övervärderade*,
- Verb och verb: VBAIMI, VBAICI, VBAIAI, VBARMI, VBAPMI, VBARCI, VBPPMI, VBDPMI, VBDPMI2, till exempel *övertala*.

För närvarande finns 417 olika morfosyntaktiska koder. Inför arbetet med uppblåsningen av lexikonet återställdes många finjusteringar. Detta för att möjliggöra den morfologiska expansionen och att göra det hela så transparent som möjligt. I stort sett alla de 13 000 poster som reducerades bort enligt ovan i den projektspecifika databasen har återställts. Till exempel har alla reduceringar med homonyma lemmaformer inom samma lexem helt återställts (från *stans2+3.nn* till *stans2.nn* och *stans3.nn*). Alla nomen-adjektiv-ordformshomonymer har återställts, dessutom har många av verb-adjektiv- och verb-verb-ordformshomonymerna återställts.

3.3.3 ScarrieLex och det intermediära lexikonformatet

Informationen i databasen ScarrieLex kan exporteras till det intermediära lexikonformatet (intermediate dictionary format, IDF). Se exempel (1). Det intermediära lexikonformatet är tabbseparatorat. Det är sex fält för varje post, se exempel (1).

(1)	Ordform Uppsala	Frekvens 59568	Stil N	Morfosyntaktisk information PNOUN,SVERIGE,SING,NIL,BASIC	Ersättning	Extrainformation Uppsala.PM PMXB
-----	--------------------	-------------------	-----------	---	------------	-------------------------------------

Informationsfälten är i ordning från vänster: ordform, frekvens, stilregisterinformation, morfosyntaktisk information, ersättningsinformation, och extrainformation. Se bihang, avsnitt A.1 på sidan 43, för ytterligare information och exempel på vad dessa fält kan få för innehåll vid export från databasen (ScarrieLex). Den morfosyntaktiska informationen används i sammansättningsanalysen och därför behöver kategorier och särdragsvärden definieras för användning i sammansättningsgrammatiken (se avsnitt 4.1). Varje ordform kan tilldelas ett eller flera stilregistervärden. Fältet för ersättningsinformation är tomt om ordformen inte ska ersättas med något annat ord eller fras. Efter exporten från databasen kompileras det intermediära lexikonformatet till ett trie-baserat format med fast poststorlek, detta för att ge snabb lexikonuppslagning vid körning av programmet. Detta körningslexikon kallas Scarrie-lexikonet. I det svenska Scarrie-lexikonet (och IDF), ligger lemma och morfosyntaktisk kod i extrainformationsfältet. I den danska och norska prototypen finns här endast morfosyntaktiska särdragsvärden.

3.4 Några ord om framtiden

Underhållet av det lexikala materialet måste vara kontinuerligt. Administratören måste snabbt kunna få överblick över materialet för att fatta beslut om tillägg eller borttagningar ur lexikonet. Sammanställningen av det lexikala material bör göras mer finindlad och behålla mer metadata så att en avdelnings-specifik jämförelse kan göras. Detta gör att till exempel enheter som kommer över en viss brytpunkt i relativ frekvens kan stilmarkeras automatiskt. Användarspecifika lexikon där nya ord och individuellt språkbruk samlas upp bidrar till variation. Men den användarspecifika informationen kan även propageras upp i lexikonhierarkin, där den kan utgöra beslutsunderlag för den språkligt ansvariga administratören i frågor som rör tillägg till centrala eller domänspecifika lexikon samt i språkpolitiskt normativa frågor. Det webbaserade lexikongränssnittet kan underlätta uppdateringar och domän eller användarspecifika tillägg till den lexikala databasen.

Systemevalueringen, se kapitel 6, visar att täckning, precision, och förslagskvalitet, kan förbättras. Särskilt gäller detta precisionen som i systemevalueringen var 41,3 %. Den morfologiska expansionen gör täckningen bättre och den generalisering som genomförts av den lexikla databasen gör att framtida tillämpningar underlättas. Även förslagskvaliteteten behöver förbättras genom att få ner andelen felmarkeringar utan ersättningsförslag.

Kapitel 4

Sammansättningsanalys

Ordbildning med sammansättningar och avledning ger upphov till ett flertal utmaningar för språkgranskning i svenskan och i andra flekterade språk. Bland utmaningarna återfinns: fogemorfemfel, stavfel i sammansättningsled, särskrivningar och ihopskrivningar. Eftersom ordbildning är en produktiv process är det omöjligt att bygga ett lexikon med komplett täckning. Detta ger behov av sammansättningsanalyser med hög täckning och precision, för att klara en större del av språkets produktiva ordbildning. För en utförlig genomgång av svensk ordbildning, se [Thorell 1981], som har varit den primära källan för den nuvarande implementationen av sammansättningsgrammatiken. Thorell identifierar 166 olika typer av sammansättningar och 117 typer av avledningar. Av dessa har jag implementerat 91 sammansättnings typer och 33 avledningstyper i Scarrie-piloten. De övriga har inte implementerats på grund av någon av de följande orsakerna:

- Information om betonade och obetonade stavelser har inte varit tillgänglig inom projektet.
- Ingen hänsyn har tagits till semantisk information.
- Typerna kan inte uttryckas otvetydigt utan att det orsakar övergenerering.

De ovan angivna orsakerna till att vissa typer inte har implementerats har inte utgjort någon större begränsning eftersom typerna har varit relativt frekventa, och därför inkluderats i lexikonet, eller väldigt sällsynta, och därför inte varit speciellt produktiva. Ändock har till någon del information från vissa oimplementerade typer kunnat användas för att markera lexikoningångar (poster) med sammansättningsbarhetsattributvärden.

De följande avsnitten presenterar sammansättningsanalysatorimplementationen i detalj.

4.1 Sammansättningsgrammatiken

Sammansättningsgrammatiken hanterar de möjliga konkateneringsmöjligheter som finns för ordformer i lexikonet utifrån de grammatiska kategorier och egenskaper som angivits. Detta enligt formalism av Vosse implementerad i Corrie-prototypen [Vosse 1994]. Kategorier och särdragsvärden för varje ordform är tillgängliga för sammansättningsanalysatorn genom interaktion med lexikonet. Notera att typer-

na kategori `category` och särdragsvärde `feature`, samt ensilda morfem `morpheme` måste definieras i grammatikdeklarationen. Grammatiken kan innehålla regler av typerna `compound`, `exception`, `suspicious` och `expand`. Exempel på användning av alla de fyra regeltyperna finns senare i kapitlet, avsnitt 4.2.2. Ett första exempel, exempel (2) och dess implementation kan ses i figur 1.2 på sidan 3. Regeln kan uttydas som att ett egennamn som inte har särdragsvärdet `genitiv`, följt av ett bindestreck och ett egennamn utan restriktioner är en giltig sammansättning. Regeln kan till exempel användas för att känna igen dubbelnamn som *Leif-Jöran*. Detta är en ganska vanlig svensk namngivningskonvention.

(2) `compound (PROPER_NOUN & !GENITIVE) HYPHEN PROPER_NOUN`

Sammansättningsgrammatikens reguljära uttryck liknar starkt standardiserade reguljära uttryck. Syntaxen för de reguljära uttrycken beskrivs enligt Backus-Naur-formatet (BNF) i tabell 4.1. Symbolnamn helt i versaler representerar terminala symboler.

```

reguljärt uttryck ::= elementsekvens | elementsekvens |
                    reguljärt uttryck

elementsekvens ::= element | element elementsekvens

element ::= baselement | baselement "*" | baselement "+" |
           baselement "?"

baselement ::= "(" reguljärt uttryck ")" | kategoriuttryck

kategoriuttryck ::= baskategori | baskategori villkor

baskategori ::= MORPHEME | CATEGORY | "[" val "]" | "."

val ::= CATEGORY | CATEGORY val

villkor ::= "&" FEATURE | "&" "!" FEATURE |
           "&" FEATURE villkor | "&" "!" FEATURE villkor

```

Tabell 4.1: Syntax för reguljära uttryck i sammansättningsgrammatiken.

Specialtecken som används i grammatiken förklaras i tabell 4.2.

Specialtecken	Förklaring
*	upprepa föregående element noll eller flera gånger
+	upprepa föregående element en eller flera gånger
?	föregående element noll eller en gång
.	vilken kategori som helst
	ELLER
[]	en av (avgränsar val)
()	grupperar val
&	med särdragsvärde
!	INTE särdragsvärde

Tabell 4.2: Specialtecken i reguljära uttryck i sammansättningsgrammatiken.

4.2 Sammansättningsregler

Sammansättningsregler har skapats där grafotaxen för godkända sammansättningar uttrycks enligt den formalism som beskrivs i avsnitt 4.1. Varken semantisk information eller uppgifter om betoning har tagits hänsyn till. Sammansättningsreglerna har skapats för att vara så transparenta och icke-överlappande som möjligt.

4.2.1 Sammansättningsbarhetsattribut

Attribut för sammansättningsbarhet har införts för att minska antalet analyser. Tabell 4.3 ger exempel på sådana sammansättningsbarhetsattribut.

Attribut	Beskrivning	Exempel
Läggs till av lexikonadministratören/användaren		
CI	sammansättningsbart som förled (260)	ned
CF	sammansättningsbart som huvudled (31 679)	<i>abonnerad</i>
CA	sammansättningsbart i alla led (18)	<i>ögon</i>
CE	icke-sammansättningsbart led (138)	<i>äldres</i>
STEM	ordstam, får ej förkomma ensamt eller som huvudled (151)	<i>gatu, medie(MN)</i>
SMS	kräver foge-s (35 682)	<i>a-kasseavgift</i>
SME	kräver foge-e (111)	<i>familj</i>
SMA	kräver foge-a (5)	<i>viking</i>
SC	sammansättningsled, trots färre än minsta antal bokstäver (886)	<i>ö</i>
NO_HYPH	tar ej bindestreck (2)	<i>anti, pro</i>
NEED_HYPH	kräver bindestreck (1)	<i>icke</i>
ABBR	förkortning (54)	<i>avd</i>
ACCR	akronym (863)	<i>ANC</i>
TIME	tidsuttryck (19)	<i>juni-juli</i>
Läggs till av Scarrie-piloten vid körning		
NUMBER	siffror	1,2, ...
SHORT	för kort för att vara sammansättningsled (om inte SC, se ovan)	3 bokstäver
NO_WORD	detta är inte ett ord	!
FREQUENT	frekvent förekommande	<i>i</i>
Tillagda till sammansättningsgrammatiken som regler		
BBB etcetera	tillåt expansion av dubbla konsonanter i ordledsgräns	<i>glasstrut</i>

Tabell 4.3: Attribut som begränsar sammansättningsbarheten i grammatiken.

Alla dessa attribut gör antalet analyser färre då de kombineras med de grammatiska särdragen för att få en finare indelning än vad de grammatiska särdragen förmår på egen hand. Till exempel avhuggningsförkortningar (ABBR) kan inte förekomma som förled, akronymer (ACCR) kräver bindestreck före eller efter ledet. Attributet TIME kan även det användas för att ange att det ska kunna vara intervall med endash emellan (januari–mars). NUMBER används för att skilja mellan fem och 5. Attributen SHORT och FREQUENT kan användas tillsammans för att acceptera för korta led som sammansättningsled, till exempel *-ko-*; *fjällko*, *kobingo* och *-ö-*; *öinnevånare*, *paradisö*. Om det finns fler än en analys, väljs den med minst antal led och längst förled, vilket ofta är den bästa lösningen [Karlsson 1990]. Denna lösning är även att föredra av effektivitetsskal.

4.2.2 Kommenterade sammansättningsregler

Särdragsvärdesnamnen presenteras ibland i längre tydligare former än de som faktiskt används i grammatiken. Den första exempelregeln i exempel (3) säger att en sammansättning kan vara en *preposition* följt av ett adjektiv eller ett nomen. Det finns några positionsrestriktioner på prepositionen och adjektivet. Ett par sammansättningar som känns igen av regeln är *över/aktiv* och *efter/behandling*.

```
(3) compound (PREPOSITION & NO_HYPHEN
             | PREPOSITION & COMPOUNDABLE_INITIAL)
           (ADJECTIVE & COMPOUNDABLE_FINAL
           | NOUN)
```

Exempel (4) gör att till exempel *fyrriotvå* accepteras som en korrekt sammansättning.

```
(4) compound (NUMERAL & !NUMBER & !COMPOUNDABLE_EMPTY) +
```

I exempel (5) visas ett exempel på en expansionsregel, vilken gör att till exempel *bb* i *snabbakad* expanderas så att *snabb+bakad* kan hittas i lexikonuppslagningen.

```
(5) expand BBB : "b" + "b" to "bb" + "b"
```

En användning av *exception* skulle kunna vara att göra undantag för begränsningar som i (6), där morfemet *an* räknas som ett ord med kategori *NUMERAL* och särdrag *COMPOUNDABLE_FINAL* för sammansättningsgrammatiken, som i regeln nedanför i exemplet, vilken skulle kunna tillåta $5 : an^1$. Men att införa alltför många undantag i sammansättningsgrammatiken skulle göra sammansättningsanalysen ogenomskådbar, då det inte rör sig om sammansättningsled från lexikonet.

```
(6) exception NUMERAL & COMPOUNDABLE_FINAL "an"
     compound NUMERAL & NUMBER COLON NUMERAL & COMPOUNDABLE_FINAL
```

Negativa regler för ihopskrivningar kan formuleras i analogi med de positiva beskrivna ovan. En exempelregel för att dela ihopskrivna ord skulle kunna vara ett substantiv i bestämd form följt av ett substantiv i obestämd form, till exempel *språketgranskning*. En negativ regel som skulle uttrycka detta vore exempel (7).

```
(7) suspicious NOUN & DEFINITE NOUN & INDEFINITE
```

4.2.3 Sammansättningsregelutvecklingsprocessen

En testprocedur utvecklades för att jämföra lexikonet med sammansättningsanalysatorn med avseende på snabbhet och minnesåtgång. En lexikonuppslagning sker i linjär tid, $O(n)^2$, medan en sammansättningsanalys i värsta fall kan ta upp till kvadratisk linjär tid, $O(n^2)$. Den längre tidsåtgången för sammansättningsanalysen, får då vägas mot att om lexikonet är större, så krävs mer minne för lagring av detta och

¹ Detta är endast ett exempel på möjligheterna med undantag. Morfemet i sig är ju ett böjningsmorfem i exemplet.

