OpenSubtitles2024: A Massively Parallel Dataset of Movie Subtitles for MT Development and Evaluation

Jörg Tiedemann, Hengyu Luo
University of Helsinki
Helsinki, Finland
{firstname.lastname}@helsinki.fi

Abstract

This paper introduces OpenSubtitles2024, a massively parallel dataset compiled from translated subtitles. The collection includes an extensive collection of aligned training data based on user-contributed subtitles derived from OpenSubtitles.org and a dedicated held-out dataset for development and evaluation of machine translation and multilingual language models. The collection provides an increased language coverage and doubles the size of the previous edition. Furthermore, a careful procedure was applied to reserve a subset of the most recent subtitles for system development and evaluation. The collection covers 92 languages and language variants, aligned in over 3,000 bitexts containing 40 billion tokens in 7.7 million subtitle files. The test set comprises 2,022 language pairs. In addition, we also provide a multi-parallel test set that refers to a subset of the held-out data with synchronized alignments across 40 languages and 15 subtitles.

Keywords: massively parallel corpus, machine translation, alignment

1. Introduction

Modern NLP is driven by large datasets and many languages are still lacking far behind resources in English and a handful of other high-resource languages. A long-standing challenge in the field is the multilingual support of language technology to better reflect the linguistic diversity in the world. This challenge includes support for translation between languages, and parallel datasets that provide aligned signals are still essential to enable a reliable translation performance. However, while high-quality corpora exist for formal domains, such as religious and legal texts or parliamentary proceedings (Christodouloupoulos and Steedman, 2015; Steinberger et al., 2006; Koehn, 2005), and large-scale (but often noisy) datasets have been mined from the web (Bañón et al., 2020; Schwenk et al., 2021), translated data for general, conversational domains is difficult to obtain and many language pairs have little or no support. Furthermore, most public datasets are English-centric and do not include other important translation directions.

In addition to training data, there is also a lack of multilingual benchmarks and even in the field of machine translation, the availability of test sets with significant language coverage is limited (Xu et al., 2025). Yet another problem is the re-use of established benchmarks that leads to gradual overfitting when models are tuned to perform well on those benchmarks. Furthermore, there is a substantial risk of contamination when large NLP models are trained on ever-growing web-crawled data collections that are difficult to control and clean (Balloccu et al., 2024; Sainz et al., 2023).

The OpenSubtitles2024 dataset contributes to efforts that tackle these challenges in two ways:

1. It provides an extended collection of 7.7 million aligned movie subtitles with a large language coverage (92 languages and language variants) that can be used to train machine translation and multilingual language models. Specifically, this collection builds upon and significantly expands previous subtitle-based corpora such as OpenSubtitles2018 (Lison and Tiedemann, 2016).

2. It includes carefully selected and protected held-out data that serve as a disjoint development set and comprehensive benchmark covering 2,022 language pairs including a large number of non-English translation tasks.

Training data is published in the public repository OPUS[1] using the standard formats used in this extensive collection. With this, integration into training procedures is straightforward and can utilize tools and libraries that are developed to access and work with parallel data from the entire collection.

The held-out data is published separately on GitHub[2] using password protected packages to reduce the risk of contamination through crawled datasets. Furthermore, we integrated the benchmarks in GlotEval (Luo et al., 2025) and LM-Evaluation-Harness (Gao et al., 2024) to make the multilingual test sets available to LLM developers and researchers.

[1] https://opus.nlpl.eu/datasets/OpenSubtitles

[2] https://github.com/Helsinki-NLP/OpenSubtitles-devtest

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 8897–8907
11–16 May 2026.

Details about the data are provided in the following section. After that, we will discuss the procedures applied to compile the collection, including information about the selection and preparation of the development and test set partitions. We will then provide some example use cases to demonstrate the application of the data and conclude with final remarks and directions for future work.

2. Dataset Overview

Scale and Composition: Similar to the OpenSubtitles2018 (Lison and Tiedemann, 2016) corpus (the largest open corpus of aligned subtitles so far), OpenSubtitles2024 is compiled from community-contributed movie and television subtitles sourced from `OpenSubtitles.org`. Compared to the former, this release increases the language coverage from 63 to 92 languages (see Table 1 for a detailed list) and more or less doubles the number of language pairs included in the training data. The size is also substantially increased, extending previous datasets from the same source by 2 billion sentences overall. This affects all languages, and does not only refer to more recent uploads to the subtitle provider, but also includes a significant number of files that have been missed out in the existing collection available through OpenSubtitles2018.

OpenSubtitles2024 leaves out subtitles that are known to be generated. OpenSubtitles.org supports generated data and keeps metadata about such content. Our release systematically excludes such files. The official training data contains alignments for each movie and TV episode in the same format as the previous release to support compatible pipelines. The original collection includes a large number of repeated entries provided by different users. The selection procedure is based on the alignment density scores proposed by Lison and Tiedemann (2016). This ensures that we avoid unnecessary duplication and align each movie only once per language pair. Altogether, the parallel training collection comprises 3,377 bitexts (language pairs) and is by now the biggest parallel corpus in the OPUS collection that hosts the data.

We also provide the alternative alignments covering all combinations of subtitle files to make it possible to use the entire dataset with all variants and their links. This is supported by the monolingual subtitle compilations for each language that contain all files available in the database dump provided by OpenSubtitles.org. Altogether, the release comprises 7.7 million subtitle files with over 40 billion tokens and 5.3 billion sentences and sentence fragments in total, which are aligned in various ways to corresponding files in other languages.

Additionally, we also provide intra-lingual alignments between alternative subtitle files in the same

