NANYANG
TECHNOLOGICAL
UNIVERSITY
SINGAPORE

*College of Computing and Data Science*

# AI6127 Group Project Report

## XML-based English to German Translation System with Legal Domain Specialization

**Supervisor:** Prof. Luu Anh Tuan

| Name | Matric Number | Email |
|---|---|---|
| Sha Long | G2402907A | c240102@e.ntu.edu.sg |
| Sudhakar Selvaraj | G2404967F | sudhakar004@e.ntu.edu.sg |
| Chai Weichen | G2403998E | weichen001@e.ntu.edu.sg |
| Lei Litian | G2405411E | litian001@e.ntu.edu.sg |
| Deng Jie | G2403273K | jie009@e.ntu.edu.sg |

**Semester 2 June 1, 2025**

# Abstract

This project investigates approach of translation from English to German in legal documents while preserve the original document structure and specialized terms. We built an XML-aware translation system based on the Helsinki-NLP/opus-mt-en-de neural translation model with XML structures and domain-specific legal terminology support. The system determines which parts of the text require translation, while carefully preserving the structure and content of elements such as references and dates. Our approach incorporates a terminology control mechanism to preserve accuracy in the translation of legal terms. The evaluation process involves BLEU, METEOR and TER metrics and the results show that our system performs outstanding in this task.

# 1 Introduction

Translating legal documents directly always faces challenges that go beyond those encountered in general machine translation tasks. In such documents the accurate rendering is not just the textual content, but strict preservation of structural elements. Although the field of neural machine translation is advancing rapidly, general-purpose systems still cannot handle data with a more structured (e.g., XML files with mixed content), where only certain elements are to be translated while tags or metadata must not change.

Current approaches to document translation mainly focus on extracting plain text for translation and then reconstructing the document structure. This approach often results in errors in document structure and inconsistent treatment to specialized terminology. In addition, legal documents are particularly demanding on terminology accuracy, as mistranslations may have serious consequences for interpretations.

Our project addresses these challenges by building a professional translation system that combines neural machine translation technology with XML document processing and legal terminology control. This system is able to intelligently preserve document structure while ensuring high-quality translation and consistent terminology usage. This work focuses on the translation of legal documents from English to German, addressing a practical need within the context of European legal and regulatory communication.

# 2 Related Work

## 2.1 Legal Translation Practices and CAT Tools

Professional legal translation remains predominantly manual, carried out by experienced translators who deeply understand both source and target legal systems. To improve consistency and efficiency, practitioners commonly employ computer-aided translation (CAT) tools such as SDL Trados Studio and MemoQ, which leverage translation memories and terminology databases to ensure uniform use of legal terms and track revisions [1, 2]. While CAT tools excel at managing glossaries and previous translations, they still require substantial human intervention and cannot fully automate the translation process or adapt to unseen terminology in new legal domains.

## 2.2 Neural Machine Translation Basics

### 2.2.1 Sequence-to-Sequence and Attention

The advent of neural machine translation (NMT) began with the sequence-to-sequence (Seq2Seq) framework, in which an encoder network compresses the source sentence into a fixed-size vector and a decoder network generates the target sentence step by step [3]. Bahdanau et al. introduced the attention mechanism, allowing the decoder to dynamically focus on different parts of the source representation at each time step, thereby significantly improving translation of long and complex sentences [4].

### 2.2.2 Transformer and Pretrained Models

Vaswani et al. revolutionized NMT with the Transformer architecture, which replaces recurrence with self-attention layers and enables fully parallel training, achieving state-of-the-art results on WMT14 English–German (28.4 BLEU) [5]. Building on this, large pretrained Seq2Seq models such as mBART and mT5 have demonstrated further gains by leveraging massive monolingual and parallel corpora before fine-tuning on downstream translation tasks, though they still require domain-specific adaptation to handle specialized vocabulary.

### 2.3 Challenges and Structured-Document Processing in Legal NMT

Applying general-purpose NMT directly to legal texts faces several obstacles:

- **Domain mismatch:** Models trained on news or web data often produce stylistically inappropriate translations for legal prose.
- **Rare terminology:** Legal documents contain many low-frequency terms and Latinisms that neural models struggle to learn consistently.
- **Long, nested sentences:** Statutes and contracts frequently exceed typical model length limits, resulting in truncated context and degraded fluency.
- **Term consistency:** Inconsistent rendering of the same legal concept can introduce ambiguity with serious interpretive consequences.

