Synthetic text images for North, South, Lule and Inari Sámi

This dataset contains synthetic line images meant for fitting OCR models for North, South, Lule and Inari Sámi. Clean line images are created using Pillow and they are subsequently distorted using Augraphy [1].

Text sources

The text in this dataset comes from [Giellatekno]'s corpus. Specifically, we used the data files of the `converted/`-directories of [2][3][4][5] (commit hashes `32f4af263cefae6ab9182638e2451ff151757adc`, `00dac0e9e74b4a89214ad7d34de27b83362b3f3a`, `4303edf80ae5eee2a036663c7b38756a0aa2a189`, `7e3437ce8c7dc7692ccbd2505412c03e9e617be6`).

Splits

The dataset is split randomly by file so 71 % of the files (307387 lines) are in the training split, 9 % of the files (40765 lines) are in the validation split and 20 % of the files (84534 lines) are in the test split. Each split has a unique set of typefaces and text/background colors.

Language distribution

The language distribution for the different languages are

Language Code	Num train lines	Num val lines	Num test lines
sma	76971	10992	21981
sme	76949	10992	21990
smj	76970	9081	20465
smn	76497	9700	20098

Using the dataset

The dataset is available as .parquet-files

It can be loaded with python and the datasets library like this:

```
from datasets import load_dataset

# load dataset dict with all splits
ds = load_dataset("path_to_parquet_files/")
train_ds = ds["train"]
val_ds = ds["validation"]
test_ds = ds["test"]
```

```
# or load data splits directly
train_ds = load_dataset("path_to_parquet_files/", split="train")
```

or with another tool of your choice.

The "image", "text", and "language_code" columns contain the image data, the text in the image and the language code, respectively.

Referencing the dataset

If you use this dataset in your research, then please cite both "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. Accepted for publication in Proceedings of the 25th Nordic Conference on Computational Linguistics (NoDaLiDa) 2025." (see the [paper repository]) and the SIKOR dataset the Sámi text is from: "SIKOR UiT The Arctic University of Norway and the Norwegian Saami Parliament's Saami text collection, http://gtweb.uit.no/korp, Version 01.12.2021 [Data set]." Also note that the SIKOR dataset to get Sámi text for the images is (CC-BY 3.0) licensed.

Dataset license

The dataset is licensed with a CC-BY 3.0 license.

Code

The code to create this dataset is available on our [GitHub Repo] (commit hash `90341bc19d6368c7848dcc2459065058486a89ea`).

Footnote links

[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
[paper repository]: https://github.com/Sprakbanken/nodalida25_sami_ocr