Democratizing MT

CORPUS

bergamot

OPUS

H P

L T

bergamot

https://hplt-project.org/

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

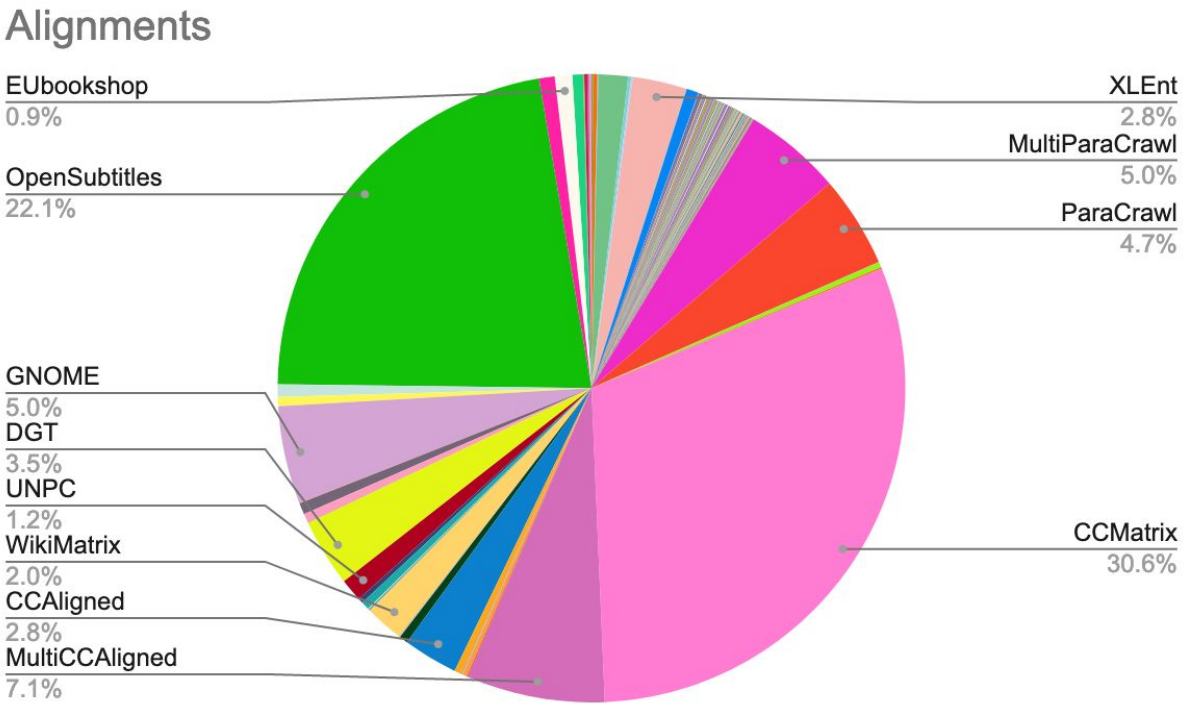
Parallel corpora are useful in a wide variety of research areas, particularly in machine translation and lexicography. However, parallel corpora have been few, often unrepresentative, and not generally available. The aim of the OPUS project is to provide a public collection of parallel corpora which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

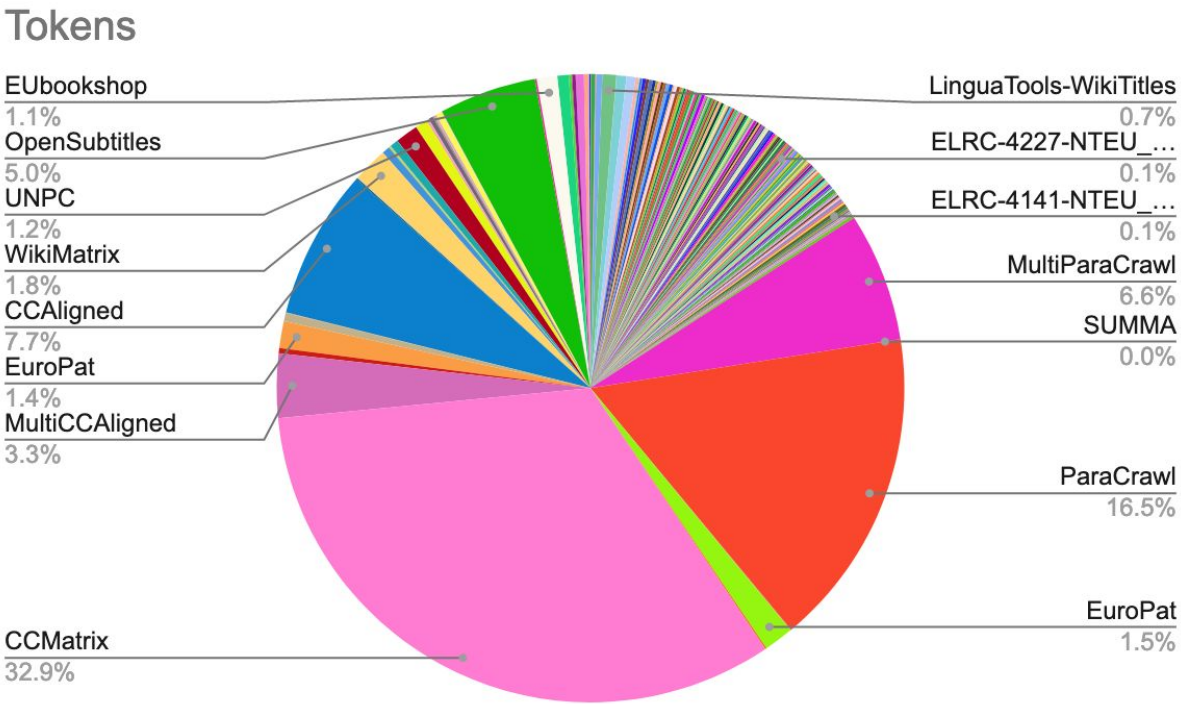
https://opus.nlpl.eu/

- 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
EUbookshop
0.9%
OpenSubtitles
22.1%
GNOME
5.0%
DGT
3.5%
UNPC
1.2%
WikiMatrix
2.0%
CCAligned
2.8%
MultiCCAligned
7.1%
XLEnt
2.8%
MultiParaCrawl
5.0%
ParaCrawl
4.7%
CCMatrix
30.6%


Tokens
EUbookshop
1.1%
OpenSubtitles
5.0%
UNPC
1.2%
WikiMatrix
1.8%
CCAligned
7.7%
EuroPat
1.4%
MultiCCAligned
3.3%
CCMatrix
32.9%
LinguaTools-WikiTitles
0.7%
ELRC-4227-NTEU_...
0.1%
ELRC-4141-NTEU_...
0.1%
MultiParaCrawl
6.6%
SUMMA
0.0%
ParaCrawl
16.5%
EuroPat
1.5%

