



UPPSALA
UNIVERSITET

Grundläggande Textanalys VT 2014

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



UPPSALA
UNIVERSITET

GRAMMATIKKONTROLL





UPPSALA
UNIVERSITET

Varför behövs grammatikkontroll?

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



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

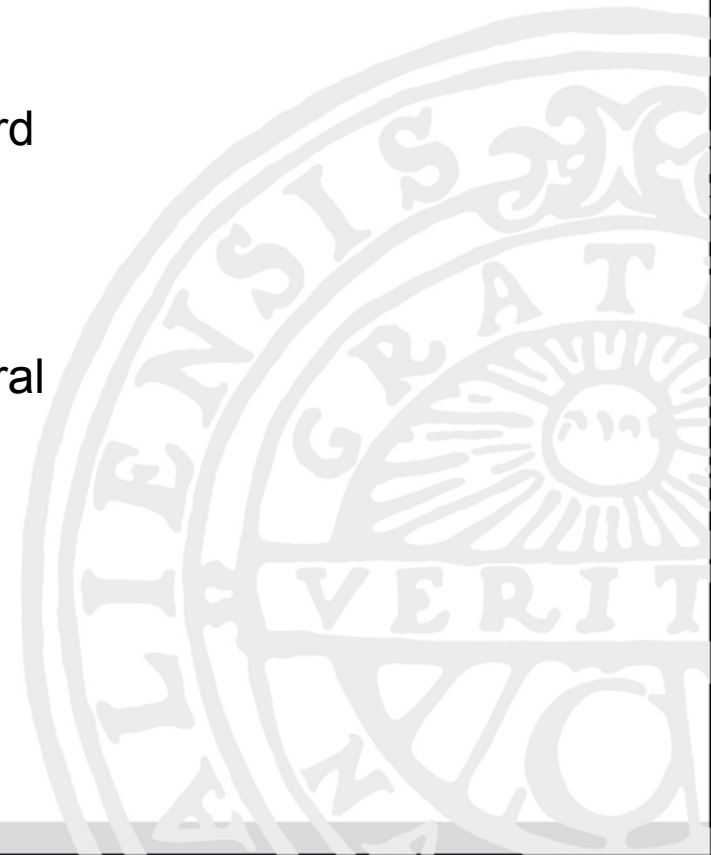
Realistiska förväntningar på grammatikkontrollprogrammet

- 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

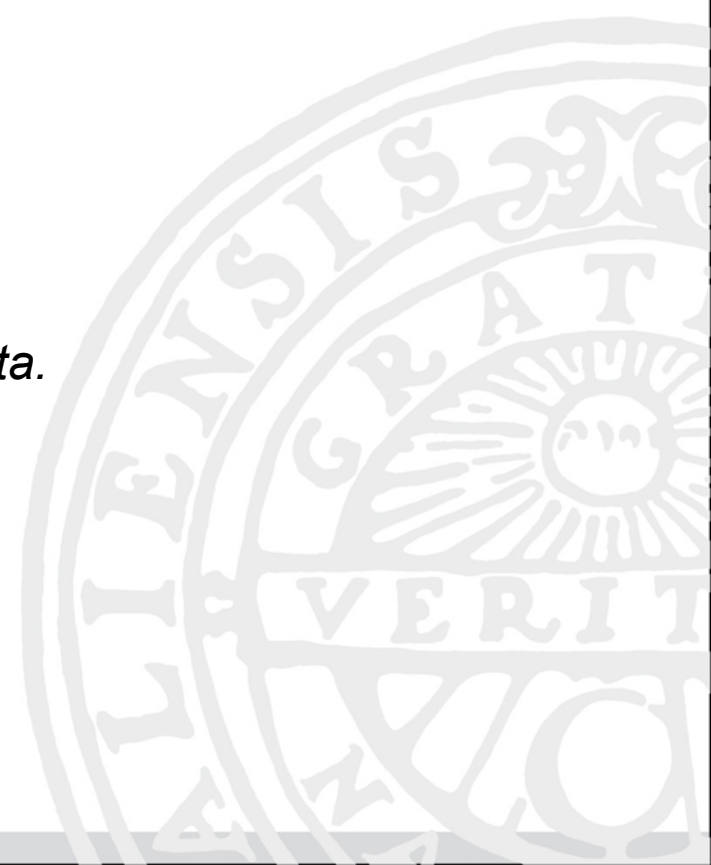
1. **Feldetektering** (*error detection*)
Fel i nominalfrasen: **det tidiga 1800-talen*
2. **Feligenkänning** (*error recognition*)
Numerusinkongruens determinerare-huvudord
3. **Feldiagnos** (*error diagnosis*)
Substantivet står i plural i stället för singular
Determineraren står i singular i stället för plural
4. **Felkorrigering** (*error correction*)
det tidiga 1800-talet
de tidiga 1800-talen