Lang. Code	ISO 639-3_ ISO 15924	Language Name	train	test	multi
af	afr_Latn	Afrikaans	✓	✗	✗
am	amh_Ethi	Amharic	✓	✗	✗
an	arg_Latn	Aragonese	✓	✗	✗
ar	ara_Arab	Arabic	✓	✓	✓
as	asm_Beng	Assamese	✓	✗	✗
ast	ast_Latn	Asturian	✓	✗	✗
az	aze_Arab	Azerbaijani	✓	✓	✗
be	bel_Cyrl	Belarusian	✓	✓	✗
bg	bul_Cyrl	Bulgarian	✓	✓	✓
bn	ben_Beng	Bengali	✓	✓	✗
br	bre_Latn	Breton	✓	✗	✗
bs	bos_Latn	Bosnian	✓	✓	✗
ca	cat_Latn	Catalan	✓	✓	✗
cs	ces_Latn	Czech	✓	✓	✓
cy	cym_Latn	Welsh	✓	✗	✗
da	dan_Latn	Danish	✓	✓	✓
de	deu_Latn	German	✓	✓	✓
el	ell_Grek	Greek	✓	✓	✓
en	eng_Latn	English	✓	✓	✓
eo	epo_Latn	Esperanto	✓	✗	✗
es	spa_Latn	Spanish	✓	✓	✓
es_419	spa_Latn	Spanish (Latin America)	✓	✓	✗
es_ES	spa_Latn	Spanish (Spain)	✓	✓	✗
et	est_Latn	Estonian	✓	✓	✓
eu	eus_Latn	Basque	✓	✓	✗
fa	fas_Arab	Persian	✓	✓	✓
fi	fin_Latn	Finnish	✓	✓	✓
fr	fra_Latn	French	✓	✓	✓
ga	gle_Latn	Irish	✓	✗	✗
gd	gla_Latn	Scottish Gaelic	✓	✗	✗
gl	glg_Latn	Galician	✓	✓	✗
he	heb_Hebr	Hebrew	✓	✓	✓
hi	hin_Deva	Hindi	✓	✓	✓
hr	hrv_Latn	Croatian	✓	✓	✓
hu	hun_Latn	Hungarian	✓	✓	✓
hy	hye_Armn	Armenian	✓	✓	✗
id	ind_Latn	Indonesian	✓	✓	✓
ig	ibo_Latn	Igbo	✓	✗	✗
is	isl_Latn	Icelandic	✓	✓	✗
it	ita_Latn	Italian	✓	✓	✓
ja	jpn_Jpan	Japanese	✓	✗	✗
ka	kat_Geor	Georgian	✓	✓	✗
kk	kaz_Cyrl	Kazakh	✓	✓	✗
km	khm_Khmr	Khmer	✓	✓	✗
kn	kan_Knda	Kannada	✓	✓	✗
ko	kor_Kore	Korean	✓	✓	✓
ku	ckb_Arab	Kurdish (Sorani)	✓	✓	✗
lb	ltz_Latn	Luxembourgish	✓	✗	✗
lt	lit_Latn	Lithuanian	✓	✓	✓
lv	lav_Latn	Latvian	✓	✓	✓
mk	mkd_Cyrl	Macedonian	✓	✓	✗
ml	mal_Mlym	Malayalam	✓	✓	✗
mn	mon_Cyrl	Mongolian	✓	✓	✗
mr	mar_Deva	Marathi	✓	✗	✗
ms	msa_Latn	Malay	✓	✓	✓
my	mya_Mymr	Burmese	✓	✓	✗
ne	nep_Deva	Nepali	✓	✗	✗
nl	nld_Latn	Dutch	✓	✓	✓
nn	nno_Latn	Norwegian Nynorsk	✓	✗	✗
no	nor_Latn	Norwegian	✓	✓	✓
oc	oci_Latn	Occitan	✓	✗	✗
or	ori_Orya	Odia	✓	✗	✗
pl	pol_Latn	Polish	✓	✓	✓
ps	pus_Arab	Pashto	✓	✓	✗
pt	por_Latn	Portuguese	✓	✓	✓
pt_BR	por_Latn	Portuguese (Brazil)	✓	✓	✓
pt_MZ	por_Latn	Portuguese (Mozambique)	✓	✓	✗
ro	ron_Latn	Romanian	✓	✓	✓
ru	rus_Cyrl	Russian	✓	✓	✓
sd	snd_Arab	Sindhi	✓	✗	✗
se	sme_Latn	Northern Sami	✓	✗	✗
si	sin_Sinh	Sinhala	✓	✓	✗
sk	slk_Latn	Slovak	✓	✓	✓
sl	slv_Latn	Slovenian	✓	✓	✓
so	som_Latn	Somali	✓	✗	✗
sq	sqi_Latn	Albanian	✓	✓	✗
sr	srp_Latn	Serbian	✓	✓	✓
sv	swe_Latn	Swedish	✓	✓	✓
sw	swa_Latn	Swahili	✓	✓	✗
ta	tam_Taml	Tamil	✓	✓	✓
te	tel_Telu	Telugu	✓	✓	✓
th	tha_Thai	Thai	✓	✗	✗
tl	tgl_Latn	Tagalog	✓	✓	✗
tr	tur_Latn	Turkish	✓	✓	✓
tt	tat_Cyrl	Tatar	✓	✓	✗
uk	ukr_Cyrl	Ukrainian	✓	✓	✓
ur	urd_Arab	Urdu	✓	✓	✗
uz	uzb_Latn	Uzbek	✓	✓	✗
vi	vie_Latn	Vietnamese	✓	✓	✓
yue	yue_Hant	Cantonese	✓	✓	✗
zh_CN	zho_Hans	Chinese (Simplified)	✓	✓	✓
zh_TW	zho_Hant	Chinese (Traditional)	✓	✓	✓

Table 1: Languages in training data (*train*), bilingual test sets (*test*), and multi-parallel test data (*multi*).

language. This enables the identification of paraphrases, spelling variations and potential errors. The procedure for categorizing intra-lingual alignments is similar to (Tiedemann, 2016).

Data Structure: The collection is organized with unique file paths specifying the language, the release year, and the corresponding movie / TV episode to make it possible to filter the data in various ways. Metadata is also included in the XML-based release, containing information that is available from OpenSubtitles.org (e.g. user ratings). Bitext alignments are stored as standoff annotation providing links between sentence IDs. This makes it highly flexible to also store alternative alignments over all language combinations without repeating the actual content that has been linked. Furthermore, those files include alignment scores that can be used for filtering and data selection. Figure 1 illustrates the encoding in the alignments files.

Besides the native XML-based release, we also provide derived data for convenience of integration into typical pipelines and workflows including monolingual plain text files, aligned plain text files (referred to as "Moses" format), and Translation Memory eXchange (TMX) files. All packages can be downloaded per language and language pair from the hosting website and through a dedicated API and Python library.

Preprocessing and Alignment: The compilation pipeline begins with subtitle normalization, sentence segmentation, and language identification. The pipeline follows the procedures outlined in Lison and Tiedemann (2016) with various adjustments and refinements, e.g., to accommodate the increased language coverage. In our work, we switched to a more reliable language identification model with better coverage.[3] Additional improvements come from refinements in the detection of character encoding, a more consistent naming strategy of individual TV series episodes, and other preprocessing steps.

At its core, the procedure aligns sentences from different language versions of the same movie or TV episode by leveraging time-overlap signals from subtitle timestamps. The alignment tool computes a score for each potential link and uses a dynamic programming algorithm to find the alignment that maximizes the overlap of linked segments. The procedure is based on Lison and Tiedemann (2016) and we use an overlap score threshold of 0.5 to remove unreliable links. Furthermore, we exclude subtitle pairs that exhibit an overall length ratio below 0.75 (in terms of duration based on the existing time information). This is important to exclude erroneously linked subtitle files or corrupted uploads that provide incomplete information – a common problem with noisy user-contributed data.

