Greppe eller søke i korpus:

grep -r 'detjegvilfinne' filnavn.txt
grep -rl 'detjegvilfinne' filnavn.txt (vis bare filnavnene)
grep **-rl** (bruk den også hvis den spør om directory):
grep -rl 'detjegvilfinne' corpus-fkv/converted/ (søk i hele converted-mappa)

Kjør en tekst og få ut ord som mangler

cat tekst.txt | preprocess| ufkv | cut -f2|cut -d"+" -f1|uniq | sort|uniq -c|sort -nr|cut -c6-|fkvnob | grep "+?" | less

vs. gammel kommando for egen fil:

cat text.txt | preprocess | ufkv | grep '+?' |cut -f1 | sort | uniq -c | sort -nr | less

cat tekst.txt | preprocess| ufkv | cut -f2|cut -d"+" -f1|uniq | sort|uniq -c|sort -nr|cut -c6-|fkvnob | grep "+?" | less **(näyttää sanat ja ufkv:n kysymysmerkit, sanojen aakkosjärjestyksessä)**

cat text.txt | preprocess | ufkv | grep '+?' |cut -f1 | sort | uniq -c | sort -nr | less **(tämä näyttää vain sanat, frekventtiyksen järjestyksessä)**

Kommandoforklaring:

```
cat ~/Downloads/mandat.txt| # ota teksti
preprocess|          # yksi sana per rivi
unob|                # norjan analysaattori
cut -f2|             # ota lemma + analyysi
cut -f1,2            # ota ensimmäinen ja toinen
cut -d"+" -f1|       # heitä analyysi pois
uniq|                # lyö tuplamuodot yhteen
sort|                # laita aakkosjärjestykseen
uniq -c|             # laske muotoja
sort -nr|            # järjestää määrän mukaan
cut -c6-|            # ota numerot pois
nobfkv|              # käännä kveeniksi
grep "?"|            # poimi tuntemattomat
cut -f1|             # muodot vain kerran
grep '[a-z]'|        # vain ne, joilla on vähintään yksi kirjain
tr '\n' ','|         # lista muutetaan yhdeksi riviksi
sed 's/,/, /g;'|     # lisätään väli pilkun jälkeen
see                  # ja heitetään subethaeditiin
```

Missinglists

Missinglists:
• *Vanha preprosessointi*
ccat -l fkv converted/fkv/|preprocess|ufkv|grep '+?'|cut -f1|sort|uniq -c | sort -nr > fkv.missing.191021
• *Uusi preprosessointi:*
ccat -l fkv converted/fkv/|hfst-tokenize -cg ~/langs/lang-fkv/tools/tokenisers/tokeniser-disamb-gt-desc.pmhfst|grep ' ?'|cut -d'"' -f2|sort|uniq -c |sort -nr | less **(funker 12.02.21)**

- ccat -l fkv converted/fkv/|preprocess|ufkv|grep '+?'|cut -f1|sort|uniq -c | sort -nr | less (i freecorpus)
- cat kvensktekst.txt|preprocess|ufkv|grep '+?'|cut -f1|sort|uniq -c | sort -nr | less (i egen fil)
- ccat -l fkv converted/fkv/|grep 'miitä'| less (finn kontekst)

fkv:ssa:
./configure --with-hfst --enable-tokenisers --enable-reversed-intersect --enable-dialects
make -j

Gammelmåte
- cd freecorpus
- ccat -l fkv converted/fkv/|preprocess|ufkv|grep '+?'|cut -f1|sort|uniq -c | sort -nr > fkv.missing.191021
- see fkv.missing.191021

Nymåte
- ccat -l fkv converted/fkv/|hfst-tokenize -cg $GTHOME/langs/fkv/tools/tokenisers/tokeniser-disamb-gt-desc.pmhfst|grep ' ?'|cut -d'"' -f2|sort|uniq -c |sort -nr > fkv.missing.191021
- see fkv.missing.191021

Tokenize må være aktivert via confi:
./configure --with-hfst --enable-tokenisers --enable-reversed-intersect --enable-dialects

```
————
lang-fkv$ hufkv
huonet
huonet   huonet+N+Sg+Nom  0,000000


lang-fkv$ hdfkv
huonet+N+Sg+Nom
huonet+N+Sg+Nom  huone    0,000000
huonet+N+Sg+Nom  huonet   0,000000
————

lang-fkv$ nobfkv
bil
bil piili    0,000000
```

