OJ4OCRMT: A Large Multilingual Dataset for OCR-MT Evaluation

Paul McNamee[1], Kevin Duh[1,2], Cameron Carpenter[2], Ron Colaianni[3], Nolan King[3], Kenton Murray[1,2]
[1]Human Language Technology Center of Excellence
[2]Department of Computer Science, Johns Hopkins University
[3]Department of Defense
Correspondence: mcnamee@jhu.edu

Abstract

We introduce *OJ4OCRMT*, an Optical Character Recognition (OCR) dataset for Machine Translation (MT). The dataset supports research on automatic extraction, recognition, and translation of text from document images. The *Official Journal of the European Union* (OJEU), is the official gazette for the EU. Tens of thousands of pages of legislative acts and regulatory notices are published annually, and parallel translations are available in each of the official languages. Due to its large size, high degree of multilinguality, and carefully produced human translations, the OJEU is a singular resource for language processing research. We have assembled a large collection of parallel pages from the OJEU and have created a dataset to support translation of document images. In this work we introduce the dataset, describe the design decisions which we undertook, and report baseline performance figures for the translation task. It is our hope that this dataset will significantly add to the comparatively few resources presently available for evaluating OCR-MT systems.

1 Introduction

Relatively few datasets exist for studying the translation of document images. The manual labor associated with obtaining suitable digital images and producing high-quality transcriptions of the source image and translations in the target language(s) is an impediment. We survey some of the available datasets in Table 1. Common limitations include being small in size, narrow in image types, restricted to a few languages, and reliance on automatic generation of images or translations.

The *Official Journal of the European Union* (OJEU) is available in digital form in the official languages of the EU and it contains content going back decades. The OJEU is in the public domain, and its quantity of data, high quality translations, and large number of supported languages covering three writing systems, make it an attractive source for developing a open source dataset to support translation of document images. The OJEU and related EU publications have previously been used as corpora in the development and evaluation of language technologies. For example, Koehn produced parallel texts from transcripts of the European Paraliament (2005). Similarly, Steinberger and colleagues at the JRC have released parallel texts such as *JRC-Acquis* (Steinberger et al., 2006) and *DGT-Acquis* (Steinberger et al., 2012), and they even foresaw the use of these collections for supporting OCR research (Steinberger et al., 2014).[1] Our present focus is in the development of corpora to support the evaluation of document image translation, which can be accomplished through pipelines of OCR and MT systems, or through use of newly available vision language models such as *Claude* (Kim et al., 2025) or *Pali Gemma* (Steiner et al., 2024).

In Section 2 we survey datasets for this task. Section 3 describes the creation of *OJ4OCRMT*, including the design choices we undertook and key characteristics of the dataset. In Section 4 we describe our experimental setup. Finally, we present and discuss our baseline results in Section 5.

2 Related Work

Datasets for OCR-MT can be classified by the type of images used (see Table 1). First, several pioneering efforts rendered images from bilingual text (bitext) corpora commonly used in text-based MT research. For example, (Mansimov et al., 2020) created images from German to English bitext in the WMT dataset; (Ignat et al., 2022) created images

[1]To our knowledge, no OCR-specific resources have been produced from these sources.

Proceedings of Machine Translation Summit XX Volume 1, pages 113–125, June 23-27, 2025

Dataset	Image Type & Domain	Translation	Language & Script
MADCAT (Song et al., 2012)	Handwritten documents	Professional	{ar, zh}→en
(Mansimov et al., 2020)	Rendered from bitext: WMT	Professional	de→en
OCR4MT (Ignat et al., 2022)	Rendered from bitext: Flores	Professional	60 languages
IIMT (Tian et al., 2023)	Rendered from bitext: WMT	Professional	de→en
OCRMT30K (Lan et al., 2023)	Natural Image, street signs	Professional	zh→en
Vistra (Salesky et al., 2024)	Natural Image, street signs	Professional	en→{de, es, ru, zh}
MIT-10M (Li et al., 2025)	Natural Image from web	MT	14 languages, 8 scripts
CAMIO (Arrigo et al., 2022)	Natural documents from web	-	35 languages, 24 scripts
DITrans (Zhang et al., 2023)	Natural PDF: newspaper, ad	Professional	en→zh
DoTA (Liang et al., 2024)	Natural PDF: scientific doc	MT:train Pro:test	en→zh
OJ4OCRMT (this work)	Natural PDF: government doc	Professional	23 languages multi-way, 3 scripts

Table 1: Comparison of OCR-MT or Multilingual OCR datasets

with the text from 60 languages in the FLORES dataset.[2] The advantage of rendering approaches is that any quantity of images can be synthesized. For example, Tian *et al.* (2023) rendered a 4 million pair image-translation training set based on WMT bitext. A notable disadvantage is that great effort is required if we want to match the variety of image layouts found in the real world.

A second approach to OCR-MT datasets is based on collecting natural images from the wild. For example, (Liang et al., 2024) take existing Chinese street sign OCR datasets like (Sun et al., 2019) and translate the text portion into English using professional human translators. In a similar vein, (Salesky et al., 2024) collected 770 natural photographs consisting of English in-scene text from the web and hired professional translators for translation into German, Spanish, Russian, and Chinese. The dataset of (Li et al., 2025) significantly increased the scale (10 million collected images) but relied on a machine translation API to generate the translated text; it covers 14 languages and 8 scripts.

A third approach focuses on collecting natural documents from the web. The distinction with the second approach is not clear-cut, but the focus here is on collecting machine-printed documents that are text-rich and sentence-like, as opposed to photographs like street signs where in-scene text may consist of short phrases. For example, (Arrigo et al., 2022) collected and annotated 70k images for bounding boxes from the web, covering 35 languages and including both scanned and machine-printed documents like newspapers, books, journals, and web pages. A subset of ∼15k images covering 13 languages were transcribed: note this is a multilingual collection where images contain different languages and scripts; it must be translated to create an OCR-MT dataset.

Most relevant to our work are DITrans (Zhang et al., 2023) and DoTA (Liang et al., 2024), which like our work, focus on natural PDFs that are text-rich documents containing a diversity of layouts. DITrans consists of political reports, newspapers, and advertisements; DoTA consists of scientific papers from arXiv. The test sets of both of these have been professionally translated from English to Chinese. Additionally, they have provided French and German translations performed by MT. Our dataset is different in that we have a larger set of languages (23 in total) with translations professionally produced by the data provider and aligned in a multi-way parallel fashion. In general, these kinds of document PDFs, when converted to images, are challenging from the OCR perspective due to diverse layouts and reading orders; they are also challenging from the MT perspective due to the richer vocabulary and syntax.

Last but not least, there is work on translation of handwritten text, c.f. (Song et al., 2012). This is a substantially different problem than scanned or born-digital machine-print documents.

3 Dataset

An ideal dataset for OCR-MT evaluation consists of three components: (a) document images; (b) ground truth transcripts in the source languages; and, (c) human-produced translations in the target languages. We downloaded PDF files for each OJEU document in the available languages and extracted images and text for each individual page. Files were obtained by crawling the EUR-Lex online portal[3], the official repository for the OJEU. We decided to focus on content from recent years because previous datasets such as *DGT-Acquis* have released translation memories that include some

[2]They also include some natural PDFs from the Universal Declaration of Human Rights database.

[3]https://eur-lex.europa.eu/oj/direct-access.html

OJEU content, and machine translation systems are often trained using these data, which are available in the popular OPUS portal (Tiedemann, 2012).

As a general rule, OJEU pages in different languages contain equivalent content for a given published page. In other words, the ith page of document D in language L_1, matches the content of the ith page of document D in language L_2.[4] We thus have page-wise alignments, and not sentence-wise alignments which are more commonly used for machine translation. We elected to work directly with page-level alignments and not perform automated alignment of text fragments. This avoids the considerable expense required in manual determination of reading order and sentence selection and alignment.

The ground truth text for each page was obtained using *pdftotext*. The extracted text contains blank lines and many broken up lines or text fragments. The order of the extracted content can vary by language, but usually only slightly. We used multiple newline characters as hard breaks between sections of text (*i.e.,* paragraphs, list items), but conjoined other text fragments and then ran automated sentence boundary detection using the multilingual sentence splitter, *ersatz* (Wicks and Post, 2021).