### 2.4 Baseline Model Selection: Opus-MT

For our baseline we adopt the open-source Helsinki-NLP/opus-mt-en-de model for four key reasons:

- **Reproducibility:** Model weights and training scripts are publicly available on HuggingFace, facilitating academic replication.
- **Proven legal performance:** On legal parallel corpora such as JRC-Acquis, Opus-MT variants outperform generic NMT baselines on BLEU and other metrics [6].
- **Easy integration:** The HuggingFace `transformers` pipeline offers one-line inference calls, allowing seamless embedding within our XML-aware processing workflow.
- **Rich community resources:** Backed by the OPUS parallel data project, Opus-MT provides extensive terminology lists and multilingual coverage, which we leverage for domain adaptation.

This combination of accessibility, strong out-of-the-box performance on legal texts, and ease of extension makes Opus-MT the most suitable choice as our baseline.

## 3 Methodology

Our XML translation system implements a specialized pipeline designed to process structured legal documents while maintaining their integrity and ensuring high-quality translations. The approach consists of several key components working together to address the unique challenges of translating structured legal content.

### 3.1 System Architecture

The system used for this assignment is structured with three major components: data extraction, fine-tuning, and XML translation. The English-German data are extracted from the XML files, then cleaned and aligned, and saved in a tab-separated format. The extracted data is then used for fine-tuning with the Helsinki-NLP/opus-mt-en-de model. In order to preserve model efficiency, the encoder layers are frozen during training. Finally, the XML translation component reads the english XML documents and translates them to German XML format.

### 3.2 XML Processing

XML processing method used ensures that when translating english XML files to German ones, the original document structure is maintained. Semantic tags such as `<title>` and `<note>`, which contain actual descriptive texts, are translated, while `<xref>` and `<date>` metadata are excluded to ensure consistency. The pre-processing system uses a chunking operation to prevent long sentences from exceeding a token threshold.

### 3.3 Legal Terminology Management

As all the data used in the experiment are legal documents, in order for the better directed processing, a strict terminology control mechanism is established in this experiment to guarantee the consistency and accuracy of the domain-specific terminology in the translation process. A glossary of selected English-German legal term pairs is included in the training. During model training and inference, the system detects and matches the corresponding terms in the glossary in real time, and then automatically performs term substitution operations. This mechanism effectively avoids the terminological ambiguity or presentation inconsistency issues that may arise in legal text translation, which plays a key role in ensuring the professionalism and reliability of legal document translation.

### 3.4 Language Attribute Handling

To address the multilingual nature of legal XML documents, a selective translation strategy based on language tag recognition is used in this assignment. Legal XML documents usually contain explicitly defined language attributes, indicated by tags such as `<p lang=\en">`.

The system first parses the document structure to recognize and extract these linguistic attribute tags. According to the preset language processing rules, translation operations are performed only on text nodes that meet the target language conversion conditions, leaving translated content or passages marked as language exempted as it is. This approach not only optimizes the allocation of computational resources, more importantly, it ensures that the completeness of the original metadata of the document and the certified translated content are not damaged.

### 3.5 Batch Processing

In order to optimize the processing efficiency, the system supports batch translation. Batch translation is mainly divided into two parts:

1. The system parses the input XML document, extracts all translatable text fragments, and intelligently groups them based on:
    (a) Length of text fragments
    (b) Positional relationship in the document hierarchy
2. After the translation is completed, the system restores the processed text to the original document structure accurately through XML path localization.

This batch strategy can significantly reduce the number of calls to the translation engine, thus shortening the overall processing time. More importantly, it ensures terminological and stylistic consistency between semantically related text segments, and strictly maintains the original markup attributes and document hierarchies.

## 4 Experiments

The experiment is designed to translate English XML files into German while preserving their structural integrity and ensuring accurate translation of domain-specific terminology, such as legal terms. It comprises two primary phases:

1. **Translation Phase:** English XML files are processed using a neural machine translation model to convert translatable content into German. The system handles specific XML tags differently, translating some (e.g., `title`, `p`) while preserving others (e.g., `date`, `xref`), and applies rules to ensure correct translation of legal terms.
2. **Evaluation Phase:** The translated (hypothesis) files are compared against reference German XML files to assess translation quality using multiple automatic metrics, including a custom metric for domain-specific term accuracy.