Problem inom grammatikkontroll

- 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





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

- Fullständig syntaktisk analys av meningen
- Relaxering för att hantera ogrammatiska strukturer
- Styrka 😊
Kan hantera strukturer där två av varandra oberoende ord inte står intill varandra
****Problemet** med den kommunala fördelningen från inkomstskatter för säsonganställda är **svår** att lösa*
- Svagheter ☹️
 - robusthet
 - effektivitet

Fenomenbaserad grammatikkontroll

- Partiell syntaktisk analys
- Styrkor 😊
 - robusthet
 - effektivitet
- Svaghet ☹️

Begränsat till felaktigheter mellan ord som står nära varandra, såsom otillåtna ordklasskombinationer

 - **kunnat kommit*
 - **en rött hus*



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

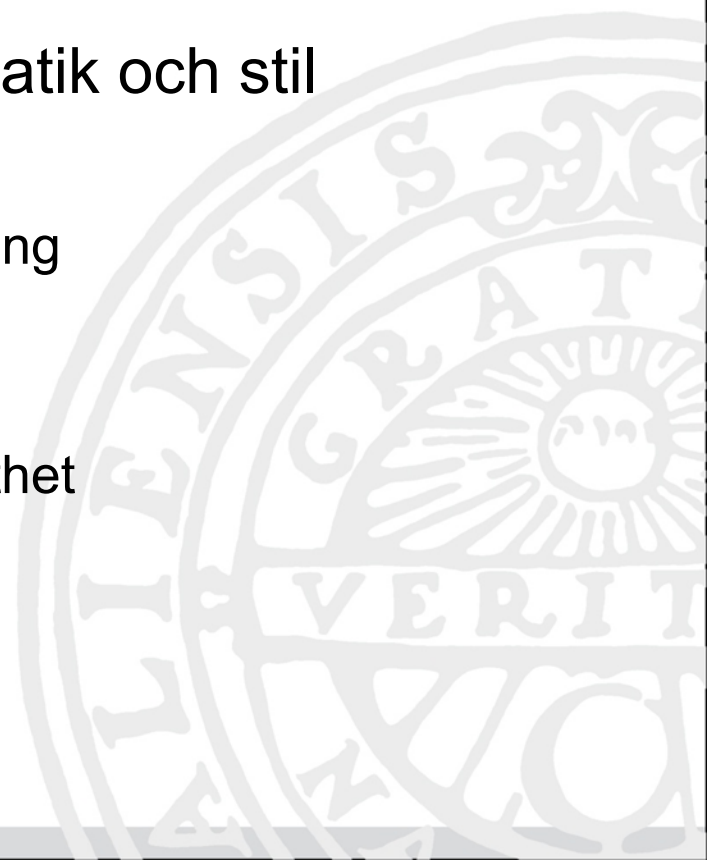
KONTROLLERAT SPRÅK





Kontrollerat språk