Generere/oppdatere nobfkv og fkvnob:

cd langs/giella-core/dicts/
rm nobfkv/bin/*.lexc
rm fkvnob/bin/*.lexc
make -f make-bildict SLANG=nob TLANG=fkv TNUM=all
make -f make-bildict SLANG=fkv TLANG=nob TNUM=all

Hvor mange ord er det i ordboka?

tell alle ord for hver ordbok, antall ord: **cat src/*.xml|grep '<l '|wc -l**

Historikk:

cat fkvnob/src/*.xml|grep '<l '|wc -l
    9139
cat nobfkv/src/*.xml|grep '<l '|wc -l
    9032
^ 24. april 2020

Tuomas:dicts thomas$ cat fkvnob/src/*.xml|grep '<l '|wc -l
    9651
Tuomas:dicts thomas$ cat nobfkv/src/*.xml|grep '<l '|wc -l
    9577
^ 23. des 2020

dicts thomas$ cat fkvnob/src/*.xml|grep '<l '|wc -l
   10189
dicts thomas$ cat nobfkv/src/*.xml|grep '<l '|wc -l
   10179
^ 17. mars 2021

thomasbk:dicts thomas$ cat fkvnob/src/*.xml|grep '<l '|wc -l
   10236
thomasbk:dicts thomas$ cat nobfkv/src/*.xml|grep '<l '|wc -l
   10231
^ 15. nov 2021

Tuomas:dicts thomas$ cat fkvnob/src/*.xml|grep '<l '|wc -l
   10273
Tuomas:dicts thomas$ cat nobfkv/src/*.xml|grep '<l '|wc -l
   10299
^31. jan 2022

thomasbk:dicts thomasbk$ cat fkvnob/src/*.xml|grep '<l '|wc -l
   10401
thomasbk:dicts thomasbk$ cat nobfkv/src/*.xml|grep '<l '|wc -l
   10397
^8. nov 2022

Tuomas:dicts thomas$ cat fkvnob/src/*.xml|grep '<l '|wc -l
   10745
Tuomas:dicts thomas$ cat nobfkv/src/*.xml|grep '<l '|wc -l
   10663
^29. nov 2022

thomasbk:dicts thomasbk$ cat fkvnob/src/*.xml|grep '<l '|wc -l
   12672
thomasbk:dicts thomasbk$ cat nobfkv/src/*.xml|grep '<l '|wc -l
   11851
^01. apr 2023

thomasbk:dicts thomasbk$ cat fkvnob/src/*.xml|grep '<l '|wc -l
   14941
thomasbk:dicts thomasbk$ cat nobfkv/src/*.xml|grep '<l '|wc -l

13389
^06. okt 2023

thomasbk:dicts thomasbk$ cat fkvnob/src/*.xml|grep '<l '|wc -l
15016
thomasbk:dicts thomasbk$ cat nobfkv/src/*.xml|grep '<l '|wc -l
13633
^11. okt 2023

thomasbk:dict-fkv-nob thomasbk$ cat src/*.xml|grep '<l '|wc -l
15220
thomasbk:dict-nob-fkv thomasbk$ cat src/*.xml|grep '<l '|wc -l
13821
^07. nov 2023

tromso-nat-guest-737:dict-fkv-nob thomasbk$ cat src/*.xml|grep '<l '|wc -l
15461
tromso-nat-guest-737:dict-nob-fkv thomasbk$ cat src/*.xml|grep '<l '|wc -l
14072
29/11-23

thomasbk:dict-fkv-nob thomasbk$ cat src/*.xml|grep '<l '|wc -l
15649
thomasbk:dict-nob-fkv thomasbk$ cat src/*.xml|grep '<l '|wc -l
14304
14/02-24

dict-fkv-nob$ cat src/*.xml|grep '<l '|wc -l
15661
dict-nob-fkv$ cat src/*.xml|grep '<l '|wc -l
14314
sommeren 2024

———————————————————

Hvor mange ord er det i korpus?

