Dealing with Foreign Language Documents in Litigation and Other Legal Matters

A Catalyst White Paper
Dealing with Foreign Language Documents in Litigation and Other Legal Matters

With many S&P 500 companies now getting over half of their revenues from foreign sales, foreign language documents are quickly becoming a routine part of the legal practice. Whether for cross-border litigation, multi-national investigations, antitrust reviews or simple due diligence, legal professionals increasingly need to search, organize and review documents in a multitude of languages.

Many of the litigation support systems and other legal document review platforms in use today were designed primarily to support the English language. These systems were built in the 1990s or even later and rely upon technology that makes it difficult, if not impossible, for them to support many foreign languages. With multiple language documents becoming increasingly common, corporations and their counsel require legal support tools that are built for the global economy.

This whitepaper provides guidance on the issues relating to reviewing foreign language documents. It starts by addressing the limitations of the ASCII code used by most current systems and shows how Unicode, the new global standard, provides a solution. It then looks at foreign language search, particularly the thorny issues relating to searching Chinese, Japanese, Korean and Thai (CJK languages), which use symbols instead of letters and often do not use punctuation or spaces to show word boundaries. Next, the paper examines the additional challenges presented by searching multi-language document sets. Finally, it addresses some of the current best practices for creating efficient review assignments in multi-language matters.

Reading Foreign Language Documents

Computers read bits and bytes, not letters or characters. The translation of letters and characters into bits and bytes occurs through a process known as “encoding.” Encoding has evolved significantly during recent years from the old English-oriented standard, ASCII, to the new global standard, Unicode.

The Limitations of ASCII

The American Standard Code for Information Interchange (ASCII), was developed in the 1960s to encode basic characters so computers could read them, then output them in a way that people could understand.

In ASCII, each character is presented to the computer not as letters, but as a byte of code. A byte consists of 8 individual bits, that are either a “1” or “0”. The original ASCII used 7 of those 8 bits, giving us 128 different combinations to work with. This was enough to cover our basic alphabet along with numbers and punctuation, but not much else.

ASCII did the job it was built for, and for more than three decades was sufficient to support text transcripts, word processing programs and the programming used in litigation support applications. However, as business communications became more global, the 128 combinations became insufficient to handle languages with special characters such as German, Russian and Hebrew, let alone Asian character-based languages.

According to Fulbright’s Third Annual Litigation Trends Survey of 422 corporations, more than half had litigation matters pending outside the United States.
Programmers extended ASCII to 8 bits, doubling the range of characters to 256, however this was a temporary solution. The constraints of ASCII led to the development of proprietary code sets to support different languages, and operating systems had to be able to recognize each one. By the 90s, people started looking for a universal code with the capability to handle thousands of language variants.

The Invention of Unicode

The development of Unicode boosted the 256 character ASCII capabilities to 1,114,112 possible characters, accommodating symbolic languages like Chinese, Japanese, Korean and Thai (commonly abbreviated CJK). Using Unicode character sets is emerging as the global standard for foreign language documents. Various forms of Unicode encoding exist in all modern operating systems and languages, including Windows XP, Mac OSX, Linux and .NET.

UTF-8 (Unicode Transformation Format) is the most common form of Unicode encoding. It can use up to four bytes to represent any possible language character. ASCII characters occupy the first byte and still use the same digits, which makes UTF-8 backward compatible with ASCII. The compatibility with 30+ years of ASCII files accounts for UTF-8’s status as the lingua franca of Unicode.

The second byte can be used to describe another 65,000 characters, which covers many foreign language needs, and CJK languages use the third byte. The fourth is largely unused — at least for now.

Why this Matters

Unless your software application understands Unicode, it will not handle foreign-language documents easily. Most off-the-shelf litigation packages and processing software in use today only understand ASCII, and, as a result, will fail to decode email and documents with foreign language characters, filling in symbols such as ☛ and ???.

In order to successfully search and review foreign language documents, you need to make sure the software used to collect and process them is Unicode compliant. Then it needs to be loaded into a database or other software that supports Unicode.

Searching Foreign Language Documents

How a Document Search Works

To speed up the search process, computer software is programmed to find all the words in a document when the document is loaded and put them into a word index. A sophisticated index will tell you which documents contain the word you’re looking for as well as the page, line number and even position in the line. This way, the computer doesn’t need to open and search every document in your repository collection, it just needs to open the index and identify which documents contain that word.

Indexing is how we can search millions of pages of case documents using Lexis or Westlaw and the same principles apply to litigation support systems.
According to the Beijing Municipal Bureau of Commerce, 185 of the Fortune world 500 top multinationals have invested US$12.26 billion on 355 projects in Beijing.

Tokenization

When computers index, they don’t actually read the words, rather they identify them by the spaces and punctuation surrounding them. This process is called tokenization. Technically speaking, each letter/number combination is called a token — whether it’s a word is irrelevant to the computer.

In most Western Languages, words are separated by spaces and punctuation, so a search engine can index a French document just as easily as an English one. However, languages that don’t use spaces or punctuation to delineate words pose a problem for most search engines. In Chinese, Japanese and Thai, for example, characters run together without any clear break points. Korean uses spaces but contains compound words that are really several words put together.

To define the individual words in a sentence and create a document index, search engines with special tokenizers, such as FASTInstream®, must be used to recognize each language and determine where words begin and end.

