UPPSALA UNIVERSITET

Grundläggande Textanalys VT 2016

Språkgranskning (2)

Eva Pettersson

eva.pettersson@lingfil.uu.se

Översikt

- Förra gången
  - Stavningskontroll
    - Allmänt om stavningskontroll
    - Feligenkänning
    - Felkorrigering
    - Samarbetsuppgift
- Denna gång
  - Grammatikkontroll
  - Stilkontroll
  - Kontrollerat språk
  - Språkgranskningssystem, med fokus på MS Word och Granska
  - Kort om labben
  - Samarbetsuppgift

UPPSALA UNIVERSITET

GRAMMATIKKONTROLL

UPPSALA UNIVERSITET

Varför behövs grammatikkontroll?

- Slarvfel (performansfel)
- Komplexa grammatiska konstruktioner (kompetensfel)
- Stilkontroll
- Andraspråksinlärning
- Dyslexi och andra skrivsvårigheter
- Stavningskontroll i kontext

UPPSALA UNIVERSITET

Vad förväntas av det ideala grammatikkontrollprogrammet?

- Känna igen och larma för alla grammatiska fel
- Känna igen och acceptera alla grammatiskt korrekta strukturer
- Ge ett korrekt ersättningsförslag för alla ogrammatiska strukturer och kontextuellt felaktiga ordformer

Realistiska förväntningar på grammatikkontrollprogrammet

- Undvika falska alarm, även om inte alla fel upptäcks
- Fokusera på typiska och/eller frekventa feltyper
  - korpusstudie
  - anpassning till olika typer av skribenter, såsom dyslektiker eller andraspråksinlärare
- Fokusera på typer av fel som är lätta att hitta

Grammatikkontrollens delar

1. **Feldetektering** (*error detection*)
   Fel i nominalfrasen: **det tidiga 1800-talen*

2. **Feligenkänning** (*error recognition*)
   Numerusinkongruens determinerare-huvudord

3. **Feldiagnos** (*error diagnosis*)
   Substantivet står i plural i stället för singular
   Determineraren står i singular i stället för plural

4. **Felkorrigering** (*error correction*)
   det tidiga 1800-talet
   de tidiga 1800-talen

Utmaningar

- Hur kan man känna igen så många fel som möjligt utan att få för många falska alarm?
- Långa och komplexa strukturer
  de i paragraf 5 stadgade **undantaget*
- Undantag från normalfallet
  Strindbergs** Röda **Rummet
- Begränsat till meningsnivå
  *Jag träffade **en flicka** igår. *De var mycket söta.*
- Hur sträng bör man vara?
  Jag insåg att det här **går inte.*
- Hur hanterar man dialektala uttryck?
  **Denna boken är bra.*

UPPSALA UNIVERSITET

Strategier för grammatikkontroll

- Grammatikbaserad grammatikkontroll
- Fenomenbaserad grammatikkontroll
- Sannolikhetsbaserad grammatikkontroll

UPPSALA UNIVERSITET

Grammatikbaserad grammatikkontroll

- Fullständig syntaktisk analys av meningen
- Relaxering för att hantera ogrammatiska strukturer
  - Accepterande regler
    en hund$_{NP}$
  - Relaxerade regler
    ett hund$_{NP\ med\ \ kongruensfel}$

Grammatikbaserad grammatikkontroll

- Styrka ☺
  Kan hantera strukturer där två av varandra oberoende ord inte står intill varandra
  Problemet* *med den kommunala fördelningen från inkomstskatter för säsongsanställda är* ***svår* *att lösa*

- Svagheter ☹
  - robusthet
  - effektivitet
  - regelskrivande är tidskrävande och kräver stor lingvistisk och programmeringsteknisk kunskap

UPPSALA UNIVERSITET

Fenomenbaserad grammatikkontroll

- Inriktning på förväntade fel
  - kongruensfel: artikel i neutrum följt av substantiv i utrum
    **ett hund*
- Partiell syntaktisk analys
- Reguljära uttryck

Fenomenbaserad grammatikkontroll

- Styrkor ☺
  - robusthet
  - effektivitet

- Svaghet ☹
  Begränsat till felaktigheter mellan ord som står nära varandra, såsom otillåtna ordklasskombinationer och kongruens inom nominalfrasen
  **kunnat kommit*
  **en rött hus*

