# Korpusværktøjet CoREST
# Manual for **STANDARDUDGAVEN**

## Version 2020

Jørg Asmussen
Det Danske Sprog- og Litteraturselskab
9. maj 2020

### Om manualen

Denne manual beskriver **Standard-udgaven** af CoREST, som alle sprog-interesserede frit kan downloade fra korpus.dsl.dk/corest. I forhold til den version, som bruges på Den Danske Ordbogs redaktion har **Standard-udgaven** nogle begrænsninger:

- ✧ Antallet af tilgængelige korpusser er begrænset
- ✧ Tekster er kun tilgængelige i form af citater og kan ikke læses i deres helhed
- ✧ Annoteringssystemet er mere begrænset
- ✧ Der er ikke adgang til Den Danske Ordbogs interne ordsamlinger.

Manualen indfører i brugen af CoREST, og den beskriver, hvordan man installerer og starter CoREST. Herefter gennemgår den samtlige funktioner i værktøjet. Til gennemgangen knytter der sig mange konkrete eksempler, som skal gøre læseren fortrolig med mulighederne i dette værktøj.

OBS!

Den foreliggende manual, korpusværktøjet CoREST og CoREST-søgbare tekstkorpusser og ordbaser samt sprogteknologiske metoder og værktøjer anvendt under processeringen af materialet og det tilhørende website korpus.dsl.dk er udarbejdet af **Jørg Asmussen, Det Danske Sprog- og Litteraturselskab**.

Læseren er velkommen til at lade sig inspirere af den foreliggende manual og dens bilag til egne faglige formål, **hvis der henvises til den inklusive weblink**. Det samme gælder citater fra dokumentationen eller dens bilag.

# Indhold

# Del I

# Hurtigt i gang

# Kapitel 1

# Hvad er CoREST?

*Forskellige CoREST-udgaver, teknikken bag, feedback*

Dette kapitel giver en introduktion til, hvad CoREST er, og hvad dette værktøj kan bruges til.



## 1.1 Funktioner

CoREST står for *Corpus Retrieval System & Tools*. CoREST er et værktøj til avancerede sproglige undersøgelser i meget store tekstsamlinger, såkaldte tekstkorpusser. Med CoREST kan du blandt andet

- ✧ udføre avancerede søgninger på enkeltord og grupper af ord
- ✧ bruge tekstoplysninger som filtre i søgningerne
- ✧ se fund som matchlister, konkordanser og grafiske præsentationer
- ✧ se, hvilke ord der hyppigt optræder sammen med fundet
- ✧ se lister over fremtrædende ord i bestemte tidsintervaller
- ✧ sortere resultater på forskellig vis
- ✧ se større kontekster omkring fund
- ✧ se tekstoplysninger som titel, forfatter m.m. – og søge på dem

- ✧ se supplerende oplysninger om de enkelte ord, fx deres grundform, ordklasse og bøjning – og søge på dem
- ✧ markere fund
- ✧ eksportere fund som tekstfil til brug i andre programmer.

## 1.2 Udgaver

CoREST findes i forskellige udgaver med forskellige funktioner og med adgang til forskellige korpusser. Der findes følgende udgaver:

**Standard** Denne udgave indeholder en række basale korpusser, som giver et indblik i dansk sprog fra omkring 1980 og frem til i dag. Til visse korpusser knytter der sig specielle lister over søgeord, fx årets ord. Søgeresultater kan forsynes med simple markeringer. Visningen af hele tekster er af ophavsretlige grunde ikke mulig. Standard-udgaven kan downloades frit fra korpus.dsl.dk/corest og kan bruges af alle med interesse for dansk sprog.

**Research** Denne udgave er identisk med Standard-udgaven og tillader derudover mere avancerede markeringer af søgeresultater (annoteringer). Den henvender sig især til brugere med forskningsinteresser, som har brug for systematisk at analysere store mængder søgeresultater. Du kan få adgang til denne udgave ved at henvende dig til korpus@dsl.dk med en kort beskrivelse af dit forskningsprojekt.

**Ømål** Denne udgave bruges af redaktionen på Ømålsordbogen[1], og den er særligt tilpasset dette projekts behov. Der er adgang til et dialektkorpus og der kan udskrives fund til ordsedler. De øvrige funktioner svarer til Research-udgavens.

**DSL** Denne udgave bruges internt på Det Danske Sprog- og Litteraturselskab, DSL, og anvendes først og fremmest ved redigeringen af Den Danske Ordbog[2]. I forhold til Research-udgaven giver DSL-udgaven adgang til en række specialkorpusser samt til visse særlige funktioner, som bruges i forbindelse med ordbogsredigering, fx kan korpustekster vises i deres helhed og der er adgang til ordbogsinterne ordsamlinger og et særligt modul til organisering af ordbogsarbejdet.

Denne manual beskriver **Standard-udgaven**. Hvis du bruger en anden udgave af CoREST, bør du skaffe dig den dertilhørende manual enten på hjemmesiden korpus.dsl.dk/corest eller ved at sende en mail til korpus@dsl.dk.

[1] https://dialekt.ku.dk/dialektforskning_i_dk/ordboeger/oemaalsordbogen/
[2] https://ordnet.dk/ddo

## 1.3 Teknik

Teknisk set består CoREST-systemet af to dele:

**Klient** Den ene del er en klient. Det er det program, som skal ligge på din lokale maskine. I klienten formulerer du søgeforespørgsler, får vist resultater, kan markere fund, eksportere resultater osv.

**Server** Over internettet står klienten i forbindelse med den anden del af CoREST-systemet, nemlig en server på Det Danske Sprog- og Litteraturselskab, som tager imod søgeforespørgslerne, udfører dem på tekstkorpusserne og sender resultater tilbage til klienten.

Populært sagt fungerer CoREST-klienten som en slags browser for det indhold, som korpusserveren stiller til rådighed. CoREST-klienten er udviklet i programmeringssproget Java og fungerer derfor på alle gængse operativsystemer. CoREST-serveren er opbygget som en REST-baseret webservice[3], ligeledes programmeret i Java, og kører som en web-applikation. CoREST-serveren bruger CQP i OpenCWB[4] som back-end-korpussøgemaskine. CQP står for *Corpus Query Processor*.

## 1.4 Feedback

CoREST og tilhørende korpusser er under stadig udvikling. Der er derfor ikke tale om et færdigt og fejlfrit produkt! Du vil helt sikkert konstatere en række fejl og mangler, og det er meget sandsynligt, at du vil savne en række funktioner.

For at kunne videreudvikle CoREST, så systemet tilgodeser brugernes behov, er dine kommentarer meget velkomne! Forbedringsforslag og flere konkrete eksempler til selve manualen er også meget velkomne!

I de følgende afsnit kan du læse om, hvordan du indsender dine kommentarer og ønsker.

### 1.4.1 Indsend meddelelser om fejl

Skriv en e-mail til korpus@dsl.dk. I emnefeltet skriver du *CoREST-fejl*. Giv selve mailen følgende opbygning:

**Fejl** *Når jeg vil lave x, sker der y* – hvor du så indsætter din beskrivelse for *x* (hvad du gør) og *y* (hvad der sker).

### 1.4.2 Vær med til at styre den fremtidige udvikling

Skriv en e-mail til korpus@dsl.dk. I emnefeltet skriver du *CoREST-feature*. Selve mailen skal have følgende opbygning:

[3] https://en.wikipedia.org/wiki/Representational_state_transfer
[4] http://cwb.sourceforge.net

**Feature** *Jeg vil gerne lave x, derfor har jeg brug for funktionalitet y* – hvor du indsætter din beskrivelse for *x* (hvilket behov du har) og *y* (hvilken funktionalitet du mangler i den forbindelse).

### 1.4.3 Hjælp med at forbedre af manualen

Finder du fejl i manualen, har du forbedringsforslag til den eller gode eksempler, som bør optages her, skriver du til korpus@dsl.dk. I emnefeltet skriver du *CoREST-manual*.

# Kapitel 2

# Hvordan får man fat i CoREST?

*Download, installation og start*

I dette kapitel beskrives, hvor du kan downloade CoREST, hvordan du installerer programmet på din computer, og hvordan du får CoREST til at virke.



**OBS!** Download, installation og brug af CoREST sker på dit eget ansvar! Det Danske Sprog- og Litteraturselskab fralægger sig ethvert ansvar for mulige konsekvenser, som brugen af CoREST måtte medføre. DSL har desværre ingen ressourcer til at hjælpe dig ved download, installation og brug af CoREST.

## 2.1 Downloade CoREST

Du kan downloade Standard-udgaven af CoREST-klienten fra CoREST's hjemmeside[1]. Du får programmet i form af en JAR-fil[2], en fil, hvis navn ender på *.jar*. Navnet på den downloadede JAR-fil har følgende struktur:

✧ `CoREST-`*`åååå-mm-dd-uuu`*`.jar`

Den midterste kursiverede del af navnet varierer afhængigt af CoREST's version og udgave: *åååå-mm-dd* oplyser datoen for den pågældende version af CoREST, mens *uuu* oplyser udgaven af CoREST i form af en kode på tre bogstaver.

[1] Se `https://korpus.dsl.dk/corest/`. Research-udgaven kan du få ved at sende en mail med en kort beskrivelse af det formål, som du vil bruge CoREST til, til korpus@dsl.dk.

[2] Se fx `https://en.wikipedia.org/wiki/JAR_(file_format)`

## 2.2 CoREST har brug for Java

CoREST er et Java-program. Det kan derfor køre på alle operativsystemer, forudsat at Java 8 (eller højere) er installeret. På langt de fleste computere bør det være tilfældet. Hvis den nødvendige Java-version af en eller anden grund ikke findes på din maskine eller ikke er sat korrekt op, vil CoREST ikke kunne køre. Du kan downloade Java fra Oracles hjemmeside[3], hvor der også er hjælp at hente til installationen af det.

## 2.3 Installere CoREST

CoREST kræver ingen særlig installation. Du kan lægge *jar*-filen et vilkårligt sted på din maskine, men mest oplagt er det at lægge filen det sted, hvor dine øvrige programmer ligger. Det kan være en idé at oprette en genvej til programfilen, fx fra Windows' skrivebord eller fra Mac'ens dock.

## 2.4 Starte CoREST

Når du har lagt *jar*-filen der, hvor du vil have den, starter du CoREST ved at dobbeltklikke på den.

### 2.4.1 Programmet vil ikke starte under Windows

Under Windows kan det forekomme, at *jar*-filen ikke bliver udført som et program, men i stedet for bliver pakket ud, når du dobbeltklikker på den. Det skyldes, at Windows i så fald ikke betragter jar-filen som et program, men som et arkiv. For i virkeligheden er en java-programfil nemlig ikke én fil, men et arkiv med mange filer, som er pakket sammen på lignende måde som en zip-fil. Hvis Windowssystemet pakker *jar*-filen ud i stedet for at starte den, er der formentlig tale om en opsætningsfejl af sammenspillet mellem Windows og java. Hjælp kan i så fald findes på internettet.[4]

### 2.4.2 »Ukendt udvikler« på Mac

Første gang, du åbner CoREST på en Mac, vil der komme en meddelelse om, at programmet stammer fra en ukendt udvikler og derfor ikke kan startes. Du starter programmet ved at højreklikke på programfilen, hvorefter der kommer en lokal menu op. Vælg *Åbn* fra denne menu. Igen bliver du advaret om, at programmet stammer fra en ukendt udvikler, men denne gang kan du starte det ved at klikke på knappen *Åbn* i den boks, der kommer op. Du skal kun igennem denne manøvre én gang, herefter vil CoREST kunne startes som alle øvrige programmer på din Mac.

[3] https://www.java.com/en/download/

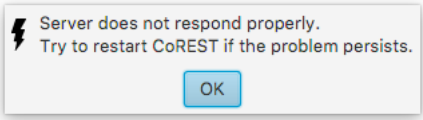
[4] Fx i denne artikel: https://www.makeuseof.com/tag/how-to-open-jar-files/

## 2.5 Kommunikation med serveren

Det er vigtigt, at CoREST kan få adgang til CoREST-serveren fra din computer. Det vil sige:

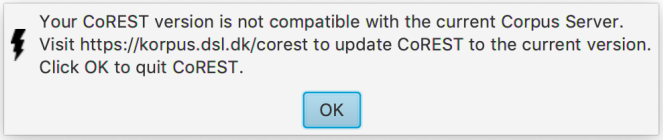
1. Din computer skal have internetforbindelse
2. Din version af CoREST skal kunne kommunikere med den aktuelle version af CoREST-serveren.

Hvis der opstår problemer med kommunikationen over nettet mellem CoREST og serveren, viser CoREST en fejlmeddelelse om, at serveren ikke svarer som forventet, se figur 2.1. Efter at du har klikket på *OK* og har sikret dig, at din netforbindelse fungerer, kan du forsøge at genstarte CoREST.



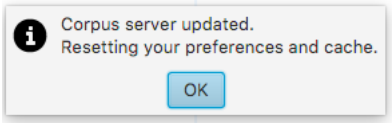

Figur 2.1: Kommunikationsfejl

Hvis din version af CoREST er forældet i forhold til serveren, kan du ikke længere bruge CoREST, før du har opdateret programmet til den nyeste version. Du vil da få en meddelelse om manglende kompatibilitet som vist i figur 2.2.




Figur 2.2: Ikke kompatibel CoREST-version

Endelig kan du få en meddelelse om, at serveren er blevet opdateret, se figur 2.3.




Figur 2.3: Korpusserver opdateret

Der er i dette tilfælde oftest tale om, at et korpus er blevet udvidet eller ændret. I dette tilfælde slettes CoREST's lokale cache på din maskine og dine indstillinger nulstilles, så CoREST ikke viser forældede eller fejlagtige resultater. Efter

en sådan opdatering opfører CoREST sig, som var den blevet startet for første gang, hvilket betyder, at du skal være opmærksom på, at korpusset KORPUS-DK automatisk er valgt, jf. kapitel 3. Du kan læse mere om cachen i afsnit 2.6

Hvis der opstår problemer, som du mener skyldes en generel fejl i CoREST, er du velkommen til at indsende en fejlrapport, se kapitel 1.

## 2.6 Cache

CoREST-klienten kommunikerer med CoREST-serveren over internettet. Udvekslingen af data over nettet kan være en tidskrævende faktor ved søgningen. For at få systemet til at reagere så hurtigt som muligt, gemmes alle data, som din CoREST-klient har hentet fra serveren, i en dertil oprettet database på din maskine, en såkaldt *cache*. Får din klient på et senere tidspunkt brug for at hente de samme data igen, undersøger den først, om de allerede ligger i cachen. Hvis det er tilfældet, hentes data direkte herfra i stedet for fra serveren.

Normalt behøver du ikke tænke over cachens tilstedeværelse. Hvis der dog optræder problemer med CoREST's funktionalitet, kan de sommetider løses ved at slette cachens indhold. Problemer kan optræde, når en ny version af et korpus eller CoREST tages i brug, eller når en søgning slog fejl. Du tømmer cachen ved at klikke på knappen »Empty Cache« i info-vinduet. Mere herom i kapitel 5.

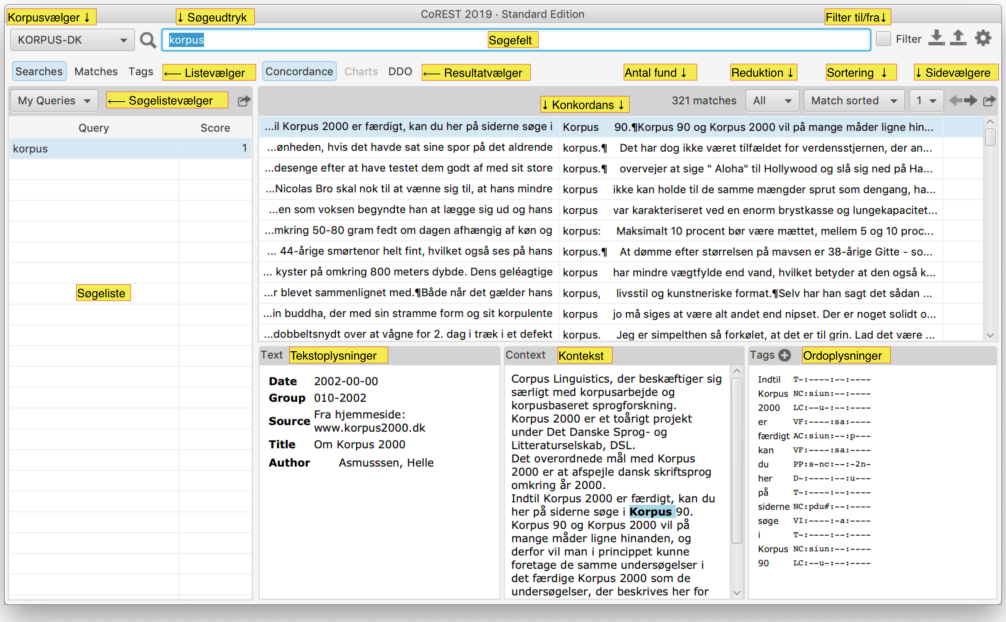
# Kapitel 3

# Brugerinterfacet

*Betjeningselementer og visning af søgeresultater*

I dette kapitel beskrives, hvordan brugerinterfacet er opbygget, og hvordan du hurtigt kan lave simple korpussøgninger. En mere detaljeret gennemgang af alle elementerne i brugerinterfacet finder du i del II af denne manual.

Figur 3.1 viser CoREST's brugerinterface efter en søgning på `korpus` i korpusset KORPUS-DK. De vigtigste elementer i CoREST's brugerinterface fremgår af de gule labels i figuren, og de vil blive gennemgået i dette kapitel.




Figur 3.1: CoREST-vinduet efter en søgning

Når du åbner CoREST for første gang på din computer, er korpusset KORPUS-DK automatisk valgt i korpusvælgeren, så det er dette korpus, CoREST som udgangspunkt søger i. Der er indsat en søgning på `korpus` i søgefeltet. Du kan nu enten taste retur eller klikke på søgeknappen – det er knappen, der ligner en lup. Herefter udfører CoREST en søgning på ordet *korpus* i KORPUS-DK og viser som resultat alle fund, hvert på en linje med søgeordet centreret og noget kontekst omkring. En sådan opstilling kaldes en *konkordans*. Efter søgningen svarer CoREST-vinduet omtrent til det, du kan se i figur 3.1.

## 3.1 Betjeningselementer

CoREST's primære funktion er det at udføre komplekse søgninger i store tekstkorpusser og vise resultaterne i form af ordlister og som lister med små tekstuddrag, såkaldte konkordanser, som behandles i detaljer i kapitel 7. Øverst i CoREST-vinduet finder du alle betjeningselementer, som bruges til at lave søgninger i et korpus med, se figur 3.2.

Fra venstre mod højre er det


Figur 3.2: Betjeningselementer til søgning

- ✧ Korpusvælgeren i form af en rullemenu, se afsnit 3.1.1
- ✧ Søgeknappen i form af en lup, se afsnit 3.1.2
- ✧ Søgefeltet i form af et indtastningsfelt, se afsnit 3.1.3
- ✧ En filterboks i form af en tjekboks, se afsnit 3.1.4
- ✧ Knapper til at gemme eller hente søgninger, se afsnit 3.1.5.

Endelig er der også en

- ✧ Knap til indstillinger og systeminfo i form af et tandhjul, se afsnit 3.1.6.

### 3.1.1 Korpusvælger

Med korpusvælgeren øverst til venstre i CoREST-vinduet vælger du det korpus, som du vil søge i. Hvilke korpusser du kan vælge, afhænger af den CoREST-udgave, du bruger. CoREST husker det valgte korpus fra gang til gang, men når du starter CoREST for første gang, eller når CoREST eller et korpus er blevet opdateret, vil CoREST altid automatisk vælge KORPUS-DK. Så efter start af CoREST bør du huske at tjekke, om CoREST faktisk bruger det korpus, du vil søge i. Alle udgaver af CoREST giver adgang til KORPUS-DK, som er det korpus, de fleste eksempler i denne manual beror på.

### 3.1.2 Søgeknap

Søgeknappen findes til højre for korpusvælgeren. Når du klikker på søgeknappen (eller taster retur) sættes en søgning i gang i det korpus, du har valgt. Der søges altid på eksempler i korpus, som matcher det, du har formuleret i dit søgeudtryk i feltet til højre for søgeknappen. Du kan læse mere om søgeudtryk i kapitel 4. Ved længerevarende søgninger forvandles søgeknappen til et roterende ventetidssymbol.

I stedet for at klikke på søgeknappen kan du også sætte en søgning i gang ved at taste retur, når søgefeltet (eller et filterfelt) har fokus.

### 3.1.3 Søgefelt

Til højre for søgeknappen er der et indtastningsfelt, hvor du indtaster det, du vil søge på, formuleret som et søgeudtryk. Her kan du indtaste to typer søgeudtryk:

*simple søgeudtryk* (med eller uden kontroltegn), se kapitel 4, eller *avancerede søgeudtryk*, se del III. Et simpelt søgeudtryk består i sin mest enkle form af et eller flere ord. I søgeudtrykket kan der også indgå såkaldte jokertegn (regulære udtryk), mere herom i kapitel 14. CoREST finder selv ud af, hvilken type søgning der er tale om ud fra det søgeudtryk, du har indtastet, men du kan ikke blande de to typer søgeudtryk i en og samme søgning.

### 3.1.4 Filterboks

Sætter du hak i filterboksen til højre for søgeknappen, åbnes der en række med forskellige filtermuligheder, som behandles nærmere i kapitel 18. Filtre giver mulighed for at begrænse søgningen til kun at omfatte tekster fra en bestemt kilde.

### 3.1.5 Gemme og hente søgninger

Til højre for filterboksen er der to knapper, der henholdsvis tjener til at gemme en søgning som en fil på din computer og senere hente den ind i CoREST igen. Ved meget komplekse søgninger med et avanceret søgeudtryk og detaljerede filtre kan det være en fordel at gemme opskriften på søgningen til genbrug ved en senere lejlighed. Endvidere kan disse søgeopskrifter også deles med andre brugere.

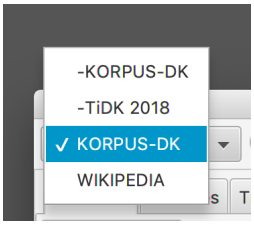
### 3.1.6 Indstillinger og systeminfo

Tandhjulsymbolet yderst til højre giver adgang til indstillinger og systemoplysninger. Se mere herom i kapitel 5.

## 3.2 En simpel søgning

### 3.2.1 Vælge korpus og søge i det

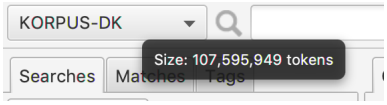
Med korpusvælgeren vælger du det korpus, som du vil søge i. Hvilke korpusser du kan vælge imellem, afhænger af den CoREST-udgave, du bruger. Som minimum skulle de følgende korpusser være tilgængelige: KORPUS-DK, TiDK og WIKIPEDIA, se figur 3.3.




Figur 3.3: Valg af korpus med korpusvælgeren

#### 3.2.1.1 Korpusstørrelse

Størrelsen på det valgte korpus kan du få oplyst ved at holde musemarkøren over korpusvælgeren, se Figure 3.4 on page 17, hvor størrelsen for KORPUS-DK vises. Størrelsen oplyses i som det samlede antal ord (tokens) i alle korpustekster.




Figur 3.4: Oplysning af korpusstørrelse

Du kan også se alle tilgængelige korpussers størrelser i sessionsloggen. Du kan læse mere om denne log i kapitel 5.

#### 3.2.1.2 Forskellige slags korpusser

I CoREST findes der tre typer af korpusser, som falder i de to grupper *sammensatte* og *ikke-sammensatte* korpusser, en mere detaljeret gennemgang finder du i kapitel 19.

Et **sammensat korpus** består af en række mindre delkorpusser. Fx er -TiDK-korpusserne sammensatte og består af ét delkorpus per år. En søgning i et sammensat korpus vil give særskilte resultater for samtlige delkorpusser, så du kan lave sammenlignende undersøgelser og også få præsenteret resultaterne i form af diagrammer. Sammensatte korpusser kan du i korpusvælgeren kende på, at deres navne begynder med en streg som fx -TiDK-korpusset i figur 3.3 på forrige side.

Et **ikke-sammensat** korpus som KORPUS-DK derimod er kun ét korpus. Det udgør en helhed og er ikke sammensat af andre, mindre korpusser. KORPUS-DK findes i øvrigt også i en sammensat version; søgninger i den sammensatte version af dette korpus demonstreres i kapitel 7. Desuden vises eksempler på søgninger i det sammensatte TiDK-korpus i kapitel 12.

I manualen vil de allerfleste eksempler bero på det ikke-sammensatte KORPUS-DK. Læs mere om korpustyper i kapitel 19.

### 3.2.2 Indtast søgeudtryk

I *søgefeltet* indtaster du det, du vil søge på, formuleret som et *søgeudtryk*. I søgefeltet kan du indtaste to typer søgeudtryk: simple og avancerede. Indtil videre holder vi os udelukkende til simple søgeudtryk, mens de avancerede søgemuligheder bliver behandlet i del III af denne manual.

Et simpelt søgeudtryk består i sin mest enkle form af et eller flere ord. I søgeudtrykket kan der også indgå såkaldte jokertegn, *regulære udtryk*, se kapitel 14. Ved at klikke på *søgeknappen* eller ved at taste retur sætter du søgningen i gang i det valgte korpus, i dette eksempel KORPUS-DK.

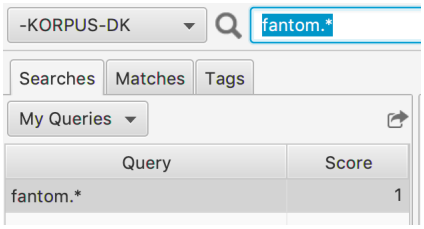
Prøv at skrive søgeudtrykket `fantom.*` i søgefeltet og tast retur eller klik på søgeknappen. Denne søgning finder alle eksempler fra korpusset, som indeholder ord, som begynder med *fantom* inklusive ordet *fantom* selv. Jokertegnet `.*` (punktum stjerne) betyder, at der kan være et vilkårligt antal (også nul) vilkårlige bogstaver på dette sted. En udførlig beskrivelse af, hvordan man bruger jokertegn i CoREST, finder du i kapitel 14.

I det følgende gennemgås de forskellige oplysninger, som CoREST giver som resultat af en søgning.

### 3.2.3 Se søgeresultater

#### 3.2.3.1 Søgeliste

Efter start af CoREST er fanebladet »Searches« i området med lister til venstre for konkordansområdet automatisk valgt, efter søgningen på `fantom.*` står dette søgeudtryk øverst på søgelisten »My Queries«, jf. Figure 3.5 on page 18.




Figur 3.5: Søgelisten efter en simpel søgning

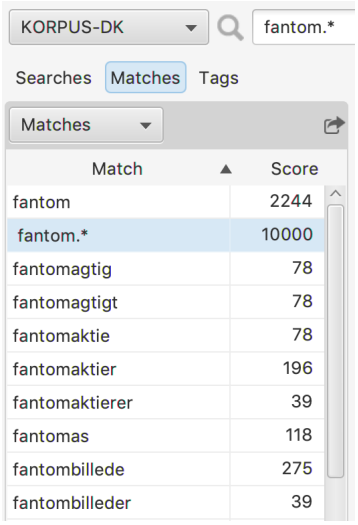
Søgelisten indeholder måske i forvejen andre søgeudtryk, som stammer fra tidligere søgninger. I søgelisten »My Queries« registreres alle dine søgninger i kolonnen »Query« sammen med det antal gange, du har udført dem, i kolonnen »Score«. Hvis du hurtigt vil vende tilbage til en tidligere søgning, kan du nøjes med at markere søgeudtrykket i søgelisten og taste *retur*.

En udførlig beskrivelse af søgelisten og andre ordlister finder du i kapitel 6.

#### 3.2.3.2 Matchliste

Til højre for fanebladet »Searches« er der et med navnet »Matches«. Klikker du på det efter søgningen på `fantom.*`, vises alle forskellige ord fra korpus, som matcher søgeudtrykket, se figur 3.6 på næste side.

Listen er som udgangspunkt sorteret alfabetisk. Blandt de første ord på listen står altid selve søgeudtrykket rykket lidt ind til højre, så det kan skelnes fra de andre ord på listen. Ud for hvert ord vises en score, der angiver, hvor hyppig pågældende ord er blandt alle ord, der matcher søgeudtrykket. Alle match tilsammen sættes altid til 10.000 (som så svarer til 100 %). En søgning på `fantom.*` giver blandt mange andre følgende match: *fantom* med en score på 2244, altså en andel på 22,44 % af samtlige match, og *fantomet* med en score på 2598 (25,98 %).




Figur 3.6: Udsnit af matchlisten efter en søgning på `fantom.*`

Disse to former udgør tilsammen altså knap halvdelen af alle match. Matchlisten kan også sorteres efter score ved at klikke på overskriften »Score« over scorekolonnen. Markerer du et match i matchlisten ved at klikke på det, fx på *fantombillede*, og derefter taster *retur*, fremhæves den første forekomst af denne form i konkordansen. En udførlig beskrivelse af matchlisten og andre ordlister finder du i kapitel 6.

#### 3.2.3.3 Annoteringer

Et tredje faneblad »Tags« er ikke relevant i denne hurtige gennemgang af funktionerne. Det beskrives i kapitel 7.

#### 3.2.3.4 Konkordans

Konkordansen viser korpusforekomster af de ord, der matcher søgeudtrykket. Hver forekomst vises centreret på en linje med nogle ords kontekst omkring, se Figure 3.7 on page 20

Konkordansen er som udgangspunkt lige som matchlisten sorteret alfabetisk på matchet. Ved en søgning på `fantom.*` i KORPUS-DK er det samtlige 254 forekomster af alle former fra matchlisten, der vises i konkordansen. En konkordans kan maksimalt vise 2000 fund fordelt på 20 sider med 100 fund per side. Hvis der er flere match i korpus, reduceres antallet af viste fund automatisk til 2000. Hvis en konkordans indeholder flere end 100 fund og dermed består af flere sider, kan du skifte fra side til side med sidevælgeren, jf. figur 3.1 på side 14. En udførlig beskrivelse af konkordanser og hvad du kan lave med dem, findes i kapitel 7.


Figur 3.7: Udsnit af konkordansen efter en søgning på `fantom.*`

#### 3.2.3.5 Konkordanslinjeoplysninger

Under konkordansen vises forskellige oplysninger om den konkordanslinje, der aktuelt er markeret – umiddelbart efter en søgning er det altid den første linje. Oplysningerne vises i de følgende tre felter fra venstre mod højre under konkordansen, se figur 3.1 på side 14:

**»Text« – Tekstoplysninger** Her vises oplysninger om den tekst, som den markerede konkordanslinje stammer fra. Tekstoplysningernes omfang og art kan variere fra korpus til korpus. Du kan læse mere om tekstoplysninger i kapitel 8.

**»Context« – Kontekst** Her vises matchet med noget mere kontekst omkring, end hvad der umiddelbart kan ses i selve konkordanslinjen. En udførlig beskrivelse af kontekstvisning finder du i kapitel 9.

**»Tags« – Ordoplysninger** Til hvert ord i korpus er der knyttet forskellige oplysninger, fx dets grundform eller dets ordklasse. Et udvalg af disse oplysninger vises i ordoplysningsfeltet. Du kan læse mere om ordoplysninger i kapitel 10.

### 3.2.4 Intet match

Hvis der i det valgte korpus ikke findes noget match på din søgning, markeres søgefeltet med en rødlig farve, som vist i Figure 3.8 on page 20.


