# ForMaT: Dataset for Visually-Grounded Multilingual PDF Translation

**Michał Ciesiółka**[1,2], **Dawid Wiśniewski**[1,3], **Adrian Charkiewicz**[1], **Kamil Guttmann**[1,2]
[1] Laniqo, Poznań, Poland
[2] Faculty of Mathematics and Computer Science, Adam Mickiewicz University, Poznań, Poland
[3] Poznań University of Technology
{name}.{surname}@laniqo.com

## Abstract

We present ForMaT (**For**mat-Preserving **M**ultilingu**a**l **T**ranslation), a parallel corpus of 3,956 PDFs across 15 language pairs that preserves original layout metadata proposed for multimodal machine translation. To ensure structural diversity in the dataset, we employ K-Medoids sampling over 45 geometric features, capturing complex elements like nested tables and formulas to focus only on visually diverse PDF documents. Our evaluation reveals that current MT systems struggle with spatial grounding and geometric synchronization, often losing the link between text and its visual context. ForMaT provides a benchmark for developing layout-aware translation models that integrate visual and textual context for high-fidelity document reconstruction.

## 1 Introduction

Modern machine translation (MT) systems increasingly leverage multimodal signals to enhance translation accuracy. In the audio domain, for instance, paralinguistic features, such as speaker identification, gender, and emotional tone, provide essential context that helps disambiguate intent and refine target-language nuances. Similarly, when processing visually rich documents like PDF files, visual and spatial cues are often indispensable for resolving lexical ambiguity and selecting the appropriate morphological forms. This shift toward context-dependent, multimodal translation has emerged as a significant frontier in MT research (Shen et al., 2024; Feng et al., 2025).

In this paper, we focus on visual context through the introduction of a new parallel corpus comprising 3,956 PDF documents, which we named ForMaT (**For**mat-Preserving **M**ultilingu**a**l **T**ranslation). The dataset spans 15 language pairs involving English, German, Spanish, French, Italian, and Polish. Unlike traditional parallel corpora that are limited to plain text, our dataset preserves the rich layout and formatting information of the original source documents.

Our contributions are threefold: first, we motivate the necessity of layout-aware resources for modern MT; second, we detail a methodology for dataset collection and provide an in-depth analysis of the corpus's structural properties; and third, we establish the dataset's diversity through a multidimensional analysis, positioning it as a benchmark for evaluating the next generation of layout-aware and document-level translation systems. To demonstrate the practical utility of this benchmark, we conclude by evaluating several state-of-the-art PDF translation systems, specifically analyzing their ability to preserve both linguistic meaning and complex document layouts.

### 1.1 Motivation

In machine translation, visual cues within a source document are often essential for generating accurate target output. When processing PDF files, several use cases demonstrate why visual and spatial context is critical:

- Image captions – Visual context helps disambiguate polysemic words and personal pronouns. For instance, an image depicting a woman can signal the use of the feminine

pronoun *she* when translating from gender-neutral languages (e.g., Turkish, Hungarian, or Basque). Similarly, short captions often lack sufficient textual context; an image can clarify whether the word *head* in a medical document refers to an anatomical body part, a team leader, or the top element of a device.

- Text position – Spatial positioning helps identify named entities and document semantics. In an invoice, for example, a company name is typically expected at the top, while numerical values in specific regions are more likely to represent monetary amounts rather than dates or quantities.

- Tables – The translation of table cells often depends on context provided by adjacent cells or headers. Depending on the layout, these headers may appear in the top rows (column-wise representation) or the flanking columns (row-wise representation). Such structures pose unique challenges (Yin et al., 2020), as the visual layout of a PDF helps a model identify that a text fragment is part of a table, allowing it to focus on the correct relational context to understand cell content.

- Text Segmentation – The spacing between words and paragraphs is a vital indicator of context. Traditional OCR tools paired with MT models often fail when text is justified with non-standard spacing, written in creative or non-linear formats (e.g., one word per line), or rotated vertically. Identifying text clusters through visual analysis allows the model to segment the document into meaningful units, preserving the intended translation context.

- Geometric Constraints – Document layout serves as a guide for translation length and copy-fitting. Because target languages may vary significantly in word length, the layout dictates how the translation must be aligned. By analyzing the visual properties of the PDF, models can select translations that better fit the available space, minimizing unnatural gaps or managing page breaks more effectively.

To address these challenges, MT datasets must evolve to include images and rich formatting (such as original PDFs). This integration enables a more robust evaluation of how translation models perform when document structure is inextricably linked to meaning.

## 2 Related works

The evolution of Machine Translation (MT) has been marked by a steady expansion of the context window, moving from isolated sentences to entire documents and, more recently, to multimodal inputs. `ForMaT` sits at the intersection of Document-level MT, Multimodal Learning, and Visually-Rich Document Understanding (VRDU).

### 2.1 Multimodal and Document-Level MT

Early efforts in Multimodal Machine Translation (MMT) focused primarily on visual grounding for image captioning, exemplified by the Multi30K (Elliott et al., 2016) dataset. These tasks used images to resolve lexical ambiguities (e.g., gender or entity type) but were limited to short, isolated sentences. Recent research has shifted toward a more context dependent approaches utilizing various modalities, where researchers leverage non-textual signals to improve translation quality in specific domains (Shen et al., 2024; Feng et al., 2025).

While Document-Level MT (DMT) addressed long-range textual dependencies, most existing benchmarks, including the massive DocHPLT (O'Brien et al., 2025), rely on "flattened" web-crawled text. These corpora lack the 2D spatial information (headers, footers, sidebars) that is vital for interpreting the logical flow of high-stakes documents like technical manuals or legal acts.

### 2.2 Visually-Rich Document Understanding (VRDU)

The field of VRDU has established that spatial coordinates and typographic cues are as important as the text itself for understanding complex layouts. The LayoutLM series (v1, v2, v3) (Huang et al., 2022) pioneered the use of 2D positional embeddings to model the relationship between text and visual structure. More recently, Layout-Aware LLMs have demonstrated that encoding document geometry as specialized tokens can significantly improve performance in information extraction and Document Visual Question Answering (VQA) (Lu et al., 2025).

Other popular models aimed at VRDU task

are: TaBERT (Yin et al., 2020), focused on the joint understanding of textual and tabular data, or DocLayout-YOLO (Zhao et al., 2024), PP-DocLayout (Sun et al., 2025a), and the PaddleOCR 3.0 (Cui et al., 2025) all focused on layout understanding.

However, a gap remains: while VRDU models excel at extraction, they are rarely evaluated on generative translation. ForMaT provides the necessary parallel data to bridge this gap, treating translation not just as a linguistic task, but as a layout-preservation task.