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

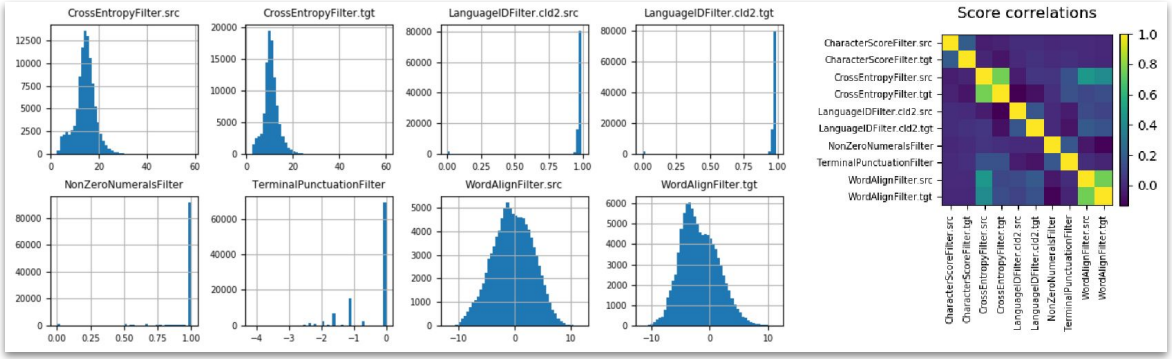
AVAILABLE FILTERS

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

```yaml
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
```

```yaml
  - 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

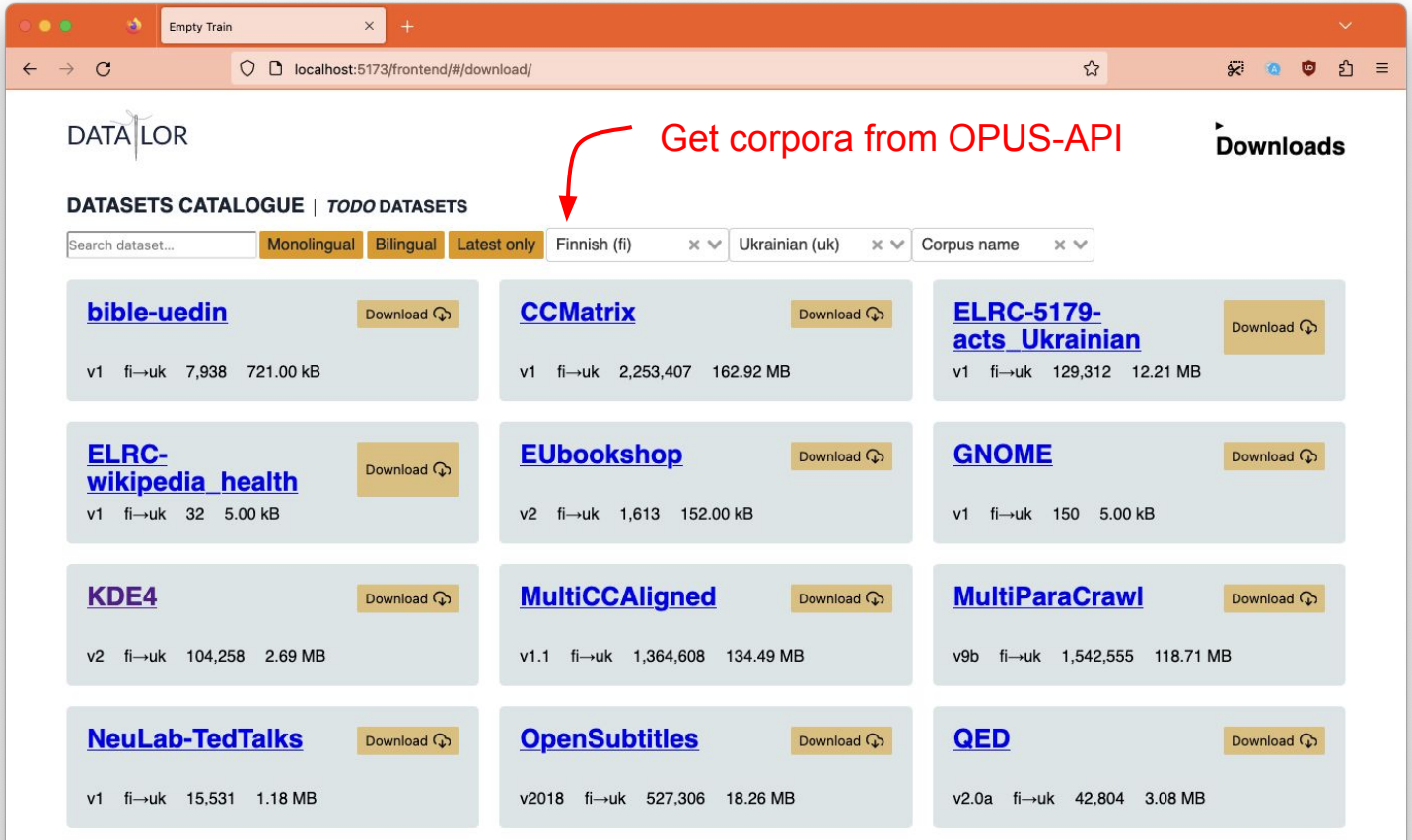
- More rule-based filters!
- Binary classifier
  - Noisy sentences scored near to 0.
  - Cleaner sentences scored near to 1.
- Classifier based on XLM-Roberta.
  - Multilingual models and zero-shot classification also!
- Downloadable trained models for 32 language pairs
  - All the 70 languages of HPLT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

Bicleaner AI: sentence-level scoring

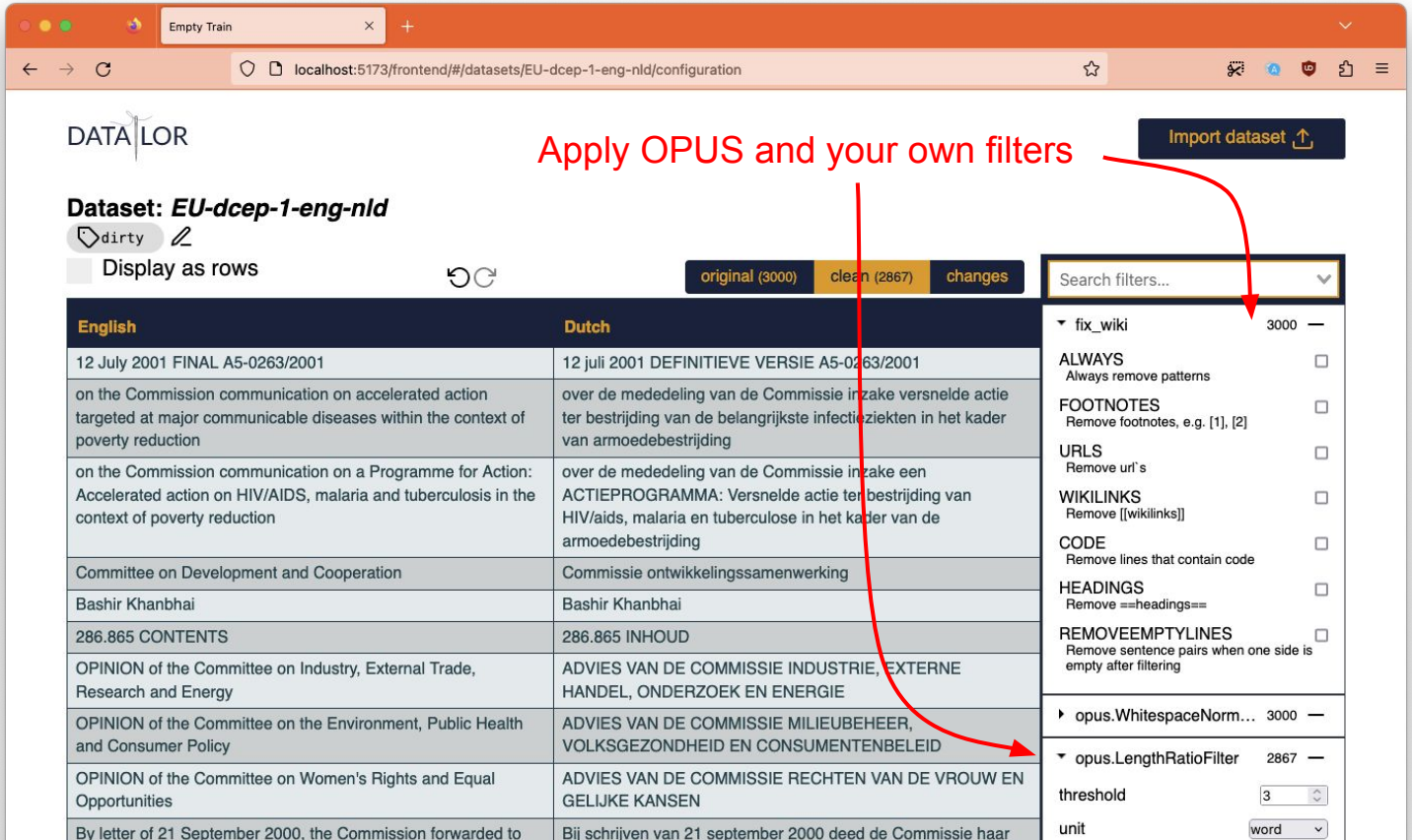
Click Run to start the installation immediately.	Kliknite Pokreni da biste odmah započeli instalaciju.	0.920
The 206L LongRanger is a stretched variant with seating for seven.	206L LongRanger je izdužena inačica s prostorom za smještaj sedam osoba.	0.981
Just go.” he said. it sucks.	Najebali smo. stvarno.."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

Effective data curation

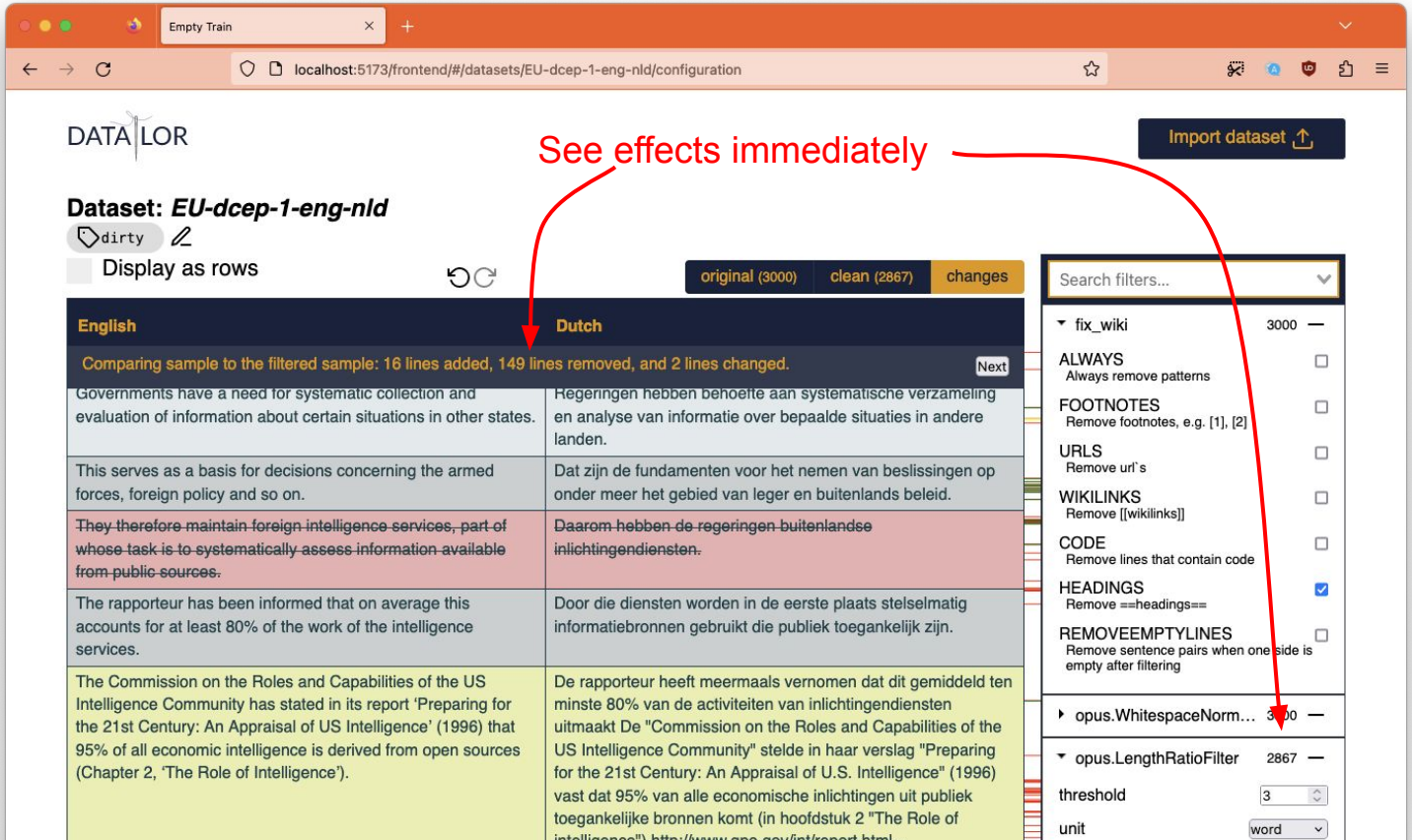
Data curation and cleaning (empty-train)



Data curation and cleaning (empty-train)



Data curation and cleaning (empty-train)



Efficient training pipelines

Efficient training pipelines (empty-train*er*)

