Att söka i Korp med CQP och Regexp – en introduktion

Klas Hjortstam 2018

MEIJERBERGS INSTITUT
FÖR
SVENSK ORDFORSKNING

Förord

Under våren 2018 genomfördes ett projekt finansierat av Meijerbergs institut för svensk ordforskning vid Göteborgs universitet. Syftet var att med språkhistorikerns ögon se över metadatan i de historiska texter som tillhandahålls av Språkbanken via korpusverktyget Korp. Under arbetets initialskede kom frågan upp om inte en handledning i grundläggande CQP kunde vara till hjälp för alla användare av Korp i allmänhet och för språkhistoriker i synnerhet. Alla användare kan ha god nytta av den större precision och kontroll som CQP tillför; för den som söker i historiskt material, med dess ortografiska variation kan skillnaden vara avgörande. Föreliggande text avser vara just en sådan handledning.

Författaren har fått fin hjälp av Anna Hannesdóttir, Joakim Lilljegren, Martin Hammarstedt, Malin Ahlberg, Anne Schumacher, Yvonne Adesam och Henrik Rosenkvist vid Institutionen för svenska språket, Göteborgs universitet, samt Lars-Olof Delsing vid Språk- och litteraturcentrum, Lunds universitet och Erik Magnusson Petzell vid Institutet för språk och folkminnen.

Med reservation för ändringar i Korp efter handledningens publicerande.

Innehållsförteckning

1 Inledning

Anledningarna till att du vill använda Språkbankens korpussökningsverktyg Korp (Borin, Forsberg & Roxendal 2012)[1] kan vara flera. Oavsett om du är allmänt språkintresserad, språkstudent, doktorand eller fullfjädrad forskare önskar du dock troligen sökresultat av så hög kvalitet som möjligt; ju fler relevanta träffar, och ju mindre "skräp" att sortera bort manuellt, desto mer användbart är det databasen ger dig. Ett verktyg för att åstadkomma detta är frågespråket *CQP* inklusive notationspraxisen *Regexp*, vilka ger dig möjlighet att inkludera och exkludera med precision.

Handledningen vänder sig både till dig som tvekar inför de avancerade sökteknikerna och dig som börjat använda dem och vill komplettera dina kunskaper.

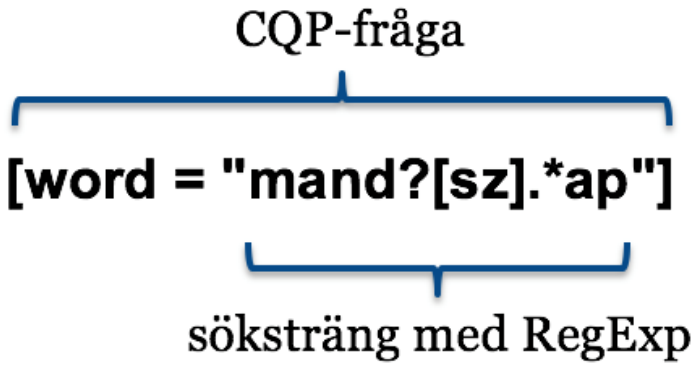
För dig som intresserar dig för Korps historiska material rekommenderas Joakim Lilljegrens *Introduktion till Språkbankens historiska material* (2018)[2] som komplement.

[1] Verktyget hittar du på http://spraakbanken.gu.se/korp
[2] Introduktionen hittar du på https://meijerbergs.hum.gu.se/digitalAssets/1694/1694144_introduktion.pdf

2 Vad är CQP och Regexp?

I varje databas finns alltid ett underliggande systemspråk som används för sökningar - ett *frågespråk* (*query language*). Frågespråket är vanligen dolt bakom ett mer användarvänligt gränssnitt, men för den som vill göra avancerade sökningar är det ofta både kraftfullare och mer praktiskt att söka direkt med frågespråket. I fallet med Språkbankens korpusdatabas *Korp* heter frågespråket *Corpus Query Processor query language* (CQP). En hel del beståndsdelar i CQP har sin grund i engelska språket, vilket förklarar ord och förkortningar som *word*, *variants*, *pos* (part of speech), *msd* (morphosyntactic description) etc.

En viktig del av den fråga du ställer med CQP är *söksträngen* (*Figur 1*) som måste skrivas i enlighet med notationspraxisen *Regular expressions* (Regexp, Regex., sv.: *reguljära uttryck*). Fortsättningsvis avses "CQP, inklusive Regexp" när förkortningen CQP används. När termen Regexp används ensam avses dock bara Regexp.



Figur 1

2.1 Tre sätt att söka

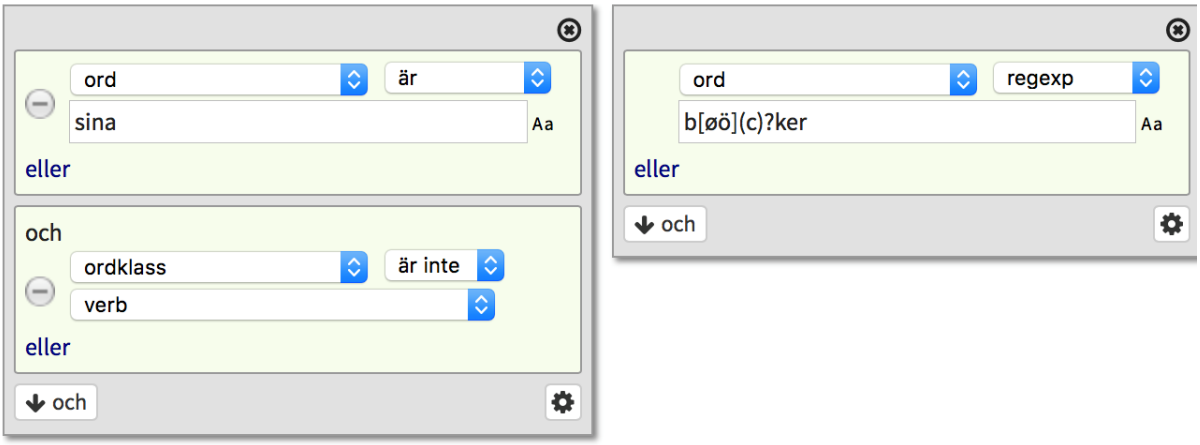
Språkbanken tillhandahåller tre alternativ för sökningar i Korp: *enkel sökning* (**2.1.1**), där frågespråket är helt dolt för användaren, *utökad sökning* (**2.1.2**), där frågespråket i viss utsträckning kan (men inte måste) användas, samt till sist *avancerad sökning* (**2.1.3**) där frågan i sin helhet skrivs med frågespråket CQP av dig som användare. Det är det sistnämnda alternativet denna handledning främst handlar om.

Emellanåt rekommenderas tekniker med en kombination av utökad och avancerad sökning (se **2.3**, flera avsnitt under **4**), varvid även gränssnittet för utökad sökning berörs. Mycket av den Regexp som presenteras kan dessutom användas som del av frågan i utökad sökning genom att *regexp* eller *inte regexp* väljs i den högra rullgardinsmenyn (**2.1.2**, *Figur 3*).

2.1.1 Enkel sökning - CQP är dolt

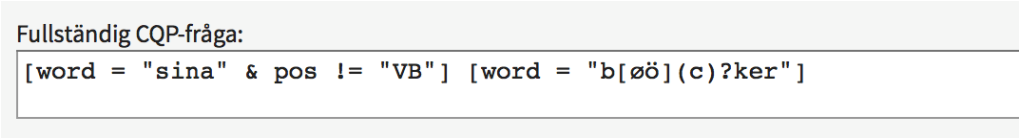
Figur 2

2.1.2 Utökad sökning - *Regexp kan användas*



Figur 3

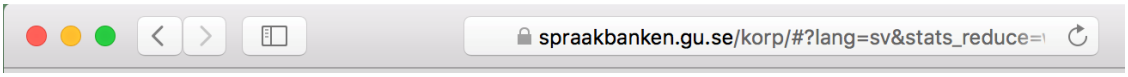
2.1.3 Avancerad sökning - *frågan skrivs i CQP*



Figur 4

2.2 Spara dina sökningar

När du väl gjort din sökning kan du spara den genom att kopiera den adress som skapats i webbläsarens adressfönster (*Figur 5*).



Figur 5

Så här kan webbadressen se ut (den kan vara avsevärt längre):

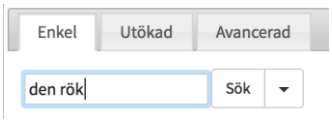
https://spraakbanken.gu.se/korp/#?lang=sv&stats_reduce=word&cqp=%5B%5D&page=0&corpus=twitter&search=word%7Csparad%20sökning

Spara adressen i ett eget dokument. Även det urval av korpusar du använde finns med i adressen. På detta vis kan du spara sökningar för återanvändning eller för att skicka till någon. Detta fungerar oavsett om du använt enkel, utökad eller avancerad sökning.

2.3 Se dina egna frågor i CQP

Ett mycket användbart pedagogiskt verktyg inbyggt i Korp är att du kan skriva en fråga i enkel (*Figur 6*) eller utökad (*Figur 7*) sökning och sedan skifta till fliken för avancerad sökning, där frågan nu presenteras konverterad till en CQP-fråga.

Fråga i enkel sökning

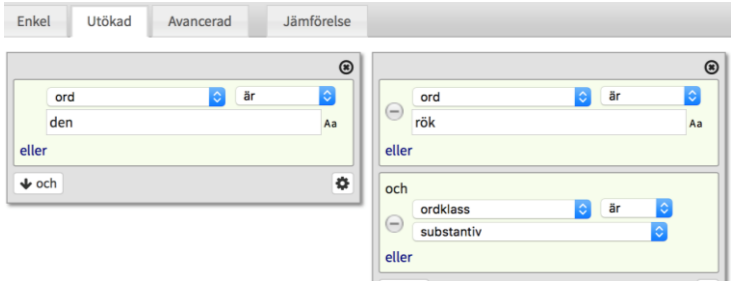


Figur 6

Samma fråga som CQP

```
[word = "den"] [word = "rök"]
```

Fråga i utökad sökning



Figur 7

Samma fråga som CQP

```
[word = "den"] [word = "rök" & pos = "NN"]
```

2.4 Vilket CQP pratar vi om?

Denna handledning beskriver inte CQP i dess helhet, utan är begränsad dels till de delar av CQP som alls kan nyttjas i Korps gränssnitt, dels till vad som rimligen kan ingå i en introduktion av detta format. Den som önskar ytterligare kunskap om CQP hänvisas till Evert m.fl. (2016).

Nämnas bör att den inkluderade notationen Regexp är en praxis med stor flexibilitet; databasutvecklaren har frihet att anpassa den till de egna förutsättningarna. Den "dialekt" av Regexp som ligger till grund för Korps variant är *PCRE - Perl Compatible Regular Expressions* (PCRE).

3 Sök ett token

När du gör en sökning i Korp söker du ett eller flera *token*, vilket är det texterna i en korpus utgörs av. Majoriteten av alla token utgörs av ett *ord* (se dock **3.7**). Enklast möjliga CQP-fråga ser ut enligt frågan nedan, som ger träffar på alla token som består av teckenföljden <l>, <a>, <g>, <o>, <m>, alltså ordet *lagom*:

```
[word = "lagom"]
```

Parametern **word** har sökströngen **lagom** som *värde*.

Enkla frågor som den ovan kan förkortas:

```
"lagom"
```

Du får alltså samma resultat med de båda frågorna ovan. Så snart frågorna blir mer komplicerade är dock hakparenteser, parameter och lika-med-tecken obligatoriska.

Relationen parameter/värde kan inverteras genom att <**=**> (*är lika med*) byts mot <**!=**> (*är inte lika med*). Denna fråga ger träff på alla token som inte är ordet *lagom*:

```
[word != "lagom"]
```

OBS: Om du glömmer citattecknen, eller om du anger en parameter som inte finns:

```
[word = lagom]
```

```
[wyrd = "lagom"]
```

svarar Korp **Antal träffar: 0**. Du får alltså inte ett felmeddelande, och risken är att du felaktigt tror att det du söker inte finns. Var alltså noga med parametern och citattecknen!

OBS: Allt i CQP-syntaxen är skiftlägeskänsligt. Parametern **word** måste t.ex. skrivas med enbart gemener - skriver du **Word** eller **WORD** får du inga träffar. Även sökströngen är skiftlägeskänslig, vilket t.ex. gör att frågan nedan inte ger träffar på *dag*, bara på *Dag*.

```
[word = "Dag"]
```

Hur du hanterar skiftlägeskänslighet i sökströngen ser du i avsnitt **3.5**.

OBS: En bokstav med diakritiskt tecken anses inte vara samma bokstav som utan. Frågan nedan får inte träffar på *cafe*, bara på *café*.

```
[word = "café"]
```

Hur du hanterar diakritiska tecken i sökströngen ser du i avsnitt **3.5**.

3.1 Hitta varianter med parenteser () och lodstreck |

När du söker ett ord med stavningsvariation, eller när du söker flera former eller avledningar av ett ord, kan du med hjälp av Regexp ange vad du är intresserad av.