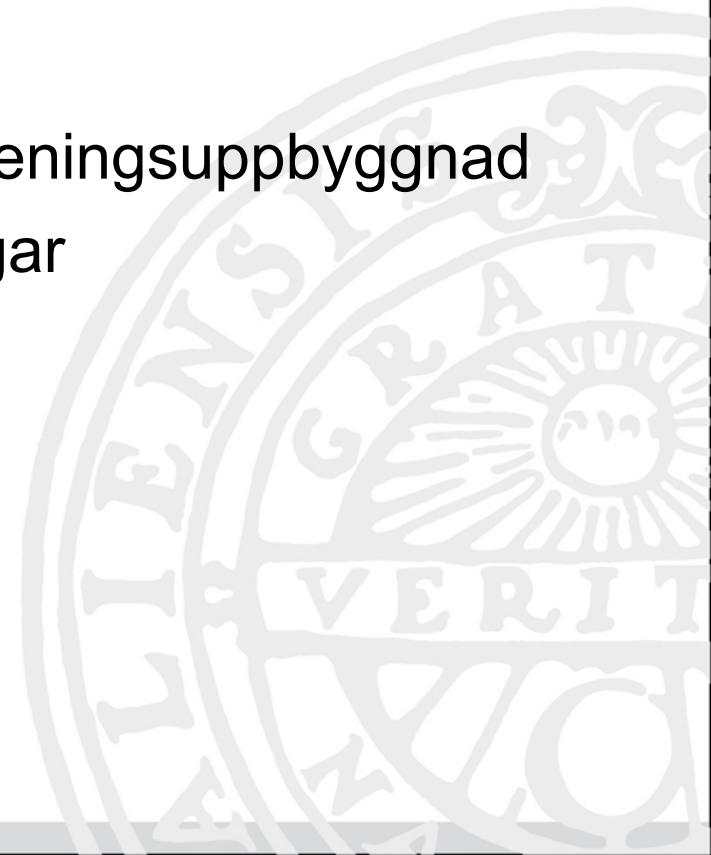
- Specialfall av stilkontroll
- Delmängd av naturligt språk
- Restriktioner på vokabulär, grammatik och stil
- Mål:
 - Enhetlig layout, terminologi och stavning
 - Ökad läsbarhet
 - Tydlighet (undvik tvetydighet)
 - Grammatisk och terminologisk korrekthet
 - Maskinöversättningsbar källtext





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
- Scania-svenska
 - Utvecklat av Uppsala universitet



Hur upplevs kontrollerat språk?

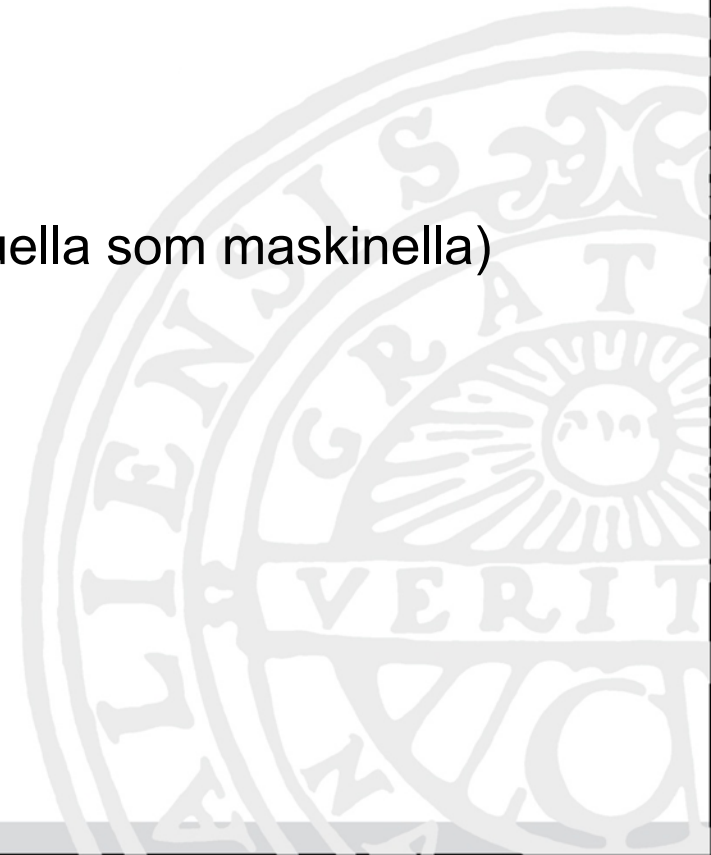
- Finns ett motstånd från användarna...
 - Krångligt och tidsödande att lära sig
 - Känner sig inte fria att skriva som de vill
 - Måste lära sig det kontrollerade språket
 - Måste lära sig den programvara som används för att se till att det kontrollerade språket följs
 - ”den tid man sparar på maskinöversättning 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

