

Democratizing MT



OPUS



bergamot

ORPUS



A bit of history ...



OPUS - an open source parallel corpus

<http://logos.uio.no/opus/>

Jörg Tiedemann
Department of Linguistics
Uppsala University
Box 527
SE-751 20 Uppsala, Sweden
joerg@stp.ling.uu.se

Lars Nygaard
Tekstlaboratoriet HF
University of Oslo
Postboks 1102 Blindern
0317 Oslo
lars.nygaard@ilf.uio.no

1 Introduction

Parallel corpora are useful in a wide variety of research areas, particularly in machine translation and lexicography. However, parallel corpora have been few, often unrepresentative, and not generally available. The aim of the OPUS project is to provide a public collection of parallel corpora which can be freely used and distributed. This makes it possible for everyone to

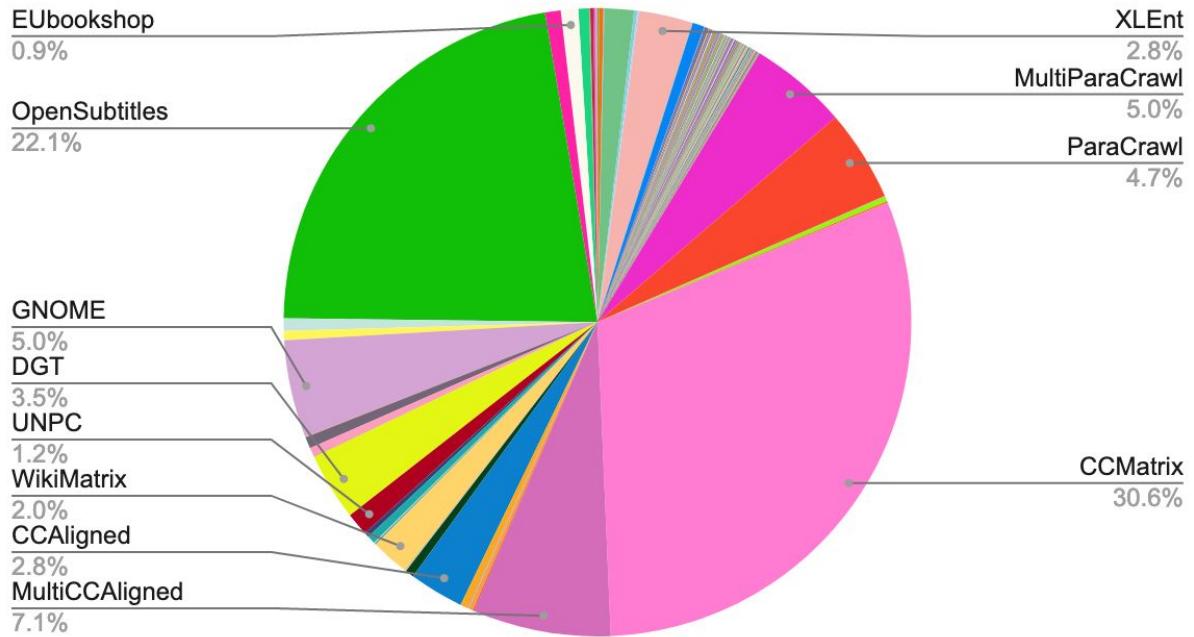


- Total release size: 12 TB
- > 600 languages (without regional variants)
- > 40,000 language pairs
- > 22 billion sentences in 31 billion alignments
- > 340 billion tokens

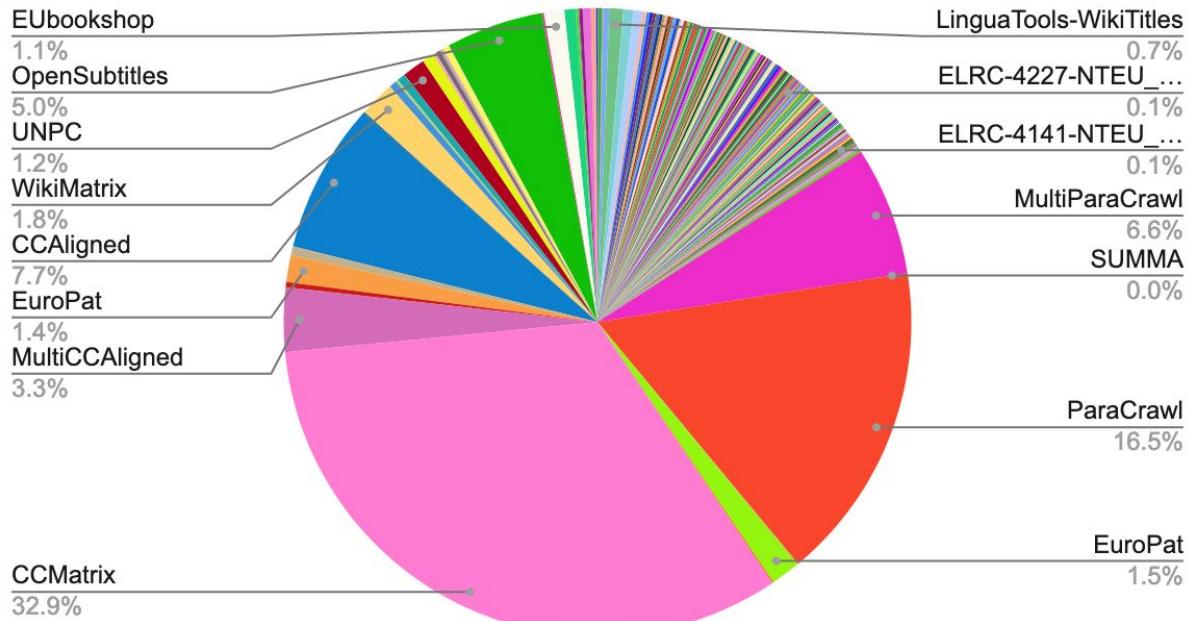
Released packages:

- Native XML format (standoff alignment)
- Plain aligned text, TMX, monolingual data
- Word/phrase alignments, frequency tables

Alignments



Tokens





OPUS-API

- Web service for searching data sets
- <https://opus.nlpl.eu/opusapi/>

OpusTools

- Find and download data sets
- Convert and extract data
- pip install opustools

Using opus-tools

OpusTools

Tools for accessing and processing OPUS data.