The experiment is implemented through a Python script that leverages libraries like `transformers` for translation and `pandas` for reporting, with the goal of producing high-quality translations for legal or official documents.

## 4.1 Data

### 4.1.1 Input Data

- **Format**: The input consists of XML files containing English text with structured elements, stored in a directory specified by `INPUT_DIR` (default path: **/content**)
- **Content**: The XML files contain both translatable tags (e.g., `title`, `head`, `p`, `note`) and non-translatable tags (e.g., `xref`, `date`, `classCode`, `bibl`), featuring domain-specific legal terminology that requires precise German equivalents (e.g., "Official Journal" → "Amtsblatt", "paper edition" → "Papierausgabe", "deemed authentic" → "verbindlich gelten").
- **Characteristics**: The data contains specialized legal language with exact translation requirements, where specific English terms must be consistently mapped to their German counterparts while maintaining the XML structure and preserving non-translatable elements.

### 4.1.2 Reference Data

The reference data consists of XML files containing German translations that serve as ground truth for evaluation, stored in a directory specified by `ref_dir` (default location: **/content/test_refs**), and are used to compute quality metrics through comparison with translated output files.

### 4.1.3 Output Data

- **Translated Files**: The translated output consists of XML files containing German translations with updated language attributes (e.g., `lang="de"`) and preserved document structure, stored in the directory specified by `OUTPUT_DIR` (default location: **/content/test_hyp**).
- **Evaluation Report**: The system generates an evaluation report in CSV format (`translation_quality.csv`) stored in **/content**, containing quality metrics including BLEU, METEOR, TER, COMET, and domain-specific term accuracy scores for comprehensive translation assessment.

## 4.2 Translation Method

The translation process is implemented through the `XMLTranslator` class, which combines a pre-trained machine translation model with custom processing for XML files. Key components include:

### 4.2.1 Model

The translation model is `Helsinki-NLP/opus-mt-en-de`, a neural machine translation model optimized for English-to-German translation, provided by the Language Technology Research Group at the University of Helsinki. It is accessed via the `transformers` library's `pipeline` function, configured with parameters such as `max_length=1000`, `clean_up_tokenization_spaces=True`, and `num_beams=5` for improved translation quality.

### 4.2.2 Process

1. **Initialization:** The `XMLTranslator` class is initialized with the model name and defines translatable tags (e.g., `title`, `head`, `p`, `note`), preserved tags (e.g., `xref`, `date`, `classCode`, `bibl`), and terminology rules, including regex patterns to replace specific legal terms (e.g., converting "Official Journal" to "Amtsblatt").
2. **Translation:** The `translate_text` method processes individual text snippets using the translation model, followed by post-processing where regex patterns are applied to ensure domain-specific terms are correctly translated.
3. **XML Processing:** The `process_xml` method parses the input XML file using xml.etree.ElementTree, updates language attributes in the root and `teiHeader` elements to "de", translates content within translatable tags while preserving non-translatable tags, and maintains whitespace formatting to ensure structural consistency.

4. **Batch Processing:** The `batch_translate_xml` function processes all `.xml` files in the input directory, saves translated files to the output directory, and prints progress updates (e.g., the number of files processed) to enhance usability for large datasets.

## 4.3 Features

- **Structural Preservation:** Non-translatable tags are kept unchanged to maintain the document's integrity.
- **Domain-Specific Accuracy:** Regex-based term replacement ensures legal terms are translated correctly.
- **Scalability:** Batch processing supports large-scale translation tasks.

## 4.4 Evaluation Methods

The evaluation is handled by the `TranslationEvaluator` class and the `generate_report` function.

### 4.4.1 Metrics

| Metric | Description | Score Range | Interpretation |
|---|---|---|---|
| **BLEU** | Measures n-gram precision between translated and reference text. | 0–1 | Higher is better |
| **METEOR** | Considers synonyms, stemming, and word order for a nuanced evaluation. | 0–1 | Higher is better |
| **TER** | Calculates the number of edits needed to match the hypothesis to the reference. | 0–$\infty$ | Lower is better |
| **COMET** | Neural-based quality estimation score using the WMT21 COMET-QE-DA model. | 0–1 | Higher is better |

Table 1: Translation evaluation metrics

### 4.4.2 Process

1. **Alignment:** The alignment process begins by using the `evaluate_file` method to match lines between reference and hypothesis files through exact string comparison. This ensures that corresponding segments from both files are directly compared for accurate metric evaluation. If no matches are detected during alignment, a warning is logged to highlight potential misalignment issues, which could arise from mismatched file structures or formatting discrepancies. This step is critical to maintain the integrity of subsequent analyses.