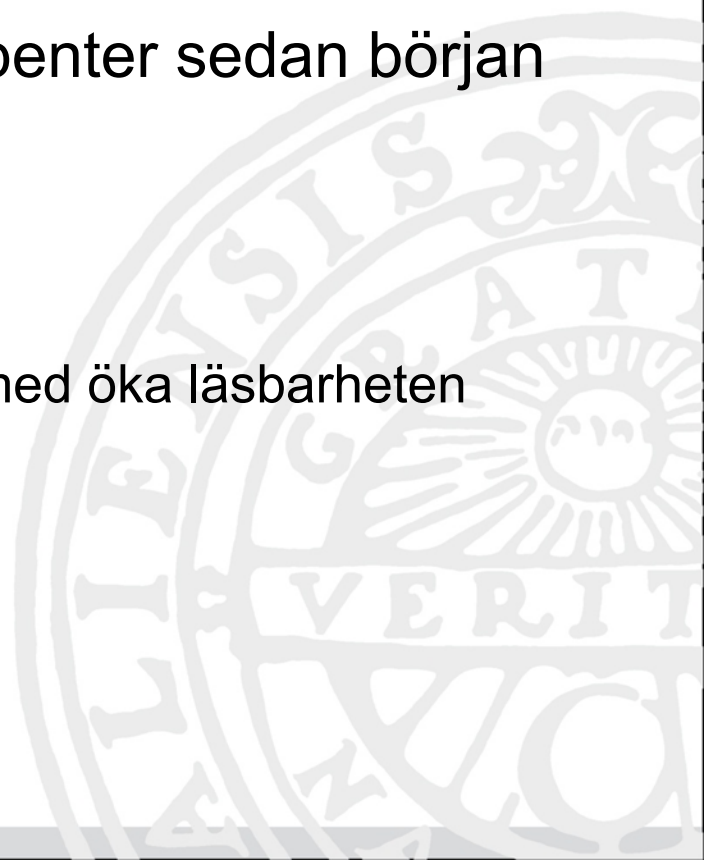
- Ordkontroll
 - Rätt terminologi
 - Rätt stavningsvariant (color/colour)
 - Akronymer (IKEA/Ikea)
- Grammatikkontroll
- Stilkontroll
 - Ord- och meningslängd
 - Datumformat
- Ersättningsförslag





Scania Checker

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



Färguppmärkning i Scania Checker

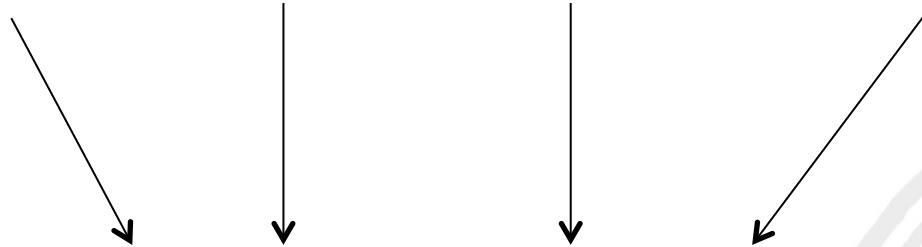
- Gulmarkering
 - Okända ord (saknas i lexikonet)
- Rödmarkering
 - Minusord med ersättningsförslag
- Grönmarkering
 - grammatikfel



