Omsetjingsminne frå Nynorsk pressekontor

Desse omsetjingsminna inneheld omsetjingar frå bokmål til nynorsk av nyhendetekst frå Norsk telegrambyrå (NTB). Tekstene er omsette av Nynorsk pressekontor, som nyttar den såkalla Nynorskroboten til automatisk omsetjing av tekster frå bokmål til nynorsk, og korrigerer feila roboten gjer manuelt før publisering. Les meir om dette hjå Nynorsk pressekontor (https://www.npk.no/).

Materialet skriv seg frå perioden februar 2011 til september 2021. Av opphavsrettslege grunnar er omsetjingseiningane randomiserte. Ei omsetjingseining svarar stort sett til eit setningspar.

Del 1 (2011-2019)

Den eldste delen av korpuset inneheld data frå perioden 2011 til februar 2019. Setningane er automatisk parallellstilte, og korpuset er konvertert til tmx-format (translation memory exchange). Denne delen av korpuset inneheld om lag 339.000 omsetjingseiningar (TU-taggar), i grove trekk svarar dette til setningar. Ein oppdatert versjon av omsetjingsminnet vart lagt til i mars 2020. Ein del feil og manglar i den førre versjonen har vorte retta opp i. Denne optimaliseringa vart utført av Vitec MV. Sjå dokumentasjonen for detaljar om dette. Den opphavelege fila er framleis mogleg å laste ned.

Del 2 (2019 og 2020-2021)

Dei to nyaste filene inneheld data på tsv-format (tab-separert tekst).

Fila *2019_tm_npk_ntb.zip* inneheld data frå 2019 (f.o.m. februar t.o.m. desember) med tre variantar av tekstane, originaltekst (bokmål), automatisk omsetjing til nynorsk, og den korrigerte, endelege versjonen av nynorskomsetjinga. Ein slik triplett opptek ei line i fila, variantane er skilde frå kvarandre med ein tab. Fila inneheld totalt 162.649 triplettar.

Fila *2020_2021_tm_npk_ntb.zip* inneheld data f.o.m. januar 2020 t.o.m. 9. september 2021. Her finst originalteksten (bokmål) og den korrigerte omsetjinga (nynorsk) på same line, skilde frå kvarandre med ein tab. Til saman finst det 196.946 omsetjingspar i denne fila.

Fil	Omsetjingar (~setningar)	Filstorleik (.zip)	Filstorleik (utpakka)	Filformat (utpakka)
2020_2021_tm_npk_ntb.zip	196.946	10.3 MB	39.4 MB	.tsv
2019_tm_npk_ntb.zip	162.649	9.3 MB	50.9 MB	.tsv
2011_2019_tm_npk_ntb_vitecmv.zip	339.000	20.0 MB	101.6 MB	.tmx
Totalt antal omsetjingar/setningar	698.595			

Fil	Dato
2020_2021_tm_npk_ntb.zip	2020-01-01 — 2021-09-09
2019_tm_npk_ntb.zip	2019-02 — 2019-12-31
2011_2019_tm_npk_ntb_vitecmv.zip	2011 — 2019-02

Nasjonalbiblioteket, 2021-11-04

Translation memories from Nynorsk News Press Agency

This corpus contains translations of news text from Norwegian Bokmål to Norwegian Nynorsk. The texts are produced by the Norwegian News Agency (https://www.ntb.no/about-ntb), and translated by Nynorsk News Press Agency (https://www.npk.no/), who translate the texts automatically (using the so called *Nynorsk Robot*), then correcting the translations manually before publication.

The material in this corpus was produced in the period from 2011 to September 2021. For copyright reasons, the translation units(in most cases corresponding to sentence pairs) have been randomized.

Part 1 (2011-2019)

The oldest part of the corpus contains data from the period 2011 to February 2019. The texts have been parallelized automatically and converted to tmx (translation memory exchange file format). This part of the corpus contains about 339,000 translation units (TU tags). An updated version of this part of the corpus was added in March 2020. Some errors and bugs in the original version were rectified. These corrections were done by Vitec MV. See the documentation (included in the download) for details. The original file can also be downloaded.

Part 2 (2019 and 2020-2021)

The newest part contains two files in tsv format (tab-separated text).

The file *2019_tm_npk_ntb.zip* contains data from 2019 (February to December) with three variants of the texts, the original text (Bokmål), the automatic translation into Nynorsk, and the corrected, final version of the Nynorsk translation. One such triplet takes up a line in the file, and the variants are separated from each other by a tab. The file contains a total of 162,649 triplets.

The file *2020_2021_tm_npk_ntb.zip* contains data from January 2020 to September 9, 2021. Here, only the original text (Bokmål) and the corrected translation (Nynorsk) is contained. The two variants occur on the same line, separated by a tab. There is a total of 196,946 sentence pairs in this file.

File	Translations (~sentences)	File size (.zip)	File size (unzipped)	File format (unzipped)
2020_2021_tm_npk_ntb.zip	196,946	10.3 MB	39.4 MB	.tsv
2019_tm_npk_ntb.zip	162,649	9.3 MB	50.9 MB	.tsv
2011_2019_tm_npk_ntb_vitecmv.zip	339,000	20.0 MB	101.6 MB	.tmx
Total	698,595			

File	Date
2020_2021_tm_npk_ntb.zip	2020-01-01 — 2021-09-09
2019_tm_npk_ntb.zip	2019-02 — 2019-12-31
2011_2019_tm_npk_ntb_vitecmv.zip	2011 — 2019-02

The National Library of Norway, 2021-11-04