² Order of $O(g(n)) = \{f : \mathbf{N} \rightarrow \mathbf{R} \mid \text{Det finns ett } n_0 \in \mathbf{N}, \text{ och ett } c \leq 0 \text{ i } \mathbf{R}, \text{ så att för alla } n \geq n_0 \text{ är } f(n) \leq c \cdot g(n)\}.$

programmet kan gå långsammare om minnestillgången är låg. Ett frekvensurval, med lägsta frekvens två, skapades från tidningskorpusen [Dahlqvist 1998]. Frekvensurvalet kördes igenom UCP-systemet. De 186 522 ordformer som varken analyserades som enkla ord eller lexikaliserade sammansättningar av UCP-systemet, benämns härmed testdata.

Initialt användes ett litet lexikon med 183 073 ingångar av sammansättningsanalysatorn. Lexikonet innehöll endast enkla ord, lexikaliserade sammansättningar och några avledningar, se SOB sidan VI [Svensk ordbok 1986].

En första referenskörning gjordes med alla sammansättningsregler aktiverade. 100 788 poster från testdata kändes igen som sammansättningar. Genom manuell inspektion var det möjligt att se att en hel del av de icke-godkända sammansättningarna inte var godkända ord, utan de var felstavningar, felsegementerade ord, uppmärkningskoder etcetera.

För att utröna vilka typer av sammansättningar som är vanligast i det svenska tidningskorpusmaterialet, lades en regel åt gången till till sammansättningsgrammatiken. Detta förfarande gav som resultat att de icke-godkända sammansättningarna markerades. De ord som inte markerats räknades som igenkända av den sammansättningsregeln som just lagts till. Resultatet av varje sammansättningsregel applicerad på testdata presenteras i tabell 4.4.

Typ ³	Förled	Huvud	Exempel	Procent	Summa
1	NOUN+	NOUN	asyl/frågan, folk/hälso/projekt	40,5 %	75 599
4	NOUN+	ADJ	bild/lösa, fack/lärary/förbundet	4,1 %	7 674
3	ADJ+	NOUN	direkt/buss, skum/rask/affär	3,5 %	6 612
7	NOUN+	VERB	bok/läsande, ex-port/order/ingången	2,6 %	4 801
17	PNOUN_HYPH	NOUN	ANC-regeringen, ERM-valuta	1,8 %	3 280
21	PNOUN	NOUN	Fyris/bion, Internet/kafé	1,7 %	3 163
25	NUM_HYPHEN?	NOUN	nionde/klassare, 45-åringen	0,9 %	1 645
2	VERB	NOUN	lyssnar/brev	0,8 %	1 542
8	ADJ+	VERB	egen/tillverkad, sär/redo/visa	0,8 %	1 536
5	ADJ+	ADJ	blek/rosa, tomt/stor/lek	0,7 %	1 246
31	NUM_!NUMBER	NOUN+	andra/fiolen, fem/dagars/resa	0,6 %	1 140
16	PNOUN_HYPH	PNOUN	Alvik-Gullmarsplan, Lars-Magnus	0,4 %	837
26	NUM_HYPHEN?	ADJ	2000-årig, 96-procentig	0,3 %	515

Tabell 4.4: Igenkända sammansättningar, del ett.

75 599 sammansättningar av typ ett i tabell 4.4, inkluderades i lexikonet eftersom de vara så markant dominerande till antalet. Detta för att undvika att en allt för stor andel av texternas sammansättningar behöver kännas igen med sammansättningsanalysatorn istället för att slå upp dem i lexikonet. Detta gav en granskningstidsförbättring på 12,2 %.

Efter detta tillägg till lexikonet gjordes en ny testkörning på testdata. Denna gång hittades 119 210 sammansättningar (42 030 igenkända av analysatorn samt 77 780 uppslagna i lexikonet).

Som sagts tidigare så är de teoretiska sammansättningsmöjligheterna i svenskan obegränsade eftersom det inte finns någon begränsning på antalet sammansättningsled. Siffrorna från körningarna på testdata indikerar dock att av de igenkända sammansättningarna dominerar tvåledssammansättningarna, som utgör 62,8 % av de igenkända sammansättningarna, 37,1 % består av tre led och 0,1 % har fyra eller fler led.

Typ	Förled	Huvud	Exempel	Procent	Summa
13	NOUN_ACCR_HYPH	NOUN	cd-skiva, pr-trick	0,2 %	330
11	PREP	NOUN	anti/mode, över/jäst	0,1 %	275
22	PNOUN	ADJ	Israel/fientliga	0,08 %	167
19	PNOUN_HYPH	VERB	IT-baserade	0,07 %	139
24	NUM_HYPHEN ?	NUM	femtio/nio	0,07 %	135
32	NUM	ADJ+	två/färgad	0,07 %	128
29	NUM_COLON	E	7:e	0,07 %	125
23	PNOUN	VERB	Singapore/baserade	0,07 %	122
6	VERB	ADJ	lyssnar/vänlig	0,06 %	117
18	PNOUN_HYPH	ADJ	VVS-tekniska	0,06 %	112
30		NUM_!CE+	tjugo/fjärde, fem/hundra/tusen	0,06 %	108
10	PREP	ADJ	efter/satta	0,05 %	102
12	PREP	VERB	efter/släpande, över/kokt	0,05 %	95
9	VERB	VERB	skriv/sugna	0,05 %	88
20	PNOUN_ACCR	NOUN	Iforstyrka	0,03 %	63
28	NUM_COLON	NUM_!CI	33:an	0,03 %	53
14	NOUN_ACCR_HYPH	ADJ	VM-aktuella	0,004 %	8
15	NOUN_ACCR_HYPH	VERB	VM-deltagande	0,001 %	3
27	NUM_COLON	NUM_DIG	11:59	0,0 % ⁴	0

Tabell 4.5: Igenkända sammansättningar, del två.

4.2.4 Finjusteringar

Både lexikonet och sammansättningsreglerna har förfinats. Några regler har överlappande täckning. Detta kan ses i de skillnader som blir mellan igenkänningsresultatet av de enskilda reglerna och hela grammatiken. Genom halvautomatiska metoder, sortering på huvudled eller förled med efterföljande manuell inspektion, har problematiska ord/morfkluster identifierats.

Några av de problem som identifierats är:

1. Många sammansättningar avvisades på grund av saknade fogemorfemmarkeringar på ord i lexikonet, till exempel *anlag*, *viking*, *stab*.
2. Nomenstammar som skiljer sig från normalformen, till exempel *A-kasse-*, *diktar-*, *hälso-*.
3. Högfrekventa ord saknades i lexikonet, till exempel *inrikes*, *sökande*, *städning*, *utrikesminister*.
4. Semantiskt felssegmenterade ordformer fanns i lexikonet. Några av dem var *vinsten* (vinst+en eller vin+sten), *stenlten* och *pastor* (pastor+0 eller past+or).
5. Några problematiska ord förekommande som icke-huvudled identifierades: *-arbetar-* (finit verb i presens sammanfallande med nomenstam), *-data-* (bestämt eller obestämt nomen i plural), *-domar-* (nomen i plural) *-fakta-* (bestämt eller obestämt nomen i plural).

Efter identifiering av problemen ökade täckningen med 21 % (från 83 513 igenkända sammansättningar till 100 788) genom följande åtgärder:

- Bindemorfemmarkeringar lades till till cirka 4 000 poster.
- Högfrekventa nomenstammar som saknades lades till.

- Andra högre frekventa ord som saknades i lexikonet lades också till.
- Felsegmenterade ordformer (utifrån semantiska aspekter) togs bort eller korrigerades.

4.3 Sammansättningsutvärdering

Efter finjusteringarna gjordes en utvärdering av effekten. En framlumpad mängd ($n=1\ 001$) av de igenkända sammansättningarna undersöktes manuellt för att utröna sammansättningsanalysatorns precision. 998 (99,7 %) av utvärderingsmängden var korrekta sammansättningar. Grammatikens täckning på en framlumpad mängd från testdata blev 62,8 %.

Informella test, dock enbart för substantivsammansättningar, som har genomförts tillsammans med Ebba Gustavii visar en täckning på cirka 85 % med bibehållen hög precision. Detta indikerar att det i viss mån går att få upp täckningen utan precisionssänkning. Ytterligare undersökningar måste göras för att se hur långt detta kan leda. Frågan är också vilken nivå användarna kan acceptera för falska alarm/missade fel.

För ytterligare utvärderingssiffror, se kapitel 6.

4.4 Morfologisk avledning

Avledning är den näst mest produktiva ordbildningsoperationen i svenskan. De implementerade avledningstyperna är prefix- och suffixmorfer, och icke-sammansättningsbara ord. Detta utökar lexikonets täckning, som i exempel (8), där ett okänt uttralt nomen med suffixet *-ant* kan accepteras om det i lexikonet finns ett verb med samma förled (kan sammafalla med men är ej samma som teknisk stam) som slutar med suffixet *-era*.

(8) VERB + *-era* → NOUN + *-ant*

Den morfologiska suffixavledningen hanteras separat med hjälp av strängmatchningsersättningar av suffixmorfer. Detta är inte sunt implementerat i Corrie-prototypen, eftersom denna lexikala information är insprängd som en statisk datastruktur i programkoden och kräver omkompilering av hela programmet för att en ändring ska kunna ta effekt. Prefixavledning hanteras i sammansättningsgrammatiken med hjälp av lexikonet. Se avsnitt 3.3.1.

Kopiering av lemmat och lemmaspecifik information är tänkt att göras i samband med matchningen av böjningsmönster och ändelser.

Detta har alltså inte implementerats, utan en morfologisk expansion av lexikonet har genomförts istället, se kapitel 3.

4.5 Några ord om framtiden

Systemevalueringen, se kapitel 6, visar att täckning och förslagskvalitet, behöver förbättras. Särskilt gäller detta täckningen som i valideringen på testsviter var 63,5 %. Detta verifieras av de siffror som

sammansättningsutvärderingen i avsnitt 4.3 ger (62,8 %). Fortsatta förfiningar enligt proceduren i både lexikon och regler kan ge bättre täckning. Justering av restriktiviteten kan vara värt att prova för att se hur långt det går att komma utan att precisionen minskar. Tester, dock endast för substantivsammansättningar, med mindre restriktiva regler har gett en täckning på cirka 85 % med bibehållen hög precision. När det gäller förslagskvaliteteten bör den förbättras genom att få ner andelen felmarkerade sammansättningar utan ersättningsförslag.

Kapitel 5

Grafem-till-fonem-överföring

Det här kapitlet beskriver framtagandet av uttalsregler för stavningskorrektur av svenska som de har implementerats av mig i Scarrie-piloten. Reglerna ska användas för att höja förslagskvaliteten. Därför ingår också en utvärdering av effekten av reglerna där jag tittar på hur bra ersättningsförslagen blir med respektive utan reglerna. Överföringsgrammatiken skapar en fonematisk representation som används i jämförelsen av den granskade och felmarkerade ordformen med de redan överförda representationerna av lexikoningångarna. Målet med den uttalsbaserade jämförelsen är inte att göra perfekta fonetiska transkriberingar utan att spegla alla symboler med samma uttal till samma fonem och att alla teckensekvenser som kan motsvara varandra speglas till samma fonemsekvenser. Detta angreppssätt hanterar både kompetens- och performansfel i stavningen.

Några autentiska exempel från de två felkategorierna är:

1. kompetensfel: jöra (*göra*), skälv (*själv*), känst (*tjänst*), gågna (*gångna*), bällen (*bollen*), ursekta (*ursäkta*), kann (*kan*), dräckt (*dräkt*), nyj (*ny*), galmal (*gammal*)
2. performansfel: diskissoin (*diskussion*), förenign (*förening*)

Se avsnitt C.2 för exempel på ersättningsalternativ för orden under kompetensfel ovan. Avståndsmåttet i antalet editeringsoperationer för ordet **diskissoin* under performansfel ovan blir två (ersättning av *i* mot *u* och omkastning av *oi* till *io*) jämfört med **förenign* som bara har avståndet ett (omkastning). Se även [Wedbjer Rambell et al 1998], för de faktiska felen i den sammanställda SvD/UNT-feldatabasen. Utifrån felen i feldatabasen kan man se att användargruppen inte gör speciellt många kompetensfel, utan det är performansfelen som dominerar starkt. Av den anledningen blir effekten av uttalsreglerna inte särskilt stor (se utvärderingssiffror nedan i avsnitt 5.2). Om användargruppen hade gjort en större mängd kompetensfel, skulle effekten blivit större. Dock måste uttalsreglerna anpassas till de förhållanden som gäller, eftersom målet med den uttalsbaserade jämförelsen är att spegla alla producerade symboler och sekvenser med samma uttal till samma fonem eller fonemsekvenser. Denna produktion kan variera mycket. Fördelarna med den uttalsbaserade korrektionen kommer ju när det krävs två eller fler editeringsoperationer för att korrigera ett fel, till exempel för sje-ljudet eller i fallet med assimilation.

Relationerna mellan fonem och grafem kan studeras ur två motsatta perspektiv:

- i) hur grafemet uttalas, *grafo-fonematiska relationer*, eller
- ii) hur ett fonem stavas i skrift *fono-grafematiska relationer*.

I överföringsregelimplementationen har jag baserat relationerna mellan fonemet och dess stavning i svenskan på [Garlén 1988]:157–162. I avsnitt 5.1 presenteras överföringsgrammatiken. Där ingår även något litet om några fonologiska processer som implementerats som metaregler. Hela avsnittet är upp-blandat med exemplifierande uttalsregler som implementerats i överföringsgrammatiken. Men först något om fonografematiska relationer.

Fonemen representeras av fonetiska tecken från IPA 1993 [IPA 1993], och för detta används TIPA-paketet [Rei 1996]. I implementationen används Sampas representation av det fonetiska alfabetet anpassat för svenska [Wells 1989]. En sammanställning av de fonografematiska relationerna i svenskan ges i tabell 5.1.

