Tema 12: Olika somaliska korpusar

Just nu finns minst fyra somaliska korpusar med mellan 3 och 79 miljoner token. Tre är öppna för allmänheten, men en av dem kräver att man registrerar sig och loggar in med lösenord. En är inte tillgänglig för allmänheten.

Bangiga Af Soomaaliga / Somaliska Korp

Vid Språkbanken på Göteborgs universitet finns den svenska national-korpusen **Korp**. Den innehåller ca 13 miljarder token på svenska, men även en del mindre samlingar av texter på andra språk.

Sedan 28 oktober 2015 finns en avdelning med somaliska texter i Korp. För närvarande uppgår de somaliska samlingarna till ca 19 miljoner token: https://spraakbanken.gu.se/korp/?mode=somali

Innehållet i somaliska Korp är till största delen nyhetstexter, men även skolböcker från Mogadishu, Hargeysa och Jigjiga utgör en viktig del.

Kaydka Af Soomaaliga / Somali Corpus

Vid universitetet i Neapel (på italienska: Napoli) har fil.dr. Jama Muse Jama byggt upp **Somali Corpus** under professor Giorgio Bantis handledning. Korpusen, som lanserads på webben i juni 2016, innehåller cirka 3 miljoner ord och är placerad vid Redsea Cultural Foundation i Hargeysa: http://www.somalicorpus.com/

För att få tillgång till **Somali Corpus** måste man registrera sig som användare genom att klicka på *Furo xubinnimo cusub*.

Somali Corpus innehåller ganska mycket skönlitteratur och poesi, men även en del nyheter liksom politiska och andra samhälleliga texter.

HaBiT / Somali Web Corpus 2016

Universiteten i Oslo (Norge), Brno (Tjeckien) och Addis Abeba (Etiopien) har skapat stora korpusar för fyra etiopiska språk. Korpusarna innehåller uteslutande texter från internet: amhariska (30 milj. token), oromo (5 milj. token), tigrinja (2,5 milj. token), men deras största korpus är den somaliska som omfattar ca 79 miljoner token:

https://corpora.fi.muni.cz/habit/run.cgi/first_form?corpname=sowac16;align=

Texterna har samlats in helt automatiskt av en dator som programmerats att söka efter texter där ett litet antal av somaliskans allra vanligaste ord förekommer i samma text.

An Crúbadán

Även projektet An Crúbadán vid universitetet i Saint Louis (USA) har skapat en somalisk korpus om ca 25 miljoner token baserad på texter från internet. Deras korpus är dock inte tillgänglig för allmänheten. Bara en frekvenslista finns att ladda ner på deras hemsida:

http://crubadan.org/languages/so

Jämförelse av de olika korpusarnas storlek:

Svenska Korp:	13 260 000 000 token	Göteborg
HaBiT Somali WaC:	79 741 231 token	Brno, Tjeckien
An Crúbadán:	24 648 653 ord	Saint Louis, USA
Somaliska Korp:	19 300 000 token	Göteborg
Somali Corpus :	3 002 198 ord	Hargeysa

Internet

Jämfört med all den text på somaliska som finns i alla böcker, tidningar och på hela internet är naturligtvis samtliga dessa korpusar ganska små. Många viktiga ord förekommer över huvud taget inte i korpusarna. För att hitta exempel på hur mindre vanliga ord används måste man ofta göra sökningar på internet med hjälp av någon sökmotor.

Olika sökmotorer fungerar på väldigt olika sätt. De söker egentligen inte ut på internet, utan man har samlat in sidor från internet som man sedan har processat och indexerat på jättestora datorer. Antalet sidor som man har indexerat på det här sättet varier stort och därför får man också väldigt olika antal träffar på med olika sökmotorer.

Google har kanske det största antalet indexerade sidor och ger därför flest träffar.

https://www.google.com/

Bing söker på lite färre sidor, men har å andra sidan en del egenskaper som gör att sökträffarna blir mera precisa.

https://www.bing.com/

Den ryska sökmotorn Yandex ger ännu färre träffar, men här finns ännu större möjligheter att göra exakta sökningar.