2. **Filtering:** Before computing metrics, a filtering step replaces missing values (`None`) in the hypothesis or reference texts with empty strings. This preprocessing measure prevents errors during calculations, ensuring that null values do not disrupt the evaluation workflow. By standardizing the input data, the system maintains consistency and reliability in metric computation.

3. **Metric Computation:** Metrics are calculated for each aligned line pair between the reference and hypothesis texts. The process incorporates error-handling mechanisms to address issues such as missing data or formatting errors, ensuring robustness. A custom *term accuracy* metric is also applied, which evaluates the presence of predefined legal terms (e.g., statutory references or jurisdictional phrases) in the translated text. This specialized metric ensures domain-specific quality assessment alongside standard measures like BLEU and METEOR scores.

4. **Reporting:** The `generate_report` function systematically processes all hypothesis files in the output directory, pairing them with their corresponding reference files for comparison. Results are aggregated into a pandas DataFrame and exported as `translation_quality.csv`, providing a structured record of scores across files. Additionally, summary statistics—including average BLEU, METEOR, and term accuracy scores—are printed to offer a concise, high-level overview of translation quality. This facilitates quick interpretation and decision-making based on key performance indicators.

### 4.5 Model settings

- model_name = "Helsinki-NLP/opus-mt-en-de"
- learning_rate = 5e-6
- bach_size = 8
- num_train_epochs = 3
- tokenizer = AutoTokenizer.from_pretrained(model_name)

In the fine-tuning process of the model "Helsinki-NLP/opus-mt-en-de", we use the Seq2SeqTrainer framework with 5e-6 learning rate and batch size 8. The tokenizer is provided by the hugging face transformer library which corresponds to this model. Training is carried out for 3 epochs with weight decay set to 0.01. And to preserve the obtained the information and reduce computational power, during the fine-tuning process, the encoder weights are not updated, and only the decoder part or the newly added modules are trained.

### 4.6 Results & Analysis

We take one document `jrcC2006#294#37-de.xml` from *Joint Research Centre Acquis Communautaire*[6] which is the collection of EU legal documents as the example. The output document generated by our system, along with the original English version, are included in the Appendix. As shown in Figure 2, some tags such as <note> and <title> are translated to German, while tags such as <bibl> and <xref> are preserved.

The Table 2 is result of comparison with baseline. It can be observed from Figure 1 that all metrics of our system are better than the baseline (the less the TER is, the better the performance is), which suggests compared with baseline, our system yields a more fluent, semantically accurate and closer to human-level translation.

| System | BLEU (%) | Avg Meteor (%) | COMET-QE (%) | TER |
|---|---|---|---|---|
| Baseline | 15.24 | 32.81 | 44 | 125 |
| Our System | 26.45 | 40.04 | 63 | 95 |

Table 2: Comparison of translation performance between baseline and our system across multiple metrics.

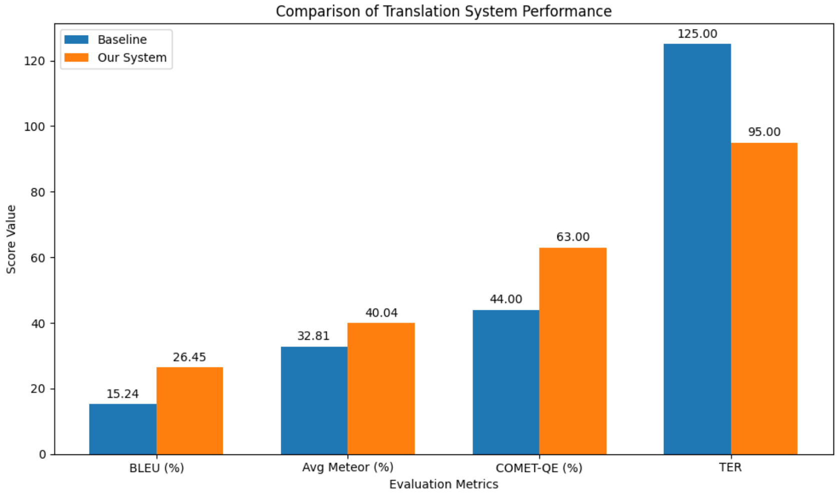


Figure 1: Metrics Comparison

# 5 Conclusion

This project demonstrates an approach for translating legal documents from English to German while maintaining the accuracy of legal language and structural formatting. The integration of neural machine translation (NMT) with XML processing allows for aligned translation of documents, which effectively solves legal translation problems.