Issues with unreliable time information are mitigated using the synchronization algorithm outlined in Tiedemann (2008). The necessary anchor points for synchronization are detected using bilingual dictionaries that are extracted from automatic word alignments derived from OpenSubtitles2018. To avoid spurious signals, we exclude words that are shorter than five characters. We use a maximum of 10 matched pairs of lexical translations from the beginning and from the end of the movie to test potential improvements through re-synchronization. This anchor-based re-alignment is tested exhaustively, and the alignment with the best overlap score overall is kept at the end. Synchronization issues appear surprisingly often, and the procedure outlined above, despite being quite expensive, is necessary to keep the quality on an acceptable level.

The collection also includes large amounts of uploads that cover the same movie or TV series. Selecting among alternative subtitle files when aligning across languages is done based on alignment density scores. Those scores provide the ratio of non-empty sentence alignments over the total number of alignments in a document pair. This metric provides a reasonable heuristics to choose subtitle pairs that match well and do not leave a lot of chronologically non-overlapping text elements that do not align easily. For convenience, those alignment density scores are also provided in the XCES Align files for each document pair (see Figure 1), allowing further filtering if necessary. The scores are also available for the complete release of alternative subtitle alignments, which we provide in a separate download package.

Access and Tooling: The training data is distributed as zipped XML files with one package per language. Alignments come in the form of compressed XCES Align files that cover the links from all document pairs for a specific language pair. We provide convenient tools to retrieve and process the aligned data in order to integrate subsets of the collection in workflows and pipelines. In particular, we maintain an online API[4] to find the resources within the broader collection. A Python package[5] provides the utilities to access the API, to retrieve the data and process the packages to extract aligned sentences and documents using dedicated command-line tools. These tools support convenient features for attribute-based filtering (for example, using the overlap scores), allowing researchers to create custom data splits and ensuring

[3] https://huggingface.co/laurievb/OpenLID-v2

[4] https://opus.nlpl.eu/opusapi

[5] https://pypi.org/project/opustools/

```
<?xml version="1.0" encoding="utf-8"?>
<cesAlign version="1.0">
<linkGrp score="0.937015503875969" targType="s"
         fromDoc="de/2024/10003600_1928307_1_1/1960968140.xml.gz"
         toDoc="en/2024/10003600_1928307_1_1/1960964267.xml.gz">
<link id="SL0" xtargets="1;" />
<link id="SL1" xtargets="2;1" overlap="0.732" />
<link id="SL2" xtargets="3;2" overlap="0.857" />
<link id="SL3" xtargets="4;3" overlap="0.857" />
<link id="SL4" xtargets="5;4" overlap="0.723" />
<link id="SL5" xtargets="6;5 6" overlap="0.757" />
<link id="SL6" xtargets="7;7" overlap="0.473" />
...
```

Figure 1: An example of a sentence alignment file in XCES Align format. Aligned documents (i.e. subtitle files) are linked using the 'linkGrp' tags. The 'score' provides the alignment density value that can be used as an overall quality score. 'link' elements align sentences using their unique IDs in 'fromDoc' and 'toDoc' XML files. For example, 'SL0' describes an empty alignment where the sentence with ID 1 is not aligned to any sentence in the target language. 'SL5' shows an example of an alignment from one source sentence (6) to two target sentences (5 and 6). The overlap value specifies the time overlap between aligned source and target segments. This value is used for the automatic alignment procedures.

reproducible data preparation pipelines.

Licensing and Ethical Considerations: The project does not claim ownership of the subtitle texts and redistributes only files believed to be legally shareable under fair use principles. A take-down policy is in place for content owners. Publications using this corpus are requested to include a link to `OpenSubtitles.org`. This information should be noted in any ethics or limitations statement.

3. Development and Test Data

A primary contribution of OpenSubtitles2024 is a new, disjoint set of development and test data created to provide a reliable benchmark. All subtitles of movies and TV shows released in 2024 were withheld from the training corpus and reserved exclusively for these sets. This held-out data is released separately from the main corpus and is also available as a dataset on the Hugging Face Data Hub[6] using the gating function to discourage automated crawling. We deliberately chose the most recent year in our collection to minimize potential contamination in other webcrawls that also include movie subtitles of the same kind.

3.1. Bilingual Benchmarks

The bilingual benchmarks are designed for standard pairwise machine translation evaluation. They consist of a development set and a test set. The test set includes the aligned subtitles with the highest alignment density scores. The development set consists of all other high-quality aligned pairs from the 2024 data pool that were not selected for the test set, ensuring it is disjoint from both training and test data. Table 7 in Appendix B provides a detailed breakdown of the number of sentence pairs for each language pair in the development and test set. In addition to Hugging Face and GitHub, we also make the dataset available through the evaluation pipeline implemented in GlotEval.[7]

Construction: For each of the 2,022 language pairs, which cover the 70 languages available in the held-out set, the test set was constructed by selecting at most 5 movies or TV episodes that exhibited the highest alignment quality. The key filtering criterion was an alignment density score threshold of ≥ 0.8. These bilingual test sets are not multi-parallel; that is, the underlying movies may differ from one language pair to another. The development set consists of all remaining high-quality aligned pairs from the 2024 data pool that pass the alignment score threshold (67 languages, 1786 language pairs in total). It is recommended not to train on development data directly but rather use even that dataset for hyper-parameter fine-tuning only, or for monitoring models during development cycles and early stopping. There should not be any overlap between test and development data, but similarities due to the connections across episodes in TV shows and other potential overlaps may happen and should be avoided by leaving out development from direct exposure in model training.

[6] https://huggingface.co/datasets/Helsinki-NLP/OpenSubtitles2024

[7] https://github.com/MaLA-LM/GlotEval

3.2. Multilingual Benchmarks

Another subset refers to a synchronized multilingual benchmark. This subset offers a fully **multi-parallel** dataset, where subtitles for a specific movie are aligned across a set of multiple languages simultaneously. As such, the structure enables use cases such as multi-target translation and cross-lingual consistency evaluation.

