KTHs morfologiska och lexikografiska verktyg och resurser

Viggo Kann
professor i datalogi vid KTH
Leksikografi og språkteknologi i Norden, januari 2010

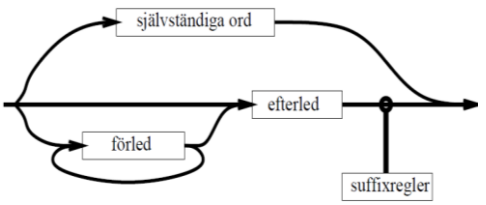
Språkteknologigruppen på KTH

Filosofi
- utveckla effektiva och resurssnåla metoder för språktekniska system
- i synnerhet för svensk text
- fritt tillgängligt om möjligt

Huvudområden
- Svensk språkgranskning
- Informationssökning och -extraktion
- Ordböcker

Stava - svensk stavningskontroll [Domeij, Hollman, Kann, Tillenius]

- Ordbas: bygger på SAOL 11
- Uppdelad i förled, efterled, självständiga ord



Stavas rättstavning

- Rangordnade rättelseförslag med hjälp av felstavningsmetrik och ordfrekvenser
- 60% av de felstavade orden får korrekt förstahandsrättelseförslag

Tillgänglighet

- *Källkod (i C) och kodade orddatabaser fritt tillgängliga*
- Stava kan köras på webben: http://www.nada.kth.se/stava

Utvidgningar av Stava [Kann]

- 1500 suffixregler märkta med lemma-form och ordklasstagg enligt SUC
- Taggstava– kan tagga alla regelbundet böjda ords alla ordformer, även sammansättningar
- Lemmatisering av dessa ordformer

Sammansättningsanalysator [Sjöbergh, Kann]

- Troligaste sammansättningspunkterna tas fram med en statistisk kombinations-metod som bygger på tre delar:
  1. antalet ordled mun-vinklarna mun-vin-klarna
  2. ordledsfrekvenser upp-rättar upprätt-ar
  3. ordledsordklasser upp-rättar upprätt-ar
- *Källkod (i C/C++) fritt tillgänglig*

Granska – svensk grammatik-kontroll

- http://skrutten.nada.kth.se
- Granskatagger – ordklasstaggare med lemmatiserare och ordböjare
  källkod (C++) och lexikon fritt tillgängliga
  [Carlberger, Kann]
- GTA - Grammatikkontroll och menings-chunkare med grammatikregelspråk
  [Bigert, Kann, Knutsson]

Grim – en interaktiv lärmiljö [Westlund, Knutsson, Sjöbergh m fl]

- http://skrutten.nada.kth.se/grim
- Liten ordbehandlare med stavnings-kontroll, regelbaserad och probabilistisk grammatikkontroll, presentation av ordklasser, sökning i lexikon mm

Lexin [Språkrådet, Kann]

- http://lexin.nada.kth.se
- 30000 svenska ord översatta till 15 olika språk: albanska, arabiska, bosniska, engelska, finska, grekiska, kroatiska, nordkurdiska, persiska, ryska, serbiska, somaliska, spanska, sydkurdiska, turkiska
- 20 miljoner uppslagningar i månaden – massor av användarstatistik samlas

Skandinavisk ordbok [Nordiska språksekretariatet, Kann]

- www.nada.kth.se/skandlexikon
- 10000 ord som skiljer mellan danska, norska och svensk

Tvärslå [projektet Nordisk nätordbok]

- http://ordbok.nada.kth.se/
- Söker i mängder av ordböcker mellan de nordiska språken och engelska.
- Ordböckerna insamlade i projektet Nordisk nätordbok och kodade i samma XML-format

Vad är en *fri* språkresurs?

