Document Translation

# Overview

Document translation is a new feature  [of the Azure Translator service](https://azure.microsoft.com/en-us/services/cognitive-services/translator/) that enables enterprises, translation agencies, and consumers who need to translate complex document volumes into one or more languages while preserving the structure and formatting of the original document. Asynchronously translate entire documents in a variety of file formats**, including text, HTML, Word, Excel, PowerPoint, Outlook, PDF, and Markdown** in any of the **111 languages and dialects** supported by the Translator service.

Standard translation offerings on the market only accept plain text or html, and limit the number of characters in a request. Users translating large documents must parse the documents to extract text, break them down into smaller sections, and translate them separately. If sentences are broken up at an unnatural breakpoint, it can lose context, resulting in suboptimal translations. Upon receiving the translation results, the client must merge the translated parts into the translated document. This involves keeping track of which translated piece corresponds to the equivalent section in the original document. The problem is complicated when customers want to translate complex documents with rich content.



Document translation makes it easier for the client to translate:

1. large document volumes,
2. documents in a variety of file formats,
3. documents that require preservation of the original presentation and formatting, and
4. documents in multiple target languages.

# User Experience

The user makes a request to the document translation service specifying the location of the source and destination documents, and the list of target languages. The service returns an identifier that allows the user to track the status of the translation. Asynchronously, document translation extracts each document from the source location, recognizes the format of the document, applies the correct parsing technique to extract the textual content from the document, and translates the textual content into the target languages. It then reconstructs the translated document while preserving the layout and formatting present in the source documents, and stores the translated document in a specified location. Document Translation updates the status of the translation at the job or document level.

Users can provide a custom model identifier created with the Custom Translators portal, custom glossaries, or both as part of the document translation request. Document translation applies such customization, preserving specific terminologies and providing domain-specific translations in the translated documents.