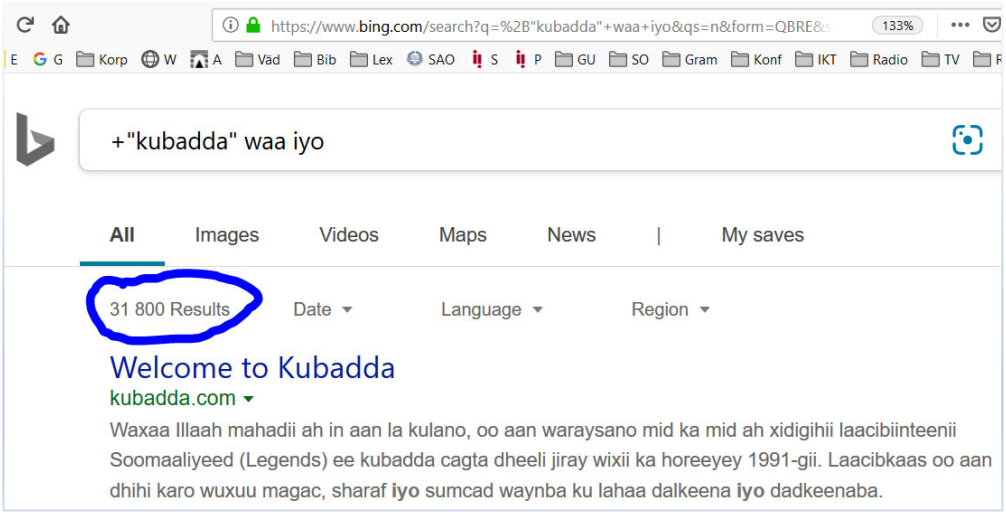
https://yandex.com/

Om man söker på en felaktig stavning med Google så är programmet gärna hjälpsamt och föreslår även träffar med den vedertagna stavningen. I Bing däremot kan man lättare styra mera över sökningarna. Om man i Bing skriver sökordet inom citationstecken och lägger till ett plustecken framför alltsammans, så ska man bara få träffar på just det man skrivit in i sökrutan. Pröva i Bing med t.ex.

+”kubada”
+”kubbada”
+”kubadda”
+”kubbadda”

Ibland kan det hända att samma ord finns på andra språk, och i så fall kan antalet träffar bli väldigt felaktigt. För att bara få träffar på somaliska sidor kan man efter sökordet inom citationstecken lägga till ett par av de allra vanligaste somaliska orden. Då msåte alla de angivna orden finnas någonstans på en och samma sida, t.ex.

+”kubada” waa iyo
eller +”kubada” soo iyo

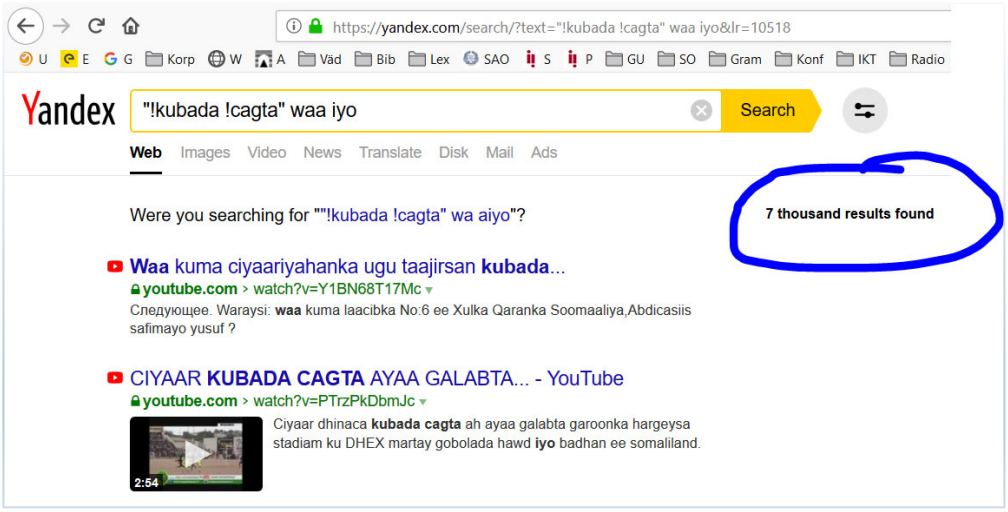


De vanliga småorden som man lägger till för att hitta rätt språk brukar kallas för **ANKARORD**. Några vanliga, användbara somaliska ord är **iyo, soo, waa, ayaa, aan, waxa**…

Om man skriver in en sådan sökning i Bing så är sannolikheten väldigt liten att man ska få exempel från andra språk än just somaliskan.

I Yandex gör man en exakt sökning genom att sätta orden man söker efter inom citattecken och sätta ett utropstecken framför varje ord i sökningen, t.ex.