```
datasets:
  clean: clean.gz
  medium: medium.gz
  dirty: dirty.gz

stages:
  - start
  - mid
  - end

modifiers:
  - uppercase 0.05
  - titlecase 0.05

seed: 31337
```

```
start:
  - clean 0.9
  - medium 0.1
  - until clean 6

mid:
  - clean 0.6
  - medium 0.3
  - dirty 0.1
  - until medium 1

end:
  - clean 0.4
  - medium 0.3
  - dirty 0.3
  - until dirty inf
```

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 …
```

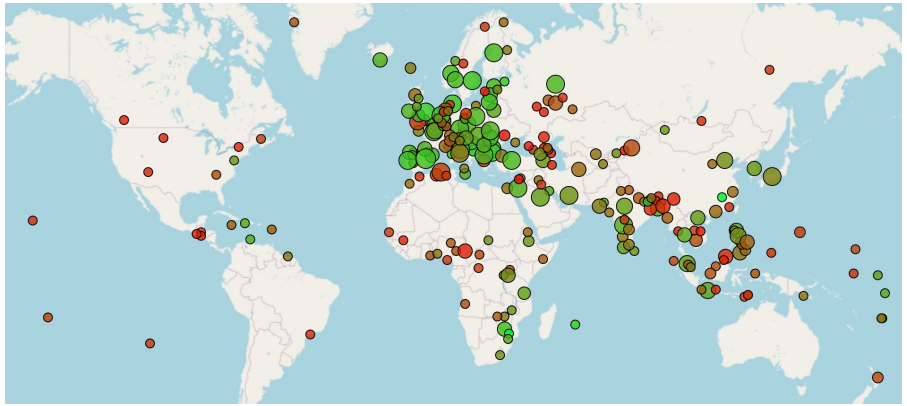
Efficient training pipelines (empty-train*est*)