### 2.3 Multimodal LLMs

Modern approaches leverage Multimodal Large Language Models (MLLMs) – language models with the ability to understand modalities other than texts (Yin et al., 2024). Recent models that focus on images, introduce various optimizations helping understand different modalities, e.g., utilizing visual instruction tuning (Liu et al., 2023) and global-local dual perception (Lu et al., 2026) for high-resolution images. Recent systems like InImageTrans (Zuo et al., 2025), TranslateGemma (Finkelstein et al., 2026), or Gemini (Anil and GeminiTeam, 2025) demonstrate the potential of LLMs to handle "visually-situated" text and translate between languages. Furthermore, research into zero-shot MMT (Futeral et al., 2025) and unimodal alignment (Zhang et al., 2025a) aims to reduce the reliance on costly supervised parallel data.

### 2.4 Benchmarks for Document Image Translation (DIMT)

The most recent frontier in the field is Document Image Machine Translation (DIMT), highlighted by the ICDAR 2025 DIMT Challenge (Zhang et al., 2025b). Current state-of-the-art benchmarks such as DIMT-WebDoc-300K and DIMT-arXiv-124K focus heavily on translating document images into Chinese, often prioritizing scale over structural variety.

Similarly, while M3T (Hsu et al., 2024) has introduced document-level multimodal machine translation benchmarks, there remains a need for a corpus sampled specifically for structural difficulty. ForMaT addresses this gap by using a rigorous K-Medoids sampling (Kaufman and Rousseeuw, 1990) methodology across 45 structural features to collect only diverse PDF documents, ensuring that the corpus serves as a stress test for the next generation of layout-aware translation models.

### 2.5 Evaluating Multimodal Models

While traditional metrics remain standard, recent findings by (Sun et al., 2025b) indicate that n-gram overlap scores like BLEU often fail to capture document-level coherence, advocating for more nuanced, multi-dimensional evaluation. This shift is reflected in the ICDAR 2025 DIMT Challenge (Zhang et al., 2025b), which underscores the persistent difficulty of translating complex layouts. Methods, such as multimodal reasoning and dual-perception architectures try to address this challenge (Huang et al., 2026; Lu et al., 2026).

In comparison to existing literature, ForMaT offers three distinct advantages:

1. **Language Diversity**: We provide 15 language pairs, with a focus on European languages often underrepresented in recent DIMT challenges.

2. **Structural Complexity**: Unlike datasets that rely on random crawling, we employ K-Medoids sampling over 45 structural features to ensure our corpus includes challenging cases like complex tables, inline formulas, multiple columns, and images with captions.

3. **High-Fidelity Metadata**: We provide raw layout metadata alongside the parallel text, enabling the development of models that can reconstruct the target PDF with pixel-perfect accuracy.

## 3 Dataset

The process of collecting the dataset consists of several phases as depicted in Figure 1. First, we identify websites providing PDF documents that meet our selection criteria, then, we sample a first, broad collection of documents using quota sampling. Finally, we filter the broad collection of documents, to leave only the most diverse and interesting documents in terms of visual complexity and composition. The final ForMaT dataset is represented by 3,956 documents.

### 3.1 Data sources

To create our dataset, we targeted two domains exhibiting distinct linguistic profiles and visual formats: legal documents and technical user manuals.

For the legal domain, we curated a corpus from international and national institutions that publish official documentation in a multilingual format. A significant portion of this data was sourced from European Union repositories, specifically focusing on three distinct legal contexts: legislative acts via EUR-Lex, parliamentary proceedings through the European Parliament portal, and judicial documentation from the European e-Justice Portal. To ensure broader geographic and administrative variety, we further incorporated the corpus with documents from the United Nations digital library, the Swiss federal law repository (Fedlex), and the U.S. Social Security Administration (SSA).

To curate the user manual domain, we utilized a global index of electronics brands[1] as a foundational blueprint for our search. To facilitate downstream document processing and alignment, we exclusively selected manufacturers that publish localized instructions as individual, single-language files rather than consolidated multilingual volumes. Although our initial search targeted electronic product domain, we expanded the scope to include major automotive manufacturers to increase the variety of instructional layouts. The final selection included documentation from Huawei, Lenovo, Philips, Nissan and Toyota, providing a diverse set of technical terminologies and visual schematics. The full set of sources for both domains is summarized in Table 1.

The choice of the domains and data sources depended on the licenses assigned to the documents; we collected only those documents that can be used for research purposes.

**Table 1:** Dataset Sources and URLs

| Source | URL |
|---|---|
| United Nations | un.org |
| SSA | ssa.gov |
| E-Justice | e-justice.europa.eu |
| Fedlex | fedlex.ch |
| European Parliament | europarl.europa.eu |
| EUR-Lex | eur-lex.europa.eu |
| Toyota | toyota-europe.com |
| Philips | philips.com |
| Nissan | nissan-techinfo.com |
| Lenovo | lenovo.com |
| Huawei | huawei.com |

For each domain, we have collected only parallel data available in multiple languages, where source and target documents are expressed in: English, French, German, Italian, Polish, and Spanish. This set of languages was found representative, as the attempts to add new languages to the set resulted in unrepresented language pairs at the sampling stage.

[1] https://en.wikipedia.org/wiki/List_of_electronics_brands

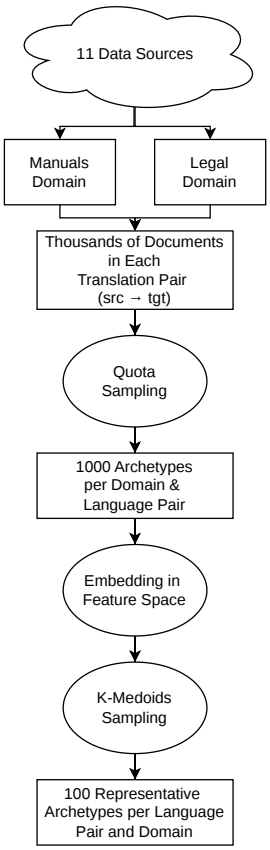



**Figure 1:** ForMaT dataset collection process. Each operation was performed independently for each language pair in both domains.

### 3.2 Data sampling

To balance data across the two primary domains and fifteen language pairs, we targeted a sample of 1,000 documents per pair in each domain.

We adopted a quota sampling strategy (Cochran, 1977) with two modifications to address imbalances in underrepresented sources. First, we grouped documents by source and language pair, then sampled them in ascending order of available documents. This ensured that underrepresented groups were fully included before larger sources were tapped. Second, when a source failed to meet its quota, the remaining capacity was dynamically redistributed to larger sources.