Lossless images in PNG format were produced in three resolutions: 72, 150, and 300 dpi. This corresponds to the historical default resolution for web images, an intermediate value for experimentation, and the current standard high-quality resolution.

Due to the fact that a number of the articles were not available in the Irish language, we made the decision to exclude it from the set of languages in the dataset. Every page in the dev and test sets is available in the other 23 EU languages.[5]

A sample image and its extracted and reconstituted text are shown in Figure 1.

3.1 Partitions

Our goal was to produce *dev* and *test* partitions consisting of at least 1,000 images (*i.e.,* pages). We used content from 2022 for a *dev* partition, and content from the first nine months of 2023 for a *test* partition. In total there are 1,656 pages in *dev* and 1,119 pages in *test*, each of which was manually vetted. These were selected from 944 documents (82,589 pages), and 661 documents (67323 pages), respectively. Statistics are given in Tables 2 and 3.

[4]This rule is sometimes broken, when there are mid-sentence breaks at page boundaries, or in transcripts of Parliamentary discourse that are intentionally left untranslated.

[5]Bulgarian, Croatian, Czech, Danish, Dutch, English, Estonian, Finnish, French, German, Greek, Hungarian, Italian, Latvian, Lithuanian, Maltese, Polish, Portuguese, Romanian, Slovak, Slovene, Spanish, and Swedish

lang	**dev**: 1,656 pages #tokens	#types	**test**: 1,119 pages #tokens	#types
bg	888,760	79,247	590,287	59,967
cs	793,944	87,327	520,563	66,126
da	793,933	84,881	525,390	62,744
de	822,715	85,955	542,173	64,159
el	916,722	80,393	609,965	60,627
en	856,815	62,476	571,423	47,068
es	989,472	67,917	662,827	51,756
et	651,124	108,658	428,058	80,893
fi	641,669	118,029	423,119	87,639
fr	964,872	69,780	644,016	52,587
hr	785,715	85,765	522,407	64,992
hu	770,588	105,837	504,316	78,924
it	903,328	72,291	599,804	54,569
lt	735,764	91,347	485,585	69,448
lv	716,048	91,169	476,038	69,336
mt	739,617	92,775	492,477	70,396
nl	887,333	76,750	589,037	57,248
pl	806,494	95,316	531,906	72,305
pt	928,921	69,558	621,720	52,568
ro	901,817	75,588	597,887	57,616
sk	784,675	91,566	519,818	69,277
sl	781,993	87,441	515,796	66,653
sv	783,356	83,188	517,883	62,072

Table 2: Data statistics: number of tokens and types

	pages	**regular**	**tables**	**figures**
dev	1,656	1,412 (85%)	193 (12%)	51 (3%)
test	1,119	979 (87%)	98 (9%)	42 (4%)

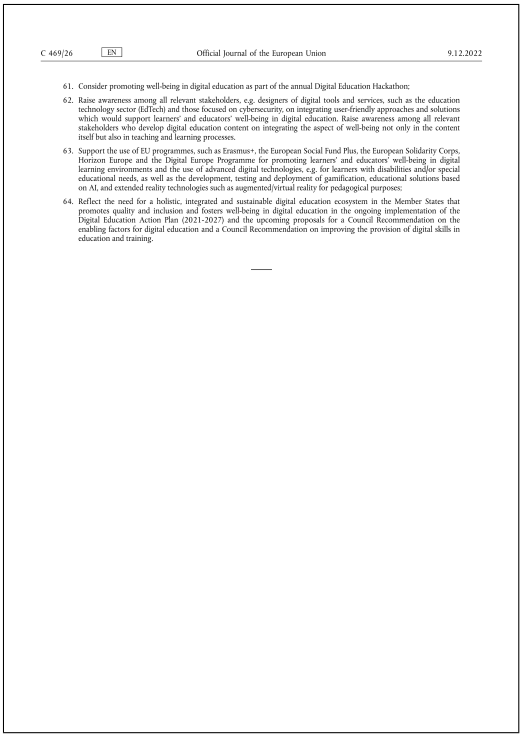
Table 3: Data statistics: partition size and numbers of pages with tables or figures.

3.2 Diversity of Content

The greater part of the dataset consists of text in narrative format (*e.g.,* letters or memoranda), or outline or enumerated list format. However, we did observe a variety of visual and textual features, including: tables of contents, tabular data, forms, scientific charts, drawings, figures, logos, signatures, and equations. A sample of pages is shown in Figure 2. Pages were tagged as *table* if they contained at least one form or table, whether in portrait or landscape orientation. Pages were tagged as *figure* if they contained a graph, logo, seal, photograph, or drawing. The remaining pages, which are the majority, are deemed *regular*. Table 3 reports the relative prevalence of tables and figures.

3.3 Quality Control

To ensure the quality of the data that we selected, we performed both automated filtering and human review. We automatically rejected pages if: (a) they



C 469/26 EN Official Journal of the European Union 9.12.2022

61. Consider promoting well-being in digital education as part of the annual Digital Education Hackathon;

62. Raise awareness among all relevant stakeholders, e.g. designers of digital tools and services, such as the education technology sector (EdTech) and those focused on cybersecurity, on integrating user-friendly approaches and solutions which would support learners' and educators' well-being in digital education. Raise awareness among all relevant stakeholders who develop digital education content on integrating the aspect of well-being not only in the content itself but also in teaching and learning processes.

63. Support the use of EU programmes, such as Erasmus+, the European Social Fund Plus, the European Solidarity Corps, Horizon Europe and the Digital Europe Programme for promoting learners' and educators' well-being in digital learning environments and the use of advanced digital technologies, e.g. for learners with disabilities and/or special educational needs, as well as the development, testing and deployment of gamification, educational solutions based on AI, and extended reality technologies such as augmented/virtual reality for pedagogical purposes;

64. Reflect the need for a holistic, integrated and sustainable digital education ecosystem in the Member States that promotes quality and inclusion and fosters well-being in digital education in the ongoing implementation of the Digital Education Action Plan (2021-2027) and the upcoming proposals for a Council Recommendation on the enabling factors for digital education and a Council Recommendation on improving the provision of digital skills in education and training.

—

C 469/26
EN
Official Journal of the European Union
9.12.2022
61.
Consider promoting well-being in digital education as part of the annual Digital Education Hackathon; 62.
Raise awareness among all relevant stakeholders, e.g. designers of digital tools and services, such as the education technology sector (EdTech) and those focused on cybersecurity, on integrating user-friendly approaches and solutions which would support learners' and educators' well-being in digital education.
Raise awareness among all relevant stakeholders who develop digital education content on integrating the aspect of well-being not only in the content itself but also in teaching and learning processes. 63.
Support the use of EU programmes, such as Erasmus+, the European Social Fund Plus, the European Solidarity Corps, Horizon Europe and the Digital Europe Programme for promoting learners' and educators' well-being in digital learning environments and the use of advanced digital technologies, e.g. for learners with disabilities and/or special educational needs, as well as the development, testing and deployment of gamification, educational solutions based on AI, and extended reality technologies such as augmented/virtual reality for pedagogical purposes; 64.
Reflect the need for a holistic, integrated and sustainable digital education ecosystem in the Member States that promotes quality and inclusion and fosters well-being in digital education in the ongoing implementation of the Digital Education Action Plan (2021-2027) and the upcoming proposals for a Council Recommendation on the enabling factors for digital education and a Council Recommendation on improving the provision of digital skills in education and training.

Figure 1: The page image for OJ:C:2022:469:FULL.en.p-28 is shown at left. The original document can be viewed at https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=OJ:C:2022:469:FULL#page=28. On the right is the extracted text from *pdftotext* which was then run through sentence boundary detection.

were blank or contained fewer than 80 alphabetic characters; (b) parallel content did not match the expected language, according to automated language ID; or, (c) content was not available in one of the 23 languages (possibly due to errors in downloading). Human review consisted of avoiding less desirable pages, such as pages with mid-sentence breaks at the top or bottom of the page, pages largely consisting of tables of numbers, and atypical language, such as long lists of names or product codes.

3.4 Limitations

In addition to the *OJ4OCRMT's* desirable properties, it also has a couple of limitations. There are no ground-truth annotations for reading order or sentence segmentation. And because the data is obtained from a single source, there is some homogeneity in both the visual properties (*e.g.,* layouts and fonts) and the textual characteristics (*e.g.,* translators may use consistent terminology and style).

