

IMAGINE IF you could work in the language of your choice, yet still retrieve conceptually-meaningful information *even when it's in another language.*

Content Analyst® software combines powerful machine learning with mathematically-based LSI algorithms that allow you to sort, classify, and retrieve information based on *concepts* with the same ease that most users associate with keyword searches.

Because it is mathematic in nature, Content Analyst is naturally multi-lingual: searches can be done in any given language, not only English, against a matching volume of data. What surprises many people not familiar with our technology is that we are also *cross-lingual*: Content Analyst can perform searches against a data volume that contains *many languages* and return relevant results *without prior translation* of the documents in the query text.

How does this work? The answer is machine learning. The engineers at Content Analyst Company have already created a series of cross-lingual parallel corpora: these are “training sets” of identical documents, all carefully translated into the identified languages. For a cross-lingual installation of Content Analyst, you merely index the required parallel corpora as part of your data collection, and the installation will “learn” the relationship between those languages and apply that knowledge across your index. (Better yet, information from those parallel corpora never “pollutes” your search results – they only serve to train Content Analyst to “think” cross-lingually.)



Three Parallel Corpora for Worldwide Requirements

Content Analyst can provide three separate parallel corpora, based on business-level documents and providing “out-of-the-box” cross-lingual functionality. You simply license the desired corpora and install it as part of your data set.

UN: Contains roughly 3,000 documents, based on general UN discussions including a wide range of topics such as global warming and terrorism. Supported Languages in this set: Arabic, English, French, Russian, Spanish, and Chinese.

Euro: Contains roughly 2,500 documents, based on European Union parliamentary proceedings. Supported Languages in this set: Danish, German, Greek, English, Spanish, Finnish, French, Italian, Dutch, Portuguese, and Swedish.

Asian: Contains roughly 1,000 documents, compiled from web news articles. Supported Languages in this set: English, Japanese, Korean, and Chinese.

Note that these are generalized corpora; for very specific business functions, it may be necessary to create your own corpora, which is easily done by simply identifying or translating a set of identical documents into the desired languages (all Unicode languages are supported).

Learn More: Call 888.349.9442/703.391.8700 or Email info@contentanalyst.com