”!kubada !cagta” waa iyo

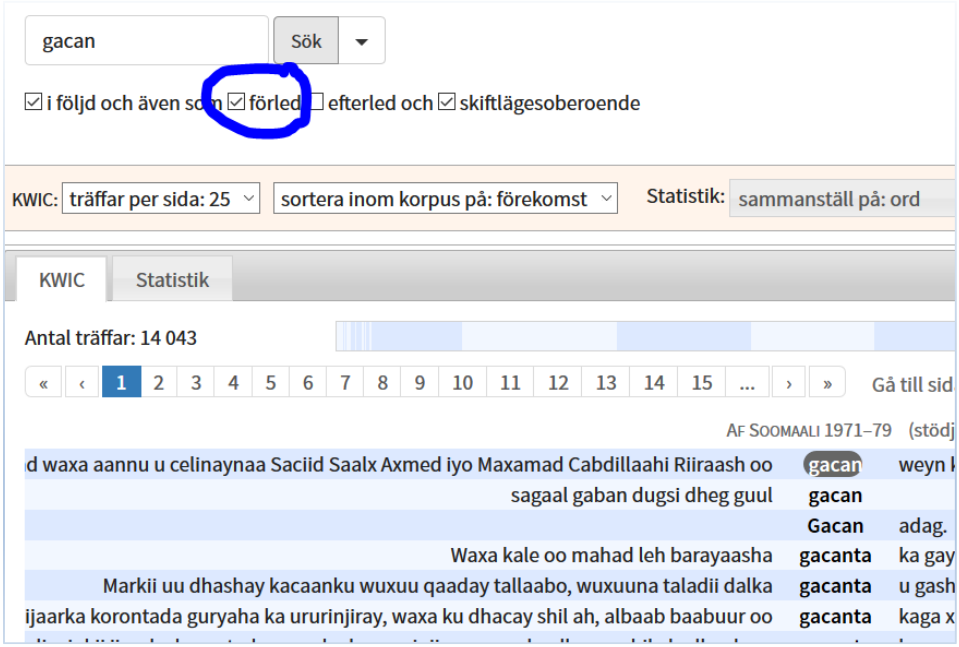


Sökinställningar

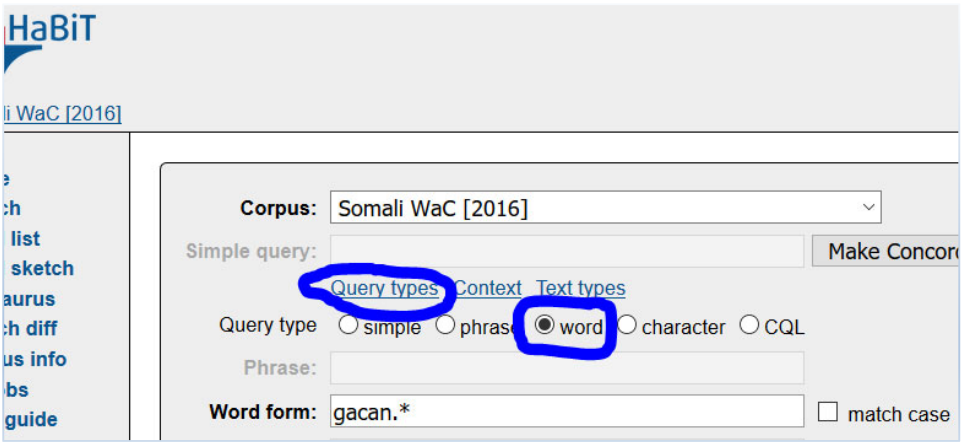
Även när man söker i korpusar kan man göra en rad olika val för att mera exakt kunna söka efter det man är intresserad av.

Om man vill hitta alla ord som börjar eller slutar på ett visst sätt kan man ange bara början eller bara slutet av dessa ord i sökrutan. I Korp markerar man sedan i de små kryssrutorna om det man skrivit i sökrutan ska vara *förled* (=början) eller *efterled* (=slutet) på de ord som man vill hitta.