To capture potential document changes in time, we sampled the EUR-Lex corpus evenly by year. Given its unique volume, we selected two search-

result pages per year from 2005 to 2025, specifically choosing sets where every document was available in all six target languages. This approach allows us to observe the evolution of official language and structure over two decades. For each document, we collected and analyzed only the first 10 pages.

### 3.3 Data retrieval

To ensure the selection of stylistically diverse PDFs, we developed a hybrid extraction pipeline that integrates computer vision for layout detection with low-level PDF parsing for precise text and metadata extraction. This approach allowed us to filter our initial pool of 1,000 documents per domain/language pair, retaining only those with the most diverse structural and formatting features.

#### 3.3.1 Visual Layout Analysis

We employed the `PaddleOCR` (Cui et al., 2025) library, specifically the `PP-DocLayoutV2` (Sun et al., 2025a) model, to analyze the visual structure of each document. Unlike standard OCR models that prioritize text character recognition, this model treats the PDF page as a visual image to identify high-level semantic regions. It segments the page into discrete categories, including headers, footers, figures, tables, and standard text blocks.

#### 3.3.2 Textual Extraction and Metadata Parsing

Complementing the visual analysis, we utilized the `pdfminer`[2] library to extract styling metadata directly from the PDF file. We chose direct parsing over OCR-based recognition to ensure the high-fidelity preservation of formatting attributes.

Our extraction module retrieves word-level styling information, including font family, size, color, and weight. These attributes are critical features for document translation pipelines: they ensure that formatting properties from the source text (e.g., a specific token marked in bold or red) are accurately mapped to the corresponding fragment in the target translation.

#### 3.3.3 Vectorization

To quantify the structural complexity of the sampled documents, each file was mapped to a 45-dimensional feature vector $\vec{v}$ (where $\vec{v} \in \mathbb{R}^{45}$), as detailed in Table 3. These features integrate entity types identified via `PaddleOCR` with formatting metadata from `pdfminer`, and are organized into three primary categories:

- **Text-Based Labels:** These features represent the average frequency of textual elements per page. This category includes structural elements such as `text`, `footer`, `paragraph_title`, and `abstract`.
- **Visual Labels:** These features capture non-textual or complex graphical entities within the layout, including technical structures (e.g., `table`, `algorithm`), graphical assets (e.g., `image`, `seal`), and specialized spatial formatting such as `vertical_text`.
- **Typographic and Stylistic Attributes:** This category captures the visual properties of the text using a combination of frequency and occurrence metrics. We record the total number of unique font weights (e.g., `bold`, `italic`) and distinct font names present in the document. To normalize variations, font sizes are rounded to the nearest 0.5 points before counting. The color profile is represented as a binary sub-vector, where specific indices (e.g., `blue`, `black`) are assigned a value of 1 if the color is present in the document text and 0 otherwise.

#### 3.3.4 Clustering

To capture maximum structural and stylistic diversity, we employed the $K$-Medoids clustering algorithm to select representative documents from our vectorized pool. By clustering the documents into $K$ groups and selecting the medoid (the most centrally located document) of each cluster, we ensured that our final selection of $K$ documents represented the full breadth of the data's stylistic variance.

We performed clustering independently for each language pair within the two domains. To represent a parallel document pair as a single entry, we computed a combined feature representation by averaging the 45-dimensional feature vectors extracted from the source and target documents. These representations were then clustered into $K$=100 distinct groups per language pair in each domain (15 pairs $\times$ 2 domains = 30 processes) using the Euclidean distance metric. To ensure a globally representative selection, we utilized the

[2] https://github.com/pdfminer/pdfminer.six

k-medoids++ initialization strategy, which spreads the initial medoids far apart in the feature space.

This approach reduces the total number of documents while preserving a broad spectrum of complexities, ranging from simple text-only reports to intricate technical schematics. The choice of $K$=100 was a heuristic intended to capture a wide range of layouts without over-fragmentation.

In this paper, we refer to the underlying content shared across translations as a document archetype, distinguishing it from a specific PDF file in a specific language, which we call a document instance. Because many documents in our source pool exist in multiple languages, the independent sampling processes occasionally selected the same document "archetype" as a representative for different language pairs.

As a result of this overlap across the 30 sampling processes, we identified 1,278 unique document archetypes. Since these archetypes do not all exist in every one of the 15 language pairs, we collected all available translations for these specific selections, resulting in a total of 3,956 unique document instances (URLs). This methodology ensures that each language pair is represented by a diverse set of layouts.

#### 3.3.5 Representativeness and Diversity Gain

Our sampling strategy produced a corpus with significantly greater average structural complexity per document than the initial pool. By selecting cluster medoids rather than random samples, we successfully amplified the presence of underrepresented structural features.

The final corpus is considerably more visually dense than the original dataset. Specifically, the relative frequency of images more than doubled (+101.2%), while the occurrence of tables increased by 60.9%. Supporting graphical elements—such as `vision_footnotes` and `figure_titles` also saw substantial growth, rising by 48% and 33%, respectively.

The selection process markedly increased color diversity by capturing stylistic variations typically overlooked by random sampling. While the initial pool was predominantly monochromatic, the final subset exhibited substantial growth in minority colors: the frequencies of Yellow and Red rose by 264% and 215%, respectively, while Purple, Teal, and Pink each increased by over 115%.

The complete impact of the clustering process on entity and color distributions is detailed in Appendix B.

### 3.4 Dataset availability

The dataset is available online: Dataset (Hugging Face) [3].

## 4 Explorative Data Analysis

The concept of multidimensional diversity serves as the foundational framework for evaluating the architectural complexity of this corpus. Rather than treating document difficulty as a singular, linear metric, we model it as a coordinate within a multi-axis space defined by structural, and stylistic features.

Modern translation systems face challenges on both text-level style and layout preservation. This multidimensional approach is essential because document difficulty is rarely uniform; a page might be linguistically simple yet stylistically complex, or visually dense while maintaining a rigid, predictable layout.