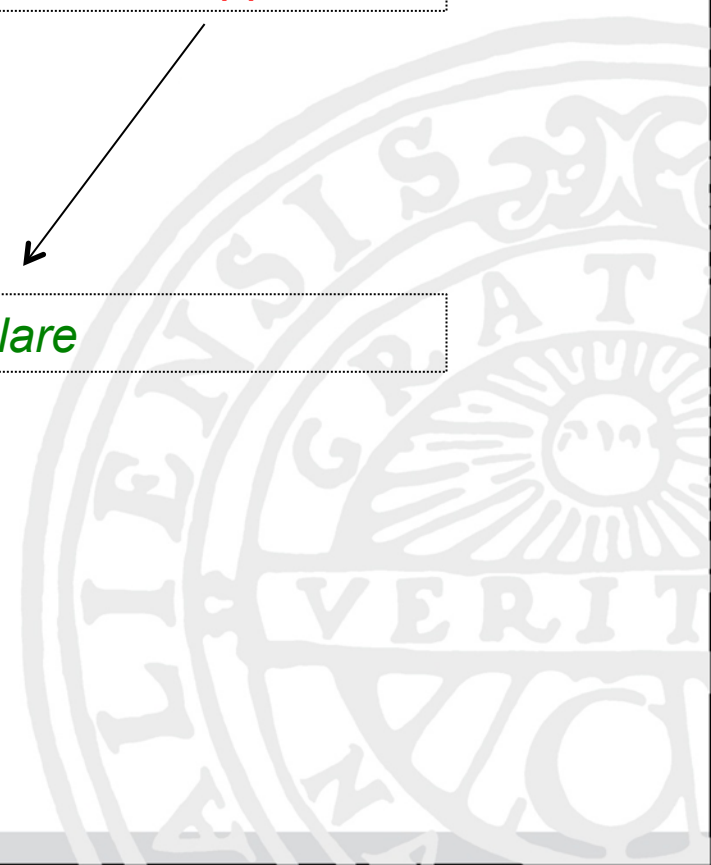


Minusord I Scania Checker

fotkontakt, golvkontakt, fotströmsträllare, fotomkopplare



Ersätt med: *golvströmställare*





UPPSALA
UNIVERSITET

SPRÅKGRANSKNINGSSYSTEM





UPPSALA
UNIVERSITET

Tidiga språkgranskningsystem

- Critique
- CORRie

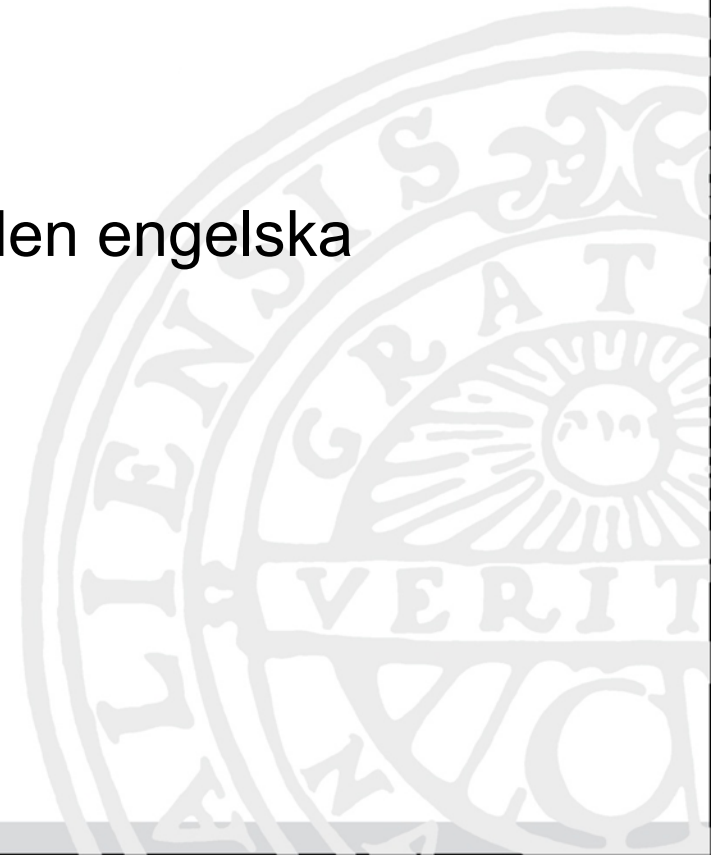




UPPSALA
UNIVERSITET

Critique

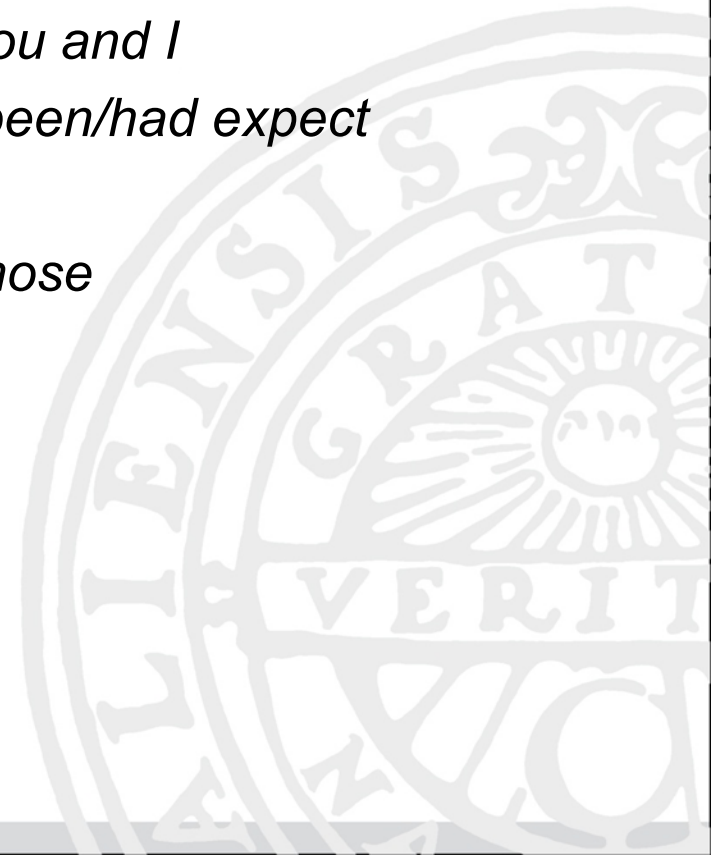
- Första praktiska grammatikbaserade språkkontrollsystemet (80-talet)
- IBM
- Engelska och franska
- Föregångare till språkkontrollen i den engelska versionen av Microsoft Word





