

European Language Grid – META-FORUM 2020
Jörg Tiedemann (University of Helsinki, Finland)

ELG Pilot Project: Open Translation Models, Tools and Services

OPUS-MT – Public NMT Models & Tools

Language Technology - University of Helsinki



<https://opus.nlpl.eu>
OPUS

corpus	docs*	sentences	tokens	XCONLL-XML	raw	TMX	Moses	mono raw	udg dic	freq	other files
EUBookshop	10,400	10,400	80,834	431,600	431,600	431,600	431,600	431,600	431,600	431,600	[sample]
MultiUN	8740	14,240	373,840	454,000	454,000	454,000	454,000	454,000	454,000	454,000	[sample]
OpenSubtitles2018	55650	45,240	363,400	336,000	336,000	336,000	336,000	336,000	336,000	336,000	[sample]
Opus-MT	20,000	20,000	20,000	20,000	20,000	20,000	20,000	20,000	20,000	20,000	[sample]
DGT	26879	3,140	72,830	68,700	68,700	68,700	68,700	68,700	68,700	68,700	[sample]
Europarl	1825	1,825	36,500	36,500	36,500	36,500	36,500	36,500	36,500	36,500	[sample]
JRC-Acquis	13000	13,000	36,000	36,000	36,000	36,000	36,000	36,000	36,000	36,000	[sample]
Wikipedia	2	0,800	23,000	17,800	17,800	17,800	17,800	17,800	17,800	17,800	[sample]
IMDb	100	100	30,000	30,000	30,000	30,000	30,000	30,000	30,000	30,000	[sample]
GlobeVoices	14501	0,340	7,000	7,000	7,000	7,000	7,000	7,000	7,000	7,000	[sample]
ECB	7398	0,240	5,700	6,500	6,500	6,500	6,500	6,500	6,500	6,500	[sample]
News-Commentary	2293	0,050	5,600	5,300	5,300	5,300	5,300	5,300	5,300	5,300	[sample]
GNOME	2	0,020	5,000	5,000	5,000	5,000	5,000	5,000	5,000	5,000	[sample]
News-Commentary (test)	278022	117,240	1,760	117,230	92,240	108,400	117,230	117,230	117,230	117,230	[sample]

Search & download resources: en (English) fr (French) zh-TW (Chinese)

Language resource click on [] to see its details | click on [] to download the file! (raw = untokenized, udg = paired with universal dependencies, slg = word alignments and phrase tables)



Sam Hardwick



Tommi Nieminen



<https://blogs.helsinki.fi/fiskmo-project>



<https://github.com/Helsinki-NLP/OPUS-CAT>

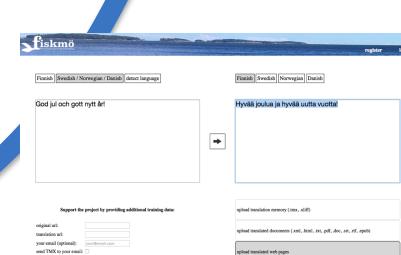
<https://opus-repository.ling.helsinki.fi>

ELG-Project:
Focus on public NMT
for European Minority
Languages

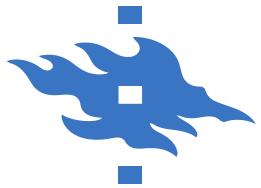


<https://github.com/Helsinki-NLP/Opus-MT>

OPUS
mt



<https://translate.ling.helsinki.fi/>



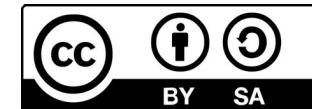
OPUS-MT - Public NMT Models & Tools

<https://github.com/Helsinki-NLP/Opus-MT>



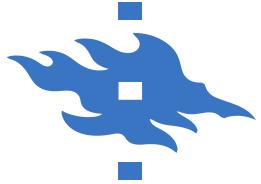
Available software:

- MT server solution based on Marian NMT
- dockerised web-app, translation interface and API
- NMT training pipeline (OPUS-MT-train)
- CAT integration (OPUS-CAT)



Pre-trained translation models:

- number of bilingual models: 1,048
- number of multilingual models: 53
- number of supported source languages: 229
- number of supported target languages: 222
- number of supported language pairs: 1,715



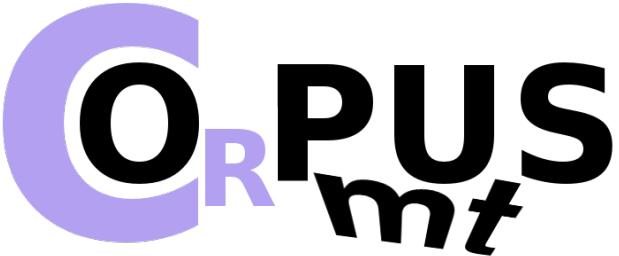
OPUS-MT – Public NMT Models & Tools

<https://github.com/Helsinki-NLP/Opus-MT>

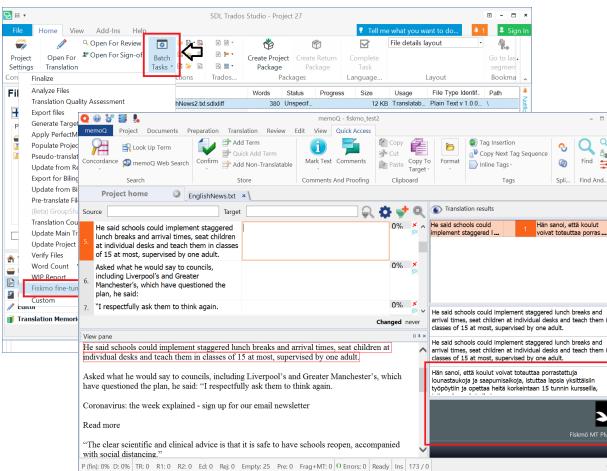


```
{
    "alignment": [
        "0-0 0-2 1-1 2-3",
        "0-0 1-1 3-2 4-3 5-4"
    ],
    "result": "How are you? The translation is fun.",
    "server": "192.168.1.18:20001",
    "source": "fi",
    "source-segments": [
        "Mit\u00e4 kuuluu ?",
        "K\u00e4\u00e4nn\u00f6 s on hauskaa ."
    ],
    "source-sentences": [
        "Mit\u00e4 kuuluu?",
        "K\u00e4\u00e4nn\u00f6 s on hauskaa."
    ],
    "target": "en",
    "target-segments": [
        "How are you ?",
        "The translation is fun ."
    ],
    "target-sentences": [
        "How are you?",
        "The translation is fun."
    ]
}
```

Development demo:
<https://translate.ling.helsinki.fi/ui/sami>



OPUS-CAT - plugins and local MT engines
<https://github.com/Helsinki-NLP/OPUS-CAT>



OPUS-MT at huggingface
<https://huggingface.co/Helsinki-NLP>

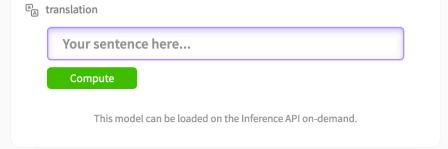


HUGGING FACE

[Back to all models](#)
Model: Helsinki-NLP/opus-mt-ROMANCE-en

pytorch rust marian lm-head seq2seq roa en translation

Hosted inference API ⓘ

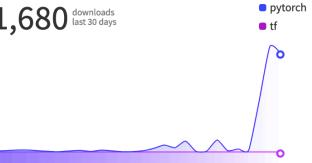


This model can be loaded on the Inference API on-demand.

Monthly model downloads

[Helsinki-NLP/opus-mt-ROMANCE-en](#)

31,680 downloads last 30 days



How to use this model directly from the [transformers](#) library:

```
from transformers import AutoTokenizer,
AutoModelWithLMHead

tokenizer = AutoTokenizer.from_pretrained("Helsinki-
NLP/opus-mt-ROMANCE-en")

model = AutoModelWithLMHead.from_pretrained("Helsinki-
NLP/opus-mt-ROMANCE-en")
```

List all files in model · See raw config file