UPPSALA UNIVERSITET

Sannolikhetsbaserad grammatikkontroll

- Sannolikhet för n-gram av ordsekvenser baserat på korpusdata
- Trigram mest effektivt
  - unigram och bigram för informationsfattiga
  - fyrgram, femgram etc leder till *data sparseness*
- Osannolika trigram byts ut mot mer sannolika trigram

  John ***came form*** *the house* → *John* ***came from*** *the house*

Sannolikhetsbaserad grammatikkontroll

- Styrkor ☺
  - (relativt) lätt att implementera
  - ingen lingvistisk kompetens krävs
- Svagheter ☹
  - alla ordformer kan inte täckas in av korpusen – data sparseness
  - sannolikhetsberäkningar kan vara resurskrävande i termer av datorkraft

UPPSALA UNIVERSITET

STILKONTROLL

Vad innefattar stilkontroll?

- Genomsnittlig ord- och meningslängd
- Minusord
  - ålderdomliga
  - byråkratiska
  - vardagliga/talspråkliga
- Vissa syntaktiska konstruktioner
  - passiv
- Konsekvenskontroll
  - *ska – skall*
  - *mej – mig*

UPPSALA UNIVERSITET

KONTROLLERAT SPRÅK

UPPSALA UNIVERSITET

Kontrollerat språk

- Specialfall av stilkontroll
- Delmängd av naturligt språk
- Restriktioner på vokabulär, grammatik och stil
- Möjliga mål:
  - Enhetlig layout, terminologi och stavning
  - Ökad läsbarhet
  - Tydlighet (undvik tvetydighet)
  - Grammatisk och terminologisk korrekthet
  - Maskinöversättningsbar källtext

Simplified English

- Undvik passiv form
- Använd om möjligt artikel i nominalfraser
- Använd enkla tempusformer
- Var konsekvent i språkbruk och meningsuppbyggnad
- Undvik flerledade sammansättningar
- Använd korta meningar

Kontrollerade språk

- AECMA Simplified English
  - Luftfartsindustri, såsom Boeing Simplified English (1990)
- Caterpillar Technical English (CTE)
  - KANT-systemet
- IBM EasyEnglish
- Scaniasvenska
  - Utvecklat av Uppsala universitet

Hur upplevs kontrollerat språk?

- Finns ett motstånd från användarna…
  - Krångligt och tidsödande att lära sig
  - Känner sig inte fria att skriva som de vill
  - Måste lära sig det kontrollerade språket
  - Måste lära sig den programvara som används för att se till att det kontrollerade språket följs
  - ”den tid man sparar på maskinöversättning äts upp av den tid det tar för skribenten att anpassa sitt skrivande till det kontrollerade språket”

UPPSALA UNIVERSITET

Uppnås målet?

- Undersökningar har visat att kontrollerat språk:
  - Ger högre kvalitet på texten
  - Ökar läsbarheten
  - Ger bättre översättningar (såväl manuella som maskinella)

Komponenter i en checker för kontrollerat språk

- Ordkontroll
  - Rätt terminologi
  - Rätt stavningsvariant (color/colour)
  - Akronymer (IKEA/Ikea)
- Grammatikkontroll
- Stilkontroll
  - Ord- och meningslängd
  - Datumformat
- Ersättningsförslag