Tecken eller grupper av tecken som är sinsemellan utbytbara åtskiljs med lodstreck: **ä|æ|e|ae**. För att skilja dem från ej utbytbara används parenteser: **H(ä|æ|e|ae)ster**. Frågan nedan returnerar alla förekomster av *Häster*, *Hæster*, *Hester* och *Haester*.

```
[word = "H(ä|æ|e|ae)ster"]
```

Frågan nedan returnerar alla förekomster av *Hestar*, *Hestars*, *Hestarna* och *Hestarnas*:

```
[word = "Hest(ar|ars|arna|arnas)"]
```

Du kan ange alternativ på fler än en position i strängen. Frågan nedan ger träff på alla förekomster av *Hästar*, *Hästars*, *Hästarna*, *Hästarnas, Hæstar*, *Hæstars*, *Hæstarna*, *Hæstarnas, Hestar*, *Hestars*, *Hestarna*. *Hestarnas, Haestar*, *Haestars*, *Haestarna* och *Haestarnas*.

```
[word = "H(ä|æ|e|ae)st(ar|ars|arna|arnas)"]
```

Om det är hela ordet som är utbytbart behövs inga parenteser, men det påverkar inte resultatet att trots det använda dem. Frågorna nedan ger identiskt resultat (träffar på *droska* och *vagn*):

```
[word = "(droska|vagn)"]
```

```
[word = "(droska)|(vagn)"]
```

```
[word = "droska|vagn"]
```

OBS: Använd endast lodstreck *mellan* alternativen; skriv inget inledande eller avslutande lodstreck. Extra lodstreck kommer att tolkas som att du inluderar *'tom sträng'* (vill säga *'ingenting'*) bland alternativen. Nedanstående fråga tolkas som att du söker *Hest+er*, *Hest+a* eller *Hest+um*. Den ger således träff på *Hester*, *Hesta* och *Hestum*.

```
[word = "Hest(er|a|um)"]
```

Nedanstående fråga (observera det avslutande lodstrecket) tolkas istället som att du söker *Hest+er*, *Hest+a*, *Hest+um* eller *Hest+'tom sträng'*. Den ger alltså träff på *Hester, Hesta, Hestum* och *Hest*:

```
[word = "Hest(er|a|um|)"]
```

3.1.1 Nästning av parenteser

Parenteser kan innehålla parenteser, som i sin tur innehåller parenteser. Detta är användbart när den variation du önskar är komplicerad. Frågan nedan ger träffar på *hästfoder*, *hæstfoder*, *hestfoder* och *vinterfoder*. Notera hur lodstreck står mellan utbytbara tecken eller grupper av tecken, och hur det som är utbytbart avskiljs från resten med parenteser:

```
[word = "((h(ä|æ|e)st)|(vinter))foder"]
```

Nästning kan expanderas i princip hur långt som helst. Vi nöjer oss i frågan nedan med att nästa ytterligare några parenteser för att tillåta även binde-a (*hästafoder*, *hæstafoder*, *hestafoder*), stavning med <*w*> (*winter*) samt stavning med <*þ*> (*foþer*):

```
[word = "((h(ä|æ|e)(st|sta))|((w|v)inter))fo(d|þ)er"]
```

Ytterligare exempel på avancerad användning av parenteser finner du bl.a. under **7.4**.

3.2 Uppsättning utbytbara tecken inom hakparentes []

Hakparentes kan användas när enstaka tecken är sinsemellan utbytbara. Lodstreck används då inte. Frågan nedan ger alla förekomster av *hästa* och *hästs* (men alltså inte *häst*).

```
[word = "häst[as]"]
```

Frågan nedan returnerar alla förekomster av *Häster, Hæster* och *Häster* (jmfr **3.1**).

```
[word = "H[äæe]ster"]
```

3.2.1 Spann inom hakparentes

Inom hakparentes kan ett spann representeras av sin start- respektive slutpunkt (gäller siffror 0-9 och bokstäver a-z). Frågan nedan returnerar *rea*, *reb*, *rec* och *red*:

```
[word = "re[a-d]"]
```

Nedanstående fråga returnerar *A1, A2, A3* och *A4*:

```
[word = "A[1-4]"]
```

3.2.2 Kombinera tecken och spann

Enstaka tecken och spann kan kombineras. Frågan nedan ger *rea, rem, ren, reo* och *rep*:

```
[word = "re[am-p]"]
```

Frågan nedan returnerar *A1, A3, A4, A5* och *A6*:

```
[word = "A[13-6]"]
```

Frågan nedan returnerar *AB, AC, AD, A1, A3, A4, A5* och *A6*:

```
[word = "A[B-D13-6]"]
```

OBS: en uppsättning inom hakparentes måste bestå av enstaka tecken eller av spann angivna med två enstaka tecken på var sin sida om <-> (bindestreck). Frågan nedan ger alltså *inte* träff på *10*, *11*, *12* och *13*.

```
[word = "[10-13]"]
```

Frågan ovan tolkas istället av Korp så:

Sök efter ett token som består av tecknet

1

ELLER av något av tecknen inom spannet

0-1

ELLER av tecknet

3

Frågan returnerar alltså träffar på 0, 1 och 3.

Önskar du träffar på *10*, *11*, *12* och *13* kan du t.ex. skriva på något av följande vis:

```
[word= "1[0-3]"]
```

```
[word= "1[0123]"]
```

```
[word= "10|11|12|13"]
```

```
[word= "1(0|1|2|3)"]
```

3.2.3 Bindestreck inom hakparentes

Om du vill att <-> (bindestreck) skall vara ett av de utbytbara tecknen måste du skriva det i en position där det inte kan tolkas som angivande av spann. Vill du t.ex. få träffar på <->, <**1**> eller <**4**> skriver du alltså *inte* så här (jmfr **3.2.1**):

```
[word = "[1-4]"]
```

utan t.ex. så här:

```
[word = "[-14]"]
```

Vill du få träffar på <-> och alla tecken inom spannet 1-4 skriver du t.ex. så här:

```
[word = "[-1-4]"]
```

3.2.4 Invertera teckenuppsättningen

Teckenuppsättningen kan inverteras med <**^**>. Du får träffar på alla tecken som inte finns i uppsättningen. Frågan nedan får träffar på *Hoster*, *Höster*, *H°ster*, *H8ster* etc., men inte *Häster, Hæster* och *Hester*.

```
[word = "H[^äæe]ster"]
```

Frågan nedan får träffar på *1866* och *1966* (även *1H66* och *1€66*) etc. men inte t.ex. *1066*.

```
[word = "1[^0-7]66"]
```

Kom i håg att en bokstav med diakritiskt tecken inte anses vara samma bokstav som utan (se **3**, **3.5**). Frågan nedan får inte träffar på Te och To, men på *Té*, *Tô* (även *Tv*, *T3*) etc.

```
[word = "T[^eo]"]
```

OBS: hakparenteser har alltså två distinkta roller. Dels kan de markera gränserna för ett uttryck som beskriver ett token:

```
[word = "Häster"]
```

Dels kan de, när de används i en Regexp-sträng, markera en uppsättning sinsemellan utbytbara tecken:

```
"H[äæe]ster"
```

3.2.5 Hakparenteser i parenteser

Hakparenteser kan inte skrivas inuti andra hakparenteser (*nästas*), men de kan skrivas inuti parenteser. Den långa nästningen under 3.1.1 kan skrivas om med hakparenteser i de positioner där enstaka tecken är sinsemellan utbytbara. Här återges både den ursprungliga frågan och den omskrivna:

```
[word = "((h(ä|æ|e)(st|sta))|((w|v)inter))fo(d|þ)er"]
```

```
[word = "((h[äæe](st|sta))|([wv]inter))fo[dþ]er"]
```

Fler exempel på avancerad användning av hakparenteser och parenteser finner du under **7.3**.

3.3 Operatorerna <?>, <*> och <+> samt klammerparentes {}

3.3.1 Operatorn <?>

Vill du att något ska vara optionellt använder du operatorn <**?**>, som betyder att det närmast föregående elementet (tecken eller grupp av tecken avgränsade med parentes eller hakparentes) ska förekomma 0 eller 1 gånger.

Frågan nedan returnerar alla förekomster av *fira* och *firas*.

```
[word = "firas?"]
```

Frågan nedan returnerar alla förekomster av *fira*, *firat*, *firad*, *firas* och *firar*.

```
[word = "fira[tdsr]?"]
```

Frågan nedan returnerar förekomster av *för*, *föra*, *förde*, *förþe*, *förþ*, *fördhe*, *fördh* och *fört*.

```
[word = "för(a|de|þe|þ|dhe|dh|t)?"]
```

När operatorn endast skall påverka ett enstaka tecken måste inte parenteser användas, men det påverkar inte resultatet att trots det använda dem. Frågorna nedan ger alltså identiskt resultat (samtliga förekomster av *lagom* och *laghom*).

```
[word = "lag(h)?om"]
```

```
[word = "lagh?om"]
```

Exempelfrågan under **3.2.5** kan ytterligare förkortas med operatorn <**?**>. Här återges de båda tidigare varianterna av frågan såväl som den senaste förkortingen med <**?**>:

```
[word = "((h(ä|æ|e)(st|sta))|((w|v)inter))fo(d|þ)er"]
```

```
[word = "((h[äæe](st|sta))|([wv]inter))fo[dþ]er"]
```

```
[word = "((h[äæe]sta?)|([wv]inter))fo[dþ]er"]
```

3.3.2 Operatorn <*>

En asterisk <*> anger att det närmast föregående elementet (tecken eller grupp av tecken avgränsade med parentes eller hakparentes) ska förekomma 0 eller fler gånger. Frågan nedan returnerar *hej*, *heja*, *hejaa*, *hejaaa* etc (hur många <*a*> som helst):

```
[word = "heja*"]
```

Frågan nedan returnerar *trala*, *tralala*, *tralalala*, *tralalalala* etc.:

```
[word = "trala(la)*"]
```

När operatorn endast skall påverka ett enstaka tecken måste inte parenteser användas, men det påverkar inte resultatet att trots det använda dem. Frågorna nedan ger alltså identiskt resultat (*hej*, *heja*, *hejaa*, *hejaaa* etc):

```
[word = "heja*"]
```

```
[word = "hej(a)*"]
```

3.3.3 Operatorn <+>

Ett plustecken <+> anger att det närmast föregående elementet (tecken eller grupp av tecken avgränsade med parentes eller hakparentes) ska förekomma 1 eller fler gånger.

Frågan nedan returnerar *tralala*, *tralalala*, *tralalalala* etc. (men alltså inte *trala*).

```
[word = "trala(la)+"]
```

När operatorn endast skall påverka ett enstaka tecken måste inte parenteser användas, men det påverkar inte resultatet att trots det använda dem. Dessa båda frågor ger alltså identiskt resultat (*heja*, *hejaa*, *hejaaa* etc., men inte *hej*).

```
[word = "heja+"]
```

```
[word = "hej(a)+"]
```

3.3.4 Klammerparenteser {}

Ett tal inom klammerparenteser **{}** anger exakt antal gånger ett tecken eller en grupp av tecken ska förekomma. Frågan nedan ger träffar endast på *heej*:

```
[word = "he{2}j"]
```

Två tal separerade med kommatecken anger ett spann för antal gånger det närmast föregående elementet (tecken eller grupp av tecken avgränsade med parentes eller hakparentes) ska förekomma. Frågan nedan ger träffar på *heej*, *heeej*, *heeeej* och *heeeeej*:

```
[word = "he{2,5}j"]
```

När det är fler än ett tecken som skall påverkas måste som vanligt parentes användas. Frågan nedan ger träffar på *kvitt*, *kviddevitt*, *kviddeviddevitt* och *kviddeviddeviddevitt*:

```
[word = "k(vidde){0,3}vitt"]
```

3.4 Vilket tecken som helst <.>

En punkt <.> motsvarar vilket tecken som helst (eng. *matchall*). Frågan nedan ger alltså träffar på *mader*, *mager*, *maler*, *maþer, ma#er* etc.:

```
[word = "ma.er"]
```

Matchall ger flexibla möjligheter tillsammans med operatorerna för repetition och olika parenteser. Frågan nedan ger bl.a. *manskap*, *mandskap, mandskaap* och *mandszkap* i Korps fornsvenska material:

```
[word = "mand?[sz].*ap"]
```

Frågan nedan, ställd till korpussamlingarna under *Historiskt*, ger t.ex. *hestar, hestanä, hestarna, hestakööp* men inte *hestafoder* (eftersom efterledet *foder* har fler än fyra tecken):

```
[word = "hesta.{1,4}"]
```

OBS: När <.> används tillsammans med operatorerna <*> och <+> eller med antal inom klammerparenteser (t.ex. **{2}**) är det inte nödvändigtvis samma tecken som upprepas. Frågan nedan kan returnera *sin*, *siin*, *sijn*, *siþan*, *siælfwæn* och många fler.

```
[word = "si.*n"]
```

3.5 Ignorera skiftläge och diakritiska tecken

I enkel och utökad sökning kan du med klick i gränssnittet ange att sökningen ska ignorera skiftläge. I avancerad sökning anges detta istället med **%c** direkt efter söksträngen - utanför citattecknen. Frågan nedan ignorerar om tecknen är versala eller gemena och ger följdaktligen träffar såväl på *Björn* som *björn* (även *bJÖrn*, *björN* etc.).

```
[word = "björn" %c]
```

Du kan även be Korp bortse från alla diakritiska tecken. Det görs med **%d** direkt efter sökströngen - utanför citattecknen. Frågan nedan ger träffar på *biörn* och *biorn* (även *bïorn*, *biôrn* etc., om det skulle finnas i texten).

```
[word = "biörn" %d]
```

De båda (**%c** och **%d**) kan kombineras, och skrivs då med endast ett procenttecken. Frågan nedan ger träffar på *biörn*, *biorn*, *Biörn*, *Biorn* etc.:

```
[word = "biörn" %cd]
```

3.6 Att söka på specialtecknen

Som du har sett finns det tecken med speciella funktioner. Parenteser, lodstreck, <**?**>, <*****> m. fl. utgör instruktioner till databasen, när du skriver dem utan någon särskild markering. Vill du istället söka dessa tecken "som sig själva" måste du använda *avbrottstecknet* (*escape character*) omvänt snedstreck <****> för att visa att tecknet i detta fall inte är ett specialtecken.

Frågan nedan ger felmeddelande i Korp. Sökströngen består av en felaktigt använd operator:

```
[word = "?"]
```

Frågan nedan ger däremot träffar på materialets frågetecken:

```
[word = "\?"]
```

De tecken som avses är <**.**>, <**?**>, <*****>, <**+**>, <**|**>, <**(**>, <**)**>, <**[**>, <**]**>, <**{**>, <**}**>, <**^**> och <**$**>.

> **OBS:** <**$**> är inte en del av Regexp i CQP, men måste ändå markeras med avbrottstecken.

