# Syntetiske tekstbilder for nord-, sør-, lule- og inaresamisk

Dette datasettet inneholder syntetiske linjebilder som kan brukes til å finjustere OCR-modeller for nord-, sør-, lule- og inaresamisk. Fremgangsmåten for å lage disse bildene er å lage 'rene' linjebilder og tilføre støy ved hjelp av Augraphy [1].

## Tekstkilder

Teksten i datasettet kommer fra [Giellatekno] sitt korpus. Spesifikt brukte vi datafilene fra `converted/`-mappene i [2][3][4][5] (commit hasher `32f4af263cefae6ab9182638e2451ff151757adc`, `00dac0e9e74b4a89214ad7d34de27b83362b3f3a`, `4303edf80ae5eee2a036663c7b38756a0aa2a189`, `7e3437ce8c7dc7692ccbd2505412c03e9e617be6`).

## Datasett-oppdeling

Datasettet er tilfeldig delt opp slik at 71% av filene (307387 linjer) er i treningsdelen, 9% av filene (40765 linjer) er i valideringsdelen og 20% av filene er i (84534 linjer) testdelen. Hver del har en unik mengde skrifttyper og tekst- og bakgrunnsfarger.

## Språkfordeling

Fordelingen av de forskjellige språkene er

| Språkkode | Antall treningslinjer | Antall valideringslinjer | Antall testlinjer |
|---|---|---|---|
| sma | 76971 | 10992 | 21981 |
| sme | 76949 | 10992 | 21990 |
| smj | 76970 | 9081 | 20465 |
| smn | 76497 | 9700 | 20098 |

## Hvordan bruke datasettet

Datasettet er tilgjengelig som .parquet-filer

Det kan lastes inn med python og datasets-biblioteket slik:

```
from datasets import load_dataset

# last inn datasett med alle datasplitter
ds = load_dataset("sti_til_mappa_parquet_files/")

# hent hver splitt
train_ds = ds["train"]
```

```
val_ds = ds["validation"]
test_ds = ds["test"]

# eller last inn datasplit direkte
train_ds = load_dataset("sti_til_mappa_parquet_files/", split="train")
```

eller med et annet verktøy du foretrekker.

Kolonnene "image", "text" og "language_code" inneholder bildedata, teksten i bildet, og språkkoden, respektivt.

### Referanse til datasettet

Hvis du bruker dette datasettet i din forskning, vennligst siter både "Enstad T, Trosterud T, Røsok MI, Beyer Y, Roald M. Comparative analysis of optical character recognition methods for Sámi texts from the National Library of Norway. Akseptert for publisering i Proceedings of the 25th Nordic Conference on Computational Linguistics (NoDaLiDa) 2025." (se [artikkel-repositoriet]) og SIKOR-datasettet som de samiske tekstene er hentet fra: "SIKOR UiT The Arctic University of Norway and the Norwegian Saami Parliament's Saami text collection, http://gtweb.uit.no/korp, Versjon 01.12.2021 [Datasett]." Merk også at SIKOR-datasettet for å få samisk tekst til bildene er lisensiert under (CC-BY 3.0).

### Datasettlisens

Datasettet er lisensiert med en CC-BY 3.0-lisens.

## Kildekode

Koden for å lage dette datasettet er tilgjengelig på vårt [GitHub Repo] (commit hash `90341bc19d6368c7848dcc2459065058486a89ea`).

## Fotnotelenker

[1]: https://github.com/sparkfish/augraphy
[2]: https://github.com/giellalt/corpus-sma
[3]: https://github.com/giellalt/corpus-sme
[4]: https://github.com/giellalt/corpus-smj
[5]: https://github.com/giellalt/corpus-smn
[Giellatekno]: https://giellatekno.uit.no/
[GitHub Repo]: https://github.com/Sprakbanken/synthetic_text_images
[artikkel-repositoriet]: https://github.com/Sprakbanken/nodalida25_sami_ocr