- opus_read: read parallel data sets and convert to different output formats
- opus_express: Create test/dev/train sets from OPUS data.
- opus_cat: extract given OPUS document from release data
- opus_get: download files from OPUS
- opus_langid: add language ids to sentences in xml files in zip archives

Installation:

```
pip install opustools
```

```
opus_read --directory RF \
    --source en \
    --target sv \
    --write en-sv.en en-sv.sv \
    --write_mode moses
```

Print XCES align format of all 1:1 sentence alignments:

```
opus_read --directory RF \
    --source en \
    --target sv \
    --src_range 1 \
    --tgt_range 1
```

Print alignments with alignment certainty greater than 1.1:

```
opus_read --directory RF \
    --source en \
    --target sv \
    --attribute certainty \
    --threshold 1.1
```

<https://github.com/Helsinki-NLP/OpusTools>

OpusFilter

AVAILABLE FUNCTIONS

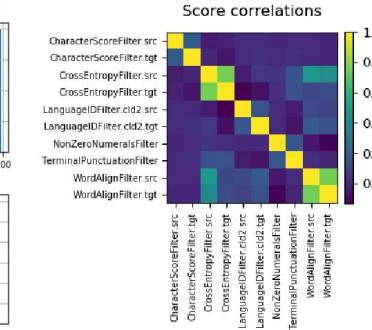
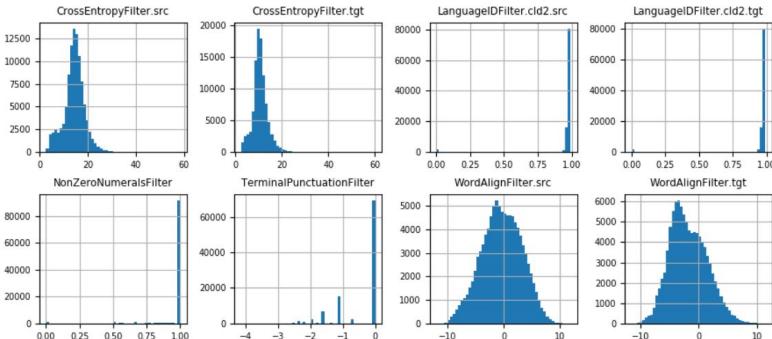
- Downloading and selecting data
- Preprocessing text
- Filtering and scoring
- Using score files
- Training language and alignment models
- Training and using classifiers

AVAILABLE FILTERS

- Length filters
- Script and language identification filters
- Special character and similarity filters
- Language model filters
- Alignment model filters
- Sentence embedding filters
- Custom filters

steps:

```
- type: opus_read  
parameters:  
  corpus_name: ParaCrawl  
  source_language: fi  
  target_language: en  
  release: v4  
  preprocessing: raw  
  src_output: paracrawl.fi.gz  
  tgt_output: paracrawl.en.gz
```



```
- type: filter  
parameters:  
  inputs:  
    - all.fi.gz  
    - all.en.gz  
  outputs:  
    - filtered.fi.gz  
    - filtered.en.gz  
  filters:  
    - LengthFilter:  
      unit: word  
      min_length: 1  
      max_length: 100  
    - LengthRatioFilter:  
      unit: word  
      threshold: 3
```

<https://github.com/Helsinki-NLP/OpusFilter>

Bicleaner AI: sentence-level scoring

- More rule-based filters!
- Binary classifier
 - Noisy sentences scored near to 0.
 - Cleaner sentences scored near to 1.
- Classifier based on XLM-Roberta.
 - Multilingual models and zero-shot classification also!
- Downloadable trained models for 32 language pairs
 - All the 70 languages of HPLT will be supported!

Bicleaner AI: sentence-level scoring

Click Run to start the installation immediately.	Kliknite Pokreni da biste odmah započeli instalaciju.	0.920
The 206L LongRanger is a stretched variant with seating for seven.	206L LongRanger je izdužena inačica s prostorom za smještaj sedam osoba.	0.981
Just go." he said. it sucks.	Najebali smo. stvarno.." rekao je Rinčica..	0.154
Kundalini will be awakened quickly if the	Ako pojedinac prodre .	0.074

Effective data curation

Data curation and cleaning (empty-train)

Empty Train

localhost:5173/frontend/#/download/

DATAFLOR

DATASETS CATALOGUE | TODO DATASETS

Search dataset... Monolingual Bilingual Latest only Finnish (fi) Ukrainian (uk) Corpus name

Get corpora from OPUS-API

Downloads

Dataset	Version	Source	Target	Count	Size
bible-uedin	v1	fi	uk	7,938	721.00 kB
CCMatrix	v1	fi	uk	2,253,407	162.92 MB
ELRC-5179-acts_Ukrainian	v1	fi	uk	129,312	12.21 MB
ELRC-wikipedia_health	v1	fi	uk	32	5.00 kB
EUbookshop	v2	fi	uk	1,613	152.00 kB
GNOME	v1	fi	uk	150	5.00 kB
KDE4	v2	fi	uk	104,258	2.69 MB
MultiCCAligned	v1.1	fi	uk	1,364,608	134.49 MB
MultiParaCrawl	v9b	fi	uk	1,542,555	118.71 MB
NeuLab-TedTalks	v1	fi	uk	15,531	1.18 MB
OpenSubtitles	v2018	fi	uk	527,306	18.26 MB
QED	v2.0a	fi	uk	42,804	3.08 MB

Data curation and cleaning (empty-train)

The screenshot shows the DATA LOR interface for the dataset **EU-dcep-1-eng-nld**. The interface compares two versions of the dataset: **original (3000)** and **clean (2867)**. A red arrow points from the text "Apply OPUS and your own filters" to the sidebar on the right.

Dataset: EU-dcep-1-eng-nld

Display as rows

English **Dutch**

12 July 2001 FINAL A5-0263/2001	12 juli 2001 DEFINITIEVE VERSIE A5-0263/2001
on the Commission communication on accelerated action targeted at major communicable diseases within the context of poverty reduction	over de mededeling van de Commissie inzake versnelde actie ter bestrijding van de belangrijkste infectieuze ziekten in het kader van armoedebestrijding
on the Commission communication on a Programme for Action: Accelerated action on HIV/AIDS, malaria and tuberculosis in the context of poverty reduction	over de mededeling van de Commissie inzake een ACTIEPROGRAMMA: Versnelde actie ter bestrijding van HIV/aids, malaria en tuberculose in het kader van de armoedebestrijding
Committee on Development and Cooperation	Commissie ontwikkelingssamenwerking
Bashir Khanbhai	Bashir Khanbhai
286.865 CONTENTS	286.865 INHOUD
OPINION of the Committee on Industry, External Trade, Research and Energy	ADVIES VAN DE COMMISSIE INDUSTRIE, EXTERNE HANDEL, ONDERZOEK EN ENERGIE
OPINION of the Committee on the Environment, Public Health and Consumer Policy	ADVIES VAN DE COMMISSIE MILIEUBEHEER, VOLKSGEZONDHEID EN CONSUMENTENBELEID
OPINION of the Committee on Women's Rights and Equal Opportunities	ADVIES VAN DE COMMISSIE RECHTEN VAN DE VROUW EN GELIJKE KANSEN
By letter of 21 September 2000, the Commission forwarded to	Bij schrijven van 21 september 2000 deed de Commissie haar

Import dataset

Search filters...

fix_wiki

- ALWAYS Always remove patterns
- FOOTNOTES Remove footnotes, e.g. [1], [2]
- URLS Remove url's
- WIKILINKS Remove [[wikilinks]]
- CODE Remove lines that contain code
- HEADINGS Remove ==headings==
- REMOVEEMPTYLINES Remove sentence pairs when one side is empty after filtering

opus.WhitespaceNorm... 3000

opus.LengthRatioFilter 2867

threshold

unit

Data curation and cleaning (empty-train)

The screenshot shows the DATAILOR web application interface for dataset configuration. At the top, the title bar says "Empty Train" and the URL is "localhost:5173/frontend/#/datasets/EU-dcep-1-eng-nld/configuration".

The main area displays a comparison between "original (3000)" and "clean (2867)" rows. A red arrow points from the text "See effects immediately" to the "changes" tab.