3.7 Skiljetecken som token

Texterna i Korp utgörs av *token*. Oftast är ett token vad vi i dagligt tal kallar ett *ord*. Ett token kan dock även bestå av ett skiljetecken - <.>, <,>, <?>, <;> etc. Att de behandlas som egna token beror på två saker:

För det första ska de kunna sökas och analyseras separat, utan att höra till ett ord. Frågan nedan ska, som redan sagts, ge träffar på skiljetecknet <?> (se **3.6**).

```
[word = "\?"]
```

För det andra ska ord ge träff även om de står med skiljetecken. Frågan nedan ska ge träff i såväl satsen *Eller **hwad** gingen j vth til at see?* som satsen *weten i **hwad**?*

```
[word = "hwad"]
```

Hade frågetecknet räknats till ordet hade frågan inte gett träff på den andra meningen.

Vill du söka ord som står med skiljetecken får du skriva det som två token i samma fråga. Denna fråga ger träffar på alla *hwad* som är följda av ett frågetecken:

```
[word = "hwad"] [word = "\?"]
```

Se mer om att söka på fler än ett token nedan under **5**.

3.7.1 Citattecken

Precis som i fallet med andra skiljetecken (**3.7**) måste du skriva varje citattecken som ett eget token. Eftersom citattecknen har funktion som avgränsare av Regexp-söksträngar: **[word = "handledning"]**, markeras de på ett särskilt sätt när de istället skall tolkas som del av söksträngen. Här används inte avbrottstecknet <****> (**3.6**), utan en speciell kombination av citattecken.

När du söker enkla citattecken avgränsar du även söksträngen med enkla citattecken. Inom dessa skriver du *två* enkla citattecken för att representera *ett*. I frågan nedan ser du alltså fyra enkla citattecken på rad, och den ger träffar på enstaka enkla citattecken:

```
[word = ""]
```

När du söker dubbla citattecken avgränsar du söksträngen med dubbla citattecken. Inom dessa skriver du sedan *två* dubbla citattecken för att representera *ett*. I frågan nedan ser du alltså fyra dubbla citattecken på rad, och den ger träffar på enstaka dubbla citattecken:

```
[word = """"]
```

Frågan nedan ger träffar på alla förekomster av *'häst'*, inklusive de enkla citattecknen:

```
[word = ""] [word = "häst"] [word = ""]
```

Frågan nedan ger träffar på alla förekomster av *"häst"*, inklusive de dubbla citattecknen:

```
[word = """"] [word = "häst"] [word = """"]
```

Se mer om att söka på fler än ett token nedan under **5**.

OBS: Taggningen av skiljetecken, i kombination med de tekniska förutsättningarna för Korp, gör att det kan vara ett blanksteg mellan skiljetecknet och det närmast liggande ordet, oavsett om där var något blanksteg i originaltexten eller inte. Du kan dock bortse från detta - frågan nedan ger träff på både *Hej!* och *Hej !*

```
[word = "Hej"] [word = "!"]
```

3.8 Fler parametrar i samma fråga

Det går bra att kombinera två parameter-/värdepar i samma fråga. Om du t.ex. konstruerat en söksträng för ett ord med stavningsvariation, men vill exkludera en möjlig stavning kan du använda parametern **word** ännu en gång fast med operatorn **!=** (*inte lika med*). Du fogar det andra parameter-/värdeparet till det första med **&**. Frågan nedan ger träffar på *Klas*, *Claes* och *Klaes* men inte *Clas*:

```
[word = "[KC]lae?s" & word != "Clas"]
```

Läs om andra multipla parameter-/värdepar under **4** nedan.

3.9 Blanksteg och radbrytning

3.9.1 Blanksteg i CQP-syntaxen

Blanksteg är insignifikanta i CQP-syntaxen (utom i söksträngen, se **3.9.3**). Det betyder att du mellan de olika beståndsdelarna i syntaxen kan skriva med eller utan blanksteg, och hur många blanksteg du vill. De tre frågorna nedan ger alla samma resultat (observera att ingen fråga har blanksteg i söksträngen innanför cittattecknen):

```
[word="forbuþ"]

[ word = "forbuþ" ]

[word    ="forbuþ"    ]
```

3.9.2 Radbrytning i CQP-syntaxen

Radbrytning är insignifikanta i CQP-syntaxen (utom i söksträngen, se **3.9.3**). Det betyder att du kan dela upp din fråga på flera rader för överskådlighet (frågan är från **7.4.1**):

```
[word = "(m[äe](þ|th)an)|([eä][fp]t[eiy]r?)|([uvw]th?an)|(at[ht]?)"]
[word = "([ij]a(k|gh?))|([tþ]h?u)|(h[aou]n)|([uw]ij?)|((þ|th)[ei])"]
[] []
[word = "o(k|c[hk]?)"]
```

3.9.3 Blanksteg och radbrytning i söksträngen

Blanksteg och radbrytning utgör aldrig egna token och är heller aldrig del av ett token. I söksträngen för token (`[word = "sökträng"]`) bör du därför aldrig använda blanksteg och radbrytning. Vill du söka på ett uttryck eller ett namn som utgörs av två ord (*for þy*, *Gustav Vasa*) måste du skriva orden som var sina token (se **5**). I ordattributen förekommer aldrig mellanslag (eller radbrytning) i värdet, så inte heller där skall sådana användas. I värden för textattribut händer det dock att blanksteg ingår - t.ex. textattributen *titel* (**4.2.1**) och *tidsintervall* (**4.2.2**).

OBS: Om du använder blanksteg i söksträngen, som i exemplet nedan:

```
[word = "Gustav Vasa"]
```

svarar Korp **Antal träffar: 0.** Du får alltså inte ett felmeddelande, och risken är att du felaktigt tror att det du söker inte finns.

Använder du blanksteg i en sträng där element är sinsemellan utbytbara får du inga träffar på de varianter där blankstegen är skrivna. Frågan nedan - notera blankstegen före <**ä**> och efter <**e**> - ger träffar på *Hæster* och *Haester* (men alltså varken på *Häster* och *Hester* eller på *H äster* och *He ster*):

```
[word = "H( ä|æ|e |ae)ster"]
```

Om du använder radbrytning i söksträngen får du ett felmeddelande.

4 Sök ett annoterat token

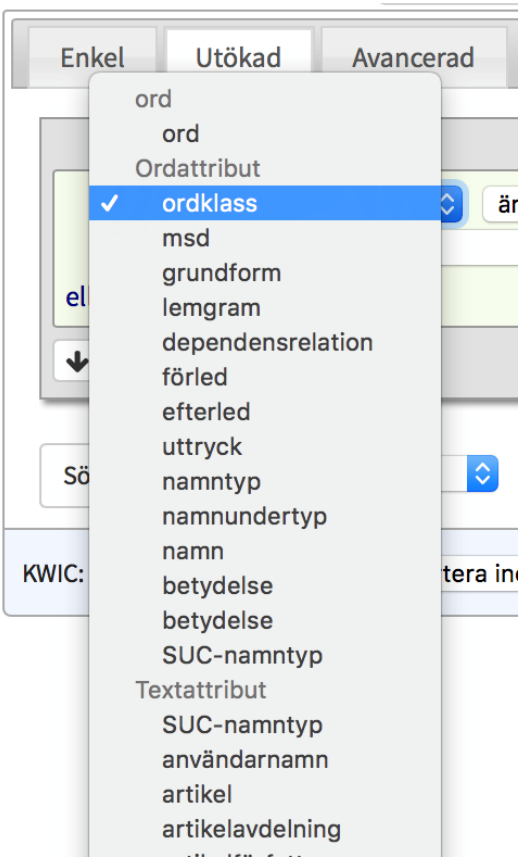
Ett token kan vara *annoterat* ("taggat") med *ordattribut* av olika slag, som *ordklass*, *dependensrelation* eller *betydelse*. Hela texten du söker i kan i sin tur vara annoterad med *textattribut* som *datum*, *bloggadress*, *issn* eller *antal sidor*. Annoteringen av Korps historiska material (som den såg ut i maj 2018) redovisas i Joakim Lilljegrens *Introduktion till Språkbankens historiska material* (2018).

OBS: Alla korpusar är inte annoterade med samma attribut. Om du söker med *attributet X* i *korpus A* och *korpus B*, där *korpus A* är annoterad med *attributet X* och *korpus B* inte är det kan problem uppstå. Du får nämligen inget felmeddelande; Korp returnerar helt enkelt de token med *attributet X* som finns i *korpus A* och nonchalerar *korpus B*. Risken är att du felaktigt tror att det du söker inte finns i *korpus B*. Var därför noga med att ta reda på vilka annoteringar som kan användas med vilken korpus.

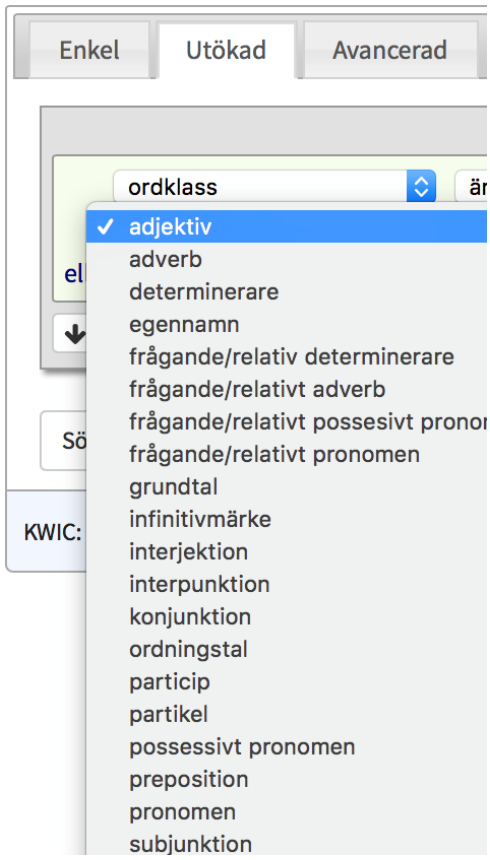
Korps annoteringar redovisas som del av sökresultatet; om du klickar på ett token bland träffarna visas dess (och dess texts) attribut i sidopanelen (*Figur 8*).

Figur 8

Du kan också använda attributen i din sökning. I utökad sökning väljs attributet (*Figur 9*) och dess värde (*Figur 10*) i rullgardinsmenyer.

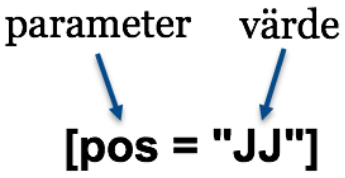


Figur 9



Figur 10

I avancerad sökning anges attribut och värde i CQP-frågan, precis som i fallet med ord (se **3**), med en parameter och ett värde (*Figur 11*). Frågan i *figur 11* ger träff på alla token som är annoterade med ordklass adjektiv - **pos** står för *part of speech* (*ordklass*), medan **JJ** står för *adjektiv*, se **4.1.1** och **4.1.2**).



Figur 11

4.1 Ordattribut

4.1.1 Ordattributet *ordklass*

Om önskat attribut är *ordklass*, noteras det som parametern **pos** (*part of speech*), med värden som t.ex. *verb*, noterat **VB**. Frågan nedan ger träffar på alla verbannoterade token:

```
[pos = "VB"]
```

Som nämnts under **3.8** kan två parametrar kombineras i samma fråga med **&**. Frågan nedan ger träff på alla förekomster av *kör* som är annoterade som verb.

```
[word = "kör" & pos = "VB"]
```

Ordklassförkortningarna för ordattributet *ordklass* finner du i **Bilaga 1**. Observera att det inte är samma som ordklassförkortningarna för ordattributet *lemgram* (**4.1.4**).

4.1.2 Ordattributet *homografmängd*

Delar av det historiska materialet i Korp är inte annoterade med *ordklass*, utan med det närliggande attributet *homografmängd*. Begreppet avser alla de ordklasser ett token och dess homografer[3] kan ha. Ordet *plåga* i *Han war dock intet bättre än de, til at **plåga** grannarne* har således *ordklass* verb men *homografmängd* verb och substantiv. Det homografa ordet *plåga* i *Jag wet ingen wärre **plåga*** har *ordklass* substantiv, *homografmängd* verb och substantiv. Ett token kan vara annoterat med endast en homogafmängdsannotering (*söter*; adjektiv) eller med flera (*far*; substantiv, verb)

Om önskat attribut är *homografmängd*, noteras det som parametern **posset** (*part of speech-set*) följt av **contains** och värden som t.ex. *verb*, noterat **VB**. Frågan nedan ger träffar på alla token annoterade med homografmängd substantiv:

```
[posset contains "NN"]
```

Frågan nedan ger träffar på alla token annoterade med både homografmängd substantiv och homografmängd verb:

```
[posset contains "NN" & posset contains "VB"]
```

För att invertera frågan skrivs **not contains**. Denna fråga ger träff på alla token som är annoterade med annan homografmängd än substantiv:

```
[posset not contains "NN"]
```

Ordklassförkortningarna för ordattributet homografmängd finner du i **Bilaga 1**. Observera att det är samma som för ordattributet *ordklass*, men inte samma som för ordattributet *lemgram*.

[3] Homograf = två ord som stavas likadant, oavsett uttal. Substantiven kors (krucifix) och kors (nöt av honkön; plural, genitiv) är homografa; värk (smärta) och verk (t.ex. konstverk) är det inte.

4.1.3 Ordattributet *grundform*

Om önskat attribut är *grundform*, noteras det med parametern **lemma** följt av **contains**. och ett värde. Frågan nedan ger träff på alla token som är annoterade med grundform *röra*, oavsett om det är verbet eller dess homonyma substantiv:

```
[lemma contains "röra"]
```

för att invertera frågan skrivs **not contains**. Denna fråga ger träff på alla token som är annoterade med annan grundform än *röra*:

```
[lemma not contains "röra"]
```

4.1.4 Ordattributet *lemgram*

Ett *lemgram* är ett ords eller ett flerordsuttrycks samtliga böjningsformer och sammansättningsformer. Annoteringen av *lemgram* (och *närliggande lemgram*, se **4.1.5**) sker maskinellt. För varje token identifierar Korp det lemma (den uppslagsform) ur Språkbankens lexikala resurs (Karp) som förefaller vara lämpligaste *lemgram*-kandidat.

