The EuroPat Corpus: A Parallel Corpus of European Patent Data

Kenneth Heafield[†], Elaine Farrow[†], Jelmer van der Linde[†],
Gema Ramírez-Sánchez[‡], Dion Wiggins[°]
[†]School of Informatics, University of Edinburgh, UK
kheafiel@inf.ed.ac.uk, Elaine.Farrow@ed.ac.uk, jelmer.vanderlinde@ed.ac.uk
[‡]Prompsit Language Engineering, SL (PLE), Spain
gramirez@prompsit.com,
[°]Omniscien Technologies (Trading) B.V., Netherlands
dion.wiggins@omniscien.com

Abstract

We present the EuroPat corpus of patent-specific parallel data for 6 official European languages paired with English: German, Spanish, French, Croatian, Norwegian, and Polish. The filtered parallel corpora range in size from 51 million sentences (Spanish-English) to 154k sentences (Croatian-English), with the unfiltered (raw) corpora being up to 2 times larger. Access to clean, high quality, parallel data in technical domains such as science, engineering, and medicine is needed for training neural machine translation systems for tasks like online dispute resolution and eProcurement. Our evaluation found that the addition of EuroPat data to a generic baseline improved the performance of machine translation systems on in-domain test data in German, Spanish, French, and Polish; and in translating patent data from Croatian to English. The corpus has been released under Creative Commons Zero, and is expected to be widely useful for training high-quality machine translation systems, and particularly for those targeting technical documents such as patents and contracts.

Keywords: Parallel data, Corpus, Patent, Legal, Technical translation

1. Introduction

As neural machine translation (MT) engines improve, they become more sensitive to their input data and perform better when they are trained with clean and high quality data (Carpuat et al., 2017; Koehn and Knowles, 2017). Patents are a rich source of technical vocabulary, product names, and person names that complement other data sources. Patent data covers domains across science, medicine, and engineering. Patent translations are high quality due to their explicit purpose of protecting intellectual property in courts. Millions of sentences are available, copyright protection is nonexistent or permissive, and published patents present no privacy concerns. However, patents in different languages are not exact translations, as each legal jurisdiction has different laws, and some parts of a patent may not be valid in all jurisdictions[1]. This can result in missing sections, and content added or modified to match the legal jurisdiction.

The goal of the EuroPat project was to mine parallel corpora from patents by aggregating and aligning patent data in order to prepare clean processed parallel corpora in the patent domain. We assembled patent-specific parallel corpora for 6 official European languages in parallel with English: German, Spanish, French, Croatian, Norwegian, and Polish. This is the first time parallel data in the patent domain has been made available for three of the language pairs (Croatian-English, Norwegian-English and Polish-English), and significantly enlarges the quantity of available data for three more language pairs (German-English, Spanish-English, and French-English).

The EuroPat corpus was released under Creative Commons Zero, which is as close to public domain as legally possible. Patents themselves are not subject to copyright, but the European Patent Office (EPO)[2] claims a database copyright on their collection. By processing raw data into a parallel corpus ourselves, we created a derivative product, which the EPO's licensing terms explicitly allowed us to release[3].

2. Related Work

The World Intellectual Property Organization (WIPO)[4] makes a parallel corpus of patent data available[5], which excludes data they received from partner patent offices. Naturally, much of the data for European languages is found at the EPO and national offices.

Many of the tools that were used in the present study were developed by the ParaCrawl project (Bañón et al., 2020), which mined the unstructured web for parallel data. While some patent translations were incidentally collected from the web in that project, EuroPat ex-

[1] https://www.epo.org/searching-for-patents/helpful-resources/raw-data.html

[2] https://www.epo.org/

[3] http://documents.epo.org/projects/babylon/eponet.nsf/0/130F16FEB85269BDC1257B1A005973B8/$FILE/Licensing_of_EPO_databases_agreement_sample_en.pdf

[4] https://www.wipo.int

[5] http://www.wipo.int/patentscope/en/data/#coppa

ploited substantial metadata to link high-quality translations, rather than noisy text found on the web by matching content. Moreover, EuroPat procured source data that is not accessible from the web in bulk due to rate limiting by the EPO. Even freely available patent data, such as from Poland[6], must be accessed via a web form that web crawlers do not fill out.

3. Methodology

A schematic of the full processing pipeline from raw data to aligned parallel sentences is shown in Figure 1.

3.1. Data Acquisition

English-language patent data was acquired from two main sources: the United States Patent Office (USPTO)[7], which makes patent data available to download in bulk at no cost; and the EPO, which charges a fee for bulk access and imposes licence conditions on the use of the data.