Fonem	Realiseras som	Exempel
/p/	p, pp	<i>apa, pappa</i>
/t/	t, tt	<i>tåt, åtta</i>
/k/	k, ck	<i>kåk, rycka</i>
/b/	b, bb	<i>bar, tub, stubbe</i>
/d/	d, dd	<i>dåd, ladda</i>
/g/	g, gg	<i>gå, aga, agg</i>
/m/	m, mm	<i>tam, mamma</i>
/n/	n, nn	<i>nå, inne</i>
/ŋ/	ng	<i>äng, sjunga</i> (/ŋn/ representerar <i>gn</i> i <i>ugn</i>)
/f/	f, ff	<i>får, soffå</i>
/s/	c, s, ss, z	<i>cykel, så, oss, zon</i>
/ç/	k, ki, kj, tj, ch	<i>kär, kiosk, kjol, tjog, check</i>
/ʒ/	ge	<i>bagage, garage, beige</i>
/ʃ/	ch, che, g, ge, gi, ige, j, je, sc, sch, sh, shi, si, sj, sk, skj, ssi, ssj, stg, sti, stj, ti	<i>chef, apache, geni, religiös, jour, damejeanne, crescendo, schack, shunt, fashionabel, division, sju, skön, skjorta, mission, ryssja, västgöte, suggestion, stjärna, station, (/kʃ/ betecknas med <i>xi</i> i <i>reflexion</i> och med <i>xj</i> i <i>Växjö</i>)</i>
/h/	h	<i>ha, hund</i>
/v/	v, vv, w, u	<i>väv, vovve, watt, Quist</i>
/j/	j, g, dj, gj, hj, lj, y	<i>jord, genast, djur, gjord, hjord, ljuga, yoga</i>
/l/	l, ll	<i>al, le, alla</i>
/r/	r, rr	<i>rå, orre</i>
/i/	i	<i>bi, sil, sill</i>
/e/	e	<i>se, sett</i>
/ɛ/	ä, e	<i>säd, sätt, berg</i>
/y/	y	<i>fyra, fylla</i>
/ø/	ö, y	<i>lös, löss, fyrtio</i> (marginellt)
/ʊ/	u	<i>ful, full</i>
/u/	o	<i>bo, ost</i>
/o/	o, å	<i>kol, hoppa, kål, åska</i>
/a/	a, e	<i>kal, katt, cendré</i>

Tabell 5.1: Fonemrepresentationer av grafem.

5.1 Grafem-till-fonem-överföringsregler

Grafem-till-fonem-överföringen baseras på en grammatik som sammanställts utifrån den angivna specifikationen av fonemrepresentation i figur 5.1. Överföringsgrammatiken bygger på en kontextberoende formalism av Vosse implementerad i Corrie, där regler opererar i tre steg med utgångspunkt i ordformssträngens grafem. De tre regeltyperna följer samma struktur:

Steg	Regelformat
1	vänsterkontext/målsträng/högerkontext \rightarrow ersättning
2	vänsterkontext/målsträng/högerkontext \Rightarrow ersättning
3	vänsterkontext/målsträng/högerkontext $\sim\rightarrow$ ersättning

De fyra regelkomponenterna, vänsterkontext, målsträng, högerkontext och ersättning kan alla vara tomma. Snedstrecken och pilarna är dock obligatoriska. Observera att pilens typ avgör i vilket steg som regeln kommer att appliceras. Regler i steg ett och två är deterministiska, och steg-tre-regler appliceras strikt vänster-till-höger. Den deterministiska regelappliceringen i steg ett och två skiljer sig åt på punkt. För steg-ett-regler gäller att om en matchande regel appliceras flyttas matchningsfönstret fram till efter matchningen, men för steg-två-regler behålls fönstrets punktposition så att alla matchande regler kan appliceras.

Testord använda under utvecklingen av överföringsreglerna och exempelkorrektioner finns i bihang C. I avsnitten 5.1.2–5.1.4 beskrivs ett antal olika överföringsregler i de olika stegen.

5.1.1 Grafonematiska relationer i svenskan

Det enda prosodiska särdrag som är systematiskt representerat i svenskans stavning är längd. Kort vokalkvantitet markeras oftast genom dubblering av konsonanten som följer en betonad kort vokal. Annars är enkel konsonant efter kort betonad vokal norm. För undantag från dessa huvudfall, se [Garlén 1988]:162–164.

En någorlunda generell regel för uttal är att ett vokaltecken i en betonad stavelse uttalas med lång kvantitet om det följs av högst en konsonant i morfemet.

Regler för att återspegla de grafematiska representationerna av fonemen ges i exemplen (9)–(35) i nästa avsnitt (5.1.2). Som tidigare sagts i kapitelinledningen, så är målet inte att göra perfekta fonetiska transkriptioner utan att återspegla alla sammanfall i uttal och fonemsekvenser för att få så bra ersättningsförslag som möjligt. Därför gäller det att lägga sig på en lagom nivå i detaljrikedom, för att inkludera så få alternativ som möjligt i sökrymden, men att de ändå är de maximalt relevanta ersättningsförslagen.

5.1.2 Steg-ett-regler

Steg-ett-regler operera på teckensträngar och är tänkta att på enklaste sätt överföra grafemen till fonem. Någon djupare analys av kombinationer av fonem och deras egenskaper sker inte i detta steg. För vokalgrafemet *a* kan kvantiteten längd och kvaliteten fram–bak urskiljas exempel (9). Observera att längdangivelser som genereras med steg-ett-regler kan ändras i senare steg.

- (9) /a/ & K & K -> a
/a/r/ -> A:

I exempel (9) säger den första regeln att ett *a* följt av två konsonanter ger ett kort främre /a/, medan ett *a* följt av r och ordslut ger ett långt bakre /a/. När det gäller konsonantgrafemen så råder en större variation.

- (10) /c/ & H -> s

Regeln ovan säger att ett *c* framför främre hög vokal (*e, i, y*) blir /s/ som i till exempel *cykel*. Dock finns mer specifika regler, som exempel (11):

- (11) /cc/ & H -> ks

Där *cc* följt av främre hög vokal blir /ks/, som i *succé*.

Grafemet *ch* har ett antal olika fonemiska realisationer, /tʃ/, /ç/ och /k/. I nästa exempel (12) står *x* för /tʃ/ som i *charm*.

- (12) /ch/ -> x

I exempel (13) är det fonemet /ç/, representerat av *C*, som *ch* realiserar, vanligtvis i engelska låneord som *charter*.

- (13) /ch/(arterleddarlipslachalilen) -> C

I ordet *och* representerar grafemet *ch* fonemet /k/.

- (14) /ch/ -> k

För övrigt överförs *c* till /k/

- (15) /c/ -> k

För ordet *djonk* blir grafemet *dj* överfört till morfemet /dj/. För övriga morfeminitiala användningar blir *dj* /j/, till exempel i *djur*.

- (16) ^/dj/ -> j

Grafemet *e* representerar vanligtvis fonemet /e/ vid lång realisation och fonemet /ɛ/ vid kort realisation eller i till exempel prefixet *er-*, se exempel (17). I några låneord från franskan, till exempel *engagemang* blir det överfört till /a/.

- (17) ^/e/r -> E

Morfeminitialt eller initialt i betonad stavelse framför vokalerna *e, i, y, ä, ö* representerar grafemet *g* fonemet /j/ med undantag av några lånord.

- (18) ^/g/[eiyäö] -> j

Grafemet *g* blir överfört till /ŋ/ mellan vokal och ett *n* i samma morfem, till exempel *lugn*.

(19) /g/n -> N

I *champagne* blir *gne* /nj/. Grafemet *ge* blir ordfinalt /z/, som i *garage*¹.

(20) /ge/\$ -> Z

Framför annat vokalgrafem blir *ge* och *gi* även de /fj/, som i *religiös* och *sergeant*.

(21) /g[ie]/&V -> x

Överföringen för *h* blir /h/ om det inte är *hj* som i *hjort* och *hjärta*, då blir överföringen till fonemet /j/. När det gäller grafemet *i* är det längd och öppenhet som är alternativen i valet av överföring. Om grafemet till exempel följs av två konsonanter väljs den korta men något öppnare realisationen:

(22) /i/&K&K -> I

I ordfinal position väljs lång slutna realisation:

(23) /i/\$ -> i:

I exempel (24) visas ett par fall där grafemet *j* blir överfört till /fj/, till exempel för *jour* och *projekt*.

(24) /jou/r -> xu:
o/j/ek -> x

Utöver detta förekommer *j* i grafemsekvenser som representerar /fj/, /ç/ och /j/. Se nedan exempel (25), (30) och (31). I övriga fall blir överföringen till /j/.

Morfeminitialt framför *e*, *i*, *y*, *ä*, *ö* blir *k* /ç/ i till exempel *kela*, *kila*, *kyla*, *kära*, *köra*. Här finns dock en del undantag, bland annat lånord, slang- och barnspråksord som *kefir*, *kex*, *kille*, *kis*, *kissa*, *kisse*.

Även sekvensen *kj* blir morfeminitialt /ç/ i ett fåtal ord: *kjol*, *kjortel*, *kjusa*.

(25) ^/k/[eiyäö] -> C
^/kj/ -> C

Grafemet *k* förekommer också i kombinationerna *sk* och *skj* vilka representerar /fj/. Se nedan. I alla andra fall blir grafemet överfört till /k/.

Där *n* och *g* hör till samma morfem blir *ng* överfört till /ŋ/ som i *slänga*:

(26) /ng/ -> N

För grafemet *o* i bland andra följande betonade slut med främmande ursprung *-fon* och *-for*, till exempel *telefon* och *metafor* väljs fonemet /o/:

(27) f/o/[nr]\$ -> O:

Däremot blir *o* överfört till /u/ i bland annat följande betonade slut av främmande ursprung: *-onisk*, till exempel *harmonisk*, *platonisk* och *-(t)ion*, till exempel *union*, *passion*, *nation*, *station*:

¹Här vill Garlén ha /fj/ vilket skulle bli x istället för Z som ersättning i regeln i exempel (20).

- (28) i/o/n -> u:
/o/nisk -> u:

Skillanden i längd kan ses i exempel (29), där till exempel *o* i *zon* och *schizofren* får lång respektive kort realisation.

- (29) z/o/n -> u:
z/o/f -> U

Grafemet *s* blir för sig självt /s/. Detta gäller även sekvensen *sc* i *scen* och *obscen*. För övrigt blir *sc* tillsammans med grafemsekvenserna *sch*, *sh*, *si*, *sj*, *sk*, *skj*, *ssi*, och *stj* överförda till /fj/.

- (30) /stj/ -> x

Variation finns för grafemet *ti*, där det i några ord som slutar på *-tion*, till exempel *auktion*, *lektion*, *station* blir /fj/, men /tj/ i andra, till exempel *motion*, *nation*, *portion*. *tj* morfeminitialt i till exempel *tjata*, *tjog*, *tjuta*, *tjära* blir /ç/:

- (31) ^tj/ -> C

I övrigt överförs *t* till /t/.

För det svenska u-ljudet /ʉ/ i till exempel *fura*, *sula*, *hus*, eller det öppnare kortare u-ljudet i *full*, *snurra* ses i exempel (32) hur de blir överförda till ett långt /ʉ/ respektive ett kort /θ/:

- (32) /u/&K&K -> 8
/u/&K&V -> }

Grafemet *x* kan resultera i /ks/, /kfj/ eller mer sällan /s/. Vanligtvis blir det överfört till /ks/ som i *ax*, *yx*:

- (33) /x/ -> ks

För grafemet *y* blir gäller att i *fyrtyo* och sammansättningar och avledningar av detta ord blir överföringen till /ø/. I ett fåtal lånord blir det /j/, till exempel i *yoga* och *yoghurt*. I övrigt blir grafemet överfört till /y/.

I lånord av tyskt eller italienskt ursprung, *mezzosopran*, *Schweiz*, *pizza*, blir grafemet *z* överfört till /ts/. I alla andra fall till /s/.

För grafemen *å*, *ä* och *ö* gäller att de blir överförda till /o/, /ε/ respektive ø med skillnader i längd till exempel *håll*, *håll*, *föl*, *föll*, där de korta realiseringarna ges i exempel (34):

- (34) /ä/&K&K -> E
/ö/&K&K -> 2

Framför *r* blir överföringen till följande allofoner /æ/ respektive /œ/:

- (35) /ä/r -> {
/ö/r -> 9

5.1.3 Steg-två-regler

I detta avsnitt visas steg-två-regler. Steg-två-regler är så kallade metaregler eftersom de opererar på grovt överförda fonemsekvenser och inte på grafemsekvenser som steg-ett-reglerna beskrivna i avsnitt 5.1.2 gör. Steg-tre-regler är i och för sig också metaregler, men kallas här efterbearbetningsregler.

Några fonologiska processer som arbetar på konsonanter i svenskan:

1. Regeln om tonlöshet
/d/ blir /t/ eller /g/ /k/ i till exempel *svid* respektive *vägt*
2. Assimilation av /n/
/n/ kan bli /m/ och /ŋ/ i *min bil* respektive *min katt*.
3. Supradentalisering
/r/ överför sin artikulationsplats till omedelbart efterföljande dental och de två blir därför sammansmälta. Till exempel kan detta gälla i kombinationerna *rt*, *rd*, *rn*, *rs* och *rl*, där de isåfall blir /t/, /d/, /ŋ/, /s/, respektive /l/.

Dessa processer påverkar den överförda fonemsekvensen på olika sätt. Regeln om tonlöshet och assimilation ger möjligen ett utbyte av ett fonem mot ett annat, medan supradentaliseringen ger ett fonem istället för två och gör därigenom fonemsekvensen kortare, till exempel det supradentala /d/ istället för /rd/. Det finns förstås undantag från supradentaliseringen, där båda fonemljuden frambringas utan sammansmältning. Detta markeras brukligt med ett bindestreck (/r-d/) mellan fonemen i sampanotationen.

En regel för att göra fonemsträngen kortare är:

(36) /Odsk/ => Osk

Där blir till exempel d:et i *brådska* utsläckt.