Om önskat attribut är *lemgram* noteras parametern **lex** följt av **contains**. Värdet skrivs enligt mönstret **"token\.\.nn\.1"**, där **token** är *lemgrammets namn*, **nn** är *ordklass* (i detta fall substantiv) och **1** är *böjningsmönster* (i detta fall nummer ett). Frågan nedan ger träff på alla token som är annoterade med lemgram *röra* (verb):

```
[lex contains "röra\.\.vb\.1"]
```

Frågan nedan ger träff på alla token som är annoterade med lemgram *röra* (substantiv):

```
[lex contains "röra\.\.nn\.1"]
```

Befintliga lemgram för den korpus/de korpusar du valt finner du i utökad sökning: Välj *lemgram*, skriv ditt ord och vänta tills en meny med valbara lemgram framträder. Menyn dyker bara upp om du skrivit en teckenföljd som motsvarar ett befintligt lemgram - du kan få prova dig fram. Ordklass och böjningsmönstrets nummer samt uppgift om det är ett historiskt lemgram (fornsvenska, 1600-tal, 1800-tal, se **4.1.4.1**) ser du i samma meny. När du valt lemgram finns frågan som vanligt i CQP under fliken *avancerat*.

OBS: Endast korrekt CQP-syntax i kombination med de i Korp befintliga lemgrammen fungerar. Använder du frågorna nedan svarar Korp **antal träffar: 0**:

```
[lex contains "skalle"]
```

```
[lex contains "zkalle\.\.nn\.1"]
```

Alltså inget felmeddelande, och risken är att du felaktigt tror att det du söker inte finns.

OBS: Det händer att token är annoterade med fler än ett lemgram.

Ordklassförkortningarna för ordattributet *lemgram* finner du i **Bilaga 2**. Observera att det inte är samma som ordklassförkortningarna för ordattributen *ordklass* och *homografmängd*.

Hur du gör frågorna mer inkluderande med Regexp ser du i **6.1**.

4.1.4.1 Historiska lemgram

För Korps historiska material finns särskilda lemgram, hämtade från de historiska lexikon som finns i Språkbankens lexikala resurs (Karp). Dessa lemgram noteras på var sina vis enligt nedan.

Från Söderwalls (1884-1918; 1953-1973) och Schlyters (1877) fornsvenska ordböcker hämtas Korps *fornsvenska lemgram*, vilka noteras med **fsvm--** före token enligt följande: **fsvm--dag\.\.nn\.1**. Frågan nedan ger träff på alla token som är annoterade med det fornsvenska lemgrammet *dag (substantiv)*.

```
[lex contains "fsvm--dag\.\.nn\.1"]
```

Från Swedbergs ordbok (ca 1725) hämtas Korps *1600-talslemgram*, vilka noteras med **swedbergm--** före token enligt följande: **swedbergm--dag\.\.nn\.1**. Frågan nedan ger träff på alla token som är annoterade med 1600-talslemgrammet *dag (substantiv)*.

```
[lex contains "swedbergm--dag\.\.nn\.1"]
```

Från Dalins ordbok (1850-1853) hämtas Korps *1800-talslemgram*, vilka noteras med **dalinm--** före token enligt följande: **dalinm--dag\.\.nn\.1**. Frågan nedan ger träff på alla token som är annoterade med 1800-talslemgrammet *dag (substantiv)*.

```
[lex contains "dalinm--dag\.\.nn\.1"]
```

4.1.5 Ordattributet närliggande lemgram

Attributet *närliggande lemgram* används i Korps fornsvenska material. Det är nära kopplat till attributet *lemgram* (**4.1.4**) och kan användas som komplement till detta, bl.a. för att hantera stavningsvariation (**7.4.2**).

Annoteringarna *lemgram* och *närliggande lemgram* sker maskinellt. För varje token identifierar Korp det lemma (den uppslagsform) ur aktuellt lexikon som förefaller vara lämpligaste *lemgram*-kandidat. På grund av stavningsvariationen i äldre material hittas inte alltid rätt ordform i lexikonet. För att i någon mån kompensera detta identifierar Korp, med hjälp av en särskilt framtagen systematisk modell (Adesam, Ahlberg, Bouma 2012), former som har ortografiska likheter med den form som identifierats som lemgram. Dessa redovisas som *närliggande lemgram* - för användaren att ta ställning till.

OBS: Maskinannotering innehåller av nödvändighet en del fel. Använd ditt omdöme!

Om önskat attribut är *närliggande lemgram* noteras parametern **variants** följt av **contains**. Se *Ordattributet lemgram* (**4.1.4**) för ytterligare information om syntaxen. Frågan nedan ger träff på alla token som är annoterade med närliggande lemgram *röra* (verb):

```
[variants contains "röra\.\.vb\.1"]
```

Frågan nedan ger träff på alla token som är annoterade med närliggande lemgram fornsvenska *maþir*.

```
[variants contains "fsvm--maþir\.\.nn\.1"]
```

Hur du gör frågorna mer inkluderande med Regexp ser du i **6.1**.

4.1.6 Ordattributen *förled* och *efterled*

Om önskat attribut är *förled* eller *efterled*, noteras parametrarna **prefix** respektive **suffix** följt av **contains**. Värdet skrivs på samma vis som i ordattributet *lemgram* (se **4.1.4**). Frågan nedan ger träffar (i Bloggmix) på bl.a. *vinnarskalle*, *tokskalle*, *döskallar*, *skinnskallarna*.

```
[suffix contains "skalle\.\.nn\.1"]
```

OBS: Alla historiska korpusar är inte taggade med *för-* och *efterled* (se Lilljegren 2018)

4.1.7 Ordattributet *dependensrelation*

Om önskat attribut är *dependensrelation*, noteras det med parametern **deprel**, medan värdet t.ex. kan vara *formellt subjekt*, noterat **FS**. Denna fråga ger träffar på alla token som är annoterade med dependensrelation formellt subjekt:

```
[deprel = "FS"]
```

Denna fråga ger träff på alla förekomster av *det* som är annoterade som formellt subjekt:

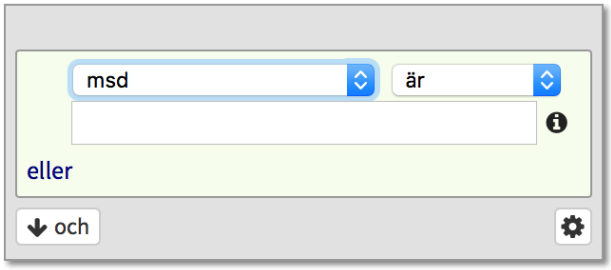
```
[word = "det" & deprel = "FS"]
```

Du hittar förkortningarna för samtliga dependensrelationer i **Bilaga 3**.

4.1.8 Ordattributet *msd*

Förkortningen *msd* står för *morphosyntactic description* (*morfosyntaktisk beskrivning*). Det är en annotering med en kedja av förkortningar (**NN.NEU.SIN.IND.NOM**) som beskriver ordets ordklass och böjning. Du finner förkortningarna i **Bilaga 4** och CQP-syntaxen i **Bilaga 5**. Du kan också använda det speciella hjälpverktyg som finns för *msd* i utökad sökning:

När du väljer ordattributet *msd* dyker en i-märkt knapp upp till höger om sökfönstret (*Figur 12*). När du klickar på den, kan du välja de morfosyntaktiska egenskaper du önskar (*Figur 13*). Byter du sedan till fliken för avancerad sökning ser du som vanligt din fråga som CQP.



Figur 12

Figur 13

Frågan nedan returnerar alla förekomster annoterade som substantiv med genus neutrum, numerus singular, species obestämd, kasus nominativ (observera avbrottstecknet ****, se **3.6**):

```
[msd = "NN\.NEU\.SIN\.IND\.NOM"]
```

Frågan nedan returnerar alla förekomster av substantivet *barn* som är taggade som substantiv med genus neutrum, numerus singular, species obestämd, kasus nominativ.

```
[word = "barn" & msd = "NN\.NEU\.SIN\.IND\.NOM"]
```

De underspecifierade taggarna UTR/NEU, SIN/PLU, IND/DEF, SUB/OBJ som du finner i **Bilaga 4** anges inte med snedstreck i Korp. Istället används plustecken <+>. Detta stöds dock inte för alla ord, och för att minska felrisken rekommenderas att du alltid väljer värde för msd i utökad sökning enligt ovan (*Figur 12 och 13*).

OBS: De omvända snedstrecken är avbrottstecken (**4.1.8**) framför specialtecknet <.> (**3.6**) och skall bara användas i avancerad sökning. Använder du utökad sökning ser msd-annotationerna ur de ovanstående två frågorna ut så:

NN.NEU.SIN.IND.NOM

4.2 Textattribut

4.2.1 Textattributet *titel*

Textattributet *titel* är praktiskt när du är intresserad av en eller ett fåtal texter ur en korpus som rymmer flera. Vilka titlar som finns att välja på hittar du om du väljer *textattribut titel* i utökad sökning. Frågan nedan ger träff på alla förekomster av ordet *huffuudh* i texten *Sancti Joannisoppenbarelse* (som återfinns i korpussamlingen *Yngre religiös prosa*).

```
[word = "huffuudh" & _.text_title = "Sancti Joannisoppenbarilse"]
```

4.2.2 Textattributet *tidsintervall*

Textattributet *tidsintervall* är användbart t.ex. för periodisk litteratur som dagstidningar. Det har en lång och komplex CQP-syntax som du enklast skapar i utökad sökning (genom att välja *tidsintervall* i menyn och välja två datum) och sedan byter till fliken avancerad. Nedanstående fråga, ställd i korpussamlingen *kubhist*, ger alla förekomster av ordet *båda* i tidningar som gavs ut mellan 1:a och 31:a januari 1900:

```
[word = "båda" & ((int(_.text_datefrom) = 19000101 & int(_.text_timefrom) >= 0) | (int(_.text_datefrom) > 19000101 & int(_.text_datefrom) <= 19000131)) & (int(_.text_dateto) < 19000131 | (int(_.text_dateto) = 19000131 & int(_.text_timeto) <= 235959))]
```

4.2.3 Textattributet *sida*

Textattributet *sida* kan vara användbart t.ex. om du letar efter förekomster på en dagstidnings förstasida. Attributet anges med **_.page_no**; värdet anges med heltal. Frågan nedan ger (i korpussamlingen *Kubhist*) alla förekomster av ordet *bassäng* på tidningarnas förstasidor:

```
[word = "bassäng" & _.page_no = "1"]
```

4.3 Fler ord- och textattribut

Det finns ytterligare ord- och textattribut. Du kan undersöka dem via attributmenyn i utökad sökning (se **4**, *Figur 9* och *10*). Menyns innehåll är dynamiskt och och visar endast de attribut som är valbara för aktuell korpus (se även Lilljeberg 2018). För att se hur de ska skrivas i CQP väljer du ett attribut och ett värde, byter till fliken för avancerad sökning och tittar under *Aktiv fråga i utökad*.

OBS: Om du i avancerad sökning skriver ett attribut som inte finns (t.ex. *yrdklass*), eller som inte stöds av den valda korpusen (t.ex. *ordklass* i korpusen *Äldre lagar*) svarar Korp **antal träffar: 0**. Du får alltså inget felmeddelande, och risken är att du felaktigt tror att det du söker inte finns. Var alltså noga med att skriva rätt attribut!

Se även OBS-rutan under **4**.

OBS: Taggningen av Korps texter sker automatiskt, vilket betyder att annoteringsfel förekommer.

5 Sök kombinationer av token

Du kan kombinera i princip hur många token som helst, avgränsade med hakparenteser. Frågan nedan ger träff på alla förekomster av ordföljden *Her byriarz laghbok væsgöta*:

```
[word = "Her"] [word = "byriarz"] [word = "laghbok"] [word = "væsgöta"]
```

Frågan nedan får träff på alla kombinationer av två ord där det första är *kör* och det andra är annoterat som verb, t.ex. *kör sjunger*:

```
[word = "kör"] [pos = "VB"]
```

Önskar du träff på vilket ord som helst (*matchall*) skriver du hakparenteserna utan någonting mellan, såhär:

```
[]
```

Frågan nedan ger träffar på alla följder om fyra ord där det första är *kör*, det andra och tredje är vilket ord som helst (*matchall*) och det fjärde är annoterat som verb:

```
[word = "kör"] [] [] [pos = "VB"]
```

I kombination med de tekniker som redovisats i tidigare avsnitt är sökningar med multipla token ett kraftfullt verktyg. Se exempel under **6** och **7** nedan.

6 Fler sätt att använda Regexp

Under avsnitten **3.1** till **3.7** ovan beskrevs hur Regexp används i söksträngar för enstaka token, t.ex. **[word = "mand?[sz].*ap"]**. Det sättet att använda Regexp kräver egentligen inte att du använder CQP i avancerad sökning; det fungerar i utökad sökning genom att du väljer *regexp* eller *ej regexp* i rullgardinsmenyn (se **2.1**). Här följer en del andra sätt att använda Regexpnotering, varav flera kräver att du skriver frågan i CQP i avancerad sökning.

6.1 Regexp i värdet när parametern är ett ord-/textattribut

När du arbetar med ord- eller textattribut kan deras värden manipuleras med Regexp. Här följer några exempel.