Feltyper i Critique

- 25 grammatiska feltyper, fördelade på fem kategorier:
 1. numerusinkongruens **many book/he go*
 2. fel pronomensform **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
 2. borttagning
 3. substitution
 4. transposition
 5. feature mismatch

tåget har **har redan gått – hus bil*

**tåget redan gått*

the **well performance*

jag tror att tåget **går inte idag*

****den** lilla huset – männen är **glad***

Språkgranskningsystem för svenska

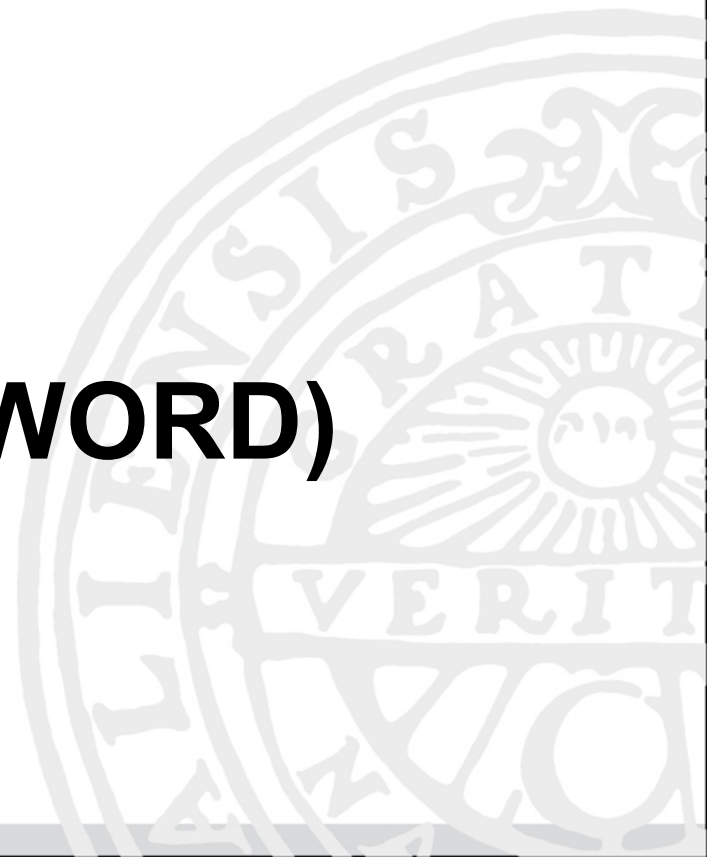
- Grammatifix, Lingsoft
- Granska, Nada, KTH
- SCARRIE, Uppsala universitet
 - Modifierad variant: Scania Checker
- Skribent, Norstedts Ordbok





UPPSALA
UNIVERSITET

GRAMMATIFIX (MS WORD)

