# Status Quo efter lektion 6

Begyndte dagen med at lave lidt det samme som før, men nu i det analyserede output.

## Finde forkortelser

- Analyser et corpus på morfologisk niveau, med trace-information

```
cd ~/work
cat Aviscorpus.txt | ./kal-analyse --morf -t > Aviscorpus-analyzed.txt
```

- Find forkortelseskandidater

```
cat Aviscorpus-analyzed.txt | grep -C5 'Abbr' | more
```

Hvert output viser om ordet er mulig forkortelse, og om den er valgt eller fjernet. Et eksempel på noget der kunne være en forkortelse, men ikke blev valgt som det:

```
"<Filmit>"
    "film" OLang/DAN N Abs Pl
;   "Fil" Prop Gram/Abbr Sg Abl REMOVE:2630:0001L
;   "film" OLang/DAN N Abs Sg 2SgPoss REMOVE:3212:0101
;   "film" OLang/DAN N Rel Pl REMOVE:3051:0072
```

---

## Finde proprier

- Tilsvarende med proprier

```
cat Aviscorpus-analyzed.txt | grep -C5 'Prop'
```

Med eksempel på noget der kunne være både Prop eller N, og er valgt som Prop:

```
"<Sisimiuni>"
        "Sisimiut" Sem/Geo Prop Lok Pl SELECT:2666:0001X
;       "sisi" MIU Der/nn N Abs Pl 4SgPoss SELECT:2666:0001X
;       "sisi" MIU Der/nn N Abs Sg 4SgPoss SELECT:2666:0001X
;       "sisi" MIU Der/nn N Lok Pl SELECT:2666:0001X
;       "sisi" MIUQ Der/nn N Abs Pl 4SgPoss REMOVE:1800:MIUQ
;       "sisi" MIUQ Der/nn N Abs Sg 4SgPoss REMOVE:1800:MIUQ
;       "sisi" MIUQ Der/nn N Lok Pl REMOVE:1800:MIUQ
```

---

## Stavefejl eller andet sprog?

Ord der ikke kunne analyseres har analysen `?`. Vi brugte en del tid på at kigge på sådanne eksempler, for at se om det mest var stavefejl, slang, dialekt, eller faktisk udvikling i sproget.

- Find ord der ikke kunne håndteres af den morfologiske analyse

```
cat Aviscorpus-analyzed.txt | grep -C5 ' ?' | more
```

I Facebook-corpuset var der flere, specielt fordi alle emojis ikke håndteres.

# Noter & Hints

En metode til at finde sætningslængde men hvor man tæller ord i stedet for bogstaver:

```
   cat Ataqqinartuaraq.txt | perl -pe 's/([.?:])/\1\n/g;' | perl -pe
'$cnt=scalar(() = $_ =~ (m/(\s+)/g))+1; print "$cnt ";' | sort -nr | more
```

Her gives et meget kort Perl-script som tæller antal spaces på linien, og lægger 1 til det. I behøver ikke forstå `$cnt=scalar(() = $_ =~ (m/(\s+)/g))+1; print "$cnt ";` - det kan gøres nemmere med andre programmeringssprog, men Perl var lige det jeg havde.

Analysatoren kommer også med en stavekontrol, som bruges således:

```
   echo '5 ilasseqatigiisitsillunilu' | hfst-ospell-office
~/langtech/kal/tools/spellcheckers/fstbased/desktop/hfst/kl.zhfst
```

Input er hvor mange forslag man gerne vil have og så ordet man gerne vil have checket, så i dette tilfælde skal ordet *ilasseqatigiisitsillunilu* checkes og hvis det er forkert vil vi gerne have *5* rettelsesforslag.