Frågan nedan ger träffar på alla token vars ordklassannotering (**4.1.1**) består av två tecken och börjar med **R**, alltså **RG** (*grundtal*) och **RO** (*ordningstal*):

```
[pos = "R."]
```

Nedanstående fråga ger token vars ordklassannotering är **JJ** (adjektiv) eller **AB** (adverb):

```
[pos = "JJ|AB"]
```

Nedanstående fråga ger träff på token som är annoterade som lemgrammen (**4.1.4**) *Svea* (egennamn) och *Sven* (egennamn) (**pm** anger att det är egennamn som söks).

```
[lex contains "Sve.\.\.pm\.1"]
```

Lägger du till ett <*> efter <.>, som i frågan nedan, får du träffar på alla token som är annoterade som egennamnslemgram och börjar på *Sve*, alltså förutom *Svea* och *Sven* även t.ex. *Sverker*, *Sverige* och *Svedala*.

```
[lex contains "Sve.*\.\.pm\.1"]
```

Nedanstående fråga använder parenteser och lodstreck, och ger träff på token som är annoterade med lemgram *röra (verb)*, *röra (substantiv 1)* och *röra (substantiv 2)*. Det finns inget lemgram *röra (verb 2)* i Korp, men hade det funnits skulle även det genererat träffar.

```
[lex contains "röra\.\.(vb|nn)\.(1|2)"]
```

6.2 Lodstreck mellan parametrar

Lodstreck <|> kan användas mellan två parameter-/värdepar. Frågan nedan ger träff på alla token som har teckenföljden *låg* ELLER är annoterat som adjektiv:

```
[word = "låg" | pos = "JJ"]
```

Notera skillnaden mot nedanstående fråga som ger träff på alla token som har teckenföljden *låg* OCH är annoterat som adjektiv:

```
[word = "låg" & pos = "JJ"]
```

6.3 Regexp över tokennivå

6.3.1 Lodstreck

Lodstreck kan även användas mellan tokenalternativ. Frågan nedan ger träff på alla token med teckenföljden *gitarr* eller teckenföljden *luta*:

```
[word = "gitarr"] | [word = "luta"]
```

alltså samma resultat som nedanstående fråga (se **3.1**):

```
[word = "gitarr|luta"]
```

Mer meningsfullt att använda lodstrecket på detta vis blir det t.ex. om multipla parametrar ska användas endast för det ena ordet:

```
[word = "gitarr"] | [word = "luta" & pos != "VB"]
```

6.3.2 Operatorer och parenteser

Operatorerna för upprepning samt klammerparentes och parentes (**3.1**-**3.3**) kan användas ovanför tokennivå. Frågan nedan använder <**?**>, som gör kommatecknet optionellt, och ger således träff på *ack huru lycklig* och *ack, huru lycklig*:

```
[word = "ack"] [word = ","]? [word = "huru"] [word = "lycklig"]
```

I frågan nedan används parenteser runt två token, vilket gör att <**?**> avser dem båda. Frågan ger träffar på *ack, huru lycklig* och *huru lycklig*:

```
([word = "ack"] [word = ","])? [word = "huru"] [word = "lycklig"]
```

Frågan nedan använder <*****> och ger träffar på *huru lycklig, ack huru lycklig, ack ack huru lycklig, ack ack ack huru lycklig* (etc., med hur många *ack* som helst):

```
[word = "ack"]* [word = "huru"] [word = "lycklig"]
```

Frågan nedan använder <**+**> och ger träffar på *ack huru lycklig, ack ack huru lycklig, ack ack ack huru lycklig* (etc., med hur många *ack* som helst):

```
[word = "ack"]+ [word = "huru"] [word = "lycklig"]
```

Frågan nedan använder klammerparenteser **{}** och ger träffar endast på *ack huru lycklig* och *ack ack huru lycklig*:

```
[word = "ack"]{1, 2} [word = "huru"] [word = "lycklig"]
```

7 Tillämpningsexempel

Med de tekniker som redogjorts för ovan har användaren stora möjligheter att skräddarsy sina egna frågor. Kombinationsmöjligheterna är för omfattande för en systematisk framställning. För att ge en vink om vad som kan göras har här samlats ett fåtal exempel.

7.1 Exemplet ordföljd - SVO

Om du önskar träffar på alla sekvenser om tre ord med satsledsföljden *subjekt*, *predikatsverb*, *direkt objekt* kan du göra på flera vis.

7.1.1 Sök med ordattribut dependensrelation

Förkortningarna för dependensrelationerna *formellt subjekt* och *subjekt (övrigt subjekt)* är **FS** respektive **SS**. Förkortning för *finit verb (predikatsverb)* är **FV**. Förkortningen för *direkt objekt* är **OO** (se alla förkortningar för ordattributet dependensrelation i **Bilaga 3**). Frågan nedan ger träff på alla sekvenser om tre token som är annoterade med i tur och ordning dependensrelationerna *subjekt (formellt subjekt* eller *subjekt [övrigt subjekt])*, *predikatsverb* och *direkt objekt*.

```
[deprel = "FS|SS"] [deprel = "FV"] [deprel = "OO"]
```

Beroende på hur väl Korps automatiska annotering lyckats, kan detta sätt att söka generera förvånansvärt få träffar, varav en relativt stor andel dessutom är felannoterade. Alla öppna korpusar under rubriken *moderna* (224 korpusar, över 13 miljarder token) gav endast 444 träffar (18-04-07). Alla korpusar under Kubhist (82 korpusar, över 1 miljard token) gav 113 träffar (18-04-25).

OBS: Alla korpusar har inte alla attribut. Exempelvis saknar flera historiska korpusar just attributet *dependensrelation*.

7.1.2 Sök med ordattribut ordklass

Förkortningarna för ordklass substantiv är NN (se **Bilaga 1**). Förkortning för verb är VB. Frågan nedan ger träffar på alla sekvenser om tre token som är annoterade i tur och ordning med ordklass *substantiv*, *verb*, *substantiv*:

```
[pos = "NN"] [pos = "VB"] [pos = "NN"]
```

Detta sätt att söka kan tvärtom generera stora mängder träffar, vilket kan bli ohanterligt om de ska gås igenom manuellt. De öppna korpusarna under *moderna* gav 31 255 618 träffar (18-04-07). Ordklassannoteringen har dock generellt färre felannoteringar än dependensrelationtaggningen.

OBS: Alla korpusar har inte alla attribut. Exempelvis saknar flera historiska korpusar just attributet *ordklass* och använder istället attributet *homografmängd* (**4.1.2**).

7.1.3 Sök med blandad teknik

Som vi konstaterat ovan kan en teknik generera få/fel träffar, medan en annan kan generera stora mängder. För den som arbetar kvantitativt kan det förstnämnda vara förödande; för den som arbetar kvalitativt kan det sistnämnda vara alltför tidsödande. Beroende på vad du söker kan du välja den ena eller andra tekniken. Du kan också blanda tekniker. Frågan nedan ger träff på alla sekvenser om tre token där det första är annoterat som subjekt, det andra är ordet *äter* och det tredje är annoterat som substantiv:

```
[deprel = "FS|SS"] [word = "äter"] [pos = "NN"]
```

Frågan ovan gav 216 233 träffar från de öppna korpusarna under *moderna* (18-04-10). Om vi antar att användaren i detta fall inte är intresserad av relativbisatser med inledande *som*, kan frågan effektiviseras ytterligare enligt nedan:

```
[deprel = "FS|SS" & word != "som"] [word = "äter"] [pos = "NN"]
```

Frågan ovan gav 192 575 träffar från de öppna korpusarna under *moderna* (18-04-10). Om det vi söker är endast fullständiga satser om tre ord med inledande versal och avslutande punkt, frågetecken eller utropstecken kan vi lägga till Regexp för detta:

```
[deprel = "FS|SS" & word = "[A-ZÅÄÖ].*" & word != "som"] [word = "äter"] [pos = "NN"]
[word = "[\.\?!]"]
```

Frågan ovan gav 697 träffar från de öppna korpusarna under *moderna* (18-04-10). (Observera hur <.> och <?> måste föregås av avbrottstecknet <\>, se **3.6**)

På detta vis kan en fråga "slipas" för att generera större andel relevanta träffar.

OBS: Du har själv ansvaret att ta hänsyn till begränsningar i Korp. Om du arbetar kvantitativt är det av största vikt att du tar reda på om alla de korpusar du valt verkligen är annoterade med de attribut du vill inkludera i frågan (se OBS-rutan under **4**) och att du bildar dig en uppfattning om i vilken grad felannoteringar kan snedvrida ditt resultat. Jämför **7.5**.

7.2 Exemplet pseudosamordning

En användare söker i ett historiskt material efter exempel på pseudosamordning med *gå och [verb]*.[4] En enkel fråga för detta ändamål kan konstrueras som nedan:

```
[word = "gå"] [word="och"] [pos = "VB"]
```

Frågan ovan missar dock ett antal förekomster pga stavningsvariation. Frågan anpassas efter variationen och blir som nedan:

```
[word = "gå|ga|gaa"] [word="och|ok|oc|åck|okk|ock|og|åg|å"] [pos = "VB"]
```

[4] Tack till Peter Andersson och Kristian Blensenius för detta autentiska exempel. Författaren har här tagit sig friheten att fabulera ett narrativ kring hur frågan kan ha växt fram.

För att få med flera tempus och modus modifieras frågan ytterligare enligt nedan. Notera hur *gå* och *gaa* "lämnas öppna" finalt med hjälp av matchall <.> och operatorn <*>, vilket inkluderar även former som *gått*, *gånga* etc. Dock inkluderas därmed samtidigt ovidkommande träffar som *gåfvor*, *gårdagen* och *gårdsinspektor*, så det är en avvägning mellan att konstruera inkluderande eller begränsande:

```
[word = "gå.*|ga|gaa.*|gick|gingo"] [word="och|ok|oc|åck|okk|ock|og|åg|å"] [pos = "VB"]
```

För att få med varianten *gå att [verb]* läggs *at* och *att* till frågan, enligt nedan:

```
[word = "gå.*|ga|gaa.*|gick|gingo"] [word="och|ok|oc|åck|okk|ock|og|åg|å|at|att"] [pos = "VB"]
```

Till sist läggs ordklassannotering **VB** till frågans första token och noll till två *matchall*-ord för att tillåta adverbial ('*han går* ***nu sannolikt*** *och funderar*'). Den färdiga frågan ser således ut som nedan:

```
[word = "gå.*|ga|gaa.*|gick|gingo" & pos = "VB"] []{0,2}
[word="och|ok|oc|åck|okk|ock|og|åg|å|at|att"] [pos = "VB"]
```

7.3 Exemplet kollokationer

Korp med CQP kan användas för att ta reda på vilka ord som är vanliga i närheten av andra ord. Vill du t.ex. veta vilka predikatsverb som är vanliga i samband med objektet *en paus* kan nedanstående enkla fråga användas (ger 6736 träffar i Bloggmix 18-04-17):

```
[] [word = "en"] [word = "paus"]
```

Ovanstående fråga har dock begränsningar. Eftersom predikatsverbet inte alltid står precis före objektet genereras "skräp" av typen *jag en paus*, *faktiskt en paus* etc. Du kan rensa bort sådana träffar genom att ange att första ordet ska vara taggat som verb, som i frågan nedan (ger 4750 träffar i Bloggmix 18-04-17):

```
[pos = "VB"] [word = "en"] [word = "paus"]
```

Om du vill tillåta att predikatsverbet står längre ifrån nominalfrasen kan du lägga in noll eller flera valfria ord, som i frågan nedan (6540 träffar i Bloggmix 18-04-17):

```
[pos = "VB"] []* [word = "en"] [word = "paus"]
```

Frågan nedan begränsar antalet träffar något genom att endast söka efter konstruktioner där det är 0 eller 1 ord mellan verbet och nominalfrasen (5919 träffar i Bloggmix 18-04-17):

```
[pos = "VB"] []{0,1} [word = "en"] [word = "paus"]
```

7.4 Exemplet historisk stavningsvariation

Under avsnitten **3.1** till **3.4** ovan redogjordes för hur stavningsvariation kan hanteras med Regexp. Under **4.1.5** nämndes att även närliggande lemgram kan användas till detta. Här följer ett exempel på vardera tekniken.

7.4.1 med *Regexp*[5]

En användare söker i historiskt material en speciell sorts pronominalkilkonstruktion - där kilen består av förstakonjunkten i en samordning (*'att han köpa ville och äta en semla'*). Användaren är beredd att manuellt granska ett stort antal träffar och vill inte använda ord-attribut pga risken att felannotering snedvrider resultatet (se **7.1**). En fråga med fem token ska konstrueras där nr 1 är en subjunktion, nr 2 ett pronomen i nominativ, nr 3 vilket ord som helst, nr 4 vilket ord som helst och nr 5 en konjunktion. Med hjälp av Söderwalls *Ordbok Öfver svenska medeltids-språket* och testkörningar i *Korp* tas ett antal subjunktioner, pronomen och en konjunktion fram inklusive stavningsvarianter.

Subjunktioner
mäþan, meþan, medhan, mädhan
efte, efter, eftir, epte, epter, epti, eptir, äfte, äfter, äfti, äftir, äftyr, äpte, äpter, äpti, äptir
utan, uthan, vtan, vthan, wtan, wthan
at, ath, att

Pronomen
iak, iag, iagh, jak, jag, jagh
tu, thu, þu
han, hon, hun
wi, wij, ui
the, þe, thi, þi

Konjunktion
ok, oc, och, ock

Subjunktionerna kan var och en förkortas med Regexp:

```
m[äe](þ|th)an
[eä][fp]t[eiy]r?
[uvw]th?an
at[ht]?
```

Subjunktionerna sätts sedan inom var sina parenteser och ställs efter varandra med lodstreck emellan för att markera att de är sinsemellan utbytbara:

```
(m[äe](þ|th)an)|([eä][fp]t[eiy]r?)|([uvw]th?an)|(at[ht]?)
```

[5] Tack till Erik Magnusson Petzell för detta autentiska exempel, inklusive det konstruerade exemplet med semlan. För framställningens skull har författaren tagit sig två friheter: dels reducerat antalet ord/ordformer, dels skapat en CQP-fråga för avancerad sökning av den ursprungliga frågan som var för utökad sökning.

