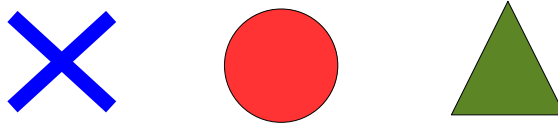
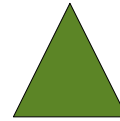
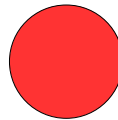
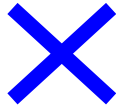


De sproglige ressourcer i OrdRet



Eckhard Bick
GrammarSoft

De sproglige ressourcer i OrdRet



Eckhard Bick
GrammarSoft

OrdRet

Grammatiske Søsterprojekter

- **VISL** (internetbaseret grammatikundervisning)
- **DanGram** (CG-tagger og -parser for dansk)
- **Korpus90/2000** (grammatisk opmærkede tekster)
- **URKAS** (Almen Sprogforståelse)

Hvorfor sproglige resurser, når man kunne nøjes med en liste?

- Brugeren har brug for en prioritering af rettelserforslagene
- Brugeren har brug for at få rettet også ellers korrektstavede ord

Ikke alle ord er lige hyppige <i>lage - læge - lagde</i>	frekvensbaserede korpusundersøgelser
Dansk tillader et uendeligt antal af sammensætninger - listen er altid for kort <i>efterlønstilhænger, kostkonsulent, ...</i>	morfologisk og kompositumsanalyse
Hvilke ord er mest lige? <i>straden - staden, straden - starten,</i> <i>straden - estraden, straden - stranden</i> <i>stdn / àè</i>	grafisk vs. fonetisk lighed bogstavudskiftninger vs. bogstavlighed stavelsesantal konsonatskelet og vokalskelet
Også ord, der ligner meget, kan være meningsløse i konteksten - <i>hun har købt en lille hæsð [hæst/hest]</i>	forkast rettelserforslag med (i konteksten) ”umulige” ordklasser
Også ellers korrektstavede ord kan være grammatisk forkerte (<i>-e/-er, -ene/-ende</i>)	tilføj kontekstuelle rettelserforslag også for velstavede ord
Listekontrollen accepterer ord i 2 dele (<i>'banegårds centret'</i>)	samkrivning foreslås ud fra bøjning og kontekst
Listekontrollen kan ikke rette noget der ikke er der (<i>sætningspunktum, manglende ord</i>)	en syntaktisk analyse kan hjælpe med at afgrænse sproglige enheder/sætninger