### 4.1 Feature Independence

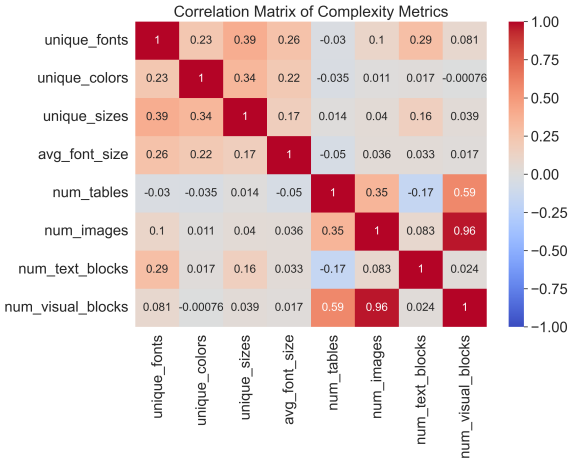



**Figure 2:** Spearman correlation matrix of document complexity metrics.

Figure 2 presents a Spearman correlation matrix of selected document attributes. The results indicate low correlation between different dimensions of document variety. Notably, stylistic attributes (such as the number of unique font colors and sizes) show minimal correlation ($r < 0.2$) with structural indicators like the number of graphical entities. This finding suggests that "visual complexity" (e.g., a page full of images and tables)

[3] https://huggingface.co/datasets/laniqo/ForMaT

and "formatting complexity" (e.g., documents with varied font colors and sizes) represent independent challenges.

However, the results also highlight a moderate coupling between typographic variety and textual density. We observe that both the number of unique fonts and the number of unique font sizes show a positive correlation with the number of text blocks ($r = 0.29$ and $r = 0.16$, respectively). This suggests that as a document is divided into more text blocks, the variety of formatting styles increases. This implies that the task of text-level style preservation becomes increasingly difficult in high-density documents, where the system must track a larger volume of independent stylistic metadata alongside the translated content.

## 4.2 Structural Variance

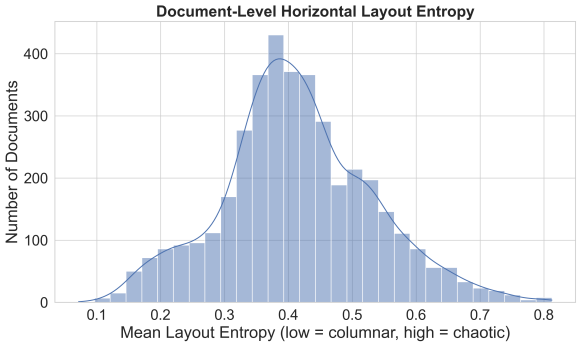



**Figure 3:** Distribution of horizontal layout entropy across documents. Low entropy indicates columnar layouts with predictable vertical alignment of text blocks, while high entropy reflects chaotic layouts with irregular spatial distribution and disrupted reading order.

Beyond simple entity counts, we measured the spatial organization of content using horizontal layout entropy ($H$) seen in Figure 3. This metric, calculated using Shannon entropy over soft-binned horizontal bounding box coordinates, measures the "predictability" of the document flow. Documents characterized by low entropy values ($H < 0.3$) typically represent the rigid, highly predictable columnar formats found in European Union legislative acts, where text blocks follow a strict, repetitive alignment. On the contrary, high entropy values ($H > 0.6$) indicate "chaotic" or non-linear layouts, which are prevalent in modern electronics manuals where the reading order is frequently interrupted by diagrams and multi-directional labels. By including high-entropy samples, we specifically challenge the translation system's ability to handle fragmented text flows without losing the logical connection between spatially distant but contextually related entities.

## 4.3 Granularity and Fragmentation

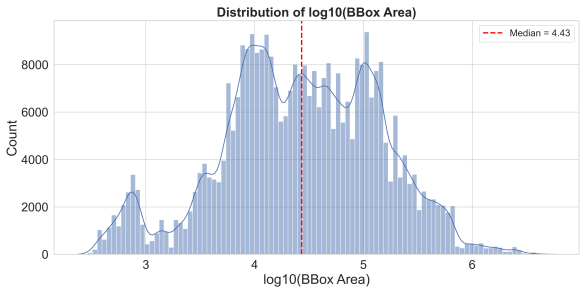



**Figure 4:** Bounding box area distribution on a logarithmic scale. The concentration of "micro-entities" (indicated by the peak at $\log_{10}$ Area $\approx 4.0$) highlights the high degree of layout fragmentation.

We analyzed the physical scale of the document components. By examining the distribution of bounding box (BBox) areas on a logarithmic scale, we identified a high degree of layout fragmentation. As shown in the BBox area analysis in Figure 4, a significant portion of the dataset consists of "micro-entities". This fragmentation serves as a stress test for layout-aware translation systems, which must maintain the logical ordering and spatial coherence of these tiny, inter-dependent elements during the translation and re-rendering process.

## 4.4 Spatial Density and Content Coverage

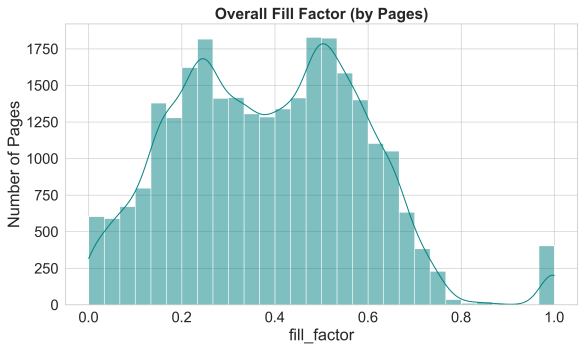



**Figure 5:** Distribution of the Overall Fill Factor across the corpus.

Finally, we quantified the physical organization of the corpus using fill factor analysis, which measures the ratio of bounding box areas to the total page area. This metric provides a macroscopic view of document saturation, allowing us to categorize the corpus into distinct layout types.

As illustrated in the multi-modal distribution of Figure 5, the dataset captures two distinct structural profiles: low-density layouts with significant margins or white space (peaking at 25% coverage) and high-density pages where content saturates the layout (peaking at 55% coverage).

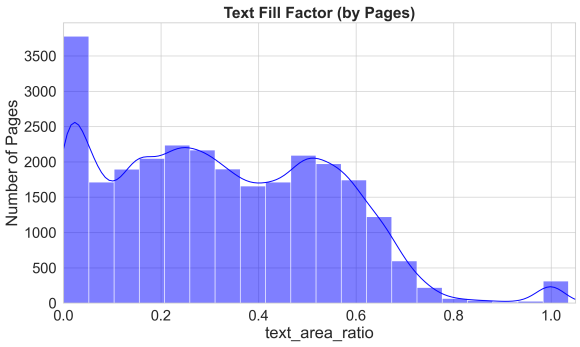



**Figure 6:** Text Area Ratio per page.

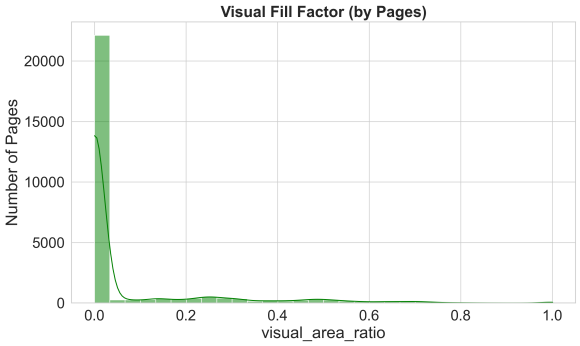



**Figure 7:** Visual Area Ratio per page.

While text remains the primary content driver (Figure 6), the Visual Fill Factor exhibits a significant long-tail distribution (Figure 7), with a dedicated subset of documents maintaining high structural occupancy (up to 60% area ratio) due to the presence of large diagrams, schematics, and complex tables.