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



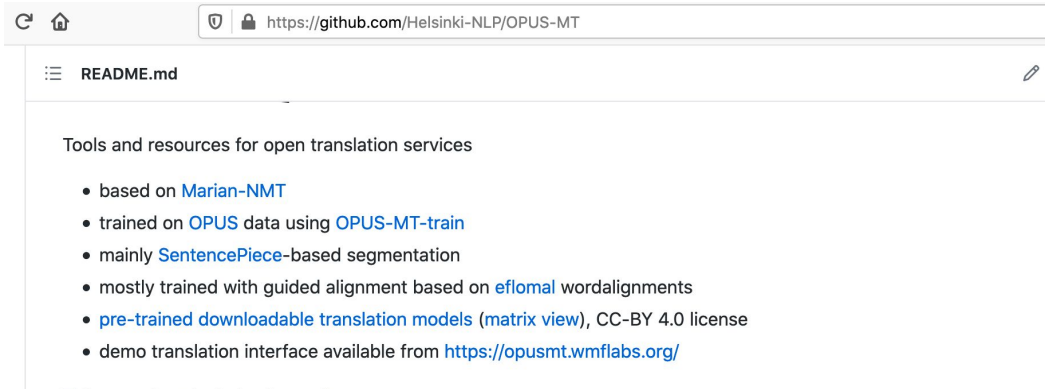
https://github.com/Helsinki-NLP/OPUS-MT

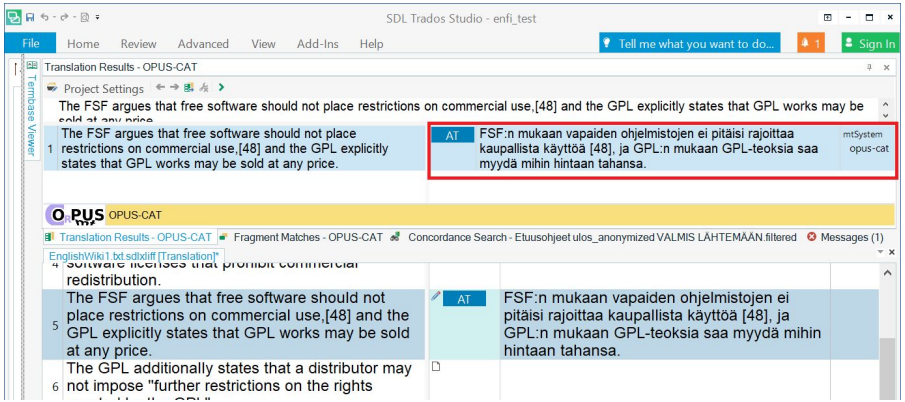
OPUS → OPUS-MT

OpusTools
OpusAPI
OpusFilter

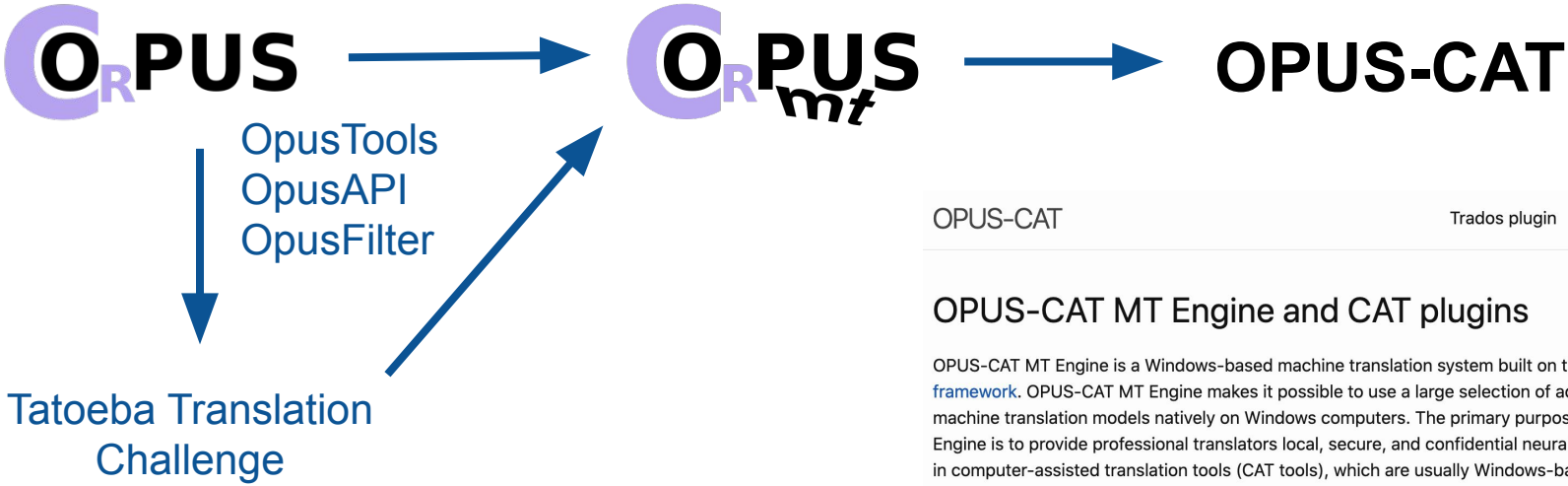
Tatoeba Translation Challenge

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





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



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).

Helsinki-NLP / **opus-mt-tc-big-fi-en** ♡ like 0

Translation | PyTorch | Transformers | en | fi | cc-by-4.0 | marian | text2text-generation | opus-mt-tc | Eval Results | AutoTrain Compatible