Construction: The multilingual sets were created from movies with subtitle files available for all languages within a given language pack. It is not a trivial task to identify subsets that align across all languages with reasonable quality. As we are aiming at a completely synchronized multilingual datasets, i.e. subtitle files for which the same content is consistently aligned across all languages in a set, we need to properly combine the pairwise alignments with their segmentation into units that may differ between each individual language pair. Depending on the alignment quality threshold, some language pairs may also be missing because of their sub-optimal alignment density scores. Hence, the extraction of such multi-parallel dataset requires several steps:

- Step 1: Identify sets of subtitles from movies and TV shows that describe a connected graph through their alignments with each other; there can be many sets spanning different subsets of languages depending on the availability of translated and linked subtitles
- Step 2: Find pivot languages for each subtitle set that can be used to connect to all other languages in the set with the best possible alignment quality and the maximum number of languages that can be covered
- Step 3: Use the alignments to the pivot language to adjust individual bilingual sentence alignments to span over the same content in all languages, i.e. merge alignments until the links are synchronized over all language files.

Note that pivoting is only used to synchronize the alignments across all languages. All files in the set are original subtitles from the source collection, and no pivoting through translation has been done in our procedures.

We considered three variants of the procedure applying different density score thresholds: 0.9, 0.8 and no threshold. In this way, we can balance the coverage and size of the selected data. After inspecting the size and language coverage properties of extracted subsets, we selected a dataset with 40 languages and 15 movies / TV episodes as the best solution for our final test set, opting for the variant with the most strict alignment threshold (0.9) to ensure quality. We published the dataset on Hugging Face[8], and the evaluation pipeline is implemented using LM-Evaluation-Harness.[9]

In fact, other subsets with the same threshold did not exceed 41 languages as their maximum coverage, but all of them showed a significant drop in size. Appendix A includes a list of other high-quality subsets of multi-parallel subsets. All of them are available through our GitHub repository.

3.3. Data Formats and Access

Both benchmark types are provided as aligned plain text (Moses format) and XCES Align XML similar to the larger training data. To prevent leakage into web-crawled training sets, the archives are password-protected. We also make the benchmarks available through a password-protected browser interface[10] that allows one to inspect and explore the data. The main motivation behind this interface is to enable community feedback on the benchmarks we provide. Even though we applied a thorough selection procedure as described above, there might still be quality concerns about the data due to their source in user-contributed subtitles.

A full evaluation is out of scope for our budget and we hope that quality feedback can be collected from the community. Figure 2 shows a screenshot of the interface with a sample taken from the multi-parallel benchmark data. Quality feedback can be provided for each individual translation unit using the five stars on the side. Alternatively, a slider can also be used in an alternative view to provide a more fine-grained assessment on a scale from 0-100. Feedback is collected per user in a simple database and an average rating will be shown and can easily be retrieved from the system. Anonymous access is also available through a guest account.

Basic guidelines for annotation are available explaining the general principles of the star-based rating system. We intentionally opt for a very simple evaluation framework in order to motivate users to provide feedback without complex procedures and heavy guidelines. We intend to collect sufficient annotations to increase the credibility of the benchmark, and it will also help to exclude erroneous parts of the data in future releases. We consider opening a similar interface for the entire training data as well, but may apply some compromises in terms of language coverage due to the sheer size of the collection.

[8] https://huggingface.co/datasets/Helsinki-NLP/OpenSubtitles2024-40-langs-15-movies

[9] https://github.com/Helsinki-NLP/lm-evaluation-harness

[10] https://opus.nlpl.eu/bench/explore/

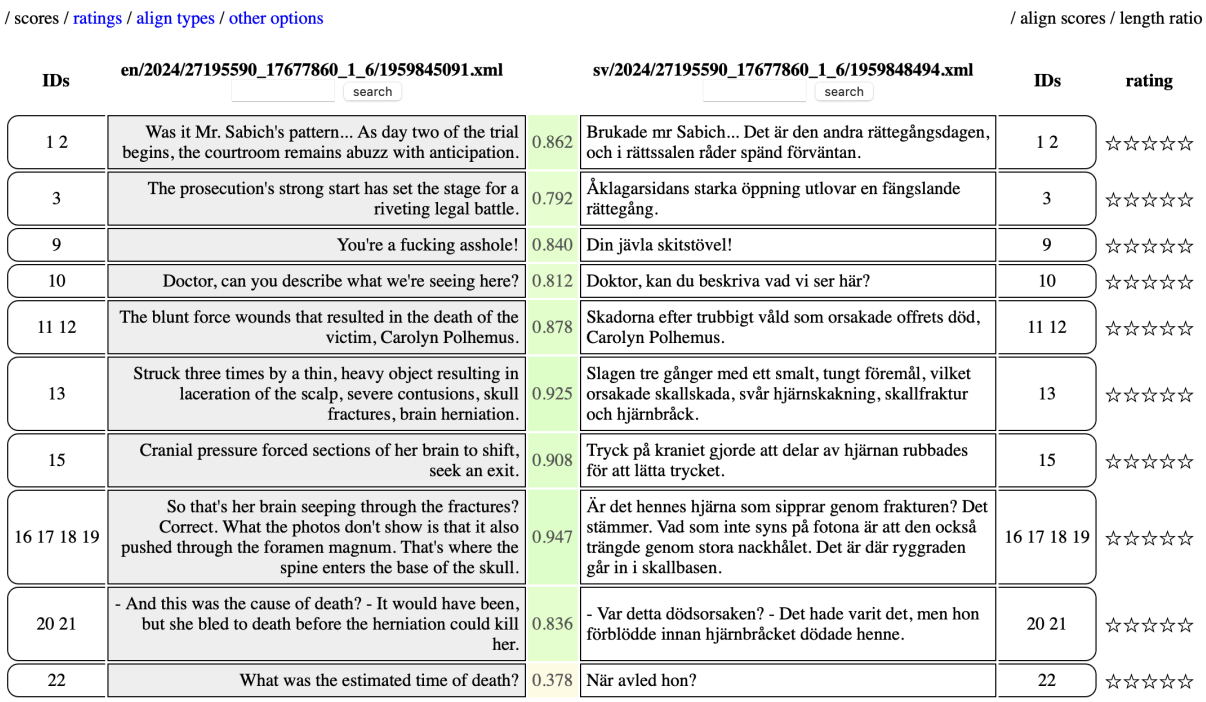

Figure 2: A screenshot of the data explorer interface with feedback function for a small sample from the multi-parallel benchmark dataset in English and Swedish.

4. Baseline Use Case: Fine-tuning a Multilingual Model for Low-Resource Translation

To demonstrate the practical value and quality of the new OpenSubtitles2024 data, we conducted a small fine-tuning experiment on a challenging, low-to-medium resource language pair: Finnish (fi) and Basque (eu). The primary goal was to measure the performance improvement when fine-tuning a general-purpose multilingual translation model on this specific language pair with data from our new corpus and to compare it with the OpenSubtitles2018 corpus.

