GÖTEBORGS UNIVERSITET
INST FÖR SPRÅK OCH LITTERATURER

Kort presentation av Korp, Sveriges nationalkorpus

Morgan Nilsson
Göteborgs universitet

19 januari 2017 vid
Avdelningen för nordiska språk, L'Orientale-universitetet i Neapel

Morgan Nilsson

- Fil.dr. i slaviska språk.
- Undervisat i slovenska, tjeckiska, polska, och ryska till 2014.
- Tidvis undervisat i introd. till språkvetenskap & allmän grammatik.
- Undervisar och forskar i somaliska sedan 2014.

- Kompendium om allmän grammatik. pdf
- Föreläsningar om allmän grammatik. YouTube

Språkbanken

- Självständig forskningenhet vid institutionen för svenska språket, Göteborgs universitet.
- Ska samla in, systematisera och ge tillgång till språkresurser åt forskare och allmänheten. De ska också bedriva egen forskning och utveckling som är till nytta för sådan användning.
- Ansvarar för den svenska nationalkorpusen KORP.

Resurser vid Språkbanken

Bland annat

- KORP – verktyg för sökning i Språkbankens textkorpusar
- KARP – lexikala infrastruktur och sökverktyg
- LÄRKA – för att lära grammatik genom korpusanalys
- SALDO – Semantiskt och morfologiskt lexikon för språkteknologi

Produkter från Lexikaliska institutet

- Svenska Akademiens Ordlista
  ny upplaga ungefär vart 10:e år
- Svensk Ordbok
  på uppdrag av Svenska Akademien.
- Svenska Lexin
  på uppdrag av Skolöverstyrelsen
- ISLEX
  från isländska till sv, nno, nob, dk, fö

Andra verk från Göteborgs universitet

- Nationalencyklopedins Ordbok
  på uppdrag av Nationalencyklopedins förlag.
- Digitalisering på uppdrag av Svenska Akademien av
  - Svenska Akademiens Ordlista
    alla äldre upplagor
  - Svenska Akademiens Ordbok
    banden A-Vedersyn
  - äldre version av SAOB, banden A-Tövla

Korp

- Korp innehåller texter med 12 463 068 634 löpord

- Lilla Korpskolan:
  en kort användarhandledning som korta videoklipp.

Olika typer av ord

- löpord = token varje enskilt ord i en text
- typord = type alla unika ord i en text
- lemma alla böjningsformer av samma ord
- lexem en specifik betydelse hos ett lemma

Välja texter i Korp

- Välja korpusar (förvalt är delar av de moderna korpusarna)
  - det finns ett antal olika samlingar och subkorpusar i varje samling,
  - i korpusväljaren syns subkorpusarnas storlek,
  - ovanför finns en tidslinje för de valda texterna,

Söka i Korp

- enkel, utökad eller avancerad sökning

- Enkel sökning på
  - ord, fras eller lemgram
  - förled eller efterled för ord = sammansättningar för lemgram
  - beroende på skiftläge (lemgram är alltid oberoende)

Sökresultat: exempel

- KWIC = keyword in context.
- Antal träffar.
- Lista med samtliga träffar. Vanligtvis visas en mening per träff.
- Träffarna är grupperade efter korpus, och vilken korpus de efterföljande träffarna kommer från står skrivet med liten text ovanför.
- Vissa korpusar ger möjlighet att välja större kontext.
- Om träffarna är många måste man bläddra mellan sidor.
- Man kan bläddra med tangenterna F och N.
- En färgad remsa visar mängden träffar i de olika korpusarna.
- Om man för musen över remsan visar korpusarnas namn.
- Man kan klicka på önskad korpus för att komma till just de träffarna.

Enkel sökning (övningar)

- Sök alla former av ordet **fågel** i **Skönlitteratur**.
- Ta dig till träffarna från korpusen **Äldre svenska romaner**.

Sökresultat: Sidopanel

- När man klickar på ett ord I ett exempel visas till höger en sidopanel.
- Den innehåller information
  - om det markerade ordet (**Ordattribut**: ordklass, grundform, sammansättningsanalys med mera) och
  - om den text som ordet ingår i (**Textattribut**: författare, utgivninsår och liknande).
- Vissa attribut är klickbara.
- Lemgram = en ny sökning på det lemgrammet.
- Saldo = en ny flik öppnas i Karps sökgränssnitt för lexikala resurser.