During the time-period of the EuroPat project, data was unfortunately not available for purchase directly from the German Patent Office; however, we were able to purchase a bulk collection of European patents from the EPO that had been filed in French, German, and English, allowing us to include German in our target set of language pairs. Patents that were originally filed with the patent offices in Spain, France, Croatia, Norway, and Poland were also available from the EPO via the Open Patent Service (OPS) API[8].

Many older patents have not yet been fully digitised, but scanned images of the original patent documents can be accessed through the EPO OPS API. We developed software to automate the process of querying the API and downloading the image files for patents in our target languages where the content was not available as machine-readable text, and used optical character recognition (OCR) to extract text from these images.

3.1.1. Machine-Readable Text: Bulk Data

Table 1 shows the number of patents we acquired in bulk as machine-readable text. Some of the same English-language patents appeared in both the USPTO and the EPO databases. The data also included multiple versions of the same patents, including applications, application updates, grants, and grant updates. We extracted the text content of the patents from the bulk files, along with relevant metadata, into a normalised text format for further processing.

3.1.2. Machine-Readable Text: API Downloads

Using the EPO OPS API, we downloaded patents that had been registered with the patent offices in Spain, France, Croatia, Norway, and Poland in the years 1800-2020. The same normalised text format was used as for the bulk data. Often, only part of the text was available through the API in a machine-readable format; for example, only the title and abstract of the patent. We indicate in Table 2 how many patents were available in full and in part as machine-readable text.

The API limits the rate at which data can be accessed, even for users with a paid subscription, in order to preserve responsiveness for all users. Therefore, we developed a download tool[9] that would respect the various rate-related responses from the API, inserting pauses as needed, and allowing us to download a large quantity of data without ongoing manual intervention. We made a decision to exclude *kind A* patents from the downloads. These are patent applications, rather than granted patents. When a patent application is granted, a new patent document (a patent grant) is published, with a different kind code, and it is these granted patents that we expected to find registered in multiple languages.

Source	Language	Patents
USPTO	en	7,758,382
EPO	en	1,465,888
EPO	de	4,608,223
EPO	fr	595,741

Table 1: Machine-readable patents acquired in bulk, by source and language.

		Patents with text	
Country	Language	Any	All
Spain	es	909,551	563,334
France	fr	20,543	0
Croatia	hr	16,113	11
Norway	no	193,793	14,608
Poland	pl	131,747	0

Table 2: Patents accessed through the API with any/ all parts as machine-readable text, by country/ language.

The correspondence between the country of filing and the language used in the patent was not straightforward. For example, patents filed in Croatia were written in Croatian, Bosnian, and sometimes English. The language of the downloaded patents was not always indicated correctly in the metadata. The XML data returned by the EPO OPS API always contained a language attribute in the metadata, but in many cases we found that it had the value `ol`, which is not a valid language code. We treated those cases as if the language was that of the major language of the country where the patent was filed, relying on language filtering later in the pipeline to remove unwanted content.