4.1. Experimental Setup

Dataset: The training and test data for this experiment were sourced exclusively from the available subtitle collections. We used two different training sets for comparison. The first was derived from the `fi-eu` parallel corpus within the OpenSubtitles2018 release, containing 581,609 sentence pairs. The second, larger set came from the new OpenSubtitles2024 data release, which comprises 818,937 sentence pairs for the same language pair. To ensure that the model learned both translation directions, we created symmetric training data for both versions by taking the original sentence pairs and adding their swapped versions into a single training set. This resulted in approximately 1.16 million training examples for the 2018 set and 1.64 million for OpenSubtitles2024. For evaluation, we used our new test set of 3,702 Finnish-Basque sentence pairs.

Models: We compared three models in this study: The baseline model is the pre-trained multilingual neural machine translation model `Helsinki-NLP/opus-mt-tc-bible-big-mul-mul`, which has not been further optimized for Finnish-Basque translations.[11] The other two models are the same baseline model after being fine-tuned for two epochs on our symmetric 'fi-eu' training data from OpenSubtitles2018 and OpenSubtitles2024, respectively.

Metrics: We report both BLEU and chrF++ scores. Given that Finnish and Basque are morphologically rich and typologically distant languages, chrF++ is used as the primary metric as it is more robust to morphological variations and better reflects translation quality in such scenarios.

4.2. Results and Analysis

The results of our experiment, summarized in Table 2, show a clear and significant improvement in translation quality after fine-tuning, with the model

[11] This is the best open model currently available for this language pair according to the MT dashboard at `https://opus.nlpl.eu`.

trained on OpenSubtitles2024 achieving the best performance.

Task	Model	BLEU	chrF++
eu → fi	Baseline	4.97	27.41
	Fine-tuned (OpenSubtitles2018)	4.59	28.04
	Fine-tuned (OpenSubtitles2024)	**7.26**	**31.51**
fi → eu	Baseline	1.70	27.43
	Fine-tuned (OpenSubtitles2018)	4.99	31.72
	Fine-tuned (OpenSubtitles2024)	**11.04**	**37.76**

Table 2: Comparison of translation quality between Basque (eu) and Finnish (fi) for the Baseline, and models fine-tuned on OpenSubtitles2018 and OpenSubtitles2024 data.

As the table illustrates, fine-tuning on either dataset improves performance over the baseline in most cases. However, the results clearly show that fine-tuning with OpenSubtitles2024 leads to significantly better results than using the OpenSubtitles2018 data. For the direction into Finnish, the chrF++ score increased by more than 4 points, a notable jump compared to the 28.04 score from the experiments with 2018 data, which is just marginally on top of the baseline.

The improvement was even more pronounced in the other direction with translations into Basque. Here, the chrF++ score jumped by ten points in chrF++ with the new data, surpassing the 31.72 achieved with the 2018 data by a large margin. In this direction, the BLEU score also saw a more than six-fold increase with the OpenSubtitles2024 data, from a very low score of 1.70 to 11.04, indicating a substantial improvement in translation quality. This result is more than double the BLEU score of 4.99 obtained with the 2018 dataset. However, BLEU scores on this range should not be over-rated, especially also with respect to the issues on highly inflectional languages.

This small-scale use case demonstrates that the new OpenSubtitles2024 corpus can lead to substantial improvements and superior translation models. We attribute this significant performance gain to two key factors. Firstly, OpenSubtitles2024 provides a considerable expansion in data size for this language pair. Secondly, the enhanced performance suggests that improvements in the data's intrinsic quality and up-to-dateness are also crucial contributors to achieving better translation models. However, further studies are needed to properly evaluate the dataset in different scenarios and tasks, but even this particular test case already show-cases the potentials very well. In the next section, we will look more broadly at the multilingual performance using the multi-parallel benchmark we also extracted in this work.

	OpenSubtitles2024			OpenSubtitles2018	
Language	Total Lines	Sampled	Ratio	Total Lines	Sampled
ar-en	87,893,588	1,000,000	1.14%	29,823,188	339,310
bg-en	54,970,271	1,000,000	1.82%	40,204,338	731,383
cs-en	64,898,138	1,000,000	1.54%	42,346,436	652,506
da-en	50,915,427	1,000,000	1.96%	14,474,569	284,286
de-en	65,673,701	1,000,000	1.52%	22,512,639	342,795
el-en	76,818,359	1,000,000	1.30%	40,492,942	527,125
es-en	105,482,431	1,000,000	0.95%	61,434,251	582,412
et-en	20,224,893	1,000,000	4.94%	12,486,898	617,402
fa-en	61,518,181	1,000,000	1.63%	6,198,109	100,752
fi-en	59,651,022	1,000,000	1.68%	27,281,566	457,352
fr-en	83,896,581	1,000,000	1.19%	41,763,488	497,797
he-en	58,814,909	1,000,000	1.70%	29,887,386	508,160
hi-en	3,004,537	1,000,000	33.28%	93,016	30,958
hr-en	57,609,799	1,000,000	1.74%	35,131,729	609,822
hu-en	67,930,371	1,000,000	1.47%	42,655,519	627,930
id-en	73,834,757	1,000,000	1.35%	9,268,181	125,525
it-en	72,430,053	1,000,000	1.38%	35,216,229	486,210
ko-en	31,052,957	1,000,000	3.22%	1,391,190	44,800
lt-en	3,283,329	1,000,000	30.46%	1,415,961	431,257
lv-en	1,647,132	1,000,000	60.71%	519,553	315,428
ms-en	17,322,086	1,000,000	5.77%	1,928,345	111,322
nl-en	78,018,084	1,000,000	1.28%	37,200,621	476,820
no-en	37,439,738	1,000,000	2.67%	8,624,996	230,370
pl-en	75,324,736	1,000,000	1.33%	41,998,942	557,571
pt-en	68,557,861	1,000,000	1.46%	33,222,606	484,592
pt_BR-en	115,123,153	1,000,000	0.87%	61,367,019	533,055
ro-en	100,540,377	1,000,000	0.99%	50,693,226	504,207
ru-en	61,544,952	1,000,000	1.62%	25,910,105	420,994
sk-en	13,595,435	1,000,000	7.36%	8,850,871	651,017
sl-en	21,291,631	1,000,000	4.70%	19,641,457	922,496
sr-en	56,609,882	1,000,000	1.77%	42,635,098	753,138
sv-en	53,218,418	1,000,000	1.88%	17,660,152	331,842
ta-en	1,691,056	1,000,000	59.13%	32,417	19,169
te-en	1,355,728	1,000,000	73.76%	27,222	20,079
tr-en	73,487,258	1,000,000	1.36%	44,986,121	612,162
uk-en	10,541,711	1,000,000	9.49%	877,780	83,267
vi-en	37,005,102	1,000,000	2.70%	3,505,276	94,724
zh_CN-en	22,394,812	1,000,000	4.47%	11,203,286	500,262
zh_TW-en	18,583,480	1,000,000	5.38%	4,772,273	256,801

Table 3: Detailed breakdown of sentence pairs for the multilingual fine-tuning experiment. For each language, we show the total available lines, the number of lines sampled, and the sampling ratio applied.