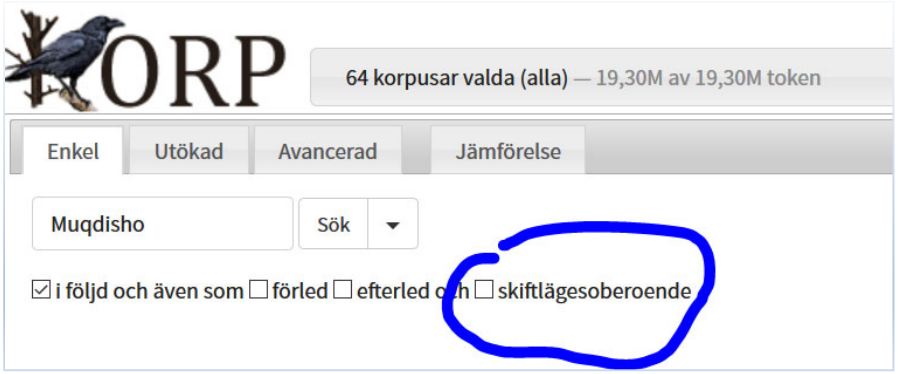
Om man skriver in **gacan** i sökfältet och anger att detta ska vara förled i sökningen så får man alltså även träffar som **gacanta, gacantii, gacantiisa, gacanka, gacanku…**



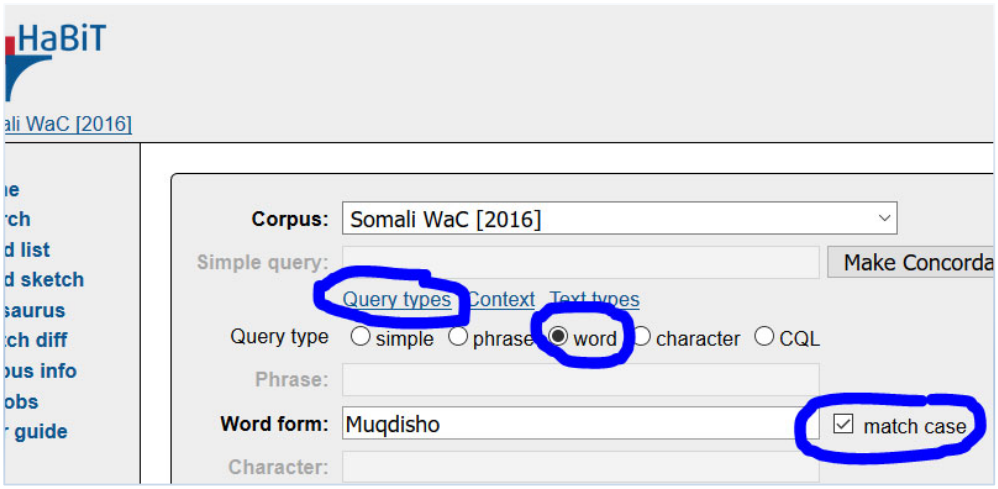
För att söka på samma sätt i HaBiT måste du klicka på *Query types*, sedan markera *word* och skriva början av ordet följt av **.*** (punkt+stjärna).



Ett annat viktigt val är om man vill ha träffar skrivna med både stora och små bokstäver, eller träffar bara på exakt det man skrivit. Om man söker på **Muqdisho** och med en bock framför *skiftlägesoberoende* så får man även träffar som t.ex. **muqdisho** och **MUQDISHO**. Detta är den förvalda inställningen i Korp. Om man bara vill få träffar på **Muqdisho** måste man ta bort markeringen i rutan *skiftlägesoberoende*.



För att på samma sätt göra en exakt sökning i HaBiT måste du klicka på ***Query types***, sedan markera ***word*** och till sist sätta en bock framför ***match case***.



Detta är bara två av de många inställningar som är möjliga i de flesta korpusverktyg idag. Nedan kommer vi att ta en närmare titt på just de inställningar som är möjliga i Korp.

Information om träffarna

Resultatet av en sökning presenteras i första hand som det totala antalet träffar i korpusen samt en lång lista med textexempel från alla de ställen i korpusen där det sökta ordet förekommer. Ordet presenteras som regel med en kontext som är maximalt en mening lång. Dessutom anges det i form av olika rubriker i vilken delkorpus som exemplet förekommer.