The "changes" tab shows a table comparing English and Dutch text samples. The English text is in light grey, and the Dutch translation is in white on a dark background. A red arrow points from the "changes" tab to the right-hand sidebar.

The right-hand sidebar lists various cleaning filters:

- fix_wiki (3000) - checked
- ALWAYS (Remove remove patterns)
- FOOTNOTES (Remove footnotes, e.g. [1], [2])
- URLS (Remove url's)
- WIKILINKS (Remove [[wikilinks]])
- CODE (Remove lines that contain code)
- HEADINGS (Remove ==headings==) - checked
- REMOVEEMPTYLINES (Remove sentence pairs when one side is empty after filtering)
- opus.WhitespaceNormalizer (3000) - checked
- opus.LengthRatioFilter (2867)
- threshold (3)
- unit (word)

Efficient training pipelines

Efficient training pipelines (empty-trainer)

```
datasets:          start:  
    clean: clean.gz      - clean 0.9  
    medium: medium.gz    - medium 0.1  
    dirty: dirty.gz      - until clean 6  
  
stages:           mid:  
    - start              - clean 0.6  
    - mid                - medium 0.3  
    - end                - dirty 0.1  
                          - until medium 1  
  
modifiers:  
    - uppercase 0.05  
    - titlecase 0.05  
  
seed: 31337
```

1. Reads, shuffles and mixes datasets
2. Augment on the fly
3. Different stages of training
... and eventually
4. See and control training process
(*a la Tensorboard*)

```
empty-trainer/trainer.py schedule.yml marian --tsv --train-sets --shuffle batches stdin ...
```

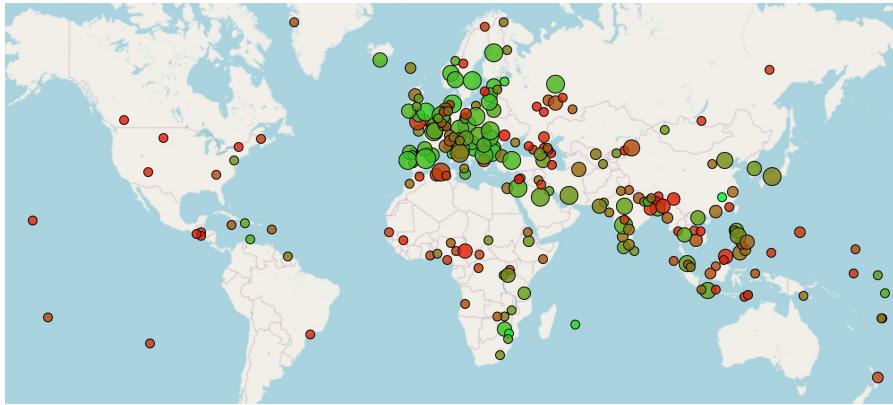
Efficient training pipelines (empty-trainest)

Not yet ready to share...

... but <https://github.com/mozilla/firefox-translations-training> exists!

This is snakemake-based pipeline to produce fast & small Marian translation models.
A product of the Bergamot project.

Scaling-up, integration and use cases



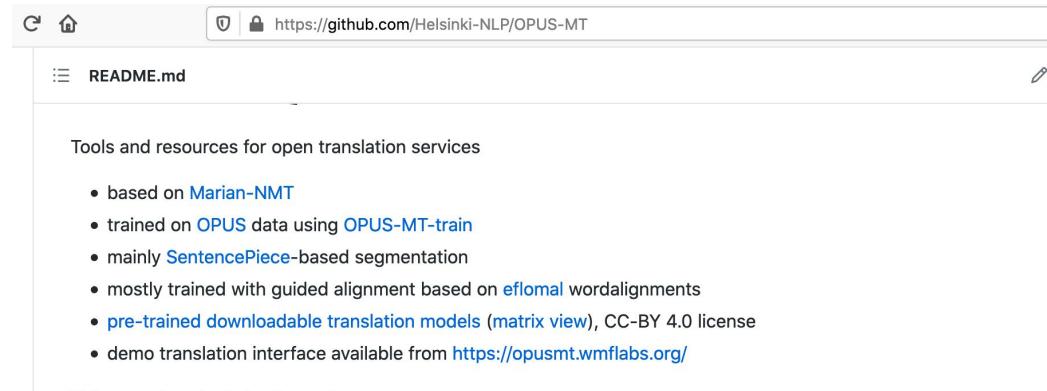
<https://github.com/Helsinki-NLP/OPUS-MT>

OPUS → **OPUS_{mt}**



Tatoeba Translation
Challenge

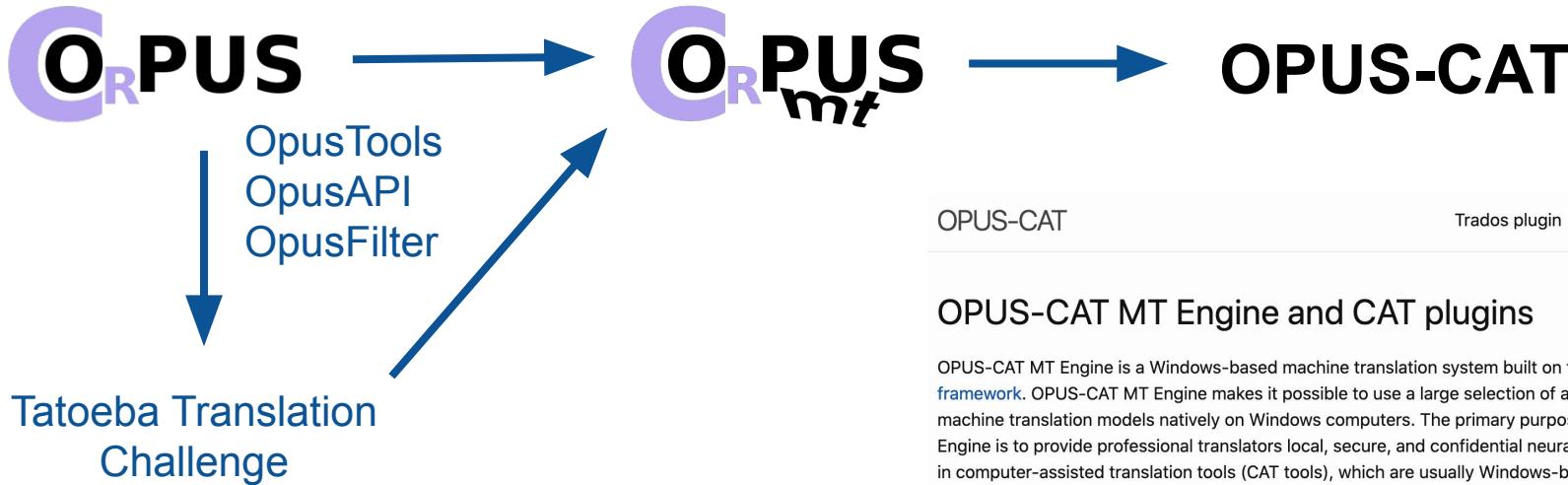
<https://github.com/Helsinki-NLP/Tatoeba-Challenge/>



The FSF argues that free software should not place restrictions on commercial use.^[48] and the GPL explicitly states that GPL works may be sold at any price.