Samma sak tillämpas på pronomina; först förkortas de med Regexp:

```
[ij]a(k|gh?)
[tþ]h?u
h[aou]n
[uw]ij?
(þ|th)[ei]
```

Sedan sätts även pronomina inom var sina parenteser och ställs efter varandra med lodstreck emellan för att markera att de är sinsemellan utbytbara:

```
([ij]a(k|gh?))|([tþ]h?u)|(h[aou]n)|([uw]ij?)|((þ|th)[ei])
```

Slutligen förkortas den enda konjunktionen:

```
o(k|c[hk]?)
```

Subjunktionerna, pronomina och konjunktionen placeras i var sin token och två tomma token (*matchall*) tillfogas. CQP-frågan är därmed klar:[6]

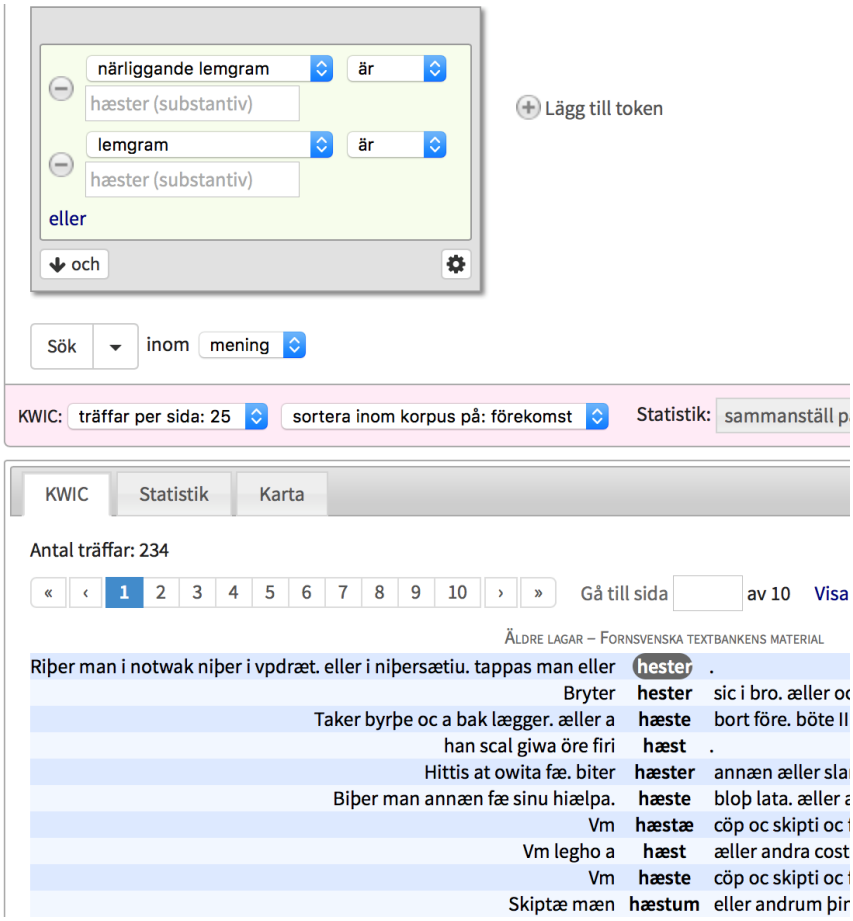
```
[word = "(m[äe](þ|th)an)|([eä][fþ]t[eiy]r?)|([uvw]th?an)|(at[ht]?)"]
[word = "([ij]a(k|gh?))|([tþ]h?u)|(h[aou]n)|([uw]ij?)|((þ|th)[ei])"]
[] []
[word = "o(k|c[hk]?)"]
```

7.4.2 med närliggande lemgram

Det precisaste sättet att hitta de ord du söker, inklusive stavningsvariation, är att använda *Regexp* (**7.4.1**), eftersom det låter dig inkludera/exkludera exakt de former du önskar. Du måste då ha kunskap om böjningsmorfem och stavningsvariation. Saknar du sådan kunskap, eller om lika hög precision inte är nödvändig, kan du istället använda en kombination av *lemgram* (**4.1.4**) och *närliggande lemgram* (**4.1.5**).

I utökad sökning, välj *ordattribut lemgram* och testa stavningar tills du hittar ett lemgram som passar din sökning. Klicka sedan på *'eller'*, välj *ordattribut närliggande lemgram* och ange där samma värde som du nyss angivit under *lemgram*. Nedanstående fråga *(Figur 14)* ger träffar på alla ord som är annoterade med *fornsvenskt lemgram hæster* eller *fornsvenskt närliggande lemgram hæster*.

[6] I framställningen ovan har CQP-frågan radbrytningar för överskådlighetens skull. Radbrytningarna får, men måste inte användas i Korp (**3.9**).



Figur 14

Samma fråga ser med CQP ut enligt nedan (för syntax, se **4.1.4** och **4.1.5**).

[(lex contains "fsvm--hæster\.\.nn\.1" | variants contains "fsvm--hæster\.\.nn\.1")]

Skrivna med CQP kan frågor av detta slag naturligtvis manipuleras med Regexp i värdet (**6.1**) och Regexp ovanför tokennivå (**6.3**).

På ovan beskrivna vis hittar du stavningsvarianter och olika böjningar av ett och samma ord. I idealfallet räcker det med att du söker med ett enda värde (t.ex *hæster*, som i fallet ovan) för att Korp ska klara av att hitta alla ortografiskt närliggande lemgram. I praktiken finns dock brister; i fallet med *Hæster* så hittar Korp t.ex. (18-06-01)[7] det närliggande lemgrammet *Hester* men inte *Häster*. Vill du öka chansen att din fråga returnerar allt du önskar måste du alltså ange fler lemgram/närliggande lemgram. Det förutsätter kunskap om stavningsvariation, och ju mer du kan om stavningsvariation desto mindre intressant blir ju denna teknik jämfört med Regexp (**3.1**, **3.4**, **7.4.1**).

OBS: Du har själv ansvaret att ta hänsyn till begränsningar i Korps annoteringar och se till att de inte förvanskar ditt resultat.

[7] Korp uppdateras kontinuerligt och i skrivande stund (2018) finns planer på förbättrande av just denna funktion.

7.5 Exemplet OCR[8]

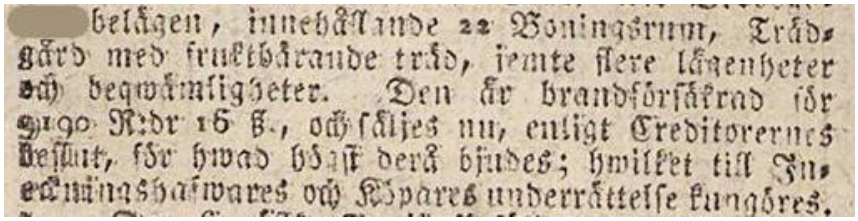
Många texter i Korp har tagits fram med hjälp av tekniken OCR (Optical character recognition).[9] Feltolkningar vid OCR-skanning är relativt vanliga. Har den tryckta förlagan fläckar, ojämnt tryck eller frakturstil kan felen förändra texten till oigenkännlighet.

OBS: Du har själv ansvaret att ta hänsyn till textkvalitetsbegränsningar i Korp. Om du arbetar kvantitativt bör du bilda dig en uppfattning om i vilken grad OCR-fel kan påverka ditt resultat. Jämför **7.1**. För uppgifter om hur enskilda historiska korpusar har digitaliserats, korrekturlästs etc., se Joakim Lilljegrens *Introduktion till Språkbankens historiska material i Korp* (2018).

Om du arbetar kvalitativt - dvs snarare är beroende av ett fungerande autentiskt exempel än av statistik - kan du prova att ta hjälp av Regexp när texten har OCR-relaterade problem. Här kommer två exempel, ett där den OCR:ade texten får anses oanvändbar (**7.5.1**) och ett där den är i tillräckligt gott skick för att en sådan regexpteknik kan vara värd att överväga (**7.5.2**).

7.5.1 Dagstidning med frakturstil

I Post och inrikes tidningar 12 maj 1821 fanns på sid 4 detta:


belägen, innehållande 22 Boningsrum, Trädgård med fruktbärande träd, jemte flere lägenheter och bequämligheter. Den är brandförsäkrad för 9190 R:dr 16 ß., och säljes nu, enligt Creditorernes beslut, för hwad höst derå bjudes; hwilket till Inrednings-hafwares och Köpares underrättelse kungöres.

Figur 15

Texten har OCR:ats och ser i Korp ut så:

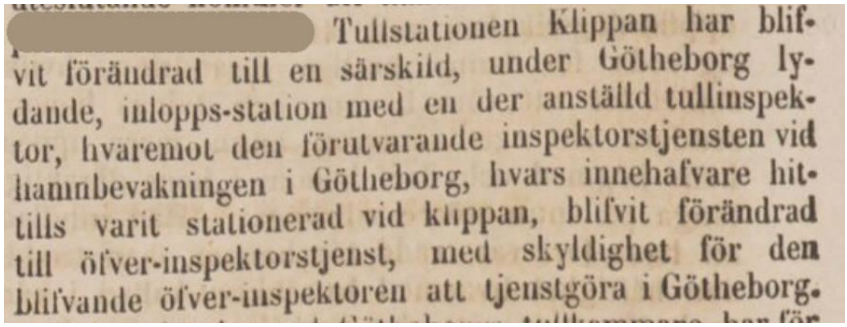
> belägen iuueblNiNde 2» V^ ' i ' N< ' !irnnl , Tiäd » gärd med ' ri !' ktdä . ' aiide träi ;, !>>i !lte ffere lägenheter » ch^ beqwämlig ', eter . De » 5r brandfursäkrad ilr «Z>^ R : dr >6 ss, , ochsä !>e4 nn , enligt Ereditorerncs iikffnl , löe hwad hö.st d^.ä bfn^es ; hwiltet till In » » snnng^halwaws och Köparet underrättelse kulöres .

[8] Exemplen är från korpussamlingen *Kubhist*. Denna omfattande samling med historiska dagstidningskorpusar från Kungliga Biblioteket skall dock (sannolikt under 2018) uppdateras med nya OCR-inläsningar.
[9] T.ex. *Kubhist*, och vissa andra historiska korpusar och korpussamlingar. För information om till vilka historiska korpusar OCR använts, se Lilljegren (2018).

Problemen är som synes omfattande, eventuellt beroende både på frakturstilen och tryckkvaliteten. OCR-tekniken har här sannolikt anpassats till det aktuella typsnittet och därmed bl.a. lyckats göra skillnad på gemena <f> och gemena 'långa <s>' i *brandfursäkrad*. Trots det är texten svåranvänd då feltolkningarna är frekventa och inte följer tydliga mönster. Den OCR:ade textens tecken <^> motsvarar t.ex. såväl originalets <2>, avstavningstecken, <o>, <n>, <9> som <e>. Den OCR:ade textens <u>, <N>, <'>, <ii>, <»> och <l> kan alla motsvara originalets <n>.

7.5.2 Dagstidning med antikva

I Post och inrikes tidningar 6 april 1864 fanns på sid 2 detta att läsa:


Tullstationen Klippan har blif-
vit förändrad till en särskild, under Götheborg ly-
dande, inlopps-station med en der anställd tullinspek-
tor, hvaremot den förutvarande inspektorstjensten vid
hamnbevakningen i Götheborg, hvars innehafvare hit-
tills varit stationerad vid klippan, blifvit förändrad
till öfver-inspektorstjenst, med skyldighet för den
blifvande öfver-inspektoren att tjenstgöra i Götheborg.

Figur 16

Texten har OCR:ats och ser i Korp ut så:

> Tullstationen Klippan har blifvit förändrad Ull en särskild , under Obtheborg lydande , inlopps-station med eu der anställd tullinspektor , bvaremoi den förutvarande mspeklorsljensten vid hamnbevakniiigen i Gölbeborg , hvars iuuebafvare hittills vant slalionerad vid klippan , blifvit förändrad till öfver-mspeklorstjensl , med skyldighet för den bhfvande öfver-inspekloren all tjenstgöra i Gölbeborg .

OCR har gjort dessa feltolkningar:

till	->	Ull
Götheborg	->	Obtheborg
en	->	eu
hvaremot	->	bvaremoi
inspektorstjensten	->	mspeklorsljensten
hamnbevakningen	->	hamnbevakniiigen
Götheborg	->	Gölbeborg
innehafvare	->	iuuebafvare
varit	->	vant
stationerad	->	slalionerad
inspektorstjenst	->	mspeklorstjensl
blifvande	->	bhfvande
inspektoren	->	inspekloren
att	->	all
Götheborg	->	Gölbeborg

Problemen är färre än i frakturexemplet (**7.5.1**). Felen inskränker sig till listan till vänster. Vi ser vissa mönster: <n> blir <u>, <t> blir <l> eller <i>, <h> blir <b>. Det finns också en tendens att två tolkas som en: <ti> blir <U>, <in> blir <m>, <ri> blir <n>, <li> blir <h>. Även motsatsen: <n> har i ett fall tolkats som <ii>.

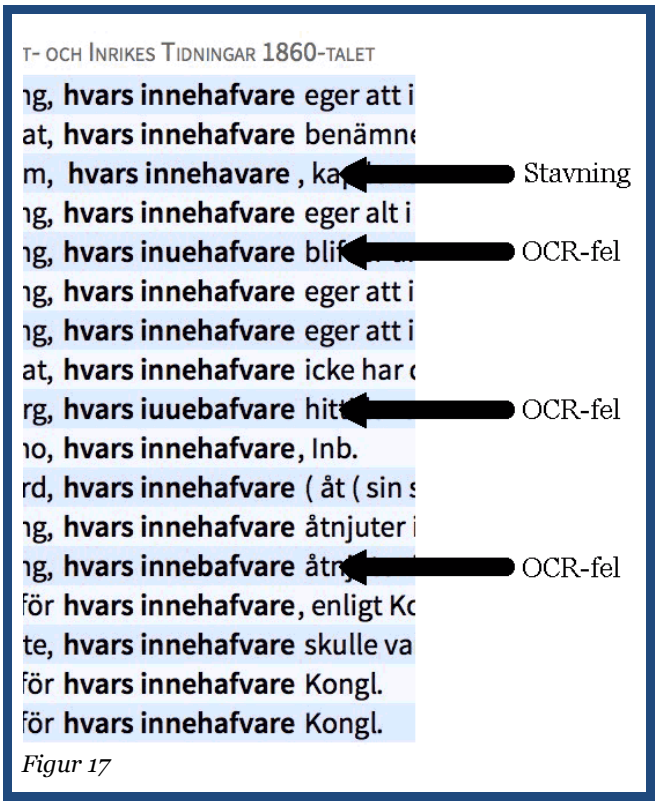
Iakttagelserna kan nyttjas i en OCR-fel-tålig fråga. Frågan nedan är konstruerad utan hänsyn till eventuella OCR-fel. Den hittar 13 förekomster i Post och Inrikes tidningar 1860-tal, men inte förekomsten i avsnittet från tidningen den 6 april 1864 ovan.