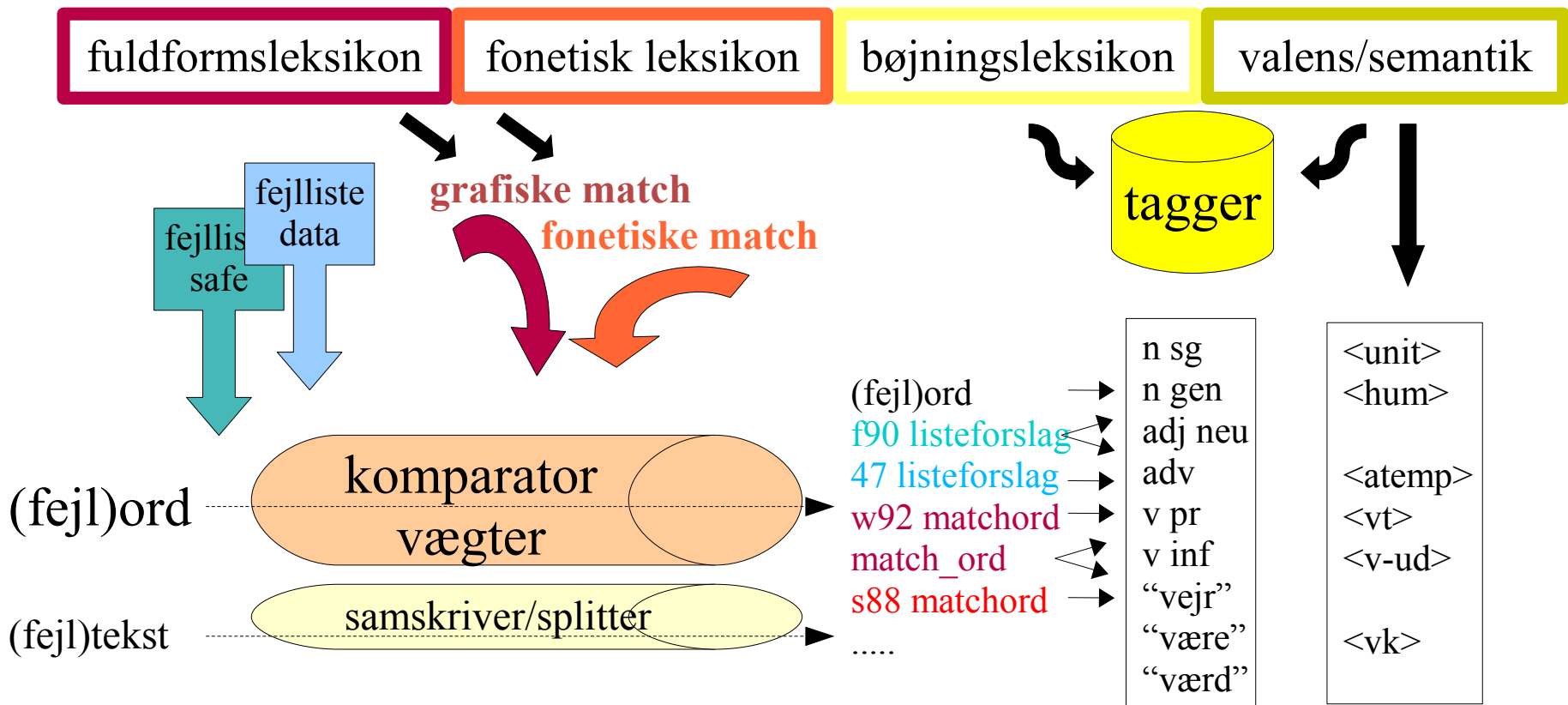
CG & stavekontrol

- Swedish: Grammatifix/Svefix/SWECG (<http://www.lingsoft.fi/svefix/>)
Arppe, A. 2000. Developing a grammar checker for Swedish. In Nordgård, T. (ed.) *Nodalida '99 Proceedings from the 12th Nordiske datalingvistikkdager*, Department of Linguistics, University of Trondheim, p. 13-27.

Birn, J. 2000. Detecting grammar errors with Lingsoft's Swedish grammar checker. In Nordgård, T. (ed.) *Nodalida '99 Proceedings from the 12th Nordiske datalingvistikkdager*, Department of Linguistics, University of Trondheim, p. 28-40.
- Norwegian:
Hagen, K., Johannesen, J.B. & Lane, P.: “Some problems related to the development of a grammar checker” (http://www.hf.uio.no/tekstlab/prosjekter/NoDaLiDa_gram.html), foredrag på NoDaLiDa 2001
Hagen, K., Lane, P. & Trosterud, T.: “En grammatikkontrol for bokmål”, i Kjell Ivar Vannebo og Helge Sandøy (eds.): *Språkknyt* 3-2001
- Orthograpix (<http://www.lingsoft.fi/orthograpix/>) exists for the above plus Danish, Finnish and Suahili, but is a spell checker without grammar/CG

OrdRets grammatiske ressourcer i tal

- Sprogteknologisk leksikon (140.000 leksemer)
bøjningsklasser, valens, semantik ...
- Morfologisk analysemaskine
ekspliciterer flertydighed
- Morfologisk Constraint Grammar (2.550 regler)
- Fejltype CG m.m. (800 regler)
- Korpusbaseret frekvensliste (1.050.000 ord)
- Fonetisk leksikon og transskriptionsmaskine
- Fejl databaser (13.300 + 5.100 former)



DanGram i OrdRet 1

CG mapping-regler

CG remove/select

bøjningsleksikon

kontekst

fejlmapper
unsafe

disambiguator
DanGram CG

fejlmapper
safe

Ret-1 Analyse 1a
Analyse 1b
Ret-2 Analyse 2
Ret-3 Analyse 3a
Analyse 3b
Ret-4 Analyse 4
Ret-5 Analyse 5a
Analyse 5b
Analyse 5c

@majuskel
@comp
@headstop
@:forslag

Analyse 1ab
Analyse 2
Analyse 3ab
Analyse 5ae
Analyse 5b
Analyse 4

@vfin
@inf
@pl
@neu

gene-
rator

Ret-5b
bøjet

1. forslag

2. forslag
3. forslag
4. forslag
.....

► komparator-vægtede restforslag

DanGram i OrdRet 2

OrdRet

Eksempler

- Hun har en opfattelse af at **kvinde** (@pl) er bedre til det **merster** (R:meste).
(ingen indefinitte entals-substantiver uden prænominaler, - undtagen <mass>)
- Han kan ikke **hører** (@inf) dig.
(hjælpeverbumskontekst)
- Han ønsker ikke **og** (@:at) forstyrre.
(verbum med infinitivvalens til venstre, infinitiv til højre)
- Min søster er **syge plejerske** (@comp).
- Hun besøgte **barndoms** (@comp-) veninden.
(ubestemt substantiv i genitiv ental før et substantiv i bestemt form)
- Glasset var **fuld** (@sc-neu).
(konkordanskrav mellem subjekt og subjektsprædikativ)
- Jeg er **træt** (@headstop) Det har **vært** (R:været) en lang dag.
(Majuskel og/eller syntaktiske tegn på ny sætning)
('været' vinder over 'vært' efter hjælpeverbum)