<https://helsinki-nlp.github.io/OPUS-CAT>

The FSF argues that free software should not place restrictions on commercial use.^[48] and the GPL explicitly states that GPL works may be sold at any price.



OPUS-CAT

Trados plugin Fine-tuning About

OPUS-CAT MT Engine and CAT plugins

OPUS-CAT MT Engine is a Windows-based machine translation system built on the [Marian NMT framework](#). OPUS-CAT MT Engine makes it possible to use a large selection of advanced neural machine translation models natively on Windows computers. The primary purpose of OPUS-CAT Engine is to provide professional translators local, secure, and confidential neural machine translation in computer-assisted translation tools (CAT tools), which are usually Windows-based. To that end, there are plugins available for two of the most popular CAT tools, SDL Trados Studio and memoQ (OPUS-CAT can also be used in the Wordfast CAT tool as a custom provider). OPUS-CAT MT Engine provides pretrained MT models for a very wide selection of language pairs, courtesy of the OPUS MT project ([listing of OPUS MT models](#)).

Translation PyTorch Transformers en fi cc-by-4.0 marian text2text-generation

opus-mt-tc Eval Results AutoTrain Compatible

Train Deploy

Model card Files

```
from transformers import pipeline  
pipe = pipeline("translation", model="Helsinki-NLP/opus-mt-tc-big-fi-en")  
print(pipe("Kolme kolmanteen on kaksikymmentä seitsemän.))
```

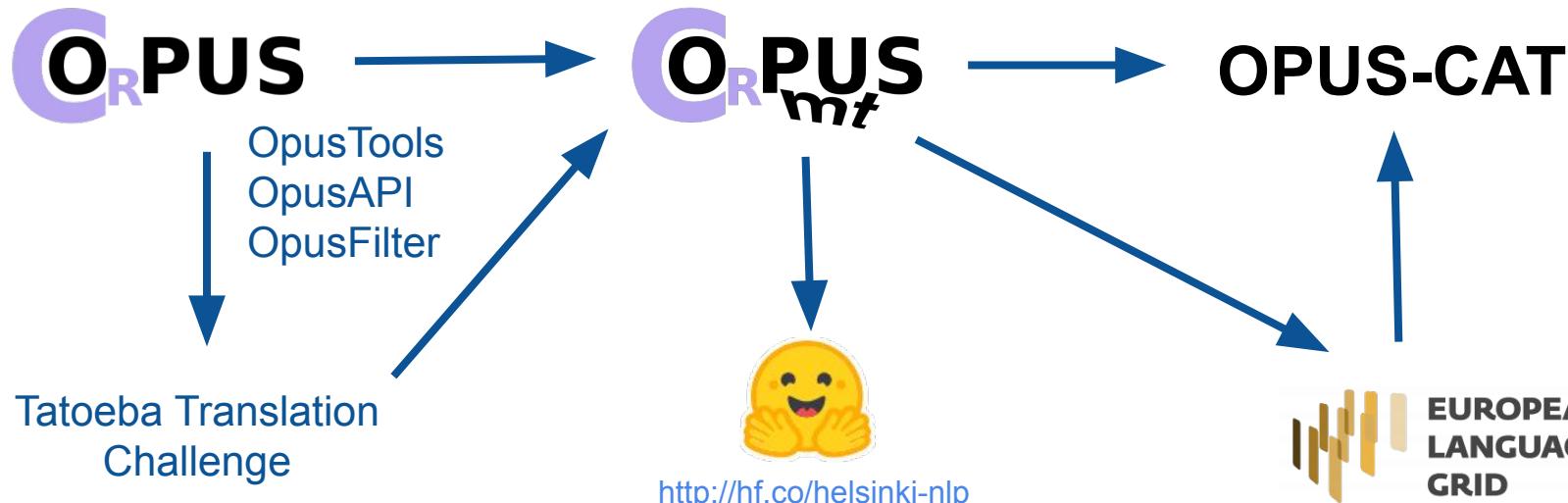
Edit model card

Downloads
last month
990

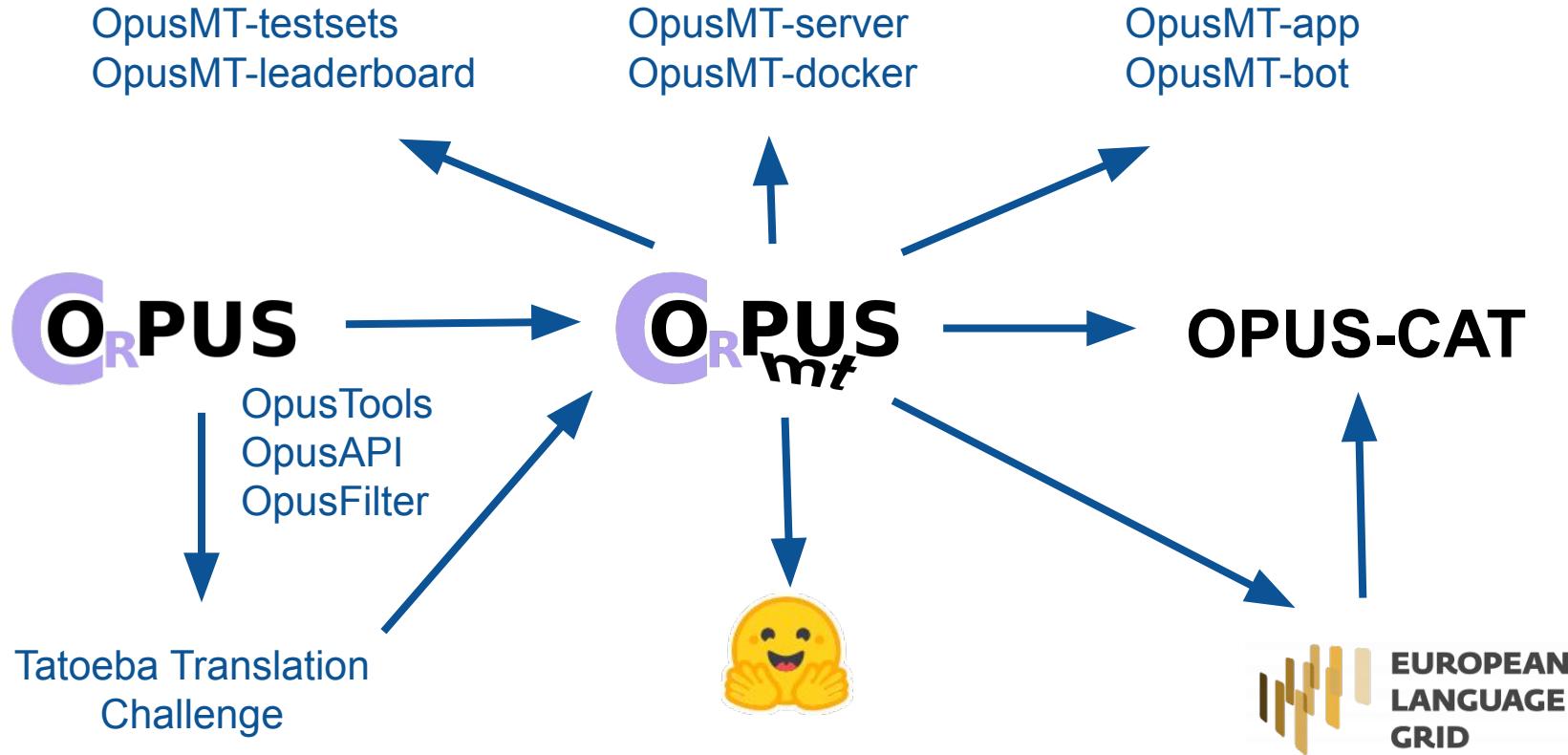
opus-mt-tc-big-fi-en



Opus HelsinkiNLP - OPUS-MT: English-Ukrainian machine translation
OPUS-MT eng-ukr
Version: 2022.3.13
ELG-compatible service