[word = "hvars"] [word = "innehafvare"]

Använder vi nedanstående fråga, hittar vi dock förekomsten. Frågan accepterar <*u*> för <*n*> och <*b*> för <*h*>. Operatorn **?** har dessutom lagts till efter <f> för att acceptera stavning med endast <v>.

```
[word = "hvars"]
[word = "i[nu][nu]e[hb]af?vare"]
```

Frågan gav (2018-05-06) 17 träffar i *Post och inrikes tidningar 1860-tal* i korpussamlingen *kubhist*. Av de 17 träffarna (*Figur 17*) är tre uppenbart fel-OCR:ade och en är stavad med <*v*> istället för <*fv*>. Jämfört med den första frågan generar den alltså ytterligare fyra användbara träffar, vilket kan vara till hjälp för den som letar enstaka exempel.



Figur 17

Göteborg juni 2018

Källor

Adesam, Ahlberg, Bouma (2012). *bokstaffua, bokstaffwa, bokstafwa, bokstaua, bokstawa. . . : Towards lexical link-up for a corpus of Old Swedish*. Proceedings of KONVENS 2012 (LThist 2012 workshop), Vienna, September 21, 2012

Användarhandledning (Korp)
https://spraakbanken.gu.se/swe/forskning/infrastruktur/korp/anvandarhandledning
Hämtad 18-04-06.

Borin, Forsberg & Roxendal (2012). *Korp – the corpus infrastructure of Språkbanken*. Proceedings of LREC 2012. Istanbul: ELRA, pages 474–478.

Dalin, A.F. (1850-1853). *Ordbok öfver svenska språket*. Stockholm.

Evert, Stefan & The CWB Development Team (2016). *The IMS Open Corpus Workbench (CWB) CQP Query Language Tutorial*. CWB Version 3.4.
http://cwb.sourceforge.net/files/CQP_Tutorial.pdf

Karp = Språkbankens lexikala resurs. https://spraakbanken.gu.se/karp

Lilljegren, Joakim (2018). *Introduktion till Språkbankens historiska material i Korp*. Göteborgs universitet.
https://meijerbergs.hum.gu.se/digitalAssets/1694/1694144_introduktion.pdf

PCRE (*Perl Compatible Regular Expressions*). https://www.pcre.org

Schlyter C.J. (1877). *Ordbok till Samlingen af Sweriges Gamla Lagar, Saml. af Sweriges Gamla Lagar 13*. (Lund)

Swedberg, Jesper (ca 1725). *Swensk ordabok*. Utgiven postumt 2008. Järpås : Skara stifts-historiska sällskap.

Söderwall, K.F. (1884-1918). *Ordbok Öfver svenska medeltids-språket. Vol I-III*. Lund.

Söderwall, K.F. (1953-1973). *Ordbok Öfver svenska medeltids-språket. Suppl. Vol IV—V*. Lund.

Bilaga 1 - förkortningar, *ordklass* i *ordklass*

AB	Adverb
DT	Determinerare, bestämningsord
HA	Frågande/relativt adverb
HD	Frågande/relativ bestämning
HP	Frågande/relativt pronomen
HS	Frågande/relativt possessivuttryck
IE	Infinitivmärke
IN	Interjektion
JJ	Adjektiv
KN	Konjunktion
NN	Substantiv
PC	Particip
PL	Partikel
PM	Egennamn
PN	Pronomen
PP	Preposition
PS	Possessivuttryck
RG	Räkneord: grundtal
RO	Räkneord: ordningstal
SN	Subjunktion
UO	Utländskt ord
VB	Verb
MAD	Meningsskiljande interpunktion
MID	Interpunktion
PAD	Interpunktion

Dessa förkortningar gäller när du söker på *ordattributen ordklass* (**4.1.1**) och *homografmängd* (**4.1.2**). Ordklassförkortningar för *ordattributet lemgram* (**4.1.4**) hittar du i **Bilaga 2**.

Data till denna tabell hämtad 18-04-20 från

https://spraakbanken.gu.se/

Bilaga 2 - förkortningar, *ordklass* i *lemgram*

nn	substantiv
av	adjektiv
vb	verb
pm	egennamn
ab	adverb
in	interjektion
pp	preposition
nl	numeral
pn	pronomen
sn	subjunktion
kn	konjunktion
al	artikel
ie	infinitivmärke
mxc	flerordsprefix
sxc	prefix
abh	adverbsuffix
avh	adjektivsuffix
nnh	substantivsuffix
nnm	substantiv, flerording
nna	substantiv, förkortning
avm	adjektiv, flerording
ava	adjektiv, förkortning
vbm	verb, flerording
vba	verb, förkortning
pmm	egennamn, flerording
pma	egennamn, förkortning
abm	adverb, flerording
aba	adverb, förkortning
pnm	pronomen, flerording
inm	interjektion, flerording
ppm	preposition, flerording
ppa	preposition, förkortning
nlm	numeral, flerording
knm	konjunktion, flerording
snm	subjunktion, flerording
kna	konjunktion, förkortning
ssm	flerording, sats
e	*ordklass saknas*

Förkortningarna hämtas från Språkbankens lexikon SALDO. För det äldre textmaterialet använder Korp dock även äldre lexikon som t.ex. Dalin och Swedberg. Därifrån kommer förkortningen ”e” som betyder ”ordklass saknas”. I det material som 2018-04-24 ligger under *äldre fornsvenska*, *yngre fornsvenska*, *profan prosa*, *verser* och *medeltidsbrev svenska* är 193 102 av materialets 4 058 032 token (4,8 %) enbart taggade med (e). Dessa ordklassförkortningar gäller när du söker på ordattributet lemgram (**4.1.4**). Ordklassförkortningar för attributen ordklass (**4.1.1**) och homografmängd (**4.1.2**) finns i **Bilaga 1**.

Data (förutom *e - ordklass saknas*) till denna tabell hämtad 18-04-20 från

https://spraakbanken.gu.se/swe/forskning/saldo/taggm%C3%A4ngd

Bilaga 3 - förkortningar, *dependensrelation*

Tagg	Betydelse	Kommentar
++	Samordnande konjunktion	skrivs **\+\+** i avancerad sökning[10]
+A	Konjunktionellt adverb	skrivs **\+A** i avancerad sökning
+F	Koordination på huvudsatsnivå	skrivs **\+F** i avancerad sökning
AA	Annat adverbial	
AG	Agent	
AN	Apposition	
AT	Framförställt attribut	
CA	Kontrastivt adverbial	
CJ	Samordnat led	
DB	Dubbel funktion	
DT	Determinerare, bestämningsord	
EF	Relativ bisats	
EO	Egentligt objekt	
ES	Egentligt subjekt	
ET	Efterställd bestämning	
FO	Formellt objekt	
FP	Fritt subjektivt predikativ (predikatsfyllnad)	
FS	Formellt subjekt	
FV	Finit verb, predikatsverb	
HD	Huvud	
I?	Frågetecken	skrivs **I\?** i avancerad sökning
IC	Citattecken	
IF	Infinitivfras, utom infinitivmärke	
IG	Övrig interpunktion	
IK	Kommatecken	
IM	Infinitivmärke	
IO	Indirekt objekt (dativobjekt)	
IP	Punkt	
IQ	Kolon	
IR	Parentes	
IS	Semikolon	
IT	Divis, bindestreck	
IU	Utropstecken	
IV	Infinit verb	
JC	Citattecken 2	
JG	Övrig interpunktion 2	
JR	Parentes 2	
JT	Divis 2, bindestreck 2	
KA	Komparativt adverbial	
MA	Satsadverbial	
MS	Makrosyntagm	
NA	Negerande adverbial	
OA	Objektsadverbial (prepositionsobjekt)	
OO	Direkt objekt (ackusativobjekt)	
OP	Objektspredikativ (objektiv predikatsfyllnad)	

[10] Se **3.6** - Att söka på specialtecknen

PA	Prepositions komplement	
PL	Verbpartikel	
PR	Preposition	
PT	Predikativt attribut	
RA	Platsadverbial	
ROOT	*Rot*	
SP	Subjektspredikativ (subjektiv predikatsfyllnad)	
SS	Subjekt (övrigt subjekt)	
TA	Tidsadverbial	
TT	Tilltalsfras	
UA	Underordnad sats (bisats), utom subjunktion	
UK	Subjunktion	
VA	Korrelativt adverbial	
VG	Verbgrupp	
VO	Objekt med infinitiv	
VS	Subjekt med infinitiv	
XA	Uttryck som "så att säga"	
XF	Fundamentsfras	
XT	Uttryck som "så kallad"	
XX	Oklassificerbar satsfunktion	
YY	Interjektionsfras	

Data till denna tabell hämtad 18-04-18 från

http://stp.lingfil.uu.se/~nivre/swedish_treebank/GF.html

förutom ROOT (Rot), som lagts till av Språkbanken för att markera roten av dependensträdet. Alla övriga ord är uppmärkta med en referens till ett annat ord (dess syntaktiska huvud) och annotationen *deprel* som anger dess relation till detta ord, förutom ordet märkt med ROOT som alltså saknar syntaktiskt huvud.

Bilaga 4 - förkortningar, *msd*

Ordklass

AB	Adverb
DT	Determinerare, bestämningsord
HA	Frågande/relativt adverb
HD	Frågande/relativ bestämning
HP	Frågande/relativt pronomen
HS	Frågande/relativt possessivuttryck
IE	Infinitivmärke
IN	Interjektion
JJ	Adjektiv
KN	Konjunktion
NN	Substantiv
PC	Particip
PL	Partikel
PM	Egennamn
PN	Pronomen
PP	Preposition
PS	Possessivuttryck
RG	Räkneord: grundtal
RO	Räkneord: ordningstal
SN	Subjunktion
UO	Utländskt ord
VB	Verb

Avseende genus

UTR	Utrum
NEU	Neutrum
MAS	Maskulinum
UTR/NEU	Underspecificerat
-	Ospecificerat

Avseende numerus

SIN	Singularis
PLU	Pluralis
SIN/PLU	Underspecificerat
-	Ospecificerat

Avseende bestämdhet

IND	Obestämd form
DEF	Bestämd form
IND/DEF	Underspecificerat
-	Ospecificerat

Avseende substantivform

NOM	Grundform
GEN	Genitiv
SMS	Sammansättning
-	Ospecificerat

Avseende komparation

POS	Positiv
KOM	Komparativ
SUV	Superlativ

Avseende satsdel

SUB	Subjekt
OBJ	Objekt
SUB/OBJ	Underspecificerat

Avseende verbformer

PRS	Presens
PRT	Preteritum (imperfekt)
INF	Infinitiv
SUP	Supinum
IMP	Imperativ
AKT	Aktiv diates
SFO	S-form (passivum, deponens)
KON	Konjunktiv
PRF	Perfekt particip

Övrigt

AN	Förkortning
MAD	Meningsskiljande interpunktion
MID	Interpunktion
PAD	Interpunktion

Data till denna tabell hämtad 18-04-20 från

https://spraakbanken.gu.se/

De underspecifierade taggarna UTR/NEU, SIN/PLU, IND/DEF, SUB/OBJ anges inte med snedstreck i Korps CQP-syntax. Istället används plustecken <+>.