5.1.4 Steg-tre-regler

Steg-tre-regler kan användas för att mjuka upp skillnader, till exempel mellan kort och lång konsonantkvantitet (37) eller mjuka upp i andra sammanhang där det blir för stora skillnader i återspegligen (38). För konsonantfonemen sker en uppmjukning av skillnad i längd genom att två förekomster efter varandra reduceras till en, det vill säga den andra förekomsten tas bort istället för att markeras kvantitativt:

(37) b/b/ ~>
...

Detta kan förstås också göras för vokalfonemen. En annan möjlig uppmjukning skulle kunna vara mellan /e/ och /ɛ/ i de fall det behövs (38):

(38) /E/ ~> @
/e/ ~> @

I fallet med nuvarande användargrupp ger detta dock ingen förbättring av förslagskvaliteten.

5.2 Utvärdering av uttalsreglernas effekt på förslagskvaliteten

En utvärdering av uttalsöverföringsreglernas effekt på förslagskvaliteten visar för en delmängd (n=2 522) av stavningsfelen i SvD/UNT-feldatabasen [Wedbjer Rambell et al 1998] att uttalsreglerna ger en svag positiv påverkan på täckningen och en markant ökning av precisionen, se tabell 5.2. Uppdelat på de två tidningarna visar det sig att denna fördel med avseende på täckningen till större del gäller UNT-materialet där förbättringen blev knappt 1,3 % jämfört med drygt 1,1 % för SvD-materialet. Med avseende på precisionen var förhållandet det omvända, en något svagare precisionsökning, drygt 10 % för UNT-materialet och för SvD-materialet gavs en något större precisionsökning på cirka 20 %. Andelen ord² utan ersättningsförslag var 34,4 % med uttalsreglerna och 19,7 % utan uttalsreglerna. Detta är en direkt effekt av kvalitetshöjningen, eftersom en större andel av de ersättningsförslag som gavs utan uttalsreglerna, där inget förslag var rätt, uteslöts vid användning av uttalsreglerna.

	Med uttalsregler	Utan uttalsregler
Antal fel som fått ersättningsförslag	1 654	2 026
Antal korrekta ersättningsförslag	1 120	1 089
Antal fel utan ersättningsförslag	868	496
Precision	67,7 %	53,8 %
Täckning	44,4 %	43,2 %

Tabell 5.2: Effekt av uttalsreglerna på förslagskvaliteten.

5.3 Några ord om framtiden

Överföringsgrammatiken ger grova fonetiska representationer och förslag för viktad korrektion av performans- och kompetensfel. Den tänkta slutanvändargruppen gör inte speciellt många kompetensfel, men i större utsträckning förekommer performansfel [Wedbjer Rambell et al 1998]. En utvärdering av uttalsreglernas påverkan på förslagskvaliteten visar på en dryg 1,2-procentig förbättring av täckningen och en nästan 14-procentig ökning av precisionen (n=2 522) för denna användargrupp. Se också utvärderingen i kapitel 6. Om slutanvändargruppen växlar måste uttalsregelgrammatiken anpassas därefter.

²Observera att i detta test har inte någon sammansättningsanalys utförts.

Kapitel 6

Systemevaluering och komponentvalidering

I detta kapitel presenteras evalueringsproceduren och de relevanta evalueringsresultaten för de komponenter som ingår i uppsatsen. I [Sågvall Hein et al 1999] och [Paggio 1999] står mer om evalueringsresultaten för den svenska Scarrie-piloten. Evalueringen har genomförts i två steg för att försäkra att den språkliga funktionaliteten följer funktions-specifikationen, först på komponentnivå och därefter övergripande för systemet som helhet. Komponentvalideringen har varit iterativt integrerad i utvecklingsprocessen, medan systemevalueringen har utförts med hjälp av testsviter och utvärderingstexter för att försäkra att pilotens funktionalitet följer den uppställda funktions-specifikationen.

6.1 Evalueringmetoder

En övergripande evalueringsuppställning presenteras i Eagles ramverk för utvärdering av språkteknologiprodukter [EAGLES 1996]. Detta har varit utgångspunkten för utvärderingen inom projektet. En uppdelad evaluering har utförts för stavnings- respektive grammatikkorrektion¹ med avseende på funktionsattribut, som följer TEMAA [Paggio & Music 1998], där riktlinjer för användarkrav och systemegenskaper modelleras. Funktionalitetstestning har varit i fokus, för att utröna om piloten gör det den är tänkt att göra. Andra kvalitetskriterier som användbarhet och underhållbarhet har inte direkt utvärderats i evalueringen, även om dessa kriterier har tagits hänsyn till och användarsynpunkter samlats in.

De övergripande utvärderade attributen är: täckning (recall), precision, förslagskvalitet (suggestion adequacy). Täckning och precision beräknas för lexikal täckning, grammatikkontrolltäckning² och feltäckning. Förslagskvaliteten är kopplad till lexikal feltäckning. Förslagskvaliteten är i sin tur uppbyggd av andra attribut, *ersättningsförslag* och *diagnos*. För *ersättningsförslag* finns underattributen *första förslag*, *i förslagslista*, *inget korrekt förslag*, *inget förslag givet*.

¹Den svenska piloten gör ingen grammatikkorrektion, utan identifierar endast fel och rapporterar dess feltyp.

²I systemet ingår en grammatikkontroll, men eftersom den inte behandlas i uppsatsen, så refereras heller inte grammatikkontrolltäckningen i detta arbete.

6.2 Validering av funktionsattribut

Valideringen av den svenska Scarrie-piloten har utförts med hjälp av testsviter baserade på det material som sammanställts i SvD/UNT-feldatabasen [Wedbjer Rambell et al 1998], samt utvärderingstexter som slutanvändarna tillhandahållit.

6.2.1 Resultat av körning på testsviterna

Detta avsnitt presenterar de resultat som är relevanta för uppsatsens komponenter från testkörningar på de svenska testsviterna [Ahlbom & Sågvall Hein 1999]. Nedan följer resultaten för lexikal täckning på standardvokabulären och sammansättningar som känts igen av sammansättningsanalysatorn, till skillnad från de sammansättningstyper, till exempel vissa särskrivningar, som hanteras av grammatikkontrollen.

Den lexikala täckningen var 97,0 % för standardvokabulären och 75,4 % för sammansättningar. Scarrie-piloten genererade alltså falska alarm i tre procent av fallen för standardvokabulären (n=101) och för sammansättningar var andelen falska alarm hela 24,6 % (n=69). Direkt överfört till löpande text skulle detta innebära cirka tjugo falska alarm per sida.

En sammanställning av resultatet av testsvitsvalideringen uppdelat på skrivfel, stavfel och fel i sammansättningar ges i tabell 6.1.

Feltyp	Funna fel	Missade fel	Korrekt ersättningsförslag
Skrivfel	94,8 %	5,2 %	69,9 %
Stavfel	100,0 %	0,0 %	66,7 %
Fel i sammansättningar	63,5 %	36,5 %	60,1 %

Tabell 6.1: Sammanställning av resultaten för testsvitsvalideringen.

Resultaten sammanställda i tabell 6.2 gäller feltäckning och förslagskvalitet för de olika representerade feltyperna. De problem som kan identifieras i tabellen är; för skrivfel och stavfel, att andelen utan ersättningsförslag är hög, vilket gäller för omkastningar och fel i lånord; för sammansättningarna är det saknade bindestreck, foge-s och särskrivningar som utskiljer sig. Orden i ett par av testsviterna har manuellt förvanskats för att få med alla feltyper som skall klaras enligt funktions-specifikationen.

De godkända ord som flaggats, det vill säga falska alarm, härstammar till största del från en specifik genre, vilken är musik/populärkultur. I denna genre kommer främmande modeord snabbt in och används omodifierade innan de efter ett tag antingen försvinner eller tar lånordets anpassningsskrud på sig. En årsproduktion i förändring mellan produktionen av korpusmaterialet, som ligger till grund för lexikonet, och testsviterna är en allt för stor tid. Variationen i ordföråd är stor mellan de olika avdelningarna i tidningarna. Större ansträngning måste därför läggas på att genomföra stilanalyser för de olika avdelningarna i tidningarna, så att specifik vokabulär kan markeras och användas i berörda avdelningar. Dessutom måste det vara enkelt att växla mellan olika stilar i granskningen, till exempel utifrån uppmärkningskod. Som det är nu sätts stil eller stilar för varje granskningskörning. Angående de flaggade sammansättningarna, så har förfinade sammansättningsregler och de lexikala förbättringar som skett, se avsnitt 4.2.4, åtgärdat vissa brister, men vissa feltyper har lämnats som de är eftersom de inte kan tas av tillräckligt specifika regler för att inte släppa igenom felstavningar. Här blir återigen behovet av snabb återkoppling med användarspecifika lexikon där även för tillfället frekventa sammansättningar som inte är lättigenkännbara

	Totalt antal fel	Flaggade fel	Korrekta förslag	Felaktiga förslag	Utan förslag
SKRIVFEL					
1.1 borttappning	23	23	13	9	1
1.2 insättning	15	15	13	0	2
1.3 omkastning	11	11	4	2	5
1.4 ersättning	9	7	5	1	1
1.5 konsonantdubbling	8	7	7	0	0
1.6 enkelkonsonant	11	10	9	1	0
Totalt	77	73	51	13	9
STAVFEL					
2.1 l-relaterade fel	3	3	1	2	0
2.2 m-relaterade fel	1	1	1	0	0
2.3 n-relaterade fel	5	5	4	1	0
2.4 r-relaterade fel	1	1	1	0	0
2.5 stumma bokstäver	2	2	2	0	0
2.6 fel i låneord	10	10	5	0	5
2.7 borttappning/upprepning av stavelse	1	1	1	0	0
2.8 bokstavsersättning	1	1	1	0	0
Totalt	24	24	16	3	5
FEL I SAMMANSÄTTNINGAR					
3.1 bindestreck saknas	6	1	0	1	0
3.2 felaktigt bindestreck	7	7	5	0	2
3.3 foge-s saknas	6	4	3	1	0
3.4 felaktigt foge-s	7	4	3	1	0
3.5 särskrivning	8	1	1	0	0
3.6 sammansättningar med egennamn	3	2	1	1	0
3.7 sammansättningar med förkortning	2	2	1	0	1
3.8 sammansättningar med felaktig stam	12	11	5	0	6
3.9 sammansättningar med felaktig avledning	1	1	1	0	0
Totalt	52	33	20	4	9

Tabell 6.2: Täckning för ordfel och förslagskvalitet för testsviterna (från [Sågvall Hein et al 1999]).

granskas och inkluderas med jämna mellanrum. Detta kräver ett aktivt och kontinuerligt underhåll av de språkliga resurserna.

6.2.2 Resultat för körning på testtext

En testtext med 233 meningar bestående av sammanlagt 3 641 löpord (1 479 ordtyper) användes för funktionsvalideringen. Av de 3 641 löporden var 3 588 godkända och 53 icke-godkända. Den lexikala täckningen som uppgår till 98,8 %, vilket är 1,8 % högre än för testsviterna. Feltäckningen är i detta fall 100 %, med en precision som är 55,2 %. Andelen korrekta ersättningsförslag ligger på 84,9 % och andelen utan ersättningsförslag är 11,3 %, vilket är 18 respektive 6,4 % bättre än för testsviterna.

6.3 Systemevaluering

Systemevalueringen genomfördes i samarbete med de två svenska slutanvändarrepresentanterna, Svenska Dagbladet (SvD) och Upsala Nya Tidning (UNT). Den språkliga funktionaliteten och systemeffek-

tiviteten var i fokus för evalueringen. Testmaterialet, som består av texter på cirka 15 000 löpord är slumpmässigt utvalt från den valideringskorpus på 694 000 löpord totalt, som tidningarna tillhandahållit för detta syfte.

Generella resultat av systemevalueringen finns i tabell 6.3, där visas täckning, precision och förslagskvalitet. Kategoriserade resultat över falska alarm, missade fel och felaktiga eller saknade ersättningsalternativ, ges i tabell 6.4.

	UNT	SvD	Summa	Procent
Täckning				
	UNT	SvD	Summa	Procent
Godkända ord	7 797	6 213	14 010	
Godkända ord accepterade	7 651	6 081	13 732	98,0
Godkända ord avvisade (oriktigt markerade)	146	128	274	2,0
Icke-godkända ord (riktigt markerade)	56	144	200	
Hittade fel (riktigt markerade)	56	137	193	96,5
Missade fel	0	7	7	3,5
Precision				
	UNT	SvD	Summa	Procent
Markeringar	202	265	467	
Riktiga markeringar	56	137	193	41,3
Oriktiga markeringar	146	128	274	58,7
Förslagskvalitet				
	UNT	SvD	Summa	Procent
Riktiga förslag	56	137	193	
Första förslag	19	37	56	29,0
I förslagslista	2	0	2	1,0
Missade (ersättningsförslag, inget korrekt)	0	18	18	9,3
Inget ersättningsförslag	35	82	117	60,6

Tabell 6.3: Generell funktionssammansättning för systemevalueringen.

I de generella resultaten utmärker sig för korrekta träffar — inget ersättningsförslag — under förslagskvaliteten. I detta fall 60,6 % jämfört med de avsevärt lägre siffrorna för testsviterna 17,7 % respektive 11,3 % för testtexten.

I den kategoriserade funktionssammansättningen för systemevalueringen kan tre problemområden kommenteras:

1. falska alarm
 - konjungerade fraser
 - sammansättningar
 - egennamn
 - rubriker
 - akronymer, förkortningar, symboler
2. missade fel (endast SvD)
 - versal/gemen
 - skrivfel
 - övriga missar
3. felaktigt eller inget ersättningsförslag (i högre grad SvD)
 - ordbildning

	UNT	SvD	Summa	Procent
Godkända ord avvisade (oriktigt markerade)				
Ordtyp	UNT	SvD	Summa	Procent
Fras (eller del av fras)	30	13	43	15,7
Sammansättning	8	31	39	14,2
Låneord	0	5	5	1,8
Sifferuttryck, datum, valuta, måttenheter	16	1	17	6,2
Egennamn	35	13	48	17,5
Akronymer, förkortningar, och symboler	12	24	36	13,1
Tekniska uttryck	1	0	1	0,4
Annat	44	41	85	31,0
Totalt	146	128	274	100,0
Missade fel				
Feltyp	UNT	SvD	Summa	Procent
Versal/gemen	0	3	3	42,9
Ordbildningsfel (bindestreck, fogemorfem, etcetera)	0	1	1	14,2
Stavfel	0	0	0	0
Skrivfel	0	3	3	42,9
Andra fel	0	0	0	0
Total	0	7	7	100,0
Felaktigt eller inget ersättningsförslag				
Feltyp	UNT	SvD	Summa	Procent
Versal/gemen	0	5	5	9,4
Ordbildningsfel (bindestreck, fogemorfem, etcetera)	0	10	10	18,9
Stavfel	0	0	0	0
Skrivfel	3	23	26	49,0
Andra fel	0	12	12	22,6
Totalt	3	50	53	100,0

Tabell 6.4: Kategoriserad funktionssammansättning för systemevalueringen.