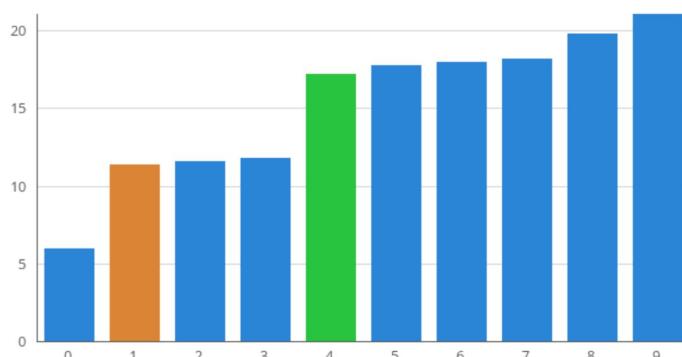
<https://live.european-language-grid.eu/>



Tracking progress with the OPUS-MT dashboard

OPUS-MT leaderboard

- Language pair: fin-ukr [top models]
- Model: all models
- Benchmark: flores101-devtest [fin-ukr] [all benchmarks]
- Metrics: bleu [chrf][comet]



Model Scores (bleu scores on the "flores101-devtest" testset)

ID	bleu	Other	Output	Model Download
9	21.0	scores	show, download	fin-zle/opusTCv20210807+xb+bt+pft+pbt_transformer-big_2022-04-27.zip
8	19.7	scores	show, download	fin-ukr/opusTCv20210807+pbt_transformer-align_2022-03-07.zip
7	18.1	scores	show, download	fin-zle/opusTCv20210807+bt_transformer-big_2022-03-23.zip
6	17.9	scores	show, download	fin-zle/opusTCv20210807+bt_transformer-big_2022-03-17.zip
5	17.7	scores	show, download	fin-zle/opus4m+btTCv20210807-2022-01-19.zip
4	17.1	scores	show, download	fin-ukr/opusTCv20210807+pbt+pft-sepvoc_transformer-tiny11-align_2022-03-16.zip
3	11.7	scores	show, download	fiu-zle/opus-2021-02-11.zip
2	11.5	scores	show, download	fiu-sla/opus-2021-02-16.zip
1	11.3	scores	show, download	fi-uk/opus-2020-01-08.zip
0	5.9	scores	show, download	tatoeba-zero/opus-2020-06-19.zip

[start] [show previous] show examples 110 - 119 [show next]

Sielta he löysivät 53-vuotiaan Saroja Balasubramanianin ruumiin, joka oli peitetty veritahraisilla huovilla.
Там вони знайшли тіло Сарожа Баласубраманіана у віці 53 років, вкрите ковдрами у плямах крові.
Там вони знайшли тіло 53-річного [-Сароджі-] {+Сароя+} Баласубраманіяна, покрите [-плямами крові.-] {+кров'ю.+}

Polisiin mukaan ruumis vaikutti olleen siellä noin vuorokauden ajan.

Поліція заявила, що тіло пролежало там близько доби.

За [-даними поліції,-] {+словами поліцейських,+} тіло пролежало там близько 24 годин.

Kuluvan kauden ensimmäiset tautitapaukset raportoitiin heinäkuun lopulla.

Про перші випадки хвороби цього сезону було повідомлено наприкінці липня.

Перші випадки захворювання були [-зареєстровані-] {+повідомлені+} в кінці липня.

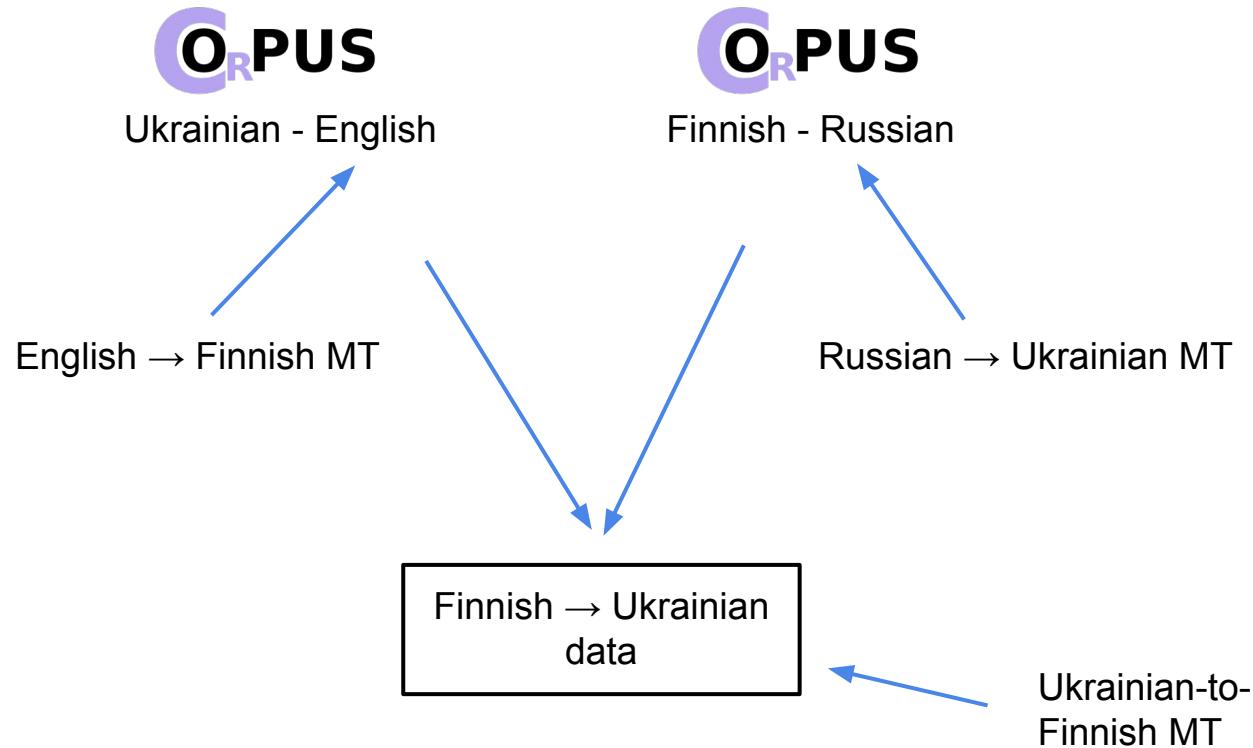
Tautia kantavat siat, joista se sitten siirryy ihmisiin hyttysten kautta.