Historikk:
freecorpus thomas$ ccat -l fkv converted/fkv/|wc -w
368315 (10. jan 2021)

freecorpus thomas$ ccat -l fkv converted/fkv/|wc -w
371121 (11. jan 2021)

freecorpus thomas$ ccat -l fkv converted/fkv/|wc -w
403588 (17. mars 2021)

thomasbk:freecorpus thomas$ ccat -l fkv converted/fkv/|wc -w
406279 (12. mai 2021)

thomasbk:freecorpus thomas$ ccat -l fkv converted/fkv/|wc -w
412784 (03. juli 2021)

thomasbk:freecorpus thomas$ ccat -l fkv converted/fkv/|wc -w
423028 (01. sept 2021)

thomasbk:freecorpus thomas$ ccat -l fkv converted/fkv/|wc -w
427847 (19. okt 2021)

dhcp3398-stud:freecorpus thomas$ ccat -l fkv converted/fkv/|wc -w

431100 (11. nov 2021)

thomasbk:freecorpus thomas$ ccat -l fkv converted/fkv/|wc -w
433773 (31. jan 2022)

thomasbk:freecorpus thomas$ ccat -l fkv converted/fkv/|wc -w
444248 (16. mai 2022)

thomasbk:freecorpus thomasbk$ ccat -l fkv converted/fkv/|wc -w
444075 (8. nov 2022 reset maskin)

thomasbk:freecorpus thomasbk$ ccat -l fkv converted/fkv/|wc -w
450729 (10. mars 2023)

thomasbk:freecorpus thomasbk$ ccat -l fkv converted/fkv/|wc -w
461821 (11. mars 2023)

thomasbk:corpus-fkv thomasbk$ ccat -l fkv converted/|wc -w
472808 (6. okt 2023)

thomasbk:corpus-fkv thomasbk$ ccat -l fkv converted/|wc -w
508837 (5. des 2023)

thomasbk:corpus-fkv thomasbk$ ccat -l fkv converted/|wc -w
495527 (14. feb 2024)

thomasbk:corpus-fkv thomasbk$ ccat -l fkv converted/|wc -w
499242 (12. jun 2024)

Vanlig korpus pluss Børselv:

thomasbk:corpus-fkv thomasbk$ ccat -l fkv converted/|wc -w
499242 (12. jun 2024)
thomasbk:corpus-fkv-x-closed thomasbk$ ccat -l fkv converted/|wc -w
616573 (Børselv lagt til her: corpus-fkv-x-closed)
Yhtheensä: **1.115.815**

———————————————————-

Finn ord som finnes i en ordbok, men mangler i den andre

cat fkvnob/src/*.xml | grep '<t '|
cut -d"<" -f2|cut -d">" -f1
| preprocess
| nobfkv | grep '?'

Original:
cat fkvnob/src/*.xml | grep '<t '| cut -d"<" -f2|cut -d">" -f1 | preprocess | nobfkv | grep '?'

12.10.23:

cat fkvnob/src/V_fkvnob.xml | grep '<t '| cut -d"<" -f2|cut -d">" -f2 | tr " " "_"| nobfkv | grep '?' |l
cat fkvnob/src/N_fkvnob.xml | grep '<t '| cut -d"<" -f2|cut -d">" -f2| tr " " "_"| nobfkv|grep '?'|l

Script som automatisk lager xml-oppsettet til ordboka fra en tabliste:

cat Klart.txt | sed "s/ /N /" | perl ~/main/words/dicts/fkvnob/scripts/c2x.pl > testing.txt

- Lage reversert (erstatt TAB med mellomrom først)

cat test.txt | sed -e 's/\([^]*\) *\([^]*\)/\2 \1/' > reversert.txt

listemedord | fkvnob | grep " " | sed "s/ /N /" | perl fkvnob/scripts/c2x.pl | less

fkvnob | grep " " | sed "s/ /N /" | perl fkvnob/scripts/c2x.pl | less

———————————————————————————————————

cat nobfkv/src/N_nobfkv.xml | grep '<t '| less (de oversatte)
cat nobfkv/src/N_nobfkv.xml | grep '<l '| less (de norske)

hele:
cat nobfkv/src/N_nobfkv.xml | grep '<t '| cut -d"<" -f2|cut -d">" -f2 | preprocess | fkvnob | grep '?'|cut -f1 | sort | uniq -c | sort -nr | cut -c6- | hunob | cut -f2 | cut -d"+" -f1 | uniq | grep "[a-z]" | fkvnob | grep "?" | cut -f1 | fkvnob | grep " " | less

- **Kjøre ordliste gjennom ordboka**: cat filnavn.txt |cut -f1|fkvnob|grep "?"|cut -f2 | l
- **Reversere**: cat filnavn.txt | sed -e 's/\([^]*\) *\([^]*\)/\2 \1/' | l
- **Lage xml**: cat filnavn.txt | sed "s/ /N /" | perl ~/main/words/dicts/**nobfkv**/scripts/c2x.pl | l **(og bytt til fkvnob)**