- Anyone can use it in an application
- Anyone can study it and modify it
- Anyone can take a copy of it
- Anyone can improve it, release the improvements to the public, so that the whole community benefits

(baserat på Four freedoms of free software, Richard Stallman)

Typiska sätt att konstruera en resurs

…om du är en språkteknolog:

- Skaffa finansiering
- Använd resurser som är tillgängliga för forskare
- Anställ lexikografer som kan göra det stora jobbet

…om du är en fri-programvaruhacker:

- Använd andra fria resurser
- Samla data från massor av människor, t ex med wiki eller webbformulär

Folkets synonymlexikon [Kann]

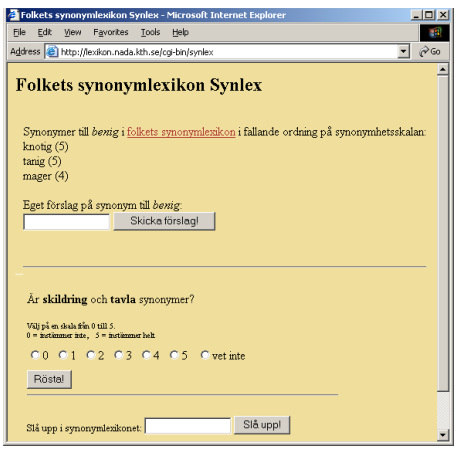
- Skapa ett svenskt synonymlexikon som en lista av synonyma ordpar.
- Jag är lat och vill inte jobba så mycket.
- Jag är snål och vill inte anställa någon.
- Det konstruerade synonymlexikonet ska bli en fri språkresurs.

Idéer

- Konstruera automatiskt en massa ordpar som kan vara synonymer.
- Använd tiotusentals människor som var och en är villig att bidra en smula utan betalning, genom att kontrollera ordpar. Använd Lexins webbplats!

Konstruktionsmetod

1. Konstruera möjliga synonympar.
2. Rensa synonymparslistan automatiskt.
3. Fråga massor av användare om paren är bra synonymer.
4. Analysera användarnas bedömningar och bestäm vilka par som behålls.



Lite statistik (januari 2010)

- 3,6 M bedömningar har gjorts
- 80 000 ordpar (bedömda ≥ 2) i lexikonet
- 123 000 användarföreslagna ordpar
- 74 000 olika användarordpar
- 24 000 av dom har accepterats
- Synonymlexikonet är fritt och laddas ner från http://lexin.nada.kth.se/synlex.html

Exempel: Synonymer till *klass*

5: rang
rank
slag

4: kategori
stånd
årskurs

3: fack
grad
grupp
kvalitet
nivå

3: sort
standard
stil

2: skikt
storleksordning
typ

1: poäng
stadga

0: uppdrag
utbilda

Hur undviks missbruk?

- Många bedömningar krävs innan ett ordpar anses vara bra.
- Ordparen som ska föreslås väljs slumpmässigt från en enorm lista.
- Ordpar som föreslås av användarna stavningskontrolleras innan dom läggs till den enorma listan.

Folkets definition av synonymitet

- Exakta betydelsen av *'synonym'* definierades inte.
- Användarna bedömer efter sin intuitiva bild av konceptet synonymitet.
- Det skapade lexikonet bygger på folkets egen definition av synonymitet, vilket förhoppningsvis är precis vad folket vill!

Folkets lexikon [Hollman, Kann]

- http://folkets-lexikon.csc.kth.se
- Pågående projekt stött av .se-stiftelsen
- Bygger på svensk-engelska Lexin
- Automatiskt framtagna översättnings-förslag bedöms av användarna
- Användarna ska själva få utvidga lexikonet och ladda ner det
- Utvidgas också med t ex Saldo

Planerat innehåll i lexikonet

- uppslagsord på svenska och engelska
- ordklass, uttal, böjningsformer
- synonymer, andra relationer
- översättningar
- definition, förklaring
- exempel, idiom, sammansättningar
- externa länkar (Wikipedia, dataterm etc)

Avstavningslexikon
[Språkrådet, Kann, Knutsson]

- Avstavningslexikon bestående av 38000 avstavade ord med böjningsformer
- Pågående projekt