## 5 Translation Systems Comparison

To demonstrate the practical usefulness of our dataset, we performed a comparative evaluation using a manual selection of PDFs that pose distinct structural and linguistic challenges based on the documents' label counts and variety. We benchmarked two industry-standard commercial engines-Google Translate[4] and DeepL[5]-as well as our internal translation system.

[4] https://translate.google.com/
[5] https://www.deepl.com/

### 5.1 Results

We group the observed problems into two categories: linguistic errors, which stem from a system's lack of document-level context, and structural errors, which arise during the construction of the output PDF file.

#### 5.1.1 Translation Errors

| Left | Center | Right |
|---|---|---|

**(a)** Source text

| Lewy | Środko-wy | Prawy |
|---|---|---|

**(b)** Original target text

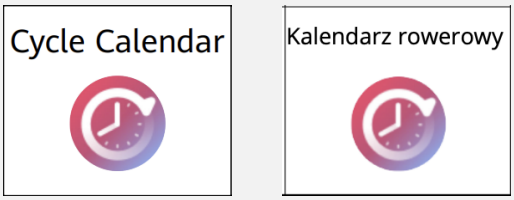
| Opuścił | Centrum | W porządku. |
|---|---|---|

**(c)** Translation system result

**Figure 8:** Table cells translation error presenting different semantic meaning to each cell.

Figure 8 illustrates a translation of three isolated table cells. In the gold-standard target text, the context refers strictly to spatial orientation: "Left," "Center," and "Right." However, one evaluated system assigned a different semantic meaning to each term. The word "Left" was mistranslated as "Opuścił" (the past tense of "to leave"), and "Right" was rendered as "W porządku" (signifying "all right" or "okay"). Furthermore, for "Center", the system opted for the noun "Centrum" instead of the required spatial adjective "Środkowy". This indicates a failure in spatial grounding, where the system lacks the table-level context.

Figure 9 illustrates a significant contextual dissonance in an image-caption translation task. The original document contains an icon labeled "Cycle Calendar" within a health-related context. However, the system maps "Cycle" to the biking domain, rendering the caption as "Kalendarz rowerowy" (Bicycle Calendar). This failure in multimodal grounding shows that the system prioritized common statistical associations over the actual visual context.

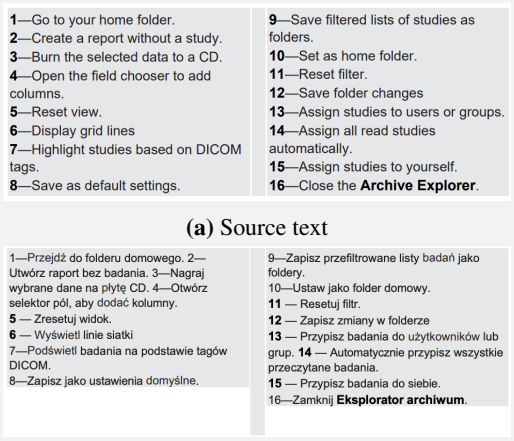



**(a)** Source text

**(b)** Translation system result

**Figure 9:** Image caption translation error introduced by missing image context.

### 5.1.2 Structural Errors



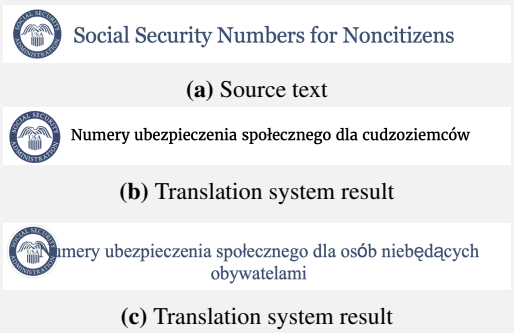

**(a)** Source text

**(b)** Translation system result

**Figure 10:** Dual-column numbered list reconstruction error. The system misplaced newline characters and resized gray background, breaking the parallel alignment.

Figure 10 illustrates a structural reconstruction failure in a dual-column numbered list. The evaluated system failed to preserve the original line breaks, resulting in segmentation errors where list indices were merged into preceding text blocks. Furthermore, the system recalculated the gray background box dimensions for each column independently. This lack of geometric synchronization produced uneven blocks, breaking the original typographic hierarchy and column alignment.

Figure 11 displays a header featuring the "Social Security Administration" logo with adjacent blue text. The evaluated systems exhibited distinct reconstruction errors. The first system failed to preserve the text color metadata, rendering the text in a default black font. The second system produced a collision, overlaying the translated text




**(a)** Source text

**(b)** Translation system result

**(c)** Translation system result

**Figure 11:** Document header reconstruction errors. One system failed to restore blue text coloring, while the second one overlayed the text over a logo and rendered some letters in different typeface.

partially onto the logo. Additionally, this system suffered from font-fallback artifacts, where the Latin-1 characters (English) and the extended Latin characters (Polish diacritics) were rendered in mismatched typefaces.

**Table 2:** System performance across structural and semantic challenges. We compared three systems: Ours: our internal translator; DeepL and GoogleT: Google Translate).

| Category | Structural Feature | Ours | DeepL | GoogleT |
|---|---|---|---|---|
| **Trans.** | **Semantic Inversion** (Table Cells) | *Fail* | *Pass* | *Partial* |
| **Trans.** | **Multimodal Grounding** (Image Captions) | *Fail* | *Pass* | *Fail* |
| **Layout** | **Metadata Preservation** (Color/Position) | *Partial* | *Partial* | *Pass* |
| **Layout** | **Lists Reconstruction** (Dual-Columns) | *Pass* | *Fail* | *Partial* |

The observed linguistic and structural limitations are synthesized in Table 2, which provides an overview of how each system managed the core challenges identified in our corpus.

## 6 Conclusion

In this work, we introduced a novel dataset comprising 3,956 PDF documents across 15 language pairs, sourced from the legal and technical domains. A defining characteristic of this corpus is its high layout and formatting complexity, which is essential for evaluating end-to-end document

translation pipelines. By utilizing a hybrid extraction pipeline and K-Medoids clustering over a set of 45 features, we ensured the dataset captures a diverse set of document layouts.