Figur 3.8: Intet match fundet i korpus

Prøv at udføre en søgning på `vxlkommen` for at få meddelelsen frem. Den samme tilbagemelding får du, hvis dit søgeudtryk indeholder formelle fejl, som bevirker, at CoREST ikke kan fortolke søgeudtrykket. Det gælder fx `velkommen]` – som

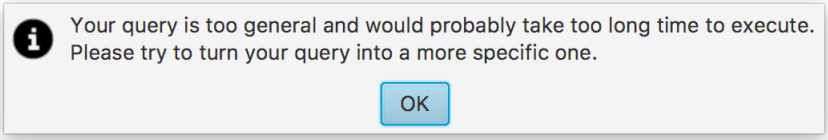
slutter på en kantet parentes. Kantede parenteser har en særlig funktion i søgeudtryk, og de skal anvendes på en særlig måde, som ikke er overholdt her.

### 3.2.5 Søgehastighed

Søgehastigheden afhænger blandt andet af det antal forekomster, som CoREST finder på baggrund af dit søgeudtryk. Hvis dit søgeudtryk søger på grupper af ord, er det antallet af match på det første ord i gruppen, der er afgørende for hastigheden. Lidt forenklet kan man sige: Jo flere match søgeudtrykket har i korpus, desto længere tid er CoREST om at finde dem frem, også selvom CoREST til sidst kun afleverer 2000 eksempler til visning i konkordansen.

Mens en længerevarende søgning bliver udført, viser CoREST et roterende ventetidssymbol på det sted, hvor søgeknappen ellers befinder sig.

Hvis du fx prøver at udføre en søgning på `.*`, som matcher alle ord i korpus, altså knap 110 millioner i KORPUS-DK, bliver søgningen afvist af CoREST, fordi den er for tidskrævende. Når CoREST blokerer bestemte søgninger, vises meddelelsen i figur Figure 3.9 on page 21. I så fald kan du komme videre ved at formulere et udtryk, som begrænser søgningen.




Figur 3.9: Meddelelse ved for generelt søgeudtryk

En gentagen søgning på samme søgeudtryk vil som regel gå hurtigere, idet resultaterne så vil blive hentet fra en lokal cache og ikke fra serveren.

## 3.3 Navigere i brugerinterfacet

Du kan gøre de forskellige områder af brugerinterfacet aktive på skift ved enten at klikke på dem eller ved at taste tabulatortasten, se figur 3.1 på side 14. Det aktive område er det, der modtager input fra tastaturet; det er det område, som har fokus. Der er følgende områder i interfacet:

**Søgefelt:** Efter en søgning har søgefeltet fokus, og søgeudtrykket er markeret med en blå baggrund.[1] CoREST er klar til at modtage et nyt søgeudtryk, som du kan indtaste umiddelbart uden forudgående klik eller tastetryk eller uden at skulle slette det sidste søgeudtryk.

**Ordliste:** Efter et tryk på tabulatortasten skifter fokus til den aktuelt valgte ordliste – som udgangspunkt den, der hedder »Searches«. Er området med filtre

[1] Baggrundsfarven er afhængig af dit operativsystem og kan derfor være en anden end blå.

åbent, får filterfelterne fokus et efter et, før ordlisten får fokus. I ordlisten kan du nu med pil-ned eller pil-op flytte markøren fra linje til linje. Hvis fanebladet »Searches« er valgt, medfører et tryk på returtasten en korpussøgning på det aktuelt markerede ord eller udtryk. Hvis fanebladet »Matches« er valgt, medfører et tryk på returtasten, at første forekomst af pågældende match bliver vist og fremhævet i den tilhørende konkordans. Fanebladet »Tags« har en mere kompleks funktion, se mere herom under næste punkt eller i kapitel 7.

**Konkordans:** Endnu et tryk på tabulatortasten rykker fokus fra matchlisten til konkordansen. Nu vedrører alle tastetryk konkordansen. Med tryk på *pil-ned* og *pil-op* kan du flytte markøren fra konkordanslinje til konkordanslinje. Konkordanslinjeoplysningerne under konkordansen opdateres løbende og viser altid de oplysninger, som vedrører den aktuelt markerede linje. Hvis konkordansen består af flere sider, skifter du mellem siderne med *pil-højre* eller *pil-venstre*, ved at klikke på knapperne *pil-højre* eller *pil-venstre* over konkordansen eller ved at vælge en side med sidevælgeren ved siden af pileknapperne. Når konkordansområdet har fokus, kan der også indtastes korte annoteringer i form af bogstaver og cifre, såkaldte mærker eller tags, som bliver synlige til højre for konkordanslinjen. Mærkerne samles under fanebladet »Tags«. Trykker du retur på et mærke fra listen, får du vist de fund i korpus, som har fået tildelt pågældende mærke.

**Tilbage til søgefeltet:** Et sidste tryk på tabulatortasten skifter fokus fra annoteringsfeltet tilbage til søgefeltet.

Pladsfordelingen mellem ordlisteområdet til venstre i CoREST-vinduet og konkordansområdet kan ændres, ved at du med musen forskyder skillelinjen mellem disse områder. Hvis du skyder denne linje helt ud til venstre, kan du skjule området med ordlister helt. Du kan tilsvarende forskyde fordelingen mellem konkordans og konkordansoplysningerne under den, og mellem konkordansoplysningerne indbyrdes, dog ikke så meget, at du kan skjule nogle af disse områder helt.

# Kapitel 4

# Standardsøgninger

*Simple søgeudtryk, kontroltegn, øvelser*

Dette kapitel giver en række eksempler på, hvordan du formulerer simple søgeudtryk. Desuden indeholder det en række øvelser, som skal træne dig i at formulere dine egne søgeudtryk.



I kapitel 3 om brugerinterfacet udførte vi to simple søgninger i KORPUS-DK, nemlig en på `korpus.*` og en på `fantom.*`. I dette afsnit vil vi udvide repertoiret af simple søgeudtryk.

## 4.1 Simple søgeudtryk

Tabel 4.1 på næste side viser en række søgeudtryk, som du bør prøve for at blive fortrolig med CoREST. Alle søgningerne er lavet i korpusset KORPUS-DK. I nogle af eksemplerne anvendes jokertegnet `.*` (punktum stjerne), som betyder nul, ét eller flere vilkårlige bogstaver på jokertegnets plads.

| Søgeudtryk | Kommentar |
|---|---|
| `smart` | Matcher alle forekomster af *smart* i korpus |
| `SMART` | Samme resultat som søgningen på `smart` |
| `smart.*` | Matcher alle ord, som begynder med *smart* |
| `smarttelefon.*` | Matcher alle ord, som begynder med *smarttelefon*. Ingen forekomster i KORPUS-DK |
| `.*telefon` | Matcher alle ord, som ender på *telefon* |
| `tele.*en` | Matcher alle ord, som begynder med *tele* og ender på *en* |
| `.*udbyder.*` | Matcher alle ord, som indeholder (eller begynder med eller slutter på) *udbyder* |
| `nye medier` | Matcher alle ordgrupper, som starter med *nye* og slutter med *medier* med op til 3 vilkårlige ord mellem *nye* og *medier* |
| `$nye medier` | Dollartegnet foran søgeudtrykket bevirker, at det kun matcher *nye medier* uden andre ord imellem |
| `ny.* medi.*` | Matcher alle ordgrupper, hvis første ord begynder med *ny*, og hvis sidste ord begynder med *medi*, og som har op til 3 vilkårlige ord imellem sig |
| `$ny.* medi.*` | Dollartegnet gør, at søgeudtrykket kun matcher ordpar, hvor det første begynder med *ny* og det sidste med *medi* |
| `$ud af boks.*` | Matcher alle de grupper på 3 ord, som begynder med ordene *ud af* og ender med et ord, der begynder med *boks* |
| `$to værelses` | Matcher både *to værelses* og *to-værelses*, idet bindestreg betragtes som ordgrænse |
| `$pc en` | Matcher både *(giv din) pc en (vitaminindsprøjtning)*, *(en) pc, en (telefon)* og *pc'en*, idet apostrof og sætningstegn fungerer som ordgrænse |
| `$2 0` | Matcher bl.a. *2.0*, *2,0*, *2-0*, da sætningstegn (herunder bindestreg) fungerer som ordgrænser |

Tabel 4.1: Simple søgeudtryk

Du kan sagtens formulere søgeudtryk, som indeholder flere end blot et enkelt ord. I konkordansen vises der i så fald fund, hvor der mellem hvert af de matchende ord kan stå op til tre andre vilkårlige ord. Hvis du vil have indflydelse på, hvor

mange ord der må stå mellem de matchende ord, eller hvis den indtastede sekvens af ord skal findes præcist, som den er indtastet, kan du bruge kontroltegn, se afsnit 4.3 på den følgende side eller avancerede søgeudtryk, som beskrives i kapitel 15.

Det er en god idé at vælge fanebladet »Matches« 3.2.3.2 blandt listefanebladene, inden du afprøver søgningerne, for her får du altid vist en oversigt over alle forskellige match til hver enkelt søgning. Du kan læse mere om de forskellige listefaneblade i kapitel 3.

Du kan finde uddybende information om søgning følgende steder i manualen:

- ✧ Søgning med jokertegn beskrives i kapitel 14.
- ✧ Avanceret søgning, som giver dig fuld kontrol over, hvad du vil finde, beskrives i kapitel 15.
- ✧ Hvordan et ord er defineret i CoREST, er beskrevet i kapitel 17.

## 4.2 Øvelse: Simple søgeudtryk

De følgende opgaver beror på søgninger i KORPUS-DK. Du kan besvare opgaverne ved at formulere passende simple søgeudtryk, udføre søgningerne og betragte de resulterende matchlister (vælg listefanebladet »Matches«, jf. kapitel 3), som det kan være en fordel at sortere efter faldende hyppighed ved at klikke to gange på feltet »Score« over listen. I opgaverne bør du kun tage hensyn til grundformerne af ordene, øvrige bøjningsformer behøver du ikke søge på.

1. Hvad er den hyppigste orddannelse med førsteleddet *smart*?
2. Hvorfor er der ingen forekomster af *smarttelefon*?
3. Hvad er de hyppigste to typer *telefoner* i korpus?
4. Hvilke orddannelser er hyppigst? Dem, der har *udbyder* som førsteled eller dem, der har *udbyder* som andetled?
5. Hvad er det hyppigste adjektiv mellem *nye* og *medier*?
6. Hvad er det hyppigste substantiv mellem de ord, der matcher `ny.*` og `medi.*`?
7. Hvad er de fire hyppigste ordgrupper, der matcher `høre.* gro`?

**Løsningsforslag** findes i appendiks A.1.

## 4.3 Kontroltegn i simple søgeudtryk

Der kan anvendes en række kontroltegn i simple søgeudtryk. Kontroltegn bevirker, at søgningen udføres på en lidt anden måde end ellers.

**$ = ingen ekstra ord** – Dette kontroltegn har du allerede lært at kende, hvis du har gennemgået eksemplerne i tabel 4.1 på side 24. Det sættes som det første tegn i et søgeudtryk. Mens søgeudtrykket `vanskelig situation` finder eksempler på både *vanskelig situation*, *vanskelig finansiel situation*, *vanskelig og sårbar situation* og en række andre (der kan være op til tre ord mellem *vanskelig* og *situation*), så finder udtrykket `$vanskelig situation` kun eksempler, hvor de to ord i søgeudtrykket står umiddelbart i forlængelse af hinanden.

**# = ét vilkårligt ord** – Beslægtet med `$` er kontroltegnet `#`. Det betyder *præcist ét ord pågældende sted*. Søgeudtrykket `vanskelig # situation` finder således blandt andet *vanskelig økonomisk situation*, *vanskelig politisk situation* og *vanskelig finansiel situation* i KORPUS-DK. Husk, at der skal være mellemrum mellem `#` og de omgivende ord – `vanskelig#situation` giver intet resultat. Kontroltegnet `#` kan også anvendes flere gange, fx finder `vanskelig ## situation` eksemplerne *vanskelig og penibel situation*, *vanskelig og sårbar situation* og *vanskelig og ubehagelig situation*, altså eksempler med præcist to ord mellem *vanskelig* og *situation*. Hvis du anvender kontroltegnet `#`, så ignoreres et eventuelt foranstillet `$` under søgningen.

Yderligere to kontroltegn vedrører fortolkningen af søgeord i søgeudtryk.

**! = skelne mellem store og små bogstaver** – Udråbstegn `!` umiddelbart foran et ord betyder, at ordet skal være stavet præcist som i søgeudtrykket, hvis det skal matche. Mens søgeudtrykket `SMART` finder eksempler på *smart*, *Smart* og *SMART*, finder `!SMART` udelukkende eksempler, der er stavet præcist som angivet i søgeudtrykket, i dette tilfælde altså eksempler på *SMART*. Tilsvarende finder søgeudtrykket `café` både *café* og *cafe* i store og små varianter. Det samme gælder for søgeudtrykket `cafe`. Derimod finder `!cafe` kun de stavninger som præcist svarer til den i søgeudtrykket, altså eksempler på *cafe*.

**@ = alle bøjningsformer af søgeordet** – Snabel-a `@` foran et søgeord betyder, at der skal søges på samtlige bøjningsformer af det. Således finder `@hændelse` alle forekomster af formerne *hændelse*, *hændelsen*, *hændelsens*, *hændelser*, *hændelserne*, *hændelsernes*, *hændelsers* og *hændelses*. De to kontroltegn `!` og `@` kan ikke begge sættes foran samme ord, et sådant søgeudtryk er ugyldigt og giver ikke noget resultat.

**% = alle bøjningsformer af søgeordet og ord, som starter med søgeordet** – Dette kontroltegn er en kombination af `.*` i slutningen af søgeordet og foranstillet `@`. Søgeudtrykket finder alle bøjningsformer af søgeordet og alle

ord, som begynder med søgeordet. Dette kontroltegn er et alternativ til `@`-søgningen, især i de tilfælde hvor en `@`-søgning ikke giver det forventede resultat. Ikke alle ordformer er identificeret korrekt i korpus, hvilket kan have indflydelse på resultatet af en `@`-søgning. Det tager noget mere tid at udføre en `%`-søgning end en `@`- eller `.*`-søgning alene.

Tabel 4.2 giver en række eksempler på søgninger med kontroltegn.

| **Søgeudtryk** | **Kommentar** |
|---|---|
| `uber` | Matcher alle forekomster af *über* og *uber* med en hvilken som helst kombination af store og små bogstaver |
| `!uber` | Matcher kun forekomster af *uber* (stavet nøjagtigt på denne måde) |
| `!Cirkus.*` | Matcher alle ord, som begynder med *Cirkus* (med stort *C*) |
| `det # !Kapel` | Matcher alle grupper på præcist tre ord, hvoraf det første er en staveform af *det*, det andet et vilkårligt ord og det tredje den præcise staveform *Kapel* (med stort *K*) |
| `@cirkus` | Matcher alle bøjningsformer af grundformen *cirkus* i søgekorpusset |
| `@bytte` | Matcher alle bøjningsformer af grundformerne *bytte* |
| `@lærer # @elev` | Matcher alle grupper på præcist 3 ord, hvoraf det første er en bøjningsform af *lærer*, det andet et vilkårligt ord og det tredje en bøjningsform af *elev* |
| `flere ## end` | Matcher alle grupper på præcist 4 ord, hvoraf det første er *flere*, det andet og tredje et vilkårligt ord og det sidste *end* |
| `@drink` | Matcher *drink*, *drinken*, *drink* og *drinksene* |
| `%drink` | Matcher som `@drink` og derudover blandt mange andre *drinkens*, *drinks* og *drinksenes* |

Tabel 4.2: Eksempler på søgninger med kontroltegn

## 4.4 Øvelse: Kontroltegn i simple søgeudtryk

De følgende opgaver beror på søgninger i KORPUS-DK, og du skal anvende kontroltegn i dine søgeudtryk og betragte de resulterende matchlister og konkordan-

ser for at kunne besvare opgaverne. OBS! I matchlisterne vises altid normaliserede former af de match, CoREST finder, mens de i selve konkordansen er gengivet i deres originale form.

1. Hvor mange forekomster er der af ordet *själ* (stavet på denne måde) i KORPUS-DK, og hvilke tekster stammer de fra?

2. Et af resultaterne af en søgning på `!uber` (altså den tyske præposition *über* uden prikker over *u*'et fulgt af et vilkårligt ord) er *uber cool*. Hvad er en mulig grund til, at sprogbrugeren bag dette fund ikke anvender den måske mere normrette stavning *übercool*?

3. Hvad er det hyppigste ord i KORPUS-DK, som begynder med *über* stavet på præcist denne måde?

4. Hvilke er de tre hyppigste navne (proprier), som begynder med *Cirkus* med stort *C*?

5. Hvad er det mest omtalte *kapel* (proprium) i KORPUS-DK? Og det næstmest omtalte?

6. Hvilke bøjningsformer af *cirkus* findes i KORPUS-DK?

7. Hvorfor findes formen *cirkuset* ikke, når man søger på `@cirkus`?

8. Hvilken tvetydighed løber man ind i ved en søgning på `@bytte`?

9. Et af matchene på `@lærer # @elev` i KORPUS-DK er *lærer slår eleverne*. Hvordan skal dette match fortolkes?

10. Hvad er den hyppigste konstruktion, som matcher `flere ### end`?

**Løsningsforslag** findes i appendiks A.2.

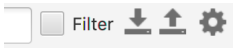
# Kapitel 5

# Indstillinger og systeminformationer

*Cache, historik, logfil og oversigt over søgninger*



I CoREST kan du skifte til et særligt vindue, hvor du dels kan ændre en række indstillinger, dels kan se en logfil, som blandt andet indeholder de søgninger, du har udført. Du skifter til dette vindue ved at klikke på tandhjulsymbolet øverst til højre i CoREST-standardvinduet, se figur 5.1.



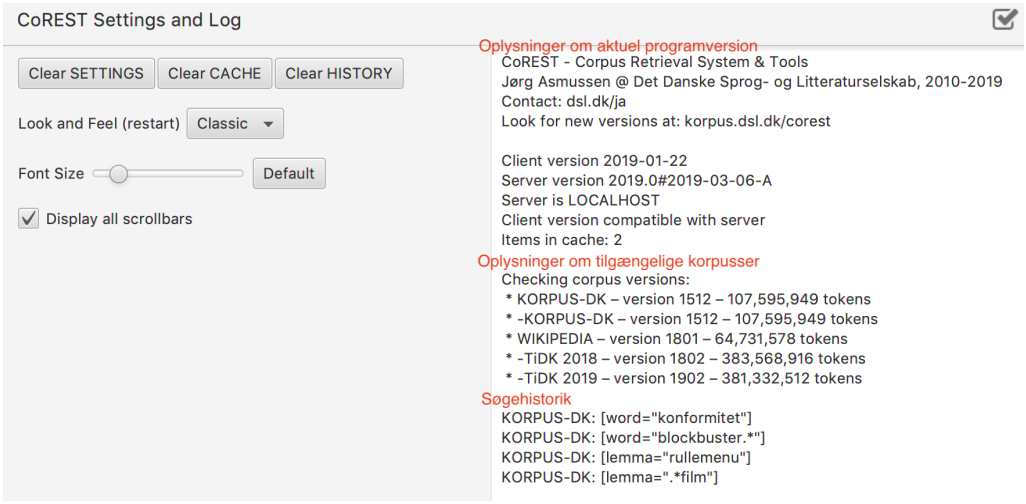

Figur 5.1: Skifte til vindue med indstillinger

Vinduet med indstillingerne og logfilen ser omtrent ud, som vist i figur 5.2 på den følgende side. I venstre halvdel findes indstillingerne, i højre logfilen, som viser, hvad CoREST har arbejdet med, siden du startede programmet.

Du kan lukke vinduet med indstillingerne ved at klikke på hakket i øverste højre hjørne, hvorefter du vender tilbage til CoREST's hovedvindue .

**OBS!** Logfilen over dine søgninger gemmes ikke, når du lukker CoREST. Hvis du vil gemme filen, kan du kopiere alt indholdet over i en editor eller et tekstbehandlingsprogram og gemme den herfra.

I området med indstillingerne i venstre halvdel af vinduet findes der øverst tre knapper »Clear SETTINGS«, »Clear CACHE« og »Clear HISTORY«, som bruges til at nulstille indstillinger, tømme cachen og tømme søgelisten over dine egne søgninger.




Figur 5.2: Vindue med indstillinger og logfil

## 5.1 Nulstille indstillinger

Du kan nulstille dine indstillinger ved at klikke på knappen »Clear SETTINGS«, se figur 5.2. Efter genstart vil CoREST bruge de standardindstillinger, programmet havde allerførste gang, det blev startet. Sletning af indstillinger medfører genstart af CoREST.

## 5.2 Tømme søgecache

Du kan tømme søgecachen på din maskine ved at klikke på knappen »Clear Cache«, se figur 5.2. Normalt skulle det ikke være nødvendigt at tømme cachen, men har du en mistanke om, at noget opfører sig unormalt, fx at søgninger ikke medfører sandsynlige resultater, kan det være en løsning at tømme cachen. Du kan læse mere om søgecachen i kapitel 2.

## 5.3 Tømme søgelisten med egne søgninger

Søgelisten med dine egne tidligere søgninger 6.1.1 sletter du ved at klikke på knappen »Clear History«, se figur 5.2. Du kan læse mere om denne søgeliste i kapitel 6.

## 5.4 Indstillinger

Med indstillingerne kan du fastlægge CoREST's udseende.

**Design:** CoREST kan køres med forskellige fremtoninger. Ud over den aktuelle versions standarddesign kan du vælge et traditionelt »Classic«-design. Desuden kan du også vælge et design fra tidligere versioner af CoREST. Ændring af design kræver genstart af CoREST.[1]

**Skriftstørrelse:** Du kan indstille skriftstørrelsen ved at flytte skydeknappen »Font Size« mod venstre (mindre skrift) eller mod højre (større skrift). Skriftstørrelsen ændres umiddelbart. Vil du vende tilbage til systemets standardskriftstørrelse, klikker du på knappen »Default« til højre for skyderen.

**Visning af rullebjælker:** De fleste rullebjælker i siden (og bunden) af visninger kan skjules (eksperimentel funktionalitet). Denne mulighed kan især være interessant, hvis du styrer din computer med et pegefelt eller lignende i stedet for en mus. Hvis du ønsker at se færre rullebjælker, kan du fjerne hakket ved »Display all scrollbars«.

## 5.5 Sessionsprotokol

Sessionsprotokollen ses i højre halvdel af vinduet, se figur 5.2 på forrige side. Her registreres CoREST-klientens kommunikation med serveren: Der gives oplysninger fx om software- og korpusversioner og om de søgeudtryk, som CoREST har sendt til serveren. Her er simple søgeudtryk altid oversat til det tilsvarende avancerede udtryk.[2] Du kan kopiere udtryk fra protokollen og indsætte dem i søgefeltet for at udføre en bestemt søgning igen, nemmere er det dog at bruge listen »My Queries« til højre for konkordansområdet. Du kan læse mere om søgelister i kapitel 6.

## 5.6 Hvor gemmes systemoplysninger?

CoREST's indstillinger og systemoplysninger gemmes i en usynlig mappe `.CoREST` i roden af brugerens mappe. På en Mac er det under

`/Users/`*`brugernavn`*`/.CoREST`

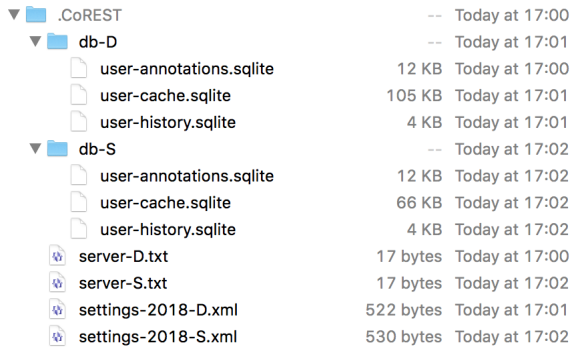
på en Linux-maskine under

`/home/`*`brugernavn`*`/.CoREST`

I din filbrowser skal visning af usynlige filer været slået til, før du kan se `.CoREST`-mappen. Figur 5.3 på næste side viser et eksempel på `.CoREST`-mappens indhold på en Mac – den er dog ganske tilsvarende opbygget i de andre styresystemer.

[1] Skærmbillederne her i manualen viser forskellige fremtoninger af CoREST, så der kan være forskel på, hvordan interfacet ser ud i din version af CoREST og illustrationerne her i manualen.

[2] Du kan læse mere om avanceret søgning i del III.




Figur 5.3: Filer med indstillinger og systemoplysninger

Nederst i strukturen ses to xml-filer:[3]

1. `settings-2018-D.xml`
2. `settings-2018-S.xml`

I de fleste tilfælde er der kun én fil med indstillinger til stede. Hvis du bruger forskellige udgaver af CoREST samtidig, fx både *DSL*- (suffiks `-D`) og *Standard*-udgaven (suffiks `-S`) eller forskellige versioner (= årgange, fx 2018, 2019, 2020) af CoREST er der indstillinger for hver version eller udgave. Det anbefales ikke, at du bruger flere versioner (= årgange) parallelt. Derimod er det uproblematisk at bruge forskellige udgaver parallelt, dog skal du være opmærksom på, at hver udgave ud over indstillingerne bruger sine egne annoteringer og sin egen søgehistorik.

[3] Årstallet i filnavnet afhænger af den CoREST-version, du bruger.

# Del II

# Brugerinterfacet i detaljer

# Kapitel 6

# Ordlister

*Søgninger, resultater, annotationer*

Dette kapitel omhandler søgelister med ord og andre søgeudtryk, som du kan udføre i CoREST, samt lister med matchene som resultat af din søgning.

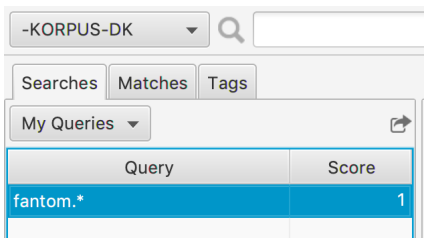
---



---

Til venstre for konkordansområdet i CoREST-vinduet finder du fanebladene »Searches« og »Matches«, se figur 3.1 på side 14, hvorunder der gemmer sig lister med henholdsvis søgninger, du kan udføre eller allerede har udført, og søgeresultater. Der findes endvidere lister med annoteringer under fanebladet »Tags«.

Du vælger et af fanebladene ved at klikke på dets overskrift, se figur 6.1, hvor fanebladet »Searches« er valgt.




Figur 6.1: Søgeliste med søgeudtrykket `fantom.*`

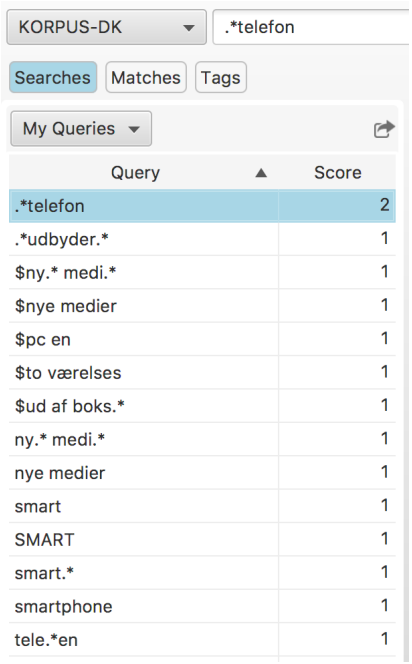
I figuren ses søgningen `fantom.*` i listen »My Queries« markeret med en blå bjælke. Når et ord eller et søgeudtryk er markeret på en liste, vil et tryk på returtasten medføre en søgning på det.

## 6.1 »Searches« – Lister med søgninger

Under fanebladet »Searches« finder du forskellige søgelister med ord eller søgeudtryk. Hvilke lister der er, afhænger af, hvilken udgave af CoREST du bruger, og hvilket korpus du har valgt. Men uanset hvilken udgave eller hvilket korpus du bruger, så vil der altid være en søgehistorik »My Queries«, som indeholder alle de søgninger, du selv har foretaget. I TiDK-korpusser er der ud over »My Queries« lister med månedens ord og årets ord.

### 6.1.1 Søgehistorik »My Queries«

Alle dine søgeudtryk, som gav et resultat, registreres i listen »My Queries« sammen med det antal gange, du har udført hver af disse søgninger, kolonnen »Score«. Normalt er listen sorteret sådan, at dine nyeste søgeudtryk står øverst på listen, mens de ældste står nederst. Du kan dog også sortere dem alfabetisk eller efter deres hyppighed ved at klikke på kolonneoverskrifterne »Query« eller »Score«. Gentagne klik skifter mellem de forskellige sorteringsmuligheder. Figur 6.2 viser en historik med simple søgeudtryk.




Figur 6.2: Søgehistorik

Historikken er sorteret alfabetisk, og det ses, at søgningen på udtrykket `.*telefon` er udført to gange, mens de øvrige kun er udført én gang hver. Sø-

gehistorikken er knyttet til det aktuelt valgte korpus. Skifter du korpus, skifter søgehistorikken også, så den altid indeholder de søgninger, som vedrører det aktuelt valgte korpus.

Du kan lave en søgning ud fra »My Queries«-listen ved at markere det ønskede søgeudtryk og taste retur. Du kan slette søgehistorikken under CoREST's indstillinger.

### 6.1.2 Fremtrædende ord

Bruger du et TiDK-korpus[1], vil der for hver måned og hvert år, hvor der har været tilstrækkeligt med tekstmateriale, være en liste over statistisk overrepræsenterede, fremtrædende, ord. Hver af disse lister indeholder typisk mellem 50 og 100 ord. Manglende måneder eller år skyldes utilstrækkelige data.

#### 6.1.2.1 Månedens ord

Månedens ord er ord, som blev brugt betydeligt hyppigere i den pågældende måned end ellers. Der er således tale om *overrepræsenterede ord*. Sommetider finder man også nye ord blandt disse ord.

Månedens ord kan også fortælle noget om årets gang. Tabel 6.1 viser som eksempel de tre mest fremtrædende ord for hver måned i 2011 og siger således noget om årets begivenheder sammenfattet i 36 stikord.

| Måned | Ord |
|---|---|
| Januar | *forbrugerminister, tunesisk, efterløn* |
| Februar | *statsløs, libysk, egypter* |
| Marts | *libysk, europagt, atomkatastrofe* |
| April | *påskeæg, skærtorsdag, påske* |
| Maj | *grænsekontrol, tryghedspakke, toldkontrol* |
| Juni | *grænseaftale, toldkontrol, folkemøde* |
| Juli | *gældsloft, skybrud, medieimperium* |
| August | *valgudskrivelse, betalingsring, vækstpakke* |
| September | *betalingsring, valgfest, regeringsforhandling* |
| Oktober | *regeringsgrundlag, rekapitalisering, løftebrud* |
| November | *narkogæld, vækstminister, finanslovsudspil* |
| December | *skattesag, uannonceret, undersøgelseskommission* |

Tabel 6.1: De tre mest prominente ord for alle månederne i 2011

[1] Du kan læse mere om de forskellige korpusser i CoREST i kapitel 20 og 21.

Hvor meget et ord er overrepræsenteret, udtrykkes ved en score, som beregnes ved at sammenligne samtlige tekster fra en bestemt måned (typisk ca. 3 mio. ords tekstmateriale fra Infomedia[2]) med samtlige tekster fra de 12 forudgående måneder. Beregningsalgoritmen er tilrettelagt sådan, at den både prøver på at ramme tidsånden og det leksikografisk interessante.