Train | Deploy | U

Model card | Files

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

Edit model card

opus-mt-tc-big-fi-en

Downloads last month
990

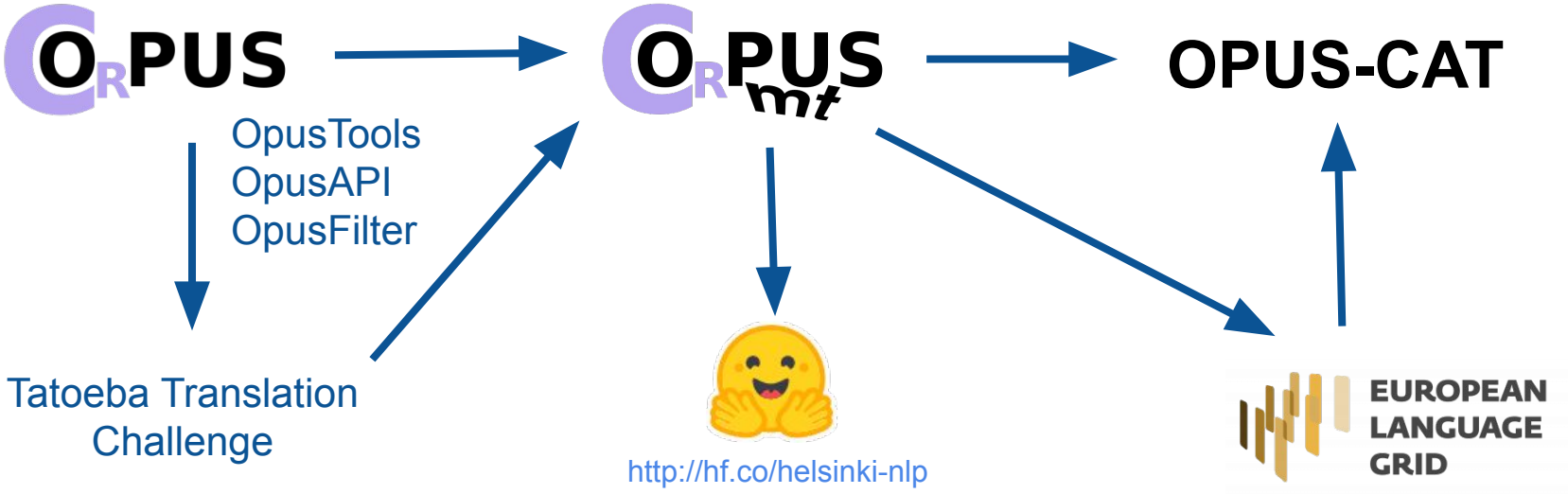
EUROPEAN LANGUAGE GRID
RELEASE 3

HelsinkiNLP - OPUS-MT: English-Ukrainian machine translation

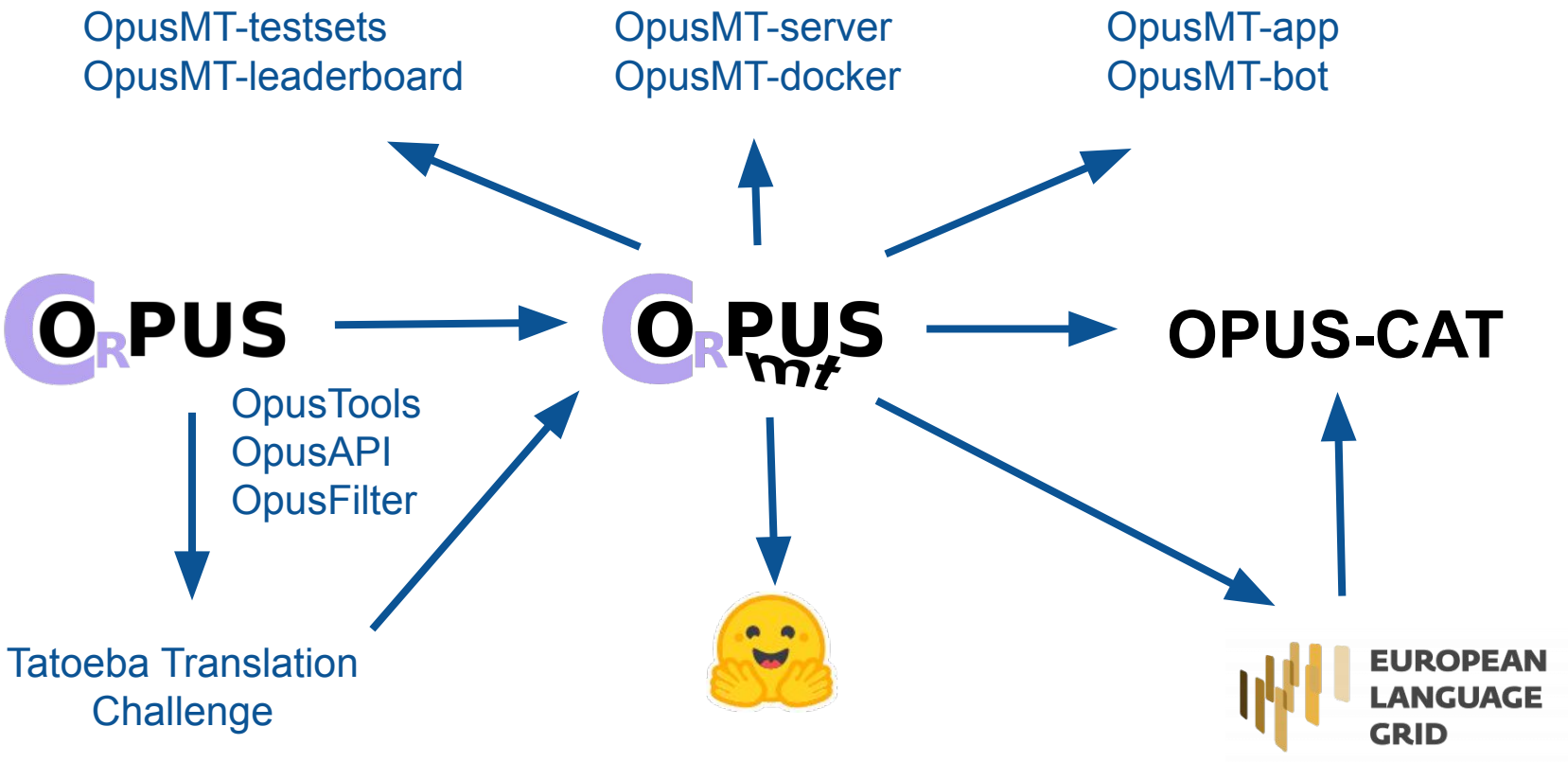
OPUS-MT eng-ukr

Version: 2022.3.13

ELG-compatible service



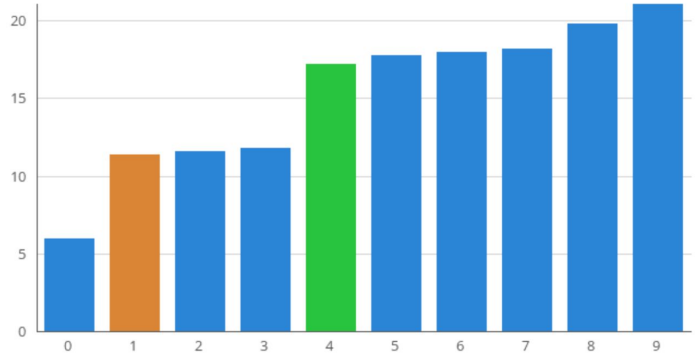
https://live.european-language-grid.eu/


OpusMT-testsets
OpusMT-leaderboard
OpusMT-server
OpusMT-docker
OpusMT-app
OpusMT-bot
OPUS
OPUS mt
OPUS-CAT
OpusTools
OpusAPI
OpusFilter
Tatoeba Translation Challenge
EUROPEAN LANGUAGE GRID

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]



- orange = OPUS-MT models, blue = Tatoeba-MT models, green = compact models

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]