In contrast to conventional plain-text datasets commonly used in machine translation, this benchmark preserves the full layout and typographic context of the original documents, thereby providing a more accurate representation of real-world translation scenarios. This addresses a critical gap in machine translation research, where visual context is frequently discarded or limited to a single reference image.

Our qualitative analysis of commercial translation systems highlights significant limitations in current architectures when processing visually-rich documents. The evaluated systems frequently fail to maintain structural and formatting integrity during the translation process.

This dataset serves as a benchmark for evaluating layout-aware and document-level translation systems. We expect it to drive future research toward models that jointly optimize for linguistic accuracy, contextual translation, and formatting preservation. We leave the development of automatic metrics for evaluating formatting preservation to future work.

## 7 Limitations

The released dataset is limited to left-to-right languages written in the Latin alphabet as we focus on main European languages. Moreover, long documents are limited to 10 pages only as the goal of this benchmark is to focus on visual context, which frequently is quite local.

## References


[Anil and GeminiTeam2025] Anil, Rohan and GeminiTeam. 2025. Gemini: A family of highly capable multimodal models.

[Cochran1977] Cochran, William G. 1977. *Sampling Techniques*. John Wiley & Sons, New York, NY, 3rd edition.

[Cui et al.2025] Cui, Cheng, Ting Sun, Manhui Lin, Tingquan Gao, Yubo Zhang, Jiaxuan Liu, Xueqing Wang, Zelun Zhang, Changda Zhou, Hongen Liu, Yue Zhang, Wenyu Lv, Kui Huang, Yichao Zhang, Jing Zhang, Jun Zhang, Yi Liu, Dianhai Yu, and Yanjun Ma. 2025. PaddleOCR 3.0 technical report. July.

[Elliott et al.2016] Elliott, Desmond, Stella Frank, Khalil Sima'an, and Lucia Specia. 2016. Multi30k: Multilingual english-german image descriptions. In *Proceedings of the 5th Workshop on Vision and Language, hosted by the 54th Annual Meeting of the Association for Computational Linguistics, VL@ACL 2016, August 12, Berlin, Germany*. The Association for Computer Linguistics.

[Feng et al.2025] Feng, Yi, Chuanyi Li, Jiatong He, Zhenyu Hou, and Vincent Ng. 2025. Multimodal neural machine translation: A survey of the state of the art. In Christodoulopoulos, Christos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025, Suzhou, China, November 4-9, 2025*, pages 22130–22147. Association for Computational Linguistics.

[Finkelstein et al.2026] Finkelstein, Mara, Isaac Caswell, Tobias Domhan, Jan-Thorsten Peter, Juraj Juraska, Parker Riley, Daniel Deutsch, Geza Kovacs, Cole Dilanni, Colin Cherry, et al. 2026. Translategemma technical report. *arXiv preprint arXiv:2601.09012*.

[Futeral et al.2025] Futeral, Matthieu, Cordelia Schmid, Benoît Sagot, and Rachel Bawden. 2025. Towards zero-shot multimodal machine translation. In Chiruzzo, Luis, Alan Ritter, and Lu Wang, editors, *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 761–778, Albuquerque, New Mexico, April. Association for Computational Linguistics.

[Hsu et al.2024] Hsu, Benjamin, Xiaoyu Liu, Huayang Li, Yoshinari Fujinuma, Maria Nadejde, Xing Niu, Ron Litman, Yair Kittenplon, and Raghavendra Reddy Pappagari. 2024. M3T: A new benchmark dataset for multi-modal document-level machine translation. In Duh, Kevin, Helena Gómez-Adorno, and Steven Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Short Papers, NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 499–507. Association for Computational Linguistics.

[Huang et al.2022] Huang, Yupan, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. 2022. Layoutlmv3: Pre-training for document ai with unified text and image masking.

[Huang et al.2026] Huang, Ailin, Chengyuan Yao, Chunrui Han, Fanqi Wan, Hangyu Guo, Haoran Lv, Hongyu Zhou, Jia Wang, Jian Zhou, Jianjian Sun, et al. 2026. Step3-vl-10b technical report. *arXiv preprint arXiv:2601.09668*.

[Kaufman and Rousseeuw1990] Kaufman, Leonard and Peter Rousseeuw. 1990. *Finding Groups in Data: An Introduction To Cluster Analysis*. 01.

[Liu et al.2023] Liu, Haotian, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In Oh, A., T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 34892–34916. Curran Associates, Inc.

[Lu et al.2025] Lu, Jinghui, Haiyang Yu, Yanjie Wang, Yongjie Ye, Jingqun Tang, Ziwei Yang, Binghong Wu, Qi Liu, Hao Feng, Han Wang, Hao Liu, and Can Huang. 2025. A bounding box is worth one token - interleaving layout and text in a large language model for document understanding. In Che, Wanxiang, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Findings of the Association for Computational Linguistics: ACL 2025*, pages 7252–7273, Vienna, Austria, July. Association for Computational Linguistics.

[Lu et al.2026] Lu, Junxin, Tengfei Song, Zhanglin Wu, Pengfei Li, Xiaowei Liang, Hui Yang, Kun Chen, Ning Xie, Yunfei Lu, Jing Zhao, Shiliang Sun, and Daimeng Wei. 2026. Global-local dual perception for mllms in high-resolution text-rich image translation.

[O'Brien et al.2025] O'Brien, Dayyán, Bhavitvya Malik, Ona de Gibert, Pinzhen Chen, Barry Haddow, and Jörg Tiedemann. 2025. Dochplt: A massively multilingual document-level translation dataset. *CoRR*, abs/2508.13079.

[Shen et al.2024] Shen, Huangjun, Liangying Shao, Wenbo Li, Zhibin Lan, Zhanyu Liu, and Jinsong Su. 2024. A survey on multi-modal machine translation: Tasks, methods and challenges.

[Sun et al.2025a] Sun, Ting, Cheng Cui, Yuning Du, and Yi Liu. 2025a. PP-DocLayout: A unified document layout detection model to accelerate large-scale data construction. March.

[Sun et al.2025b] Sun, Yirong, Dawei Zhu, Yanjun Chen, Erjia Xiao, Xinghao Chen, and Xiaoyu Shen. 2025b. Fine-grained and multi-dimensional metrics for document-level machine translation. In Ebrahimi, Abteen, Samar Haider, Emmy Liu, Sammar Haider, Maria Leonor Pacheco, and Shira Wein, editors, *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop)*, pages 1–17, Albuquerque, USA, April. Association for Computational Linguistics.

[Yin et al.2020] Yin, Pengcheng, Graham Neubig, Wentau Yih, and Sebastian Riedel. 2020. TaBERT: Pretraining for joint understanding of textual and tabular data. In Jurafsky, Dan, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8413–8426, Online, July. Association for Computational Linguistics.