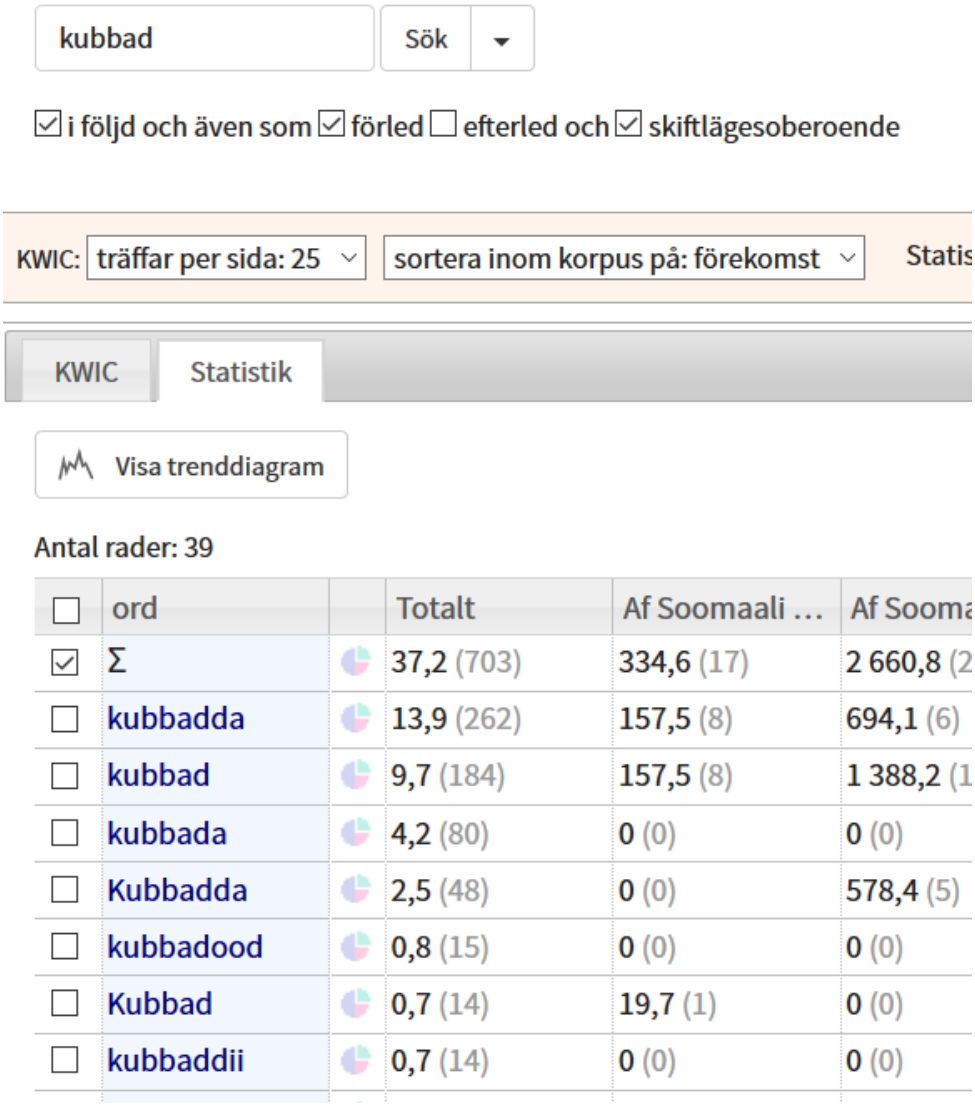
Om man i Korp klickar på ett av exemplen får man i högerkanten mera information om just det exemplet. Oftast får man veta delkorpusens namn, textens titel, författarens namn, publiceringsår och ibland även vilken sida exemplet finns på. Om texten finns tillgänglig på internet finns det en klickbar länk till texten i fråga. Eftersom länkar ibland förändras kan en del länkar ha slutat fungera, men i så fall hittar man oftast texten genom att googla på några ord i exempelraden.

I HaBiT anges alltid webbplatsen där exemplet är hämtat i vänsterkanten av varje rad. Om man klickar på en rad öppnas exemplet i en ny ruta med lite mera kontext, vilket kan underlätta för att förstå exemplet ordentligt.

Statistik

Om man har gjort en mera komplex sökning, t.ex. genom att bara skriva in början av det man vill söka på, då är det väldigt nyttigt att titta på den statistik som finns om sökträffarna.

I Korp finns en flik som heter just ***Statistik***. Under den fliken hittar man en lista över alla de olika former och stavningar som förekommer i resultatet av den sökning som man har gjort, i det här fallet **kubbad**. tillsammans med markeringen *förled*.