Sökresultat: visningsalternativ

- För KWIC
  - antalet träffar per sida,
  - sorteringsordning
    - efter höger- eller vänsterkontext, eller på själva träffen i sig.

    Sorteringen sker enbart inom varje korpus.

- För statistiken

    vilket attribut statistiken ska sammanställas på.

Sökresultat: Statistik

- Tabell med en kolumn för varje korpus och en radför varje form.
- Bland visningsalternativen kan man välja ansra sätt att sammanställa statistiken, till exempel ordklass eller något textattribut.
- Man kan även välja om sammanställningen ska vara skiftlägesberoende eller ej.
- Tabellens visar antalet förekomster i absoluta och relativa tal.
- De relativa talen visar antal träffar per en miljon löpord.
- Man kan sortera varje kolumn i stigande eller fallande ordning.
- En liten ikon öppnar ett cirkeldiagram med fördelningen av träffarna mellan de olika korpusarna. Man kan välja absoluta eller relativa tal.
- Längst ner på sidan finns möjligheten att exportera statistiktabellen.

Statistik (övningar)

- Sök på ordet **nåt**.
- Testa att sortera resultaten efter olika kolumner.
- Klicka fram några cirkeldiagram.

- Välj att söka i **Svenska partiprogram och valmanifest** 1887-2010.
- Välja att sammanställa på på **parti** eller **år**.
- Sök på något lemgram (t.ex. frihet, jämlikhet).
- Kan man se några skillnader?

Sökresultat: Trenddiagram

- Trenddiagrammet utgår från rader i statistiktabellen.
- Det visar dessa raders relativa frekvens över tid som antalet träffar per en miljon löpord för varje tidsenhet.
- Först väljer man en eller flera rader med kryssrutorna till vänster.
- Därefter klickar man på Visa trenddiagram.
- Till höger går det att välja vilka linjer man vill visa.
- Om man klickar på linjen visas alla träffar för just den tidpunkten.
- Under diagrammet finns en mindre version av diagrammet.
- Där kan man välja att zooma in i det stora diagrammet.

Trenddiagram (övningar)

- Välj att söka i **Svenska partiprogram och valmanifest** 1887-2010.
- Sök i utökad sökning på lemgram är **fred eller** lemgram är **krig**.
- Välj ut dessa rader i statistiken och klicka på **Visa trenddiagram**.
- Kan man se någon trend?

Sökresultat: Ordbild

- Ordbild måste aktiveras före sökningen.
- Den fungerar bara för Enkel sökning av ett ord eller ett lemgram.
- Ordbild visa vilka andra ord som har syntaktiska relationer till det sökta ordet.
- För verb visas de vanligaste subjekten och objekten.
- För substantiv de vanligaste attributen, de verb som substantivet är subjekt och objekt.
- 15 ord visas för varje grammtisk relation
- Siffran abger hur många gånger relationen finns i de valda texterna.
- Ikonen vid varje ord leder en lista med träffar där vald relation förekommer.

Ordbild (övning)

- Välj GP-korpusarna.
- Vad gör en hund ofta och vad gör man ofta med en hund?
- Ta fram alla meningar där hundar skäller.

- Vilka är de vanligaste objekten till verben äta och dricka i Strindbergs texter och i Göteborgsposten?

Sökresultat: Karta

- Kartfunktionen måste aktiveras före sökningen.
- För Bloggmix 1998-2004 och 2014-2015 visas en karta baserad på bloggarens hemort.
- För Press 95-98 och GP 2012-2013 visas en karta baserad på platser som förekommer i samma stycke som träffarna.
- Om man på kartan klickar på en ort visas de träffar som har anknytning till orten.

Karta (övning)

- Välj Bloggmix och gå till **Utökad sökning**.
- Sök på **lemgram är palt** eller **lemgram är kroppkaka**.
- Sammanställ på **lemgram**.
- Markera i statstikfliken alla rader och klicka på **Visa karta**.
- Hur fördelar sig träffarna?
- Klicka på en plats på kartan för att se träffar från den platsen.

- Försök att hitta fenomen som borde vara vanligare i t.ex. Göteborg än i Stockholm.