4 Experimental Setup

4.1 OCR Engines

In our benchmark experiments we used two open-source OCR engines as part of OCR-MT cascades, and one commercial end-to-end OCR-MT model.

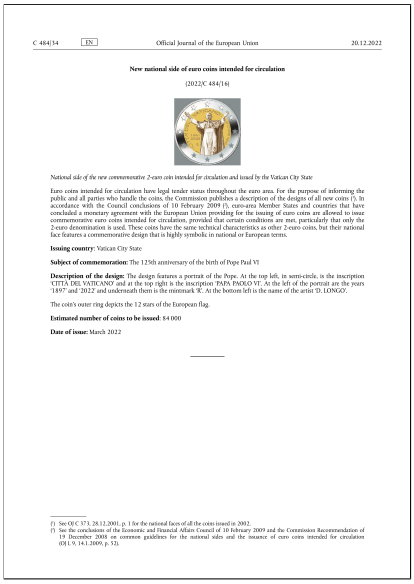
EasyOCR is an open-source, python-based multilingual OCR engine.[6] We used version 1.7.1 which is released under the Apache 2.0 license. The tool does not support the Finnish or Greek languages, but we did run Latin-based decodes for those languages anyway; we obtained reasonable results for Finnish, and meaningless results for Greek (as one would expect).

Another open source tool used was *Tesseract* (version 5.5.0), which was applied to the images from the dataset at each of the three resolutions. For each language's portion of the data, the corresponding LSTM-based pretrained Tesseract model was applied. (Smith et al., 2009). *Tesseract* was run using the Unix `parallel` utility for increased CPU throughput (Tange, 2024).

For the end-to-end OCR-MT commercial system evaluation, we used the API service hosted by Anthropic. All results used the model identifier `claude-3-5-sonnet-20241022`. We used the following prompt structure:

```
system: "You are a highly skilled
translator and interpreter with
expertise in many languages. Your
```

[6]Available from: https://github.com/JaidedAI/EasyOCR.



C 484/14 EN Official Journal of the European Union 20.12.2022

New national side of euro coins intended for circulation

(2022/C 484/16)

National side of the new commemorative 2-euro coin intended for circulation and issued by the Vatican City State

Euro coins intended for circulation have legal tender status throughout the euro area. For the purpose of informing the public and all parties who handle the coins, the Commission publishes a description of the designs of all new coins [1]. In accordance with the Council conclusions of 10 February 2009 [2], euro-area Member States and countries that have concluded a monetary agreement with the European Union providing for the issuing of euro coins are allowed to issue commemorative euro coins intended for circulation, provided that certain conditions are met, particularly that only the 2-euro denomination is used. These coins have the same technical characteristics as other 2-euro coins, but their national face features a commemorative design that is highly symbolic in national or European terms.

Issuing country: Vatican City State

Subject of commemoration: The 125th anniversary of the birth of Pope Paul VI

Description of the design: The design features a portrait of the Pope. At the top left, in semi-circle, is the inscription 'CITTÀ DEL VATICANO' and at the top right is the inscription 'PAPA PAOLO VI'. At the left of the portrait are the years '1897' and '2022' and underneath them is the mintmark 'R'. At the bottom left is the name of the artist 'D. LONGO'.

The coin's outer ring depicts the 12 stars of the European flag.

Estimated number of coins to be issued: 84 000

Date of issue: March 2022

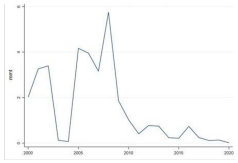
(1) See OJ C 373, 28.12.2001, p. 1 for the national faces of all the coins issued in 2002.
(2) See the conclusions of the Economic and Financial Affairs Council of 10 February 2009 and the Commission Recommendation of 19 December 2008 on common guidelines for the national sides and the issuance of euro coins intended for circulation (OJ L 9, 14.1.2009, p. 52).

16.3.2022 EN Official Journal of the European Union L 88/53

(250) Other export restrictions include licencing requirements and a qualified exporters list (in practice limited to the state trading enterprise Metal and Mineral Trading Corporation) for certain grades of the chromium ore, as well as a congestion surcharge levied on the Base Freight Rates of the Indian Railways in the rail transport of chromium ore to Bangladesh and Pakistan (being the only countries with rail connection with India).

(251) The most visible effect of the policy was the impressive reduction of the volume of export of chromium ore following the introduction of the targeted export restraints in 2008 and then again following their expansion in 2012. The development of chromium ore export is illustrated by the below graph:

IND, Chromium exports (HS 261000)



(252) Hence, the targeted export restraints achieved the goal pursued by the GOI of discouraging exports and keeping chromium ore available for the domestic downstream industry.

(253) These export volumes have to be seen against the background of the domestic consumption of the chromium ore. The GOI provided in its reply only limited data concerning export of the chromium ore (Financial Year (FY) 18/19 (*[3]*) to FY20/21), while figures with regard to domestic production and consumption were provided for FY16/17 to FY19/20. Both sets of those data are provided as confidential. Nevertheless, it is possible to make on this basis a comparison of export volumes and domestic consumption for the FY19/20 (covering 9 months of the IP) and the preceding FY18/19. In both financial years, export volumes did not exceed 1,5 % of domestic consumption. It should be noted that share of imports in the domestic market was also insignificant.

(254) The Indian market also showed a constant and irrational overcapacity of production compared to the sum of domestic consumption and exports minus imports. This overcapacity could not be explained by the GOI. Such overcapacity can thus only be explained by the fact that the mining companies, including OMC in the chromium ore sector, exercised governmental functions (in particular, ensuring the adequate supply of chromium ore in line with the GOI's policy objectives to support the downstream industry and add value to the supply chain).

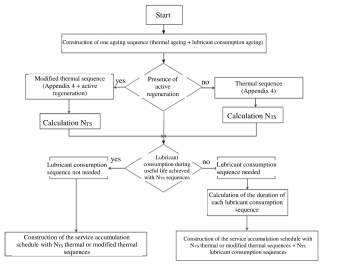
(255) Second, the Commission found that according to the GOI Minerals Concession Rules of 2016, captive mines have a right of first refusal when their mining lease expires. Article 8a of the Mines and Mineral Development Act of 1957 (as amended in 2015), also provides for extended periods of mining leases for captive users and reiterates the right of first refusal in auctions of mining leases for captive users. This allows mining leases to remain predominantly in the hands of captive downstream ferrochrome producers, in line with the statements in recital (244) above. *De facto*, the Commission also found, as mentioned in recital (236) above, that all the private chromium ore mining owners were

(*[3]*) 'FYx/x+1' covers period from 1 April of the year x to 31 March of the year x+1.

L 14/458 EN Official Journal of the European Union 16.1.2023

Annex 13 - Appendix 9

Flowchart on the performance of the service accumulation schedule



20.5.2022 EN Official Journal of the European Union C 205/85

Thursday 11 November 2021

Amendments 35 and 253

Proposal for a regulation

Recital 18

Text proposed by the Commission	*Amendment*
(18) The recognition of public health emergency situations and the legal effects of this recognition provided by Decision No 1082/2013/EU should be broadened. To this end, this Regulation should allow for the Commission to formally recognise a public health emergency at Union level. In order to recognise such an emergency situation, the Commission should establish an independent advisory committee that will provide expertise on whether a threat constitutes a public health emergency at Union level, and advise on public health response measures and on the termination of this emergency recognition. The advisory committee should consist of independent experts, selected by the Commission from the fields of expertise and experience most relevant to the specific threat that is occurring, representatives of the ECDC, of the EMA, and of other Union bodies or agencies as observers. Recognition of a public health emergency at Union level will provide the basis for introducing operational public health measures for medical products and medical devices, flexible mechanisms to develop, procure, manage and deploy medical countermeasures as well as the activation of support from the ECDC to mobilise and deploy outbreak assistance teams, known as 'EU Health Task Force'.	(18) The recognition of public health emergency situations and the legal effects of this recognition provided by Decision No 1082/2013/EU should be broadened. To this end, this Regulation should allow for the Commission to formally recognise a public health emergency at Union level. In order to recognise such an emergency situation, the Commission should establish an independent advisory committee that will provide expertise on whether a threat constitutes a public health emergency at Union level, and advise on public health response measures and on the termination of this emergency recognition. The advisory committee should consist of independent experts, ***representatives of health and care workers, including nurses and medical doctors, and representatives of civil society***, selected by the Commission from the fields of expertise and experience most relevant to the specific threat that is occurring, representatives of the ECDC, of the EMA, ***of the HERA*** and of other Union bodies or agencies as observers. ***All members of the Advisory Committee should provide declarations of interest. The advisory committee should work in close cooperation with national advisory bodies.*** Recognition of a public health emergency at Union level will provide the basis for introducing operational public health measures for medical products and medical devices, ***Union expert control mechanisms,*** flexible mechanisms to develop, procure, manage and deploy medical countermeasures ***through the HERA*** as well as the activation of support from the ECDC to mobilise and deploy outbreak assistance teams, known as 'EU Health Task Force'. ***The recognition of a public health emergency may trigger the activation of the framework set out in a Council Regulation on a framework of measures for ensuring the supply of crisis-relevant medical countermeasures in the event of a public health emergency at Union level. That framework should remain operational for an initial period of 6 months, renewable as long as the public health emergency exists.***