[6] http://pubserv.uprp.gov.pl/PublicationServer/index.php?jezyk=en

[7] https://www.uspto.gov/

[8] https://www.epo.org/searching-for-patents/data/web-services/ops.html

[9] https://github.com/paracrawl/europat-scripts

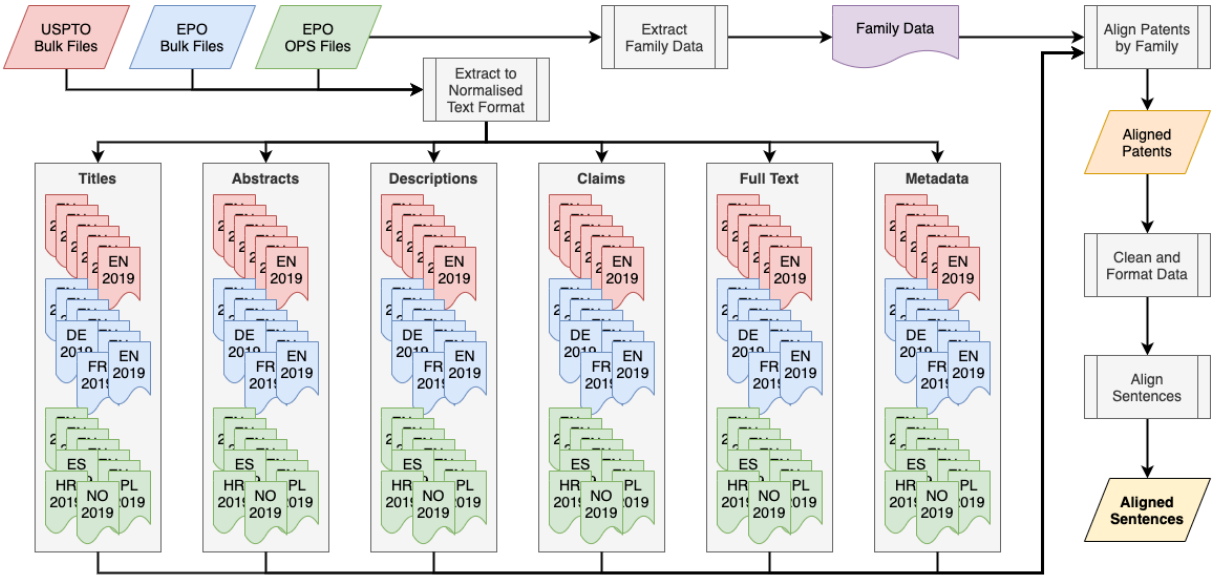


Figure 1: Schematic of the processing pipeline from raw text to aligned sentences.

3.1.3. Scanned Image Files of Original Patents

Many of the patents the EPO offered in languages other than English were not available to download as machine-readable text but only as scans of the original documents. The download tool we developed was used to access images of original patent documents that had been registered with the patent offices in Spain, France, Croatia, Norway, and Poland. Just as the number of published patents with machine-readable text differed between the target languages (Tables 1 and 2), so the number of patents available as images from relevant patent offices also differed (Table 3). To avoid duplication of effort, if all the text of a given patent had already been extracted as machine-readable text, the tool did not download images from that patent. In addition, during early testing, we found that longer patents typically contained a lot of chemical formulas or gene sequences, which are not useful as parallel data. The tool was updated so that it only downloaded patent images with no more than 25 pages (87% of those available).

		Scanned patent images	
Country	Language	$\leq$ 25 pages	$>$ 25 pages
Spain	es	35,997	10,080
France	fr	55,312	259
Croatia	hr	16,143	320
Norway	no	82,316	11,446
Poland	pl	56,692	15,050

Table 3: Scanned patent images by country/ language. Patents with $\leq$ 25 pages were processed with OCR.

3.2. OCR and Text Extraction Pipeline

The downloaded image documents were PDF files, each containing a graphical image of a single page. We used the open-source Tesseract OCR library (Kay, 2007) to extract the text from the images. Whereas the metadata for machine-readable text patents sometimes indicated the source language, downloaded patent images were supplied with no language information at all. Therefore it was necessary to provide multiple languages to Tesseract (for example, Croatian, Bosnian, and English), allowing it to determine the best match for each document.

Substantial further processing was needed after the initial text extraction, for example to remove page numbers and to rejoin sentences broken by soft returns and page breaks. Therefore, the scanned patent image files were passed through a pipeline (Figure 2) involving several steps and different tools, described below, to generate a clean patent document in text format.

The patent image data was often low quality, with issues such as marks on the page, unclear/blurred text, irrelevant reference text, extra formatting text, official stamps, row and column markers, text on an angle, and low resolution scans (Figure 3). In some cases, particularly with older documents, the data could not be processed because the image quality was too poor, such that processing would not result in meaningful data.

At the start of the pipeline, the images from the downloaded PDF files were extracted and passed to Tesseract to produce an HOCR file. Tesseract was configured to produce confidence scores at the character level so that they could be used later in the process to improve OCR quality and to fine-tune the output text. The raw HOCR output needed a considerable amount of additional custom processing to improve the quality.

After combining the HOCR files generated from each page into a single document for each patent, the next step in the process addressed issues with spelling: separating glued words, correcting the spelling of words