Sökresultat: Jämförelse

- För en jämförelse behöver man först spara två sökningar.
- Detta gör man genom att klicka på pilen till höger om Sök-knappen.
- Då sparas sökningen i stället för att utföras.
- När man har sparat två sökningar kan man välja Jämförelse.
- Man väljer vilket attribut som ska jämföras.
- Ett exempel på två sökningar:
- alla substantiv i romaner
- alla substantiv i nyhetstexter.
- Jämförelsen presenterar en kolumn med de träffar som är mest utmärkande för sökning #1 och en kolumn för sökning #2.
- Siffrorna visar absolut frekvens.

Jämförelse (övning 1)

- Välj att söka i **Svenska partiprogram och valmanifest** 1887-2010.
- Mata in din sökning på lemgrammet **frihet**.
- Istället för att söka ska du spara sökningen genom att trycka på den lilla pilen till höger om Sök-knappen.
- Gör samma sak för **jämlikhet**.
- Dina sökfrågor kommer sedan att finnas tillgängliga i Jämförelse-fliken.
- Gå dit och utför jämförelsen.

Jämförelse (övning 2)

- Välja Bloggmix och GP-korpusarna.
- Sök på ett ord som borde vara vanligare i tidningar är på bloggar.
- Se om du kan verifiera din idé.
- Sök gärna på ditt ord som lemgram och välj även för- och efterled för att få fler träffar.

Utökad sökning

- Tillåter mer avancerade sökningar.
- Varje box motsvarar ett löpord.
- För varje löpord kan man ge flera kriterier.
- För fler boxar använder man +-knappen till höger.
- Man kan byta ordning genom att dra en box med musen.
- Man väljer vilket attribut man vill söka på, t.ex. ord, ordklass, lemgram.
- Till höger en lista med villkor.
- Symbolen "Aa“ stänga av skiftlägesberoende för det aktuella fältet.
- Skiftlägesoberoende sökning går betydligt långsammare.
- Om man väljer "ord" och lämnar textfältet tomt, så motsvarar det vilket ord som helst.

Utökad sökning (forts.)

- Upprepning, meningsbörjan och meningsslut
- I varje box finns det en knapp med ett kugghjul som visar en meny.
- Upprepa möjliggör att ordet upprepas ett visst antal gånger.
- Till exempel “Vilket ord som helst” och 1 till 3 gånger ger en sökning på ord på olika avstånd till varandra.
- Meningsbörjan ger sökning på ett ord bara först i meningar.
- Meningsslut ger sökning på ett ord bara sist i meningar.

Utökad sökning (övningar)

- GP-korpusarna: sök efter minst två adjektiv följt av ordet **katt**.

Betydelsedisambiguering

- När man söker på vissa ord som har många betydelser får man fler träffar än man önskar.
- Genom att söka på en specifik betydelse kan man få träffar som är mer relvanta.
- Sådan betydelsedisambiguering finns i
- Press 95-98
- GP 2012-2013
- Sociala medier, Bloggmix 1998-2004 och 2014-2015.

Betydelsedisambiguering (övning)

- Sök på ordet **fil**.
- Sammanställ på betydelse.
- När man klickar på ett ord i ett exempel kan man se ordets betydelseannotering i sidopanelen.
- Om man klickar på visa fler får man se alla alternativ tillsammans med deras sannolikheter.
- Kolla även statistikfliken där man kan se frekvens för de olika betydelserna.
- Gå till **Utökad sökning**. Välj **betydelse**.
- Skriv **fil** och det kommer en meny där du kan välja vilken betydelse du vill söka på.

Parallellkorpusar

- Innehåller samma text på två språk och är länkade på meningsnivå.
- Sökresultatet innehåller par av meningar, en för varje språk.
- För väljer man "Parallella" högst upp på sidan.
- Parallellsökning kan bara göras som Utökad sökning.
- Man anger vilken språkversion man vill söka i en särskild språkmeny.
- Det går att söka i båda språken genom att trycka på "Fler språk“
- Då måste sökkriterierna uppfyllas i båda språken samtidigt.
- Man kan till exempel välja bort en viss översätting genom att söka efter träffar där svenskan måste innehålla “farbror", medan man för italienska kryssar i rutan "Innehåller inte" och anger ordet “zio".

Fördjupning

- Användarhandledning för Korp
- Kursmaterial: Corpus methods in linguistics

Andra språk

- Finska Språkbanken: Kielipankki
- Indgangen til det danske sprog – sproget.dk

Kuriosa

- Liten finsk talspråkskorpus med video och audio