L 147/12 EN Official Journal of the European Union 7.6.2023

(2) the corresponding weighted average import prices of the product under review from India from Eurostat, established on a CIF basis, excluding the anti-dumping and countervailing duties, with appropriate adjustments for customs duties and post-importation costs. In the absence of any other information, these costs were estimated at 1 % of the CIF value.

(51) This resulted in the average undercutting margin of 33,2 %. These prices were also considered as a reasonable indicator of future possible price levels should measures be repealed.

(52) After disclosure the Government of India (GOI), argued that the undercutting margin calculated by the Commission is not representative as it is not based on actual market prices in the Union. In GOI's view the fact that sales sourced from GrafTech Iberica were made pursuant to long-term contracts (LTAs) have resulted in an artificially high selling price which is not indicative of the market price of GES in the Union.

(53) The Commission disagrees with this assessment. LTAs are not an uncommon commercial practice and a business decision in which a customer accepts to tie itself to a particular price level in exchange for a security of supply. Considering that LTAs were being used in dealings on the relevant market their presence in the sample does not render the sample not representative. Moreover, Only part of the sales made by GrafTech were covered by LTAs while the sales of the other two sampled Union producers in the review investigation period were not covered by similar LTAs. Therefore, in the Commission's view, whilst representative of a part of the market, overall the LTAs used by GrafTech Iberica did not significantly affected the undercutting calculation. Therefore, this claim was dismissed.

5.3.3. *Imports from third countries other than India*

(54) The imports of the product under review from other third countries were mainly from China, Mexico and Russia.

(55) The volume of imports from other third countries, as well as the market share and price trends developed over the period considered as follows:

Table 4

Imports from third countries other than India

Country		2018	2019	2020	Review investigation period
PRC	Volume (tonnes)	22 054	19 284	20 074	26 065
	Index	100	87	91	118
	Market share (%)	14	16	20	19
	Index	100	111	139	131
	Average price	10 875	5 253	2 337	2 614
	Index	100	48	21	24
Russia	Volume (tonnes)	1 076	3 229	780	3 371
	Index	100	300	72	313
	Market share (%)	1	3	1	2
	Index	100	381	111	348

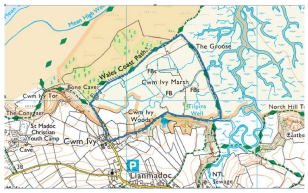
L 188/54 EN Official Journal of the European Union 27.7.2023

3) Cwm Ivy saltmarsh - as shown outlined dark blue on Map B and Map C below.

Co-ordinates
OS X (Eastings) 244220
OS Y (Northings) 194094
Nearest Post Code SA3 1DL
Lat (WGS84) N51:37:27 (51.624043)
Long (WGS84) W4:1:1:07 (-4.251832)
Lat,Long 51.624043,-4.251832
Nat Grid SS4442940 / SS4422094094

Map C

Cwm Ivy salt marsh



5. **Link with the geographical area**

'Gower Salt Marsh Lamb' has the reputation of being a quality product that is born reared and slaughtered on the Gower peninsular in South Wales. It is a natural seasonal product available from June until the end of December.

The lamb is produced from an extensive traditional system of farming where the characteristics and qualities of the final product are influenced by the following 2 key contributory factors:

— The lamb's natural diet which consists of grazing the unique saltmarsh vegetation found on the north Gower coastline.

— Knowledge and skill base of producers which has developed and remained relatively unchanged for generations.

These factors provide a strong link between the geographical area and the final product and contribute to the unique flavour and characteristics of the product.

Figure 2: Examples of the visual diversity in the collection, including embedded images, scientific charts, figures such as flowcharts, multicolumn text, outline format with tables, and images such as maps.

task is to accurately translate the document I provide into English while preserving the structure and meaning of the original text as literally as possible."

user: <image>

user: "Translate all of the text in this document into English, including the text of any headers, body text, figures, tables, and footnotes. Non-linguistic text like proper names, numbers, identifiers, and punctuation should be preserved as much as possible but transliterated into Latin characters if necessary. Output only the text of the document exactly as it appears, but translated so that a person who only knows English can understand it."

This prompt was created using guidance from the Anthropic documentation with manual adjustment based on observed initial failure cases (such as omitting header and footer text).[7] There is considerable room for continued prompt tuning in future work. In particular, we note that the prompt does not specify the source language even though this information was available for each document, and we did not perform a rigorous search or evaluation of many prompt alternatives, which can greatly affect the performance of LLMs.

In order to keep the images within the size limits supported by the service, we used the pre-computed 300dpi renderings but resized the longest edge to 1280 pixels before uploading, for an effective resolution of approximately 110dpi.

We limited all decodings to a maximum of 2048 tokens. All examples from the test set fit within this limit. Furthermore, we set the decoding temperature to 0.2 following existing machine translation examples from Anthropic.

4.2 MT Systems

In our benchmark studies we relied on NLLB-200, a multilingual translation system from Meta (NLLB Team et al., 2022). Specifically, we employed the NLLB-200 3.3 billion parameter model that is quantized to 8int for fast inference with *ctranslate2*[8]. We chose NLLB because it is supports the languages in the dataset: a single model simplifies the implementation, but we note that it may be possible to further improve the MT system by doing language-specific fine-tuning (Tang et al., 2021) or domain adaptation (Verma et al., 2022).

4.3 Metrics

Conventional MT evaluation metrics such as BLEU (Papineni et al., 2002) are not directly applicable to our dataset because the atomic unit of operation is an entire page, not a sentence. Specifically, for a given page, the ground-truth reference extracted by *pdftotext* may contain n lines, whereas the output of an OCR-MT system may be m lines. Different OCR-MT systems may obtain different numbers of lines. It is non-trivial to automatically re-stitch lines in OCR-MT output into linguistically-valid sentences and align to the n reference lines.

Therefore, we propose to use Page-Level BLEU, where all lines from each page are concatenated and treated as a single long "sentence" for the purpose of alignment between reference and hypothesis. If there are k pages in a dataset ($k = 1119$ for our testset), then we first re-organize the n lines of reference and m lines of hypothesis both into k long lines. Then we run the standard BLEU metric using *SacreBLEU* (Post, 2018), treating each page as if it were a sentence. Pseudocode for this processing is shown in Algorithm 1.

This kind of page-level scoring is also employed in other OCR tasks like reading order detection (Wang et al., 2021). Some researchers[9] use the term "Document-level BLEU" to refer to what we call "Page-Level BLEU." We think they are interchangeable terms but we prefer "page" to emphasize the fact that single pages rather than full-length multi-page documents are being scored. Other page-level translation metrics based on COMET or TER are also conceivable, but they would require substantial computation to calculate due to the long lines.

While Page-Level BLEU is our primary metric for evaluating OCR-MT systems, we propose to use Page-Level Character F-score (chrF) to eval-

[7]Despite our best efforts, the API refused to decode one page in the test partition for 6 of the 23 languages. As this amounts to less than 1/1000th of the data we considered this inconsequential.

