# Opsætning og forklaring

# OCR-behandling af billeder fra NAS og gem OCR-PDF'er i samme mapper

Vi bruger:

- **Tesseract OCR for Windows**
- Et simpelt **Batch-script (ocr_billeder.bat)** til at køre hele processen
- OCR-resultater gemmes som `.pdf` i samme mapper
- Logger filerne du har behandlet
- Springer over eksisterende `.pdf`-filer

---

# 🔧 Trin 1: Installer Tesseract OCR på Windows

1. Download og installer: 🔗 https://github.com/UB-Mannheim/tesseract/wiki
2. Under installation:
    - Husk at vælge dansk sprog (`dan`) i sproglisten! Måske også Tysk og Tysk (fraktur)
    - Vælg at tilføje til PATH (så du kan kalde `tesseract` globalt)
3. Tjek at det virker: Åbn en **Kommandoprompt** og skriv:

    ```
    tesseract -v
    ```

    Du bør få en version tilbage.

**********

# Tilføj manuelt til PATH bagefter

Hvis du **allerede har installeret** Tesseract uden at tilføje til PATH, gør sådan her:

## 1. Find Tesseract's installationsmappe

Typisk er den her:

```
C:\Program Files\Tesseract-OCR
```

## 2. Tilføj til systemets PATH

**Windows 10 eller 11:**

1. Tryk `Win + S` og søg: `miljøvariabler` eller `Environment Variables`
2. Klik: **Rediger systemmiljøvariabler**
3. Klik på **Miljøvariabler...**
4. Find "Path" under "Systemvariabler" og klik "Rediger"

5. Klik "Ny" og indsæt:

   ```
   C:\Program Files\Tesseract-OCR
   ```

6. Tryk OK, OK, OK

### 3. Genstart evt. kommandoprompten

Åbn en ny CMD og test med:

```
tesseract -v
```

Hvis du ser en versionstekst, er det klar! 🎉

**********

## ☑ "Detected 33 diacritics"

Det er **ikke en fejl** – bare information.

### 🔍 Det betyder:

Tesseract har fundet 33 **diakritiske tegn** (som ø, å, é, ü, accenter osv.) i billedet under OCR-behandlingen.

> **Kort sagt:** Det er et godt tegn. Det betyder, at OCR'en fandt tegn som typisk findes i dansk tekst.

## ⚠ "Error in boxClipToRectangle: box outside rectangle"

## ⚠ "Error in pixScanForForeground: invalid box"

Disse er **advarsler fra Leptonica**, ikke Tesseract direkte.

### 🔍 Det betyder:

Tesseract forsøger at finde tekstområder (bounding boxes), men en eller flere af dem **ligger udenfor selve billedet**, eller **billedet har beskadigede eller tomme områder**.

### Mulige årsager:

- Billedet har **meget sort kant** eller tom margin
- Billedet er **forvrænget** eller indeholder fejl i formatet
- Filen er **korrupt** eller **forkert kodet**
- Tesseract tror, der er noget tekst i et område, der **faktisk ikke findes**

### 🔧 Men OCR'en virker stadig ofte:

Selv med disse fejlmeddelelser kan Tesseract som regel stadig udtrække tekst korrekt. Men hvis du får *meget få eller ingen resultater* i PDF'en, så er det et tegn på, at noget skal renses eller konverteres.

Brug evt. Scantailor til at efterbehandle billedfilerne først – det giver ofte bedre resultat.