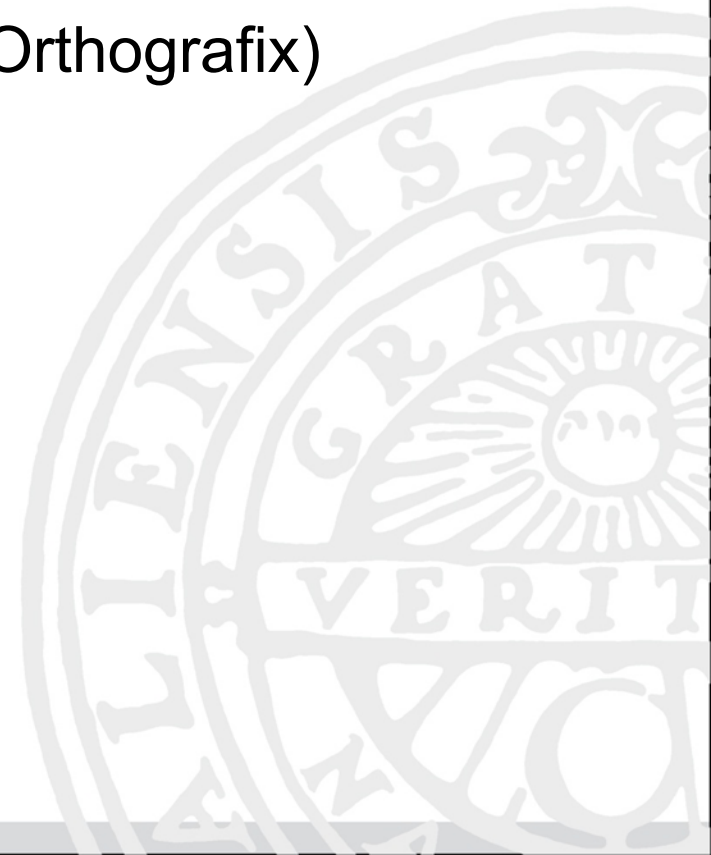




UPPSALA
UNIVERSITET

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

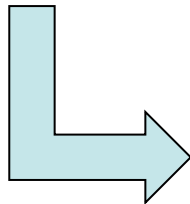




Steg 1: Förbehandling

- 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

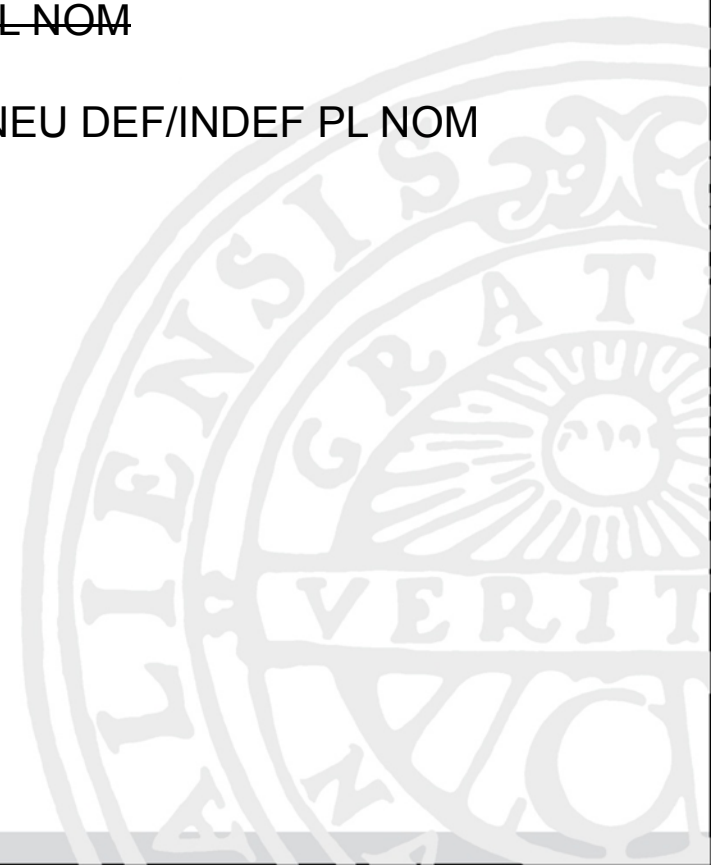
~~"dämpa" <PCP2> A NEU INDEF SG NOM~~

"<explosionen>"

"explosion" N UTR DEF SG NOM

"<\$.>"

"\$." CLB <PUNCT>



Steg 4: Morfosyntaktisk mappning

- Normalt I SWECG: syntaktiska funktionstaggar
- I Grammatifix: 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 *särsla, tillförne*
 - Brott mot skrivkonventioner (14)
 - Parenteser, datumformatering, interpunktion mm



UPPSALA
UNIVERSITET

GRANSKA





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



Fokus på tre feltyper

- Särskrivna sammansättningar
*Han orkade inte flytta **sten bumlingarna**.*
- Inkongruens i nominalfrasen
***ett** villa*
- Inkongruens i predikativ
*Äktenskapet är **baserad** på kärlek.*