[Yin et al.2024] Yin, Shukang, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2024. A survey on multimodal large language models. *National Science Review*, 11(12):nwae403, 11.

[Zhang et al.2025a] Zhang, Le, Qian Yang, and Aishwarya Agrawal. 2025a. Assessing and learning alignment of unimodal vision and language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 14604–14614. Computer Vision Foundation / IEEE.

[Zhang et al.2025b] Zhang, Yaping, Yupu Liang, Zhiyang Zhang, Zhiyuan Chen, Lu Xiang, Yang Zhao, Yu Zhou, and Chengqing Zong. 2025b. Icdar 2025 competition on end-to-end document image machine translation towards complex layouts. In *International Conference on Document Analysis and Recognition*, pages 505–522. Springer.

[Zhao et al.2024] Zhao, Zhiyuan, Hengrui Kang, Bin Wang, and Conghui He. 2024. Doclayout-yolo: Enhancing document layout analysis through diverse synthetic data and global-to-local adaptive perception.

[Zuo et al.2025] Zuo, Fei, Kehai Chen, Yu Zhang, Zhengshan Xue, and Min Zhang. 2025. InImageTrans: Multimodal LLM-based text image machine translation. In Che, Wanxiang, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Findings of the Association for Computational Linguistics: ACL 2025*, pages 20256–20277, Vienna, Austria, July. Association for Computational Linguistics.

## Appendix A. Vector Indices Mapping

**Table 3:** Full enumeration of the 45-dimensional document feature vector mapping.

| Idx | Entity / Attribute | Idx | Entity / Attribute | Idx | Entity / Attribute |
|---|---|---|---|---|---|
| 0 | algorithm | 15 | footer_image | 30 | diff_font_names |
| 1 | ref_content | 16 | text | 31 | diff_font_sizes |
| 2 | doc_title | 17 | number | 32 | black |
| 3 | vision_footnote | 18 | chart | 33 | white |
| 4 | inline_formula | 19 | header_image | 34 | gray |
| 5 | display_formula | 20 | image | 35 | pink |
| 6 | footer | 21 | abstract | 36 | beige |
| 7 | reference | 22 | content | 37 | brown |
| 8 | seal | 23 | paragraph_title | 38 | red |
| 9 | table | 24 | figure_title | 39 | orange |
| 10 | aside_text | 25 | vertical | 40 | yellow |
| 11 | formula_num | 26 | normal (weight) | 41 | green |
| 12 | footnote | 27 | bold (weight) | 42 | teal_cyan |
| 13 | vertical_text | 28 | italic (weight) | 43 | blue |
| 14 | header | 29 | bolditalic (weight) | 44 | purple |

## Appendix B. Clustering Impact

**Table 4:** Impact of the clustering methodology on typographic color distribution.

| Color | Before (%) | After (%) | Change (%) |
|---|---|---|---|
| Yellow | 0.00062743 | 0.00228579 | +264.31% |
| Red | 0.01930891 | 0.06094932 | +215.65% |
| Purple | 0.01514131 | 0.03878174 | +156.13% |
| White | 0.774143 | 1.74502507 | +125.41% |
| Teal_Cyan | 0.0067147 | 0.01469656 | +118.87% |
| Pink | 4.986e-05 | 0.00010739 | +115.38% |
| Orange | 0.0006856 | 0.0014267 | +108.10% |
| Green | 0.19981624 | 0.38189586 | +91.12% |
| Black | 96.60691349 | 96.30528719 | -0.31% |
| Gray | 1.92919544 | 1.21493631 | -37.02% |
| Blue | 0.44740404 | 0.23460807 | -47.56% |

**Table 5:** Impact of the clustering methodology on entity labels distribution.

| Label | Before (%) | After (%) | Change (%) |
|---|---|---|---|
| aside_text | 0.65733778 | 1.58451011 | +141.05% |
| inline_formula | 0.02724798 | 0.05804066 | +113.01% |
| content | 2.44546664 | 5.1188963 | +109.32% |
| image | 2.06501872 | 4.15455068 | +101.19% |
| table | 0.96620069 | 1.55432896 | +60.87% |
| vision_footnote | 0.20388733 | 0.30239186 | +48.31% |
| chart | 0.00834765 | 0.01160813 | +39.06% |
| figure_title | 0.14120123 | 0.18776155 | +32.97% |
| paragraph_title | 12.50579807 | 14.15292554 | +13.17% |
| number | 5.94651679 | 6.30495725 | +6.03% |
| footer | 5.25318963 | 4.90675767 | -6.59% |
| reference | 0.00094502 | 0.00087061 | -7.87% |
| header | 5.69340985 | 5.24455433 | -7.88% |
| text | 57.56836211 | 51.29604801 | -10.90% |
| header_image | 1.11582832 | 0.9536081 | -14.54% |
| reference_content | 0.19420091 | 0.15670979 | -19.31% |
| footnote | 2.23606672 | 1.76472637 | -21.08% |
| doc_title | 1.21261377 | 0.93938814 | -22.53% |
| footer_image | 1.63763497 | 1.25483914 | -23.37% |
| seal | 0.00039376 | 0.0002902 | -26.30% |
| abstract | 0.08284645 | 0.04643253 | -43.95% |
| formula_number | 0.00102377 | 0.0002902 | -71.65% |
| algorithm | 0.02661797 | 0.00406285 | -84.74% |
| display_formula | 0.00984392 | 0.00145102 | -85.26% |

## Appendix C. Additional Translation Errors Examples

| | | |
|---|---|---|
| Seating position suitable for universal belted (Yes/No)* | Yes<br>Forward-facing only | Yes<br>Forward-facing only |

**(a)** Source text

| | | |
|---|---|---|
| Miejsce odpowiednie do zamocowania pasem bezpieczeństwa fotelika uniwersalnego (Tak/Nie)* | Tak<br>Przodem do kierunku jazdy | Tak<br>Przodem do kierunku jazdy |

**(b)** Original target text

| | | |
|---|---|---|
| Pozycja siedząca odpowiednia dla uniwersalnych pasów bezpieczeństwa (Tak/Nie)* | Tak<br>Tylko tyłem do kierunku jazdy | Tak<br>Tylko w kierunku jazdy |

**(c)** Translation system result

**Figure 12:** (Table cells translation error. In the center column, the system produces a critical semantic inversion, incorrectly translating the instruction as "Tylko tyłem" (Rear-facing only). This error represents a significant user safety risk and demonstrates the danger of translating table cells without structural context.