Bilaga 5 - MSD-syntax

adjektiv	
adjektiv förkortning	JJ\.AN
adjektiv komparativ utrum/neutrum singularis/pluralis obestämd/bestämd genitiv	JJ\.KOM\.UTR\+NEU\.SIN\+PLU\.IND\+DEF\.GEN
adjektiv komparativ utrum/neutrum singularis/pluralis obestämd/bestämd nominativ	JJ\.KOM\.UTR\+NEU\.SIN\+PLU\.IND\+DEF\.NOM
adjektiv komparativ utrum/neutrum singularis/pluralis obestämd/bestämd sammansättningsform	JJ\.KOM\.UTR\+NEU\.SIN\+PLU\.IND\+DEF\.SMS
adjektiv positiv maskulinum singularis bestämd genitiv	JJ\.POS\.MAS\.SIN\.DEF\.GEN
adjektiv positiv maskulinum singularis bestämd nominativ	JJ\.POS\.MAS\.SIN\.DEF\.NOM
adjektiv positiv neutrum singularis obestämd genitiv	JJ\.POS\.NEU\.SIN\.IND\.GEN
adjektiv positiv neutrum singularis obestämd nominativ	JJ\.POS\.NEU\.SIN\.IND\.NOM
adjektiv positiv neutrum singularis obestämd/bestämd nominativ	JJ\.POS\.NEU\.SIN\.IND\+DEF\.NOM
adjektiv positiv utrum sammansättningsform	JJ\.POS\.UTR\.-\.-\.SMS
adjektiv positiv utrum singularis obestämd genitiv	JJ\.POS\.UTR\.SIN\.IND\.GEN
adjektiv positiv utrum singularis obestämd nominativ	JJ\.POS\.UTR\.SIN\.IND\.NOM
adjektiv positiv utrum singularis obestämd/bestämd nominativ	JJ\.POS\.UTR\.SIN\.IND\+DEF\.NOM
adjektiv positiv utrum/neutrum pluralis obestämd nominativ	JJ\.POS\.UTR\+NEU\.PLU\.IND\.NOM
adjektiv positiv utrum/neutrum pluralis obestämd/bestämd genitiv	JJ\.POS\.UTR\+NEU\.PLU\.IND\+DEF\.GEN
adjektiv positiv utrum/neutrum pluralis obestämd/bestämd nominativ	JJ\.POS\.UTR\+NEU\.PLU\.IND\+DEF\.NOM
adjektiv positiv utrum/neutrum sammansättningsform	JJ\.POS\.UTR\+NEU\.-\.-\.SMS
adjektiv positiv utrum/neutrum singularis bestämd genitiv	JJ\.POS\.UTR\+NEU\.SIN\.DEF\.GEN
adjektiv positiv utrum/neutrum singularis bestämd nominativ	JJ\.POS\.UTR\+NEU\.SIN\.DEF\.NOM
adjektiv positiv utrum/neutrum singularis/pluralis obestämd nominativ	JJ\.POS\.UTR\+NEU\.SIN\+PLU\.IND\.NOM
adjektiv positiv utrum/neutrum singularis/pluralis obestämd/bestämd nominativ	JJ\.POS\.UTR\+NEU\.SIN\+PLU\.IND\+DEF\.NOM
adjektiv superlativ maskulinum singularis bestämd genitiv	JJ\.SUV\.MAS\.SIN\.DEF\.GEN
adjektiv superlativ maskulinum singularis bestämd nominativ	JJ\.SUV\.MAS\.SIN\.DEF\.NOM
adjektiv superlativ utrum/neutrum pluralis bestämd nominativ	JJ\.SUV\.UTR\+NEU\.PLU\.DEF\.NOM
adjektiv superlativ utrum/neutrum pluralis obestämd nominativ	JJ\.SUV\.UTR\+NEU\.PLU\.IND\.NOM
adjektiv superlativ utrum/neutrum singularis/pluralis bestämd nominativ	JJ\.SUV\.UTR\+NEU\.SIN\+PLU\.DEF\.NOM
adjektiv superlativ utrum/neutrum singularis/pluralis obestämd nominativ	JJ\.SUV\.UTR\+NEU\.SIN\+PLU\.IND\.NOM
adverb	AB
adverb förkortning	AB\.AN
adverb positiv	AB\.POS
adverb komparativ	AB\.KOM
adverb sammansättningsform	AB\.SMS
adverb superlativ	AB\.SUV
determinerare	
determinerare förkortning	DT\.AN
determinerare maskulinum singularis bestämd	DT\.MAS\.SIN\.DEF
determinerare maskulinum singularis obestämd	DT\.MAS\.SIN\.IND
determinerare neutrum singularis bestämd	DT\.NEU\.SIN\.DEF
determinerare neutrum singularis obestämd	DT\.NEU\.SIN\.IND
determinerare neutrum singularis obestämd/bestämd	DT\.NEU\.SIN\.IND\+DEF
determinerare utrum singularis bestämd	DT\.UTR\.SIN\.DEF
determinerare utrum singularis obestämd	DT\.UTR\.SIN\.IND

determinerare utrum singularis obestämd/bestämd	DT\.UTR\.SIN\.IND\+DEF
determinerare utrum/neutrum pluralis bestämd	DT\.UTR\+NEU\.PLU\.DEF
determinerare utrum/neutrum pluralis obestämd	DT\.UTR\+NEU\.PLU\.IND
determinerare utrum/neutrum pluralis obestämd/bestämd	DT\.UTR\+NEU\.PLU\.IND\+DEF
determinerare utrum/neutrum singularis bestämd	DT\.UTR\+NEU\.SIN\.DEF
determinerare utrum/neutrum singularis obestämd	DT\.UTR\+NEU\.SIN\.IND
determinerare utrum/neutrum singularis/pluralis obestämd	DT\.UTR\+NEU\.SIN\+PLU\.IND
egennamn	
egennamn genitiv	PM\.GEN
egennamn nominativ	PM\.NOM
egennamn sammansättningsform	PM\.SMS
fråge-/relativuttryck	
frågande/relativ determinerare neutrum singularis obestämd	HD\.NEU\.SIN\.IND
frågande/relativ determinerare utrum singularis obestämd	HD\.UTR\.SIN\.IND
frågande/relativ determinerare utrum/neutrum pluralis obestämd	HD\.UTR\+NEU\.PLU\.IND
frågande/relativt adverb	HA
frågande/relativt possesivt pronomen bestämd	HS\.DEF
frågande/relativt pronomen neutrum singularis obestämd sammansättningsform	HP\.NEU\.SIN\.IND\.SMS
frågande/relativt pronomen neutrum singularis obestämd	HP\.NEU\.SIN\.IND
frågande/relativt pronomen utrum singularis obestämd	HP\.UTR\.SIN\.IND
frågande/relativt pronomen utrum/neutrum pluralis obestämd	HP\.UTR\+NEU\.PLU\.IND
frågande/relativt pronomen	HP\.-\.-\.-
infinitivmärke	IE
interjektion	IN
interpunktion	
meningsskiljande interpunktion	MAD
annan avskiljande interpunktion	MID
omslutande interpunktion	PAD
konjunktion	KN
konjunktion förkortning	KN\.AN
particip	
particip förkortning	PC\.AN
particip perfekt maskulinum singularis bestämd genitiv	PC\.PRF\.MAS\.SIN\.DEF\.GEN
particip perfekt maskulinum singularis bestämd nominativ	PC\.PRF\.MAS\.SIN\.DEF\.NOM
particip perfekt neutrum singularis obestämd nominativ	PC\.PRF\.NEU\.SIN\.IND\.NOM
particip perfekt utrum singularis obestämd genitiv	PC\.PRF\.UTR\.SIN\.IND\.GEN
particip perfekt utrum singularis obestämd nominativ	PC\.PRF\.UTR\.SIN\.IND\.NOM
particip perfekt utrum/neutrum pluralis obestämd/bestämd genitiv	PC\.PRF\.UTR\+NEU\.PLU\.IND\+DEF\.GEN
particip perfekt utrum/neutrum pluralis obestämd/bestämd nominativ	PC\.PRF\.UTR\+NEU\.PLU\.IND\+DEF\.NOM
particip perfekt utrum/neutrum singularis bestämd genitiv	PC\.PRF\.UTR\+NEU\.SIN\.DEF\.GEN
particip perfekt utrum/neutrum singularis bestämd nominativ	PC\.PRF\.UTR\+NEU\.SIN\.DEF\.NOM
particip presens utrum/neutrum singularis/pluralis obestämd/bestämd genitiv	PC\.PRS\.UTR\+NEU\.SIN\+PLU\.IND\+DEF\.GEN
particip presens utrum/neutrum singularis/pluralis obestämd/bestämd nominativ	PC\.PRS\.UTR\+NEU\.SIN\+PLU\.IND\+DEF\.NOM
partikel	PL
partikel sammansättningsform	PL\.SMS

preposition	PP
preposition förkortning	PP\.AN
preposition sammansättningsform	PP\.SMS
pronomen	
pronomen maskulinum singularis bestämd subjektsform/objektsform	PN\.MAS\.SIN\.DEF\.SUB\+OBJ
pronomen neutrum singularis bestämd subjektsform/objektsform	PN\.NEU\.SIN\.DEF\.SUB\+OBJ
pronomen neutrum singularis obestämd subjektsform/objektsform	PN\.NEU\.SIN\.IND\.SUB\+OBJ
pronomen utrum pluralis bestämd objektsform	PN\.UTR\.PLU\.DEF\.OBJ
pronomen utrum pluralis bestämd subjektsform	PN\.UTR\.PLU\.DEF\.SUB
pronomen utrum singularis bestämd objektsform	PN\.UTR\.SIN\.DEF\.OBJ
pronomen utrum singularis bestämd subjektsform	PN\.UTR\.SIN\.DEF\.SUB
pronomen utrum singularis bestämd subjektsform/objektsform	PN\.UTR\.SIN\.DEF\.SUB\+OBJ
pronomen utrum singularis obestämd subjektsform	PN\.UTR\.SIN\.IND\.SUB
pronomen utrum singularis obestämd subjektsform/objektsform	PN\.UTR\.SIN\.IND\.SUB\+OBJ
pronomen utrum/neutrum pluralis bestämd objektsform	PN\.UTR\+NEU\.PLU\.DEF\.OBJ
pronomen utrum/neutrum pluralis bestämd subjektsform	PN\.UTR\+NEU\.PLU\.DEF\.SUB
pronomen utrum/neutrum pluralis bestämd subjektsform/objektsform	PN\.UTR\+NEU\.PLU\.DEF\.SUB\+OBJ
pronomen utrum/neutrum pluralis obestämd subjektsform/objektsform	PN\.UTR\+NEU\.PLU\.IND\.SUB\+OBJ
pronomen utrum/neutrum singularis/pluralis bestämd objektsform	PN\.UTR\+NEU\.SIN\+PLU\.DEF\.OBJ
possesivt pronomen förkortning	PS\.AN
possesivt pronomen neutrum singularis bestämd	PS\.NEU\.SIN\.DEF
possesivt pronomen utrum singularis bestämd	PS\.UTR\.SIN\.DEF
possesivt pronomen utrum/neutrum pluralis bestämd	PS\.UTR\+NEU\.PLU\.DEF
possesivt pronomen utrum/neutrum singularis/pluralis bestämd	PS\.UTR\+NEU\.SIN\+PLU\.DEF
grundtal / ordningstal	
grundtal genitiv	RG\.GEN
grundtal neutrum singularis obestämd nominativ	RG\.NEU\.SIN\.IND\.NOM
grundtal nominativ	RG\.NOM
grundtal sammansättningsform	RG\.SMS
grundtal singularis bestämd nominativ	RG\.MAS\.SIN\.DEF\.NOM
grundtal utrum singularis obestämd nominativ	RG\.UTR\.SIN\.IND\.NOM
grundtal utrum/neutrum singularis bestämd nominativ	RG\.UTR\+NEU\.SIN\.DEF\.NOM
ordningstal genitiv	RO\.GEN"
ordningstal maskulinum singularis obestämd/bestämd genitiv	RO\.MAS\.SIN\.IND\+DEF\.GEN
ordningstal maskulinum singularis obestämd/bestämd nominativ	RO\.MAS\.SIN\.IND\+DEF\.NOM
ordningstal nominativ	RO\.NOM
ordningstal utrum/neutrum singularis/pluralis obestämd/bestämd sammansättningsform	RO\.UTR\+NEU\.SIN\+PLU\.IND\+DEF\.SMS
subjunktion	SN
substantiv	NN\.-\.-\.-\.-
substantiv förkortning	NN\.AN
substantiv neutrum pluralis bestämd genitiv	NN\.NEU\.PLU\.DEF\.GEN
substantiv neutrum pluralis bestämd nominativ	NN\.NEU\.PLU\.DEF\.NOM
substantiv neutrum pluralis obestämd genitiv	NN\.NEU\.PLU\.IND\.GEN
substantiv neutrum pluralis obestämd nominativ	NN\.NEU\.PLU\.IND\.NOM
substantiv neutrum sammansättningsform	NN\.NEU\.-\.-\.SMS
substantiv neutrum singularis bestämd genitiv	NN\.NEU\.SIN\.DEF\.GEN
substantiv neutrum singularis bestämd nominativ	NN\.NEU\.SIN\.DEF\.NOM

substantiv neutrum singularis obestämd genitiv	NN\.NEU\.SIN\.IND\.GEN
substantiv neutrum singularis obestämd nominativ	NN\.NEU\.SIN\.IND\.NOM
substantiv neutrum	NN\.NEU\.-\.-\.-
substantiv sammansättningsform	NN\.-\.-\.-\.SMS
substantiv utrum pluralis bestämd genitiv	NN\.UTR\.PLU\.DEF\.GEN
substantiv utrum pluralis bestämd nominativ	NN\.UTR\.PLU\.DEF\.NOM
substantiv utrum pluralis obestämd genitiv	NN\.UTR\.PLU\.IND\.GEN
substantiv utrum pluralis obestämd nominativ	NN\.UTR\.PLU\.IND\.NOM
substantiv utrum sammansättningsform	NN\.UTR\.-\.-\.SMS
substantiv utrum singularis bestämd genitiv	NN\.UTR\.SIN\.DEF\.GEN
substantiv utrum singularis bestämd nominativ	NN\.UTR\.SIN\.DEF\.NOM
substantiv utrum singularis obestämd genitiv	NN\.UTR\.SIN\.IND\.GEN
substantiv utrum singularis obestämd nominativ	NN\.UTR\.SIN\.IND\.NOM
substantiv utrum	NN\.UTR\.-\.-\.-
utländskt ord	UO
verb	
verb förkortning	VB\.AN
verb imperativ aktiv	VB\.IMP\.AKT
verb imperativ s-form	VB\.IMP\.SFO
verb infinitiv aktiv	VB\.INF\.AKT
verb infinitiv s-form	VB\.INF\.SFO
verb konjunktiv presens aktiv	VB\.KON\.PRS\.AKT
verb konjunktiv preteritum aktiv	VB\.KON\.PRT\.AKT
verb konjunktiv preteritum s-form	VB\.KON\.PRT\.SFO
verb presens aktiv	VB\.PRS\.AKT
verb presens s-form	VB\.PRS\.SFO
verb preteritum aktiv	VB\.PRT\.AKT
verb preteritum s-form	VB\.PRT\.SFO
verb sammansättningsform	VB\.SMS
verb supinum aktiv	VB\.SUP\.AKT
verb supinum s-form	VB\.SUP\.SFO

Data till denna tabell hämtad 18-04-10 direkt från Korps utökade sökningsgränssnitt. Länken nedan leder till sagda gränssnitt med msd valt i attributmenyn för korpusen *Litteraturbanken*:

https://spraakbanken.gu.se/korp/?mode=lb#?lang=sv&stats_reduce=word&cqp=%5Bmsd%20%3D%20%22%22%5D&search_tab=1