Scania Checker

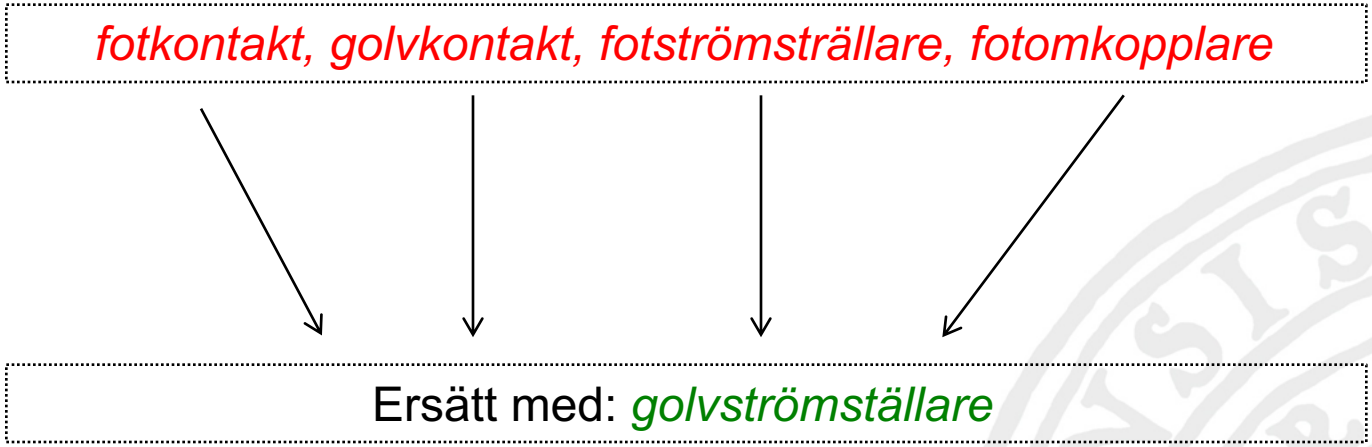
- Utvecklat vid institutionen för lingvistik, Uppsala universitet
- Används dagligen av Scanias skribenter sedan början av 2000-talet
- Mål:
  - Minska översättningskostnaden
  - Minska antalet termvarianter och därmed öka läsbarheten
  - Korrekt stavning och grammatik

Färguppmärkning i Scania Checker

- Gulmarkering
  - Okända ord (saknas i lexikonet)
- Rödmarkering
  - Minusord med ersättningsförslag
- Grönmarkering
  - grammatikfel

UPPSALA UNIVERSITET

Minusord I Scania Checker



UPPSALA UNIVERSITET

SPRÅKGRANSKNINGSSYSTEM

Tidiga språkgranskningssystem

- Critique
- CORRie

UPPSALA UNIVERSITET

Critique

- Första praktiska grammatikbaserade språkkontrollsystemet (80-talet)
- IBM
- Engelska och franska
- Föregångare till språkkontrollen i den engelska versionen av Microsoft Word

Feltyper i Critique

- 25 grammatiska feltyper, fördelade på fem kategorier:
  1. numerusinkongruens **many book/he go*
  2. fel pronomenform **between you and I*
  3. fel verbform **seems to been/had expect*
  4. skiljeteckenfel
  5. sammanblandningar **who's – whose*

- 85 stilistiska fel

CORRie

- Utvecklat för holländska
- Fem generella feltyper

1.	insättning	**tåget har **har** redan gått – hus bil*
2.	borttagning	**tåget redan gått*
3.	substitution	**the **well** performance*
4.	transposition	**jag tror att tåget **går inte** idag*
5.	feature mismatch	****den** lilla huset – männen är **glad***

Språkgranskningssystem för svenska

- Grammatifix, Lingsoft
- Granska, Nada, KTH
- SCARRIE, Uppsala universitet
  - Modifierad variant: Scania Checker
- Skribent, Norstedts Ordbok

UPPSALA UNIVERSITET

GRAMMATIFIX (MS WORD)

UPPSALA UNIVERSITET

Grammatifix

- Används för språkkontroll i den svenska versionen av Microsoft Word
- Stavfel hanteras i separat modul (Orthografix)

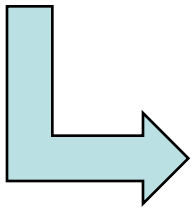
Grammatikkontroll i Grammatifix

- Parsning med SWECG (Swedish Constraint Grammar) i fem steg:
  1. Förbehandling
  2. Morfologisk analys
  3. Morfologisk disambiguering
  4. Morfosyntaktisk mappning
  5. Syntaktisk disambiguering

Steg 1: Förbehandling

- Tokenisering
  - Cirka 150 fraser
  - Cirka 5 000 förkortningar

Dessa entreprenöriella faktorer hade än så länge dämpat explosionen.



*dessa
entreprenöriella
faktorer
hade
än_så_länge
dämpat
explosionen
$.

