Korp

Övningar
Språkbankens höstworkshop 2016

https://spraakbanken.gu.se/korplabb/
sb-korp@svenska.gu.se

17 oktober 2016

ÖVERSIKT

Korp är Språkbankens korpusverktyg och en väsentlig del av vår korpusinfrastruktur. Med hjälp av Korp kan man utforska flera hundra olika korpusar i både modern och äldre svenska och även i några andra språk.

Om du har använt Korp tidigare - hoppa direkt till övning 6 för att testa ny funktionalitet.

ATT SÖKA I KORP

För att göra relevanta sökningar så måste man först välja de korpusar som man vill söka i. Längst upp till vänster ser man att man först kommer in i Korps moderna mode. Där hittar vi en mängd olika typer av material, t.ex. sociala medier, skönlitteratur, Wikipedia och myndighetstexter. För att komma åt äldre material eller material på andra språk så får man välja ett annat mode, t.ex. Kubhist, som innehåller äldre tidningstexter. När man har valt mode så kan man välja ut vilka korpusar man ska söka i med hjälp av korpusväljaren.

I Korp finns tre olika typer av sökning: Enkel, utökad och avancerad. I övningsuppgifterna kommer vi endast att använda enkel och utökad sökning. Det finns också tre resultatlägen: KWIC, statistik och ordbild. Från statistikvyn kan man för vissa material få fram ett trenddiagram och en karta.

För mer funktioner se användarmanualen:
`https://spraakbanken.gu.se/eng/research/infrastructure/korp/user-manual`

ÖVNING 1 - ENKEL SÖKNING OCH STATISTIK

I enkel sökning kan man antingen söka på ett ord, fraser eller ett lemgram. När man söker på lemgram så söker man på alla ordformer som ingår i lemgrammets böjningstabell. När du matar in ett ord så kommer det upp en meny där du kan välja vilket lemgram du vill söka på.

1.1 Gör en sökning på alla ordformer av ordet **fågel** i kategorin **Skönlitteratur**.

1.2 Försök att enkelt ta dig till träffsidan där träffarna från korpusen **Äldre svenska romaner** börjar.

1.3 Bekanta dig med statistikfliken. Testa att sortera resultaten efter olika kolumner och klicka fram något cirkeldiagram.

ÖVNING 2 - UTÖKAD SÖKNING

I den utökade sökningen kan man skapa avancerade sökuttryck. Man kan söka på alla attribut som stödjs av de valda korpusarna. Nästan alla korpusar stödjer sökning på ordklass, grundform, lemgram, betydelse, förled och efterled. Vissa korpusar har också textattribut, som titel, författare eller när texten är skriven.

En låda representerar ett token i texten (t.ex. ett ord eller en punkt). För varje token kan man sätta flera krav. T.ex. **ordform är blomma och/eller ordklass är substantiv**. Man kan lägga till fler lådor för att göra sökningar som **alla substantiv följt av verbet gå**. Det går också att säga att ett token ska upprepas flera gånger om man vill hitta t.ex. flera adjektiv följt på varandra. Mer detaljerade instruktioner finns i användarmanualen.

2.1 I GP-korpusarna, försök hitta alla träffar där det är minst två adjektiv följt av ordet **katt**.

ÖVNING 3 - UTÖKAD SÖKNING OCH SAMMANSTÄLLNING I STATISTIKEN

I statistiktabellen kan man välja att sammanställa på alla attribut som de valda korpusarna stödjer. **Svenska partiprogram och valmanifest** innehåller valmanifest från 1887-2010. Genom att välja att sammanställa på på t.ex. parti så får man en rad per parti vars valmanifest nämner det som du sökt på. Man får även totala antalet träffar och relativt antal träffar per parti, där relativa antalet betyder antalet träffar per en miljon token.

3.1 Sök på något lemgram (t.ex. frihet, jämlikhet) och sammanställ på parti eller år. Se om du kan se några skillnader på hur mycket olika partier använder uttrycken eller under vilka år de var vanliga.