[8]Model: https://huggingface.co/OpenNMT/nllb-200-3.3B-ct2-int8, Example: https://forum.opennmt.net/t/nllb-200-with-ctranslate2/5090

[9]See: ICDAR25 Competition on End-to-End Document Image MT: https://cip-documentai.github.io

Algorithm 1 Page-Level BLEU

Require: Reference_File ▷ n lines from our dataset
Require: Hypothesis_File ▷ m lines from OCR-MT system

```
procedure ConcatLines(File)
    L = { }                                  ▷ initialize dictionary
    for all line in File do
        i = GetPageId(line) ▷ which page the line belongs
        L[i] = StringConcat(L[i], line)
    end for
    ids = sort(L.keys())          ▷ Get sorted list of page ids
    return list([L[i] for i in ids])   ▷ k lines, k =len(ids)
end procedure
Ref_Lines = ConcatLines(Reference_File)
Hyp_Lines = ConcatLines(Hypothesis_File)
return SacreBLEU(Ref_Lines, Hyp_Lines)
```

uate the accuracy of the OCR component. The computation is similar to Page-Level BLEU, except that BLEU is swapped with chrF (Popović, 2015) which focuses more on character matching. Other metrics like page-level character error rate also conceivable. chrF is defined as:

$$chrF = (1 + \beta^2)\frac{chrP \cdot chrR}{\beta^2 chrP + chrR} \quad (1)$$

where $chrP$ is percentage of character n-grams in the hypothesis which match reference and $chrR$ is the percentage of character n-grams in the hypothesis which are also in the hypothesis. We set $n = 6$ and $\beta = 2$.[10]

5 Experimental Results

To demonstrate the utility of *OJ4OCRMT* and to document the performance attainable by contemporary OCR-MT systems we report several experimental results. We studied: (a) translation into English (Section 5.1); (b) OCR-MT performance using images of differing quality (Section 5.2); and, (c) multilingual translation between any source/target language pair (Section 5.3).

5.1 Primary Benchmark: xx→en

We encourage researchers to focus on translation into English (xx→en) as the main benchmark for this dataset. This is for two reasons:

1. With fewer resources for training OCR models in non-English documents, this task is more challenging and deserves more research.
2. Translation into the same English side in this multi-parallel dataset facilitates comparison across test sets. For example, we can compare the Page-Level BLEU scores of the fr→en testset with that of the de→en testset because they are based on the same reference.

Table 4 shows the Page-Level BLEU scores of various OCR-MT systems. We compare 4 systems:

(a) Reference transcription translated by NLLB

(b) Cascade: Tesseract OCR + NLLB MT

(c) Cascade: EasyOCR + NLLB MT

(d) End-to-End: Direct translation by Claude

For example, in the bg→en task, translating the ground truth Bulgarian reference using NLLB gives 49.5 Page-Level BLEU, whereas using the same translation model on Tesseract OCR outputs in a cascaded fashion gives 38.8 Page-Level BLEU; the EasyOCR+NLLB cascade gives 22.5 Page-Level BLEU and the degradation can be attributed to OCR performance differences. The end-to-end Claude system gives very strong 49.3 Page-Level BLEU.

For all language pairs, we observe a performance degradation when using automatic OCR in cascades, suggesting that this is an interesting dataset for understanding the impact of OCR errors on MT.[11] Generally, Tesseract cascades appear to perform better than EasyOCR cascades, but there is still a sizeable gap when compared with the OCR reference translation. The end-to-end system achieves very competitive scores and sometimes even outperforms reference translation (*e.g.,* 49.6 vs. 44.6 for cs→en). There are two hypotheses: (a) Claude has strong OCR, MT, or OCR-MT performance in this domain, or (b) Claude may have been exposed to similar kinds of governmental documents during training.

5.2 Different Image Quality

We are also interested in understanding how degradation in image quality impacts OCR-MT. As previously mentioned, we converted the PDFs into images at 300, 150, and 72dpi. For each resolution we ran the two cascades: systems (b) and (c) described above. We then measure the performance

[10] We use the SacreBLEU toolkit, with signatures:
BLEU= `nrefs:1|case:lc|eff:yes|tok:13a|smooth:exp|version:2.4.0`
chrF= `nrefs:1|case:lc|eff:yes|nc:6|nw:0|space:no|version:2.4.0`

[11] We note the lv→en results are low across the board. This appears to be an issue in the translation model (rather than the OCR component), which severely hallucinates on lv input.

	bg	cs	da	de	el	es	et	fi	fr	hr	hu	it	lt	lv	mt	nl	pl	pt	ro	sk	sl	sv
(a)	49.5	44.6	47.4	45.4	47.4	54.9	45.7	44.9	52.4	40.0	44.2	51.0	44.0	12.7	57.9	49.8	44.6	53.1	50.2	46.8	47.4	51.9
(b)	38.8	39.7	40.6	37.4	37.5	47.6	37.0	35.8	43.2	29.5	37.3	42.5	35.8	2.4	50.1	41.0	40.3	44.2	43.8	40.7	40.1	44.9
(c)	22.5	35.6	35.4	32.5	–	41.0	32.7	–	36.8	27.7	33.7	36.5	31.6	1.4	42.8	36.5	34.0	40.3	39.5	36.5	36.1	39.0
(d)	49.3	49.6	48.8	49.5	49.5	53.7	35.7	39.8	52.1	50.6	43.9	53.4	38.1	38.6	46.3	52.3	48.8	53.9	52.8	50.3	48.5	47.0

Table 4: Page-level BLEU results for the main benchmark: Translate into English, 300dpi setting. System (a) is the result of translating reference transcripts in the source language with the NLLB model. System (b) is a cascade of Tesseract and NLLB. System (c) is a cascade of EasyOCR and NLLB. System (d) is a VLM, Claude, run in an end-to-end fashion to directly translate into English from images.

	bg	cs	da	de	el	en	es	et	fi	fr	hr	hu	it	lt	lv	mt	nl	pl	pt	ro	sk	sl	sv
bg	–	34.4	36.2	31.6	38.7	49.5	39.1	26.7	25.8	34.2	21.2	23.1	38.3	30.3	3.1	28.4	34.1	31.5	40.6	40.2	33.6	34.1	32.8
cs	35.6	–	32.8	28.1	35.2	44.6	34.7	23.6	23.6	30.6	19.5	21.0	34.3	27.2	2.7	25.5	30.8	27.4	35.4	34.8	35.0	32.3	27.9
da	36.0	32.4	–	28.8	36.3	47.4	36.4	24.4	24.7	32.2	21.3	21.5	35.3	27.7	3.5	26.6	33.5	28.1	34.7	36.2	30.3	30.2	31.5
de	35.7	32.3	33.1	–	33.6	45.4	34.9	24.8	24.0	30.9	20.5	22.5	34.7	27.3	3.4	25.4	32.8	27.3	34.8	34.0	30.7	30.8	29.8
el	37.1	32.2	34.1	29.0	–	47.4	36.8	24.0	23.8	32.6	19.6	20.9	36.8	27.6	2.9	28.0	32.7	28.0	37.3	37.0	31.9	31.0	29.1
en	54.2	48.9	52.7	46.8	52.3	–	57.0	38.8	36.6	47.4	37.9	42.7	53.4	41.2	2.7	49.1	51.5	47.3	56.7	56.7	51.0	48.8	49.6
es	40.8	35.1	37.3	32.3	40.7	54.9	–	26.8	26.0	38.4	21.8	25.1	43.4	30.4	3.7	31.8	37.1	30.8	45.0	43.1	34.7	35.2	34.5
et	36.3	32.0	35.3	30.2	35.8	45.7	36.1	–	26.5	31.6	21.2	24.0	35.0	29.6	3.3	27.1	33.3	28.9	36.3	35.2	32.1	32.0	31.4
fi	34.7	31.1	33.3	29.3	34.6	44.9	35.1	26.4	–	31.5	20.4	23.7	34.1	27.7	3.5	26.2	32.1	27.6	34.7	33.9	30.5	31.0	31.0
fr	40.3	35.4	38.1	33.6	39.4	52.4	45.1	27.9	26.6	–	25.7	26.8	43.0	30.9	4.7	30.3	38.0	32.5	43.6	43.8	34.6	35.5	34.2
hr	27.6	21.2	24.0	18.3	27.0	40.0	24.6	16.2	17.5	22.6	–	10.3	26.2	18.8	1.7	18.5	20.2	16.5	25.2	24.6	20.0	23.3	18.4
hu	34.2	30.4	32.5	28.5	34.3	44.2	33.7	24.1	23.8	29.5	17.8	–	32.8	26.4	3.0	25.3	30.7	26.3	34.3	33.1	28.8	30.4	29.1
it	38.5	33.7	35.3	30.2	39.4	51.0	41.5	25.5	25.0	37.7	22.4	22.6	–	28.7	4.1	29.3	35.7	30.6	40.9	42.0	3.9	33.0	31.4
lt	35.6	31.3	32.2	27.6	34.1	44.0	34.7	24.7	24.0	30.5	20.2	22.3	34.2	–	3.3	25.5	31.0	27.6	34.8	34.2	30.8	29.9	28.3
lv	0.8	0.6	0.7	0.7	0.8	12.7	1.2	0.5	22.2	26.8	25.5	23.9	0.9	26.1	–	21.0	27.5	25.8	29.6	28.6	26.0	25.7	25.3
mt	40.4	34.2	38.2	31.3	40.5	57.9	41.6	25.4	25.2	36.8	19.0	22.5	40.9	28.6	4.1	–	34.4	30.2	42.0	42.4	34.7	35.4	33.4
nl	38.1	33.5	35.1	31.2	36.9	49.8	38.8	25.7	25.3	34.4	22.5	23.4	38.6	28.9	3.5	28.7	–	29.8	37.7	39.5	32.1	32.3	33.5
pl	36.3	33.0	32.8	28.9	35.4	44.6	35.3	24.3	23.5	30.9	21.0	21.8	35.4	28.0	2.9	25.7	31.3	–	35.0	35.2	31.4	32.1	28.4
pt	40.8	34.6	37.5	32.7	40.2	53.1	46.4	26.6	26.4	38.0	19.8	23.9	43.6	29.6	3.3	31.8	36.1	30.9	–	42.9	34.2	35.2	33.9
ro	38.9	32.4	35.3	28.7	39.0	50.2	39.5	24.1	23.8	35.3	18.6	19.5	39.6	27.8	2.7	30.1	32.3	28.4	40.2	–	32.1	31.6	29.8
sk	37.2	37.6	34.2	28.2	35.9	46.8	35.5	24.0	23.6	31.5	20.8	21.4	34.5	27.9	2.9	26.6	31.4	28.6	35.3	36.1	–	32.5	28.5
sl	37.2	33.7	35.0	30.3	36.8	47.4	36.1	25.1	24.8	32.5	23.8	21.5	36.9	28.9	3.5	27.4	32.7	29.9	36.2	36.9	33.7	–	30.5
sv	38.7	33.6	39.0	31.9	38.2	51.9	39.3	24.2	26.5	35.1	20.7	23.6	38.2	28.5	3.2	29.3	35.1	28.0	38.0	37.8	32.7	33.8	–