3. In view of the above, the Permanent Representatives Committee is invited to suggest that the Council:
   – adopt the abovementioned Implementing Decision as finalised by the legal/linguistic experts and set out in 7163/22 as an 'A' item on the agenda of a forthcoming meeting;
   – agree on the publication of the abovementioned Implementing Decision in the Official Journal.

**(a)** Source text

3. W związku z powyższym Komitet Stałych Przedstawicieli jest proszony o zaproponowanie Radzie, by:
   – przyjęła – jako punkt A porządku jednego z najbliższych posiedzeń – wyżej wymienioną decyzję wykonawczą w wersji ostatecznie zredagowanej przez prawników lingwistów i zamieszczonej w dokumencie 7163/22;
   – postanowiła o opublikowaniu tej decyzji wykonawczej w Dzienniku Urzędowym.

**(b)** Original target text

3. W związku z powyższym Komitet Stałych Przedstawicieli jest proszony o zasugerowanie, aby:
Rada:
– przyjąć wyżej wymienioną decyzję wykonawczą w brzmieniu ostatecznie zatwierdzonym przez prawników/językowców ekspertów i zamieszczono w 7163/22 jako punkt „A" porządku obrad nadchodzącego posiedzenia;
– wyrazić zgodę na opublikowanie w Dzienniku Urzędowym wyżej wymienionej Decyzji Wykonawczej Dziennik.

**(c)** Translation system result

**Figure 13:** Translation system losing semantic translation context between lines and misplacing the text underline.

| | |
|---|---|
| Subject: | Council Implementing Decision amending Implementing Decision (EU) 2019/310 as regards the authorisation granted to Poland to continue to apply the special measure derogating from Article 226 of Directive 2006/112/EC on the common system of value added tax<br>- Adoption |

**(a)** Source text

| | |
|---|---|
| Dotyczy: | Decyzja wykonawcza Rady w sprawie zmiany decyzji wykonawczej (UE) 2019/310 w odniesieniu do przyznanego Polsce zezwolenia na dalsze stosowanie szczególnego środka stanowiącego odstępstwo od art. 226 dyrektywy 2006/112/WE w sprawie wspólnego systemu podatku od wartości dodanej<br>– Przyjęcie |

**(b)** Original target text

| | |
|---|---|
| Temat: | Decyzja wykonawcza Rady zmieniająca decyzję wykonawczą (UE) 2019/310 w odniesieniu do upoważnienia udzielonego Polsce do dalszego stosowania szczególnego środka stanowiącego odstępstwo od art. 226 dyrektywy 2006/112/WE w sprawie wspólnego systemu podatku od wartości dodanej<br>- Adopcja |

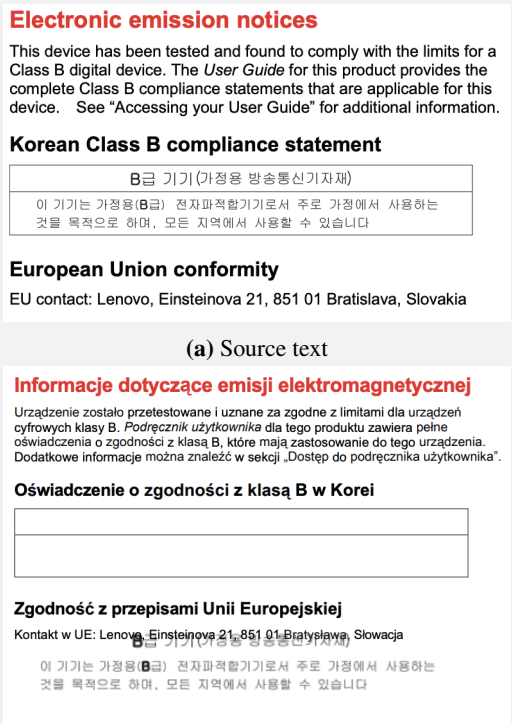
**(c)** Translation system result

| | |
|---|---|
| Temat: | Dyrektywa wykonawcza Rady zmieniająca decyzję wykonawczą (UE) 2019/310w odniesieniu do upoważnienia udzielonego Polsce do dalszego kontynuowania stosować środek szczególny stanowiący odstępstwo od art. 226 dyrektywy 2006/112/WE w sprawie wspólnego systemu podatku od wartości dodanej<br>- Adopcja |

**(d)** Translation system result

**Figure 14:** The ground-truth translation correctly renders "Adoption" as "Przyjęcie" (Legal Adoption/Approval) to match the legislative subject matter. However, the tested systems exhibit a significant domain mismatch, mistranslating the term as "Adopcja" (Biological/Family Adoption). This error stems from a loss of contextual continuity between layout elements.

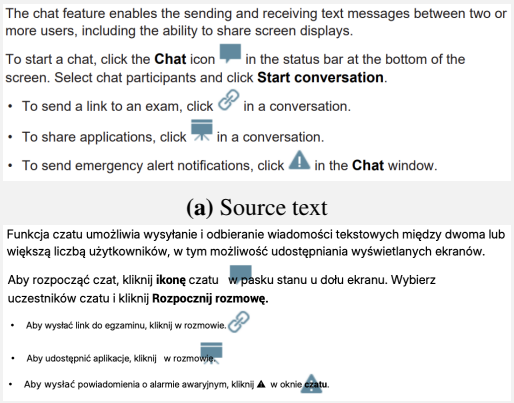
# Appendix D. Additional Reconstruction Errors Examples




**(a)** Source text

**(b)** Translation system result

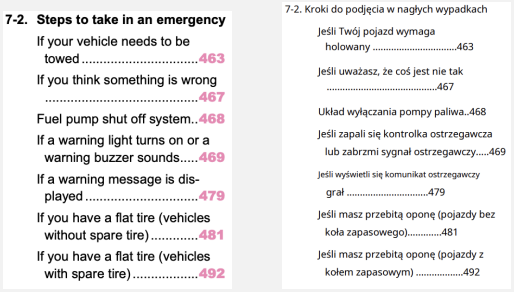
**Figure 15:** Geometric synchronization failure and layer detachment. In the original document, the compliance statement is properly encapsulated within a table structure. However, the translation system fails to maintain the link between the bounding box and its textual content.




**(a)** Source text

**(b)** Translation system result

**Figure 16:** Inline asset collision and anchor failure. The original document features functional icons embedded within the text flow. The translation system fails to account for these inline graphical assets during the reconstruction phase.




**(a)** Source text

**(b)** Translation system result

**Figure 17:** Failure in structural reconstruction and stylistic preservation. In the reconstructed output, the system fails to preserve the typographic weight and the pink-colored indices, rendering all elements in a default black font. Furthermore, the translation exhibits a significant vertical alignment drift.