På listerne optræder ordene i lemmatiseret form, det vil sige i deres grundform, men deres score er beregnet på baggrund af alle deres bøjningsformer, som forekommer i materialet. Der er ord, som lemmatiseringsalgoritmen ikke kan genkende, enten fordi de er for nye eller for afvigende. Disse ord optræder ofte ulemmatiseret på listerne. Der kan også være ord, som af samme grund er lemmatiseret forkert, fx var *uannonceret* fra december måned i oversigten ovenfor oprindelig fejlagtigt (men til en vis grad logisk) blevet lemmatiseret som *(at) uannoncere* – disse fejl er forsøgt rettet manuelt i listerne, men der kan være enkelte tilbage. Typisk drejer det sig om følgende fejltyper:

**Egennavne og tal:** Selvom beregningsalgoritmen i princippet sørger for at se bort fra egennavne og tal, kan de sommetider alligevel slippe igennem, hvis de er lemmatiseret forkert i materialet.

**Orddele:** For enkelte måneder optræder der ord, som egentlig er dele af ord fremkommet ved orddeling. De kan fx stamme fra scannede og utilstrækkeligt efterbehandlede tekster.

**Tyske ord:** Sommetider kan der også optræde tyskklingende ord på listerne, som typisk stammer fra den tosprogede Flensborg Avis[3], som også leveres gennem Infomedia. Under korpusopbygningen forsøges det automatisk at fjerne tyske tekster fra materialet, men ikke alle kan identificeres i denne proces.

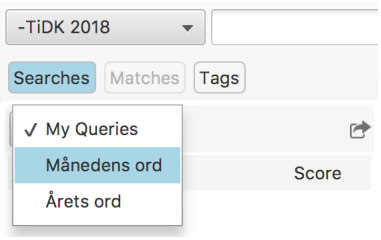
**Forkortelser:** Endelig kan der forekomme uforklarlige forkortelser, fx journalistinitialer.

For at se listerne med månedens ord i CoREST, skal du først vælge et korpus, der har denne type lister tilknyttet. Det gælder TiDK i alle udgaver af CoREST . Under fanebladet »Searches« klikker du på »My Queries« og vælger »Månedens ord« i rullemenuen, se figur 6.3.

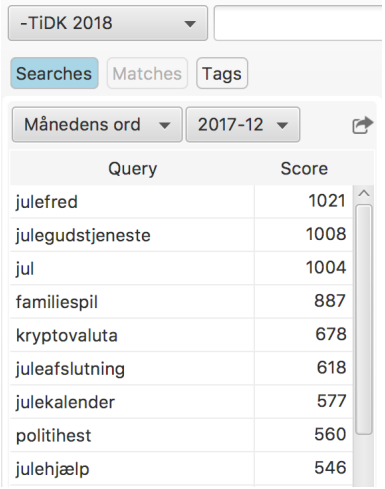
Du får automatisk vist den seneste måned, som der er beregnet fremtrædende ord for i pågældende korpus, men du kan vælge andre måneder i vælgeren, der bliver synlig til højre for ordliste-rullemenuen. Figur 6.4 viser toppen af listen for de mest fremtrædende ord for december 2017.

Du kan markere et ord i listerne ved at klikke på det. Taster du herefter retur, får du vist en konkordans, og du kan også se hyppigheden af ordet over tid. Konkordanser opstilles altid på baggrund af hele korpus, ikke kun den måned, som listen vedrører. Du kan læse mere om konkordanser i kapitel 7 og om grafiske oversigter i kapitel 12.

[2] https://infomedia.dk/
[3] https://www.fla.de




Figur 6.3: Vælge en liste med månedens ord




Figur 6.4: Fremtrædende ord for en bestemt måned

#### 6.1.2.2 Årets ord

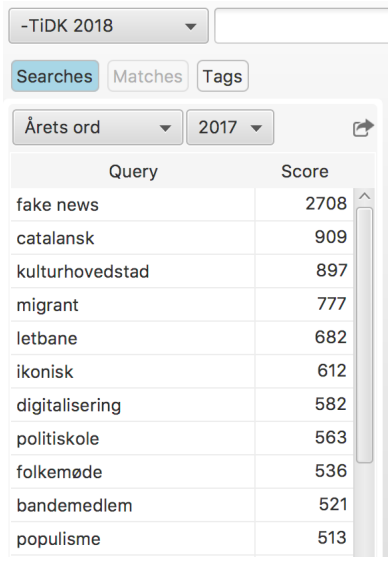
I TiDK findes der lister over årets ord for de år, som er med i korpusset. Du finder listen som et særligt punkt i søgeliste-rullemenuen, se figur 6.3. Årets ord beregnes efter de samme principper som månedens ord og kan derfor have de samme mangler. Figur 6.5 på næste side viser årets ord for 2017 i korpusset TiDK.

## 6.2 »Matches« – Lister med fund

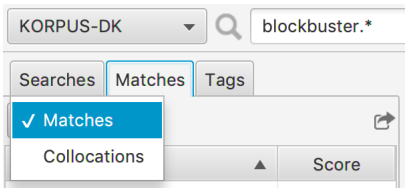
Under fanebladet »Matches« finder du dels listen over ord og udtryk, som matcher dit søgeudtryk, dels lister over hyppige naboord i konteksten omkring dine match i korpus. Du vælger mellem disse lister ved hjælp af *matchlistevælgeren*, en rullemenu, som du ser foldet ud i figur 6.6 på den følgende side.

Følgende lister er tilgængelige:

- ✧ Matchliste – vælg »Matches«, se afsnit 6.2.1 på næste side.
- ✧ Naboord – vælg »Collocations«, se afsnit 6.2.2 på side 40.




Figur 6.5: Fremtrædende ord for et bestemt år




Figur 6.6: Vælge blandt forskellige lister med fund

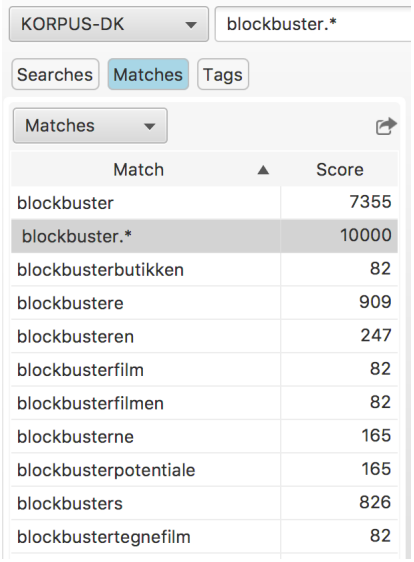
**OBS!** I listerne vises fundene altid med forenklet ortografi, dvs. at store bogstaver er omsat til små og accenttegn er fjernet.

Når du taster retur, mens en af listerne har fokus, vises korpusforekomster af det markerede fund i konkordansfeltet.

### 6.2.1 Matchliste

Matchliste-visningen får du frem ved at vælge fanebladet »Matches« i listeområdet, se figur 6.7 på næste side.

I matchlisten vises alle forskellige ord eller grupper af ord fra korpus, som matcher dit søgeudtryk. Blandt de første ord på listen står selve søgeudtrykket rykket lidt ind til højre, så det kan skelnes fra de andre ord på listen. Ud for hvert match vises en score, der angiver, hvor hyppigt pågældende match er blandt alle match. Alle match tilsammen sættes altid til 10.000 (svarende til 100 %). Prøv fx en søgning på `blockbuster.*` i KORPUS-DK og se, hvordan matchlisten kommer til at se ud.




Figur 6.7: Matchliste for `blockbuster.*`

**Oplysninger:** Søgeudtrykket `blockbuster.*` matcher alle forekomster af ord, der begynder med *blockbuster* inklusive *blockbuster* selv. Der er 121 forekomster i KORPUS-DK. Ordet *blockbuster* udgør 73,55 % af alle match (score 7355) svarende til 89 forekomster i korpus, mens fx *blockbusterbutikken* har 0,82 % af alle match (score 82), som i dette tilfælde svarer til én forekomst i korpus.

**Sortering:** Som udgangspunkt er matchlisten sorteret alfabetisk på match. Du kan sortere den efter score eller i en anden orden (stigende eller faldende) ved at klikke (evt. gentagne gange) på »Match« eller »Score« i titelbjælken over listen.

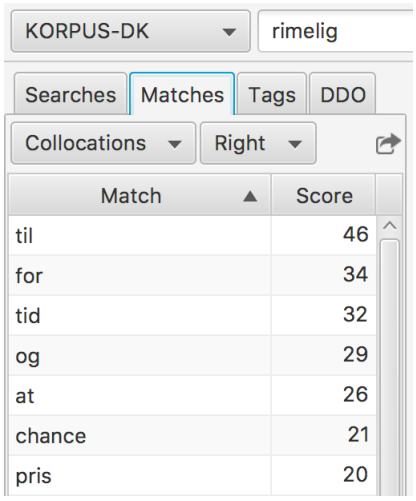
**Eksempler i konkordansen** Når matchlisten har fokus, kan du navigere fra linje til linje ved at trykke på tasterne *pil-op* eller *pil-ned*. Taster du *retur*, vises første forekomst af det markerede match i konkordansen, hvis konkordansen er sorteret på »Match«.

**OBS!** I listerne vises fundene altid med forenklet ortografi, dvs. at store bogstaver er omsat til små og accenttegn er fjernet.

### 6.2.2 Naboord

Et naboord er et ord, der optræder *hyppigt* i den umiddelbare omgivelse af et match. Vi kalder matchet sammen med dets hyppige naboord for *kollokation*. Den umiddelbare omgivelse af et match er i CoREST fastsat til to ord til højre for og to ord til venstre for matchet

En liste over naboord til dit aktuelle søgeudtryk får du frem ved, at du vælger »Collocations« fra matchlistevælgeren, se figur 6.6 på side 39. Når du stiller listevælgeren på »Collocations«, kommer der umiddelbart til højre for den en rullemenu frem, hvor du kan vælge den kontekst (»Right« eller »Left«), hvor den hyppige nabo skal stå, se figur 6.8.




Figur 6.8: Hyppige naboord til højre for *rimelig*

Som udgangspunkt står kollokationstypevælgeren på »Left«, hvilket betyder, at de hyppigst optrædende ord én til to pladser *til venstre* for matchet vises. I figur 6.8 er kollokationstypevælgeren imidlertid stillet på »Right«, her vises altså de hyppigst optrædende ord én til to pladser *til højre* for matchet *rimelig*. Som udgangspunkt vises kollokationslisten sorteret efter faldende score. Hyppighedsscoren oplyser ikke den absolutte hyppighed i korpus, men er et relativt tal ligesom i matchlisten.

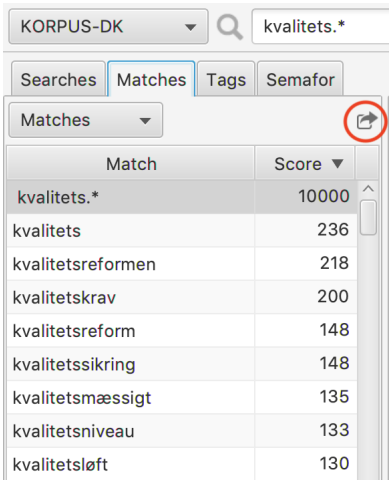
Når et ord er markeret i listen, og du taster *retur*, opretter CoREST en konkordans på kollokationen, fx *rimelig chance*, hvis *chance* var markeret i listen vist i figur 6.8.

## 6.3 »Tags« – Brugerannoteringer

Under »Tags« finder du lister med dine egne annoteringer af konkordanser i form af særlige markeringer, som du kan give dine fund. Du kan læse mere om annoteringer i kapitel 7 .

## 6.4 Eksportere ordlister

CoREST kan eksportere »Searches«- og »Matches«-lister som tekstfiler. Hertil klikker du på eksportknappen til højre over listen, jf. figur 6.9 på den følgende side, og vælger et filnavn og en placering i den dialogboks, der popper op.




Figur 6.9: Eksportere lister

En eksporteret liste består af to kolonner, den ene med selve ordet, den anden med den talværdi, der knytter sig til ordet. De to kolonner er adskilt med et tabulatortegn. Eksporterede lister kan nemt importeres i andre programmer, fx tekstbehandlere eller regneark.

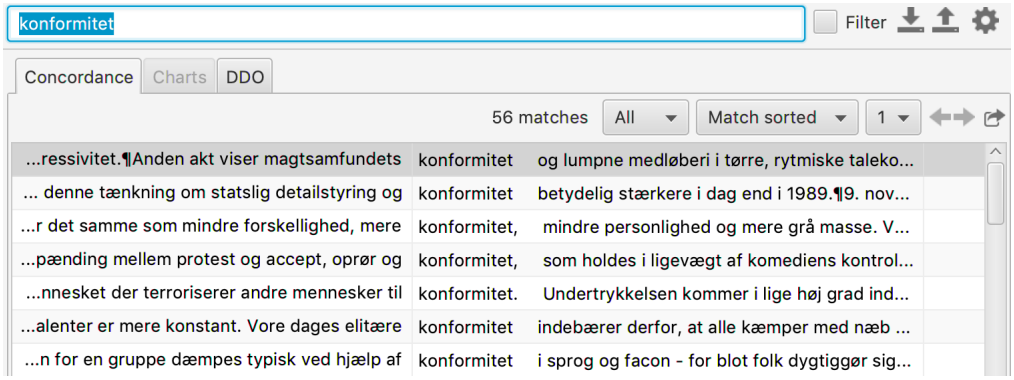
# Kapitel 7

# Konkordanser

*Gennemsyn, sortering, reduktion, markering, sletning, eksport*



En konkordans er i de fleste tilfælde det vigtigste resultat af en korpussøgning. En konkordans viser fund, der matcher søgeudtrykket, med noget kontekst omkring. Konkordanser vises i konkordansområdet, som er den centrale og største del af CoREST-vinduet til højre for området med ordlister. Hvert fund vises i midten af en linje med nogle ords kontekst til venstre og til højre for. Konkordansen er som udgangspunkt lige som matchlisten sorteret alfabetisk på matchet. I en konkordans kan der maksimalt vises 2000 fund fordelt på 20 sider med 100 fund per side. Hvis der er flere match i korpus, reduceres antallet af viste fund automatisk til 2000. Hvis en konkordans indeholder flere end 100 fund og dermed består af flere sider, kan du skifte fra side til side med sidevælgeren, se figur 7.4 på side 46. I konkordanser vises eksemplerne altid i en ortografi, som er så tæt på den oprindelige som mulig – i modsætning til lister med fund, som vises i en forenklet ortografi. Figur 7.1 på den følgende side viser et udsnit af en konkordans med fund, som matcher søgeudtrykket `konformitet` i KORPUS-DK.




Figur 7.1: Eksempel på en konkordans

Til konkordanser knytter der sig forskellige funktioner:

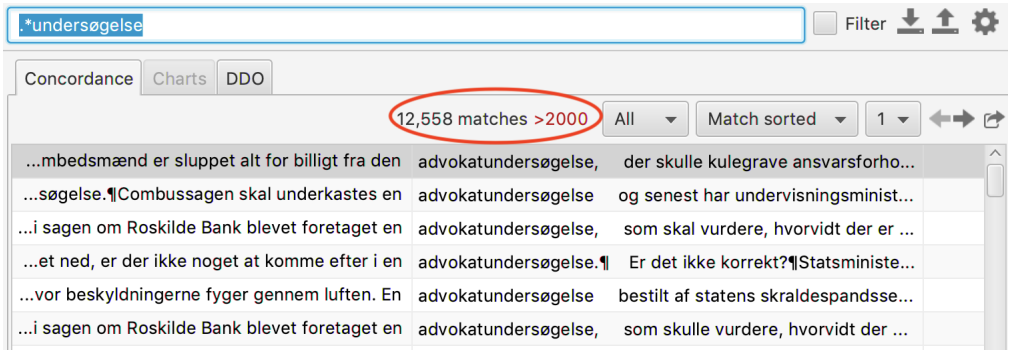
- **Automatisk reduktion** af antallet af konkordanslinjer, se afsnit 7.1
- **Manuel reduktion** af antallet af konkordanslinjer, se afsnit 7.2 på den følgende side
- **Gennemsyn** af konkordanser, se afsnit 7.3 på næste side
- **Sortering** af konkordanslinjer, se afsnit 7.4 på side 46
- **Markering og sletning** af konkordanslinjer, se afsnit 7.5 på side 47
- **Visning af markerede fund**, se afsnit 7.5 på side 47
- **Eksport** af konkordanser til brug uden for CoREST, se afsnit 7.6 på side 49
- Konkordanser **i sammensatte korpusser**, se afsnit 7.7 på side 50

## 7.1 Automatisk reduktion af en konkordans

I en konkordans kan der maksimalt vises 2000 fund fordelt på 20 sider med 100 fund per side. Hvis der er flere match i korpus, reduceres antallet af viste fund automatisk til 2000. Antallet af fund og en eventuel reduktion oplyses i bjælken over konkordansen. Figur 7.2 på næste side viser en konkordans fremkommet ved en søgning på `.*undersøgelse`.

I bjælken over konkordansen vises, at der er 12.558 match, men eftersom der maksimalt kan vises 2000 i CoREST, er konkordansen blevet automatisk reduceret til 2000 linjer. Dette oplyses ved tilføjelsen *>2000* med rødt.

En reduktion laves jævnt fordelt over hele korpus. Her er det med andre ord altså ikke kun de 2000 første fund, du får vist, men et udpluk taget fra hele korpus. Reduktionen udføres efter et tilfældighedsprincip, der dog er det samme for gentagne søgninger. På denne måde vil flere uafhængigt udførte søgninger altid medføre samme resultat.




Figur 7.2: Automatisk reduktion af konkordans

Hvis du vil se samtlige match, som dit søgeudtryk har i korpus, må du indsnævre søgningen. De første linjer i konkordansen i figur 7.2 er eksempler på brugen af formen *advokatundersøgelse*. Hvis du vil se flere eller samtlige forekomster af denne form i korpus, kan du fx søge på udtrykkene `advokatundersøgelse` eller `@advokatundersøgelse`, og du vil da få vist 38 hhv. 51 fund, altså samtlige forekomster i korpus, da der nu er under 2000 i alt.

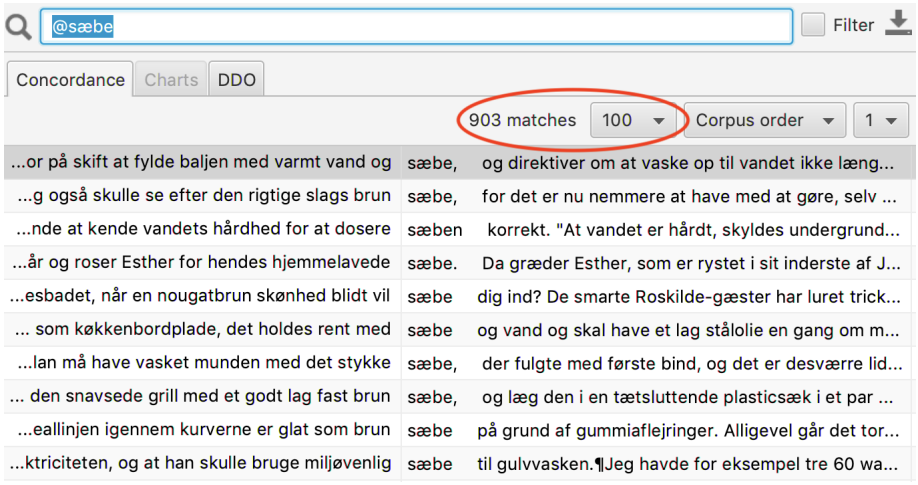
## 7.2 Manuel reduktion af en konkordans

Du kan også reducere en konkordans manuelt til 100, 200, 300, 400 eller 500 tilfældige linjer ved hjælp af en særlig rullemenu over konkordansen. Linjerne bliver ganske vist tilfældigt udvalgt, men det er det samme tilfælde, der er på spil for hver søgning, så resultatet bliver reproducerbart på samme måde som ved den automatiske reduktion, sml. afsnit 7.1. I meget store konkordanser kan du på denne måde nøjes med at få vist et overskueligt antal eksempler, som udtrækkes jævnt fordelt over hele korpusset. Figur 7.3 på næste side viser et udsnit af en konkordans på `@sæbe`, som er manuelt reduceret til 100 eksempler.

## 7.3 Gennemsyn af en konkordans

Du kan blade frem og tilbage fra side til side i konkordansen ved at klikke på pileknapperne øverst til højre over konkordansen, eller du kan springe direkte til en bestemt side ved at vælge den i rullemenuen umiddelbart til højre for pileknapperne, jf. figur 7.4 på den følgende side.

I rullemenuen oplyses den aktuelt viste side. Når konkordansområdet har fokus, kan du også blade frem og tilbage ved at bruge *pil-venstre* og *pil-højre* på dit tastatur.




Figur 7.3: Manuel reduktion af konkordans


Figur 7.4: Blade i konkordans

## 7.4 Sortering af en konkordans

Linjerne i en konkordans kan sorteres i forskellige rækkefølger. Du kan vælge sorteringen i sorteringsrullemenuen, som er markeret med rødt i figur 7.5:


Figur 7.5: Valg af sorteringskriteriet

Som udgangspunkt er linjerne sorteret alfabetisk på sorteringskriteriet *match* (»Match sorted«), altså på samme måde som matchlisten. Andre sorteringskriterier er »Corpus order«, »Left sorted« og »Right sorted«.

**»Match sorted«** Alfabetisk sortering på matchet. Dette er standardsorteringen.

**»Corpus order«** Med denne sortering vises eksemplerne i den rækkefølge, som de har i korpusset. Rækkefølgen afspejler dermed opbygningen af korpus. For alle CoREST-korpusser er rækkefølgen kronologisk, således at de nyeste eksempler står øverst, og eksemplerne bliver ældre, jo længere du ruller ned gennem konkordansen.

**»Left sorted«** Sortering på venstre kontekst. Sorteringen sker fra højre mod venstre, altså baglæns, således at ord, der ender på *a*, kommer før ord, der ender på *b*, osv.

**»Right sorted«** Sortering på højre kontekst. Sorteringen sker fra venstre mod højre, således at ord, der begynder med *a*, kommer før ord, der begynder med *b*, osv.

## 7.5 Markering og sletning

Du kan give matchene i konkordansen egne markeringer. På denne måde kan du fx

- ✧ holde **forskellige betydninger** af et match ud fra hinanden eller de forskellige sproglige konstruktioner, som matchet indgår i,
- ✧ markere **gode eksempler** eller sproglige fejl af forskellig art.

I det følgende kalder vi disse markeringer *mærker*.

### 7.5.1 Markere match i konkordansen

Du markerer et match i konkordansen ved at give det et *mærke*. Mærker kan være et bogstav (*A-Z*) eller et ciffer (*0-9*), og de vises ud for konkordanslinjerne i kolonnen til højre for konkordansen. Hvilke bogstaver eller cifre du vil bruge til at markere bestemte egenskaber ved konkordanslinjen med, er op til dig. Fx kan du give gode eksempler et *A* som mærke og give forskellige betydninger af et ord cifrene *0–9*. Der kan dog kun sættes ét mærke på hvert match. Sætter du et mærke på et match, der har et i forvejen, bliver det oprindelige erstattet af det nye.

**Sætte mærker** Klik på konkordansområdet for at give det fokus eller brug tabulatortasten. Gå med piletasterne til den linje i konkordansen, som du vil forsyne med et mærke, eller klik på linjen med musen. Tast dernæst et bogstav eller et ciffer. Matchet i konkordanslinjen er nu mærket.

**Slette konkordanslinjer** Et særligt mærke er bindestregen (*minus*), som 'sletter' eller snarere undertrykker konkordanslinjen. En undertrykt konkordanslinje vises nedtonet. Funktionen er især nyttig ved eksport af konkordanslinjer: Linjer, der er mærket med en streg, kommer nemlig ikke med i eksporten. Se mere om eksport af konkordanser i afsnit 7.6 på side 49.

**Fjerne mærker** Du kan fjerne et mærke igen ved at markere pågældende linje og taste mellemrum. På denne måde kan du også 'genoplive' undertrykte linjer.

Figur 7.6 på næste side viser et lille udsnit af en konkordans på `blad`, hvor match i betydningen 'tidsskrift' er markeret med *1*, mens *2* markerer match i betydningen 'plantedel'. Læg mærke til, at tre linjer er blevet undertrykt.

Hvis dit match omfatter flere ord, knyttes mærket altid til hele matchet. For eksempel finder søgeudtrykket `@reducere .*forbrug` blandt andet matchet *reducere bygningens energiforbrug* i KORPUS-DK. Hvis du nu sætter et mærke, for

| | | | |
|---|---|---|---|
| ...e at han havde stukket dem et Anders And- | blad, | men der var nu bøger nok at kigge i - to sangbø... | 1 |
| ...alsekretæren fx var klar til at sætte det nye | blad | for frivillige i søen, måtte han lige vente på, at 25... | 1 |
| ...r noget skal undersøges. I artiklen i forrige | blad | blev der ikke én gang refereret til en kvinde. Find... | 1 |
| ...talte hun åbenhjertlig i et interview sit eget | blad | O Magazine.¶Syv år efter sin runde fødselsdag e... | 1 |
| ...ER HAM I FREDERICIA, SAVNER DET GODE | BLAD.¶ | Lykke Petersen, fra Hus Forbis Facebookprofil... | - |
| ...ren. På et tidspunkt vil livets træ slippe det | blad, | der er mig. Og man kan ikke sætte et blad på ige... | 2 |
| ... blad, der er mig. Og man kan ikke sætte et | blad | på igen, så jeg vil svæve ind i den store helhed. U... | 2 |
| ...en.¶På et tidspunkt vil livets træ slippe det | blad, | der er mig.¶Og man kan ikke sætte et blad på ig... | - |
| ...blad, der er mig.¶Og man kan ikke sætte et | blad | på igen, så jeg vil svæve ind i den store helhed.¶... | - |
| ... Aarhus-kontoret som redaktør af Merkurs | blad | Pengevirke - et blad om bæredygtig økonomi og l... | 1 |

Figur 7.6: Konkordansudsnit med mærker

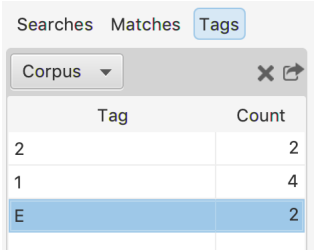
eksempel *E* ud for pågældende konkordanslinje, vedrører det kun dette match. I en efterfølgende søgning på `$bygningens energiforbrug` vil du se, at der ikke længere er et mærke ud for pågældende konkordanslinje, fordi *reducere* nu ikke længere er en del af matchet. Det samme gælder for eksempel for en efterfølgende søgning på `reducere bygningens`. Først når *reducere bygningens energiforbrug* matcher i sin helhed, vises mærket igen, uanset hvilket søgeudtryk der ellers medfører dette match.

**OBS!** Markeringer gemmes på din lokale computer. Derfor bør du altid arbejde med samme computer eller som samme bruger, når du anvender markeringer. Markeringer kan blive ugyldige i forbindelse med større korpusopdateringer, hvorfor det kan være en fordel at eksportere konkordanser 7.6 med mærker i forbindelse med større og længerevarende projekter!

### 7.5.2 Vise konkordanslinjer med markerede match

Samtlige mærker, som du bruger i dine konkordanser, samles i en liste, som du kan se ved at vælge listefanebladet »Tags«. Inden vi ser nærmere på denne liste, vil vi opstille endnu en konkordans, dennegang på søgeudtrykket `@øge .*forbrug`. Lidt længere nede i konkordansen findes matchet *øge effektiviteten i vores energiforbrug* – markér linjen og mærk den med *E*, altså med samme mærke som i matchet *reducere bygningens energiforbrug* fra forrige afsnit. Nu vælger du »Tags« i listefanebladet. Du ser nu en liste som vist i figur 7.7 på den følgende side.

På listen ser du som udgangspunkt mærkerne for hele korpus. Hvis der er tale om et sammensat korpus, jf. kapitel 3 og kapitel 19, vises dog kun mærkerne for det valgte delkorpus, ikke det samlede korpus. I figur 7.7 på næste side er der kun mærkerne *2*, *1* og *E*, som vi indtil nu har brugt i eksemplerne ovenfor. Taster du retur, når et mærke er markeret på tag-listen (fx *E* i figur 7.7 på den følgende side), vises alle konkordanslinjer, som er mærket med det pågældende mærke. På denne måde er det muligt at se beslægtede (det vil sige identisk mærkede)




Figur 7.7: Liste over mærker

konkordanslinjer samlet, selv hvis de oprindeligt stammer fra forskellige konkordanser fremkommet ved forskellige søgninger, se figur 7.8.


Figur 7.8: Konkordanslinjer med samme mærke

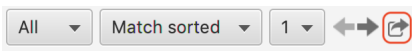
Ud over markeringer for hele korpus kan annoteringslisten også nøjes med at vise markeringer, som kun gælder den aktuelle konkordans. Du ændrer visningen fra *hele korpus* til *aktuel konkordans* ved at vælge »KWIC« i stedet for »Corpus« i vælgeren til venstre over listen, se figur 7.7 og 7.8.

### 7.5.3 Særligt om markerede match i sammensatte korpusser

Hvert delkorpus i et sammensat korpus, jf. kapitel 3 og kapitel 19, udgør under mærkningen et selvstændigt korpus. Visningen »Corpus«, som viser samtlige annoteringer for et korpus, viser i dette tilfælde altså ikke mærkerne for det samlede sammensatte korpus, men kun for det aktuelt valgte delkorpus.

## 7.6 Eksport af konkordanser

Du kan eksportere en konkordans i form af en tekstfil, idet du klikker på eksportknappen yderst til højre i bjælken over konkordansen, se figur 7.9, hvor knappen er markeret med en rød ramme.




Figur 7.9: Eksport af en konkordans

Tekstfilen er struktureret sådan, at hver linje svarer til en konkordanslinje, hvor matchet altid er omgivet af tabulatortegn. Konteksten til højre og til venstre for matchet er begrænset til 20 ord. Linjen (højre kontekst) afsluttes med et tabulatortegn evt. efterfulgt af et mærke, jf. afsnit 7.5 på side 47. Vær opmærksom på, at konkordanslinjer, som er markeret med en streg (undertrykte konkordanslinjer) ikke kommer med i eksporten. De er med andre ord slettet fra den eksporterede konkordans. En linje i en eksporteret konkordans er struktureret på følgende måde:

```
[VENSTRE KONTEKST] <TAB> [MATCH] <TAB> [HØJRE KONTEKST] <TAB> [EVT. MÆRKE]
```

Eksporterede konkordansfiler kan nemt importeres i tekstbehandlingssystemer, regneark, eller de kan behandles videre på anden vis. Figur 7.10 viser det mærkede konkordansafsnit fra figur 7.6 på side 48 importeret i et regneark.

| | A | B | C | D |
|---|---|---|---|---|
| 1 | mio. kr. til pædagogerne - men kostede forbundet BUPL 347 mio. kr., fr | blad | Børn & Unge. Af de 347 mio. kr. har BUPL's medlemmer ekstraordinært beta | 1 |
| 2 | stedet meget optaget af udviklingsrummet.¶Da generalsekretæren fx v | blad | for frivillige i søen, måtte han lige vente på, at 250 databaser ude i lokalafdel | 1 |
| 3 | at Højskolebladet orienterer sig efter mandlige synspunkter, når noget s | blad | blev der ikke én gang refereret til en kvinde. Findes der ikke kvindelige forstar | 1 |
| 4 | jeg sulten efter balance og intet andet, fortalte hun åbenhjertlig i et inter | blad | O Magazine.¶Syv år efter sin runde fødselsdag er Oprah større end længe m | 1 |
| 5 | efterår og på vej ind i vinteren. På et tidspunkt vil livets træ slippe det | blad, | der er mig. Og man kan ikke sætte et blad på igen, så jeg vil | 2 |
| 6 | vil livets træ slippe det blad, der er mig. Og man kan ikke sætte et | blad | på igen, så jeg vil sværve ind i den store helhed. Uden tanke.¶- UNDEF Nu | 2 |
| 7 | som et mål i sig selv. Platz arbejder fra Aarhus-kontoret som redaktør af | blad | Pengevirke - et blad om bæredygtig økonomi og levevis, som sendes ud til ar | 1 |

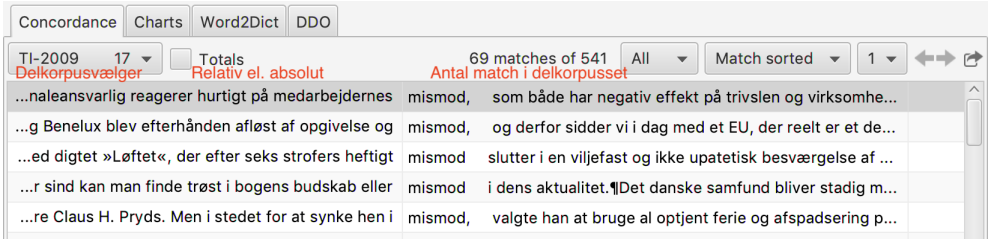
Figur 7.10: Eksporteret konkordansudsnit i regneark

## 7.7 Konkordanser i sammensatte korpusser

CoREST kan søge i flere typer korpusser: unikorpusser (standard) og multikorpusser, som er sammensatte af flere unikorpusser, jf. kapitel 19. Multikorpusser kan i CoREST kendes på, at deres navn begynder med en bindestreg. Konkordanser i multikorpusser vises på en lidt anden måde end i unikorpusser, idet der vises én særskilt konkordans for hvert delkorpus, som multikorpusset er sammensat af. Som det første resultat vises altid konkordansen for det delkorpus, som i forhold til dets størrelse har flest forekomster af matchet. Du skifter til et andet delkorpus ved at vælge det i delkorpusvælgeren til venstre i værkstøjsbjælken over konkordansen. Hvis der mangler delkorpusser i vælgeren, skyldes det, at der ikke er forekomster af det søgte i pågældende delkorpus.