Legge til ord fra sanakirja til lexc

fkvnob thomas$ cat src/N_fkvnob.xml |grep '<l' |cut -d">" -f2|cut -d"<" -f1|grep -v " " |ufkv|grep "?"|cut -f1,2|tr '\t' ':' | rev|cut -c3-|sort|rev|less

fkvnob thomas$ cat src/N_fkvnob.xml |grep '<l' |cut -d">" -f2|cut -d"<" -f1|grep -v " " |ufkv|grep "?"|cut -f1,2|tr '\t' ':' | rev|sort|rev|less **<--- denne er bra, og ganske klar til å legge til i lecx**

trond$ cat src/V_fkvnob.xml |grep '<l '|cut -d">" -f2|cut -d"<" -f1|grep -v ' '|ufkvNorm| grep "?"
dette er den gamle

=

xfst: cat ~thomas/main/words/dicts/fkvnob/src/N_fkvnob.xml |grep '<l' |cut -d">" -f2|cut -d"<" -f1|grep -v " " |ufkv|grep "?"|cut -f1,2|tr '\t' ':' | rev|sort|rev|less

hfst: cat ~thomas/main/words/dicts/fkvnob/src/N_fkvnob.xml |grep '<l' |cut -d">" -f2|cut -d"<" -f1|grep -v " " |hufkv|grep "?"|cut -f1,2|tr '\t' ':' | rev|cut -c3-|sort|rev|less

Tastatursnarveier

find next: cmd+g
replace and find next: ctrl+cmd+g

dubletter - doble, flere enn
uniq -d anna mulle jos on enemmæn kuin yksi d
uniq -c | sort -nr (enemmæn kuin kaks)

grep -A10 (vis 10 linjer etterpå)
grep -A10 haug nobfkv/src/*_nobfkv.xml

grep -v Prop (uten navn)

ctrl+K (fjern alt etter kursor)
ctrl+D (fjern enkeltbokstav etter kursor)

Delete previous character = Delete or Control+H
Delete next character = Forward Delete key, or FN+Delete or Control+D
Delete previous word = Option+Delete
Delete next word = Option+FN+Delete
Delete to start of line = Command+Delete
Delete remainder of paragraph = Control+K
Select back to the start of paragraph and delete = Shift+Option+Up then Delete, or Shift+Control+A then Delete
Select to the end of paragraph and delete = Shift+Option+Down then Delete, or Shift+Control+E then Delete
Cut = Command+X

Hvordan klone repositorie (sjekke ut til egen maskin)

To måter å gjøre det på:

git clone https://github.com/giellalt/dict-fkv-nob

og

git clone git@github.com:giellalt/dict-fkv-nob

gamle svn up = tre kommandoer for å gjøre akkurat det samme som gamle svn up:
git stash, **git pull —rebase**, og **git stash pop**

git add

git commit

angre commit (fjern den siste kommittede)
git reset --hard HEAD^1

git push

thomasbk:~ thomasbk$ cd dict-fkv-nob/
thomasbk:dict-fkv-nob thomasbk$ open -a tower .
thomasbk:dict-fkv-nob thomasbk$

Ved Pull-knappen: huk av på **Rebase instead of merge**

1. git clone https://github.com/giellalt/corpus-fkv-orig-x-closed
2. cd corpus-fkv-orig-x-closed
3. open -a tower .

Kjør disse hvis sjekket ut nytt språk:
1. ./autogen.sh
2. ./configure
3. make -j
4. make install (viktig for korpus)

Korpuskommandoer

kjør make install i fkv (ikke bare autogen, configure og make)

add_files_to_corpus -d orig/fkv/news/ruijankaiku --name ruijan-kaiku.no_laft-bak-loe.html https://www.ruijan-kaiku.no/laft-bak-loe-hirsinurkka-kajan-takana/

Added /Users/thomas/freecorpus/orig/fkv/news/ruijankaiku/ruijan-kaiku.no_laft-bak-loe.html

add_files_to_corpus -p /Users/thomas/freecorpus/orig/fkv/news/ruijankaiku/ruijan-kaiku.no_laft-bak-loe.html -l nob --name ruijan-kaiku.no_laft-bak-loe.html https://www.ruijan-kaiku.no/laft-bak-loe-hirsinurkka-kajan-takana/

freecorpus thomas$ convert2xml /Users/thomas/freecorpus/orig/**fkv**/news/ruijankaiku/ruijan-kaiku.no_laft-bak-loe.html
freecorpus thomas$ convert2xml /Users/thomas/freecorpus/orig/**nob**/news/ruijankaiku/ruijan-kaiku.no_laft-bak-loe.html

parallelize -l2 fkv converted/<den norske fila>
parallelize -l2 fkv converted/nob/news/ruijankaiku/ruijan-kaiku-25-ar.html.xml