Table 5: Page-level BLEU results for all language pairs using System (a): Reference transcript with NLLB translation. Rows are source languages and columns are target languages. 300dpi setting.

	bg	cs	da	de	el	en	es	et	fi	fr	hr	hu	it	lt	lv	mt	nl	pl	pt	ro	sk	sl	sv
bg	–	27.2	30.5	24.8	28.9	38.8	35.2	22.2	19.9	27.5	19.8	18.7	30.4	24.6	1.5	21.5	28.5	25.4	32.9	32.5	28.5	29.8	27.7
cs	33.3	–	32.0	27.6	30.3	39.7	35.7	23.5	21.9	28.3	20.0	19.4	31.7	26.0	2.6	22.5	30.6	27.1	34.5	33.3	35.0	32.2	29.2
da	32.9	30.3	–	27.9	30.1	40.6	36.3	24.7	23.7	28.5	21.6	20.1	33.0	26.7	3.2	24.0	31.9	27.0	35.2	34.0	30.6	30.7	32.7
de	31.7	29.5	32.4	–	29.0	37.4	34.3	23.8	22.6	26.8	19.9	19.9	31.1	25.1	3.2	22.0	30.8	25.8	33.4	32.0	29.3	30.5	29.4
el	30.6	26.5	27.4	23.6	–	37.5	33.3	18.2	18.6	27.3	19.6	16.6	30.0	21.9	1.1	21.2	27.1	22.9	30.9	31.3	25.6	25.8	24.1
en	45.5	41.6	45.9	38.0	42.7	–	51.8	31.0	31.0	40.5	29.2	32.2	46.0	34.9	2.7	40.5	43.5	37.8	49.6	48.2	43.9	43.1	42.6
es	37.8	33.3	36.4	31.0	36.1	47.6	–	25.0	26.6	35.2	21.9	23.8	40.8	28.6	3.6	30.3	35.6	30.8	43.1	40.8	33.5	34.1	33.7
et	30.0	27.6	31.4	25.6	28.1	37.0	32.8	–	23.0	25.7	19.9	22.2	29.8	26.5	3.4	21.0	28.9	25.5	31.5	29.9	28.6	29.6	28.2
fi	29.3	27.0	30.6	25.1	27.3	35.8	32.2	23.9	–	26.0	20.1	21.6	29.1	24.8	3.4	21.0	27.9	24.7	30.9	29.2	27.5	28.9	28.4
fr	34.3	31.0	34.8	29.2	32.3	43.2	41.6	26.3	24.2	–	25.0	24.1	37.9	27.7	4.8	27.4	33.7	29.7	40.2	37.9	32.2	33.3	31.7
hr	26.6	20.2	22.5	17.8	24.0	29.5	24.8	16.0	17.0	20.4	–	8.7	24.2	18.4	1.6	16.7	19.8	15.3	25.1	23.6	18.6	22.2	18.3
hu	30.6	27.8	30.6	25.5	28.4	37.3	32.5	23.9	22.2	25.7	17.6	–	29.9	25.1	3.0	21.7	28.2	24.8	32.2	30.5	27.3	28.8	27.0
it	34.2	31.2	34.3	28.5	32.2	42.5	40.1	25.4	23.9	32.1	24.2	22.1	–	27.3	4.5	26.2	33.2	28.5	38.8	37.2	32.2	33.5	31.5
lt	28.7	25.6	28.4	23.1	25.6	35.8	31.8	22.0	19.9	25.0	19.2	19.7	27.5	–	2.4	20.4	26.7	23.5	30.9	28.2	26.3	27.2	25.2
lv	1.1	0.8	1.0	1.0	1.0	2.4	1.8	0.6	0.6	1.6	0.9	0.8	1.1	0.7	–	0.7	1.3	0.8	1.4	1.1	0.8	0.9	0.9
mt	34.3	29.5	34.2	27.2	33.2	50.1	39.7	23.4	22.0	30.9	17.5	19.2	35.3	26.0	2.3	–	30.8	26.9	38.0	36.5	31.1	32.7	30.8
nl	32.4	29.4	33.8	28.5	29.7	41.0	36.6	24.2	23.1	29.3	21.8	22.0	33.0	25.9	3.1	23.7	–	27.1	34.8	33.7	30.3	31.4	30.9
pl	31.7	27.9	28.2	27.5	30.5	40.3	34.6	20.4	20.9	29.3	19.3	17.9	29.8	23.3	3.2	22.0	27.9	–	32.2	30.4	26.2	26.4	25.9
pt	34.9	30.2	34.8	28.6	32.9	44.2	41.9	25.1	22.9	32.9	22.2	22.7	37.8	27.2	3.2	26.6	31.9	28.2	–	37.3	32.3	33.1	31.0
ro	35.3	30.7	34.4	28.1	32.7	43.8	39.8	24.2	22.5	32.0	19.7	19.7	36.6	26.8	2.7	26.4	31.8	27.4	39.0	–	32.2	33.0	30.9
sk	33.1	34.0	32.8	27.0	30.3	40.7	36.0	23.7	21.2	28.1	21.2	20.2	31.7	25.7	2.5	22.5	29.9	27.2	34.7	33.0	–	32.9	28.8
sl	33.2	30.4	31.7	27.7	30.4	40.1	35.0	23.4	22.6	27.3	23.3	20.1	31.3	26.5	3.2	22.6	30.1	26.8	34.5	32.7	31.8	–	29.6
sv	35.0	31.9	39.0	30.1	33.1	44.9	38.7	24.3	25.5	31.8	22.0	22.4	35.6	27.9	3.1	26.1	34.0	28.4	37.2	35.6	32.4	33.9	–

Table 6: Page-level BLEU results for all language pairs using System (b): Tesseract OCR with NLLB translation. Rows are source languages and columns are target languages. 300dpi setting.