Steg 2: Morfologisk analys

SWETWOL, cirka 75 000 grundformer

```
”<*dessa>”
        ”denna” <**c> <DEM><MD>DET UTR/NEU DEF PL NOM @DN
        ”denna” <**c> <DEM>PRON UTR/NEU DEF PL NOM
”<entreprenöriella>”
”<faktorer>”
        ”faktor” N UTR INDEF PL NOM
”<hade>”
        ”ha” <AUX> V ACT PAST
”<än_så_länge>”
        ”än_så_länge” <COLLOCATION> ADV
”<dämpat>”
        ”dämpa” V ACT SUPINE
        ”dämpa” <PCP2> A NEU INDEF SG NOM
”<explosionen>”
        ”explosion” N UTR DEF SG NOM
”<$.>
        ”$.” CLB <PUNCT>
```

Morfologisk analys av okända ord

- För ord som saknas I SWETWOL-lexikonet
  - Ett 60-tal regler baserade på suffix
  - Okända ord som inte matchas av någon regel antas vara substantiv

```
”<entreprenöriella>”
        ”entreprenöriella” <NON-SWETWOL> A UTR/NEU DEF SG NOM
        ”entreprenöriella” <NON-SWETWOL> A UTR/NEU DEF/INDEF PL NOM
```

Steg 3: Morfologisk disambiguering

Cirka 2 100 morfologiska disambigueringsregler

”<*dessa>”
”denna” <**c> <DEM><MD>DET UTR/NEU DEF PL NOM @DN
~~”denna” <**c> <DEM>PRON UTR/NEU DEF PL NOM~~
”<entreprenöriella>”
”entreprenöriella” <NON-SWETWOL> A UTR/NEU DEF/INDEF PL NOM
”<faktorer>”
”faktor” N UTR INDEF PL NOM
”<hade>”
”ha” <AUX> V ACT PAST
”<än_så_länge>”
”än_så_länge” <COLLOCATION> ADV
”<dämpat>”
”dämpa” V ACT SUPINE
~~”dämpa” <PCP2> A NEU INDEF SG NOM~~
”<explosionen>”
”explosion” N UTR DEF SG NOM
”<$.>
”$.” CLB <PUNCT>

Steg 4: Morfosyntaktisk mappning

- Normalt I SWECG: syntaktiska funktionstaggar
- I Grammatifix trivialt steg: Alla ord tilldelas båda taggarna @OK och @ERR

```
”<*dessa>”
        ”denna” <**c> <DEM><MD>DET UTR/NEU DEF PL NOM @DN
        @OK @ERR
”<entreprenöriella>”
        ”entreprenöriella” <NON-SWETWOL> A UTR/NEU DEF/INDEF PL NOM
        @OK @ERR
”<faktorer>”
        ”faktor” N UTR INDEF PL NOM
        @OK @ERR
```

Steg 5: Syntaktisk disambiguering

Feldetekteringsregler väljer om @OK-taggen eller @ERR-taggen ska gälla

```
(@w =s! (@ERR) ; För ett ord (@w) väl (=s!) feltaggen (@ERR)
   (0 N-DEF)       ; om ordet själv (0) är ett substantiv i bestämd form (N-DEF) och
   (-2 GEN)        ; andra ordet till vänster (-2) är en genitiv (GEN) och
   (-1 A-DEF)      ; det första ordet till vänster är ett adjektiv i bestämd form (A-DEF)
```

unionens snabba utvidgningen

Kontrollera ordformen ***utvidgningen***. Om ett substantiv styrs av en genitiv, t.ex. *unionens*, bör det stå i obestämd form.

Feltyper i Grammatifix

- 43 feltyper fördelade på tre kategorier
  - Grammatiska fel (26)
    - Fokus på verbkedjor och kongruens i nominalfraser
  - Stilbrott på ordnivå (3)
    - Byråkratiska ord *hemställan, vederlag*
    - Talspråkliga ord *dej, redaktörn*
    - Ålderdomliga ord *särla, tillförne*
  - Brott mot skrivkonventioner (14)
    - Parenteser, datumformatering, interpunktion mm

UPPSALA UNIVERSITET

GRANSKA

UPPSALA UNIVERSITET

Granska

- Utvecklat av Nada, KTH
- Grundar sig på tidigare forskningsprojekt
  - Plita
  - Stava
- Partiell parsning i kombination med hjälpregler
  - Omfattande analys endast om meningen verkar vara felaktig
- Särskilt inriktat mot användare med svenska som andraspråk