Siffrorna i kolumnen Total är av två slag. Den första siffran anger den RELATIVA frekvens *per miljon ord* (**pmw**, *per million words*), medan den andra siffran anger den ABSOLUTA frekvensen i korpusen. I det här fallet förekommer olika former som alla börjar med **kubbad** 703 gånger i hela korpusen. Eftersom korpusen innehåller 18,87 miljoner token blir den relativa frekvensen 703 / 18,87 = 37,2 pmw. Den siffran är användbar om man vill jämföra reslutat i Krop med resultat från andra korpusar.

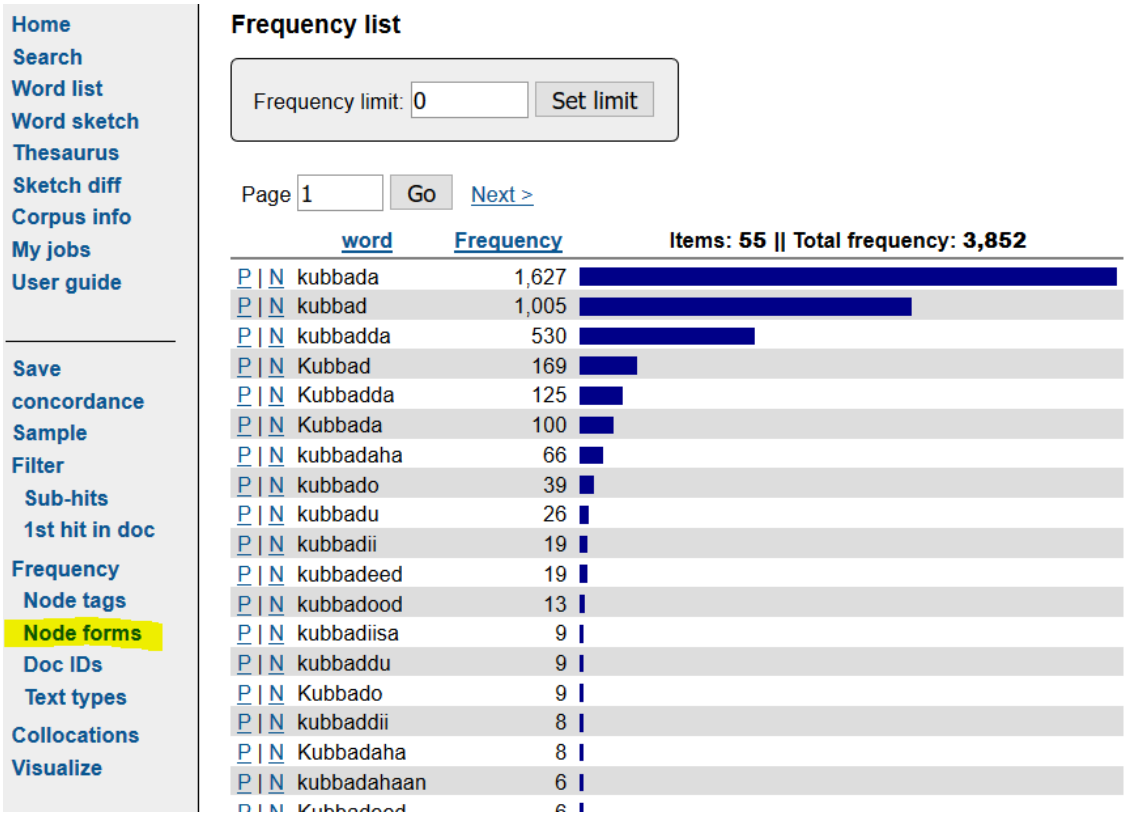
För att få fram jämförbara siffror i HaBiT måste man alltså skriva in söksträngen **kubbad.*** där tecknen **.*** anger att vilka bokstäver som helst eller inga bokstäver alls kan förekomma.

Om man går till https://corpora.fi.muni.cz/habit/run.cgi/first?corpname=sowac16 och skriver sökströngen **kubbad.*** i sökfältet så får man 3 852 träffar eller 48,31 pwm (per million words). Denna siffra räknas ut genom att ta antalet träffar delas med korpusens storlek (79,74 miljoner), alltså

$$3852 / 79{,}74 = 48{,}31$$

Nu kan man jämföra 37,2 wpm i Korp med 48,3 wpm i WaC och konstatera att olika former av kubbad är ganska mycket vanligare i WaC än i Korp. Sannolikt beror det på vilken typ av innehåll som finns i texterna i de båda korpusarna. I HaBiT finns det kanske många sidor om sport från internet.