- skrivfel
- övrigt

Frashantering och fraskorrigeringen från Corrie-prototypen måste göras om. Som det är nu är den varken tillräckligt robust eller flexibel. När det gäller sammansättningarna, har fler sammansättningsattribut börjat appliceras, för att ge fler begränsningar i sammansättningsbarhet och i viss mån klara fler typer. Egennamn i piloten bygger dels på en stor mängd frekventa namn i lexikon och en statistisk motor för att avgöra om det är ett egennamn utan att finnas i lexikonet. Egennamn är dessutom i stor utsträckning en färskvara som behöver kontinuerlig tillsyn. Formatkontroll och korrigering av numeriska uttryck var oimplementerad vid tiden för systemevalueringen. Detta behöver implementeras. Under projektets gång har tidningarna genomfört normativa förändringar för akronymers ordbildning. Detta har dock till stor del tagits om hand om i sammansättningsanalysen. I lexikonet har till viss del ersättningsposter lagts in för att peka på rätt form. Att placera ut bindestreck, för icke-lexikaliserade sammansättningar, i ersättningsförslag är dock problematiskt i vissa fall eftersom inte tillräckligt med information finns för att avgöra tvetydigheter. I sådana fall måste fler sammansättningar läggas in i användarlexikonet. Rubriker, ofta i versaler, ställer till problem eftersom den uppmärkningsinformation som finns i produktionsmiljön inte följer med till granskaren. Granskaren behöver all tillgänglig information som grund för beslut i granskningsprocessen. Ersättningsförslag måste ges i större utsträckning. Möjligen skulle en ökad viktning för omkastning ge bättre resultat. Detta måste undersökas på en större mängd aktuella felskrivningar, eftersom det är en avvägningsfråga hur intrimningen av parametrar påverkar helhetsresultatet.

6.4 Några ord om framtiden

Generellt kan sägas att den lexikala täckningen (98,0 %) och täckningen för ordfel (96,5 %) är mycket tillfredsställande om man ser till att identifiera fel. I den jämförelse som görs med Microsoft Word 97 i [Dahlqvist 1999], sammanfattad i [Paggio 1999], är resultaten fördelaktiga. Täckningen är avsevärt mycket bättre, 100 % jämfört med Words 24,6 %, även med avseende på precision är resultatet väldigt fördelaktigt, 55,2 % jämfört med 20,0 %. Det finns däremot några svaga punkter: att upptäcka saknade bindestreck och särskrivningar, samt att ge ersättningsförslag.

Vissa typer av särskrivningar hanteras i grammatikkontrollen, men den systematiska hanteringen i lexikon som var tänkt från början har inte fungerat. De 400 särskrivningsfelposter som lagts in i lexikonet fått tas bort på grund av undermålig prestanda och robusthet.

Större ansträngning måste läggas på att genomföra stilanalyser för de olika avdelningarna i tidningarna, så att specifik vokabulär kan markeras. Den snabba föränderlighet som kan uppstå i det moderna språkbruket kan dessutom påtalas då ett par språkliga reformer hunnit genomföras sedan materialet i korpusen producerades, till exempel förkortningspraxis och böjning och sammansättning med akronymer. När det gäller igenkänning av saknade bindestreck, så förväntas ytterligare förfiningar av reglerna och tilldelning av sammansättningsbarhetsattribut öka ordfelsigenkänningen och förslagskvaliteten markant. För de sammansättningstyper där inte otvetydiga regler kan skapas för att undvika att släppa igenom felaktiga sammansättningar kan dessa sammansättningar läggas in i användarspecifika lexikon. Frasersättnings- och fraskorrektionshanteringen måste arbetas om.

Att ge ersättningsförslag för fler fel är också något som måste arbetas med. I viss mån går det att trimma enskilda parametrar, som till exempel att öka vikten för omkastningar som jag föreslog ovan, men för att det ska fungera generellt behövs en undersökning av aktuella felskrivningar så att en intrimning av hela korrektionsparametersystemet kan genomföras. Värt att notera är dock att förslagskvaliteten med avseende på korrekta ersättningsförslag, det vill säga täckningen, och precisionen i utvärderingen [Paggio 1999] är mycket bättre för den svenska prototypen; förslagstäckning (84,9 %) jämfört med den danska (36,0 %), med en precision på 95,7 % respektive 60,0 %. Sätter man siffrorna i kontrast med vad MS Word 97 presterade i samma test, en förslagstäckning på 53,8 % med en precision på 70,0 % för den svenska, och 78,2 % respektive 85,7 % för den danska, är resultatet väldigt övertygande.

Kapitel 7

Slutsatser

Målet med projektet var att utveckla ett högkvalitativt korrekturläsningshjälpmedel för den skandinaviska publiceringsindustrin. Syftet med uppsatsen har varit att beskriva de tre lexikala resurser jag tagit fram för svenska (inom ramen för projektet) och en utvärdering av dessa.

Lexikon Ett korpusbaserat ordformslexikon skulle tas fram avsett för korrekturläsning av tidningstext. Trots att det lexikala urvalet har varit användarcentrerat gör vokabulärskiftningar och ett ständigt föränderligt språkbruk användarspecifika ordformslexikon nödvändiga. Dessutom behövs centrala användarlexikon för att klara anpassningar till nya normativa riktlinjer som uppstått efter att korpusmaterialet sammanställts. För att stödja språkgranskningsprocessen krävs att administratören snabbt kan överblicka de tillgängliga resurserna, att det är enkelt att fatta beslut om tillägg och borttagningar och att det går snabbt att få in ändringar i produktionsmiljön. Processen från insamlat material till färdiga lexikoningångar som används i produktion måste snabbas upp. När slutanvändarna på allvar börjar använda Scarrie-piloten, kommer lexikongränsnittet till ScarrieLex kunna stödja användartillägg till och uppdateringar av de användarspecifika lexikonerna som används tillsammans med Scarrie master dictionary.

Frasersättningshanteringen och fraskorrekturen från Corrie-prototypen fungerade inte robust nog och problemet löstes inte tillfredsställande under projektiden. Frasersättningshantering och korrektion inom fraser behöver därför arbetas om, om den ska vara användbar.

Sammansättningsanalys En sammansättningsgrammatik skulle utvecklas för att ge täckning för den produktiva ordbildningsprocessen av icke-lexikaliserade sammansättningar i svenskan. Dessutom skulle avledning genom prefix och suffix hanteras. En sammansättningsgrammatik har utvecklats, som utökar den lexikala täckningen som lexikonet tillhandahåller på ett kontrollerat sätt. Prefixavledning fungerar bra i samverkan med lexikonet. Suffixavledning däremot måste implementeras på ett annat sätt än det som tillhandahålls i Corrie-prototypen.

Uttalsbaserad korrektion En uttalsgrammatik skulle utvecklas för att förbättra förslagskvaliteten för korrigerande av kompetensfel i stavningskontrollen genom att kombinera uttalsbaserad jämförelse med editeringsavstånds-baserad. En överföringsgrammatik som överför uttalslika grafem och grafemsekvenser till samma fonemiska representation har utvecklats. Användargruppen gör inte speciellt många kom-

petensfel, så effekten på förslagskvaliteten visar sig främst i en markant precisionsökning genom att felaktiga ersättningsförslag utesluts. En ytterst måttlig ökning av korrekta ersättningsförslag, det vill säga täckningen, ges också.

Evalueringsjämförelse En utvärdering har genomförts för att säkerställa att piloten och dess komponenter uppfyller funktionsspecifikationen. Som evalueringen visar, så är den lexikala täckningen 98 % med en feltäckning på 96,5 %, vilket är mycket bra. Precisionen i systemevalueringen ligger på 41,2 %. I en jämförelse med MS Word 97 ser siffrorna lovande ut, se tabell 7.1. Trots detta så har Scarrie-piloten

Attribut	Svenska Scarrie	Svenska MS Word	Danska Scarrie	Danska MS Word
Antal löpord	3 641	3 641	3 462	3 462
Lexikal täckning	98,8 %	98,6 %	95,3 %	95,3 %
Feltäckning	100,0 %	24,6 %	58,1 %	53,5 %
Precision	55,2 %	20,0 %	13,5 %	12,4 %
Antal fel	53	53	43	43
Förslagstäckning	84,9 %	53,8 %	36,0 %	78,2 %
Förslagsprecision	95,7 %	70,0 %	60,0 %	85,7 %

Tabell 7.1: Evalueringsjämförelse med MS Word 97.

sina svagheter. Svagheter är att ge förslag för saknade bindestreck, hanteringen av särskrivningar och att ersättningsförslag ges i för liten omfattning. Några av svagheterna har börjat åtgärdas. Till exempel precisionen för vissa typer av sammansättningar. Möjligheten att öka täckningen för sammansättningar har också undersökts.

Styrkor är de lexikala resurserna som har utvecklats utgör en stabil grund för en språkgranskare baserad på ordformslexikon. Sammansättningsanalysatorn utökar den lexikala täckningen på ett kontrollerat sätt. Den uttalsbaserade korrektionen fungerar och ger ett tillskott i förslagskvalitet, om än svagt avseende förslagstäckning. I fråga om precision är förbättringen stor, cirka 14 % bättre ersättningsförslag. Förutom sammansättningsanalysatorns täckningsökning, har morfologisk expansion genomförts på det ursprungliga frekvensbaserade lexikala materialet. Detta för att höja täckningen ännu mer och att göra grunden att bygga viddare på starkare, då attributet lexikonstorlek har betydelse för andelen falska alarm, vilken av naturliga skäl bör vara så låg som möjligt.

Framtiden En ökad flexibilitet i granskningen där dokumenten inte behöver tas ur sitt sammanhang med vidhängande metainformation eller där uppmärkningsinformation inte går förlorad måste införas. Detta skulle kunna uppnås om de strukturerade format som används av de olika tidningarna tilläts ingå i granskningsprocessen och därigenom göra värdefulla metadata tillgängliga. Att göra separata dataanalyser för de olika avdelningarna i tidningarna och bygga upp lexikon med specifik vokabulär som kan användas tillsammans med det allmänspråkliga vore ett steg framåt. Även möjligheten att från metadata anpassa granskningen med användning av individuella inställningar för till exempel feltyper och korrektionsparametrar skulle öka kvaliteten på granskningen.

Arbetet med projektet visar på att behovet av språkteknologi-API:n är stort. Inte minst av ekonomiska skäl. Att utveckla en språkteknologiproduct som enbart passar ett specifikt program på marknaden är inte något som fungerar för ett mindre företag. Den ökade internationaliseringen ställer krav på flerspråkiga

produkter och textproduktion. Tillgängligheten till infrastruktur för språkteknologiska behov i moderna datormiljöer som Windows, Mac och Unix, är mycket låg. Det finns knappt stöd för en så central lingvistisk analysenhet som ord, utan de betraktas mest som sekvenser av bytes som varje språkteknologisk modul får hantera fristående från andra relaterade moduler. Detta behöver åtgärdas för att underlätta en större marknad för språkgranskningsprodukter. För korrekturläsning och andra språkgranskningsuppgifter behövs stöd för syntaktisk och semantisk information på menings- och textnivå, samt diskursinformation på dokument- och samlingsnivå. Med uppsvinget för öppna och fria kontorsprogram skulle konkurrensen kunna generera det moment som behövs för en förändring.

De producerade lexikala resurserna kan nyttjas som grundstenar för andra tillämpningar. I ScarrieLex är det förberett med fält för semantisk information. En första jämkning med övriga lexikala resurser framtagna vid institutionen har även skett, vilket underlättar användning i framtida tillämpningar. ScarrieLex och sammansättningsgrammatiken har i experimentella tester använts i trädbankskörningar. Sammansättningsanalysatorn har även använts för informella tester i Fastyprojektet.

Litteraturförteckning

- [Ahlbom & Sågvalld Hein 1999] Viktoria Ahlbom and Anna Sågvalld Hein. “Test Suites Covering the Functional Specifications of the Sub-components of the Swedish Prototype”. Technical report, Department of Linguistics, Uppsala University, Uppsala, 1999. SCARRIE Project Report, Del. 7.1.3.
- [Dahlqvist 1998] Bengt Dahlqvist. “A Swedish Text Corpus for Generating Dictionaries”. Technical report, Department of Linguistics, Uppsala University, Uppsala, 1998. SCARRIE Project Report, Del. 3.1.3.
- [Dahlqvist 1999] Bengt Dahlqvist. “Protokoll över stavningskontroll med MS Word 97 på testtext fredag4.txt”. Uppsala universitet, Institutionen för lingvistik, arbetsrapport, 1999.
- [Damerau 1964] F.J. Damerau. “A technique for computer detection and correction of spelling errors”. *Communications of the ACM*, 7:171–176, 1964.
- [EAGLES 1996] EAGLES. *Evaluation of natural language processing systems. Final Report*. EAGLES Document EAG-EWG-PR.2, ISBN: 87-90708-00-8, 1996.
- [Esselink 1998] Bert Esselink. *A Practical Guide to Software Localization*. John Benjamins Publishing, ISBN: 90-272-1954-0, 1998.
- [General Dictionary of UCP 1997] Department of Linguistics, Uppsala University, Uppsala. *The General Dictionary of UCP*, 1997.
- [Garlén 1988] Claes Garlén. *Svenskans fonologi*. Studentlitteratur, ISBN: 91-44281-51-X, 1988.
- [IEEE 1993] IEEE. *9945-2: 1993 (ISO/IEC) [IEEE/ANSI Std 1003.2-1992 and IEEE/ANSI 1003.2a-1992] Information Technology-Portable Operating System Interface (POSIX®) — Part 2: Shell and Utilities*. IEEE, New York, NY, USA, 1993.
- [IPA 1993] IPA 1993. “Council Actions on Revisions of the IPA”. *Journal of the International Phonetic Association*, 23(1):32–34, 1993.
- [Karlsson 1990] Fred Karlsson. “Constraint Grammar as a framework for parsing running text”. In H Karlgren, editor, *Papers presented to the 13th International Conference on Computational Linguistics*, volume 3, pp. 168–173, 1990. Helsinki.
- [Kukich 1992] K. Kukich. “Techniques for automatically correcting words in text”. *ACM Computing Surveys*, 24(4):377–439, December 1992.
- [Olsson 2003] Leif-Jöran Olsson. “Generalisering och utökning av den lexikala databasen Scarri-eLex”. Institutionen för lingvistik, Uppsala universitet, arbetsrapport, 2003.