5. Multilingual Use Case: Evaluations with the Multi-Parallel Benchmark

Inspired by the strong results from the previous section, we broadened our investigation to see if the performance gains from the OpenSubtitles2024 corpus generalize across a wider multilingual setting. To do this, we extended the experiment to an English-centric machine translation evaluation, using the multi-parallel benchmark introduced in Section 3.2.

5.1. Experimental Setup

Training Data: Fine-tuning on the full dataset for all 40 English-centric language pairs is computationally expensive. To manage resources while ensuring a meaningful comparison, we adopted a sampling strategy to create our training data.

For each language pair, we randomly sampled a fixed set of one million sentence pairs from the new OpenSubtitles2024 corpus. To create a comparable training set from OpenSubtitles2018 that accounts for the size difference between the corpora, we applied a proportional sampling rate. This

	ar	bg	cs	da	de	el	es	et	fa	fi	fr	he	hi
Baseline	28.43	33.95	32.63	43.67	39.74	40.14	44.34	36.15	22.59	37.23	37.60	40.51	19.28
OS18-prop	-0.02	+1.01	+0.96	-1.08	-0.31	+1.27	-0.29	+1.56	-1.56	-0.44	+0.86	+1.91	+0.18
OS18-uni	+0.38	**+1.49**	**+1.41**	-0.19	**+0.74**	+1.99	+0.08	+2.12	-0.82	+0.15	**+1.55**	**+2.69**	+0.08
OS24	**+1.90**	+1.42	+1.27	**+0.38**	-0.13	**+2.23**	**+0.19**	**+2.25**	**+6.34**	**+0.43**	+1.40	+1.85	**+10.93**
	hr	hu	id	it	ko	lt	lv	ms	nl	no	pl	pt	pt_BR
Baseline	42.53	32.23	43.88	41.60	2.71	33.39	34.47	36.25	38.96	46.14	36.01	43.63	44.38
OS18-prop	+0.13	+0.73	+1.60	+0.40	+5.86	+1.13	+1.52	+6.11	-0.26	-0.90	+0.09	+3.07	+0.88
OS18-uni	+0.67	**+1.14**	+3.86	+0.95	+11.44	+2.26	+2.31	+8.10	+0.48	+0.64	+0.46	**+4.15**	+1.38
OS24	**+0.92**	+1.10	**+4.72**	**+1.39**	**+12.55**	**+2.28**	**+3.79**	**+10.53**	**+0.49**	**+0.82**	**+0.58**	+4.10	**+1.73**
	ro	ru	sk	sl	sr	sv	ta	te	tr	uk	vi	zh_CN	zh_TW
Baseline	42.41	34.34	33.44	35.12	2.78	45.64	18.21	15.75	38.43	31.37	36.50	3.47	2.72
OS18-prop	+0.39	+0.19	+5.02	+1.66	+37.73	-1.38	+0.97	+2.80	+0.58	-5.42	-0.21	+9.32	+8.56
OS18-uni	+0.85	**+0.98**	+5.50	+1.63	+38.17	-0.34	+1.30	+2.66	+1.25	-5.22	+1.83	+9.98	+9.68
OS24	**+0.97**	+0.80	**+5.57**	**+1.69**	**+38.61**	**+0.16**	**+13.08**	**+10.68**	**+1.36**	**+1.01**	**+2.70**	**+10.36**	**+11.38**

Table 4: Comparison of translation quality (ChrF++) for the en→X direction, showing the baseline model versus models fine-tuned on OpenSubtitles2018 (OS18) and OpenSubtitles2024 (OS24) data. Deltas from the baseline are presented for fine-tuned models. *OS18-prop* refers to data sampled with a proportional sample size. *OS18-uni* refers to data with the same sample size as *OS24*. Best performance is in **bold**.

	ar	bg	cs	da	de	el	es	et	fa	fi	fr	he	hi
Baseline	**41.17**	32.04	36.22	41.96	41.20	39.64	**43.39**	37.95	35.06	35.85	33.94	42.62	17.30
OS18-prop	-1.11	+0.28	-0.02	+0.03	-0.41	+0.05	-0.82	+0.87	+0.09	+0.53	+0.25	-0.39	+2.05
OS18-uni	-0.22	+0.44	+0.44	+1.21	+0.68	+0.62	-0.31	+1.49	+2.39	+1.07	+0.67	+0.25	+2.14
OS24	-0.03	**+0.52**	**+0.55**	**+1.61**	**+0.75**	**+1.10**	-0.25	**+1.91**	**+2.89**	**+1.52**	**+0.99**	**+0.49**	**+19.44**
	hr	hu	id	it	ko	lt	lv	ms	nl	no	pl	pt	pt_BR
Baseline	42.86	34.28	39.67	42.27	12.09	35.75	37.51	37.54	36.28	42.84	37.67	44.68	43.04
OS18-prop	+0.01	+0.42	-0.60	-0.85	+5.35	+0.76	+0.79	-0.06	+0.19	-0.32	-0.77	-0.07	-0.18
OS18-uni	+0.28	+0.77	+0.92	-0.07	+13.97	+1.83	+1.32	+1.90	+0.71	+0.93	-0.23	+0.47	+0.57
OS24	**+0.51**	**+0.81**	**+1.74**	**+0.06**	**+15.36**	**+2.45**	**+2.97**	**+2.67**	**+1.02**	**+1.31**	**+0.30**	**+0.81**	**+0.65**
	ro	ru	sk	sl	sr	sv	ta	te	tr	uk	vi	zh_CN	zh_TW
Baseline	43.04	36.11	38.89	33.89	43.10	46.15	13.43	16.23	39.43	36.08	38.13	20.52	19.83
OS18-prop	-0.30	-0.58	+2.05	+1.17	+0.96	-0.43	+3.00	+3.00	+0.94	-2.30	-1.71	+8.78	+8.53
OS18-uni	-0.04	+0.09	**+2.26**	+1.09	**+1.18**	+0.69	+3.03	+2.83	+1.27	-1.04	-0.05	+10.83	+12.03
OS24	**+0.80**	**+0.45**	+2.25	**+1.26**	+0.97	**+0.88**	**+18.40**	**+16.01**	**+1.73**	**+0.05**	**+0.63**	**+10.90**	**+12.24**

Table 5: Comparison of translation quality (ChrF++) for the X→en direction, showing the baseline model versus models fine-tuned on OpenSubtitles2018 (OS18) and OpenSubtitles2024 (OS24) data. Deltas from the baseline are presented for fine-tuned models. *OS18-prop* refers to data sampled with a proportional sample size. *OS18-uni* refers to data with the same sample size as *OS24*. Best performance is in **bold**.

rate was calculated for each language pair based on the ratio of our sample size (one million) to the total number of parallel sentences available in its OpenSubtitles2024 set. We then applied this same rate to the total lines in the OpenSubtitles2018 corpus to determine its sample size. This method ensures that the sampling is representative of the data distribution between the two corpora. The exact number of total and sampled sentence pairs, along with the calculated sampling rate for each language, is detailed in Table 3.