I CoREST har du adgang til en række korpusser, hvis navne er sammensat af -TiDK (med bindestreg foran) og et årstal. TiDK-korpusserne er multikorpusser, som hvert består af ti delkorpusser, ét for hvert af de ti år forud for årstallet, som er anført i korpussets navn. Prøv at vælge et TiDK-korpus med korpusvælgeren øverst til venstre i CoREST-vinduet, fx -TiDK 2018. Udfør nu en søgning på søgeudtrykket `@mismod` (alle bøjningsformer af grundformen *mismod*). Hvis du har valgt multikorpusset -TiDK 2018, vil der som umiddelbart resultat blive vist en konkordans fra delkorpusset TI-2009, se figur 7.11 på den følgende side.

Dette delkorpus indeholder udelukkende tekster fra 2009. Konkordansen bliver lavet på dette delkorpus, fordi der her er relativt flest forekomster af det søgte. Tallet *17* i delkorpusvælgeren fortæller, hvor tit ordet optræder per 10 millioner ords tekst i det pågældende delkorpus (altså i året 2009); dette er ordets relative




Figur 7.11: Konkordansudsnit fra delkorpus i et sammensat korpus

hyppighed, og den svarer til 69 faktiske forekomster i dette delkorpus, da det er på ca. 39,5 millioner ord. Antallet af faktiske forekomster vises lidt længere mod højre i bjælken over konkordansen i oplysningen *69 matches of 541* – der er 541 forekomster i det samlede -TiDK 2018 fordelt over samtlige delkorpusser.

Du kan klikke på delkorpusvælgeren og her få et hurtigt overblik over, hvordan udbredelsen af match på dit søgeudtryk har været igennem de seneste ti år, og du kan vælge et andet år og se eksempler herfra.

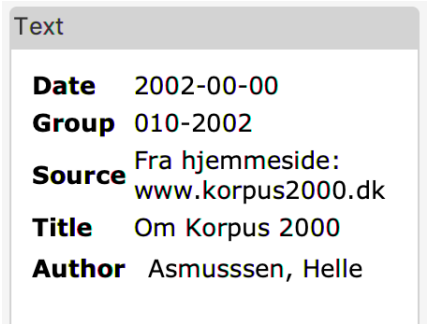
I stedet for som udgangspunkt at få vist eksempler fra det delkorpus, hvor der er relativt flest forekomster af et match, altså hvor frekvensen er højest, kan du også få vist eksempler fra det delkorpus, hvor antallet af faktiske forekomster er højest. Til dette skal du sætte hak i tjekboksen »Totals«. I eksemplet her vil du da få vist eksemplerne fra TI-2012 som det første.

# Kapitel 8

# Tekstoplysninger

*Forfatter, titel, udgivelsesdato m.m.*

Tekstoplysninger vises i det første af de tre felter under konkordansen, se området *Tekstoplysninger* i figur 3.1 på side 14. Her vises altid oplysninger om den tekst, som den aktuelt markerede konkordanslinje stammer fra. Figur 8.1 viser de oplysninger, som vedrører den markerede (første) konkordanslinje i konkordansen i figur 3.1 på side 14.




Figur 8.1: Tekstoplysninger

Tabel 8.1 på den følgende side forklarer de forskellige oplysningstyper nærmere, som du i øvrigt også kan bruge i dine søgninger, se kapitel 18.

| **Oplysning** | **Beskrivelse** |
|---|---|
| Date | Udgivelsesdatoen for teksten i formatet *åååå-mm-dd*. Hvis måned og dag er sat til *00-00*, er det kun årstallet, der er relevant. |
| Group | Tekstgruppe, som denne tekst tilhører. Teksterne i et korpus er inddelt i grupper af beslægtede tekster. |
| Source | Kilden, som teksten stammer fra, fx navnet på en hjemmeside eller en avis. |
| Title | Tekstens titel. |
| Author | Tekstens forfatter. |

Tabel 8.1: Oversigt over centrale tekstoplysninger

# Kapitel 9

# Kontekst

*Se mere eller mindre kontekst omkring fundet*

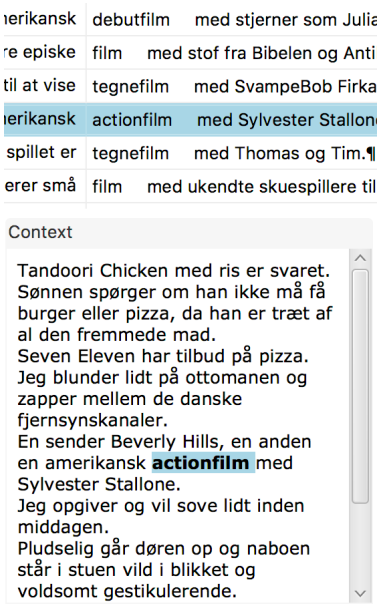


Konkordanslinjer er altid bygget op, sådan at selve matchet vises i midten af linjen med nogle ords kontekst omkring, se figur 9.1, som viser et udsnit af konkordansen (sorteret på højre kontekst, jf. kapitel 7) efter en søgning på på `@.*film` i KORPUS-DK.


Figur 9.1: Kontekster i en konkordans

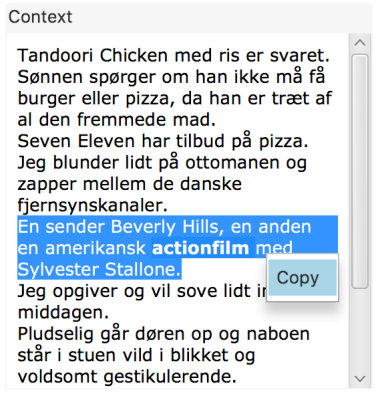
## 9.1 Se mere kontekst

Noget mere kontekst, end hvad der kan vises i selve konkordanslinjen, får du i området med kontekstoplysningerne under konkordansen, i det andet felt fra venstre, se området *Kontekst* i figur 3.1 på side 14. Selve matchet i dette felt er markeret med en særlig farve og står som regel mere eller mindre i midten af feltet, se eksemplet *actionfilm* i figur 9.2 på næste side.




Figur 9.2: Se mere kontekst til markeret linje

Fra dette felt kan du kopiere en tekstpassage på samme måde, som du kan i de fleste andre programmer på din computer, fx ved at markere det område, du vil have med, højreklikke med musen og vælge »Copy« i den lokale menu, der popper op. Du kan herefter indsætte den kopierede passage i andre programmer, fx i en tekstbehandling eller et ordbogsredigeringssystem. Figur 9.3 viser det med en passage omkring matchet *actionfilm* (i dette tilfælde efter en søgning på `@.*film` i KORPUS-DK).




Figur 9.3: Kopiering fra kontekstfeltet

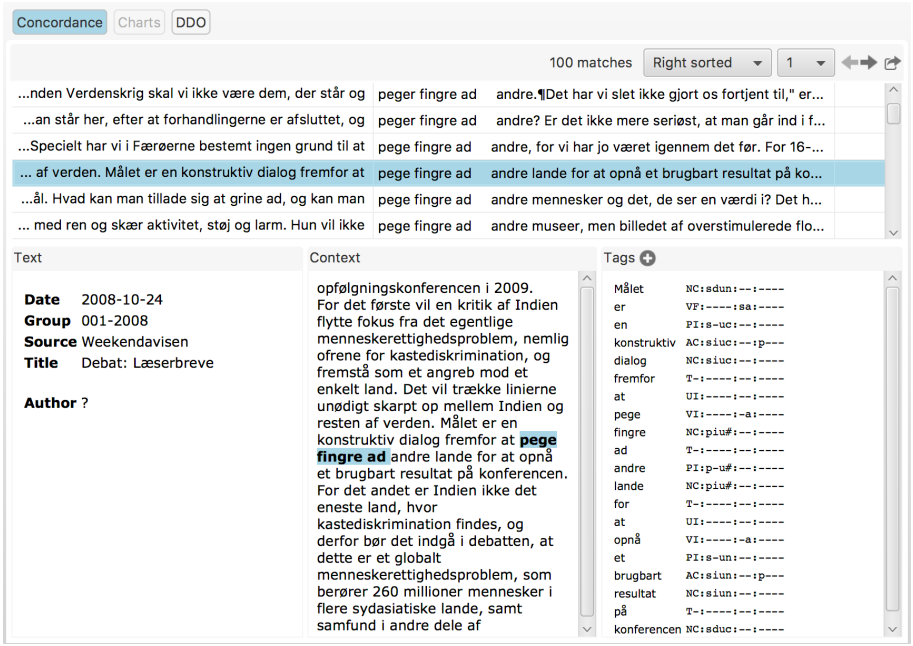
# Kapitel 10

# Ordoplysninger

*Lemma, ordklasse og bøjningsoplysninger*



Til hvert ord i et korpus er der knyttet en række oplysninger, som du også kan bruge i dine søgninger, se kapitel 15. Ordoplysningerne står i feltet med overskriften »Tags« under konkordansen. Figur 10.1 viser et eksempel efter en søgning på `$@pege fingre ad` i KORPUS-DK .




Figur 10.1: Ordoplysninger om ordklasse og bøjning

Her vises alle ord fra den sætning, som fundet i den aktuelt markerede konkordanslinje hører til, som ét ord per linje, sammen med ordklasse- og bøjningsoplysninger. Ordklasse- og bøjningsoplysningerne gives som en kode, et såkaldt *tag*, med følgende opbygning:

```
OU:nnnn:vv:dddd
```

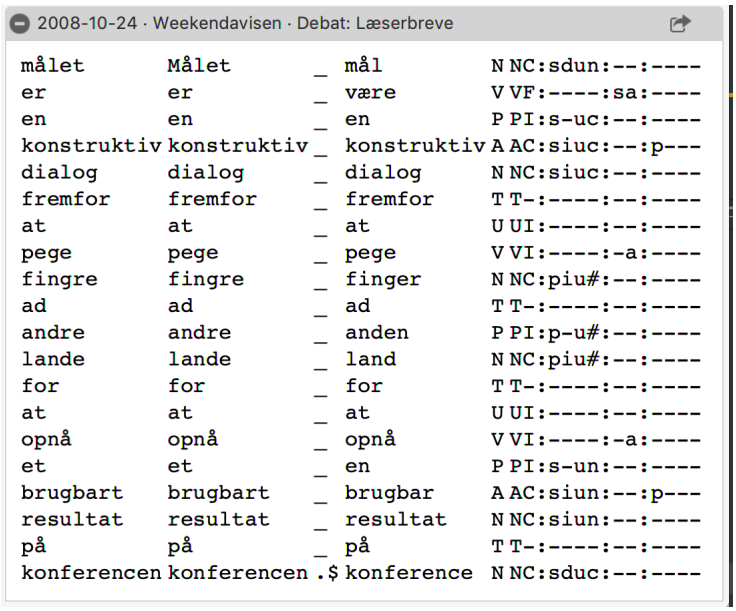
Her har de enkelte dele følgende betydninger:

- `O` – Ordklassen, fx `N` for substantiv, `V` for verbum, `A` for adjektiv.
- `U` – Underklasse til ordklassen, hvis det er relevant (ellers står der blot en streg på denne plads); ved verber fx `F` for finit, `I` for infinitiv, `M` for imperativ.
- `nnnn` – På de fire pladser betegnet `n` står der oplysninger, der vedrører den nominale bøjning af ordet.
- `vv` – De to `v`-pladser giver oplysninger om den verbale bøjning.
- `dddd` – De fire afsluttende `d`-pladser bruges til forskellige oplysninger, fx om adjektivers gradbøjning eller pronomeners person.

En detaljeret gennemgang af ordklassetaggets opbygning finder du i kapitel 16. Her beskrives også, hvordan du kan bruge disse oplysninger i dine søgninger.

## 10.1 Udvidede ordoplysninger

Klikker du på *plus*-symbolet i panelet over ordoplysningsfeltet »Tags«, se figur 10.1 på foregående side, får du vist samtlige ordoplysninger. Dette er vist i figur 10.2. Oplysningerne falder i følgende fem kolonner:




Figur 10.2: Ordoplysninger i CoREST-korpusser

**1. kolonne:** Ordet i forenklet ortografi; det er denne form, som der som udgangspunkt søges på i CoREST, og som vises i resultatlisterne, se kapitel 6.

**2. kolonne:** Ordet så tæt på den ortografi, det havde i den oprindelige korpustekst.

**3. kolonne:** Tegn, der adskiller ordet fra det følgende ord til højre for i teksten; som regel er der blot tale om et mellemrum, som gengives som en understregningsstreg i visningen, men ofte vil der også være sætningstegn, således følger der *.$* (punktum og dollartegn) efter ordet *konferencen* i eksemplet i figur 10.2 på foregående side. *$*-tegnet betyder linjeskift, det indikerer altså som regel afslutningen på et afsnit.

**4. kolonne:** Ordet i dets grundform (*lemmaform*), dvs. den form, man vil slå det op under i en ordbog.

**5. kolonne:** Ordets ordklasse; der bruges forkortelser bestående af ét stort bogstav. En komplet liste over anvendte forkortelser og deres betydninger findes i kapitel 16.

**6. kolonne:** Udvidede ordklasse- og bøjningsoplysninger. Læs mere herom i kapitel 16.

Du kan eksportere de udvidede ordklasseoplysninger i form af et html-dokument ved at klikke på eksportknappen yderst til højre i panelet.

Du kan læse mere om, hvordan et ord er defineret i CoREST i kapitel 17, og hvordan du kan bruge ordoplysningerne i avancerede søgninger i kapitel 15.

# Kapitel 11

# Brugerannoteringer

*Annotering af fund i DSL- og Research-udgaven*

I CoREST kan du markere fund i konkordanser med selvvalgte bogstaver eller cifre, se kapitel 7. På denne måde kan du bedre skelne de konkordanslinjer, der har noget tilfælles, fra de øvrige. En mere avanceret måde at markere bestemte match på er at bruge annoteringer. Annotering er kun mulig i DSL- og Research-udgaverne af CoREST. Du kan læse mere om annotering i den manual, som hører til DSL-udgaven af CoREST.

# Kapitel 12

# Grafiske oversigter

*Hyppigheder gennem tiden*



Når du søger i et sammensat korpus – nærmere bestemt en særlig type sammensatte korpusser, som vi kalder *multikorpusser*[1] – kan du få en grafisk visning af fordelingen af match på de forskellige delkorpusser, som pågældende multikorpus består af. Grafiske oversigter kan vises for alle typer søgninger, også når der er brugt jokertegn, se kapitel 14, eller avancerede søgeudtryk, se kapitel 15, herunder filtre, som beskrives i kapitel 18.

Både -KORPUS-DK[2] og de forskellige årgange af TiDK-korpusserne er multikorpusser. TiDK-korpusserne, som du kan læse mere om i kapitel 20, afspejler sproget gennem de forudgående ti år i forhold til årstallet i det pågældende korpus' navn. Version 2020 af dette korpus indeholder således materiale fra årene 2010 til 2019.[3] -KORPUS-DK, som du også kan læse mere om i kapitel 20, er sammensat af tre korpusser, nemlig KORPUS-90 med tekster fra omkring 1990, KORPUS-2000 med tekster fra omkring år 2000 og KORPUS-2010 med tekster fra omkring 2010. Delkorpusser kan normalt ikke vælges som selvstændige søgekorpusser i korpusvælgeren.

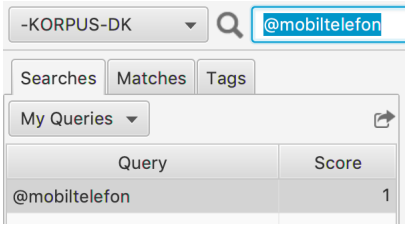
## 12.1 Eksempler

Lad os se på et eksempel i multikorpusset -KORPUS-DK (med streg foran), som du vælger ved hjælp af korpusvælgeren øverst til venstre i CoREST-vinduet. Prøv at udføre en søgning på `@mobiltelefon`, som vist i figur 12.1 på næste side.

[1] Du kan læse mere om de forskellige korpustyper i kapitel 19.

[2] Bindestregen foran navnet indikerer, at der er tale om et sammensat korpus, se afsnit 3.2.1.2 på side 17.

[3] Man kalder et korpus, som til enhver tid viser et *tidsbegrænset* udsnit af sproget, fx et fast antal år før nu, for et *monitorkorpus*.




Figur 12.1: Søgning i sammensat korpus

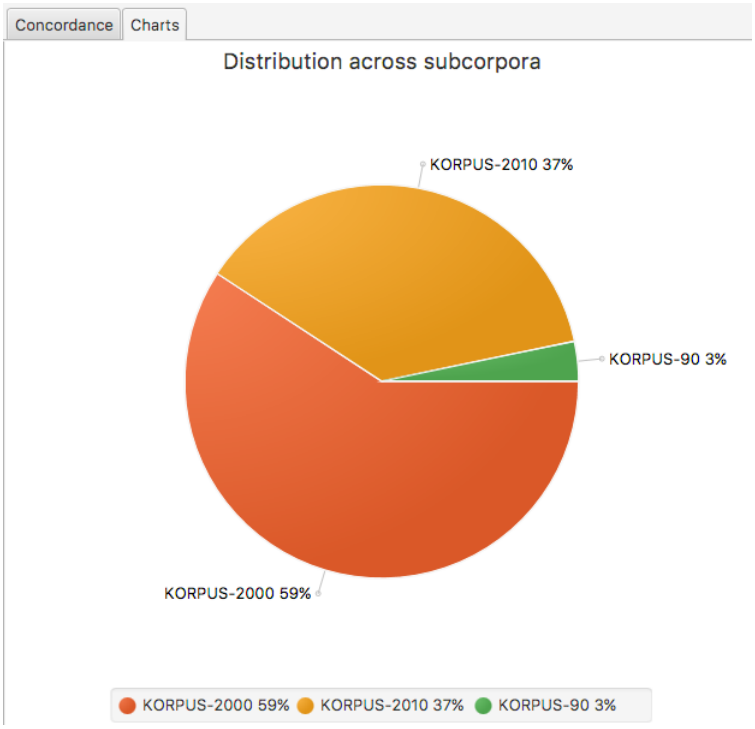
CoREST viser som resultat en konkordans fra delkorpusset KORPUS-2000, hvor der i forhold til delkorpussets størrelse er flest forekomster af dette ord, nemlig 506 per 10 millioner ord svarende til 1530 faktiske forekomster. Klikker du på delkorpusvælgeren, vil du se tallene for alle tre delkorpusser, som -KORPUS-DK er sammensat af, se figur 12.2.


Figur 12.2: Delkorpusvælger: Fordeling af forekomster

Et bedre overblik får du ved at klikke på fanebladet »Charts« over konkordansen. Under dette faneblad får du fordelingen af det relative antal forekomster i de enkelte korpusser vist grafisk i form af et lagkagediagram, jf. figur 12.3 på næste side.

På figuren ser man hurtigt, at mobiltelefoner åbenbart var mest udbredte i tekster fra omkring år 2000 med hele 59 % af samtlige relative forekomster. Det samlede procenttal for lagkagediagrammet kan sommetider afvige fra 100 på grund af afrundingen af de enkelte tal. Farverne i lagkagediagrammet vælges automatisk og kan ændre sig fra søgning til søgning. Farverne kan ikke fastlægges af brugeren. Fanebladet »Charts« kan kun vælges, hvis du arbejder med et sammensat korpus, ellers er det gråned ud og ude af funktion.

Mens du i multikorpusset -KORPUS-DK altid får vist lagkagediagrammer, der viser fordelingen af det relative antal af forekomster på de tre delkorpusser, består TiDK-korpusserne altid af ti delkorpusser, nemlig ét for hvert af de seneste ti år forud for det årstal, som er anført i TiDK-korpussets navn. Prøv at vælge et TiDK-korpus med korpusvælgeren, fx -TiDK-2019, og udfør dernæst en søgning på `tablet`. Klik herefter på fanebladet »Charts«. Du får nu vist en graf, som for hvert delkorpus, altså for hvert af de seneste ti år ($x$-aksen) viser antal forekomster per 10 millioner løbende ord ($y$-aksen). Du vil se, at der er en markant stigning i antallet af forekomster i årene fra 2010 til 2013. Mens `tablet` forekom 8 gange per 10 millioner ords løbende tekst i 2010, lå tallet på 83 for 2013, altså en tidobling, hvorefter hyppigheden begynder at falde igen. Prøv nu at udføre




Figur 12.3: Diagram over fordelingen af `@mobiltelefon` på tre delkorpusser

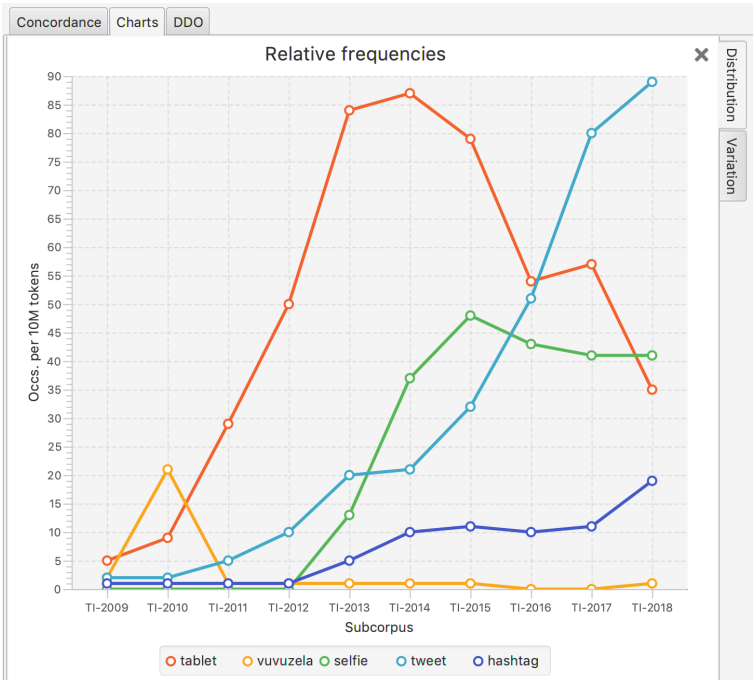
følgende søgninger: `vuvuzela`, `selfie`, `tweet` og `hashtag`. Du vil se, at de nye grafer tilføjes til de allerede eksisterende og at koordinatsystemet skaleres løbende, så alle grafer kan vises. Efter søgningen på disse ord ser grafen omtrent ud, som vist i figur 12.4 på den følgende side

Kurvernes farver fastlægges automatisk af CoREST og kan ikke ændres af brugeren. Du kan slette kurverne fra oversigten ved at klikke på sletteknappen (i form af et *X*) øverst til højre over koordinatsystemet.

Under »Charts« kan du vælge mellem to slags grafiske visninger ved at klikke på et af fanebladene »Distribution« eller »Variation« til højre for grafen:

**Distribution** viser hyppigheden af matchet i hvert delkorpus som en kurve. Da der er ét delkorpus for hvert år, og da delkorpusserne er ordnet kronologisk, viser kurven udviklingen i hyppigheden af pågældende match gennem årene. Hyppigheden vises som antal forekomster per 10 millioner ord i pågældende korpus.

**Variation** viser for hvert delkorpus afvigelsen til den gennemsnitlige hyppighed for hele korpus. Afvigelsen vil typisk svinge omkring nul. Værdien nul repræsenterer gennemsnittet for alle delkorpusser, hvori fundet optræder. Ligger værdien over nul for en årgang, er hyppigheden af matchet over gennemsnittet, ligger den under nul, er hyppigheden under gennemsnittet. Er der slet ingen forekomster i pågældende delkorpus, sættes værdien ligele-



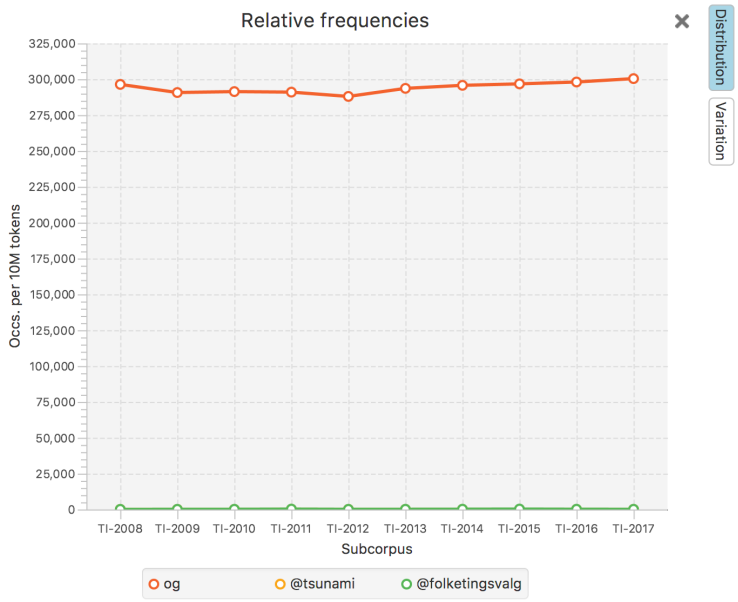

Figur 12.4: Frekvensdiagram for en række ord

des til nul. Værdien nul betyder altså enten slet ingen forekomster i pågældende korpus eller præcist gennemsnitlig mange. De interessante værdier er dem, der er over eller under nul.

Du kan få et indtryk af, hvad variationsdiagrammer kan vise, ved at betragte det hyppighedsdiagram fra TiDK 2019, som er vist i figur 12.5 på næste side.

Her er der i ét og samme diagram vist hyppighedsudviklingen over tid for ordene *og*, *tsunami* og *folketingsvalg*. I forhold til *og* ser *tsunami* og *folketingsvalg* i diagrammet ud til at være yderst sjældne i sproget: Mens *og* år for år forekommer knap 300.000 gange per 10 millioner ord, ligger udbredelsen af de to andre ord temmelig tæt på nul, så man hverken kan se udsving fra år til år eller skelne graferne for disse ord fra hinanden. Eksemplet viser, at det kan være problematisk at kombinere grafer for høj- og lavfrekvente ord i ét diagram. Hvis man undlod at vise udbredelsen af det højfrekvente *og* sammen med det lavfrekvente *tsunami* og *folketingsvalg* i ét og samme diagram, ville de to sidstnævnte vise betydeligt mere markante udsving.

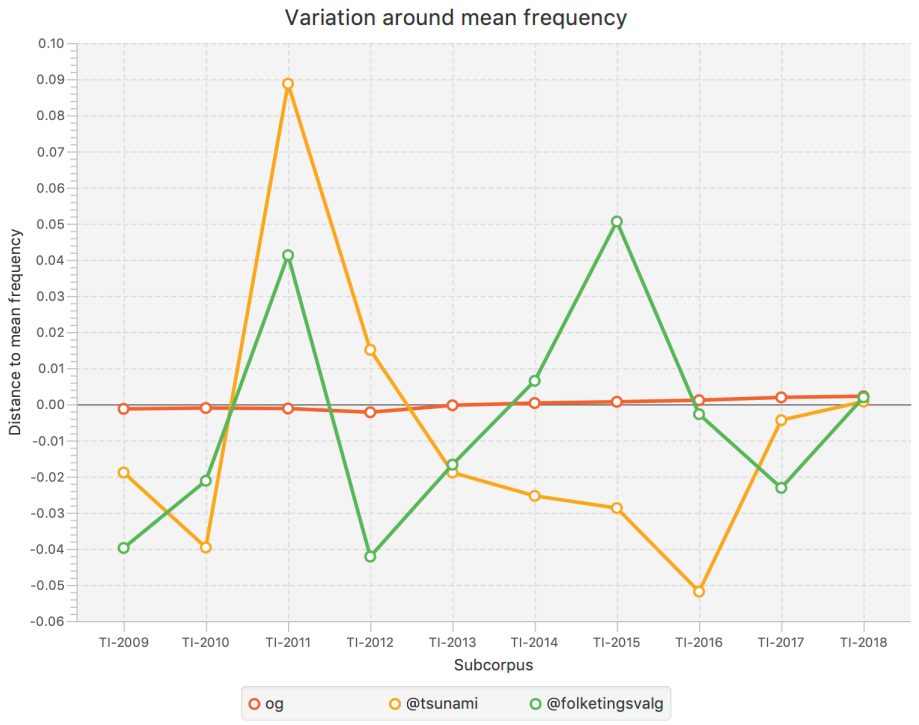
Hvis du nu betragter de samme tre ord, altså det udbredte *og* sammen med de mere sjældne *tsunami* og *folketingsvalg*, i et variationsdiagram som vist i figur 12.6 på side 65, vil du se, at *og* år for år har en fuldstændig gennemsnitlig hyppighed, hvilket ikke overrasker, da der normalt ikke sker udsving i hyppigheden af sprogets mest udbredte ord. Variationsdiagrammet viser samtidig, at udbredelsen af ordet *tsunami* ligger markant over gennemsnittet i 2011 (Fukushima-



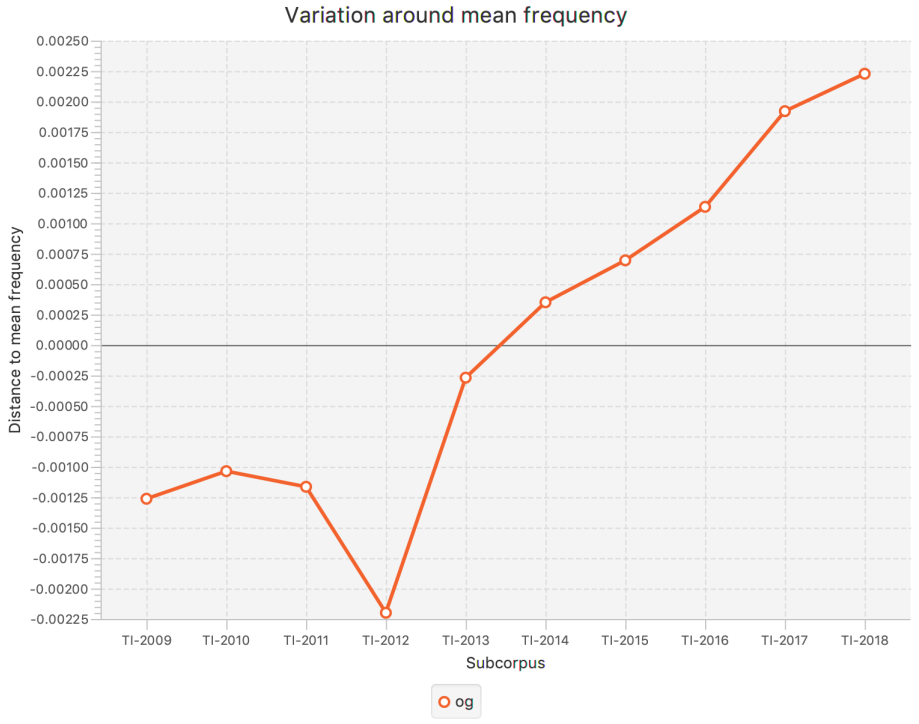

Figur 12.5: Diagram over sjældne og hyppige ords forekomster over tid

uheldet i Japan), mens *folketingsvalg* er markant overrepræsenteret i de år, hvor der afholdtes folketingsvalg, nemlig i 2011 og 2015.