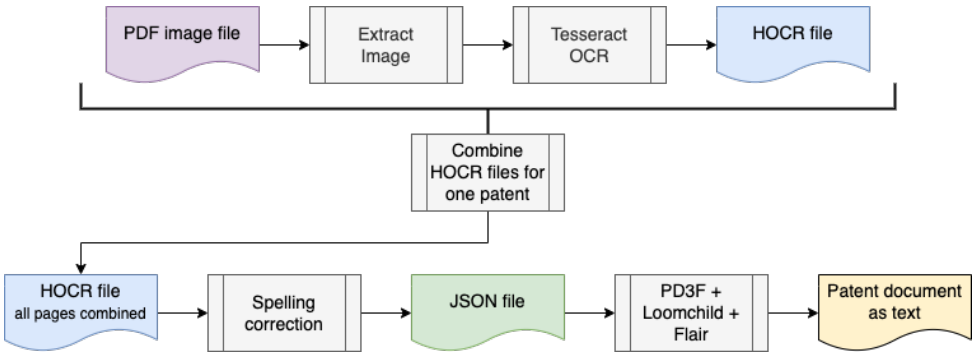
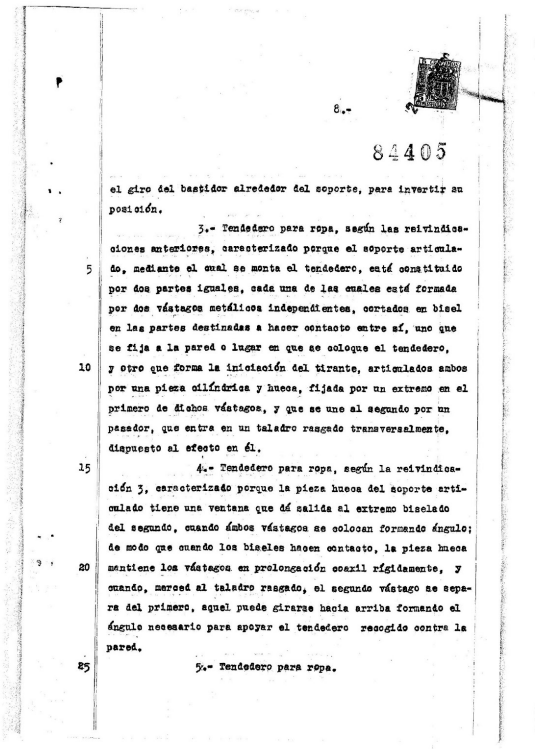


Figure 2: Schematic of the processing pipeline from multiple image files to a single combined text document.



8.-

84405

el giro del bastidor alrededor del soporte, para invertir su posición.

3.- Tendedero para ropa, según las reivindicaciones anteriores, caracterizado porque el soporte articulado, mediante el cual se monta el tendedero, está constituido por dos partes iguales, cada una de las cuales está formada por dos vástagos metálicos independientes, cortados en bisel en las partes destinadas a hacer contacto entre sí, uno que se fija a la pared o lugar en que se coloque el tendedero, y otro que forma la iniciación del tirante, articulados ambos por una pieza cilíndrica y hueca, fijada por un extremo en el primero de dichos vástagos, y que se une al segundo por un pasador, que entra en un taladro rasgado transversalmente, dispuesto al efecto en él.

4.- Tendedero para ropa, según la reivindicación 3, caracterizado porque la pieza hueca del soporte articulado tiene una ventana que dá salida al extremo biselado del segundo, cuando ambos vástagos se colocan formando ángulo; de modo que cuando los biseles hacen contacto, la pieza hueca mantiene los vástagos en prolongación coaxil rígidamente, y cuando, merced al taladro rasgado, el segundo vástago se separa del primero, aquel puede girarse hacia arriba formando el ángulo necesario para apoyar el tendedero recogido contra la pared.

5.- Tendedero para ropa.

Figure 3: Example OCR image showing noise, reference numbers, and text on angle.

that were incorrectly recognised by the OCR, and correcting missing or extra diacritic marks. In some cases the diacritics were missing in the original image, while in other cases they were lost during the OCR process.

We developed custom patent-specific dictionaries in each language for Hunspell[10] and enhanced the dictionaries further by using ngrams from Google Books[11] where data was available. We used the suggestion feature in Hunspell to identify glued words and split them into multiple words. If the spell-check still failed for a given word then different diacritics were used to generate alternatives for the lower-confidence characters. KenLM[12] models were trained and used to determine which of the spelling suggestions were most likely to be correct. This resolved most issues. We also evaluated Sublime Diacritic Fixer[13] but found that the results were no more accurate, and in some cases less accurate, than using the Hunspell approach.

The output after spell-checking and correction was a JSON file, which was then processed via a modified version of PD3F[14], making use of features such as sentence join and junk removal in order to identify text in columns and across a full page. Sentences in patents are generally more complex than typical sentences found in other texts. We used Loomchild[15] for sentence segmentation, with custom, patent-optimised SRX format rules[16] to segment sentences in the correct place and avoid over-splitting of sentences.

Many of the source documents had hard end-of-line boundaries, with sentences fragmented into two or more parts. We evaluated various approaches for rejoining these sentence fragments and selected Flair[17] as the best approach. We trained custom Flair sequence tagging models that were optimised for patents.

The final output of the pipeline was a single text file for each patent, containing all the available text, in a standardised format similar to that used for the machine-readable text.

3.2.1. Evaluation of the OCR Pipeline

The accuracy of the OCR process was evaluated using a sample of patents from the year 2010, where some of the text was available directly from the OPS API, from three jurisdictions: Spain, Croatia, and Norway. The sample only included patents where the scanned image had 25 pages or less.