Хвороба переноситься свинями, а потім через комарів передається до людей.

[-Хвороба переноситься свинями,-]

{+Захворюють свині,+} від яких [-потім передається людям-] {+вони передаються людям+} через комарів.

OPUS-MT and Ukrainian LT



English-centric data

MultiParaCrawl:
Finnish → Ukrainian
data through English
as a pivot language



Ukrainian Wikipedia

Democratizing NLP with



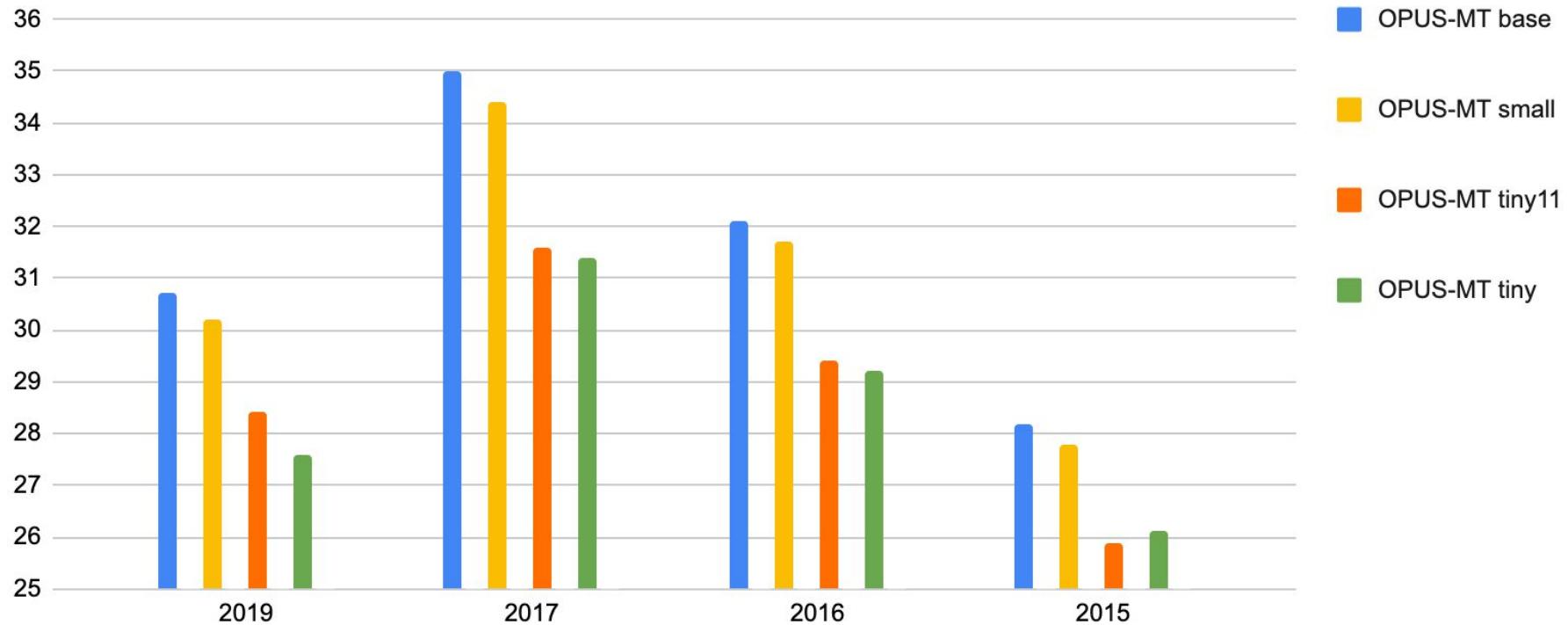
extra slides

base	small	tiny11	tiny
6 x encoder 6 x decoder	6 x encoder 2 x SSRU dec	6 x encoder 2 x SSRU dec emb. 256, ffn 1536	3 x encoder 2 x SSRU dec emb. 256

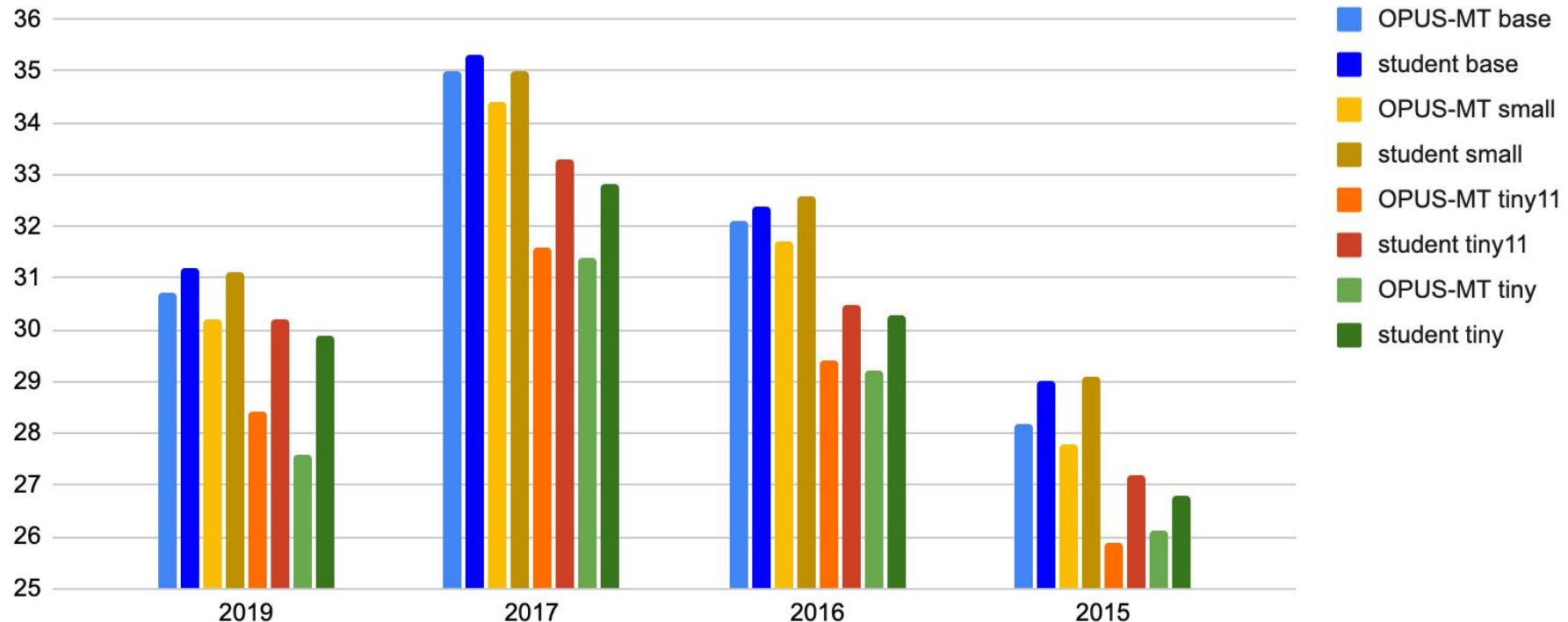
Knowledge Distillation



Performance of different model types (WMT fin-eng)



Performance of different model types (WMT fin-eng)