- [Paggio 1999] Patrizia Paggio. “Evaluation Report”. Technical report, Centre for Language Technology (CST), Copenhagen, 1999. SCARRIE Project Report, Del. 7.2.
- [Paggio & Music 1998] Patrizia Paggio and Bradley Music. “Evaluation in the SCARRIE Project”. In *First International Conference on Language Resources & Evaluation*, pp. 277–282, 1998.
- [Wedbjer Rambell 1998a] Olga Wedbjer Rambell. “Error Typology for Automatic Proof-reading Purposes”. Technical report, Department of Linguistics, Uppsala University, Uppsala, 1998. SCARRIE Project Report, Del. 2.1.
- [Wedbjer Rambell et al 1998] Olga Wedbjer Rambell, Bengt Dahlqvist, Erik Tjong Kim Sang, and Nils Hein. “Error Database of Swedish”. Technical report, Department of Linguistics, Uppsala University, Uppsala, 1998. SCARRIE Project Report, Del. 2.1.3.2.
- [Wedbjer Rambell 1998c] Olga Wedbjer Rambell. “Multi-word Expressions for Swedish”. Technical report, Department of Linguistics, Uppsala University, Uppsala, 1998. SCARRIE Project Report, Del. 5.3.3.
- [Rei 1996] Fukui Rei 1996. “TIPA: A system for processing phonetic symbols in L^AT_EX”. *TUGboat*, 17(2):102–114, jun 1996.
- [Svensk ordbok 1986] Språkdata. *Svensk ordbok*. Esselte Studium AB, ISBN: 91-24-35307-8, 1986.
- [Sågvall Hein 1990] Anna Sågvall Hein. “Parsing by means of Uppsala Chart Processor (UCP)”. Technical report, Department of Linguistics, Uppsala University, Uppsala, 1990. Version 1.0.
- [Sågvall Hein 1998] Anna Sågvall Hein. “De morfologiska beskrivningarna i Sve.Ucp”. Technical report, Department of Linguistics, Uppsala University, Uppsala, 1998. Version 1.0.
- [Sågvall Hein 1999] Anna Sågvall Hein. “A Grammar Checking Module for Swedish”. Technical report, Department of Linguistics, Uppsala University, Uppsala, 1999. SCARRIE Project Report, Del. 6.6.3.
- [Sågvall Hein et al 1999] Anna Sågvall Hein, Leif-Jöran Olsson, Bengt Dahlqvist, and Erik Mats. “Evaluation Report for the Swedish Prototype”. Technical report, Department of Linguistics, Uppsala University, Uppsala, 1999. SCARRIE Project Report, Del. 8.1.3.
- [Scarrie TA 1996] Project Programme Scarrie. “Scarrie — Scandinavian Proof-reading Tools”. LE3-4239, Annex I, version 2.0, June 1996.
- [Thorell 1981] Olof Thorell. *Svensk ordbildningslära*. Esselte Studium AB 1981, ISBN: 91-24-30652-5, 1981.
- [Tiedemann 1999] Jörg Tiedemann. “ScarrieLex System Documentation”. Technical report, Department of Linguistics, Uppsala University, Uppsala, 1999.
- [Tiedemann 2002] Jörg Tiedemann. “MatsLex - a Multilingual Lexical Database for Machine Translation”. In Manuel González Rodríguez and Carmen Paz Suárez Araujo, editors, *Proceedings of the third International Conference on Language Resources and Evaluation*, volume VI, pp. 1909–1912. LREC, ELRA, ISBN: 2-9517408-0-8, 2002.

- [Tjong Kim Sang 1997] Erik Tjong Kim Sang. “Testing Corrie for Scarrie”. Technical report, Department of Linguistics, Uppsala University, Uppsala, 1997. SCARRIE Project Report, Del. 1.2.
- [UCP] Department of Linguistics, Uppsala University, Uppsala. *Uppsala Chart Processor System Documentation*.
- [Vosse 1994] Theo G. Vosse. *The Word Connection*. Enschede Uitgeverij, ISBN: 90-75296-01-0, 1994.
- [Véronis & Ide 1996] Jean Véronis and Nancy Ide. “Guidelines for Linguistic Software Development”. <http://www.lpl.univ-aix.fr/projects/multext/LSD/LSD2.html>.
- [Wells 1989] J. Wells. “Computer-coded phonemic notation of individual languages of the European Community”. *Journal of the International Phonetic Association*, 19(1):31–54, 1989.
- [Weijnitz 2002] Per Weijnitz. “Uppsala Chart Parser Light - Improving Efficiency in a Chart Parser”. Master’s thesis, Department of Linguistics, Uppsala University, August 2002.

Bilaga A

ScarrieLex

I detta kapitel ges en beskrivning av alla tillgängliga fält i den lexikala databasen ScarrieLex, vilka typer av data som kan finnas i varje fält och så vidare. Se kapitel 3 för mer bakgrundsinformation. Här finns också en användarhandledning för handhavandet av ScarrieLex, det vill säga utföra sökningar, uppdateringar och borttagningar.

A.1 Struktur

Rubrikerna ger en kort beskrivning av fältinnehållet samt fältnamnet i databasen. Under varje rubrik ges sedan exempel och specifika anvisningar för fältinnehåll.

A.1.1 Lemma — svlemma

Lemmafältet innehåller ett lemma enligt definition i SOB [Svensk ordbok 1986]:XII. Till exempel böjningsformerna *abbot*, *abboten* and *abbotarna* tillhör alla lemmat *abbot.nn*, där *nn* står för *NomeN*. För sammansättningar som har analyserats som sådana av UCP-systemet, har ett plustecken lagts till, till exempel *abortdebatt abort.nn+debatt.nn*. De möjliga ordklassförkortningarna är *nn* (*NomeN*), *pm* (*PropriuM*), *av* (*AdjektiV*), *pn* (*PronomeN*), *vb* (*VerB*), *ab* (*AdverB*), *al* (*ArtikeL*), *nl* (*Numeral*), *pp* (*PrepositioN*), *cn* (*konjunktion*), *sn* (*SubjunktioN*), *ie* (*Infinitivmärke*).

A.1.2 Lexem — svlexeme

I lexemfältet lagras ett lexemnummer, även det enligt definition i SOB [Svensk ordbok 1986]:XVI, för varje post, såväl enkla ord som fraser. Skönsvärde är 1.

A.1.3 Stam — svstem

I stamfältet lagras den tekniska stammen för varje post, såväl enkla ord som fraser.

A.1.4 Affix — svaffix

I affixfältet lagras affix som tillsammans med ett morfomvandlingsmönster kopplas till en teknisk stam via böjningsmönstret för varje post, såväl enkla ord som fraser.

A.1.5 Morfomvandlingsmönster — svpattern

I morfomvandlingsmönsterfältet lagras ett reguljärt uttryck eller en sträng som fogas tillsammans med affixet till den tekniska stammen för varje post. Detta gäller såväl enkla ord som fraser. Skönsvärde är \$ för ingen omvandling, det vill säga direktkonkatenering av stam och affix:

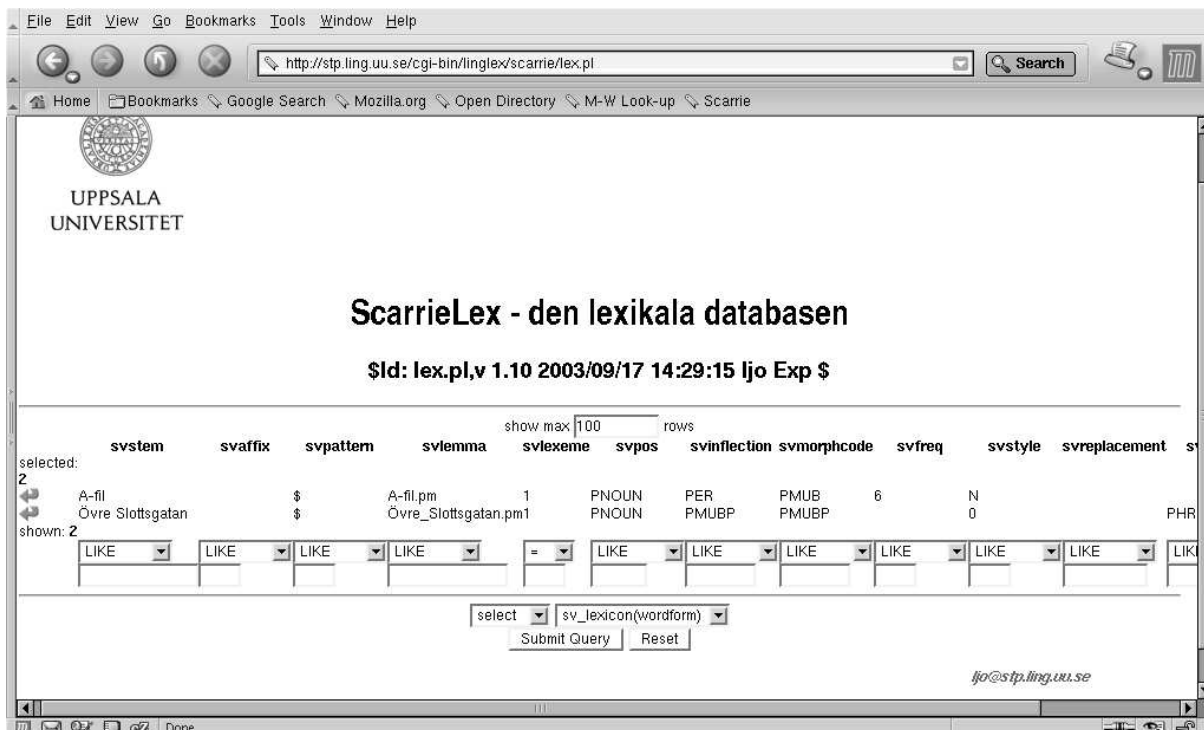
(39)	Stam <i>bord</i>	Morfomvandlingsmönster \$	affix <i>et</i>	Resulterande ytform <i>bordet</i>
------	---------------------	------------------------------	--------------------	--------------------------------------

Ett annat exempel, nu med en fras:

(40)	Stam <i>vilken som</i>	Morfomvandlingsmönster (<i>vil</i> k)	affix <i>\$Ia</i>	Resulterande ytform <i>vilka som</i>
------	---------------------------	---	----------------------	---

A.1.6 Ordklass — svpos

Fältet för ordklass innehåller information om ordklassstillhörighet. Möjliga ordklasser är: NOUN - *nomen*, PNOUN - *egennamn*, ADJ - *adjektiv*, PRON - *pronomen*, VERB - *verb*, ADV - *adverb*, ART - *artikel*, NUM -



Figur A.1: Exempel på information i fältet svsystem.

numeral, PREP - *preposition*, CONJ - *konjunktion*, och IE - *infinitivmärke*. Jämför med ordklassförkortningarna som används i lemmafältet ovan (A.1.1). För fraserna finns här även NP, VP, PP, PV, AD, QM, QU, samt SEP.

A.1.7 Böjningsmönster — svinflexion

Fältet för böjningsmönster innehåller ett mönsterord eller -fras, som indikerar vilket eller vilka böjningsparadigm lemmat tillhör. Till exempel har posten för ordformen *abbot*, STOL som mönsterord. För en mer utförlig beskrivning, se [Sågvall Hein 1998].

A.1.8 Morfosyntaktisk kod — svmorphcode

Fältet för morfosyntaktisk kod innehåller koder som uttrycker morfosyntaktiska egenskaper och några semantiska aspekter¹. Exempel på semantiska aspekter är kvantifiering, materia eller tidsuttryck. Alltso-
mallt 417 olika koder finns för tillfället.

A.1.9 Frekvens på ordformsnivå — svfreq

Frekvensfältet innehåller frekvensuppgiften från SvD/UNT-korpusen för postens ytform. Då Corrie-prototypen inte kunde hantera frekvensinformation för fraser, har den heller inte lagts in (se figur A.1). Skönsvärde är 1.