3.2 När man sammanställer på år så får man ofta väldigt många rader i tabellen. Det kan vara svårt att dra några slutsatser över tid. I trenddiagrammet kan man jämföra rader i tabellen över tid. Sök istället i utökad sökning på lemgram är **fred eller** lemgram är **krig**. Genom att sammanställa på lemgram får vi en rad för fred och en för krig. Välj ut dessa rader och klicka på **Visa trenddiagram**. Kan man se någon trend?

I **Svenska partiprogram och valmanifest** finns texterna uppmärkta med år, men trenddiagrammet stödjer även sökningar ner på minutnivå. Det kan man se i Twitter-korpusarna.

3.2 Det visade sig också vara ganska krångligt att jämföra partiernas användning av uttrycken **frihet** och **jämlikhet**. Korp har en jämförelsefunktion som lämpar sig väl för den typen av frågor. Börja med att mata in din sökning på lemgrammet **frihet**. Istället för att söka kan du spara sökningen genom att trycka på den lilla pilen till höger på Sök-knappen. Gör samma sak för **jämlikhet**. Dina sökfrågor kommer då att finnas tillgängliga i Jämförelse-fliken. När du går dit kan du utföra jämförelsen.

ÖVNING 4 - ORDBILDEN

I ordbilden visas det sökta ordet tillsammans med ord som det har olika syntaktiska relationer till i materialet, grupperat efter relation. För ett verb visas till exempel vanliga subjekt och objekt, och för ett substantiv visas vanliga attribut, och vanliga verb som substantivet är subjekt och objekt till.

För att söka i ordbilden måste du först klicka i **Visa ordbild**.

4.1 Enligt GP-korpusarna, vad gör en hund ofta och vad gör man ofta med en hund?

4.2 Från ordbilden, plocka fram alla meningar där hundar skäller.

ÖVNING 5 - JÄMFÖRA OLIKA KORPUSAR

5.1 Testa att välja Bloggmix och GP-korpusarna. Försök att komma på ett ord som borde vara vanligare i dagstidningarna är i bloggarna och se om du kan verifiera din idé.

Tips: Sök på ditt ord som lemgram och välj även för- och efterled för att få bättre resultat.

ÖVNING 6 - ORDBETYDELSEDISAMBIGUERING

När man söker på visssa ord som har många betydelser kan man få fler träffar än man önskar sig. Genom att söka på specifika betydelser kan man få en mindre träffmängd där träffarna är mer relvanta.

Den nya betydelsedisambigueringen finns i tidningarna Press 95-98 och GP 2012-2013 samt i Sociala medier, Bloggmix 1998-2004 och 2014-2015.

6.1 Sök på ordet **fil**. Sammanställ på betydelse. Notera att när du klickar på ett ord i KWIC-vyn så kan man se ordets betydelseannotering i sidopanelen. När man klickar på visa fler så får man se alla alternativ tillsammans med deras sannolikheter. Kolla även statistikfliken där man kan se frekvens för de olika betydelserna.

6.2 Gå till **Utökad sökning**. Välj **betydelse**. Om du skriver **fil** så kommer en meny fram där du kan välja vilken betydelse du vill söka på. När du söker får du fram alla ord som är uppmärkta med betydelsen som valts.

ÖVNING 7 - KARTA

Med Korps nya kartfunktion kommer det att finnas möjlighet att visa en karta baserat på textattribut så som författarens hemort eller plats där en tweet skrevs. För nurvarande kan man för Bloggmix 1998-2004 och 2014-2015 visa en karta baserat på bloggarens hemort. I Press 95-98 och GP 2012-2013 kan man visa en karta baserat på vilka platser som samförekommer i träffarnas stycke.

7.1 Välj Bloggmix och gå till **Utökad sökning**. Sök på **lemgram är palt** eller **lemgram är kroppkaka**. Sammanställ på **lemgram**. I statstikfliken kan du markera alla rader och klicka på **Visa karta**. Du ser direkt att träffarna fördelas så som man förväntar sig på kartan.

7.2 Genom att klicka på en träff och sedan klicka på rutan som kommer upp kan du få fram en KWIC-vy över träffar på den plats som du valt.

7.3 Försök att hitta andra fenomen som borde vara vanligare i t.ex. Göteborg, Stockholm eller någon annan av storstäderna.

7.4 Gör samma som ovan i GP 2012-2013 med samförekomst.