Evaluation shows strong performance in multiple metrics, particularly in language accuracy. This affirms the precision of the system in the legal language. XML constituent meta-data identification and translation capture and maintain the specific document structure and meaning, which is fundamental to the entire document.

There is still room to improve this system with additional languages, an expanded terminology lexicon, and more document management or compliance workflows. Further training on the translation model with legal parallel corpora can refine the focus on specialized content.

In summary, this system provides a valuable tool for organizations that frequently deal with multilingual legal documents, especially in areas such as regulatory compliance, legal cases, and international trade, where accurate translation is extremely important.

# 6 Team Contributions

| **Team Member** | **Contribution (%)** | **Responsibilities** |
|---|---|---|
| **Sha Long, G2402907A** | 20% | Preliminary Research, Planning, Work Separation & Distribution, Slides & Presentation, Report. |
| **Sudhakar Selvaraj, G2404967F** | 20% | Model Design and Development, Performance test on the dataset, Report |
| **Chai Weichen, G2403998E** | 20% | Model Development, Fine-tuning of model, experimental testing and analysis, report. |
| **Lei Litian, G2405411E** | 20% | Model Development, Fine-tuning of model, experimental testing and analysis, report. |
| **Deng Jie, G2403273K** | 20% | Model Development, Report |

Table 3: Team Members' Contributions

# References

[1] SDL. Sdl trados studio. `https://www.sdl.com/software-and-services/translation-software/sdl-trados-studio/`, 2025.

[2] Kilgray Translation Technologies. memoq translator pro. `https://www.memoq.com/`, 2025.

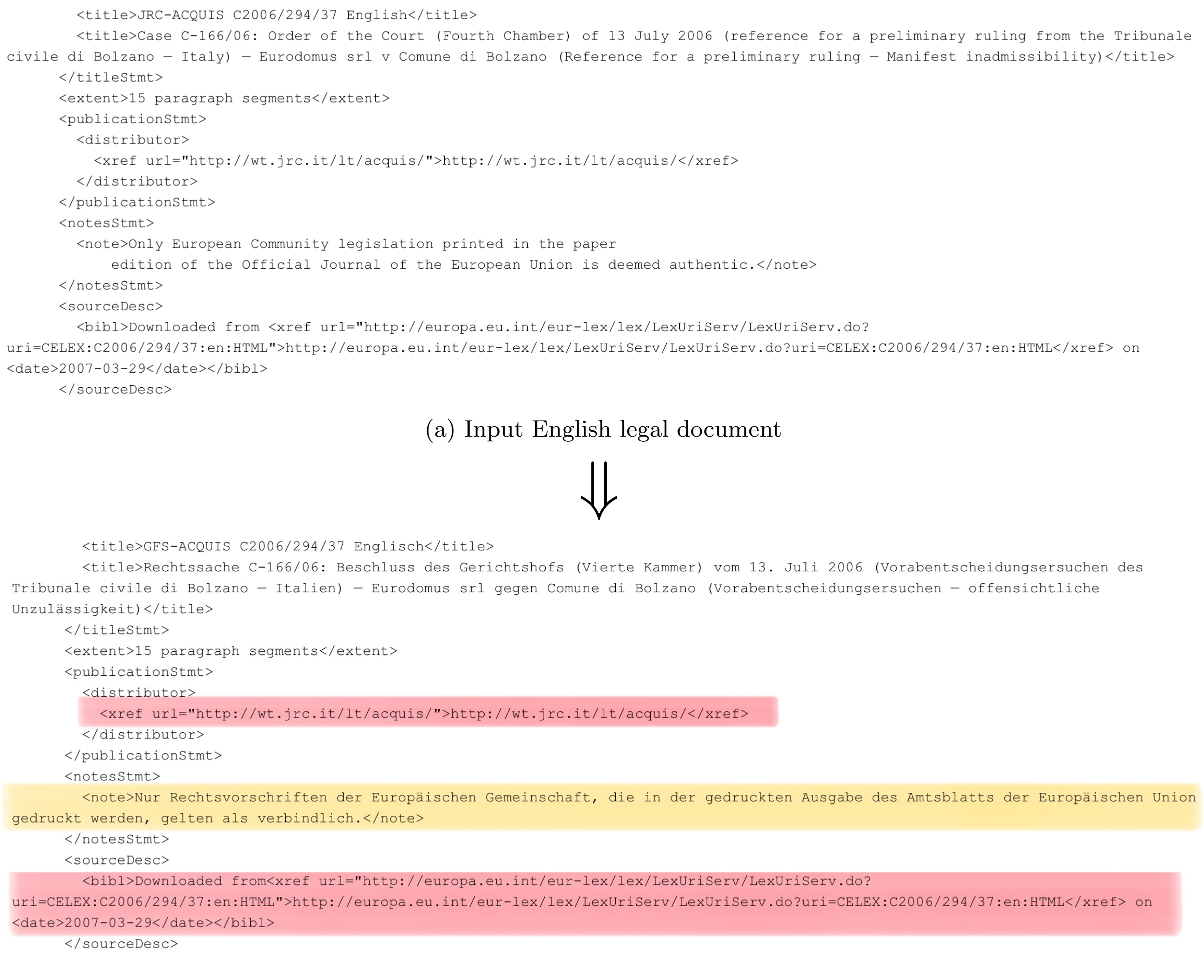
[3] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*, volume 27, 2014.