A.1.10 Stilinformation på ordformsnivå — svstyle

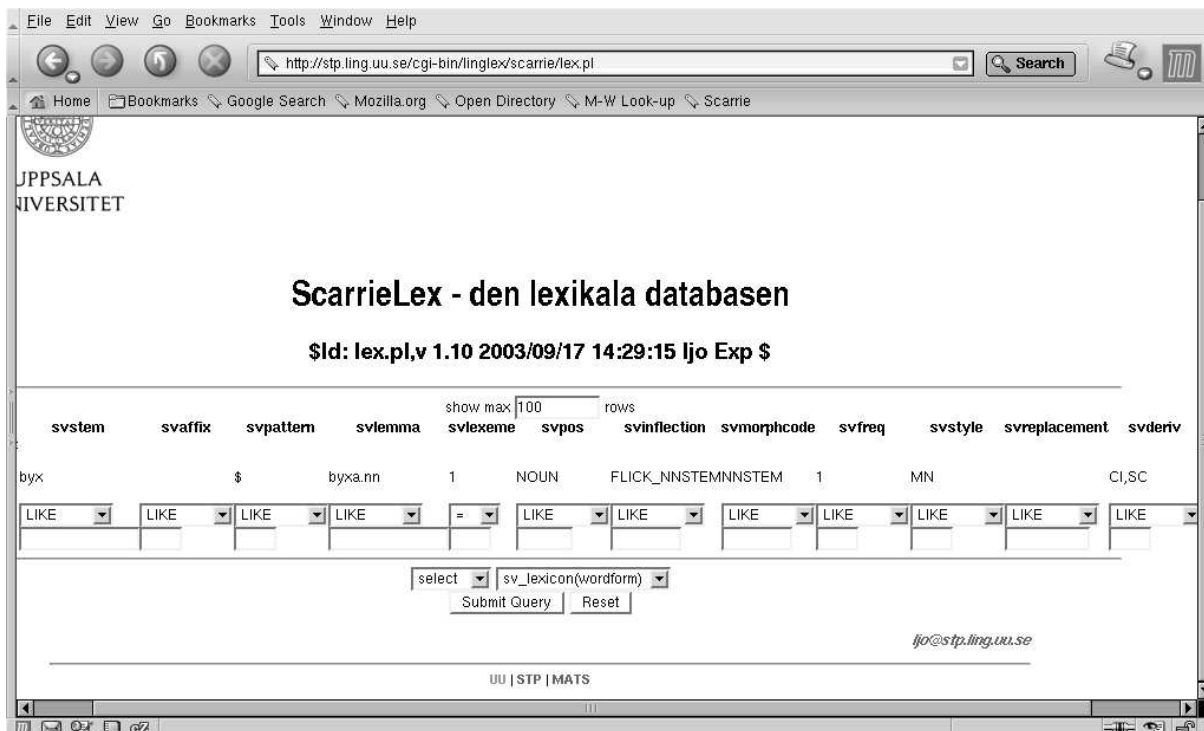
Stilinformation för posten kan vara något av:

- N - formen är godkänd under alla stilar,
- R - formen ersätts under alla stilar med ersättningsfältets innehåll (svreplacement),
- S[1–5]+ - formen är endast godkänd under stil [1–5], stil två används för slang och stil fyra och fem används för de två tidningarnas specifika språkbruk,
- C[1–5]+ - formen ersätts under angiven stil [1–5] med ersättningsfältets innehåll (svreplacement),
- E - formen är en flerordsenhet, vilket betyder att den inte används utanför de fraser som finns i lexikonet,
- I - formen är inte godkänd och flaggas som icke-godkänd om den hittas,
- M[NIS[1-5]+C[1-5]]+ - används tillsammans med stilmarkörerna N, S och C och betecknar ett morfem eller affix som måste vara del av en sammansättning för att vara godkänt,
- PN - enskild form av fras är godkänd under alla stilar,

¹Koderna har tagits fram av Olga Wedbjer Rambell.

- PR - enskild icke-godkänd form av fras ersätts under alla stilar med innehållet i ersättningsfältet (svreplacement).

Se även under rubrik för stilinformation på lemmanivå (A.1.13).



Figur A.2: Stilinformationsexempel ett.

A.1.11 Sammansättningsegenskaper på ordformsnivå — svderiv

Fältet för sammansättningsegenskaper innehåller information om sammansättningsegenskaper på ordformsnivå som används i sammansättningsgrammatiken: CI - godkänt som förled (compoundable in initial position), CA - godkänt i alla led (compoundable in all positions), CF - godkänt som efterled (compoundable in final position), CE - inte godkänt som sammansättningsled (not allowed in compounds), SC - kort element, godkänt i sammansättningar (short compound element). För ett exempel, se figur A.2. För en aktuell lista av attribut, se avsnitt 4.2.1, speciellt tabell 4.3.

A.1.12 Ersättningsalternativ på ordformsnivå — svreplacement

Om postens stilinformation för en ordform är R eller C[1-5] eller för en fras är PR, innehåller detta fält ett ersättningsalternativ, antingen en ordform eller en fras. Se figur A.3 för ett ordformsersättnings-exempel eller figur A.4 för ett frasersättnings-exempel. Se även ersättningsalternativ på lemmanivå samt figur A.5 nedan.

The screenshot shows the ScarrieLex web interface. The browser address bar contains `http://stp.ling.uu.se/cgi-bin/linglex/scarrie/lex.pl`. The page title is "UPPSALA UNIVERSITET" and the main heading is "ScarrieLex - den lexikala databasen". Below the heading, it says "\$Id: lex.pl,v 1.10 2003/09/17 14:29:15 ljo Exp \$".

The search results are displayed in a table with the following columns: svlemma, svlexeme, svpos, svinflection, svstem, svpattern, svaffix, svmorphcode, svfreq, svstyle, svreplacement. The results show three entries for the lemma "du.pn":

svlemma	svlexeme	svpos	svinflection	svstem	svpattern	svaffix	svmorphcode	svfreq	svstyle	svreplacement
du.pn	1		DIG	dej	\$		PNUSO	273	C13	dig
du.pn	1		DIG	dej	\$		PNUSO	273	S2	
du.pn	1		DIG	dig	\$		PNUSO	6793	N	

Below the table, there are search filters and a "Submit Query" button. The search criteria are: LIKE, =, LIKE, LIKE, LIKE, LIKE, LIKE, LIKE, LIKE, LIKE, LIKE.

Figur A.3: Stilinformationsexempel två.

The screenshot shows the ScarrieLex web interface. The browser address bar contains `http://stp.ling.uu.se/cgi-bin/linglex/scarrie/lex.pl`. The page title is "UPPSALA UNIVERSITET" and the main heading is "ScarrieLex - den lexikala databasen". Below the heading, it says "\$Id: lex.pl,v 1.10 2003/09/17 14:29:15 ljo Exp \$".

The search results are displayed in a table with the following columns: svstem, svaffix, svpattern, svlemma, svlexeme, svpos, svinflection, svmorphcode, svfreq, svstyle, svreplacement. The results show one entry for the phrase "en passent":

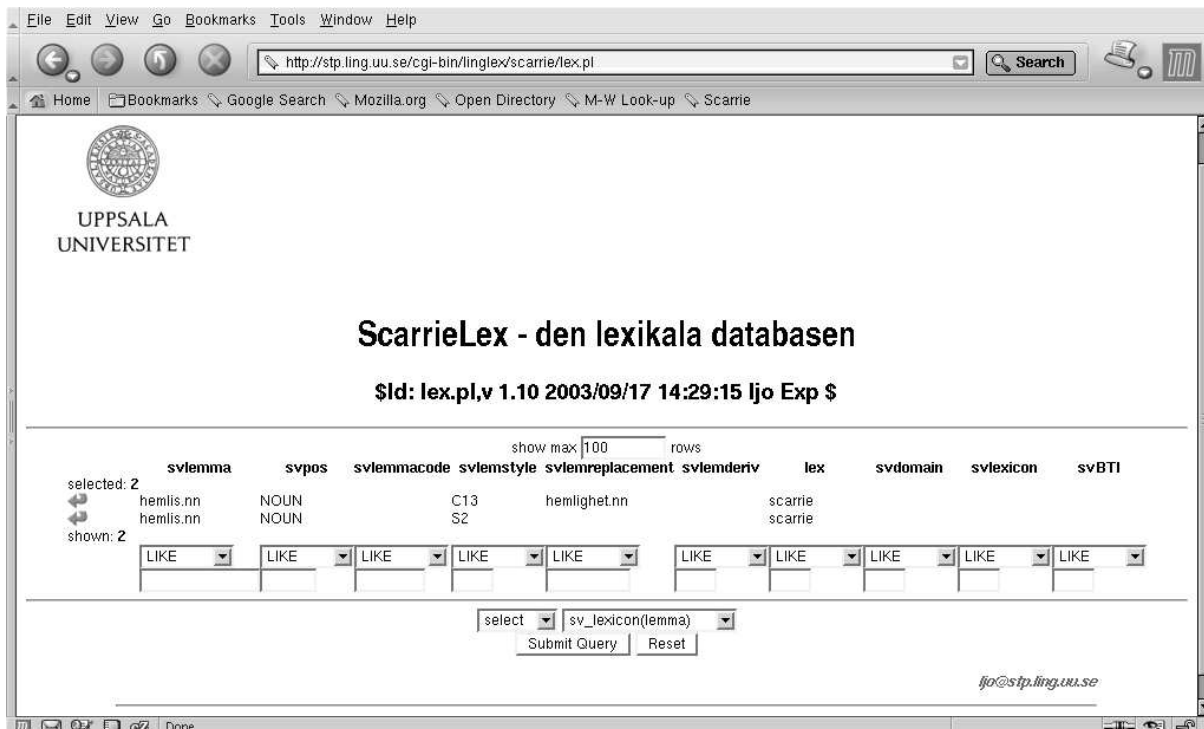
svstem	svaffix	svpattern	svlemma	svlexeme	svpos	svinflection	svmorphcode	svfreq	svstyle	svreplacement
en passent		\$	en_passent.ad	1	AD	en_passent.ad	FE		PR	i förbigående

Below the table, there are search filters and a "Submit Query" button. The search criteria are: LIKE, LIKE, LIKE, LIKE, =, LIKE, LIKE, LIKE, LIKE, LIKE, LIKE.

Figur A.4: Ett frasersättningsexempel.

A.1.13 Stilinformation på lemmanivå — svlemstyle

Stilinformation för lemmat posten är knuten till, vilket kan vara något av de stilar som anges under rubriken stilinformation på ordformsnivå ovan. Där svreplacement anges som ersättningsfält på ordformsnivå gäller i detta fall på lemmanivå fältet svlemreplacement.



Figur A.5: Exempel på lemmastilinformation och ersättning av alla former för hela lemmat.

A.1.14 Sammansättningsegenskaper på lemmanivå — svlemderiv

Fältet för sammansättningsegenskaper på lemmanivå innehåller information om sammansättningsegenskaper gemensamma för lemmat som används i sammansättningsgrammatiken: NO_HYPH - inget bindestreck (no hyphen), NEED_HYPH - kräver bindestreck (needs hyphen), SMS - kräver foge-s, SME - kräver foge-e, ABBR - förkortning (abbreviation), ACCR - akronym (acronym), NO_S - inget foge-s (no binding-s), NO_E - inget foge-e (no binding-e), TIME - godkänt i tidsuttryck (allowed in time span compounds). Se avsnitt 4.2.1, speciellt tabell 4.3 för en aktuell lista av attribut. För ett exempel med attribut på lemmanivå, se figur A.5.

A.1.15 Ersättningsalternativ på lemmanivå — svlemreplacement

Om postens stilinformation för lemmat är antingen R, PR eller C [1–5] innehåller detta fält ett ersättningsalternativ, antingen en ordform eller en fras som gäller som ersättning för lemmats alla former. Se figur A.5 för ett ersättningsexempel.

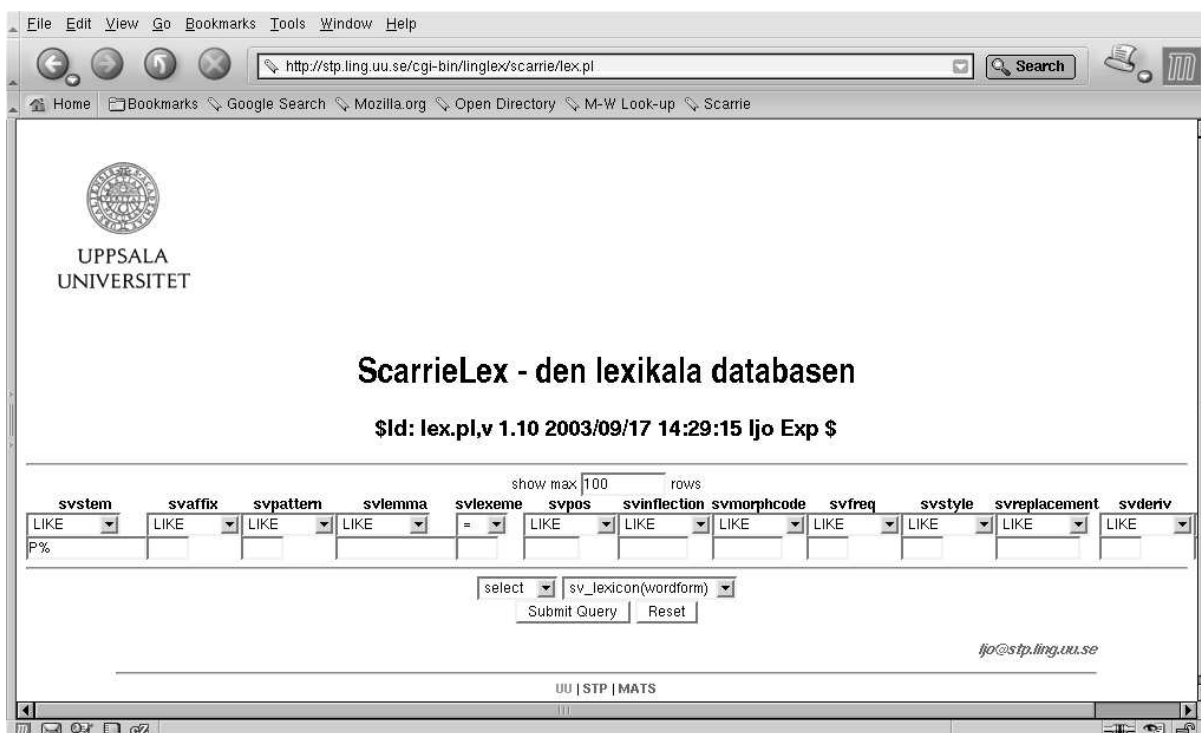
A.1.16 lexikon — lex

Detta fält anger vilket lexikon posten tillhör. Värdet `scarrie` används för de poster som utvunnits ur SvD/UNT-koruset (master dictionary). Här kan uppgift om användarspecifika lexikon läggas till.

A.2 Användning av ScarrieLex

De operationer som kan utföras på innehållet i ScarrieLex är de SQL-lika `select`, `add`, `delete` och `update`. För varje fält kan man välja begränsningar på operationen med hjälp av någon av operatorerna `LIKE`, `=`, `<>` eller `REGEXP` från droplistorna som finns ovanför varje fält i webbgränssnittet.

Om operatoren `LIKE` används kan två jokertecken användas i fältet, `%` (procenttecken) och `_` (understreck). `%` står för valfritt antal valfria tecken, se figur A.6. `_` står för valfritt tecken i en position, se figur A.7.