(reparallelize converterer til xml og kjører parallelize, så hvis du allerede har kjørt convert2xml etter endringene, kan du bare kjøre parallelize på nytt i stedet for reparallelize):
reparallelize kjøres på .html-fila, og den finner tilbake til originalfilene og viser dem øverst (**detektivarbeideren!**)

reparallelize prestable/tmx/nob2fkv/news/ruijankaiku/ruijan-kaiku_aikamoinen_kartano.html.tmx.html

1. parallellize
2. redigere noe
3. reparallelize

Greppe ord som mangler i ordboka uten å miste oversettelsene / resten av tekstlinja:

- cat ~/Desktop/sannoi_3.txt |cut -f1|fkvnob|grep "?"|cut -f1|sort|uniq > inc/puuttuvat-sannoi_230220.csv
- for i in `cat inc/puuttuvat-sannoi_230220.csv` ; do grep "^$i" ~/Desktop/sannoi_3.txt >> inc/puuttuvat-sannoi_230220.dict.csv ; done

Script
cat fil.txt | sed "s/ / /" | perl ~/main/words/dicts/fkvnob/scripts/c2x.pl |l

antiword sannoi.doc |./sannoi.sed|tr "\n" ™|sed "s/™ / /g;"|tr ™ "\n" > sannoi_3.txt

cat filnavn.txt | tr "- \n" " "| l (bytt ut alle linjeskift med mellomrom)

Speller-arbeid:

nuvviDspeller (finn info i Online speller)
https://divvun.org/proofing/online-speller.html

ll tools/spellcheckers/*.txt (sjekk filer som kan redigeres)

Se på forslag:
thomasbk$ **hfst-ospell -S -n 5 tools/spellcheckers/fkv.zhfst**
kirjalinen
"kirjalinen" is NOT in the lexicon:
Corrections for "kirjalinen":
viralinen 24.244946
kirjaluinen 25.324387

Sjekk for et spesifikt ord (kalaa):
e kalaa |hfst-ospell -S -n 10 tools/spellcheckers/fkv.zhfst
"kalaa" is NOT in the lexicon:
Corrections for "kalaa":
kansa 7.182723
alas 8.165504
alla 8.422414
kallaa 8.727074
alan 8.843147
ala 9.650238

Farge-TV:

- **sh devtools/test_ospell-office_suggestions.sh**
- legg til skrivefeil i typos.txt
- endre verdier i strings.default.txt
- lagre og kjør make, før du gjør første punkt på nytt

(hvis ord som får rødstrek mangler i analysatoren, er det bra å legge til! Mari fikk rødstrek på «puolikova» i online speller)

(de viktigste filene man kan redigere er **typos.txt**, **editdist.default.txt**, **final_strings.default.txt** og **words.default.txt**)

542 ls -l tools/spellcheckers/fit.zhfst
544 e kalaa |hfst-ospell -S -n 10 tools/spellcheckers/fkv.zhfst
546 see tools/spellcheckers/spellercorpus.sort.txt
547 see tools/spellcheckers/weights/spellercorpus.raw.txt
548 see tools/spellcheckers/weights/spellercorpus.clean.txt

———————

Spellercorpus

see tools/spellcheckers/spellercorpus.*
———————
Sjekk for et spesifikt ord (kalaa):
e kalaa |hfst-ospell -S -n 10 tools/spellcheckers/fkv.zhfst
"kalaa" is NOT in the lexicon:
Corrections for "kalaa":
kansa 7.182723
alas 8.165504
alla 8.422414
kallaa 8.727074
alan 8.843147
ala 9.650238

———————————————————————————————
Sjekke opp tall: (e = echo)

lang-fit thomasbk$ e 12 | hfst-lookup src/fst/transcriptions/transcriptor-numbers-digit2text.filtered.lookup.hfstol
> **12 kakstoista**

lang-fit thomasbk$ e 7.3. | hfst-lookup src/fst/transcriptions/transcriptor-date-digit2text.filtered.lookup.hfstol
> **7.3. maaliskuun seitsemäs päivä**