The sample images were processed through the OCR

[10] http://hunspell.github.io

[11] https://storage.googleapis.com/books/ngrams/books/datasetsv3.html

[12] https://github.com/kpu/kenlm

[13] https://github.com/samnung/sublime_diacritic_fixer

[14] https://github.com/pd3f

[15] https://github.com/loomchild/segment

[16] https://en.wikipedia.org/wiki/Segmentation_Rules_eXchange

[17] https://github.com/flairNLP/flair

pipeline and the extracted text was aligned with the machine-readable text, using the industry-standard tool bleualign (Sennrich and Volk, 2010a). Performance was evaluated using the Word Error Rate (WER) metric, which counts the substitutions, insertions, and deletions needed to transform one text into another, normalised by length. The results are shown in Table 4.

		Word Error Rate		
Country	Patents	median	mean	SD
Spain	262	0.02	0.08	0.14
Croatia	411	0.18	0.19	0.12
Norway	545	0.04	0.11	0.16

Table 4: Sentence-level WER results from OCR.

The relatively large difference between median and mean values for Spanish and Norwegian patents indicated the presence of individual patents with high WER. Manual inspection of these files revealed many misrecognised characters related to chemical formulas and gene sequences. The other major source of errors related to inaccurate sentence boundaries, rather than word identification errors in the OCR pipeline.

3.3. Data Cleaning

Neural MT has proven to be particularly sensitive to noise, and much of the patent data we obtained was noisy. Specific cleaning steps were therefore applied to the standardised text files to generate high-quality parallel corpora.

The acquired patent data varied in format, language, quality, and completeness. The format and layout of the patent documents varied across time periods and between jurisdictions. The patents that were available as structured machine-readable text distinguished between the **title**, the **abstract**, the **description**, and the **claims** of the patent (Figure 1). This allowed us to compare abstracts against abstracts, and so on – although not all sections were present in every text-based patent. In contrast, the patents that were only available as scanned image files had no such structure and could only be compared at the level of complete documents.

Patent documents of every type were processed into a common plain-text format, but before the data was ready for sentence alignment, the text needed to be cleaned. We extended existing open-source tools that were designed to target typical errors from crawled websites to adapt them for cleaning parallel corpora. Software modules were improved to cope with patent-specific cleaning, leading to improved versions of Bifixer[18], a tool to fix and tag duplicates; and Bicleaner[19], a tool to filter noise from parallel text (Ramírez-Sánchez et al., 2020).

Bifixer was adapted to better remove noise coming from OCRed patents. In particular, common typos were added to replacement lists, specific configurations for sentence splitting were added to Bifixer's segmenter, and the limits for sentence length were reviewed to adapt to the nature of patent content. Hashing for duplicates was adapted to make the process tolerant to patent numbering entities in the near-duplicates hashing mode. Finally, Bifixer was adapted to handle the input format from the EuroPat pipeline. Making use of Bifixer, we recovered damaged data from patents, added better sentence splitting, and were able to better identify duplicates.

Bicleaner was also adapted in order to clean patent-specific noise and to handle the specific format used in the EuroPat pipeline. The set of rules for removing obvious noise was fully reviewed and tested against patent corpora. Some rules were removed or relaxed to accommodate the type of noise present in the texts. Patent content was used to adapt language modelling filtering. Options to disable the default set of steps in Bicleaner were implemented leading to new configuration arguments for the tool.

Classifiers were trained using patent data, and thresholds used to distinguish between good and bad parallel sentences were fully reviewed and adapted. All supported pairs of languages were customised for this project. Two different classifiers were adapted and used as they became available in Bicleaner: one based on features (probabilistic bilingual dictionaries, monolingual frequencies and length ratios) and Extremely Randomized Trees (trained on parallel sentences and using synthetic noise made by misaligning sentences, by replacing words in sentences with other words and by omitting words); and another one based on a neural classifier using XLM-RoBERTa[20] and introducing a 1:10 positive/negative sentences sampling ratio. The neural classifier proved to have the best performance and was thus used for the final version of the EuroPat corpus. Thanks to Bicleaner, we removed all sentences that were identified as containing unacceptable noise. All modifications to Bifixer and Bicleaner have been integrated into the software available on Github.

3.4. Document Alignment

Patents lodged in different jurisdictions are related to one another in so-called *patent families*. The patent family is stored as metadata for each patent and can be leveraged to match identical or near proximity translations of documents across languages. Based on the family metadata, we can discover patent families that identify the same invention filed with different authorities and that are, therefore, likely to be translations of each other. Patents can thus be aligned using metadata references to filings in other jurisdictions. However, a patent filed in another jurisdiction is not simply a like-for-like translation, due to differences in what each country allows, or amendments made following

[18] https://github.com/bitextor/bifixer

[19] https://github.com/bitextor/bicleaner

[20] https://huggingface.co/docs/transformers/model_doc/xlmroberta

examiner responses. Filings have several stages, some of which are revisions of text, while others have irrelevant technical information. Although it might seem that we could simply take the final versions of two documents, translation accuracy may in fact be lower as claims are reworded or removed to fit different jurisdictions. For example, United States patents contain software algorithms that do not appear in European patents. Hence, older revisions may in fact be more accurate translations. We therefore compared multiple versions of patent text where available.