Figur A.6: Begränsning av sökning med hjälp av operatoren `LIKE` och jokertecken `%`.

Som ses i figur A.8, är operatoren `LIKE` dessutom storleksberoende i sina matchningar i vissa fält. Att skriva in någon av strängarna `ÖSA`, `ösa` eller `öSa` i fältet ger samma resultat. De fält som inte är storleksberoende är `svsystem`, `svlemma`, `svreplacement` och `svlemreplacement`.

Operatoren `=` är däremot storleksberoende och hela fältets innehåll måste matcha exakt (se figur A.9).

För att göra disjunktiva urval kan operatoren `<>`, skilt från, användas, se figur A.10.

För mer avancerade matchningar kan operatoren `REGEXP` användas. Då tolkas fältinnehållet som ett re-

Uppsala Universitet logo and name.

ScarrieLex - den lexikala databasen

\$Id: lex.pl,v 1.10 2003/09/17 14:29:15 ljo Exp \$

show max 100 rows

system	svaffix	svpattern	svlemma	svlexeme	svpos	svinflection	svmorphcode	svfreq	svstyle	svreplacement	svderiv
LIKE	LIKE	LIKE	LIKE	=	LIKE	LIKE	LIKE	LIKE	LIKE	LIKE	LIKE
P_A											

select sv_lexicon(wordform)

Submit Query Reset

ljo@stp.ling.uu.se

UU | STP | MATS

Figur A.7: Begränsning av sökning med hjälp av operatorn LIKE och jokertecken _.

Uppsala Universitet logo and name.

ScarrieLex - den lexikala databasen

\$Id: lex.pl,v 1.10 2003/09/17 14:29:15 ljo Exp \$

show max 100 rows

system	svaffix	svpattern	svlemma	svlexeme	svpos	svinflection	svmorphcode	svfreq	svstyle	svreplacement	svderiv
LIKE	LIKE	LIKE	LIKE	=	LIKE	LIKE	LIKE	LIKE	LIKE	LIKE	LI
			arb%		Verb						

select sv_lexicon(wordform)

Submit Query Reset

ljo@stp.ling.uu.se

UU | STP | MATS

Figur A.8: Begränsning av sökning med hjälp av operatorn LIKE (storleksberoende).

Uppsala Universitet logo and name.

ScarrieLex - den lexikala databasen

\$Id: lex.pl,v 1.10 2003/09/17 14:29:15 ljo Exp \$

show max 100 rows

system	svaffix	svpattern	svlemma	svlexeme	svpos	svinflection	svmorphcode	svfreq	svstyle	svreplacement	svderiv
LIKE	LIKE	LIKE	LIKE	=	=	LIKE	LIKE	LIKE	LIKE	LIKE	LIKE
			arb%		VERB						

select sv_lexicon(wordform)

Submit Query Reset

ljo@stp.ling.uu.se

UU | STP | MATS

Figur A.9: Begränsning av sökning med hjälp av operatorm = (identitet, storleksberoende).

Uppsala Universitet logo and name.

ScarrieLex - den lexikala databasen

\$Id: lex.pl,v 1.10 2003/09/17 14:29:15 ljo Exp \$

show max 100 rows

system	svaffix	svpattern	svlemma	svlexeme	svpos	svinflection	svmorphcode	svfreq	svstyle	svreplacement	svderiv
LIKE	LIKE	LIKE	LIKE	=	<>	LIKE	LIKE	LIKE	LIKE	LIKE	LI
			arb%		NOUN						

select sv_lexicon(wordform)

Submit Query Reset

ljo@stp.ling.uu.se

UU | STP | MATS

Figur A.10: Begränsning av sökning med hjälp av skilt från-operatorm <> (not).

guljärt uttryck ([IEEE 1993] avsnitt Regular Expression Notation), se figur A.11.

The screenshot shows a web browser window with the URL `http://stp.ling.uu.se/cgi-bin/linglex/scarrie/lex.pl`. The page header includes the Uppsala University logo and the text "UPPSALA UNIVERSITET". The main heading is "ScarrieLex - den lexikala databasen" with a subtitle "\$Id: lex.pl,v 1.10 2003/09/17 14:29:15 ljo Exp \$".

The search interface features a table of columns with dropdown menus for operators and a text input field. The columns are: `system`, `svaffix`, `svpattem`, `svlemma`, `svlexeme`, `svpos`, `svinflection`, `svmorphcode`, `svfreq`, `svstyle`, `svreplacement`, and `svderiv`. The `system` column is set to "REGEXP" and the input field contains "[PQ].+". Other columns are set to "LIKE".

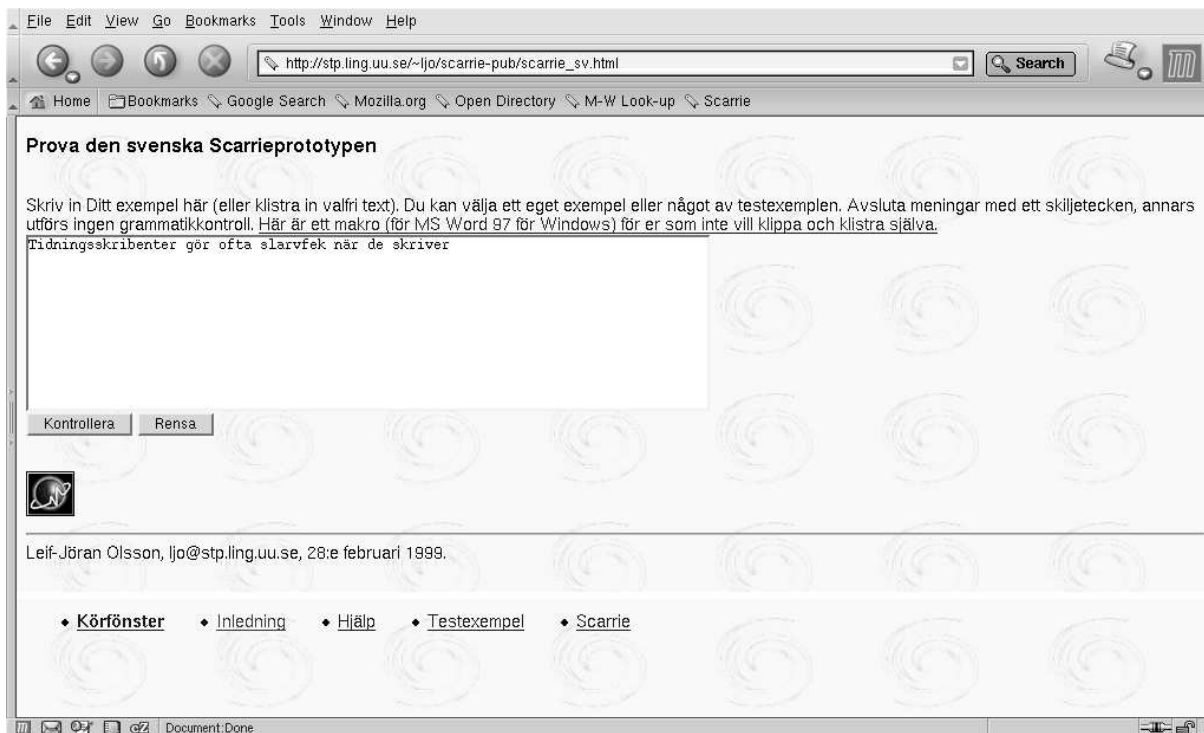
Below the table, there is a "show max 100 rows" label, a "select" dropdown menu, and a "sv_lexicon(wordform)" dropdown menu. There are "Submit Query" and "Reset" buttons. The footer includes the email address `ljo@stp.ling.uu.se` and the text "UU | STP | MATS".

Figur A.11: Begränsning av sökning med hjälp av operatör REGEXP.

Bilaga B

Användning av den svenska Scarrie-demonstratorn

Demonstratorn för Scarrie-piloten, finns tillgänglig för allmänheten via URL:
<http://stp.ling.uu.se/~ljo/scarrie-pub/>

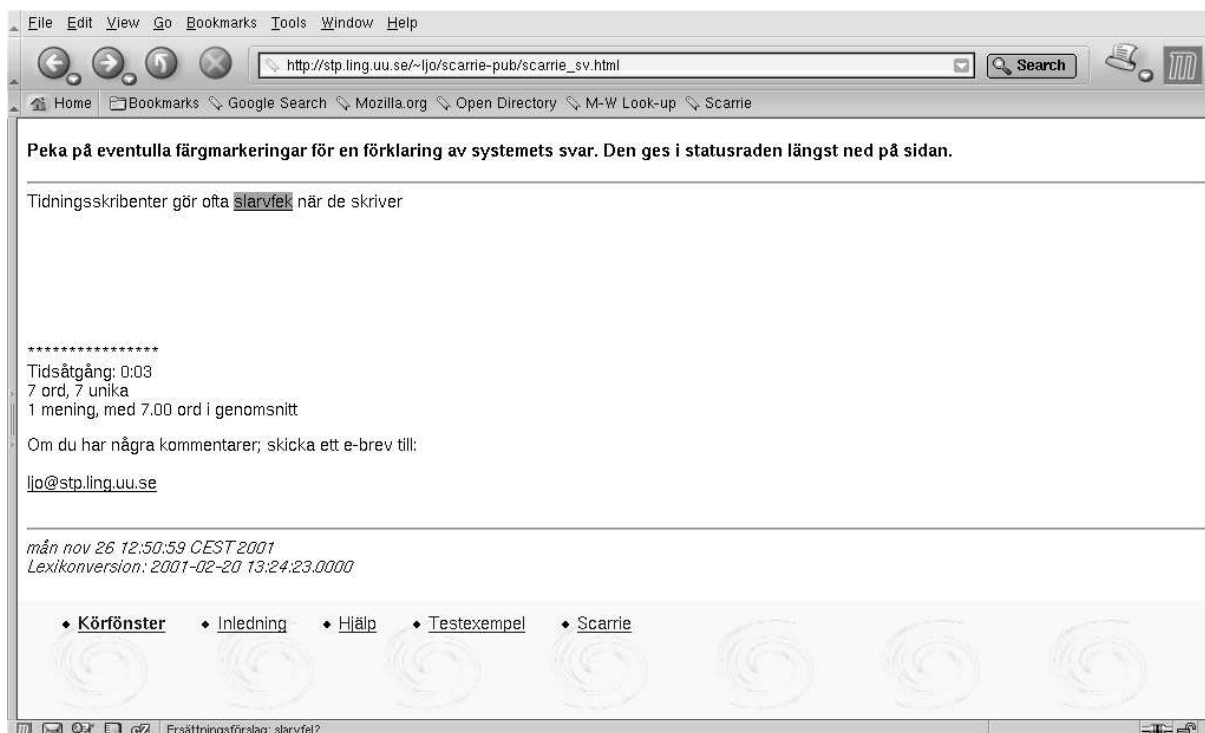


Figur B.1: Webbgränssnittet för den svenska Scarrie-demonstratorn.

B.1 Input

Välj exempelmeningar för att se vilka grammatikfel som Scarrie-piloten kan hantera. Se figur B.1. Börja med att köra några testexempel. Det ger en känsla för vad Scarrie-piloten klarar av. Kom ihåg att varje mening måste avslutas med interpunktionstecken, annars utförs ingen grammatikkontroll.

För enklare tillgång kan ett word-makro laddas ner och installeras under MS Windows. Det finns också ett snarlikt javascript-bokmärke om markera-och-klicka-för-kontroll önskas i webbläsaren.



Figur B.2: Granskningsrapport från webbgränssnittet.

B.2 Output

I granskningsrapporten, se figur B.2, är felen markerade med olika färger för att indikera feltyp. Färgkodsguide:

- Röd används för ordfel
- Grön används för grammatikfel
- Gul används för typografiska fel (till exempel dubbla mellanslag) och misstänkta sammansättningar (till exempel fyrvåningshus).

B.2.1 Diagnos/ersättningsförslag

Peka på det färgmarkerade avsnittet för att se en beskrivning i statusraden. Om du hittar några fel, så bör du köra igen efter att ha korrigerat felet, eftersom fel kan ge upphov till följdfel eller överlappa varandra. En del webbläsare kan inte separera och presentera två intilliggande markeringar ordentligt. Ingen webbläsare kan presentera överlappande markeringar på ett vettigt sätt. Eftersom detta oftast uppstår vid följdfel, löser en korrektion av det felet och en omkörning problemet.

Bilaga C

Testord och exempelkorrektioner använda under utvecklingen av grafem-till-fonem-överföringsreglerna

C.1 Testord

Följande ord har använts för utprovning av de svenska grafem-till-fonem-överföringsreglerna under utvecklingen av Scarrie-piloten.

sju	sked	skicka	tjock	kyrka	kälke
jama	jumper	joddla	tjuv	kung	bank
göra	gädda	jämföra	dygn	många	omtänksam
balja	arg	torg	blinka	gagna	välsigna
dvärg	svälja	gick	köksbord	häxa	razzia
sjal	sjutton	själva	analys	scen	precis
addition	pension	diskussion	lova	tjog	konung
garage	energi	giraff	blått	skärm	penna
justera	ljus	djup			

Tabell C.1: Testord använda under utvecklingen av grafem-till-fonem-reglerna.

C.2 Exempelkorrektioner

Följande exempel illustrerar exempelkorrektioner från den uttalsbaserade ordkorrektionen, där vinkelparentesen > indikerar användarinmatning och de numrerade orden är möjliga korrektioner föreslagna av korrektionsmodulen i Scarrie-piloten. Observera att alla felexempel har fått det korrekta ordet som bästa korrektionsförslag.

- > jöra
 - 1. göra [900]
 - 2. gör [700]
 - 3. för [600]
 - 4. före [600]
 - 5. förra [600]
 - 6. fyra [600]

- > grött
 - 1. gått [600]

- > kann
 - 1. kan [700]
 - 2. fanns [600]
 - 3. hans [600]

- > shu
 - 1. sju [900]
 - 2. så [600]

- > jup
 - 1. djup [900]

- > tjåg
 - 1. tjog [900]

- > jystera
 - 1. justera [800]

- > diskution
 - 1. diskussion [800]

- > diskission
 - 1. diskussion [800]

- > konugn
 - 1. konung [800]