Test Data: For the evaluation, we use the most representative subset from our held-out multilingual test set, which supports 40 languages and contains 6005 multi-parallel sentences from 15 movies (as shown in Appendix B). We then study the effect of fine-tuning on our new data in comparison to the performance of the baseline model and a model that has been fine-tuned on OpenSubtitles2018 data.

Models: Th three models in this study follow the same setup as in the previous use case. The baseline model is the pre-trained multilingual NMT model `Helsinki-NLP/opus-mt-tc-bible-big-mul-mul`. The other two models are this same baseline after being fine-tuned on our sampled training data from OpenSubtitles2018 and OpenSubtitles2024, respectively.

5.2. Results and Analysis

The results of our fine-tuning experiments are summarized in Table 4 for the en→X direction and Table 5 for the X→en direction.

We observe significant improvements in nearly all cases when fine-tuning on OpenSubtitles2024 data, both compared to the baseline and to the model fine-tuned on OpenSubtitles2018. Notably, fine-tuning with the proportionally smaller OpenSubtitles2018 (OS18-prop) data sometimes leads to degrading performance (e.g., for `de`, `sv`, `uk`),

which could be attributed to data quality issues or a domain mismatch, as the older corpus lacks data from more recent movies, whereas the test set is composed from the most recent movies from the entire collection.

On average, chrF++ scores improve 4.46 points for translations from English to other languages using OpenSubtitles2024 compared to an average increase of 2.17 points when fine-tuning on OpenSubtitles2018. For translations into English, the average improvement is 3.30 points for OpenSubtitles2024 and only 0.75 for OpenSubtitles2018.

We noted that the baseline model performs very poorly for several languages, especially **Korean (ko), Serbian (sr), and both Chinese variants (zh_CN, zh_TW)**, with ChrF++ scores near or below 3 points. As shown in Table 4 and Table 5, fine-tuning on either dataset effectively mitigates this issue, yielding dramatic improvements. For instance, in the en→sr direction, the score jumps from a baseline of 2.78 to over 40 after fine-tuning, demonstrating the model's ability to learn translation for these languages effectively from the new data.

For completeness, we also conducted a supplementary experiment where both fine-tuning sets were uniformly sampled to the same size (one million sentence pairs),[12] see rows denoted by *OS18-uni* in Tables 4 and 5. Even those results confirm that OpenSubtitles2024 is superior to OpenSubtitles2018. For translations from English, fine-tuning on comparable sizes achieves an average improvement of 3.03 chrF++ scores with OpenSubtitles2018, which is still significantly below the 4.46 points improvement measured with OpenSubtitles2024 fine-tuning. For translations into English, OpenSubtitles2018 fine-tuning raises the scores by 1.75 points compared to an improvement of 3.30 with OpenSubtitles2024.

6. Conclusions

The paper introduces OpenSubtitles2024, a new edition of a massively multilingual dataset. Compared to previous releases, it comes with an almost 50% increase in language coverage and doubles the size of earlier versions. In addition to the aligned training data, we also provide a dedicated held-out dataset with 2,022 language pairs in the package. We applied a careful procedure to extract reliable subtitle pairs for a wide-coverage test set and provide the remaining portions as a disjoint development set. Furthermore, we compiled a subset of synchronized multi-parallel subtitles with consistent alignments across all languages involved. All data is publicly available and can be used to train, tune, and evaluate multilingual language and translation models. In our use cases, we demonstrate that the new benchmark provides a challenging multilingual translation task with little risk of data contamination. We can also show that the training data can push translation quality for under-resourced language pairs and translation directions. The release is currently the biggest corpus in the comprehensive bitext collection OPUS and, with that, forms a major new open resource in the development of multilingual technology.

[12] Note, that OpenSubtitles2018 does not include one million sentences for Hindi (hi), Latvian (lv), Tamil (ta) and Telugu (te). For those languages, we used all available data in OpenSubtitles2018, and the substantial drop compared to OpenSubtitles2024 can be partially explained by the difference in size (see Table 3)

7. Acknowledgments

This work was supported by the European Union's Horizon Europe research and innovation programme under grant agreement No 101070350, by the OpenEuroLLM project, co-funded by the Digital Europe Programme under GA no. 101195233, and by the AI-DOC program hosted at the Finnish Center of Artificial Intelligence (decision number VN/3137/2024-OKM-6). The authors also thank CSC IT Center for Science, Finland, and LUMI supercomputers, owned by the EuroHPC Joint Undertaking, for providing computational resources.

8. Bibliographical References

Simone Balloccu, Patrícia Schmidtová, Mateusz Lango, and Ondrej Dusek. 2024. Leak, cheat, repeat: Data contamination and evaluation malpractices in closed-source LLMs. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 67–93, St. Julian's, Malta. Association for Computational Linguistics.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2024. The language model evaluation harness.

Hengyu Luo, Zihao Li, Joseph Attieh, Sawal Devkota, Ona de Gibert, Shaoxiong Ji, Peiqin Lin, Bhavani Sai Praneeth Varma Mantina, Ananda Sreenidhi, Raúl Vázquez, Mengjie Wang, Samea Yusofi, and Jörg Tiedemann. 2025. Gloteval: A

test suite for massively multilingual evaluation of large language models.

Oscar Sainz, Jon Campos, Iker García-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, and Eneko Agirre. 2023. NLP evaluation in trouble: On the need to measure LLM data contamination for each benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10776–10787, Singapore. Association for Computational Linguistics.

Jörg Tiedemann. 2008. Synchronizing translated movie subtitles. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Jörg Tiedemann. 2016. Finding alternative translations in a large corpus of movie subtitle. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3518–3522, Portorož, Slovenia. European Language Resources Association (ELRA).

Yuemei Xu, Ling Hu, Jiayi Zhao, Zihan Qiu, Kexin Xu, Yuqi Ye, and Hanwen Gu. 2025. A survey on multilingual large language models: Corpora, alignment, and bias. *Frontiers of Computer Science*, 19(11):1911362.