The family data for each patent was obtained from the EPO OPS API and imported into a database, along with the unique ID for the patent. Our matching tool used the family data to identify all possible patent ID pairs. Many patents do not have a match as they were never translated. A matched pair of patent documents in two languages can have differing content, since some sections of a patent might not have been translated. For this reason, the quantity of matched documents can differ considerably by section (Table 5).

Language Pair	Section	Pairs
de-en	Title	854,159
	Abstract	192,847
	Description	232,886
	Claim	389,230
fr-en	Title	503,804
	Abstract	62,988
	Description	138,978
	Claim	268,807

Table 5: Example: matched pairs by patent section.

3.5. Sentence Alignment

The goal of sentence alignment is to identify high-quality matching translated sentences within the paired documents. Sentences that match sufficiently well are de-duplicated, scored, and filtered on quality.

To find parallel sentences across documents in different languages, we first translated the non-English document into English and then attempted to find matching sentences in both documents. Starting with release 2 of the EuroPat corpus, we trained bespoke machine translation models to translate the patent text to English. For release 1 we used off-the-shelf models that required some additional preprocessing steps.

Baseline models were built for 5 language pairs:

- German→English,
- Spanish→English,
- French→English,
- Croatian→English, and
- Polish→English.

The model for Norwegian→English was reused from the ParaCrawl project. Each of the models was built using MarianNMT (Junczys-Dowmunt et al., 2018) transformer base models with embedding size 512, FFN size 2048, 6-layer encoder, 6-layer decoder, 8 attention heads, swish activations, sentence piece vocabulary size of $32,000$ and a maximum sentence length of 250.

From data release 2 onward, the text extracted from the documents was split into sentences using the MOSES sentence splitter[21] based on punctuation. This process was optimised for the patent domain and for each language. MOSES has a list of known exceptions, such as abbreviations, and we extended this list to include abbreviations that we noticed to commonly occur in patents. Line breaks that already existed in the extracted text were kept; for example, because Tesseract deemed there to be a paragraph end.

Alignment and matching of the most probable bilingual sentences was carried out using Bleualign-cpp (Sennrich and Volk, 2010b; Sennrich and Volk, 2011) to match sentences between documents. The sentences from the translated document were used by Bleualign to identify which original sentence, and which English sentence in the matched document, had the best match. This process also took into account the order of the sentences in the document, allowing good matches around a sentence to contribute to the acceptance of a sentence that might otherwise have been skipped. For this reason, the quality of a single sentence may be poor; but as long as the overall quality of the translation of the document is good, all translated sentences will be found. When a sentence match was found, the pair of original unprocessed sentences was added to the corpus.

For EuroPat releases 1 and 2 we only matched sentences within corresponding patent sections (for example, title against title) and used section-specific cleaning methods. Title sections were all converted to lowercase, and description sections were stripped of chemical formulas. The same preprocessing methods were applied to both documents.

The OCRed documents that were added in EuroPat release 3 did not contain section information. For these, matching was performed by first concatenating the cleaned patent sections from the English-language patent in the order they occurred in the original patent, and then using the full document to match against sentences in the OCRed document. This approach avoided introducing errors by attempting to split the OCRed documents into sections.

3.6. Domain Classification

Patents are classified using one or more International Patent Classification (IPC) identifiers[22]. By annotating

[21] https://github.com/luismsgomes/mosestokenizer

[22] https://www.wipo.int/classifications/ipc/en/

the published sentence pairs with labels based on the IPC codes that were defined for the source patents, end-users are able to filter and select the pairs from domains that are relevant to their own use-case.
The IPC classification list is extensive, with thousands of classifications and sub-classifications, which are too granular for everyday use. Patents can belong to many IPCs and thus many classifications. However, the IPC is a hierarchical system, and IPC codes can be clustered into meaningful higher level domains. A review of the IPC classifications was performed and a set of coarse domain labels was defined as grouped IPC domain categories derived from the very granular sets of IPC classifications (Table 6):

I. General

II. Computing, Science and Tech

III. Biotechnology and Chemical

IV. Engineering and Manufacturing

V. Daily life

To help end-users make use of the domain classification, we added IPC codes and grouping codes to each sentence pair. For each sentence pair that ended up in the corpus, we took care to track which documents they originated from. These document references were added to each of the sentences when the translation memory TMX format files were created. After that, tmxutil[23] was used to add the IPC codes and the coarse IPC groups based on the pattern or prefix of each of the IPC codes.