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

Poliisin 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 siirtyy ihmisiin hyttysten kautta.
Хвороба переноситься свинями, а потім через комарів передається до людей.
[-Хвороба переноситься свинями,-]
{+Захворюють свині,+} від яких [-потім передається людині-] {+вони передаються людям+} через комарів.
```

https://opus.nlpl.eu/leaderboard/

OPUS-MT and Ukrainian LT

Ukrainian - English

Finnish - Russian

English-centric data

English → Finnish MT

Russian → Ukrainian MT

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

Finnish → Ukrainian data

Ukrainian-to-Finnish MT

Ukrainian Wikipedia

https://github.com/Helsinki-NLP/UkrainianLT/

Democratizing NLP with

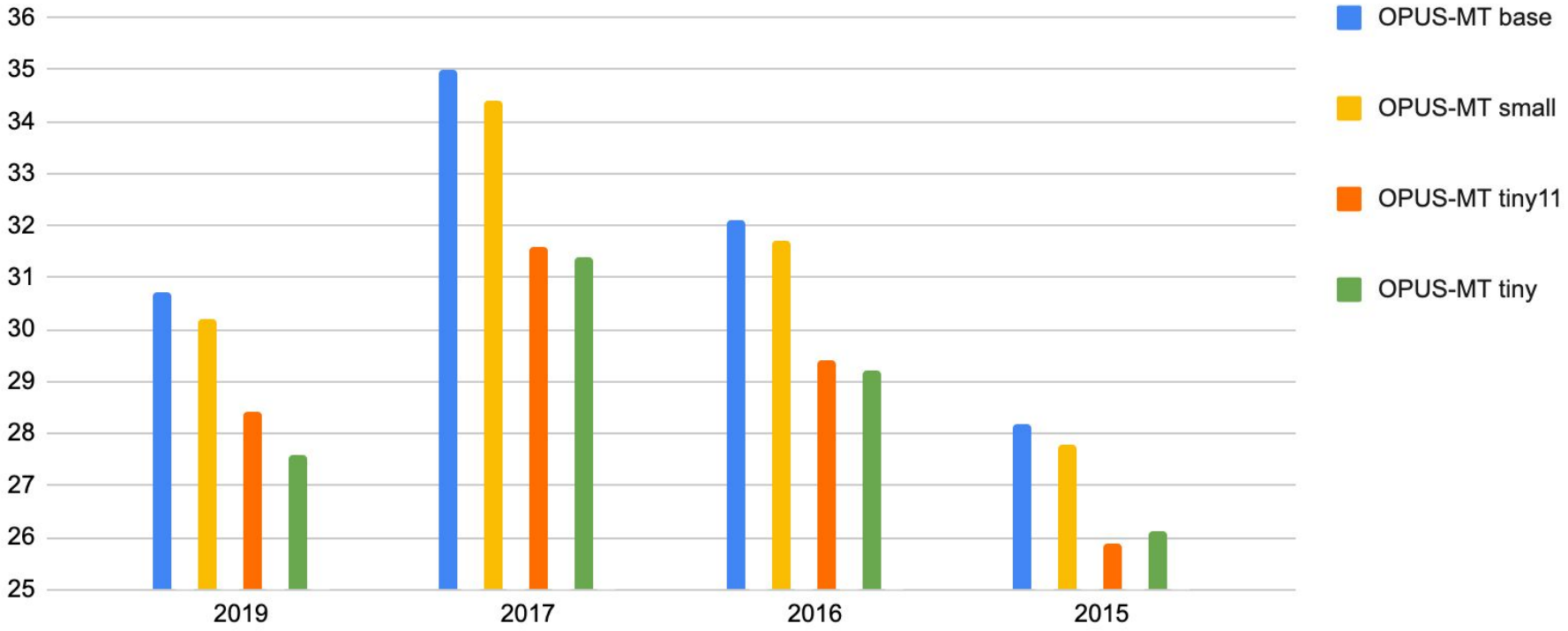
extra slides

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

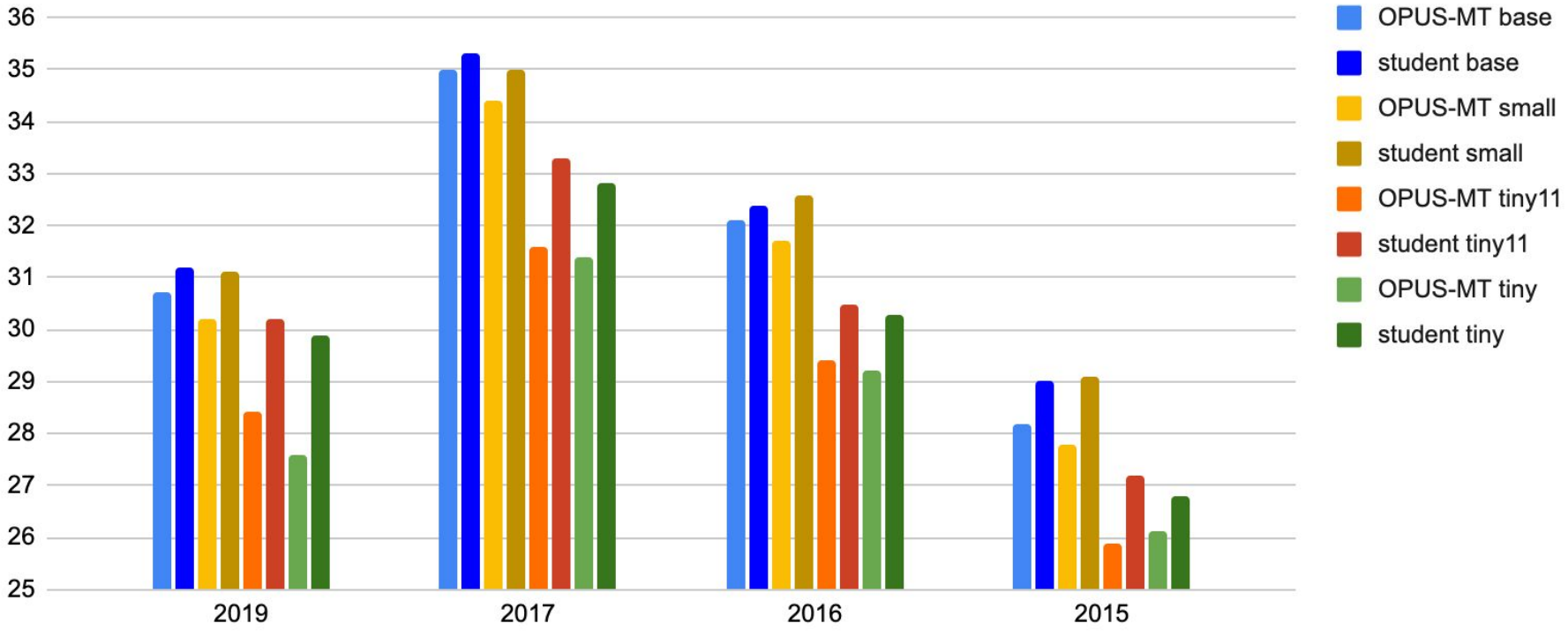
Knowledge Distillation

Simpler Simple Recurrent Units (SSRU; Kim et al., 2019)

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**

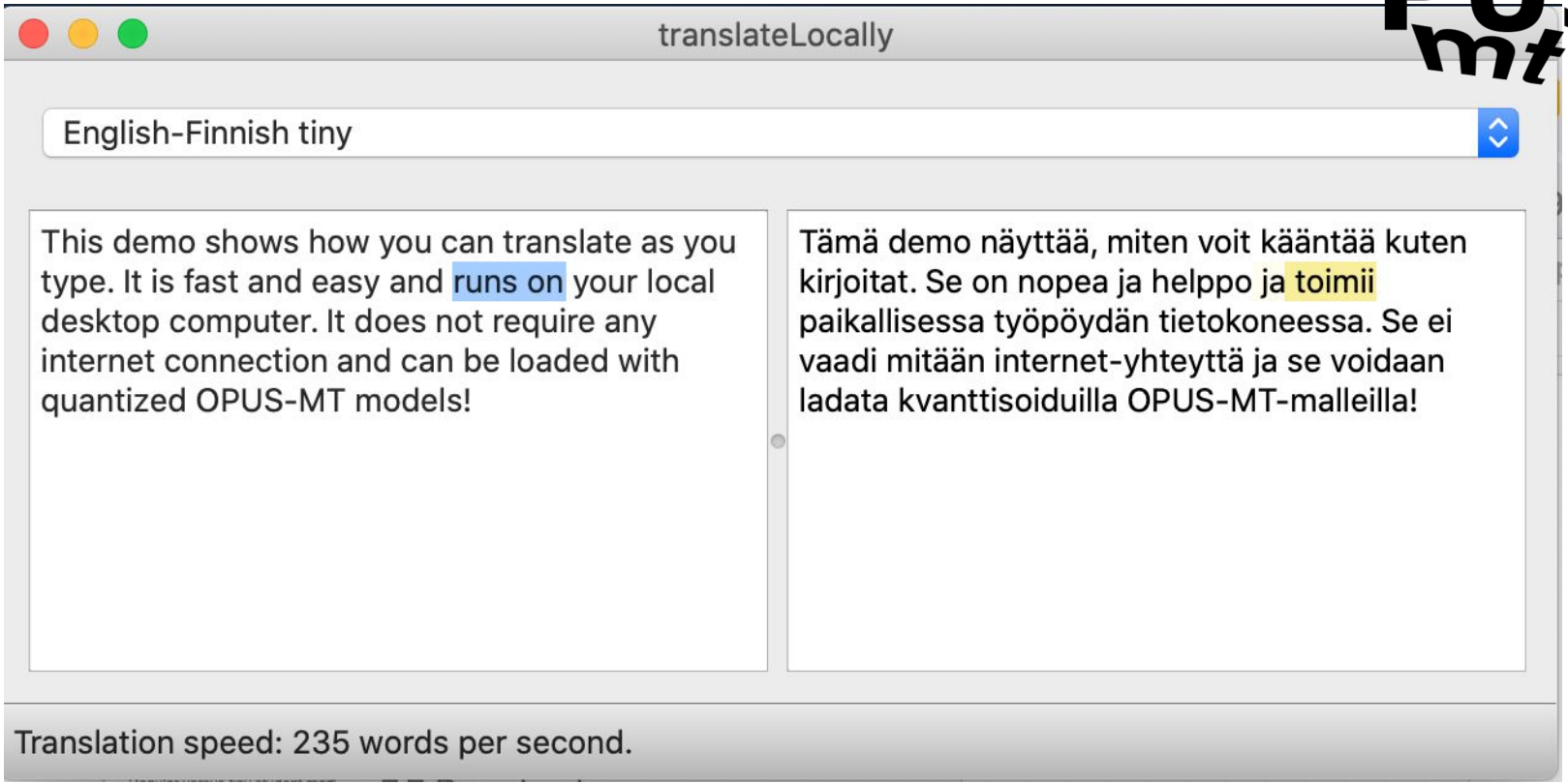
Tatoeba-test: 10,000 sentences, 48,684 words