9. Language Resource References

Marta Bañón, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis, Mikel L. Forcada, Amir Kamran, Faheem Kirefu, Philipp Koehn, Sergio Ortiz Rojas, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Elsa Sarrías, Marek Strelec, Brian Thompson, William Waites, Dion Wiggins, and Jaume Zaragoza. 2020. ParaCrawl: Web-scale acquisition of parallel corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4555–4567, Online. Association for Computational Linguistics.

Christos Christodouloupoulos and Mark Steedman. 2015. A massively parallel corpus: the bible in 100 languages. *Lang. Resour. Eval.*, 49(2):375–395.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86, Phuket, Thailand.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia. European Language Resources Association (ELRA).

Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, Armand Joulin, and Angela Fan. 2021. CCMatrix: Mining billions of high-quality parallel sentences on the web. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6490–6500, Online. Association for Computational Linguistics.

Ralf Steinberger, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaž Erjavec, Dan Tufiş, and Dániel Varga. 2006. The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

A. Multilingual Benchmark Statistics

The procedures for extracting multi-parallel test sets have been described in detail in Section 3.2. Table 6 lists statistics for the highest-quality sets (alignment threshold of 0.9) that contain 20 or more languages. From these collections, we selected the third option listed in the table as the official release, representing the best size and language coverage among the alternatives available to us.

B. Bilingual Benchmark Statistics

Table 7 lists the size of bilingual development and test data for each language pair in the collection. The table shows the number of sentences in each dataset. For some language pairs, the development set is quite substantial but should still not be used for training as the data is reserved for validation purposes and may also be used to extract additional sub-sets for benchmarking MT models in the future.

langs	movies	lines	languages IDs	avglen (lang)
41	5	2,512	ar be bg cs da de el en es et fa fi fr he hi hr hu id it ko lt lv ms nl no pl pt pt_BR ro ru sk sl sr sv ta te tr uk vi zh_CN zh_TW	8.38 (en)
41	5	2,012	ar bg cs da de el en es es_419 et fa fi fr he hi hr hu id it ko lt lv ms nl no pl pt pt_BR ro ru sk sl sr sv ta te tr uk vi zh_CN zh_TW	10.24 (en)
40	15	6,005	ar bg cs da de el en es et fa fi fr he hi hr hu id it ko lt lv ms nl no pl pt pt_BR ro ru sk sl sr sv ta te tr uk vi zh_CN zh_TW	9.98 (en)
39	7	3,104	ar bg cs da de el en es et fa fi fr he hi hr hu id it ko lt lv ms nl no pl pt pt_BR ro ru sk sl sv ta te tr uk vi zh_CN zh_TW	8.64 (en)
35	6	2,331	ar bg bn ca da de el en es et eu fa fi fr gl he hr id it ko ms nl no pl pt pt_BR ro ru sv tl tr uk vi zh_CN zh_TW	8.97 (en)
31	5	2,669	ar cs da de el en es es_419 fa fi fr hr hu id it ko ms nl no pl pt pt_BR ro ru sv tl tr uk vi zh_CN zh_TW	7.61 (en)
30	5	1,442	ar cs da de el en es fa fi fr he hr hu id it ko nl no pl pt pt_BR ro ru sk sr sv tr vi zh_CN zh_TW	9.2 (en)
30	5	1,743	ar cs da de el en es_419 es_ES fi fr he hr hu id it ms nl no pl pt pt_BR ro ru sv tl tr uk vi zh_CN zh_TW	9.87 (en)
30	6	2,967	ar cs da de el en es_419 es_ES fi fr hr hu id it ko ms nl no pl pt pt_BR ro ru sv tl tr uk vi zh_CN zh_TW	7.14 (en)
29	9	5,088	ar cs da de el en es_419 es_ES fi fr hr hu id it ko ms nl no pl pt pt_BR ro sv tl tr uk vi zh_CN zh_TW	7.03 (en)
29	6	2,967	ar cs da de el en es_ES fi fr hr hu id it ko ms nl no pl pt pt_BR ro ru sv tl tr uk vi zh_CN zh_TW	7.6 (en)
28	6	1,760	ar cs da de el en es fa fi fr he hr hu id it ko ms nl no pl pt pt_BR ro tl tr uk vi zh_CN	9.22 (en)
27	6	1,160	ar cs da de el es fi fr he hr hu it ko ms nl no pl pt pt_BR ro sv tl tr uk vi zh_CN zh_TW	9.16 (pt)
26	6	5,109	ar da de el en es fa fi fr hr id it ko ms nl no pl pt pt_BR ro ru si sv tr uk vi	7.41 (en)
25	6	3,444	ar bg cs da de el en es fa fi fr hr hu id it nl no pl pt pt_BR ru sv tr uk vi	9.36 (en)
25	6	5,102	ar cs da de el en es fi fr hu id it ms no pl pt pt_BR ro ru sv tl uk vi zh_CN zh_TW	6.51 (en)
25	6	3,694	ar cs da de el en fi hr hu id it ko ms nl no pl pt ro ru sv tr uk vi zh_CN zh_TW	7.27 (en)
23	5	1,867	ar bn cs da de el en es fa fi fr id ko ms nl pl pt pt_BR ro si sv vi zh_TW	6.56 (en)
23	5	1,375	ar cs da de el es_419 es_ES fa fi fr he hr hu is it nl no pl pt_BR ro ru sv tr	7.59 (pt_BR)
23	15	7,434	cs da de el en es es_419 fi fr hu id it ms nl no pl pt pt_BR ro sk sv tr zh_TW	7.4 (en)
22	7	1,632	ar cs da de el en fi he hi hu id it ko ms nl no pl pt ro sv tr zh_TW	12.19 (en)
21	8	3,131	ar da de el en es et fa fi he hr id nl pl pt pt_BR ro ru sv tr uk	7.19 (en)
20	7	3,164	ar bg da de el en es fa fi he hr hu id nl pl pt pt_BR sr sv tr	8.11 (en)
20	7	2,804	ar de el en es es_419 fi hr hu nl no pl pt pt_BR ro ru sv tr uk vi	7.91 (en)

Table 6: Statistics of the Multilingual OpenSubtitles Test Sets (Alignment Threshold = 0.9, only including sets with ≥ 20 languages). *langs* provides the number of languages included in the set, *movies* shows the number of movies / TV shows covered, *lines* shows the number of lines per language that are aligned across all languages, *language IDs* lists the included languages by their ISO-8859-1 codes, and *avglen* provides the average length in tokens per line in the set for a selected languages shown in brackets.

Table 7: Number of sentences in bilingual development (upper triangle) and test sets (lower triangle.