4. Corpus Data Releases

The EuroPat corpus data has been released under Creative Commons Zero in several formats and on different platforms to make it useful for the maximum number of stakeholders. Formats include

- sentence aligned data, with one tab-separated sentence pair per line, for training machine translation systems (RAW and TXT files); and
- the translation memory standard TMX format, to aid use in translation memory tools.

There were three releases of the EuroPat corpus:

- **Release 1** included German and French paired with English, using data extracted from the USPTO and EPO bulk files.
- **Release 2** included all six languages, additionally using data downloaded from the EPO OPS API.
- **Release 3** included all six languages, and additionally included data extracted from scanned patent image files.

[23] https://github.com/paracrawl/tmxutil

The patent data used to create each release was a superset of the data used in the previous release. However, the releases themselves are not strictly supersets since the pipeline processes used for cleaning, aligning, and filtering the data were improved between releases.
The corpus data was uploaded to the ELRC-SHARE repository[24] to maximize compatibility with eTranslation; and was also distributed through the existing free language resource platform OPUS[25], which supports multiple formats.

5. Corpus Quality Assessment

A two-fold strategy was followed to assess the quality of the final corpora: human and automatic. End users can also construct their own quality filters based on the corpus metadata.

5.1. Metadata to Support Quality Filters

Sentence pairs in the corpora were automatically annotated with rich metadata to allow users to create their own quality filters. When relevant, length ratios, number matching and other scores were added, following Section 4.2.2.1 of the European Language Resource Coordination (ELRC)[26] validation guidelines. Besides ELRC scores, metadata related to cleaning scores from Bifixer and Bicleaner was included, along with IPC codes for both fine-grained IPC categories and coarse-grained IPC groups. The document type and source or number of tokens were also added. Official release versions are subsets of the raw versions, filtered by a minimum threshold of 0.5 score by Bicleaner.

5.2. Human Evaluation

At a very early stage of the project, after doing an internal human evaluation of 100 sentences before the first EuroPat release, we realised that patent translation content from documents was very high quality and that most of the errors defined in ELRC guidelines were not at all relevant for this scenario. We also noted that some of the typical errors expected when dealing with patents (OCR issues, formulae, overly long sentences, differences in numbering) were not in the ELRC-SHARE list of errors. For this reason, we decided against carrying out a standard ELRC-SHARE style human evaluation. Instead, we prioritised manually looking at the output of each step of the pipeline to discover and address issues from patent-specific content – a task which requires technical expertise. Human quality assessment for releases 2 and 3 of the EuroPat corpus was additionally opened up to professional translators and MT developers using the Corset tool. We focused our human evaluation based on linguistic searches in that tool.

[24] https://elrc-share.eu
[25] https://opus.nlpl.eu
[26] https://www.lr-coordination.eu

Group ID	Description	Top Level Sub-Classifications
I	General (default)	10
II	Computing, Science and Tech (science, photography, optics, cryptography, communications)	16
III	Biotechnology and Chemical (food, biotech, nanotech, chemistry)	47
IV	Engineering and Manufacturing (engines, nuclear physics, agriculture, forestry, aviation)	101
V	Daily life (household, music, arts, clothing, jewelry, sports and decorating)	35

Table 6: Grouped International Patent Classification domain categories. There are thousands of child classifications under each of the top level sub-classifications shown.

5.3. Automatic Evaluation

Extrinsic evaluation was carried out using machine translation. At a very early stage, baseline generic neural MT systems were trained using publicly available data. For each version of the corpus, we compared the generic systems against MT systems enriched with EuroPat parallel corpus data. Systems were tested on patent-specific data sets that we compiled for each language pair. Internal evaluation used the automatic metrics BLEU (Papineni et al., 2001) and COMET (Rei et al., 2020). In total, 70 systems were built for 9 translation directions using the official corpora and different strategies for cleaning (neural or non-neural version of Bicleaner, different cleaning thresholds) and for creating the in-domain neural systems (concatenation, fine-tuning, concatenation + fine-tuning) applying state-of-the-art techniques and software such as MarianNMT.
We present the corpus sizes and BLEU scores for the MT systems trained during evaluation of the final EuroPat data release (release 3) in Tables 7 and 8. German, Spanish, and French used a test set of patent data provided by the WIPO. For the other languages, non-public test sets were used. We were not able to find a test set for Norwegian in the patent domain. Improvements thanks to the addition of EuroPat data were obtained for all language pairs except no-en and en-hr. For no-en, this could be due to the test set being out of domain. For en-hr, this could be due to the relatively small amount of Croatian patent data that was obtained.