Size and speed (fin-eng)

model	size original
big	891 MB
base	294 MB
small	226 MB
tiny11	96 MB
tiny	89 MB

Size and speed (fin-eng)

model	size original	size quantized
big	891 MB	224 MB
base	294 MB	74 MB
small	226 MB	57 MB
tiny11	96 MB	25 MB
tiny	89 MB	23 MB

Size and speed (fin-eng)

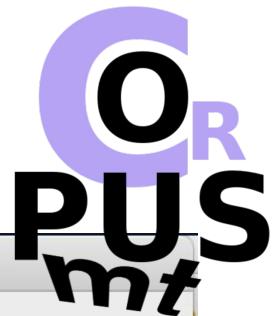
model	size original	size quantized	speed 4 CPUs
big	891 MB	224 MB	
base	294 MB	74 MB	46.36s
small	226 MB	57 MB	24.07s
tiny11	96 MB	25 MB	10.98s
tiny	89 MB	23 MB	9.90s

Tatoeba-test: 10,000 sentences, 48,684 words

Size and speed (fin-eng)

model	size original	size quantized	speed 4 CPUs	speed (+shortlist)
big	891 MB	224 MB		
base	294 MB	74 MB	46.36s	40.47s
small	226 MB	57 MB	24.07s	17.89s
tiny11	96 MB	25 MB	10.98s	7.24s
tiny	89 MB	23 MB	9.90s	6.22s

Tatoeba-test: 10,000 sentences, 48,684 words



OPUS-MT-app - based on translateLocally

The screenshot shows a Mac OS X-style application window titled "translateLocally". The title bar includes standard red, yellow, and green window control buttons. Below the title bar is a toolbar with a dropdown menu set to "English-Finnish tiny". The main area contains two text boxes. The left text box contains the following text:

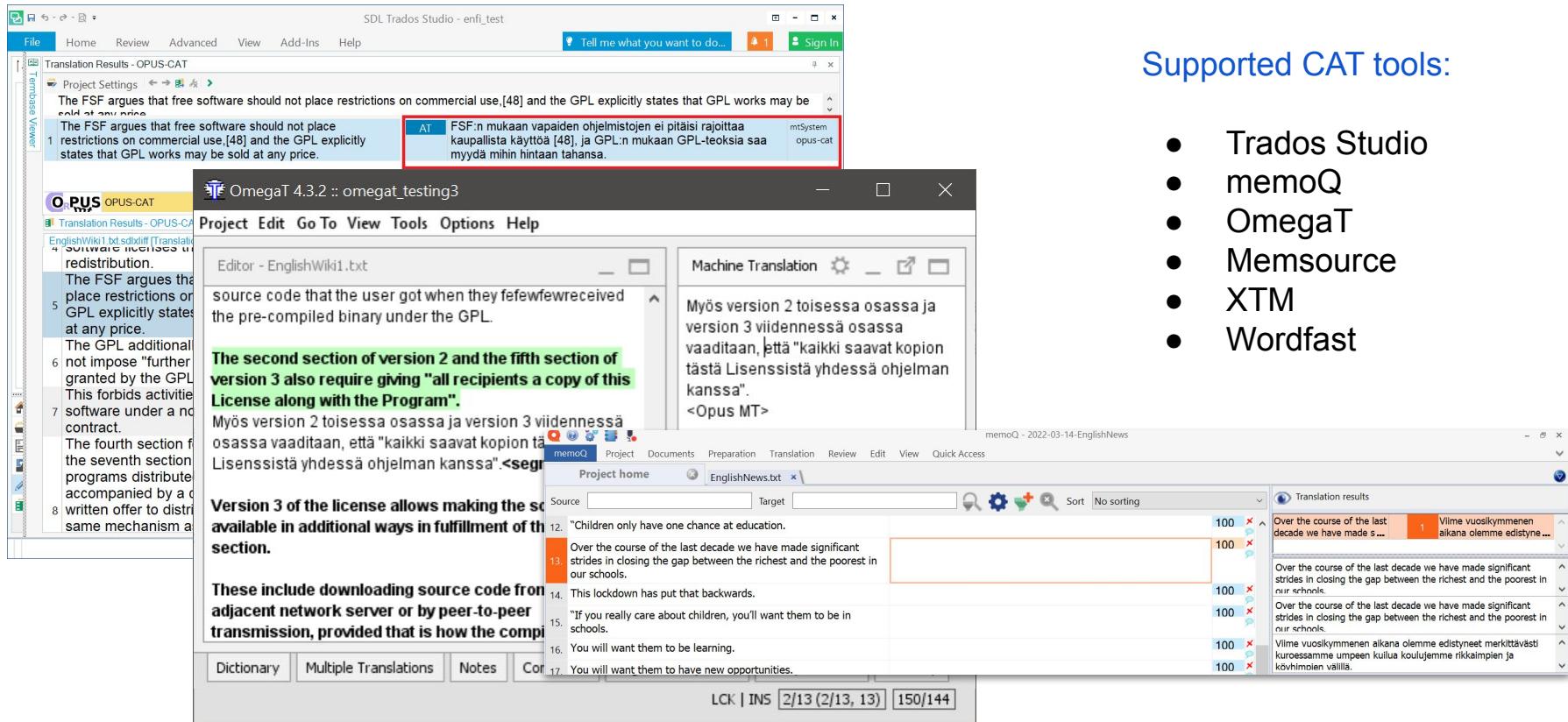
This demo shows how you can translate as you type. It is fast and easy and runs on your local desktop computer. It does not require any internet connection and can be loaded with quantized OPUS-MT models!

The right text box contains the Finnish translation:

Tämä demo näyttää, miten voit kään்�tää kuten kirjoitat. Se on nopea ja helppo ja toimii paikallisessa työpöydän tietokoneessa. Se ei vaadi mitään internet-yhteyttä ja se voidaan ladata kvanttisoiduilla OPUS-MT-malleilla!

At the bottom of the window, a status bar displays the text "Translation speed: 235 words per second."

OPUS-CAT: Support for professional translators



OPUS-CAT: Support for professional translators

The screenshot shows the OPUS-CAT MT Engine interface. The window title is "OPUS-CAT MT Engine v1.1.0.7". The menu bar includes "Models", "Settings", and "Online models". The main area is titled "Downloadable online models" and contains a table with columns: "Source languages", "Target languages", "Model name", and "Installation progress". The table lists various language pairs and model names, such as Argentine Sign L to Spanish (opus-2020-01-15), Afrikaans to German (opus-2020-01-19), and Arabic to English (opus-2019-12-18). A total of 1821 models are listed. At the bottom, it says "Total amount of filtered models: 1821". On the right side of the table, there is a filter panel with checkboxes for "OPUS-MT models", "Tatoeba models", "Newest only", "Bilingual models", and "Multilingual models (usually lower quality than bilingual models)". Below the filter panel is a button labeled "Install model locally".