[4] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *Proceedings of the International Conference on Learning Representations*, 2015.

[5] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[6] Joint Research Centre. Acquis communautaire, 2006. `https://joint-research-centre.ec.europa.eu/language-technology-resources/jrc-acquis_en`.

# A Appendix



```
        <title>JRC-ACQUIS C2006/294/37 English</title>
        <title>Case C-166/06: Order of the Court (Fourth Chamber) of 13 July 2006 (reference for a preliminary ruling from the Tribunale
civile di Bolzano — Italy) — Eurodomus srl v Comune di Bolzano (Reference for a preliminary ruling — Manifest inadmissibility)</title>
      </titleStmt>
      <extent>15 paragraph segments</extent>
      <publicationStmt>
        <distributor>
          <xref url="http://wt.jrc.it/lt/acquis/">http://wt.jrc.it/lt/acquis/</xref>
        </distributor>
      </publicationStmt>
      <notesStmt>
        <note>Only European Community legislation printed in the paper
            edition of the Official Journal of the European Union is deemed authentic.</note>
      </notesStmt>
      <sourceDesc>
        <bibl>Downloaded from <xref url="http://europa.eu.int/eur-lex/lex/LexUriServ/LexUriServ.do?
uri=CELEX:C2006/294/37:en:HTML">http://europa.eu.int/eur-lex/lex/LexUriServ/LexUriServ.do?uri=CELEX:C2006/294/37:en:HTML</xref> on
<date>2007-03-29</date></bibl>
      </sourceDesc>
```

(a) Input English legal document

⇓

```
        <title>GFS-ACQUIS C2006/294/37 Englisch</title>
        <title>Rechtssache C-166/06: Beschluss des Gerichtshofs (Vierte Kammer) vom 13. Juli 2006 (Vorabentscheidungsersuchen des
Tribunale civile di Bolzano — Italien) — Eurodomus srl gegen Comune di Bolzano (Vorabentscheidungsersuchen — offensichtliche
Unzulässigkeit)</title>
      </titleStmt>
      <extent>15 paragraph segments</extent>
      <publicationStmt>
        <distributor>
          <xref url="http://wt.jrc.it/lt/acquis/">http://wt.jrc.it/lt/acquis/</xref>
        </distributor>
      </publicationStmt>
      <notesStmt>
        <note>Nur Rechtsvorschriften der Europäischen Gemeinschaft, die in der gedruckten Ausgabe des Amtsblatts der Europäischen Union
gedruckt werden, gelten als verbindlich.</note>
      </notesStmt>
      <sourceDesc>
        <bibl>Downloaded from<xref url="http://europa.eu.int/eur-lex/lex/LexUriServ/LexUriServ.do?
uri=CELEX:C2006/294/37:en:HTML">http://europa.eu.int/eur-lex/lex/LexUriServ/LexUriServ.do?uri=CELEX:C2006/294/37:en:HTML</xref> on
<date>2007-03-29</date></bibl>
      </sourceDesc>
```

(b) Output German legal document

Figure 2: Translation from English version to German version.