Variationsdiagrammer giver især mening, hvis du vil sammenligne hyppighedsudsving for flere ord eller vendinger. Hvis du kun ser på et enkelt ord, skal du være særligt opmærksom på skaleringen af *y*-aksen. Figur 12.7 på næste side viser variationen af *og* alene. Ser man kun på grafens form, kan man forledes til at tro, at der er store udsving i hyppigheden af dette ord år for år. Aflæser man værdierne på *y*-aksen, viser det sig dog, at udsvingene er forstørret ganske gevaldigt op – og i virkeligheden ganske ubetydelige.



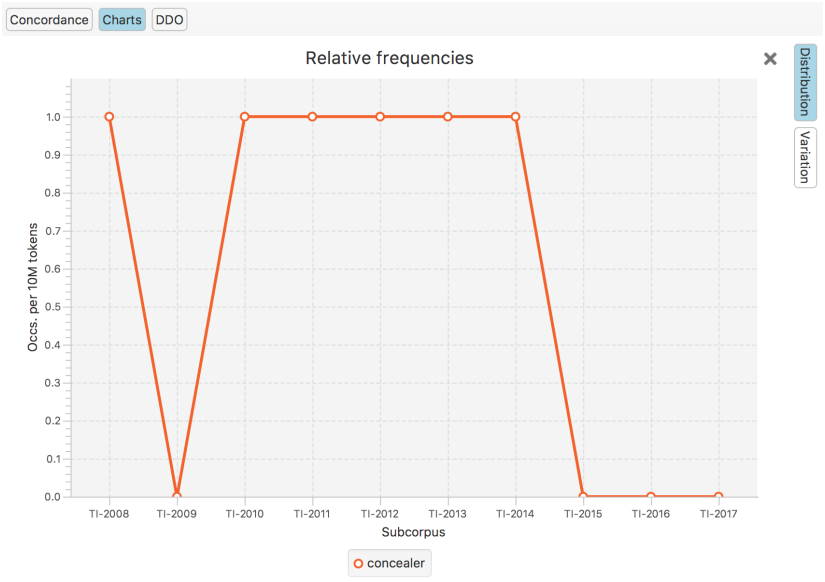

Figur 12.6: Variationsdiagram med frekvensafvigelser fra gennemsnittet *0*




Figur 12.7: Tilsyneladende store udsving i hyppigheden af *og*

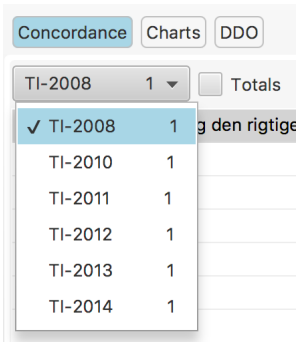
## 12.2 Sjældne ord

En søgning på `concealer` i -TiDK 2018 resulterer i en graf over ordets hyppighed for de enkelte årgange, som vist i figur 12.8.



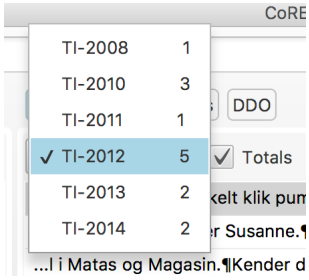

Figur 12.8: Hyppighedsfordeling af et sjældent ord

Som du ser, svinger grafen mellem 0,0 og 1,0 forekomster per 10 millioner ords tekst for de enkelte år. Du kan observere det samme, hvis du ser på den udfoldede delkorpusvælger over konkordansen, som vist i figur 12.9. For de år, som ikke er med i denne liste, er der 0 forekomster – delkorpusser med ingen forekomster af det søgte vises ikke i delkorpusvælgeren.




Figur 12.9: Relative hyppigheder af det sjældne *concealer* i delkorpusvælgeren

Sætter du hak ved »Totals« til højre for delkorpusvælgeren, viser den ikke længere de relative forekomsttal (per 10 millioner ords tekst), men de absolutte, altså det faktiske antal forekomster. I vores eksempel ser det nu sådan ud, som vist i figur 12.10 på den følgende side.




Figur 12.10: Absolutte hyppigheder af det sjældne *concealer* i delkorpusvælgeren

Som du ser, er der fem gange så mange forekomster af *concealer* i delkorpusset TI-2012 (med tekster fra 2012) som i TI-2008 og TI-2011. Betyder det nu, at TI-2012-korpusset er fem gange større end de to andre, eftersom deres relative forekomsttal er ens? Nej. Årsagen for dette misforhold er, at ordet er meget sjældent.

Tabel 12.1 viser størrelsen for de enkelte delkorpusser (i millioner ords tekst) sammen med de faktiske forekomsttal for *concealer*. På baggrund af disse tal er det relative forekomsttal beregnet i næstsidste kolonne.

| Delkorpus | Størrelse | Absolut | Relativ | Viser |
|---|---|---|---|---|
| TI-2008 | 28,9 | 1 | 0,3 | 1 |
| TI-2009 | 39,5 | 0 | 0,0 | 0 |
| TI-2010 | 42,6 | 3 | 0,7 | 1 |
| TI-2011 | 42,3 | 1 | 0,2 | 1 |
| TI-2012 | 43,3 | 5 | 0,1 | 1 |
| TI-2013 | 39,5 | 2 | 0,5 | 1 |
| TI-2014 | 39,5 | 2 | 0,5 | 1 |
| TI-2015 | 33,0 | 0 | 0,0 | 0 |
| TI-2016 | 37,5 | 0 | 0,0 | 0 |
| TI-2017 | 37,5 | 0 | 0,0 | 0 |

Tabel 12.1: Korpusstørrelse og forekomsttal for det sjældne ord *concealer*

Tabellen viser, at størrelsen for de enkelte delkorpusser ikke varierer så meget, som man ud fra observationerne ovenfor umiddelbart skulle tro, og den viser, at der er ret forskellige relative værdier på spil (næstsidste kolonne), og ikke kun 0 og 1, som CoREST ellers påstår er de eneste (sidste kolonne). Det er iøjnefaldende, at der er en faktor 7 til forskel på de relative forekomster i TI-2012 og TI-2010, og man kan med rette undre sig over, at dette ikke afspejles i grafen.

Forklaringen på denne unøjagtige visning er, at ordet *concealer* optræder meget sjældent. CoREST viser relative forekomster per 10 mio. ord kun som heltal.

Derfor sættes alle relative forekomsttal større end 0 og mindre end 1 til 1. Så et relativt forekomsttal på 1 i CoREST betyder, at ordet ganske vist forekommer i korpus, men med højst 1 forekomst per 10 millioner ords tekst. Ord, som er så sjældne, bør du være yderst varsom med at konkludere noget om, og de bør som udgangspunkt ignoreres i sprogvidenskabelige undersøgelser!

# Kapitel 13

# Word2dict

*Oversigter over beslægtede ord*

Kapitlet tilføjes i en senere version af manualen. Det vil beskrive en særlig funktionalitet i DSL-udgaven af CoREST.

# Del III

# Avancerede søgemuligheder

# Kapitel 14

# Søgning med regulære udtryk

*Jokertegn og andre symboler til udvidelse af søgeudtryk*



## 14.1 Regulære udtryk

I CoREST kan du udføre søgninger med jokertegn, altså tegn, der kan stå i stedet for forskellige andre tegn. Ved at bruge jokertegn laver du søgeudtryk, der har en vis variabilitet med hensyn til, hvad de matcher i korpus. Man kalder denne type variable søgeudtryk for *regulære udtryk*.[1] I kapitel 4 brugte vi allerede `.*` (punktum stjerne) i betydningen *nul, ét eller flere bogstaver må matche på dette sted*. Således matchede `.*telefon` alle de ord i korpus, som ender på *telefon* inklusive ordet *telefon* selv. Både `.*` og `.*telefon` er regulære udtryk. Efter en lidt mere formel definition af, hvad et regulært udtryk er, skal vi i dette kapitel se på, hvad disse udtryk præcist kan bruges til, når du søger i et korpus.

**Definition af *regulært udtryk*** Et regulært udtryk er en række af *tegn (standardtegn* eller *jokertegn*) eller *grupper* – eller en blanding af disse komponenter); tegn kan kombineres med *operatorer*.

**standardtegn** et bogstav, ciffer eller et sætningstegn, fx *a*, *b*, *Å*, *q*, *9* eller *;* (semikolon). Ikke alle sætningstegn kan umiddelbart bruges som standardtegn

[1] Se mere under `http://en.wikipedia.org/wiki/Regular_expression`.

**jokertegn** et særligt tegn (punktum), som kan stå for et vilkårligt standardtegn

**gruppe** samler (en del af) et regulært udtryk til en enhed, der funktionelt betragtes, som var den ét *tegn*

**operator** beskriver en særlig funktion, fx gentagelse, som skal gælde for et *tegn* eller en *gruppe*

## 14.2 Et regulært udtryks komponenter

Den følgende gennemgang af de enkelte komponenter i regulære udtryk kan måske forekomme en anelse vanskelig, men meningen burde senest blive klar, når vi når til eksemplerne i afsnit 14.3 på side 74.

### 14.2.1 Standardtegn

Standardtegn er bogstaver, cifre og visse interpunktionstegn. Standardtegn tages altid bogstaveligt i en søgning: De betyder det, deres pålydende er. Således finder søgeudtrykket `akryl` alle eksempler på *akryl* i korpus og intet andet, fordi tegnene *a*, *k*, *r*, *y* og *l* tages fuldstændig bogstaveligt under søgningen.

### 14.2.2 Jokertegnet *punktum*

Der findes ét jokertegn, nemlig *punktum*. Et punktum i søgeudtrykket betyder, at der på punktummets plads skal matches præcist ét vilkårligt standardtegn. Tabel 14.1 viser en række eksempler med tilhørende match i KORPUS-DK.

| Udtryk | Match i KORPUS-DK (forenklet ortografi) |
|---|---|
| `a.ryl` | *acryl*, *akryl*, *arryl* (fra propriet *Arryl House*) |
| `...sætte` | *frisætte*, *hensætte*, *indsætte*, *knæsætte*, *modsætte*, *målsætte*, *nedsætte*, *nulsætte*, *tilsætte*, *undsætte*, ... |
| `c....ing` | *coaching*, *crossing*, *catering*, *changing*, *cityring*, *charming*, *clearing*, *clubbing*, *crawling*, ... |

Tabel 14.1: Eksempler på brug af jokertegnet *punktum*

### 14.2.3 Gentagelsesoperatorer

Tabel 14.2 på den følgende side lister operatorer, der bruges til at bestemme, hvor tit det forudgående tegn eller den forudgående gruppe skal eller kan gentages. En gruppe oprettes ved at samle tegn og operatorer mellem runde parenteser, jf. oversigten med eksempler i afsnit 14.3 på side 74.

| Symbol | Betydning |
|---|---|
| `?` | ingen eller én forekomst af forudgående tegn eller gruppe |
| `*` | ingen, én eller flere forekomster af forudgående tegn eller gruppe |
| `+` | én eller flere forekomster af forudgående tegn eller gruppe |
| `{n}` | præcist *n* forekomster af forudgående tegn eller gruppe |
| `{n,m}` | mellem *n* og *m* forekomster af forudgående tegn eller gruppe |
| `{n,}` | mindst *n* forekomster af forudgående tegn eller gruppe |
| ~~`{,m}`~~ | Findes ikke. Brug `{1,m}` i stedet. |

Tabel 14.2: Gentagelsessymboler

### 14.2.4 Symboler med særlig betydning

Tabel 14.3 på den følgende side lister symboler med en særlig betydning i regulære udtryk.

| Symbol | Betydning |
|---|---|
| (*u*) | Runde parenteser samler (en del af) et regulært udtryk *u* til en **gruppe** pågældende sted i udtrykket. |
| [$t_1\,t_2\,t_3\ldots t_n$] | Kantede parenteser definerer en **mængde af standardtegn** $t_1\,t_2\,t_3\ldots t_n$, hvoraf ét skal matche pågældende sted i udtrykket; der må ikke være mellemrum mellem tegnene, ellers regnes det med i mængden. |
| [^$t_1\,t_2\,t_3\ldots t_n$] | Kantede parenteser med cirkumfleks-tegnet (^) umiddelbart efter åbningsparentesen definerer en **mængde af standardtegn** $t_1\,t_2\,t_3\ldots t_n$, hvoraf *intet* må matche pågældende sted i udtrykket; der må ikke være mellemrum mellem tegnene, ellers regnes det med i mængden. |
| $u_1$ \| $u_2$ | Lodret streg mellem to regulære udtryk (herunder tegn eller grupper) $u_1$ og $u_2$ betyder: $u_1$ **eller** $u_2$ skal matche pågældende sted i udtrykket. |
| $t_1$-$t_2$ | Bindestreg mellem to tegn $t_1$ og $t_2$ betyder: Alle tegn i det alfabetiske interval **fra** tegn $t_1$ **til** $t_2$. |
| \ | Den omvendte skråstreg betyder, at det efterfølgende tegn, som normalt har en særlig betydning i regulære udtryk (fx punktum, bindestreg, parentes), skal opfattes bogstaveligt i udtrykket, altså at det skal fortolkes som et **standardtegn**. |

Tabel 14.3: Symboler med særlig betydning

## 14.3 Eksempler på brugen af regulære udtryk

Prøv at udføre søgningerne fra tabel 14.4 på næste side i KORPUS-DK. Vær opmærksom på, at samme søgeresultat tit kan opnås med forskellige regulære udtryk.

| Regulært udtryk | Udtrykkets match | Eksempler |
|---|---|---|
| `cente?ret` | *e*'et foran ? er valgfrit | matcher *centeret* og *centret* |
| `hi(hi)*` | mindst ét *hi*, som kan gentages vilkårligt mange gange | matcher *hi*, *hihi*, *hihihi* osv. |
| `.+undersøgelse` | matcher ord, der ender på *undersøgelse* (men ikke *undersøgelse* alene) | *forundersøgelse*, *spørgeskemaundersøgelse*, *mentalundersøgelse*, ... |
| `.+undersøgelse(n\|r\|rne)?s?` | matcher samme som ovenfor og derudover også alle bøjningsformer af *undersøgelse* | *miljøundersøgelser*, *forundersøgelser*, *spørgeskemaundersøgelsen*, ... |
| `(hej){2}` | gruppen *hej* skal forekomme to gange | *hejhej* |
| `(ha){2,4}` | gruppen *ha* skal forekomme mellem to og fire gange | *haha*, *hahaha* og *hahahaha* |
| `å{2,}r?h?` | matcher ord, som begynder med mindst to *å*'er, evt. efterfulgt af ét *r* eller ét *h* eller både *r* og *h* | *ååh*, *åååh*, *ååååh*, *åååååh*, *ååååårh* og *åååååårh*, ... |
| `.*[aeiouyæøå]{4,}.*` | matcher ord, som indeholder en række af mindst fire vokaler | *niveauet/-r*, *bureauet/-r*, *layout*, *tableauet/-r*, *hawaiianer*, ... |
| `[aeiouyæøå]{3,}` | matcher ord på mindst tre bogstaver, der alle skal være vokaler | *you*, *iii*, *eye*, *eau*, ... |
| `ni[ck]ola[ij]` | *ni* efterfulgt af enten *c* eller *k* efterfulgt af *ola*, efterfulgt af enten *i* eller *j* | *Nicolai*, *Nikolai*, *Nicolaj*, *Nikolaj* |
| `mor(far\|mor)` | *mor* efterfulgt af enten *far* eller *mor* | *morfar*, *mormor* |
| `(far\|mor){2}` | *far* eller *mor* to gange efter hinanden | *farfar*, *farmor*, *morfar*, *mormor* |
| `19[0-9]{2}` | *19* efterfulgt af et af cifrene mellem *0* og *9* to gange efter hinanden | *1990*, *1992*, *1997*, *1998*, *1999*, *1960*, *1970*, ... |
| `.*[bcdf-hj-np-tv-xz]{7,}.*` | ord med mindst syv konsonanter efter hinanden | *bratschstjerne*, *danskschweizisk*, *angstskrig*, ... |
| `\*[^\*].*\*` | ord, som begynder med og slutter på en stjerne, men som ellers ikke indeholder stjerner | **suk**, **host**, **lol**, **fniz**, ... |

Tabel 14.4: Eksempler på regulære udtryk

# Kapitel 15

# Søgning på ordoplysninger

*Søge på grundformer, ordklasser og bøjningsformer*



## 15.1 Hvad er ordoplysninger?

Et CoREST-korpus er organiseret som en lodret liste af ord. Til hvert ord i denne liste knytter der sig flere oplysninger, som du ud over selve ordet også kan søge på. Vi kalder disse oplysninger for ordoplysninger eller ord-*attributter*.

Ud over de ortografisk forenklede standardsøgeord indeholder alle CoREST-korpusser oplysninger om ordenes originale stavning og de mellemrum og sætningstegn, der danner grænsen mod det følgende ord. Disse tre oplysninger betegnes `word`, `ortho` og `bound`:

`word` Ordet i en lettere forenklet ortografisk form

`ortho` Ordet i tilnærmelsesvis præcis ortografisk gengivelse

`bound` De teksttegn, der tilsammen udgør grænsen til det følgende ord

Herudover kan der findes ordoplysninger, som er specielle for et korpus. I CoREST findes der således i alle korpusser yderligere følgende tre specielle oplysninger, som kaldes hhv. `lemma`, `pos` og `epos`:

`lemma` Ordets grundform, dvs. den form, det typisk kan slås op under i en ordbog

`pos` Ordets ordklasse

`epos` En udvidet ordklasse- og bøjningsoplysning

Ved hjælp af særlige søgeudtryk er det muligt at inddrage alle disse ordoplysninger i søgningen.

### 15.1.1 Et eksempel

Tabel 15.1 viser den CoREST-interne repræsentation af et »korpus«, som i dette eksempel kun består af sætningen *Jakob byder endnu engang en masse lækre singler til fest.*

| word | ortho | bound | lemma | pos | epos |
|---|---|---|---|---|---|
| jakob | Jakob | _ | Jakob | N | NP:siu#:--:---- |
| byder | byder | _ | byde | V | VF:----:sa:---- |
| endnu | endnu | _ | endnu | D | D-:----:--:u--- |
| engang | engang | _ | engang | D | D-:----:--:u--- |
| en | en | _ | en | P | PI:s-uc:--:---- |
| masse | masse | _ | masse | N | NC:siuc:--:---- |
| lækre | lækre | _ | lækker | A | AC:p#u#:--:p--- |
| singler | singler | _ | single | N | NC:piu#:--:---- |
| til | til | _ | til | T | T-:----:--:---- |
| fest | fest | .$ | fest | N | NC:siuc:--:---- |

Tabel 15.1: Ordattributter i CoREST-korpusser

I tabellen ses de nævnte ordattributter og deres værdier for dette eksempel. I det følgende vises, hvordan du kan inddrage de forskellige ordoplysninger i dine søgninger ved hjælp af avancerede søgeudtryk.

## 15.2 Brug af ordoplysninger i søgeudtryk

Med simple søgeudtryk, se kapitel 4, kan du kun søge på ordoplysningerne `word`, `ortho` og `lemma`. Vil du lade de andre oplysninger indgå i dine søgeudtryk, skal

du bruge **avancerede søgeudtryk**. Et avanceret søgeudtryk bygger på følgende grundstruktur:

```
ordoplysning = "X".
```

I et konkret søgeudtryk vil der i stedet for `ordoplysning` stå et af ordattributterne `word`, `ortho`, `bound`, `lemma`, `pos` eller `epos`; `ordoplysning` skal matche det, der står mellem anførselstegnene, altså det, der vil stå i stedet for `X` i grundstrukturen ovenfor. Det udtryk, der står på `X`'s plads kan indeholde regulære udtryk, jf. kapitel 14.

Du kan kombinere forskellige grundstrukturer med hinanden ved hjælp af `&` (»og«) og `|` (»eller«), se eksemplerne nedenfor. CoREST kan selv finde ud af, om du har indtastet et simpelt eller avanceret søgeudtryk og vil tilpasse søgningen herefter.[1]

En variant af grundstrukturen er

```
ordoplysning != "X".
```

Den eneste forskel i forhold til den grundstruktur, vi lige har set, er, at der er sat et udråbstegn foran lighedstegnet. Det betyder, at ordoplysningen *ikke* må svare til værdien mellem anførselstegnene.

**Pas på!** Brug denne type kun i søgeudtryk, hvor du kombinerer flere grundstrukturer med hinanden ved at bruge `&` (»og«) og `|` (»eller«). Ellers kan du nemt komme galt afsted med denne type, fx betyder `[word!="cafe"]` »alle ord i korpus, som ikke er en stavevariant af *cafe*«, og det er næsten samtlige ord i korpus undtagen netop stavevarianter af *cafe* – et meningsløst resultat altså, som det tilmed tager en del tid at finde frem til!

**OBS! Kantede parenteser:** Et avanceret søgeudtryk, hvad enten det kun består af en enkelt grundstruktur eller flere kombineret med hinanden, skal være omgivet af kantede parenteser. Dette gælder søgeudtryk, der vedrører enkeltord. For søgeudtryk, som søger på grupper af ord, er situationen en lidt anden, se kapitel 17.

I det følgende gives en række eksempler på søgeudtryk for de forskellige ordoplysninger.

### 15.2.1 `word`

`[word="cafe"]` matcher de ortografiske former *cafe*, *café*, *Café* osv. Søgeudtrykket svarer til at skrive `cafe` i den simple notation. I stedet for `[word="cafe"]` kan du også nøjes med at skrive `"cafe"` i avancerede søgeudtryk: ordoplysningen `word` er den oplysning, CoREST altid vil søge på, når du kun anvender anførselstegn, og når ingen anden ordoplysning er nævnt i søgeudtrykket.

[1] Faktisk er det sådan, at CoREST internt altid oversætter simple søgeudtryk til avancerede. I sessionsloggen, se kapitel 5, registreres derfor også kun de oversatte avancerede søgeudtryk.

**Pas på!** `[word="Café"]` matcher ingenting i korpus, idet `word` altid repræsenterer korpusordene i en forenklet ortografi, hvor store bogstaver er omsat til små og accenttegn er fjernet.

### 15.2.2 `ortho`

`[ortho="cafe"]` matcher kun tilfælde, hvor den originale stavning i korpus præcist svarer til den form, der står mellem anførselstegnene. Således matcher hverken *café* eller *Café*. Dette søgeudtryk svarer til `!cafe` i den simple notation. Ordoplysningen `ortho` gør det muligt at finde staveformer med accenttegn, således finder `[ortho="café"]` kun eksempler på *café*. Dette svarer til at skrive `!café` i den simple notation. `[word="cafe" & ortho!="Caf[eé]"]` finder alle stavevarianter af *cafe*, som ikke er stavet med stort *C* som første bogstav.

### 15.2.3 `bound`

`[bound=";_"]` matcher alle ord i korpus, som har et semikolon til højre for efterfulgt af et mellemrum. Mellemrum angives som understreg i søgeudtrykket. Du kan kombinere de forskellige ordoplysninger med hinanden til et mere komplekst søgeudtryk. Således matcher `[word="cafe" & bound="-"]` alle ortografiske former af *cafe*, som følges af en bindestreg.

### 15.2.4 `lemma`

`[lemma="mand"]` matcher samtlige bøjningsformer af grundformen *mand* og en række af deres ortografiske varianter. Dette søgeudtryk svarer til `@mand` i den simple notation. Eksemplet matcher således bl.a. *mand*, *mands*, *mænd*, *manden* og *Manden*. `[lemma="tale" | lemma="lytte"]` finder alle bøjningsformer af grundformerne *tale* og *lytte*. Søgningen kan også udtrykkes således: `[lemma="tale|lytte"]`.

### 15.2.5 `pos`

`[pos="I"]` matcher udråbsord (interjektioner) i korpus, fx *nej*, *Ja*, *Nej*, *Fy*, *Basta*, *Føj* etc. `[lemma="tale" & pos="N"]` matcher kun de former af lemmaet *tale*, som er markeret som navneord (substantiver) i korpus. Det vil sige, at former af *tale*, som er markeret som verber, ikke matcher. Eksempler: *(på) tale*, *(daglig) tale*, *(dansk) tale*, *(nogle) taler (er poetiske)*.

`[(lemma="tale" | lemma="lytte") & pos="V"]` finder alle former af verberne *tale* og *lytte*, men ikke former af substantivet *tale*. Bemærk parenteserne til gruppering af de enkelte logiske dele i det sammensatte udtryk. Udtrykket kan forenkles til `[lemma="tale|lytte" & pos="V"]`. En beskrivelse af, hvilke værdier oplysningstypen `pos` kan antage, findes i kapitel 16.

### 15.2.6 epos

`[epos="..:..g.:..:...."]` matcher ord i korpus, som har kasusmærket *genitiv*, uanset ordklasse eller øvrig bøjning. En nærmere beskrivelse af opbygningen af `epos` findes i kapitel 16.

## 15.3 Avancerede over for simple søgeudtryk

Alle hidtil viste søgninger i manualen forud for dette kapitel beror på simple søgeudtryk – selvom regulære udtryk, som blev behandlet i kapitel 14, godt kan virke komplicerede. Simple søgeudtryk er tilstrækkelige i mange situationer, men der er tilfælde, hvor det er nødvendigt at udføre søgninger, hvori der indgår flere oplysninger. Her kommer avancerede søgeudtryk ind i billedet.

Tabel 15.2 sammenfatter de avancerede søgeudtryk, som gør brug af ordoplysninger, og den giver de tilsvarende simple udtryk i de tilfælde, hvor det er muligt at udføre en søgning med et simpelt udtryk også.

| **Avanceret udtryk** | **Simpelt udtryk** | **Match-eksempler** |
|---|---|---|
| `[word="cafe"]` eller `"cafe"` | `cafe` | *cafe, café, Café, ...* |
| `[ortho="cafe"]` | `!cafe` | *cafe* |
| `[ortho="café"]` | `!café` | *café* |
| `[word="cafe" & ortho!="Caf[eé]"]` | ej muligt | stavevarianter af *cafe*, som ikke er stavet med stort |
| `[bound=";_"]` | ej muligt | ord efterfulgt af semikolon og mellemrum |
| `[word="cafe" & bound="-"]` | ej muligt | *cafe-, café-, Café-, ...* |
| `[lemma="mand"]` | `@mand` | *mand, mands, mænd, manden, Manden, ...* |
| `[lemma="tale" \| lemma="lytte"]` | `@tale\|lytte` | former af *tale* og *lytte* |
| `[pos="I"]` | ej muligt | udråbsord og lydord: *nej, Ja, Nej, Fy, Basta, Føj, ...* |
| `[lemma="tale" & pos="N"]` | ej muligt | former af substantivet *tale*: *(på) tale, (daglig) tale, (dansk) tale, (nogle) taler (er poetiske), ...* |
| `[lemma="tale\|lytte" & pos="V"]` | ej muligt | former af verberne *tale* og *lytte* |
| `[epos="..:..g.:..:...."]` | ej muligt | ord i genitiv |

Tabel 15.2: Avancerede over for simple søgeudtryk

### 15.3.1 Søgninger med kontroltegnet %

I simple søgeudtryk kan du bruge kontroltegnene $, #, !, @ og %. Når du søger på baggrund af en søgeliste, se kapitel 3, vil CoREST søge på et simpelt søgeudtryk, som består af det markerede ord på søgelisten med foranstillet %. Et søgeudtryk som %zika anvendt på et af de seneste TiDK-korpusser får CoREST til at søge på alle former af lemmaet *zika* og desuden på samtlige ortografiske ord, som begynder med *zika* inklusive ordet *zika* selv.I søgeresultatet er der således ikke kun eksempler på *zika*, men også eksempler på ord, der begynder med *zika*. Du kan danne dig et overblik i matchlisten[2].

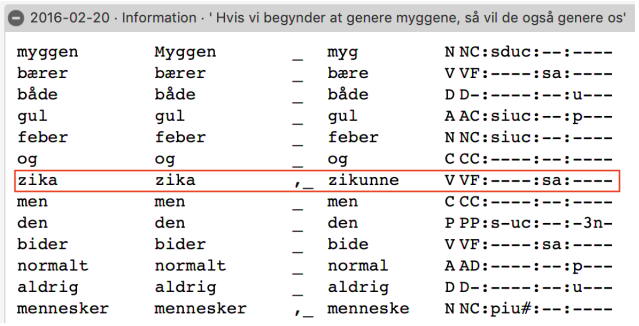
Hvis du udfører samme søgning med kontroltegnet @ i stedet for %, er der færre forekomster af (alle bøjningsformer af) af (lemmaet) *zika*. Åbenbart findes der slet ingen bøjningsformer af lemmaet *zika*, end ikke genitiven *zikas*.

Årsagen til, at en søgning på lemmaet *zika* giver så få resultater, skal findes i, at ordklassetaggeren ikke har kunnet bestemme grundformen for *zika* – og i en hel del tilfælde end ikke for noget som helst ord i sætninger, hvor *zika* indgår. Det kan du se i figur 15.1, som viser oversigten med ordoplysningerne for en af konkordanslinjerne.


Figur 15.1: Ordet *zika* kan ikke ordklassebestemmes af ePOS-taggeren

Af figuren fremgår det, at *zika* er blevet tagget med ordklasseoplysningen *XY* – altså som et ukendt ord. Derfor er lemmaformen sat til *zika*, og der er kun én bøjningsform, nemlig lemmaformen *zika* selv. Dette forklarer imidlertid ikke, at der for søgeudtrykket @zika kun er så få forekomster, mens selve den ortografiske form *zika* har ganske mange forekomster i korpus. Det er, som om der er (mindst) et andet lemma på spil end *zika, XY*. Under de udvidede ordoplysninger, jf. kapitel 10, kan du ud over ordklassetagget også se lemmaformerne for ordene. Hvis du nu gentager %zika-søgningen, vil du konstatere, at en del af *zika*-forekomsterne er blevet lemmatiseret som det ikke-eksisterende verbum *zikunne*, se figur 15.2 på næste side.

[2] Du kan læse mere om matchlister i kapitel 6.