Dealing with Pictorial Languages

As indicated above, CJK languages use pictorial characters called logograms. In order to perform a successful search, the search engine must figure out how to group the pictorial characters into words, but, perhaps more challenging, it must recognize how the characters are combined, which affects their meaning.

For example, the Traditional Chinese word for “Chinese” consists of three logograms that directly translate to “middle country people”: 中國人

If you wanted to say China rather than Chinese, you use two characters that represent “middle country”: 中國

When characters are all lumped together, typical search engine functionality would make it impossible to discern whether all three characters should be read together or if each one should be read individually.

To overcome this challenge, linguists have designed software that can read characters and understand their context sufficiently to create an index. Other languages pose other challenges: in Korean, the verb comes at the end, but the words preceding it can be in any order you choose. Arabic uses spaces to delineate words, but they use different grammatical forms tied into them.

Bottom line, to accurately search foreign language documents, you need a search engine with a tokenizer that is capable of handling the peculiarities of the world’s major languages.

Enabling Computers to Deal with Multiple Language Document Collections

To effectively handle foreign language documents, your indexing engine must first determine which language is being used. After identifying the language, the engine then applies the appropriate tokenizer to accurately identify word boundaries. This task becomes more complex when a document set or even an individual document (like an e-mail) contains more than one language, each of which has its own tokenization rules.
Many of the older document review platforms deliver foreign language support through language packs. The system applies tokenization rules for a single language to the entire collection of documents. This means that searches run against the collection may miss applicable terms in documents that are in another language, jeopardizing search accuracy.

The common workaround for these problems when using an older system calls for setting up a separate repository for each language in the document collection (assuming you know what all of those languages are), running your search against each repository, and then manually combining the results.

More modern review platforms take a different approach. They accept the concept that document collections, and even individual documents, can contain multiple languages. They analyze documents in parts, identify the primary and secondary language in each part, and apply the appropriate tokenizing schemes to each part of the document. With this sort of platform, a multi-language document set can be indexed, appropriately tokenized and searched in a single repository.

## Enabling Efficient Review Assignments

Attorneys take two different approaches toward reviewing foreign language documents. The first is to organize documents by language and assign them to reviewers with the requisite linguistic capabilities. The second is to translate documents and review them in English. Each approach presents its own challenges.

### Organizing Documents by Language

Organizing documents by language requires a system that can identify the primary and secondary languages of each document and sort them accordingly. Where a document contains multiple languages, it may need to be placed in multiple folders. Many older systems lack this capability.

Once sorted, the documents can then be culled further using sophisticated key term searches. This approach enables review teams to both receive languages they can read and enjoy the efficiencies that flow from reviewing document sets that have a common theme to them.

At any stage during the review, hot documents can be translated for review by others who lack fluency in the native language.

### Translating Documents into English for Review

The main alternative to using bilingual review teams is to translate documents into English for review. The volumes of data involved in many e-discovery cases makes manual translation prohibitively expensive and time consuming, so attorneys have turned to machine translations.

Machine translations have the benefit of being relatively quick and fairly inexpensive. The downside of the current translation software packages is the accuracy of the translation which, generally good, falls short of that of a professional manual translation.

This poses a couple of issues. First, it makes it difficult to use sophisticated English searches to cull the documents reliably before making them available for review. One approach to this issue is to cull using a multi-language platform and then machine translate the culled documents for review.
Second, it means creating mechanisms to verify the accuracy of translations rapidly. One way of doing this is to set up the review so that the original document is tied to its translation as a duplicate. Modern systems allow this setup and enable a bilingual attorney doing quality control to toggle quickly between the original document and its translation.

**What To Look For**

If you haven’t encountered foreign language documents in a legal matter yet, you will. If you’re in the market for a document review platform, save yourself the headache of a system that will soon become outdated and look for these attributes.

- A system that handles Unicode
- An indexer that tokenizes non-Western languages
- A search engine that supports multiple languages within a single repository
- A system that can batch documents by language
- A system that allows you to toggle quickly between original documents and their translations

**About Catalyst**

Catalyst builds and hosts software that allows cross-enterprise teams to review massive data sets and manage complex legal matters. Our flagship product, Catalyst CR, is an integrated, web-based review platform supporting the full range of cases for corporations and law firms. Its interface is modeled after Outlook, so review teams can get up to speed and be productive quickly.

Built on a grid of servers, Catalyst CR is designed to handle even the largest matters and allows your team to search, review and analyze documents in over 80 foreign languages with speed, simplicity and security. This means users can work with multilingual documents without the need to set up separate repositories for each language, run parallel queries and reconcile the results. With Catalyst CR you can search across millions of documents in multiple languages with a single, combined query using Catalyst’s powerful search engine. Results are displayed in the original languages for both document text and document metadata. Catalyst CR supports all western languages, as well as symbolic languages like Chinese, Japanese, Korean and Thai.

Catalyst handles data coming from the EU in accordance with Safe Harbor Principles and is registered with the U.S. Department of Commerce as a member of the Safe Harbor network.

---

*Between 1993 and 2003, the number of foreign Am Law 200 lawyers exploded (2,061 to 9,969; +383.7 percent).*

- The Conglomerate
- Nov. 17, 2005