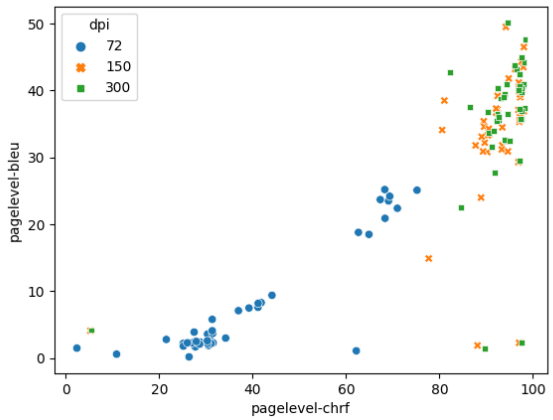


Figure 3: Scatterplot of Cascade systems under different image inputs (300dpi, 150dpi, and 72dpi). The x-axis is the chrF of the OCR engine, and the y-axis is the BLEU score representing final OCR-MT performance.

of the OCR component using Page-Level chrF and the final OCR-MT performance using Page-Level BLEU. A scatterplot is shown in Figure 3.

As can be seen, there is a strong correlation between the OCR chrF and OCR-MT BLEU scores. For example, TesseractNLLB for cs→en gives 97.6 chrF and 39.7 BLEU at 300dpi, drops to 97.3 chrF and 39.0 BLEU at 150dpi, and then further decreases to 69.1 chrF and 23.5 BLEU at 72dpi. We believe that releasing images with different resolutions will foster additional OCR-MT research.

Note there are also several failure cases in the 72dpi setting, which appear to be difficult for both Tesseract and EasyOCR. For those systems with OCR chrF under 50, the OCR transcripts are basically unreadable and the MT component hallucinates. This explains the Page-Level BLEUs under 10 in Figure 3.

5.3 Multilingual Evaluation

Since our dataset is multi-way parallel, a massively multilingual evaluation for all pairs of $23 \times 22 = 506$ language directions is feasible. We report Page-level BLEU results for System (a) in Table 5 and for System (b) in Table 6.[12] We hope this will encourage research that is not English-centric.

5.4 Error Analysis

We present examples in Table 7 to illustrate the successes and failures of one of the cascade systems. The English reference for the three examples can be seen in Figure 2. Due to space limitations, we only show an excerpt of the Tesseract OCR output and NLLB MT output for each page.

In Example 1, we observe some critical OCR errors in the lower dpi case: “Stammt aus” was mistranscribed as “53mm us” and “Landwirtschaftssystem” was mistaken as “Landwirtschaftssyster”, and the error propagation resulted in an incomprehensible translation. In contrast, in Example 2, there are also critical mistakes in OCR, but interestingly the translation still contained some of the gist. Mistranscription of “Drittländern” into “Drinländer” changed the translation from “third countries” to “non-member countries.” In terms of BLEU n-gram calculation, the main noun “countries" was translated correctly.

Example 3 is the page with the flowchart in Figure 2. It contains some complications due to layout analysis and reading order. For both 72 and 300 dpi examples, we observe that the header (“L 14/438 Amtsblatt der Europäischen Union...”) has not been sentence-split from the following caption of the flowchart (“Ablaufdiagramm für das ...” / “Flowchart on the ...”). These kinds of sentence splitting issues can impact MT significantly, especially if it expects well-formed sentences. Interestingly, the header is entirely ignored by NLLB in the 300dpi case. Additionally, the 72dpi version does not output lines in the same order as the 300dpi version, in particular jumping to generate the box containing the word “Berechnung” soon after the “Start” box of the flowchart. This resulted in significantly different translations between the two image resolutions.

We can also analyze translation quality according to page type, since the dataset contains annotations indicating which pages contain figures or tables. These more complex layouts may present additional challenges to an OCR-MT system. Table 8 shows the Page-Level BLEU of the Tesseract-NLLB cascade broken down by different subsets. For example, in bg→en, we observe that the performance on pages with tables (24.8) is 14.0 points lower than the performance on the full testset; similarly, performance on pages with figures degrades 14.6 points. Generally, across all language pairs, we observe that pages with tables or figures tend to be substantially more challenging.

[12]We did not run the Claude model for all these pairs due to the computational expense.

DPI	Tesseract output (transcript or translation)
Example 1: https://eur-lex.europa.eu/legal-content/DE/TXT/PDF/?uri=OJ:L:2023:188:FULL#page=36	
72	**OCR:** Das Lamunfleisch 53mm us einem extensiven traditionellen Landwirtschaftssyster. **MT:** The meat 53mm us an extensive traditional agricultural sister.
300	**OCR:** Das Lammfleisch Stammt aus einem extensiven traditionellen Landwirtschaftssystem. **MT:** Lamb comes from an extensive traditional farming system.
Example 2: https://eur-lex.europa.eu/legal-content/DE/TXT/PDF/?uri=OJ:L:2023:147:FULL#page=14	
72	**OCR:** Einführen aus anderen Drinländer als Im Die Einfuhren der überprüften Ware aus anderen Drinlindern stammten hauptsächlich aus China, Mexiko und Russland **MT:** Imports from non-member countries of the three countries The imports of the product under review from non-member countries of the three countries were mainly from China, Mexico and Russia
300	**OCR:** Einfuhren aus anderen Drittländern als Indien (54) Die Einfuhren der überprüften Ware aus anderen Drittländern stammten hauptsächlich aus China, Mexiko und Russland. **MT:** Imports from third countries other than India (54) Imports of the product under review from other third countries were mainly from China, Mexico and Russia.
Example 3: https://eur-lex.europa.eu/legal-content/DE/TXT/PDF/?uri=OJ:L:2023:014:FULL#page=440	
72	**OCR:** Lies Amwbia der Europäischen Union 16.2023 Ablaufdiagramm für das Betriebsakkumulationsprogramm Start Berechnung Nrs Abfolge (Anlage 4 + aktive... **MT:** 16.1.2023 European Union Amw blue Example of a farm accumulation programme with chemical, lubricant consumption and engine speed and cabinet regeneration data Example of a farm ...
300	**OCR:** L 14/438 Amtsblatt der Europäischen Union 16.1.2023 Anhang 13 - Anlage 9 Ablaufdiagramm für das Betriebsakkumulationsprogramm Start ; Aufbau einer Alterungsabfolge (thermische Alterung + Schmiermittelverbrauchsalterung) ... **MT:** Annex 13 - Appendix 9 - Schedule of the operational accumulation programme Start; construction of an ageing sequence (thermal ageing + lubricant consumption ageing) ...

Table 7: Example outputs from OCR component and OCR-MT cascade for de→en. The English version of the images are in Figure 2. The original German PDF can be accessed online with the provided URL. Interesting errors are highlighted in red and discussed in Section 5.4.

	all	table	Δ	figure	Δ
bg	38.8	24.8	(-14.0)	24.2	(-14.6)
cs	39.7	25.1	(-14.6)	30.5	(-9.2)
da	40.6	33.0	(-7.5)	37.8	(-2.7)
de	37.4	27.3	(-10.1)	30.5	(-6.9)
el	37.5	28.9	(-8.6)	22.3	(-15.2)
es	47.6	41.3	(-6.3)	33.9	(-13.7)
et	37.0	32.8	(-4.2)	26.9	(-10.1)
fi	35.8	28.0	(-7.8)	25.6	(-10.1)
fr	43.2	29.5	(-13.7)	30.7	(-12.5)
hr	29.5	27.2	(-2.3)	24.2	(-5.3)
hu	37.3	26.8	(-10.4)	27.0	(-10.2)
it	42.5	35.2	(-7.3)	31.7	(-10.9)
lt	35.8	22.1	(-13.6)	23.4	(-12.4)
lv	2.4	4.6	(+2.2)	1.2	(-1.1)
mt	50.1	29.1	(-21.0)	41.2	(-8.9)
nl	41.0	25.5	(-15.5)	32.5	(-8.4)
pl	40.3	35.7	(-4.5)	28.2	(-12.1)
pt	44.2	26.7	(-17.5)	34.0	(-10.2)
ro	43.8	26.1	(-17.6)	34.6	(-9.2)
sk	40.7	25.2	(-15.5)	33.2	(-7.6)
sl	40.1	31.7	(-8.4)	29.5	(-10.6)
sv	44.9	24.3	(-20.5)	33.0	(-11.9)
avg.	38.6	27.7	(-10.9)	28.9	(-9.7)

Table 8: Page-level BLEU breakdown by page type (pages with tables or figures), for the Tesseract-NLLB cascade. xx→en, 300dpi setting. Δ shows difference in Page-Level BLEU when compared to the all pages in the testset.