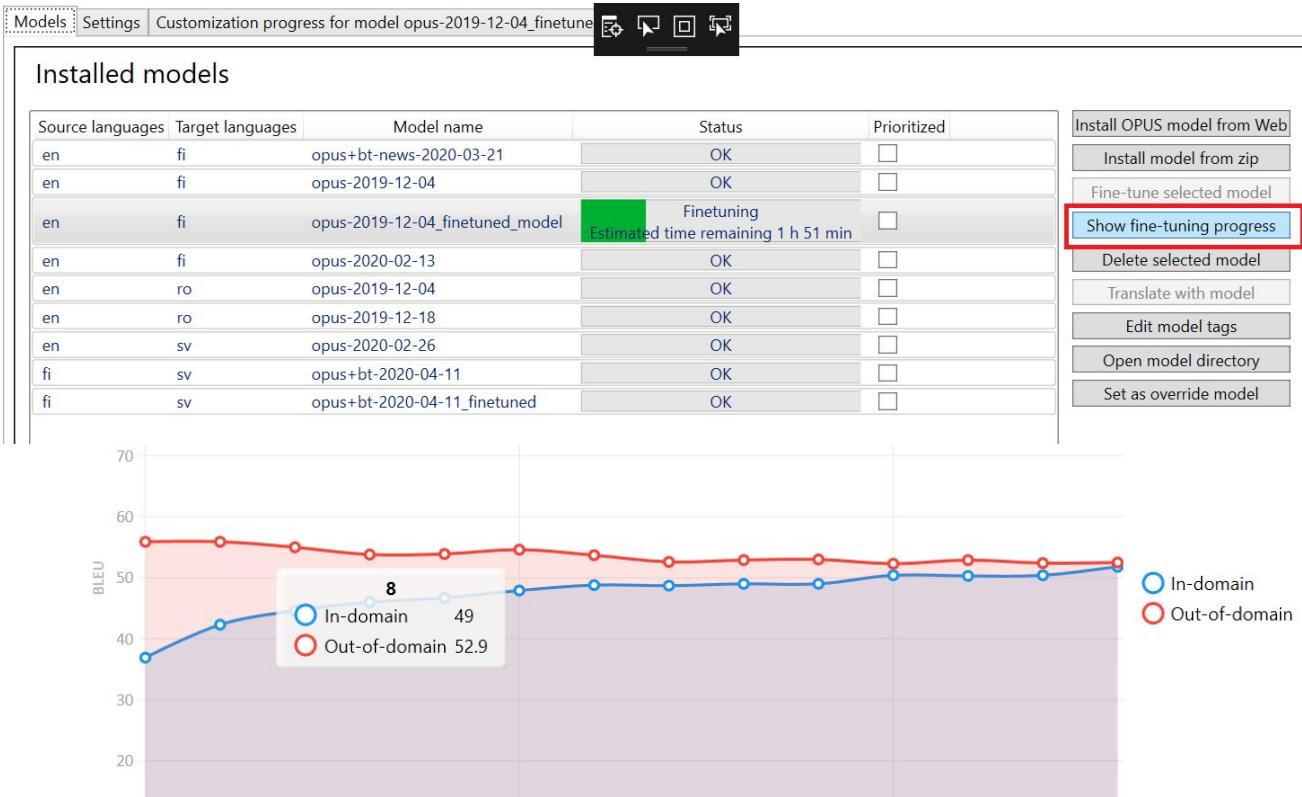
Local MT engine

- Data security
- Low latency
- Off-line mode
- No dependencies on external services

Extensible and open

- All of OPUS-MT
- No costs
- API for easy integration

OPUS-CAT: Support for professional translators



Fine-tuning:

- Personal TMX
- Runs locally
- Any number of fine-tuned models

Optimizes:

- Terminology
- Placeholder tags
- Domain-specific translations

OPUS-CAT: Improving MT usability in professional translation

The screenshot displays the OPUS-CAT interface with three main sections:

- Rule collection This rule lower-cases all but first word:** Shows 'Input to rule collection: Source text' with 'ORGANIZATIONS SUCH AS NATO, EU AND WHO' and 'Edited source text' with 'Organizations such as nato, eu and who'.
- Rule collection This rule collection upper-cases abbreviations:** Shows 'Input to rule collection: Output from This rule lower-cases all but first word' with 'Organizations such as nato, EU and WHO' and '(?i)(nato|eu|who)'. A tooltip 'TOOLTIP SHOWS APPLIED RULE FOR MATCH' points to the input field. 'Edited source text' shows 'Organizations such as NATO, EU and WHO'.
- Rule collection This rule collection upper-cases the output from translation provider:** Shows 'Source text' with 'ORGANIZATIONS SUCH AS NATO, EU AND WHO' and 'Input to rule collection: MT output' with 'Naton, EU:n ja WHO:n kaltaiset järjestöt'. A tooltip 'POSTEDIT RULE USES TRANSLATION PROVIDER OUTPUT AS ITS INPUT' points to the input field. 'Post-edited MT output' shows 'NATON, EU:N JA WHO:N KALTAISET JÄRJESTÖT'.

To the right, a list of aligned segments from a concordance search is shown, illustrating how tags are handled across different parts of the sentence.

Tag handling:

- Tag injection
- Tag restoration based on alignment

Terminology:

- Regex rules for processing MT input/output
- Planned: models with support for term lists

Practical ways of using OPUS-MT models

```
from transformers import pipeline  
pipe = pipeline("translation", model="Helsinki-NLP/opus-mt-en-fi")  
print(pipe('Please, translate this text.'))
```



<http://hf.co/helsinki-nlp>



```
from elg import Service  
service = Service.from_id(4863)  
result = service(request_input="translate this!", request_type="text")
```

```
docker run -p 8888:8888 helsinkinlp/opus-mt-elg-fin-eng:1.0  
curl -X POST -H "Content-Type: application/json" -d  
'{"content":"Translate this!"}' "localhost:8888/elg/translate/en/fi"
```



OPUS Data and Corpus Structure

Corpus Structure

Europarl
Europarl/xml
Europarl/xml/en
Europarl/xml/en/ep-10-07-05-004.xml.gz
Europarl/xml/fr
Europarl/xml/fr/ep-10-07-05-004.xml.gz
...
Europarl/xml/en-fr.xml.gz
...
Europarl/raw
Europarl/raw/en
Europarl/xml/en/ep-10-07-05-004.xml.gz
Europarl/raw/fr
Europarl/xml/fr/ep-10-07-05-004.xml.gz
...

```
<?xml version="1.0" encoding="utf-8"?>
<document>
  <CHAPTER ID="0">
    <P id="1"></P>
    <SPEAKER ID="1" LANGUAGE="DE" NAME="Rübig">
      <P id="2">
        <s id="1">Madam President, I saw a few boats landing at Parliament this
week and notified the security service.</s>
        <s id="2">Not only were there language difficulties; the telephone line was so
poor that it was almost impossible to communicate.</s>
        <s id="3">I would be most obliged if the number on which the security service
can be reached could also be clearly displayed in the House, so that if anyone
wants to report an incident, they can do so quickly and efficiently.</s>
      </P>
    ...
  
```

```
<?xml version="1.0" encoding="utf-8"?>
<!DOCTYPE cesAlign PUBLIC "-//CES//DTD XML cesAlign//EN" "">
<cesAlign version="1.0">
  <linkGrp targType="s" fromDoc="en/ep-00-01-17.xml.gz"
toDoc="fr/ep-00-01-17.xml.gz">
    <link xtargets="1;1" />
    <link xtargets="2;2" />
    <link xtargets="3;3 4" />
```

Sentences