UPPSALA UNIVERSITET

Fokus på tre feltyper

1. Särskrivna sammansättningar
   *Han orkade inte flytta **sten bumlingarna**.*

2. Inkongruens i nominalfrasen
   ***ett** villa*

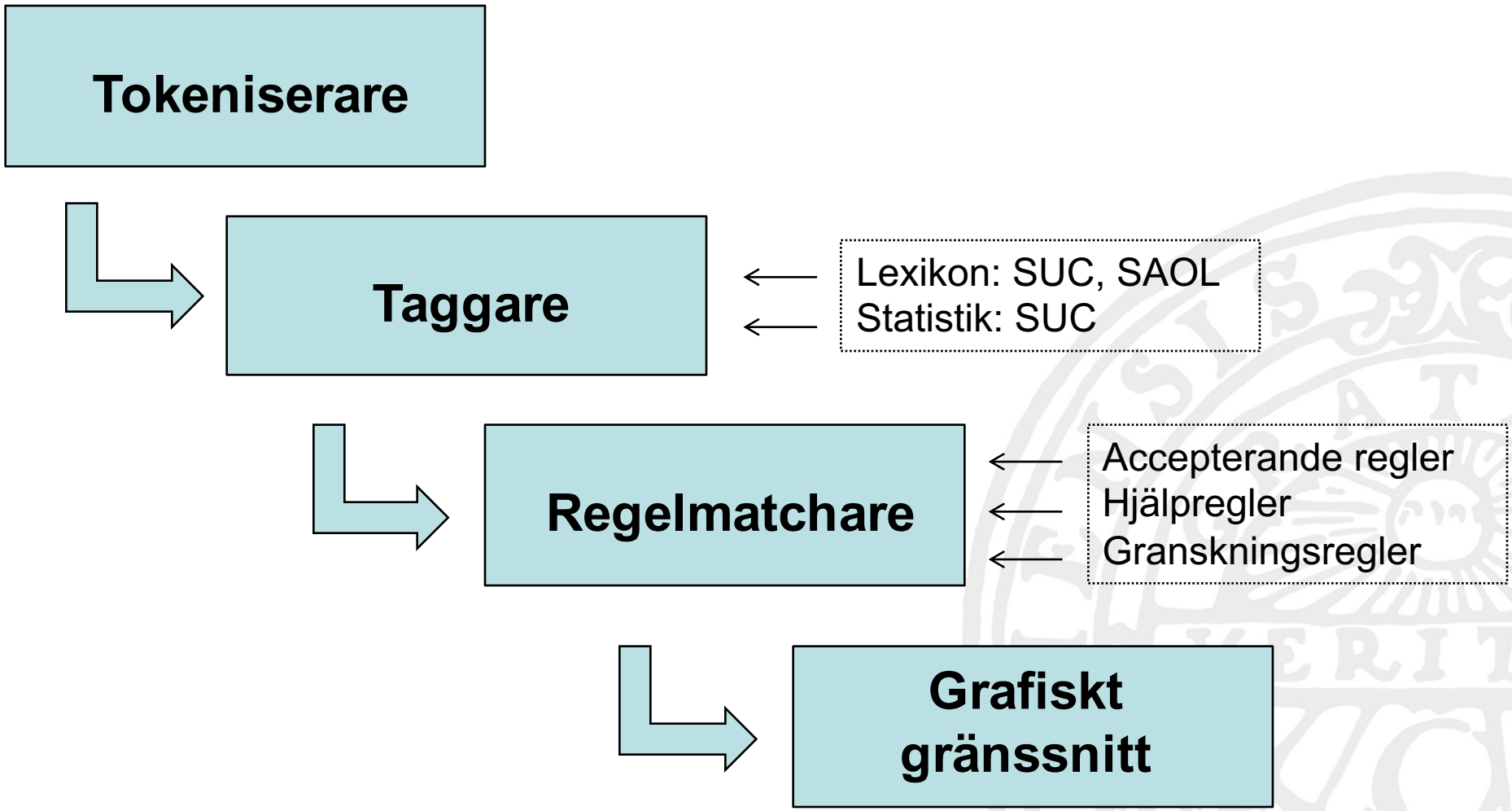
3. Inkongruens i predikativ
   *Äktenskapet är **baserad** på kärlek.*