6 Conclusions

We have created a new dataset, *OJ4OCRMT*, which adds to the set of resources available for assessing the performance of document image translation systems. The dataset is large, supports 23 languages and 3 writing systems, and contains interesting visual layouts of natural documents in the government domain. We reported benchmark experiments on this translation task using two cascaded systems and one VLM-based end-to-end system. Some of our findings include: (a) the VLM system (Claude) generally outperformed the cascaded systems; (b) Tesseract generally outperformed EasyOCR; (c) the OCR models performed poorly on 72dpi images; and, (d) the presence of tables or figures in images led to poorer translation quality.

The dataset can be obtained from https://huggingface.co/hltcoe.

Sustainability statement

As the focus of this paper is on developing a dataset and demonstrating its utility for evaluating document image translation, it was not necessary to train models. Consequently, our electrical consumption was fairly small for the work described in this paper. By compiling and releasing a reusable dataset we hope to save other researchers effort.

We will nevertheless attempt to estimate carbon footprint associated with this project. For the end-to-end translation experiments using Anthropic's Claude model, we note that the Claude 3 Model Card claims that Anthropic purchases sufficient carbon credits to offset their consumption each year (Anthropic, 2024).

For the cascade systems, our NLLB inference on V100 GPU's takes approximately 0.5 hours on each test set. We estimate 1200 test decodes, so that is 600 GPU-hours in total. The EasyOCR decodes cost around 300 GPU-hours. If we use 250 watts as the rating for a V100, then given a total 900 GPU-hours that is 0.23 MWh of electricity usage. If we assume a CO_2e emission of 432 kg/MWh and data center power usage effectiveness (PUE) of 1.5, then the CO_2e emission is guesstimated to be: $1.5 \times \frac{0.23\,\text{MWh}}{1} \times \frac{432\,\text{kg}}{\text{MWh}} = 150\,\text{kg}$. In addition, our CPU usage for ersatz sentence splitting and Tesseract OCR is estimated to be at 200 hours, corresponding to 39kg CO_2e in total.

References

Anthropic. 2024. Claude 3 model card. Technical report, Anthropic. Last updated: October 22, 2024. Accessed: January 30, 2025.

Michael Arrigo, Stephanie Strassel, Nolan King, Thao Tran, and Lisa Mason. 2022. CAMIO: A corpus for OCR in multiple languages. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1209–1216, Marseille, France. European Language Resources Association.

Oana Ignat, Jean Maillard, Vishrav Chaudhary, and Francisco Guzmán. 2022. OCR improves machine translation for low-resource languages. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1164–1174, Dublin, Ireland. Association for Computational Linguistics.

Seorin Kim, Julien Baudru, Wouter Ryckbosch, Hugues Bersini, and Vincent Ginis. 2025. Early evidence of how llms outperform traditional systems on ocr/htr tasks for historical records. *Preprint*, arXiv:2501.11623.

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

Zhibin Lan, Jiawei Yu, Xiang Li, Wen Zhang, Jian Luan, Bin Wang, Degen Huang, and Jinsong Su. 2023. Exploring better text image translation with multimodal codebook. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3479–3491, Toronto, Canada. Association for Computational Linguistics.

Bo Li, Shaolin Zhu, and Lijie Wen. 2025. MIT-10M: A large scale parallel corpus of multilingual image translation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5154–5167, Abu Dhabi, UAE. Association for Computational Linguistics.

Yupu Liang, Yaping Zhang, Cong Ma, Zhiyang Zhang, Yang Zhao, Lu Xiang, Chengqing Zong, and Yu Zhou. 2024. Document image machine translation with dynamic multi-pre-trained models assembling. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7084–7095, Mexico City, Mexico. Association for Computational Linguistics.

Elman Mansimov, Mitchell Stern, Mia Chen, Orhan Firat, Jakob Uszkoreit, and Puneet Jain. 2020. Towards end-to-end in-image neural machine translation. In *Proceedings of the First International Workshop on Natural Language Processing Beyond Text*, pages 70–74, Online. Association for Computational Linguistics.

NLLB Team, Marta Ruiz Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield,

Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Alison Youngblood, Bapi Akula, Loïc Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon L. Spruit, C. Tran, Pierre Yves Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *ArXiv*, abs/2207.04672.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Elizabeth Salesky, Philipp Koehn, and Matt Post. 2024. Benchmarking visually-situated translation of text in natural images. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1167–1182, Miami, Florida, USA. Association for Computational Linguistics.

Ray Smith, Daria Antonova, and Dar-Shyang Lee. 2009. Adapting the tesseract open source ocr engine for multilingual ocr. In *MOCR '09: Proceedings of the International Workshop on Multilingual OCR*, ACM International Conference Proceeding Series, pages 1–8. ACM.

Zhiyi Song, Safa Ismael, Stephen Grimes, David Doermann, and Stephanie Strassel. 2012. Linguistic resources for handwriting recognition and translation evaluation. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 3951–3955, Istanbul, Turkey. European Language Resources Association (ELRA).

Ralf Steinberger, Mohamed Ebrahim, Alexandros Poulis, Manuel Carrasco-Benitez, Patrick Schlüter, Marek Przybyszewski, and Signe Gilbro. 2014. An overview of the european union's highly multilingual parallel corpora. *Lang. Resour. Eval.*, 48(4):679–707.

Ralf Steinberger, Andreas Eisele, Szymon Klocek, Spyridon Pilos, and Patrick Schlüter. 2012. DGT-TM: A freely available translation memory in 22 languages. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 454–459, Istanbul, Turkey. European Language Resources Association (ELRA).

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).

Andreas Steiner, André Susano Pinto, Michael Tschannen, Daniel Keysers, Xiao Wang, Yonatan Bitton, Alexey Gritsenko, Matthias Minderer, Anthony Sherbondy, Shangbang Long, Siyang Qin, Reeve Ingle, Emanuele Bugliarello, Sahar Kazemzadeh, Thomas Mesnard, Ibrahim Alabdulmohsin, Lucas Beyer, and Xiaohua Zhai. 2024. Paligemma 2: A family of versatile vlms for transfer. *Preprint*, arXiv:2412.03555.

Yipeng Sun, Jiaming Liu, Wei Liu, Junyu Han, Errui Ding, and Jingtuo Liu. 2019. Chinese street view text: Large-scale chinese text reading with partially supervised learning. In *Proceedings of ICCV*.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2021. Multilingual translation from denoising pre-training. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3450–3466, Online. Association for Computational Linguistics.

Ole Tange. 2024. Gnu parallel 20241222 ('bashar'). GNU Parallel is a general parallelizer to run multiple serial command line programs in parallel without changing them.

Yanzhi Tian, Xiang Li, Zeming Liu, Yuhang Guo, and Bin Wang. 2023. In-image neural machine translation with segmented pixel sequence-to-sequence model. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15046–15057, Singapore. Association for Computational Linguistics.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Neha Verma, Kenton Murray, and Kevin Duh. 2022. Strategies for adapting multilingual pre-training for domain-specific machine translation. In *Proceedings of the 15th biennial conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 31–44, Orlando, USA. Association for Machine Translation in the Americas.

Zilong Wang, Yiheng Xu, Lei Cui, Jingbo Shang, and Furu Wei. 2021. LayoutReader: Pre-training of text and layout for reading order detection. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4735–4744, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Rachel Wicks and Matt Post. 2021. A unified approach to sentence segmentation of punctuated text in many languages. 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 3995–4007, Online. Association for Computational Linguistics.

Zhiyang Zhang, Yaping Zhang, Yupu Liang, Lu Xiang, Yang Zhao, Yu Zhou, and Chengqing Zong. 2023. LayoutDIT: Layout-aware end-to-end document image translation with multi-step conductive decoder. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10043–10053, Singapore. Association for Computational Linguistics.