Figur 15.2: Ordet *zika* fejlagtigt tagget som et ikke-eksisterende verbum *zikunne*

Problemerne med at lemmatisere *zika* korrekt skyldes, at ordet *zika* mangler i taggerens leksikon, da det er et forholdsvis nyt ord. I sådanne tilfælde begynder taggeren at gætte. En af gættestrategierne er at fjerne det første bogstav i et ord så længe, til der opstår et genkendeligt ord. Det, -taggeren endte med at læse ind i *zika*, var en dannelse af *zi* og *ka*, hvor *ka* blev genkendt som en variant af *ka*' altså *kan*, og lemmaformen for præsens aktiv-formen *kan* er *kunne*, derfor antog taggeren et lemma sammensat af *zi* og *kunne* altså *zikunne*. Logisk nok måske, men alligevel forkert. En søgning på `@zikunne` giver derfor mange resultater af formen *zika*.

Da ordlister i TiDK-korpusserne ofte indeholder nye ord, som kan være tagget forkert, udfører CoREST søgninger med udgangspunkt i ordlister med kontroltegnet `%` (i stedet for det mere hensigtsmæssige `@`). Herved omgås nogle af de problemer, en fejlagtig lemmatisering medfører, nemlig at der ikke vises samtlige forekomster af et sådant ord i søgeresultatet. Alle simple søgeudtryk bliver imidlertid internt omsat til avancerede søgeudtryk, som er det eneste, CoREST i virkeligheden behersker. Du kan se, hvilke avancerede søgeudtryk de omsættes til, ved at kigge i sessionsprotokollen, jf. kapitel 5:

- ✧ `@zika` omsættes til `[lemma="zika"]`
- ✧ `%zika` omsættes til `[lemma="zika" | ortho="zika.*"]`

I dine egne søgninger bør du ikke bruge kontroltegnet `%`, da `%`-søgninger ofte skyder langt over målet og medfører støj i form af for mange forekomster, som intet har med det søgte at gøre. Kontroltegnet `%` bør være forbeholdt søgninger, hvor lemmatiseringen kan antages at være fejlbehæftet.

# Kapitel 16

# Søgning på ordklasse og bøjning

*Opbygningen af ordklasse- og bøjningsoplysninger*



Samtlige ord i CoREST-standardkorpusserne er beriget med oplysninger om deres ordklasse og bøjning, de er *tagget*. Oplysningerne er tilføjet automatisk ved hjælp af ordklassetaggeren *ePOS* udviklet af Jørg Asmussen, DSL. ePOS-taggeren tager sig også af lemmatiseringen af ord, det vil sige, at den bestemmer grundformen for hvert ord i korpus.

ePOS-ordklassetaggeren anvender et bestemt inventar af markeringer for ordklasser og bøjninger, et såkaldt *tagsæt*. Inventaret beror på en grammatisk fortolkning af, hvordan de forskellige ordklasser og bøjninger kan defineres hensigtsmæssigt. ePOS-taggerens grammatiske syn og dermed dets *tagsæt* bygger videre på konventionerne fra det danske PAROLE-korpus, som er resultatet af et EU-projekt, hvor der blandt andet skulle etableres fælles tagsæt for en række europæiske sprog. I ePOS-taggeren anvendes en videreudvikling af PAROLE's *tagsæt* med en række forbedringer og forenklinger. En redegørelse for ePOS-taggeren og dens tagsæt findes i Asmussen (2013) og Asmussen (2015).

I CoREST kan du søge på ordklasse- og bøjningsoplysningerne, sådan som de kommer til udtryk gennem ordklassetaggene i korpus. Dette forudsætter naturligvis, at det korpus, som du arbejder med, i det hele taget er tagget med ePOS-taggeren. Dette er tilfældet for alle korpusser i CoREST. Søgninger på ordklasse-

og bøjningsoplysninger udfører du ved at anvende ordoplysningerne `pos` (**p**art **o**f **s**peech = ordklasse) eller `epos` (*extended* part of speech) i dine søgeudtryk. Du kan læse mere om søgning med ordoplysninger i kapitel 15.

Attributtet `pos` indeholder altid en simpel ordklasseoplysning, mens attributtet `epos` giver mere udførlige ordklasse- og bøjningsoplysninger.

## 16.1 Ordoplysningen `pos`

Tabel 16.1 viser de forskellige værdier, som `pos`-tagget kan have (tagsættet, 1. kolonne) samt hvilke ordklasser de svarer til (2. kolonne).

| Tag | Ordklasse |
|---|---|
| `V` | verbum (udsagnsord) |
| `A` | adjektiv (tillægsord) |
| `L` | numerale (talord) |
| `N` | substantiv (navneord) |
| `P` | pronomen (stedord) |
| `D` | adverbium (biord) |
| `I` | interjektion (udråbsord) eller lydord |
| `T` | præposition (forholdsord) |
| `C` | konjunktion (bindeord) |
| `U` | *som*, *der* eller infinitivmarkøren *at* |
| `E` | leksikalsk element |
| `M` | løsrevet bøjningsendelse |
| `X` | symboler, fremmede ord eller fejl |

Tabel 16.1: Ordklassetags

Du kan søge på disse ordklasseoplysninger ved at bruge ordoplysningen `pos` i dine søgeudtryk. Således finder søgeudtrykket `[lemma="tale" & pos="V"]` eksempler på anvendelsen af verbet *tale* i korpus, mens `[pos="I"]` finder eksempler på interjektioner (udråbsord) herunder lydord. Fx *abracadabra*, *basta*, *ciao*, *ding*, *ehm*, *fuck*, *halleluja*, *ja*, *muh*, *nej* osv.

De fleste af ordklasserne på listen er almindeligt kendte og beskrives derfor ikke nærmere i denne manual. Der er dog et antal særtilfælde. Således mangler ordklassen *artikel* (kendeord) i ePOS, fordi artikler betragtes som pronomener. Endvidere er der et antal lidt specielle klasser, nemlig `U`, `E`, `M` og `X`, som beskrives nærmere i det følgende.

### 16.1.1 Tagget U – *som*, *der* og *at*

Da ePOS-tagsættet beror på PAROLE, har det i mange tilfælde været nødvendigt at overtage PAROLE's valg, selvom de umiddelbart kan forekomme ejendommelige. Det gælder for eksempel brugen af tagget U (»unique«) til markering af bestemte typer ord.

***som*** Ordformen *som* kan anvendes som relativpronomen, konjunktion (eller præposition); ePOS skelner ikke mellem disse anvendelser og tagger alle forekomster af *som* med et U.

***der*** Det samme tag U bruges til markering af forekomster af *der*, som kan optræde som formelt subjekt eller som relativpronomen.

***at*** Endelig bruges U til at tagge ordet *at*, når det anvendes som infinitivmarkør i forbindelse med et infinit verbum.

### 16.1.2 Tagget E – leksikalsk element

Leksikalske elementer er dele af ord, som på grund af den anvendte ordopdeling i CoREST, se kapitel 17, får status af selvstændige ord. Da for eksempel bindestreger fungerer som ordgrænser, vil ord med bindestreg blive splittet op. For eksempel består *art deco-stil* af tre, *anti-amerikansk* af to, *Saudi-Arabien* af to, *be- eller afkræfte* af tre og *co-producer* af to ord. På denne måde opstår der særlige ord, som normalt ikke optræder selvstændigt, men kun som del af et andet ord: *art*, *deco*, *anti*, *Saudi*, *be* og *co*. Disse særlige ord betragter ePOS som leksikalske elementer, og de tagges med E. Du kan danne dig et indtryk af hyppige leksikalske elementer i KORPUS-DK ved at søge på `[pos="E"]` og sortere den resulterende matchliste efter faldende hyppighed.

### 16.1.3 Tagget M – løsrevet bøjningsendelse

Løsrevne bøjningsendelser minder lidt om leksikalske elementer, se afsnit 16.1.2, og skyldes ligeledes ordopdelingsprincippet, som du kan læse mere om i kapitel 17. Mens leksikalske elementer typisk bliver løsrevet fra deres basis på grund af en bindestreg, som betragtes som ordgrænse, så løsrives bøjningsendelser på grund af apostroffer, der også fungerer som ordgrænser.

Eksempler på løsrevne bøjningsendelser er *s* fra *FN's*, *erne* fra *1950'erne*, *er* fra *cd'er* og *en* fra *dj'en*.

Løsrevne bøjningsendelser tagges med et M (»morpheme«). Den konstruerede 'lemmaform' til disse endelser er altid selve endelsen med et foranstillet @, altså fx *@s*, *@erne*, *@er* eller *@en*. Du kan danne dig et indtryk af, hvilke løsrevne bøjningsendelser der findes i KORPUS-DK ved at udføre en søgning på udtrykket `[pos="M"]` og sortere den resulterende matchliste efter faldende hyppighed.

### 16.1.4 Tagget X – symboler m.m.

Symboler, fremmede ord og fejl tagges med X. Eksempler på symboler er tegn som % og & eller forkortelser som *BBC*, *com*, *dk*, *EUD*, *A* og *S* i *A/S* (skråstregen er ordskilletegn). Eksempler på fremmede ord er *five*, *greatest* eller *high*.

Desuden er der tilfælde, hvor ePOS ikke har været i stand til at tagge ord eller endda hele sætninger. Disse tilfælde er også tagget med X.

## 16.2 Ordoplysningen epos

Ordoplysningen epos (= »extended pos«) er en udvidet udgave af pos. Den indeholder ud over selve ordklasseoplysningen en række yderligere oplysninger. Tagget består af fire oplysningskategorier adskilt ved koloner:

**klasse : nominal : verbal : diverse**

**klasse** Denne oplysningskategori består typisk af to store bogstaver, sommetider af ét stort bogstav efterfulgt af en streg. Det første bogstav betegner ordklassen, sådan som det er beskrevet i tabel 16.1 på side 84, mens det andet betegner en underklasse til ordklassen. Ikke alle ordklasser har underklasser: Hvis der ikke er underklasser, markeres det med en streg. De forskellige underklasser og dertil hørende bøjningsparadigmer er beskrevet i afsnit 16.2.4 på næste side.

**nominal** Denne oplysningskategori består af fire tegn, som beskriver ordets nominale (substantiviske) bøjning, se afsnit 16.2.1.

**verbal** Denne oplysningskategori omfatter to tegn, der beskriver ordets verbale bøjning, se afsnit 16.2.2 på næste side.

**diverse** Denne oplysningskategori bruges til forskellige bøjningsoplysninger for adjektiver, pronomener og visse adverbier, se afsnit 16.2.3 på den følgende side.

Der er tilfælde, hvor en bestemt bøjningsoplysning ikke er relevant. I så fald sættes en streg pågældende sted. Endvidere er der tilfælde, hvor en bestemt bøjning ikke kan afgøres entydig; dette markeres ved hjælp af tegnet #. Der er også tilfælde, hvor bøjningen i princippet godt kan afgøres, men hvor det ikke er gjort. Dette er en reminiscens fra ordklasseopmærkningen i det oprindelige PAROLE-korpus. I disse tilfælde sættes tegnet § pågældende sted i tagget.

### 16.2.1 Nominale oplysninger

Nominale bøjningsoplysninger er kendetegnende for substantiver eller ord, som også kan bøjes substantivisk, fx adjektiver eller verber i visse former. Der er altid fire nominale bøjningsmarkører til stede i ePOS-tagget. Her gennemgår vi dem i den rækkefølge, de står i i tagget. Rækkefølgen ligger fuldstændig fast.

1. **Tal**: ental `s` eller flertal `p`
2. **Bestemthed**: ubestemt `i` eller bestemt `d`
3. **Kasus**: umarkeret `u`, ejefald `g`, fossileret `f` og – kun for de personlige pronomeners vedkommende – nominativ `n`. Akkusativ er identisk med umarkeret i disse tilfælde og tagges med `u`
4. **Køn**: fælleskøn `c` eller intetkøn `n`

I tilfælde, hvor en bestemt bøjningsoplysning ikke er relevant, sættes en streg pågældende sted. I tilfælde, hvor en bestemt bøjning ikke kan afgøres entydigt, markeres det ved hjælp af tegnet '#' eller '§'.

Se eksempler på taggede ord i tabel 16.2 på side 91.

### 16.2.2 Verbale oplysninger

Der er to verbale bøjningsmarkører til stede i ePOS-tagget. I den rækkefølge, de optræder i, er det:

1. **Tid**: nutid `s` eller datid `t`
2. **Diatese**: aktiv `a` eller passiv `p`

Se eksempler på taggede ord i tabel 16.2 på side 91.

### 16.2.3 Diverse oplysninger

Der er fire ekstra bøjningsmarkører til stede i ePOS-tagget. Her gennemgår vi dem i den rækkefølge, de står i i tagget:

1. **Grad** (adjektiver og visse adverbier): positiv `p`, komparativ `c`, superlativ `s` eller absolut superlativ `a`
2. **Person** (personlige og possessive pronomener): første `1`, anden `2` eller tredje `3`
3. **Refleksivitet** (personlige og possessive pronomener): ja `y` eller nej `n`
4. **Possessor** (possessive pronomener): ental `s` eller flertal `p`

Se eksempler på taggede ord i tabel 16.2 på side 91.

### 16.2.4 Underklasser og bøjningsmønstre

I det følgende gives en oversigt over de forskellige underklasser, som ordklasserne kan opdeles i. Til hver underklasse knytter der sig et særligt bøjningsmønster. Et bøjningsmønster er karakteriseret ved, at kun bestemte bøjningsoplysninger (= positioner i ePOS-tagget) er i brug, mens de øvrige ikke bruges og derfor er

markeret med en streg - i tagget. I oversigterne nedenfor er de positioner i tagget, som er i brug, markeret med en bolle •.

Grammatikken bag opdelingen i underklasserne behandles ikke i denne manual. Der henvises til generelle grammatiske beskrivelser af ordenes bøjning på dansk.

**Verber**

| Underklasse | Mønster |
|---|---|
| I infinitiv | VI:----:-•:---- |
| F finit | VF:----:••:---- |
| M imperativ | VM:----:--:---- |
| G gerundium | VG:••••:--:---- |
| P participium | VP:••••:•-:---- |
| T perf. part. | VT:siu#:•-:---- |
| D adv. part. | VD:----:•-:---- |

**Adjektiver**

| Underklasse | Mønster |
|---|---|
| C almindelige | AC:••••:--:•--- |
| D adverbielle | AD:----:--:•--- |

**Numeraler**

| Underklasse | Mønster |
|---|---|
| C kardinale | LC:--•-:--:---- |
| O ordinale | LO:--••:--:---- |

**Substantiver**

| Underklasse | Mønster |
|---|---|
| C appellativ | NC:••••:--:---- |
| P proprium | NP:••••:--:---- |

**Pronomener**

| Underklasse | Mønster |
|---|---|
| `C` reciprokke | `PC:•-•-:--:----` |
| `M` demonstrative | `PM:•-••:--:----` |
| `I` indefinitte | `PI:•-••:--:----` |
| `O` possessive | `PO:•--•:--:-•••` |
| `P` personlige | `PP:•-••:--:-••-` |
| `R` relative | `PR:•-••:--:----` |

**Adverbier**

| Underklasse | Mønster |
|---|---|
| `-` ingen | `D-:----:--:•---` |

**Interjektioner**

| Underklasse | Mønster |
|---|---|
| `-` ingen | `I-:----:--:----` |

**Præpositioner**

| Underklasse | Mønster |
|---|---|
| `-` ingen | `T-:----:--:----` |

**Konjunktioner**

| Underklasse | Mønster |
|---|---|
| `C` sideordnende | `CC:----:--:----` |
| `S` underordnende | `CS:----:--:----` |

**Infinitivmarkør og *som*/*der***

| Underklasse | Mønster |
|---|---|
| `I` infinitivmarkør | `UI:----:--:----` |
| `S` *som*/*der* | `US:----:--:----` |

**Leksikalske elementer**

| Underklasse | Mønster |
|---|---|
| `W` orddannelse | `EW:----:--:----` |

**Løsrevne bøjningsendelser**

| Underklasse | Mønster |
|---|---|
| `N` fra substantiv | `MN:••••:--:----` |
| `V` fra verbum | `MV:----:••:----` |
| `A` fra adjektiv | `MA:••••:--:•---` |

**Symboler m.m.**

| Underklasse | Mønster |
|---|---|
| `S` symbol | `XS:----:--:----` |
| `F` fremmed | `XF:----:--:----` |
| `Y` taggefejl | `XY:----:--:----` |

### 16.2.5 Eksempler på taggede ord

For at få illustrere brugen af ePOS-tags betragter vi følgende sætning fra KORPUS-DK:

- ✧ *De fleste af ulandenes egne beslutningstagere er økonomer, uddannet i neo-liberal økonomi på amerikanske eller i hvert fald vestlige universiteter.*

I tabel 16.2 på den følgende side er sætningen opstillet med ét ord per række med tilhørende grundform (lemma) og ePOS-tag.

| # | word | lemma | epos |
|---|---|---|---|
| 1 | de | den | PM:p-u#:--:---- |
| 2 | fleste | mange | AC:p#u#:--:s--- |
| 3 | af | af | T-:----:--:---- |
| 4 | ulandenes | uland | NC:pdg#:--:---- |
| 5 | egne | egen | AC:p#u#:--:p--- |
| 6 | beslutningstagere | beslutningstager | NC:piu#:--:---- |
| 7 | er | være | VF:----:sa:---- |
| 8 | økonomer | økonom | NC:piu#:--:---- |
| 9 | uddannet | uddanne | VT:siu#:t-:---- |
| 10 | i | i | T-:----:--:---- |
| 11 | neoliberal | neoliberal | AC:siuc:--:p--- |
| 12 | økonomi | økonomi | NC:siuc:--:---- |
| 13 | på | på | T-:----:--:---- |
| 14 | amerikanske | amerikansk | AC:p#u#:--:p--- |
| 15 | eller | eller | CC:----:--:---- |
| 16 | i | i | T-:----:--:---- |
| 17 | hvert | hver | PI:s-un:--:---- |
| 18 | fald | fald | NC:siun:--:---- |
| 19 | vestlige | vestlig | AC:p#u#:--:p--- |
| 20 | universiteter | universitet | NC:piu#:--:---- |

Tabel 16.2: Sætning med ePOS-tags

## Kommentarer til tabel 16.2

***de* (#1)** er tagget som et *demonstrativt pronomen* PM. Ofte bliver et sådant *de* også betragtet som en bestemt artikel, men ordklassen artikel findes ikke i ePOS, i stedet for bruges pronomen. Til pronomenet er i dette tilfælde knyttet en række nominale bøjningsoplysninger p-u#. Numerusoplysningen på førstepladsen angiver *flertal* p (pluralis). Demonstrative pronomener bøjes ikke i bestemthed, så derfor er der på bestemthedspladsen (plads 2) en streg. Kasus (plads 3) er *umarkeret* u. Da pronomenet optræder i flertal (modsat ental, hvor man kan skelne mellem *den* og *det*), kan kønnet ikke afgøres, så derfor ser vi tagget # på fjerdepladsen, hvor kønnet oplyses.

***fleste* (#2)** er tagget som *almindeligt adjektiv* AC, dvs. et adjektiv, der er brugt at-

tributivt eller prædikativt, men ikke adverbielt. Der gives følgende nominale bøjningsoplysninger i tagget `p#u#`. Tal er *flertal* `p` (pluralis), bestemthed oplyst på andenpladsen er ikke entydig `#`, kasusoplysningen på tredjepladsen er *umarkeret* `u`, og kønnet på fjerdepladsen er ikke entydigt `#`. Adjektiver bruger også gradoplysningen på førstepladsen i diverse-kategorien af tagget: `s---`. Graden er i dette tilfælde oplyst som *superlativ* `s`.

***af, i, på* (#3, #10, #13, #16)** er tagget som *præpositioner* `T-`. Præpositioner inddeles ikke i underklasser, derfor har vi en streg som underklasseoplysning. Præpositioner bøjes heller ikke, derfor er alle bøjningsoplysningerne sat til streger.

***ulandenes* (#4)** er tagget som *appellativisk substantiv* `NC`. Substantiver bøjes i tal (her *flertal* `p`), bestemthed (*bestemt* `d`) og kasus (*ejefald* `g`) oplyst på plads 1, 2 og 3 i den nominale del af tagget. Desuden oplyses deres køn på plads 4 i den nominale del. Kønnet fremgår kun direkte af substantivet, hvis det optræder i bestemt form ental, for eksempel *ulandets*, og indirekte af sammenhængen, hvis det optræder i ubestemt form ental, for eksempel *et uland*. I dette eksempel optræder det dog i flertal, hvorfor kønnet ikke kan afgøres, hverken direkte på baggrund af ordets form eller indirekte af konteksten. Derfor er køn oplyst som `#`.

***egne* (#5)** er tagget som et *almindeligt adjektiv* `AC`. Den nominale del af tagget `p#u#` oplyser, at adjektivet her optræder i *flertal* `p`, at det ud fra formen ikke kan afgøres, om den er bestemt eller ubestemt `#`, at kasus er *umarkeret* `u`, og at køn ikke kan afgøres `#`. Endvidere er gradbøjningen sat til *positiv* `p` i diverse-kategorien af tagget.

***beslutningstagere, økonomer* (#6, #8)** er tagget som *appellativiske substantiver* `NC`, som her optræder i *flertal* `p`, er i *ubestemt* form `i` med *umarkeret* kasus `u`. Køn kan ikke afgøres, da der er tale om en flertalsform `#`.

***er* (#7)** er tagget som et *finit verbum* `VF`. Ved verberne er den verbale del af tagget i brug – her oplyses om tid og diatese. I dette tilfælde er tid *nutid* `s` (præsens) og diatese er *aktiv* `a`. Finite verber er karakteriseret ved, at de kun har disse to bøjningskategorier.

***uddannet* (#9)** er tagget som et *verbum*, der i dette tilfælde fungerer som *perfektum participium* `VT`. Alle verber i denne funktion, hvor de er del af et komplekst verbal *er/blive uddannet*, er i deres nominale og verbale del låst fast til formerne `siu#` hhv. `t-` (*præteritum*, datid). Participiumsformer, som ikke indgår i verbaler, for eksempel *en uddannet økonom*, er tagget `VP`. Disse er ikke låst til bestemte former i nominal- og verbaldelen af bøjningsmønsteret.

# Kapitel 17

# Søgning på grupper af ord

*Hvordan søger man på flere ord – og hvad er et ord i det hele taget?*



## 17.1 Ordbeskrivelser i kantede parenteser

Hidtil har vi kun set på avancerede søgeudtryk, som var begrænset til søgninger på et enkelt ord. Således finder udtrykket

(17.1) `[lemma=".+arbejde" & pos="N"]`

alle ord, som er en form af et substantiv, hvis grundform ender på *arbejde*. Det, der står mellem kantede parenteser, er altid en opskrift på, hvilke egenskaber et ord skal opfylde, i dette tilfælde altså, at det skal være en form af et lemma, der ender på *arbejde*, som skal være tagget som substantiv.

## 17.2 Flere ordbeskrivelser efter hinanden

Vi kan udvide udtryk 17.1 til at omfatte flere ord, fx finder

(17.2)
```
[lemma=".+arbejde" & pos="N"] [word="med"]
[word="henblik"] [word="på"]
```

alle former af substantiviske lemmaer, som ender på *arbejde* efterfulgt af de tre ord *med*, *henblik* og *på*. Hvis der mellem to kantede parenteser kun står en opskrift på, hvordan ordoplysningen `word` skal se ud, fx `[word="på"]` eller `[word="med.+"]`, kan du nøjes med at skrive `"på"` eller `"med.+"` pågældende sted i udtrykket. Husk, at `word` altid er den ordoplysning, CoREST automatisk søger på, hvis intet andet er oplyst, og attributtet `word` matcher korpusord i forenklet ortografi. Således kan udtryk 17.2 forenkles til

(17.3) `[lemma=".+arbejde" & pos="N"] "med" "henblik" "på"`.

Er du ligeglad med, om formerne af lemmaet er substantiver eller ej, kan udtryk 17.3 også gengives med et simpelt søgeudtryk, nemlig

(17.4) `$@.+arbejde med henblik på`.

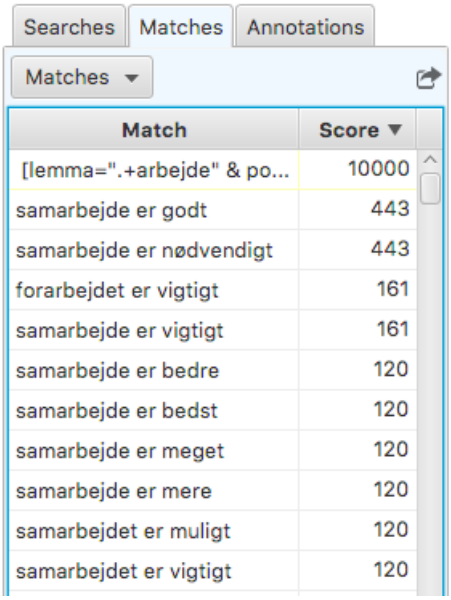
I dette tilfælde får du dog også verber med *arbejde* med i dit resultat. Avancerede søgeudtryk giver en højere grad af præcision end simple, men omkostningen er et mere komplekst søgeudtryk.

Vi vil nu ændre vores eksempel, så vi kan finde former af et substantivisk lemma, der ender på *arbejde* efterfulgt af en form af verbet *være* efterfulgt af et adjektiv. Søgeudtrykket, vi skal formulere i dette tilfælde, ser sådan ud:

(17.5)
```
[lemma=".+arbejde" & pos="N"] [lemma="være"]
[pos="A"].
```

De hyppigst forekommende eksempler i KORPUS-DK er vist i figur 17.1 på næste side. Der forekommer også matchet *samarbejde er meget*, hvor *meget* ganske vist er et adjektiv (lemmaform *megen*), dog anvendes det adverbielt i dette tilfælde, som `epos`-tagget viser, idet underklassen er oplyst som `D`, som står for et adjektiv i adverbiel anvendelse, jf. afsnit 16.2.4 på side 87. Det følgende søgeudtryk begrænser søgningen i udtryk 17.5 til adjektiver i »almindelig« attributiv anvendelse:

(17.6)
```
[lemma=".+arbejde" & pos="N"] [lemma="være"]
[epos="AC.*"].
```




Figur 17.1: Match på grupper af ord

## 17.3 Jokertegn på ord

Et problem med søgeudtrykkene 17.5 og 17.6 er, at de ikke fanger tilfælde, hvor der mellem formen af *være* og adjektivet står adverbielle størrelser, som fx *samarbejde er meget vigtigt*. Du kunne ændre dit søgeudtryk til

(17.7)
```
[lemma=".+arbejde" & pos="N"] [lemma="være"]
[epos="AD.*"] [epos="AC.*"],
```

men så ville det udelukke de eksempler, du fandt med søgeudtrykkene 17.5 og 17.6. Du skal altså oplyse, at ordet beskrevet ved `[epos="AD.*"]` i udtryk 17.7 også kan udelades. Dette gør du ved at anvende regulære udtryk, se kapitel 14, direkte på ord. Således betyder `[epos="AD.*"]?` nul eller én forekomst af ordet, altså netop det, du har brug for. Søgeudtrykket

(17.8)
```
[lemma=".+arbejde" & pos="N"] [lemma="være"]
[epos="AD.*"]? [epos="AC.*"]
```

kombinerer således søgeudtrykkene 17.6 og 17.7 til ét. Du kan også skrive `[epos="AD.*"]*` i udtryk 17.8, som betyder nul, én eller flere forekomster af adverbielt brugte adjektiver pågældende sted i søgeudtrykket. Endelig kan du oplyse antal-intervaller i krøllede parenteser, altså fx `[epos="AD.*"]{0,1}` eller `[epos="AD.*"]{0,}`, hvor `{0,1}` betyder »mindst nul og højst 1«, mens `{0,}` betyder »mindst nul«, hvilket jo henholdsvis svarer til `?` og `*`. Selvfølgelig kan der også bruges andre tal end 0 og 1 i intervalangivelserne, dog skal de være positive, og det første tal skal være mindre end eller lig med det andet.

Du kan i dine søgeudtryk også bruge kantede parenteser uden yderligere beskrivelse imellem: `[]`. Det betyder »et vilkårligt ord på dette sted« og svarer således til `#` i simple søgeudtryk. Disse vilkårlige ord kan ligeledes udvides med angivelserne `?`, `*` eller intervalangivelser i krøllede parenteser, som du i øvrigt også kan bruge efter `#` i simple søgeudtryk.

## 17.4 Hvad er et ord?

I CoREST bruges en ret primitiv regel for opdelingen af en tekst i ord:

**Ord** = en række af bogstaver, der står mellem to mellemrum eller sætningstegn.

### 17.4.1 Bindestreg, apostrof og skråstreg

Som sætningstegn og dermed ordskilletegn tæller også bindestreg, apostrof og skråstreg. Derfor er »naturlige« ord som *e-butik*, *to-værelses* eller *pc'en* i CoREST's verden hver især to ord. Hvis du vil søge på dem, må du formulere dine søgeudtryk tilsvarende. Du kan for eksempel finde de ovenstående eksempler ved hjælp af de følgende simple søgeudtryk:

(17.9)
```
$e butik
$to værelses
$pc en
```

Ellers kan du finde sætningstegn ved hjælp af ordoplysningen *bound*, fx finder søgeudtrykket

(17.10) `[word="[a-zøæå]{2,}"& bound="'"]`

alle ord på mindst to bogstaver efterfulgt af apostrof.

Da skråstregen også fungerer som ordskilletegn, er ord som *A/S* eller *km/t* splittet op i to, altså *A* og *S*, *km* og *t*. I konkordanser vises skråstreger altid normalt, jf. fx *A/S*. Internt i CoREST er skråstregen imidlertid af tekniske grunde repræsenteret som omvendt skråstreg: \. Du kan søge på skråstreger ved hjælp af søgeudtrykket

(17.11) `[bound="\\"]`.

I søgeudtryk skal en skråstreg altid angives som `\\`.

At bindestreg, apostrof og skråstreg betragtes som sætningstegn har konsekvenser for ordklasseopmærkningen og for søgningen af ord, der indeholder bindestreg eller semikolon, læs mere herom i kapitel 16.

### 17.4.2 Sætningstegn og mellemrum

Et sætningstegn mellem to ord er nok til at skille dem ad, der behøver ikke være et mellemrum. Derfor opfattes *2,0*, *2.0*, *2-0*, *bl.a.*, *m.fl.*, *cand.scient.* alle som to ord. Disse eksempler kan findes med de følgende simple søgeudtryk:

(17.12)
```
$2 0
$bl a
$m fl
$cand scient.
```

Søgeudtrykkene finder også varianter, hvor der ud over selve det adskillende sætningstegn også skulle have sneget sig yderligere mellemrum ind som for eksempel i *bl.␣a.*[1] eller *cand.␣scient.*, men matcher også *cand.scient.*, når det for eksempel indgår i *cand.scient.pol*. Vil du i din søgning differentiere mellem tilfælde med eller uden ekstra mellemrum eller om sætningstegnet er bindestreg, komma, punktum eller andet, må du anvende avancerede søgeudtryk med ordoplysningen *bound* på grupper af ord.