Size and speed (fin-eng)

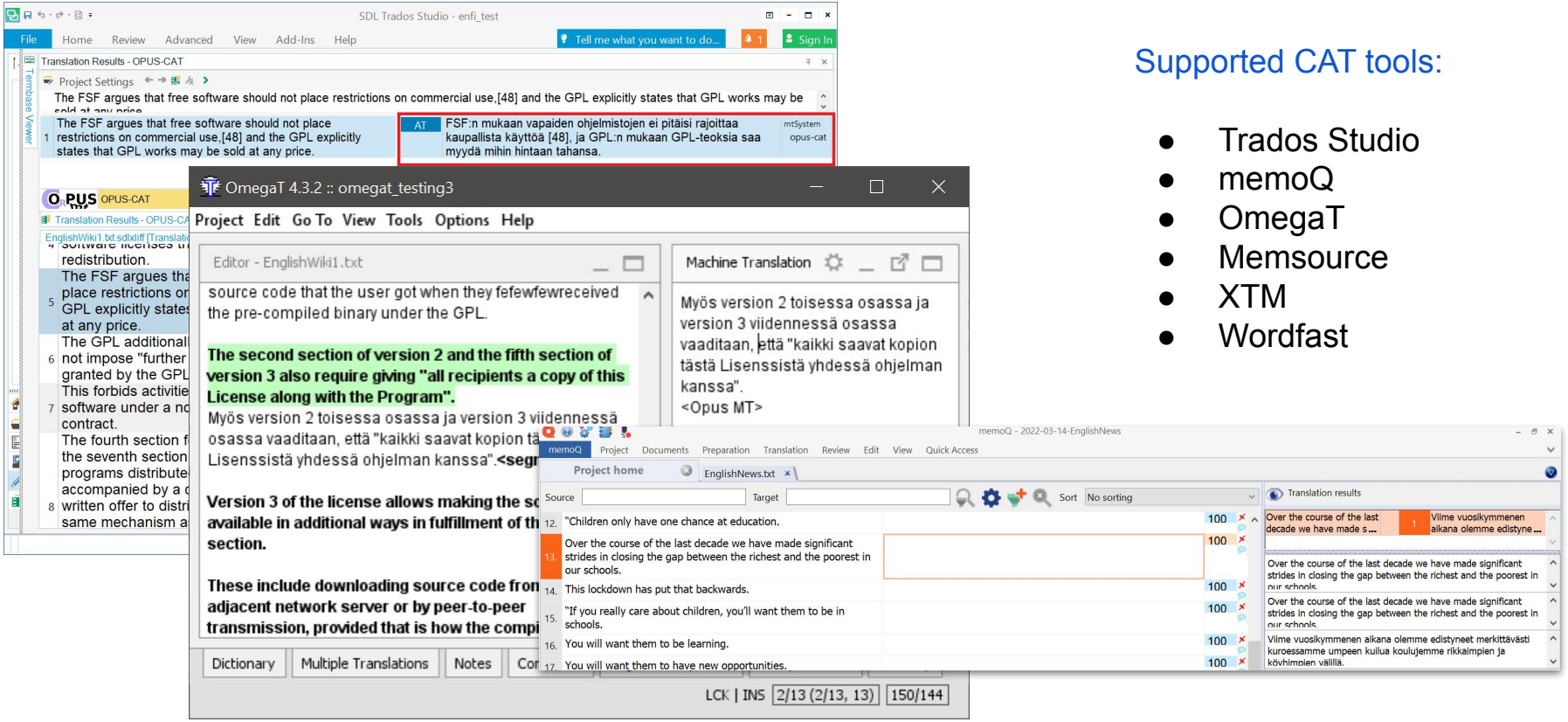
model	size original	size quantized	speed 4 CPUs	speed (+shortlist)
big	891 MB	224 MB		
base	294 MB	74 MB	46.36s	40.47s
small	226 MB	57 MB	24.07s	17.89s
tiny11	96 MB	25 MB	10.98s	**7.24s**
tiny	89 MB	23 MB	9.90s	**6.22s**

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

OPUS-MT-app - based on translateLocally



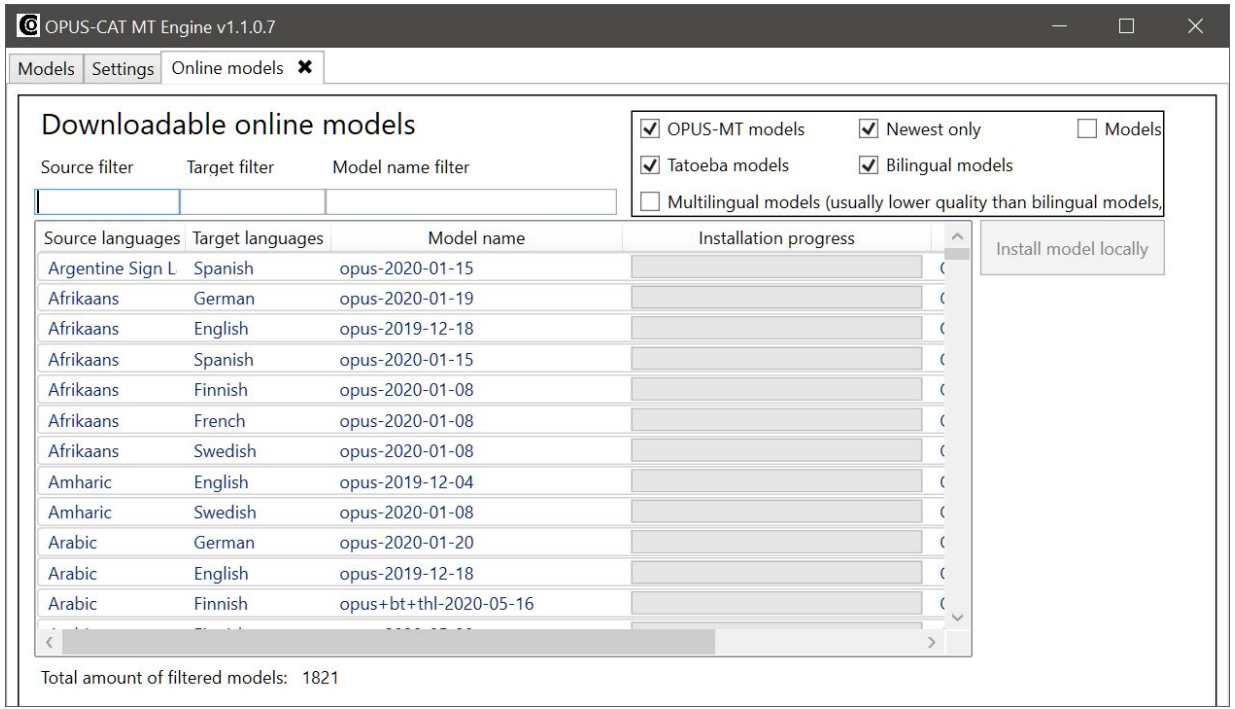
OPUS-CAT: Support for professional translators



Supported CAT tools:

- Trados Studio
- memoQ
- OmegaT
- Memsource
- XTM
- Wordfast

OPUS-CAT: Support for professional translators



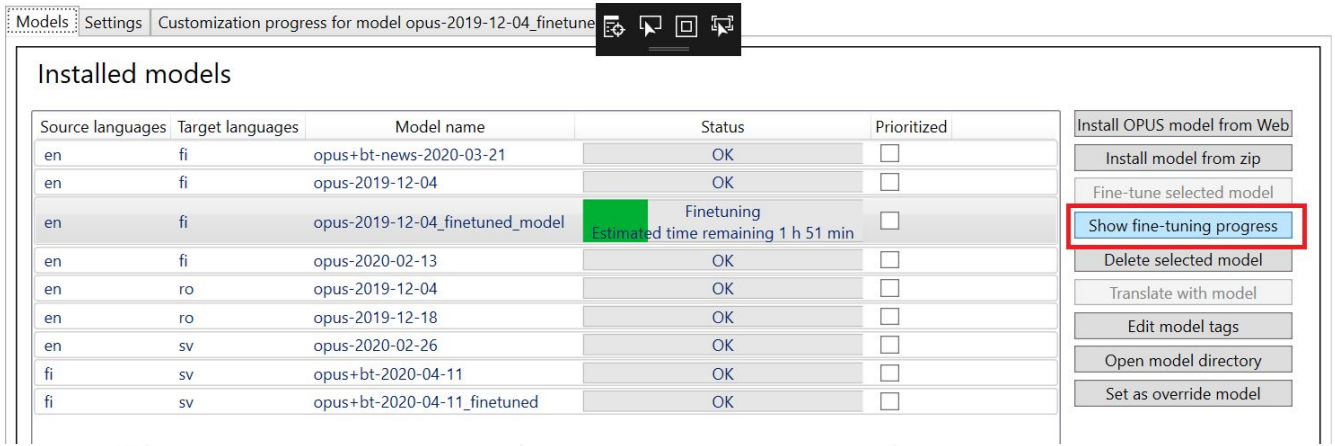
Local MT engine

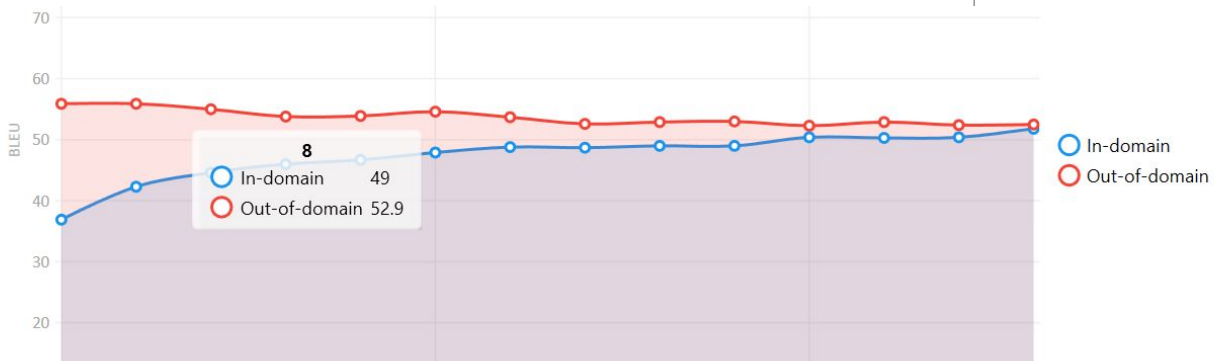
- Data security
- Low latency
- Off-line mode
- No dependencies on external services

Extensible and open

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

OPUS-CAT: Support for professional translators





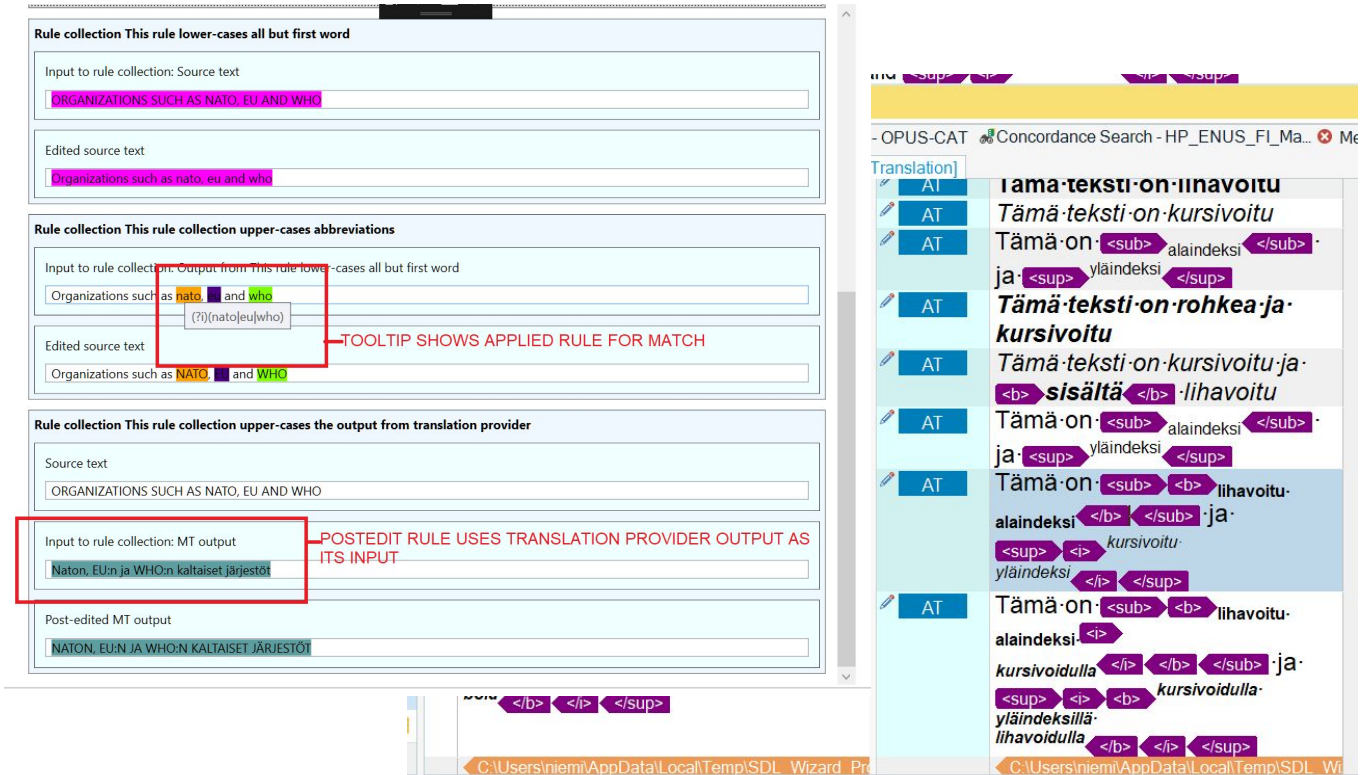
Fine-tuning:

- 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



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

EUROPEAN LANGUAGE GRID