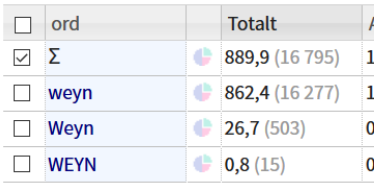
Om man sedan vill se frekvensen för varje form av **kubbad.*** i HaBiT så klickar man på *Node Forms* i vänstermenyn.



Övningar: Tema 12

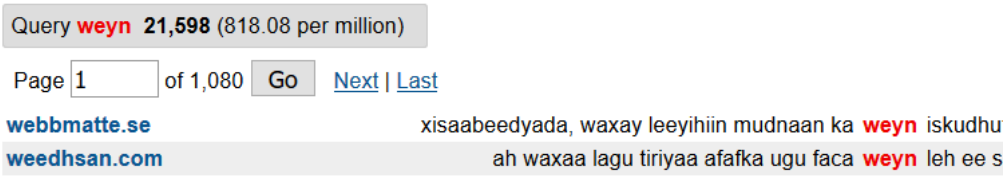
1. Gör en enkel sökning på följande ord och former utan att ändra några inställningar. Anteckna antalet förekomster (träffar) i den första tabellen. Anteckna sedan antalet förekomster per miljon ord (pmw) i den andra tabellen.

För *Somaliska Korp* hittar du båda siffrorna om du klickar på fliken *Statistik*. Då anges förekomster per miljoner ord för alla stavningar (med stor och liten bokstav) direkt efter tecknet **Σ**, och sedan anges det absoluta antalet inom parentes.



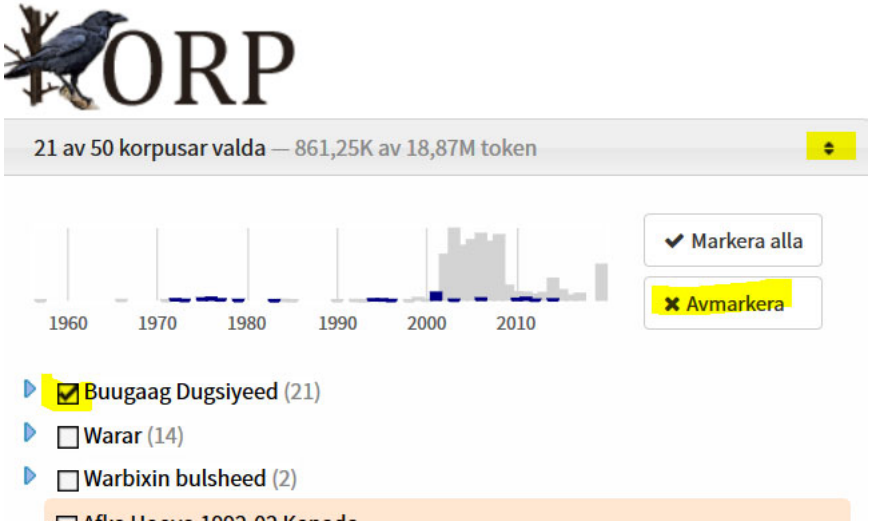
ord	Totalt	
Σ	889,9 (16 795)	1
weyn	862,4 (16 277)	1
Weyn	26,7 (503)	0
WEYN	0,8 (15)	0

För *Somali Web Corpus* hittar du båda siffrorna ovanför listan med exempel: först det absoluta antalet, sedan antalet per miljon ord inom parentes.



I *Kaydka Af Soomaaliga* kan man tyvärr inte få fram antalet förekomster för alla ord och former, så denna korpus kan inte inkluderas i den här övningen.

Gör till sist om sökningen i Korp, men välj att bara söka i skolböckerna. Det gör du genom att först klicka på de två små pilarna där det står *50 av 50 kurpusar valda*. Klicka sedan på *Avmarkera* och sätt sedan en bock framför *Buugaag Dugsiyeed*.



ABSOLUT ANTAL	Korp	Web Corpus	*Buugaag Dugsiyeed*
weyn	16 795	21 598	
wayn			
kubbad			
kubad			
kubbadda			
kubbada			
kubadda			
kubada			
PMW	Korp	Web Corpus	*Buugaag Dugsiyeed*
weyn	889,9	818,08	
wayn			
kubbad			
kubad			
kubbadda			
kubbada			
kubadda			
kubada			

Räkna till sist ut den procentuella fördelningen mellan de olika stavningarna av samma form.