Vil du i din søgning skelne mellem tilfælde med og uden ekstra mellemrum eller mellem sætningstegn som bindestreg, komma, punktum, kan du ved hjælp af et avanceret søgeudtryk med ordoplysningen *bound* søge på grupper af ord, fx

(17.13) `[word="2" & bound="\."] "0"`

for at matche *2.0* eller

(17.14) `[word="cand" & bound="\."] [word="scient" & bound="\._"]`

for at matche *cand.scient.␣* – hvilket udelukker *cand.scient.pol.*, men ikke fx *cand.scient.␣pol.* eller *cand.scient.␣soc.* Hvis *cand.scient.pol.'erne* og *-soc.'erne* skal udelukkes helt, kan du bruge udtrykket

(17.15)
```
[word="cand" & bound="\."] [word="scient" & bound="\._"]
[word!="pol|soc"].
```

### 17.4.3 Sær- og samskrivning

Særskrevne ord (fx *i går*, *ad hoc*, *happy hour*, *per se*, *status quo*, *over for*, *hen imod*, *hokus pokus*, *her til lands*, *tids nok*) betragtes altid som flere ord i CoREST, for eksempel er *over for* to ord nemlig *over* og *for*. Nogle af de nævnte ord kan også samskrives (*overfor*, *henimod*, *hokuspokus*, *hertillands*, *tidsnok*), mens andre ofte samskrives, selvom man ikke bør, for eksempel *igår*. Når ord af denne type samskrives, betragtes de som ét ord. Denne mulighed for variation skal du tage højde for i dine søgninger. Hvis du fx vil finde eksempler på *over for* i to ord, kan

[1] Tegnet ␣ bruges her i manualen til at gøre det tydeligt, at der på denne plads står et mellemrum. Det kan ikke bruges i søgeudtryk og bliver heller ikke vist i CoREST.

du indtaste det simple søgeudtryk `$over for`, hvis du derimod vil finde *overfor* i ét ord, indtaster du det simple søgeudtryk `overfor`. Du kan finde begge varianter ved hjælp af ét avanceret søgeudtryk, nemlig

(17.16) `("over""for") | "overfor"`.

Sær- eller samskrivning har også konsekvenser for ordklasseoplysningerne, for eksempel vil *hertillands* blive markeret som ét adverbium, mens *her til lands* vil blive markeret som et adverbium (*her*), efterfulgt af en præposition (*til*), efterfulgt af et substantiv (i genitiv): *lands*.

### 17.4.4 Alfabet

CoREST kan kun håndtere bogstaver fra det latinske alfabet. Eventuelt forekommende bogstaver fra andre alfabeter er enten udeladt eller omsat til særlige tegnfølger, typisk en række af spørgsmålstegn.

## 17.5 Søge inden for sætningsgrænser

Et problem ved søgeudtryk, der matcher grupper af ord, er, at de ignorerer sætningsgrænser og bare matcher hen over dem. Således finder udtrykket 17.8 på side 95 også eksemplet:

- ✧ *Siden har han modtaget flere andre priser for sit* ***revyarbejde. Er kunstnerisk*** *leder, instruktør og skuespiller*

med *revyarbejde* som en form af substantivet *revyarbejde*, *Er* (i begyndelsen af en ny sætning) som en form af *være* og *kunstnerisk* som et ikke adverbielt brugt adjektiv. Blot er dette ikke et eksempel på det forventede, idet matchet går hen over en sætningsgrænse. Man kunne udelukke denne type match ved at tilføje `& bound!="\..*"` til første, andet og tredje ord i søgeudtrykket. Nemmere er det dog at tvinge CoREST til kun at matche inden for en sætning ved at tilføje `within s` til søgeudtrykket:

(17.17)
```
[lemma=".+arbejde" & pos="N"] [lemma="være"]
[epos="AD.*"]? [epos="AC.*"] within s
```

Udtrykket 17.17 giver således det ønskede resultat. Da opsplitningen af korpusteksterne i sætninger er sket automatisk, og der her sagtens kan være sket fejl, er tilføjelsen af `within s` ikke altid helt pålidelig.

# Kapitel 18

# Søgning på tekstoplysninger

*Brug af tekstoplysninger som filtre i søgninger*



Som regel er der i et korpus knyttet en række tekstoplysninger (metadata) til hver enkelt tekst. Nogle af disse oplysninger kan du bruge til at indsnævre dine søgninger med, altså »filtrere« dem, så du kun får de resultater at se, hvor tekstoplysningerne svarer til dem, som du har fastlagt. Hvilke metadata du kan bruge som filtre i dine søgninger, afhænger principielt af det korpus, du har valgt. I CoREST bruges imidlertid de samme metadata for alle standardkorpusser.

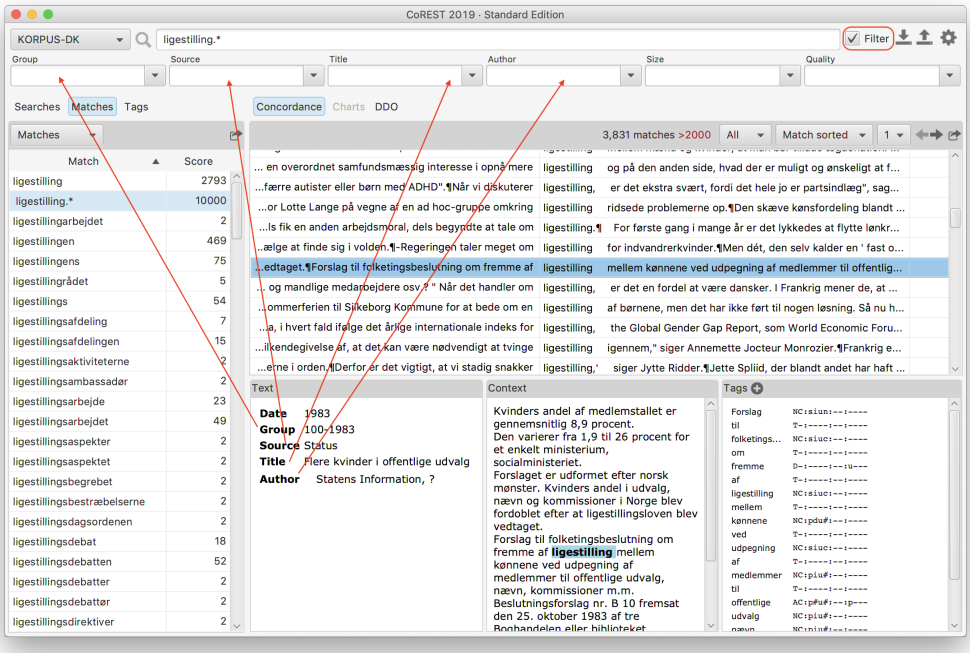
## 18.1 Brug af filterfunktionen

Du slår filterfunktionen til ved at sætte et hak i »Filter«-boksen til højre for søgefeltet, se figur 18.1 på næste side. Under søgefeltet åbner der sig en række af seks felter, hvor du kan specificere de tekstfiltre, der skal gælde for søgningen.

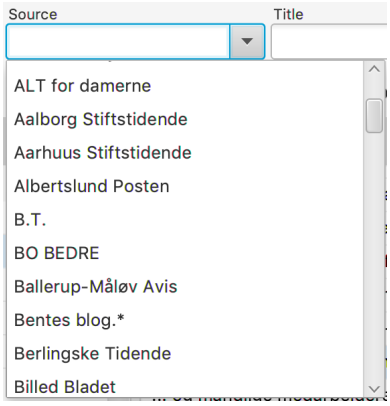
Du kan enten indtaste et filterudtryk i et af felterne eller vælge et fra listen, som kommer frem, når du klikker på den lille trekant til højre i hvert af felterne, se figur 18.2 på den følgende side med eksempler for tekstoplysningstypen »Source«. Listerne indeholder et udpluk af de hyppigste værdier, som forekommer i pågældende korpus, men der er ikke altid defineret værdier for alle felter og ikke nødvendigvis for alle korpusser. Typisk vil der mangle prædefinerede værdier for felterne »Title« og »Author«.

I filterudtrykkene kan du bruge de sædvanlige jokertegn, jf. kapitel 14, dog skelnes der aldrig mellem store og små bogstaver. En bindestreg (»minus«) som første tegn i et filterfelt betyder, at filteret ikke må være opfyldt. Filtrene »Group«, »Source«, »Title« og »Author« er identiske med de oplysninger, du finder i området med tekstoplysningerne under konkordansen, se figur 18.1 på næste side, og de er beskrevet i tabel 8.1 på side 53.

Du kan lade dig inspirere af de værdier, som vises i området med tekstoplysningerne, til at formulere dine egne filterværdier, idet du fx overtager avisnavne




Figur 18.1: Sammenhængen mellem tekstoplysninger og filtre




Figur 18.2: Prædefinerede værdier for tekstoplysningstypen »Source«

(fx `Politiken`), bestemte ord fra titler (fx `.*sanktion.*`) osv. i dine filtre. Særligt for feltet »Source« kan du fx bruge en af kilderne under hovedkildefortegnelserne i appendiks B på side 123, men ofte er det lettest blot at vælge noget fra de prædefinerede lister.

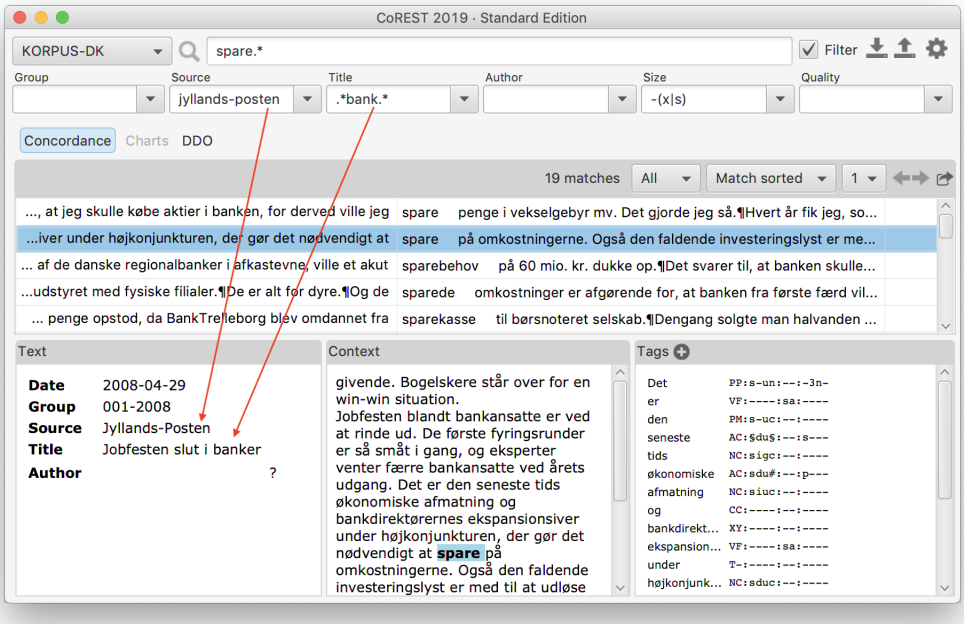
I KORPUS-DK kan du bruge filtret »Group« til at begrænse din søgning til bestemte dele af korpus. Du finder en beskrivelse af tekstgrupper i appendiks B.

Ud over de nævnte fire filtre, som umiddelbart hænger sammen med tekstoplysningerne, findes der yderligere to, nemlig »Size« og »Quality«.

**Size** fastlægger tekstens omfang. Der er fire kategorier: `L` (mindst 10.000 ord), `M` (1.000–9.999 ord), `S` (200–999 ord) og `XS` (under 200 ord). Mindre tekststørrelser vil overvejende være avisartikler, mens længere tekster kan komme fra kilder, som ikke nødvendigvis er nyhedsstof.

**Quality** fastlægger ikke teksters, men sætningers sproglige kvalitet (og er derfor ikke metadata på tekstniveau, men på sætningsniveau). Der er to kvalitetstrin: `hi` og `lo`. Opdelingen af sætninger i en formodet højere eller lavere sproglig kvalitet er beregnet automatisk og er af eksperimentel karakter.

Figur 18.3 viser et eksempel, hvor der er søgt på søgeudtrykket `spare.*` i KORPUS-DK i tekster fra Jyllands-Posten, hvor der i avisartiklens overskrift indgår strengen *bank*, og hvor de fundne eksempler skal stamme fra en tekst, som ikke må have under 1000 ord: `-(xs|s)`. Bemærk, at figuren viser CoREST-vinduet med usynligt ordlisteområde, jf. kapitel Chapter 3.




Figur 18.3: Søgning med tekstfiltre

## 18.2 Tekstoplysninger som del af søgeudtrykket

CoREST omsætter samtlige søgeudtryk til en særlig formalisme, som kan fortolkes af søgemaskinen CQP, der ligger til grund for CoREST-systemet, jf. kapitel Chapter 1. Du kan se de formelle CQP-søgeudtryk, som CoREST anvender internt, i sessionsprotokollen, jf. kapitel 5. Den samme formalisme kan du også anvende umiddelbart i de søgeudtryk, som du indtaster i CoREST's søgefelt. De seks

filtre, som du kan fastlægge i filterområdet, betragtes internt som ordoplysninger på linje med `ortho`, `bound`, `lemma` etc., jf. kapitel 10. Filtrenes attributnavne fremgår af tabel 18.1.

| Filter | Attributnavn |
|---|---|
| »Group« | `att0` |
| »Source« | `att1` |
| »Title« | `att2` |
| »Author« | `att3` |
| »Size« | `att4` |
| »Quality« | `att5` |

Tabel 18.1: CoREST-interne attributnavne for tekstfiltre

Vil du søge med tekstfiltre og definere filtrene som en del af dit søgeudtryk, kan du altså nøjes med at tilføje de ønskede tekstoplysninger med `&`-tegnet til ét af ordene i søgeudtrykket. Således fører søgeudtrykket

(18.1)
```
[word="spare.*" & att1="jyllands-posten" & att2=".*bank.*"
& att5!="(xs|s)"]
```

til samme resultat som eksemplet i figur 18.3 på forrige side. Når du udfører søgninger med tekstfiltre i selve søgeudtrykket, skal du være opmærksom på følgende tre ting:

1. Tekstattributternes værdier skal altid oplyses med små bogstaver; i eksempel 18.1 ville `att1="Jyllands-Posten"` ikke medføre noget resultat. Mellemrum skal oplyses som understreg, fx `att1="berlingske_tidende"`. I filterfelterne, se figur 18.3 på forrige side, må du derimod godt bruge både store bogstaver og mellemrum, da de automatisk omsættes til små bogstaver og understreger inden søgningen.

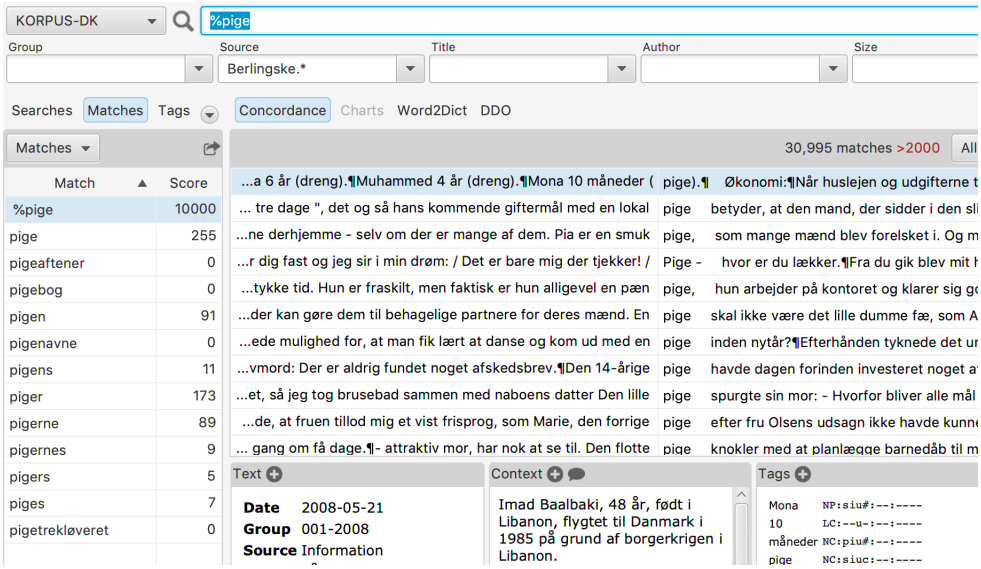
2. Der bør ikke være hak i »Filter«-boksen til højre for søgefeltet, se figur 18.3 på foregående side, da du ellers også ville søge på de øvrige definerede tekstfiltre og ikke kun dem, som du har oplyst i dit søgeudtryk.

3. Hvis du formulerer en flerordssøgning, jf. kapitel 17, bør du knytte alle dine tekstfiltre til kun ét ord i søgeudtrykket, fx det første. Internt behandles tekstoplysninger som ordoplysninger, hvilket strengt taget ikke er logisk, fordi det åbner for selvmodsigende udtryk som det følgende, som ikke giver noget resultat:

   (18.2)
   ```
   [word="spare.*" & att1="jyllands-posten"] [word="op" &
   att1="berlingske_tidende"]
   ```

## 18.3 Kontroltegnet % kombineret med tekstoplysninger

Som udgangspunkt bør du i dine søgninger undgå kontroltegnet %. I stedet for bør du enten bruge kontroltegnet @ eller søge med efterstillet .*.[1] Søgninger med kontroltegnet % vil endvidere medføre fejl, hvis du kombinerer dem med tekstfiltre. Dette skyldes en fejl i fortolkningen af søgeudtrykket. Fejlen vil blive rettet i en fremtidig version af CoREST. Hvis du fx søger på %pige med tekstfilteret »Source« sat til Berlingske.*, vil du få et resultat, hvor mange af forekomsterne ikke stammer fra en *Berlingske*-kilde, se figur 18.4, hvor »Source« for den første konkordanslinje er oplyst som *Information*.




Figur 18.4: Søgning med kontroltegnet % med tekstfilter giver fejlagtigt resultat

Når der sker den slags mystiske ting, er det altid en god idé at kigge i sessionsloggen, se kapitel 5. I sessionsloggen vil du se, at søgeudtrykket %pige blev omsat til søgeudtryk 18.3 inden søgningen:

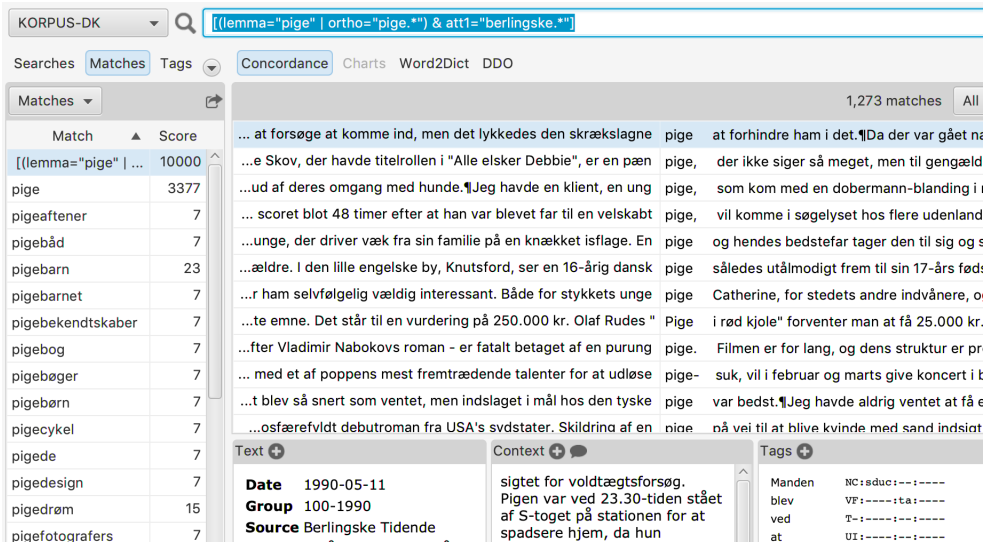
(18.3) `[lemma="pige" | ortho="pige.*" & att1="berlingske.*"]`.

Da et logisk **OG** (&) har en højere prioritet end et logisk **ELLER** (|), finder dette udtryk først alle ortografiske former, som begynder med *pige* **OG** optræder i *Berlingske*. Herefter kombineres det fundne med **ELLER**-delen af udtrykket, nemlig alle former af lemmaet *pige* (uanset kilde). Fejlen skyldes, at der mangler et par parenteser i udtrykket. Den korrekte form ser du i søgeudtryk 18.4:

(18.4) `[(lemma="pige" | ortho="pige.*") & att1="berlingske.*"]`.

[1] Du kan læse mere om kontroltegn i kapitel 4.

Prøv at fjerne hakket under »Filter« i CoREST og udfør en søgning på udtryk 18.4 på foregående side, altså med parenteser indsat. Du vil da se, at resultatet er som forventet, se figur 18.5.




Figur 18.5: Søgning med avanceret tekstfilterudtryk

# Del IV

# Korpusser

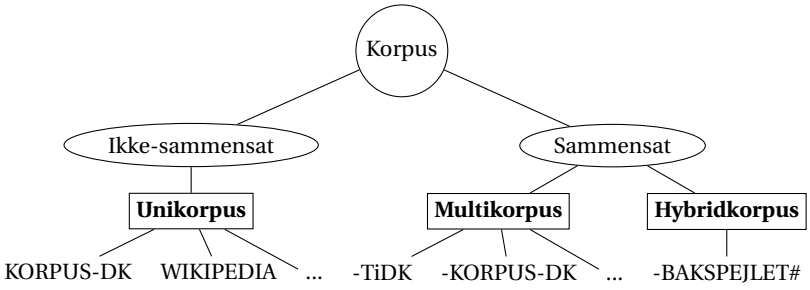
# Kapitel 19

# Typer af korpusser

*Om uni-, multi- og hybridkorpusser*



I CoREST findes der tre typer korpusser, nemlig *uni-*, *multi-* og *hybrid*-korpusser. Multi- og hybridkorpusser er beslægtede, idet der i begge tilfælde er tale om *sammensatte* korpusser, mens unikorpusser er *ikke-sammensatte*. Hybridkorpusser findes kun i DSL-udgaven af CoREST. Der knytter sig en særlig navnekonvention til hver af de forskellige typer korpusser. Generelt består korpusnavne altid af store bogstaver – med undtagelse af *i*'et i TiDK. Navnene kan indeholde bindestreger og cifre. Multikorpussers navne *begynder* derudover altid med en bindestreg, fx -KORPUS-DK, mes hybridkorpussers navne altid *begynder* med en bindestreg og *slutter* med en havelåge, fx -BAKSPEJLET#[1], se figur 19.1, som giver et overblik over de tre typer, og under dem nogle eksempler på korpusser i CoREST.[2]




Figur 19.1: Typer af korpusser

I de følgende afsnit beskrives de forskellige typer korpusser nærmere, dog beskrives hybridkorpusser kun i DSL-udgaven af denne manual.

[1] Dette korpus findes kun i DSL-udgaven af CoREST.
[2] En oversigt over samtlige CoREST-korpusser findes i tabel 19.1 på side 108.

## 19.1 Unikorpus

Et unikorpus udgør én enhed, korpusset er ikke sammensat af enkelte mindre korpusser, men er selvstændigt. Unikorpusser er standardkorpustypen i CoREST. En søgning i et unikorpus gennemsøger altid hele korpusset på én gang og frembringer som resultater én matchliste og én konkordans osv. for hele korpus. KORPUS-DK er et eksempel på et unikorpus. Her i manualen beror alle eksempler på unikorpusset KORPUS-DK, med mindre andet er oplyst.

## 19.2 Multikorpus

Et multikorpus består af flere, omtrent lige store delkorpusser, som hvert især i princippet også kunne være et særskilt unikorpus i CoREST, men i den nuværende version af CoREST ikke er det. Kriteriet for opdelingen i delkorpusser er i den nuværende version af CoREST tilblivelsestidspunktet for teksterne. Således vil et delkorpus typisk indeholde materiale fra en bestemt periode, fx en årrække eller et enkelt år. I korpusvælgeren kan et multikorpus kendes på, at dets navn begynder med en bindestreg. Multikorpusset -TiDK 2020 består af ét delkorpus per år for perioden 2010-2019 med delkorpusnavnene TI-2010 til TI-2019. -KORPUS-DK er en multikorpusudgave af KORPUS-DK og består af tre delkorpusser: KORPUS-90, KORPUS-2000 og KORPUS-2010.[3] Når du søger i et multikorpus, vil du få vist resultater for hvert delkorpus for sig, altså én konkordans per delkorpus, én ordliste osv. Ved hjælp af delkorpusvælgeren over konkordansen kan du vælge, hvilket delkorpus du ønsker at se resultater fra. Opsplitningen af et stort korpus i delkorpusser gør det lettere at lave sammenlignende undersøgelser af resultaterne for de forskellige delkorpusser. Resultater kan også præsenteres i form af diagrammer.

## 19.3 *Uni* over for *multi* – fordele og ulemper

Fordele og ulemper illustreres her med udgangspunkt i multikorpusser.

✧ Fordele ved multikorpusser:

1. Mulighed for grafiske oversigter, der kan vise et sprogligt fænomens udbredelse over tid.
2. Mulighed for visning af flere resultater: 2000 per delkorpus i stedet for 2000 i alt.
3. Nem vedligeholdelse, idet korpusadministratoren ved ændringer i korpus kun skal genindeksere det ændrede delkorpus og ikke det hele.

[3] Den rigtige stavning for disse korpusnavne er *Korpus 90*, *Korpus 2000* og *Korpus 2010*. Her er korpusnavnene omsat til CoREST's konvention, som normalt ikke bruger små bogstaver og mellemrum i navnene.

✧ Ulemper ved multikorpusser:

1. Resultater (matchlister og konkordanser) kan kun vises for ét delkorpus ad gangen, hvilket betyder, at du skal blade fra delkorpus til delkorpus med delkorpusvælgeren for at se resultaterne for de enkelte delkorpusser. Der er ingen mulighed for at se resultaterne samlet, hvilket især kan være problematisk ved ordlister.
2. Brugerannoteringer vedrører altid kun det aktuelt valgte delkorpus. Man kan altså ikke se og vælge annoteringer for hele korpus.

For unikorpusser forholder det sig omvendt med fordelene og ulemperne: Ingen grafiske oversigter, tidskrævende vedligeholdelse af meget store korpusser. Dette står over for visning af samtlige resultater og brugerannoteringer på én gang.

## 19.4 Oversigt over samtlige korpusser i CoREST

Tabel 19.1 giver et overblik over samtlige korpusser, som er tilgængelige i de forskellige udgaver af CoREST. Af tabellen fremgår korpussets type og dets struktur. Med korpusstrukturen menes opbygningen af et korpus med hensyn til, hvilke ordoplysninger og tekstoplysninger der findes. Svarer de til det princip, som gælder for KORPUS-DK og dermed er beskrevet i manualen her, er der tale om en CoREST-struktur. Er der andre ord- og tekstoplysninger på spil, har pågældende korpus en speciel struktur.

| Navn | Type | Struktur | Standard | Research | Ømål | DSL |
|---|---|---|---|---|---|---|
| KORPUS-DK | uni | CoREST | ✓ | ✓ | ✓ | ✓ |
| -KORPUS-DK | multi | CoREST | ✓ | ✓ | ✓ | ✓ |
| -TiDK-20xx | multi | CoREST | ✓ | ✓ | ✓ | ✓ |
| WIKIPEDIA | uni | CoREST | ✓ | ✓ | ✓ | ✓ |
| SDEWAC | uni | speciel | | ✓ | | ✓ |
| DIAKO | uni | speciel | | | ✓ | ✓ |
| -BAKSPEJLET | multi | CoREST | | | | ✓ |
| -BAKSPEJLET# | hybrid | CoREST | | | | ✓ |
| DDO-classic | uni | CoREST | | | | ✓ |
| S-90 | uni | speciel | | | | ✓ |
| SPORDHUND | uni | CoREST | | | | ✓ |

Tabel 19.1: Oversigt over korpusser i CoREST

# Kapitel 20

# Standardkorpusser

*Om de korpusser, som findes i alle udgaver af CoREST*



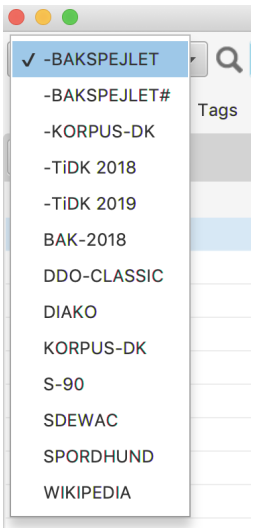
I dette kapitel beskrives alle korpusser, som er tilgængelige i Standard-udgaven og dermed i samtlige udgaver af CoREST, nemlig KORPUS-DK, TiDK og WIKIPEDIA, se også tabel 19.1 på foregående side. I kapitel 21 beskrives korpusserne i specialudgaverne af CoREST.

Alle korpusser, som beskrives i dette og det følgende kapitel, aktiveres i CoREST ved hjælp af korpusvælgeren, se figur 20.1 på den følgende side. I figuren vises korpusser i DSL-udgaven af CoREST, i andre udgaver vil der være færre korpusser at vælge imellem.

## 20.1 KORPUS-DK

KORPUS-DK er det centrale korpus i alle udgaver af CoREST, og det er det korpus, som ligger til grund for de fleste eksempler her i manualen. Korpusset er et unikorpus, jf. kapitel 19, som samler de oprindeligt selvstændige korpusser Korpus 90 med tekster fra omkring 1990, Korpus 2000 med tekster fra omkring år 2000 og Korpus 2010 med tekster fra omkring 2010 til ét unikorpus. De tre oprindelige korpusser findes ikke som selvstændige unikorpusser i CoREST.

KORPUS-DK er på over 100 millioner ord i alt, det er ePOS-tagget, jf. kapitel 16, og det tillader brugen af en række tekstoplysninger i søgningerne, jf. kapitel 18. En delmængde af KORPUS-DK, nemlig Korpus 90 og Korpus 2000 er også




Figur 20.1: Korpusser i korpusvælgeren (DSL-udgaven)

tilgængelig under navnet *KorpusDK* på ordnet.dk-sitet. I det følgende gennemgås sammensætningen i de tre oprindeligt selvstændige korpusser, som er samlet i KORPUS-DK, i detaljer.

### 20.1.1 Korpus 90

Korpus 90 er en del af Den Danske Ordbogs korpus, som blev udarbejdet i 1991-1993 til brug for redigeringen af den trykte udgave af Den Danske Ordbog (DDO (2005), DDO).

Forbilledet for Den Danske Ordbogs korpus var British National Corpus[1], dvs. man tilstræbte at sammensætte DDO's korpus tilsvarende varieret. Der indgik både skrift- og talesprog. Skriftsprog er både nyhedsmedier, skøn og faglitteratur, private breve, dagbøger, reklamer m.m. Talesprog omfatter diverse interviews, transskriberet materiale fra radio- og tv-udsendelser, forskellige protokoller m.m. I Korpus 90 er talesprogstekster fjernet helt, desuden er tekster, som leverandørerne har belagt med særlige brugsrestriktioner, også fjernet. Korpussets oprindelige tilblivelse og sammensætning er beskrevet af Norling-Christensen og Asmussen (1998).[2] Korpus 90 omfatter cirka tre fjerdedele af det oprindelige DDO-korpus.

Korpus 90 indeholder godt 32 millioner ords tekst. Du kan læse flere detaljer om korpussets sammensætning i appendiks B.1.

[1] Se `http://www.natcorp.ox.ac.uk/`

[2] ja.dsl.dk/papers/Lexikos 1998/lexikos8.pdf

### 20.1.2 Korpus 2000

Korpus 2000 blev udarbejdet under DSL-projektet af samme navn, som blev gennemført i årene 2000 til 2002. Projektet blev støttet af *År 2000 Fonden*, som var oprettet til at støtte projekter til markering af årtusindskiftet. Formålet med Korpus 2000-projektet var at opbygge et tekstkorpus, der skulle dokumentere alment dansk skriftsprog i årene omkring år 2000, og korpusset skulle gøres offentlig tilgængelig på internettet. Forbilledet for dette korpus var skriftsprogsdelen af DDO's korpus, jf. afsnit 20.1.1 på foregående side. Da projektressourcerne medførte en række begrænsninger, endte Korpus 2000 med at have en overvægt af avistekster, som det er nemmest at få i store mængder. Forskellige aspekter af Korpus 2000 er beskrevet i Asmussen (2001)[3] og Andersen et al. (2002)[4].