```
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"
```

docker hub

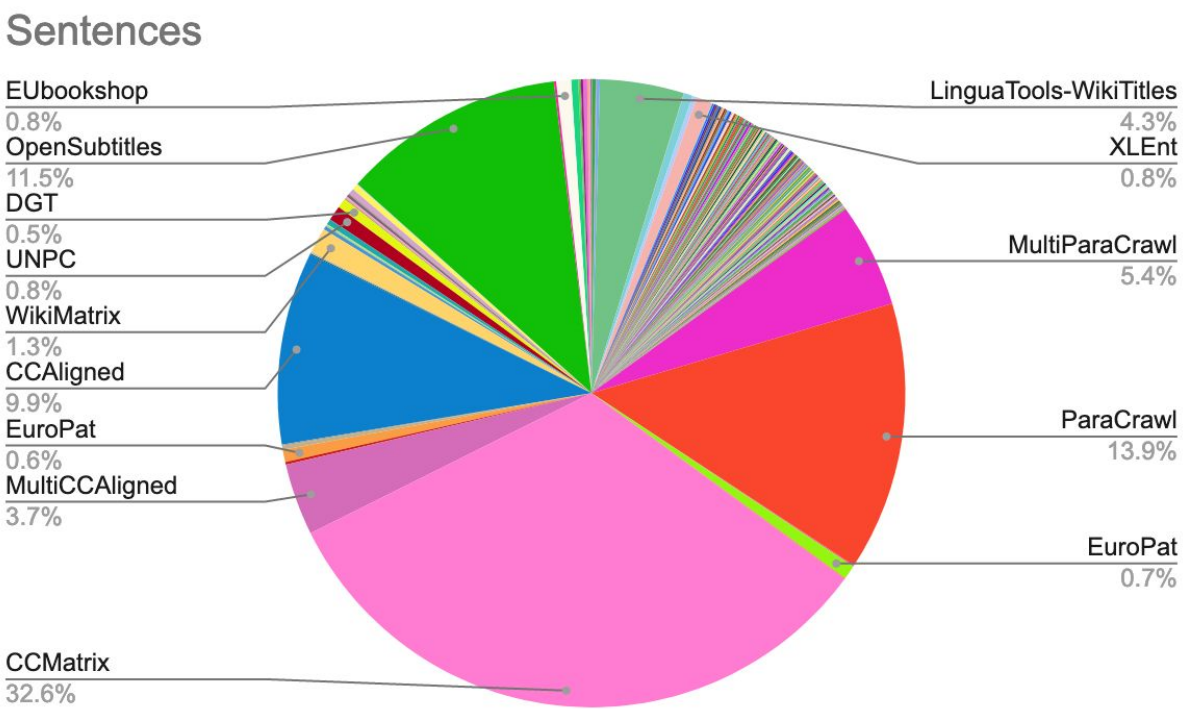
OPUS Data and Corpus Structure

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
EUbookshop
0.8%
OpenSubtitles
11.5%
DGT
0.5%
UNPC
0.8%
WikiMatrix
1.3%
CCAligned
9.9%
EuroPat
0.6%
MultiCCAligned
3.7%
CCMatrix
32.6%
LinguaTools-WikiTitles
4.3%
XLEnt
0.8%
MultiParaCrawl
5.4%
ParaCrawl
13.9%
EuroPat
0.7%