Räkna först samman det totala antalet förekomster, t.ex.
weyn/wayn i Korp: 16795 + 2478 = 19273.

Dela sedan antalet för vardera stavningen med det totala antalet och multiplicera med 100:
weyn: 16795 / 19273 x 100 = 87%
wayn: 2478 / 19273 x 100 = 13%

Tänk på att avrunda på ett korrekt sätt!

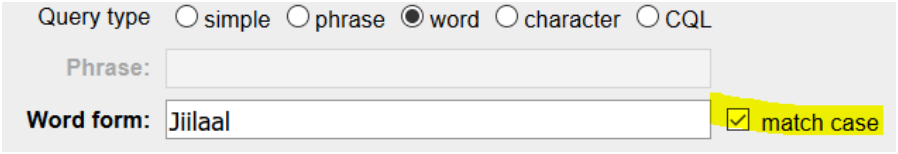
%	Korp	Web Corpus	*Buugaag Dugsiyeed*
weyn	87%		
wayn	13%		
kubbad			
kubad			
kubbadda			
kubbada			
kubadda			
kubada			

Hur skiljer sig de olika korpusarna åt?

2. Gör nu en undersökning av stavningen med stor och liten begynnelse-bokstav i följande ord.

I Korp hittar du siffrorna för olika stavningar under fliken *Statistik*.

I Somali Web Corpus hittar du siffrorna enklast siffror om olika stavningar genom att göra en sökning på varje stavning och markera *match case*.



PMW	Korp	Web Corpus	*Buugaag Dugsiyeed*
Jiilaal			
jiilaal			
Sabti			
sabti			
Soomaali			
soomaali			
Diseembar			
diseembar			
%	Korp	Web Corpus	*Buugaag Dugsiyeed*
Jiilaal			
jiilaal			
Sabti			
sabti			
Soomaali			
soomaali			
Diseembar			
diseembar			

Lösningsförslag till Tema 12

Eftersom Korp har vuxit lite under det senaste året så har siffror blivit lite större än de siffror som redovisas här nedan…

ABSOLUT ANTAL	Korp	Web Corpus	*Buugaag Dugsiyeed*
weyn	16 795	21 598	911
wayn	2 478	19 325	51
kubbad	198	1 174	84
kubad	415	2 738	8
kubbadda	310	659	150
kubbada	94	1 727	21
kubadda	2 287	13 089	7
kubada	437	8 097	10
PMW	Korp	Web Corpus	*Buugaag Dugsiyeed*
weyn	889,9	818,08	1057,8
wayn	131,3	242,35	59,2
kubbad	10,5	14,72	97,5
kubad	22,0	34,34	9,3
kubbadda	16,4	8,26	174,2
kubbada	5,0	21,66	24,4
kubadda	121,2	164,14	8,1
kubada	23,2	101,54	11,6

%	Korp	Web Corpus	*Buugaag Dugsiyeed*
weyn	87%	53%	95%
wayn	13%	47%	5%
kubbad	32%	30%	91%
kubad	68%	70%	9%
kubbadda	10%	3%	80%
kubbada	3%	7%	11%
kubadda	73%	56%	4%
kubada	14%	34%	5%

I Somali Web Corpus är weyn och wayn ungefär lika vanliga, men i Korp dominerar weyn tydligt, i synnerhet i skolböckerna. Detta beror troligen på att Korp innehåller mycket böcker och tidningar som granskats av personer som är vana att skriva.

När det gäller kubbad är roten med två bb klart vanligast i skolböckerna, men inte i någon av korpusarna som helhet.

I bestämd artikel dominerar dock skrivningen med två dd i båda korpusarna och formen med bb och dd dominerar stort i skolböckerna.

2.

PMW	Korp	Web Corpus	*Buugaag Dugsiyeed*
Jiilaal	4,7	0,80	16,3
jiilaal	3,1	1,42	10,4
Sabti	6,6	14,65	59,2
sabti	3,8	4,73	7,0
Soomaali	129,3	246,95	113,8
soomaali	15,9	40,97	8,1
Diseembar	3,2	1,47	5,8
diseembar	0,7	0,13	0
%	Korp	Web Corpus	*Buugaag Dugsiyeed*
Jiilaal	60%	36%	61%
jiilaal	40%	64%	39%
Sabti	63%	76%	89%
sabti	37%	24%	11%
Soomaali	89%	86%	93%
soomaali	11%	14%	7%
Diseembar	82%	92%	100%
diseembar	18%	8%	0%