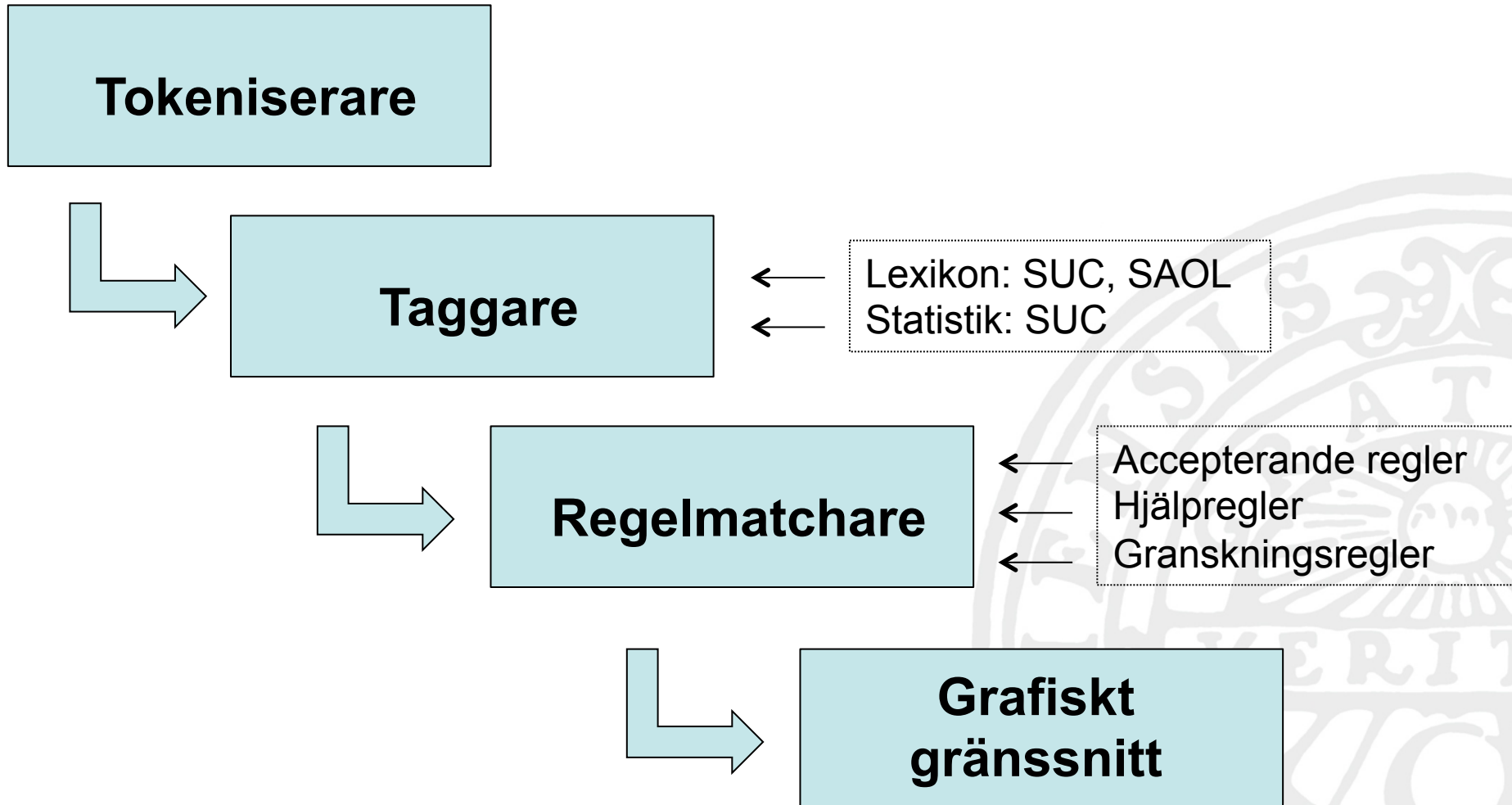


Fler feltyper

- 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*



Granskasystemet





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





Regelfilens struktur

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

```
regelnamn@regler
```

```
{
```

```
X()
```



Vänsterled (strängen man vill matcha)

```
-->
```

```
action(scrutinizing)
```



Högerled (hur man vill rätta strängen)

```
}
```



Några vanliga särdrag

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





Några vanliga särdragsvärden

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





Några logiska operatörer

- **=**
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)*



Operatörer 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: inkongruens i nominalfras

```
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

en villa

inkongruens i nominalfrasen (regel1@regler) exempelregel

Regelexempel 2: särskrivna sammansättningar

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

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

vinter tid särskrivna sammansättning (regel2@regler) exempelregel
vintertid



Labben

- Utvärdering av stavnings- och grammatikkontrollen i svenska Microsoft Word
- Testkörning och utvärdering av grammatikkontrollen i Granska
- Skriva egna regler i Granska (VG-uppgift)
- Deadline för labbrapport: fredag 23 maj till eva.pettersson@lingfil.uu.se



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*