	Baseline size	EuroPat r3 size
en-de	5.9	19.734
en-es	29	51.352
en-fr	14.5	11.098
en-hr	53	0.154
en-no	40	4.341
en-pl	34	0.332

Table 7: Corpus sizes for EuroPat release 3 and for the baseline MT systems used in evaluation, in millions of sentences.

		Baseline + EuroPat r3 data		
	Baseline alone	fine-tune	concat	concat + fine-tune
en-de	23.7	29.5	31.0	**31.1**
de-en	49.4	59.7	**60.7**	60.6
en-es	41.3	**43.9**	42.0	43.8
es-en	39.2	44.7	46.0	**46.1**
en-fr	49.7	50.8	**52.4**	52.3
fr-en	49.7	50.6	**52.1**	51.9
en-hr	**38**	**38**	-	-
hr-en	39.7	**41.4**	-	-
no-en	**33.7**	31.6	-	-
en-pl	26.4	28.6	**29.6**	29.2
pl-en	31.2	38.2	37.5	**39.2**

Table 8: Evaluation BLEU scores. Best scores for each language pair are shown in **bold**.

6. Conclusion

The EuroPat corpus is expected to be widely useful for training high-quality machine translation systems, and particularly for those targeting technical documents such as patents and contracts. With this in mind, the data, along with rich metadata allowing for specific filtering, was formatted to be compatible with existing industry standards (RAW, TXT, and TMX) and released through OPUS[27] and ELRC-SHARE[28], as well as from the project's own website[29].

7. Acknowledgements

We offer our grateful thanks to all those who worked with us on the EuroPat project, particularly William Waites, Minas Sifakis, Jaume Zaragoza-Bernabeu, and Marta Bañón. The EuroPat project was funded by the Connecting Europe Facility of the Innovation and Networks Executive Agency of the European Commission (Agreement number INEA/CEF/ICT/A2018/1761979). The contents of this publication are the sole responsibility of the implementing partners and do not necessarily reflect the opinion of the European Union.

[27] https://opus.nlpl.eu/EuroPat.php
[28] https://elrc-share.eu/repository/search/?q=europat
[29] https://europat.net/

8. Bibliographical References

Bañón, M., Chen, P., Haddow, B., Heafield, K., Hoang, H., Esplà-Gomis, M., Forcada, M. L., Kamran, A., Kirefu, F., Koehn, P., Ortiz Rojas, S., Pla Sempere, L., Ramírez-Sánchez, G., Sarrías, E., Strelec, M., Thompson, B., Waites, W., Wiggins, D., and Zaragoza, J. (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, July. Association for Computational Linguistics.

Carpuat, M., Vyas, Y., and Niu, X. (2017). Detecting cross-lingual semantic divergence for neural machine translation. In *ACL Workshop on Neural Machine Translation Workshop 2017*.

Junczys-Dowmunt, M., Grundkiewicz, R., Dwojak, T., Hoang, H., Heafield, K., Neckermann, T., Seide, F., Germann, U., Fikri Aji, A., Bogoychev, N., Martins, A. F. T., and Birch, A. (2018). Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, Melbourne, Australia.

Kay, A. (2007). Tesseract: An open-source optical character recognition engine. *Linux Journal*, 2007(159):2.

Koehn, P. and Knowles, R. (2017). Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver, Canada.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2001). BLEU: a method for automatic evaluation of machine translation. Technical Report RC22176(W0109-022), IBM Research Report, September 17.

Ramírez-Sánchez, G., Zaragoza-Bernabeu, J., Bañón, M., and Ortiz-Rojas, S. (2020). Bifixer and bicleaner: two open-source tools to clean your parallel data. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 291–298, Lisboa, Portugal, November. European Association for Machine Translation.

Rei, R., Stewart, C., Farinha, A. C., and Lavie, A. (2020). COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online, November. Association for Computational Linguistics.

Sennrich, R. and Volk, M. (2010a). MT-based sentence alignment for OCR-generated parallel texts. 11.

Sennrich, R. and Volk, M. (2010b). MT-based sentence alignment for OCR-generated parallel texts. In *Proceedings of the 9th Conference of the Association for Machine Translation in the Americas: Research Papers*, Denver, Colorado, USA, October 31-November 4. Association for Machine Translation in the Americas.

Sennrich, R. and Volk, M. (2011). Iterative, MT-based sentence alignment of parallel texts. In *NODALIDA 2011, Nordic Conference of Computational Linguistics*. Northern European Association for Language Technology (NEALT). The 18th Nordic Conference of Computational Linguistics ; Conference date: 11-05-2011 through 13-05-2011.