Fler feltyper

- Felaktiga sammansättningar (Svenska skrivregler, 1991)
  undertiden
- Sammanblandningar
  tillhör en av (ska vara *är en av* eller *tillhör*)
- Felaktig preposition (svensk handordbok 1966)
  med utgångspunkt ***från***
- Objektsform efter preposition
  Han skickade en rad vänliga brev till ***de***
- Tautologi
  orsaken *till* ***anledningen***
- Böjningsfel i verbfras
  Vi kommer spela en låt av Ebba Grön
- Ordföljdsfel
  Han sa att han ***sparkade inte*** *bollen*

UPPSALA UNIVERSITET

Granskasystemet



UPPSALA UNIVERSITET

Regeluppsättning

- Accepterande (positiva) regler
  - Hanterar korrekta konstruktioner som måste undantas för att undvika falska alarm
    en massa arbetare
- Granskningsregler (negativa)
  - Detekterar och korrigerar grammatiska felaktigheter
- Hjälpregler
  - Utökad, lokal frasstrukturanalys
  - Anropas endast vid behov

UPPSALA UNIVERSITET

Regelfilens struktur

```
category regler {
info (”exempelregler”)
link(”www” ”namn på länk”)
}

regelnamn@regler
{
X()
-->
action(scrutinizing)
}
```

Vänsterled (strängen man vill matcha)

Högerled (hur man vill rätta strängen)

UPPSALA UNIVERSITET

Några vanliga särdrag

- **wordcl** ordklass
- **gender** genus
- **num** numerus
- **spec** species
- **vbf** verbform
- **case** kasus
- **text** ordsträngen
- **lemma** lemma

UPPSALA UNIVERSITET

Några vanliga särdragsvärden

- **dt** determinerare
- **jj** adjektiv
- **nn** substantiv (nomen)
- **def** bestämd form (definit)
- **utr** utrum
- **vb** verb
- **pn** pronomen
- **pc** particip
- **ie** infinitivmärke

UPPSALA UNIVERSITET

Några logiska operatorer

- **=**
  lika med — ex: *wordcl = nn*
- **!=**
  inte lika med — ex: *wordcl != nn*
- **&**
  logiskt *och* mellan villkor — ex: *wordcl = nn & gender = utr*
- **|**
  logiskt *eller* mellan villkor — ex: *wordcl = nn | wordcl = pn*
- **:=**
  tilldelning — ex: X.form(gender := Y.gender)

UPPSALA UNIVERSITET

Operatorer för ersättningsförslag

- **join**
  X.join(Y.text) sätter ihop X och Y till ett ord
- **delete**
  X.delete() tar bort ordet X
- **replace**
  X.replace(Y.text) ersätter ordet X med ordet Y

Regelexempel 1

```
regel1@regler
{
X(text=”ett”),
Y(wordcl=nn & gender=utr)
-->
mark(X Y)
corr(X.form(gender:=Y.gender))
info(”inkongruens i nominalfrasen”)
action(scrutinizing)
}
```

Han köpte ett villa

ett villa inkongruens i nominalfrasen (regel1@regler) exempelregel

en villa

Regelexempel 2

```
regel2@regler
{
X(wordcl=nn & spec=ind & case=nom),
Y(wordcl=nn)
-->
mark(X Y)
corr(X.join(Y.text))
info(”särskriven sammansättning”)
action(scrutinizing)
}
```

Människans ingrepp döljs dock bättre vinter tid och kan rekomenderas till den känslige naturvännen.

vinter tid särskiven sammansättning (regel2@regler) exempelregel

vintertid

UPPSALA UNIVERSITET

Referenser

- Antti Arppe, 1999, *Developing a Grammar Checker for Swedish*
- Juhani Birn, 1999, *Detecting Grammar Errors with Lingsoft's Swedish Grammar Checker*
- Rickard Domeij, 2005, *Datorn granskar språket* (kapitel 6-7)
- Ola Knutsson, 2001, *Automatisk språkgranskning av svensk text*
- Theodorus Vosse, 1994, *The Word Connection: Grammar-based Spelling Error Correction in Dutch*