Korpus 2000 indeholder godt 30 millioner ords tekst. Du kan læse flere detaljer om korpussets sammensætning i appendiks B.2.

### 20.1.3 Korpus 2010

Korpus 2010 er et »referencekorpus« over dansk almensprog omkring 2010. Det blev opbygget som led i det danske CLARIN-projekt[5], som er dokumenteret under korpus.dsl.dk. Også dette korpus endte med en overvægt af nyhedsstof, men der findes også blog- og forummateriale i det og noget materiale fra Wikipedia. Metadata- og tekstformatering er beskrevet i Asmussen og Halskov (2009).[6] Tekster kan muligvis downloades fra CLARIN-projektets repositorium under clarin.dk. Adgang forudsætter dog et særligt login, som kan være vanskeligt at få.

Korpus 2010 indeholder omtrent 45 millioner løbende ord. Du kan læse flere detaljer om korpussets sammensætning i appendiks B.3 på side 125.

### 20.1.4 Søge i bestemte dele af KORPUS-DK

I KORPUS-DK bruges nogle særlige tekstgruppeoplysninger, som faktisk gør det muligt at søge direkte i de tre oprindelige korpusser Korpus 90, Korpus 2000 og Korpus 2010, selvom KORPUS-DK er et unikorpus og dermed slet ikke består af selvstændige delkorpusser. Det er derudover muligt at søge i en hvilken som helst kombination af de oprindelige korpusser. Tekstgruppeoplysningerne i KORPUS-DK består af to tal adskilt ved bindestreg, fx *100-1986* i figur 20.2 på næste side ud for tekstattributtet »Group«.

[3] ja.dsl.dk/papers/nys_2001/nys-30.pdf

[4] ja.dsl.dk/papers/euralex_2002/The Project of Korpus 2000 Going Public.pdf

[5] CLARIN: Common Language Resources and Technology Infrastructure, 2008-2011. Betegnelsen »referencekorpus« for dette korpus tager sit udgangspunkt i CLARIN-projektbeskrivelsen, jævnfør korpus.dsl.dk/clarin/dkclarin-project-desc.pdf, WP 2.1, men betegnelsen beskriver det faktiske resultat, Korpus 2010, ikke helt korrekt. Da opbygningen af et referencekorpus – altså et varieret sammensat korpus, der afspejler flest mulige varianter af sproget – er meget ressourcekrævende, var betegnelsen fra første færd misvisende, da projektet kun havde stærkt begrænsede midler til rådighed.

[6] ja.dsl.dk/papers/corpus_linguistics_2009/287_FullPaper.html


Figur 20.2: Eksempel på en tekstgruppeoplysning i KORPUS-DK

Tallet efter bindestregen er årstallet for udgivelsen af teksten, *1986* i figuren, mens det trecifrede tal foran bindestregen er korpuskoden, *100* i figuren. Korpuskoden er en kode for det oprindelige korpus, som teksten stammer fra. Tabel 20.1 viser koderne for de tre korpusser, som KORPUS-DK er sammensat af.

| Kode | Delkorpus | Årstal |
|---|---|---|
| *100* | Korpus 90 | 1983-1992 |
| *010* | Korpus 2000 | 1993-2002 |
| *001* | Korpus 2010 | 2004-2011 |

Tabel 20.1: Korpuskoder i »Group«-oplysningstypen

Ved at bruge korpuskoderne i »Group«-tekstfiltret kan du tilpasse din søgning til kun at vedrøre et eller flere af de tre oprindelige korpusser, som er indgået i KORPUS-DK. I tabel 20.2 på næste side kan du se, hvad du skal indtaste i filterfeltet »Group« for at søge i de forskellige delkorpusser eller kombinationer af dem. Alle indtastede værdier i »Group«-feltet skal afsluttes med jokertegnet **.*** som vist i tabellen, fordi gruppebetegnelsen oplyser udgivelsesåret efter bindestregen, som vi ikke tager højde for i dette eksempel. Naturligvis kan du også inddrage årstal, hvis du ønsker det.

| Indtast | Søger i |
|---|---|
| `100-.*` | Korpus 90 |
| `010-.*` | Korpus 2000 |
| `001-.*` | Korpus 2010 |
| `..0-.*` | Korpus 90+2000 |
| `.0.-.*` | Korpus 90+2010 |
| `0..-.*` | Korpus 2000+2010 |

Tabel 20.2: Søgning i delkorpusser

### 20.1.5 KORPUS-DK som multikorpus

KORPUS-DK findes også i en multikorpusversion. Du aktiverer den i CoREST ved at vælge -KORPUS-DK (med bindestreg foran) i korpusvælgeren. I multikorpusversionen består KORPUS-DK af de tre delkorpusser KORPUS-90, KORPUS-2000 og KORPUS-2010, og du kan i delkorpusvælgeren over konkordansen vælge, hvilket af disse tre delkorpusser du vil resultater fra.

### 20.1.6 Sammenlignelighed

Hvis du vil bruge korpusser fra forskellige perioder til sammenlignende tidsstudier, skal du sikre dig, at korpusserne er ensartet sammensat. Dette er kun delvist tilfældet for de tre korpusser, der indgår i KORPUS-DK. Et korpus' sammensætning kan beskrives ved hjælp af en række parametre. En af dem er *kildevariation*.

#### Definition af kildevariation

Når en hovedkilde er defineret som en kilde, som har bidraget med mindst én promille til korpussets omfang, kan man definere et optimalt varieret korpus som et, der kun består af hovedkilder, hvoraf ingen dominerer de andre, som altså har bidraget med lige meget hver. Et sådant korpus ville bestå af 1000 hovedkilder, der tilsammen dækker 1000 promille af omfanget. Vi kan nu definere et variationsindeks $v = \frac{s}{e} \cdot 1000$, hvor $s$ er antallet af hovedkilder og $e$ er deres samlede omfang af korpus i promille. Lavest tænkelige kildevariation har indeks 1 og indtræffer, hvis hele korpus ($e = 1000$) kun udgøres af én kilde ($s$=1), det højest tænkelige variationsindeks 1000 får vi, hvis korpusset er optimalt varieret (1000 kilder à 1 promille).

I appendiks B vises lister over hovedkilderne i Korpus 90, Korpus 2000 og Korpus 2010. Korpus 90 har 77 hovedkilder, som tilsammen bidrager med 608 promille til det samlede korpus, hvilket giver et kildevariationsindeks på 127. Korpus 2000 har 45 hovedkilder, som tilsammen bidrager med 969 promille til det samlede korpus svarende til et kildevariationsindeks på 46. Korpus 2000 er med

andre ord mindre varieret sammensat end Korpus 90. Korpus 2010 har 29 hovedkilder som tilsammen bidrager med 996 promille til det samlede korpus, og som derfor har et kildevariationsindeks på 29. Korpus 2010 er altså endnu mindre varieret sammensat end Korpus 2000.[7]

Du kan læse mere om sammenligning af korpusser og sammenlignende undersøgelser i Asmussen (2004) og Asmussen (2006).

## 20.2 TiDK

TiDK er tilgængelig i alle udgaver af CoREST. Det er et multikorpus bestående af ét korpus for hvert af de seneste ti år. TiDK-korpusserne er monitorkorpusser, der giver et billede af sproget gennem de sidste ti år. Der udgives et nyt TiDK-korpus én gang om året, og korpusset er derfor altid udstyret med et årstal, der oplyser udgivelsesåret. Korpusset er ellers opbygget på samme måde som KORPUS-DK, dvs. at der anvendes de samme tekstoplysninger, og at ordklassetaggingen er den samme.

## 20.3 WIKIPEDIA

CoREST indeholder et korpus baseret på den danske Wikipedia, sådan som den så ud den 20.08.2017. WIKIPEDIA-korpusset er både tilgængelig som selvstændigt korpus i alle udgaver af CoREST. WIKIPEDIA-korpusset består af over 200.000 tekster med sammenlagt knap 65 millioner ord. Det indeholder ikke blot Wikipedia-artiklerne, men også diskussions- og hjælpesider fra Wikipedia. Korpusset indeholder den rå tekst fra Wikipedia, dvs. at tabeller, lister, oversigter, billeder m.m. er udeladt. Udeladelser er markeret med *{...}* i korpusset. Wikipedia-korpusset er ePOS-tagget.

[7] På baggrund af disse indices er Korpus 90 tættest på at være et referencekorpus, mens Korpus 2010 er længst væk fra at være et referencekorpus.

# Kapitel 21

# Specialkorpusser

*Om de korpusser, som findes i specialudgaverne af CoREST*



## 21.1 BAKSPEJLET

Dette korpus er kun tilgængelig i DSL-udgaven af CoREST. BAKSPEJLET er et sammensat korpus, som både findes som multikorpus og hybridkorpus. Du kan læse mere om korpustyper i kapitel 19.[1] Oplysningstyperne i BAKSPEJLET (tekstfiltre, tagging etc.) er de samme som i KORPUS-DK og TiDK. BAKSPEJLET består af over en milliard ords løbende tekst og er det primære korpus for det redaktionelle arbejde på Den Danske Ordbog.

-BAKSPEJLET er et kronologisk sammensat korpus, dog ikke med ét delkorpus pr. år som TiDK, men med kronologiske enheder med ikke alt for begrænsede tekstmængder. Således er der et delkorpus *før 1985* samt delkorpusser for femårsperioderne *1985-1989*, *1990-1994* og *2000-2004*. De øvrige delkorpusser indeholder selvstændige årgange. Ulempen ved at have samlet flere årgange i én kronologisk enhed er, at disse perioder i de grafiske oversigter vises som én samlet værdi, man får altså inden for disse perioder ikke længere et differentieret billede for de enkelte år. På den anden side ville disse billeder meget ofte være stærkt fortegnede dér, hvor der er for lidt materiale i en årgang. Tabel 21.1 på den følgende side viser de enkelte delkorpusser, som BAKSPEJLET er sammensat af.

[1] I CoREST hedder multikorpusversionen -BAKSPEJLET (med foranstillet bindestreg), mens hybridkorpusversionen hedder -BAKSPEJLET# (med foranstillet bindestreg og efterstillet havelåge). Her i kapitlet bruges for nemheds skyld navnet BAKSPEJLET om begge versioner af dette korpus.

| Periode | Navn | Antal ord |
|---|---|---|
| 1930-1984 | T-1930-84 | 3,8 mio. |
| 1985-1989 | T-1985-89 | 16,2 mio. |
| 1990-1994 | T-1990-94 | 23,5 mio. |
| 1995 | S-1995 | 51,8 mio. |
| 1996 | S-1996 | 58,5 mio. |
| 1997 | S-1997 | 33,2 mio. |
| 1998 | S-1998 | 36,4 mio. |
| 1999 | S-1999 | 36,7 mio. |
| 2000-2004 | T-2000-04 | 44,1 mio. |
| 2005 | S-2005 | 46,2 mio. |
| 2006 | S-2006 | 45,9 mio. |
| 2007 | S-2007 | 53,9 mio. |
| 2008 | T-2008X | 110,6 mio. |
| 2009 | T-2009X | 104,0 mio. |
| 2010 | S-2010 | 42,6 mio. |
| 2011 | T-2011 | 42,7 mio. |
| 2012 | T-2012 | 44,0 mio. |
| 2013 | T-2013 | 42,6 mio. |
| 2014 | T-2014 | 44,0 mio. |
| 2015 | T-2015 | 41,2 mio. |
| 2016 | T-2016 | 51,3 mio. |
| 2017 | T-2017 | 70,0 mio. |
| 2018 | T-2018 | 26,6 mio. |
| 2019 | T-2019 | 33,2 mio. |

Tabel 21.1: -BAKSPEJLET's delkorpusser

Som det fremgår af oversigten, er der en del mere materiale for årgangene 2008, 2009 og noget mere for 2007 end der er i de fleste af de øvrige årgange. Det skyldes, at disse tre årgange (2007, 2008 og 2009) er udvidet med yderligere Infomedia-materiale.

Som det endvidere fremgår af oversigten, er der også et delkorpus, som dækker alt materiale fra 1930-1984. Der er dog kun sporadisk materiale fra før 1983, og det stammer primært fra fra SpORDhund-brugerforslag. Du kan læse mere om dette materiale i afsnit 21.5 på næste side.

I BAKSPEJLET er der desuden inkluderet en række forlagstekster fra Lindhardt & Ringhof 2006, Aschehoug 2002 og 2003, Borgen 2003, 2004 – 2007 og Gad 2005.

Endelig er WIKIPEDIA-korpusset, jf. kapitel 20, indeholdt i BAKSPEJLET. Det indeholder den danske Wikipedia, sådan som den så ud den 20.08.2017. Dette forklarer, at delkorpusset for 2017 er forholdsvist stort – det samme gælder korpusserne i årene op til.

BAKSPEJLET indeholder endvidere *Korpus 2000*-materiale fra før 1997 samt Den Danske Ordbogs Korpus, jf. afsnit 21.4.

Endelig indeholder BAKSPEJLET hele KORPUS-DK, dog uden det udvalg af Wikipedia-tekster, som ellers er en del af KORPUS-DK. Dette for at undgå, at der bliver for mange dubletter mellem dette materiale og det fulde WIKIPEDIA-korpus, som jo i forvejen allerede er med i BAKSPEJLET.

## 21.2 Dialektkorpusset DIAKO

Dette korpus er kun tilgængelig i Ømål- og DSL-udgaven af CoREST. Korpusset består af 200 transskriberede interviews med dialekttalende fra forskellige egne i Danmark. Korpusset har et omfang på i alt godt 1,3 millioner løbende ord og indgår i det redaktionelle arbejde på Ømålsordbogen. Visse typer søgninger og funktioner i CoREST vedrører specifikt DIAKO-korpusset og er kun tilgængelige, når dette korpus er valgt, herunder udskrift af fund til ordsedler. De særlige funktionaliteter for dette korpus beskrives i en særskilt dokumentation, som udarbejdes af Ømålsordbogen.

## 21.3 Tysk SDEWAC-korpus

Dette korpus er kun tilgængelig i DSL- og Research-udgaven af CoREST. SDEWAC er et ordklassetagget korpus på 770 millioner ords tekst bestående af uddrag fra tysksprogede websider. Det er et resultat af Web as Corpus-gruppens[2] arbejde. Det er med for at kunne teste CoREST med et korpus på et andet sprog end dansk.

## 21.4 DDO-CLASSIC

Dette korpus var hovedkorpusset under udarbejdelsen af den trykte version af Den Danske Ordbog. Korpusset er beskrevet i detaljer i Norling-Christensen og

[2] https://wacky.sslmit.unibo.it/doku.php

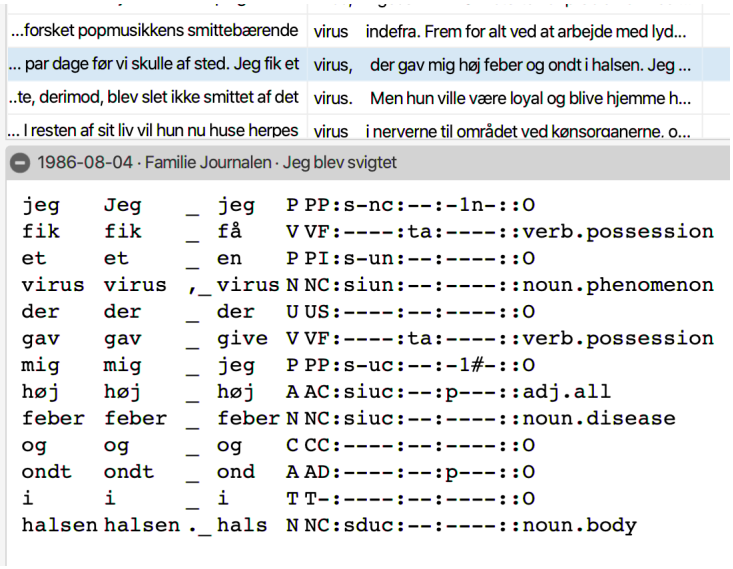
Asmussen (1998)[3]. CoREST-versionen er ePOS-tagget. De oprindelige metadata er afbildet på det sæt, som gælder for standardkorpusser i CoREST.

## 21.5 SPORDHUND

I DDO-udgaven af CoREST er der et særligt korpus over ordforslag fra SpORDhunde ved navn SPORDHUND. Under arbejdet med den trykte udgave af Den Danske Ordbog blev der indsamlet sproglige observationer af frivillige informanter, såkaldte SpORDhunde.

## 21.6 S-90

S-90 er en variant af KORPUS-2010 med forsøgsvis semantisk opmærkning med »super senses« i form af udvidede ePOS-tags. Den semantiske opmærkning er udført af Nicolai Hartvig Sørensen[4], jf. desuden Martínez Alonso et al. (2015). Figur 21.1 viser et eksempel på en sætning tagget med disse udvidede ePOS-tags, hvor »super sense«-delen er tilføjet efter et dobbelt kolon.

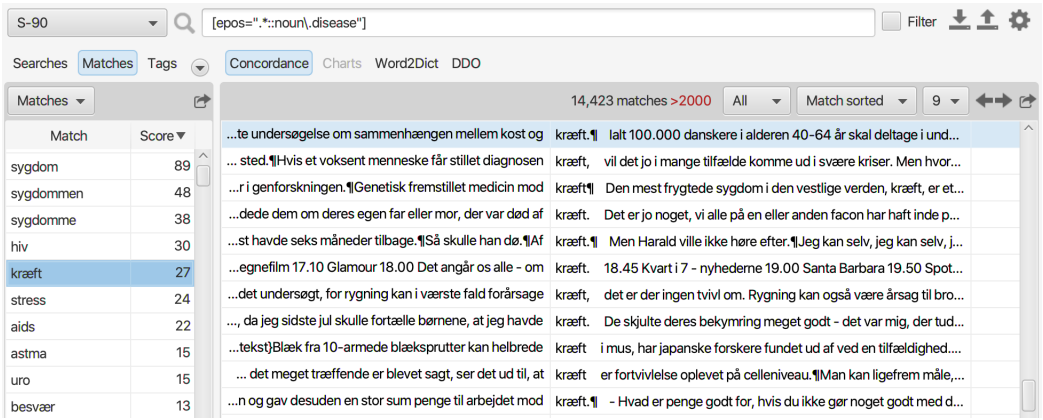



Figur 21.1: Eksempel på en sætning tagget med »super senses«

Substantiver, der betegner en sygdom, vil eksempelvis kunne findes ved hjælp af søgeudtrykket `epos=".*::noun\.disease"`. Figur 21.2 på den følgende side viser et udsnit af matchlisten og en del af konkordansen, du vil få frem efter en sådan søgning.

OBS! Den semantiske tagging er udelukkende af eksperimentel karakter og er behæftet med mange fejl.

[3] ja.dsl.dk/papers/Lexikos 1998/lexikos8.pdf
[4] Mere under dsl.dk/medarbejdere/nhs.




Figur 21.2: Udsnit af matchliste og konkordans med ord, der betegner sygdomme

# Appendiks

# Bilag A

# Løsningsforslag til opgaverne

## A.1 Øvelse: Simple søgeudtryk

1. *smartphone* – repræsenteret ved formerne *smartphone* og *smartphones*
2. Fordi *smartphone* åbenbart er en betydelig mere udbredt betegnelse end synonymet *smarttelefon*.
3. *mobiltelefon* og *biltelefon* (kun søgt på formen `.+telefon`)
4. andetled
5. *elektroniske*
6. *type* i *ny type medicin*
7. *høre græsset gro, høre blomsterne gro, høre det gro, høre ulden gro*

## A.2 Øvelse: Kontroltegn i simple søgeudtryk

1. Der er 3 forekomster, to er svenske salmetitler fra wikipedia-tekster, og en er fra en tekst om omsorg.
2. Eksemplet stammer fra en (ældre) forumtekst. Det er sandsynligt, at der her er tale om et indlån fra (amerikansk) engelsk, hvor stavningen af *über* med *u* i stedet for omlyd i denne type orddannelser er udbredt.
3. *übercool* med 7 forekomster
4. *Cirkusrevyen* (hhv. *Cirkusrevy*), *Cirkusbygningen*, *Cirkuskroen*, hvis du søger på `!Cirkus.+` – søger du derimod på `!Cirkus #` er det *Cirkus Arena*, *Cirkus Benneweis* og *Cirkus Tværs*.
5. *Det Kongelige Kapel* har flest forekomster, fulgt af *Det Sixtinske Kapel*.
6. *cirkus, cirkussens, cirkusser* og *cirkusset*

7. Fordi *cirkuset* er en uofficiel form og derfor ikke er blevet genkendt som en form af *cirkus* af lemmatiseringsalgoritmen. Det ville være kommet med i søgeudtrykket `%cirkus` – dog sammen med mange andre ord, som begynder med *cirkus*, og ikke er bøjningsformer af cirkus.

8. `@bytte` finder både alle bøjningsformerne af substantivet *(et) bytte* samt af verbet *(at) bytte*. Vil du begrænse resultaterne til en bestemt ordklasse, fx kun verberne, må du formulere et avanceret søgeudtryk, se kapitel 15.

9. Kigger man kun på matchlisten med forekomsten *lærer slår eleverne*, kan man forledes til at tro, at det er en lærer, der udfører handlingen at slå elever, altså at lærer er agens og eleverne patiens. Imidlertid forholder det sig omvendt, som konteksten i konkordansen afslører: *For en god lærer slår eleverne vel ikke på?*

10. *flere penge mellem hænderne(,) end* med 5 forekomster. Hvis søgningen skal tage hensyn til, om der er et komma til stede i matchet eller ej, må du formulere et avanceret søgeudtryk, se kapitel 15.

# Bilag B

# Korpussammensætninger

Læs om sammenlignelighed af korpusser i kapitel 20.

## B.1 Korpus 90

### B.1.1 Hovedkilder

Den følgende liste indeholder hovedkilderne for Korpus 90. En hovedkilde er defineret ved, at den står for mindst 1 promille af antallet af ord i korpus. Korpus 90 har 77 hovedkilder som tilsammen bidrager med 608 promille til det samlede korpus. Den følgende liste giver hovedkilderne samt hver deres promilleandel i Korpus 90. Listen er sorteret efter promilleandele i ikkestigende orden. Du kan begrænse dine søgninger til en (eller flere) af hovedkilderne ved at skrive dens navn(e) i filterfeltet »Source«, se kapitel 18.

```
Berlingske Tidende 184.0
B.T. 89.4
Morgenavisen Jyllands-Posten 25.2
Information 18.8
Weekendavisen 14.9
Fakta 14.1
Ude og Hjemme 13.9
Hjemmet 13.6
Familie Journalen 13.3
Se og Hør 13.2
Femina 11.6
Politiken 9.8
Alt for damerne 9.7
Ekstra Bladet 8.9
Radioavisen 7.4
Samvirke 7.4
Status 6.7
Bibelen. Det Gamle Testamente 6.0
Hendes Verden 6.0
Bibelen. Det Nye Testamente 5.5
Aarhuus Stiftstidende 4.9
Fyens Stiftstidende 4.9
Chili 4.6
Det fri Aktuelt 4.5
SKALK 4.4
PROXIMA 4.3
Månedsbladet PRESS 4.2
lokal-avisen Fredensborg-Humlebæk ... 4.1
```

```
BO BEDRE 4.0
Illustreret Videnskab 4.0
Midtjyllands Avis 3.6
Ny Dag 3.6
Billed Bladet 3.4
Cupido 3.2
Danskopgave 3.2
Rapport 3.2
Søndags-B.T. 3.2
Nyhedsmagasinet 3.0
Skive Folkeblad 2.9
natur og miljø 2.8
Helse 2.4
Jyllands-Posten 2.4
idé-nyt 2.3
Bådnyt 2.2
Ballerup-Måløv Avis 2.0
Motor 2.0
Børsens Nyhedsmagasin 1.9
Familie-Journalen 1.9
Albertslund Posten 1.7
Aalborg Stiftstidende 1.5
KOKOO 1.5
Blå Bog 1986 Rungsted Gymnasium 1.4
Frederiksborg Amts Avis 1.4
Kristeligt Dagblad 1.4
Aktuelt 1.3
Farum Nyt 1.3
Historieopgave 1.3
Lolland-Falsters Folketidende 1.3
NEONGUIDEN 1.3
Vestkysten 1.3
Børsen 1.2
Dansk litteraturhistorie. 8 1.2
Dansk Teaterhistorie. 1 1.2
det grønne område 1.2
Gladsaxe Bladet 1.2
Gør Det Selv 1.2
I FORM 1.2
Jydske Tidende 1.2
Vi Unge 1.2
Debatten Om Gensplejsning Af Ole Terney 1.1
Frederikssund Avis/Folkebladet ... 1.1
Danmarks historie 1.0
Gensplejsning Og Forskning 1.0
Hvidovre Avis 1.0
Jyllandsposten 1.0
Kosmologisk Information 1.0
Søborg Bladet 1.0
```

## B.2 Korpus 2000

### B.2.1 Hovedkilder

Den følgende liste indeholder hovedkilderne for Korpus 2000. En hovedkilde er defineret ved, at den står for mindst 1 promille af antallet af ord i korpus. Korpus 2000 har 45 hovedkilder som tilsammen bidrager med 969 promille til det samlede korpus. Tallene viser, at Korpus 2000 er betydelig mindre varieret sammensat end Korpus 90. Den følgende liste giver hovedkilderne samt hver deres promilleandel i Korpus 90. Listen er sorteret efter promilleandele i ikkestigende orden. Du kan begrænse dine søgninger til en (eller flere) af hovedkilderne ved at

skrive dens navn(e) i filterfeltet "Source", se kapitel 18.

```
Jyllands-Posten 345.9
Politiken 303.8
Dagbladet Aktuelt 79.3
www.folketinget.dk 38.0
empty 34.7
fyldepennen.dk 34.7
Tekster fra Forlaget Carlsen 24.2
Magisterbladet 10.6
Nyhedsmagasinet Danske Kommuner 9.7
Tekster fra leverandør-id 13001 7.3
Ahorn Science Univers 6.8
Nyt Aspekt 6.0
Familie Journalen 5.8
Krop og Fysik (Fysioterapeuternes blad) 4.5
KvaN 4.0
Opgaver fra Vordingborg Gymnasium 3.5
Fra hjemmeside: Malene Thyssen 3.4
Main Magazine 3.3
Ung Bladet 2.9
Psykiatri - sygepleje og sygdomslære 2.4
De forhenværende 2.2
Prædikener fra provst Mogens Jensen 2.2
Tidens Kvinder 2.1
Connery.dk 2.0
Områdeavisen Nordfyn 2.0
Sydvest Folkeblad 1.9
Ansvar og værdier, Ansvar 2000 (udg. i forlaget Centrum) 1.8
Fra hjemmeside: Bjergsnæs Efterskole 1.8
Forældre og Børn 1.7
Tekster fra ForfatterNet (litteratursiden.dk) 1.6
Fra hjemmeside: Svend Faarvangs rejsedagbog 1.5
HK-magasinet Delta 1.5
www.information.dk 1.5
Bevidsthedens niveauer 1.4
Miljøsk 1.4
Smagsprøver fra www.cicero.dk 1.4
Historieopgaver fra gymnasium eller HF (Skanderborg?) 1.3
Ingeniøren 1.3
NewsLetter, YFU-Danmark 1.3
Ordet rundt - Bibelen ved årtusindskiftet 1.3
Dansk Smede-Tidende 1.1
Rummeteren 1.1
Tekster fra DR Multimedie 1.1
Tekster fra Spurvelundskolen, Odense 1.1
Natur og Miljø 1.0
```

## B.3 Korpus 2010

### B.3.1 Hovedkilder

Den følgende liste indeholder hovedkilderne i Korpus 2010. En hovedkilde er defineret ved, at den står for mindst 1 promille af antallet af ord i korpus. Korpus 2010 har 29 hovedkilder som tilsammen bidrager med 997 promille til det samlede korpus. Tallene viser, at Korpus 2010 er endnu mindre varieret sammensat end Korpus 2000. Den følgende liste giver hovedkilderne samt hver deres promilleandel i Korpus 2010. Listen er sorteret efter promilleandele i ikkestigende orden. Du kan begrænse dine søgninger til en (eller flere) af hovedkilderne ved at skrive dens navn(e) i filterfeltet "Source", se kapitel 18.

```
Politiken 164.1
Fyens Stiftstidende 151.7
Jyllands-Posten 133.4
Berlingske Tidende 128.3
Weekendavisen 101.0
Ekstra Bladet 74.8
Information 60.7
Flensborg Avis 50.0
JP Århus 31.4
Se og Hør 17.9
Penge.dk 16.9
Erhvervsbladet 9.9
djøfbladet 8.2
Folkeskolen 8.2
Magasinet Ejendom 4.5
Juristen 4.2
Universitetsavisen 4.0
Hus Forbi 3.5
Højskolebladet 3.2
Samfundsøkonomen 3.2
Helse 2.7
Kommunalbladet 2.5
Ud og Se 2.2
Ældre Sagen NU 2.2
Bentes blog. Strøtanker fra en skribents skrivebord 1.9
Pernille - Øjeblikke fra hverdagen 1.9
Samvirke 1.9
Blogsbjerg. Lidt bedre end andre gode koste 1.8
Dansk historie 1.2
```

# Figurer

# Tabeller

# Litteratur


Andersen, M. S., Asmussen, H. og Asmussen, J. (2002). The project of Korpus 2000 Going Public. I Braasch, A. og Povlsen, C., redaktører, *Proceedings of the 10th EURALEX International Congress*, bind 1, side 291-299, Copenhagen. Euralex. 111

Asmussen, J. (2001). Korpus 2000. Et overblik over projektets baggrund, fremgangsmåder og perspektiver. *NYS*, 30. 111

Asmussen, J. (2004). Korpus 2000 – til hvilken nytte? Muligheder og grænser for empiriske sprogundersøgelser. I Duncker, D., redaktør, *Studier i Nordisk 2002-2003*, København. Selskab for Nordisk Filologi. 114

Asmussen, J. (2006). At måle og veje korpusser – et aspekt af arbejdet bag de store almensproglige korpusser for dansk. I Braasch, A., Navaretta, C., Nimb, S., Olsen, S., Paggio, P., Pedersen, B. S. og Wedekind, J., redaktører, *Sprogteknologi i dansk perspektiv*, side 157-177. C.A. Reitzels Forlag, København. 114

Asmussen, J. (2013). Survey of POS taggers. Rapport, DK-CLARIN, korpus.dsl.dk/clarin/corpus-doc/pos-survey.pdf. 83

Asmussen, J. (2015). Design of the ePOS tagger. Rapport, DK-CLARIN, korpus.dsl.dk/clarin/corpus-doc/pos-design.pdf. 83

Asmussen, J. og Halskov, J. (2009). Compiling and annotating corpora in DK-CLARIN. Interpreting and tweaking TEI P5. I Michaela Mahlberg, Victorina Gonzáles-Díaz, C. S., redaktør, *Proceedings of the Corpus Linguistics Conference CL2009*, nr. Article 287 (13 pages). University of Liverpool, UK. 111

DDO (2003–2005). *Den Danske Ordbog 1–6*. DSL & Gyldendal, København. 110

Martínez Alonso, H., Johannsen, A., Olsen, S., Nimb, S., Hartvig Sørensen, N., Braasch, A., Søgaard, A. og Sandford Pedersen, B. (2015). Supersense tagging for Danish. I *Proceedings of the 20th Nordic Conference of Computational Linguistics (NODALIDA 2015)*, side 21-29, Vilnius, Lithuania. Linköping University Electronic Press, Sweden. 118

Norling-Christensen, O. og Asmussen, J. (1998). The Corpus of The Danish Dictionary. *Lexikos. Afrilex Series*, 8:223-242. 110, 117