

# Wrapping Your Head Around Multi-Language E-Discovery

Al-Karim Makhani, Vice President of Consulting and Aedan McCrossan, Account Executive



The rate information is generated continues to increase exponentially. By 2020, the digital universe will be 44 zettabytes large, meaning there will be “40 times more bytes than there are stars in the observable universe.” [i] Concurrently, traditional information borders disintegrate or cease to exist all together. The result? Foreign language permeates legal matters more than it ever has before. The ability to deal with the language components of Big Data issues is fast becoming a valued asset to law firms and corporations alike.

Technology is an increasingly important string to the lawyer’s bow. Collaboration with technologists is key to overcoming new challenges, developing bespoke workflows, and bedding in new technology. E-discovery itself has progressed at breakneck speed to incorporate AI into everyday review.

However, there is still a lag in knowledge and preparedness where e-discovery encompasses multilingual data.

Both translation and legal industries, like many others, are faced with an unavoidable reality: how effectively are they leveraging machine learning? While each industry appears progressive in silo, integrating each to the other could be the key to interrogating documents sourced from numerous jurisdictions in a variety of languages.

How do these issues manifest? What problems can they cause? Clients face challenges ranging from one to all components of the e-discovery process: forensic collection, processing, analytics, human translation, machine translation (MT), and/or managed review. Working on the language in isolation causes problems.



@TransPerfectLegalSolutions



@TLSEgal

A common and straightforward example is seeing work duplicated—translating numerous versions of a document in a data set, which are substantively the same but not technically duplicates. Done once could be a three-figure problem—done across a multi-TB universe, it can become a seven-figure problem. We will look at potential pitfalls along the e-discovery journey and explain how we would go about fixing them.

Firstly, to prepare budgets and resources for a case, how do the lawyers confirm there are multiple languages in the data set, and in what proportion? This can be key to building the best teams across the right markets. Advanced language detection ensures this step is completed as accurately as possible. Whether it means four languages or twenty, or understanding that most documents are in a mixture of two languages and bilingual review is a necessity.

Next, for scanned documents it's important that the text is searchable. Most workflows include optical character recognition (OCR). Where foreign language is involved, it makes an already imperfect process much more risky. Unless the OCR tool can automatically detect the language, it will likely deploy the wrong character set. For example, Russian, Arabic, and Chinese—if processed with an English OCR tool—will come out as gibberish. Search terms may not be effectively directed, and important documents can be missed.

Once the data is rendered searchable it can be properly analyzed and organized. At this phase, we can gather a doc-by-doc breakdown of the dominant languages. Let's take a recent Arabic matter as an example.

The surprise that 20% of email data is Arabic can flag a few points, such as the possibility that all email disclaimers are in Arabic. On investigation, if that is the case, then this can be reliably discounted from the review or translation pool.

If there is relevant Arabic data, we will develop Arabic search terms. Even where a firm has language resources, search terms can be complex. Add in mixed, multiple languages and the syntax to go with them, and there is huge room for error. One of the trickiest aspects of multi-language e-discovery is ensuring that “proximity searching” is translated correctly, considering the language expansion or contraction rates after translation. For example, Italian to English translation will contract by 15%. Practically, this means that if you were searching for “ball” within two words of “bat” in English, the same search in Italian would have to be “mazza” within three words of “palla” to gather the same responses.

The multi-language search terms have now returned “X” number of Arabic and English hits to review. What becomes of analytics in non-English data? It's true that most tools have been developed for English document sets. Since analytics have the potential to transform the scope of a review exercise, what are the ways to leverage them? Can the machine reliably extrapolate tags across multiple languages? The short answer is “yes.” The longer answer is “yes but only with some expertise on the back end”. In regard to TAR 1.0, the question is historically whether to bifurcate by language or train the machine on a mixed-language set. Fortunately for newer CAL-based models such as Brainspace, we have the ability to deploy “language agnostic” analytics from the start.

The technology learns and improves by understanding concepts, not the order of letters and words created in distinct languages. So, for the relevant English content, Brainspace analytics can recognize the same concept in Arabic.

Lastly, what to do with the documents that still remain in a different language? Human translate them all for review? Have multilingual contract lawyers review them all? Turn to an online text translation tool? No. Technology exists to leverage MT within review platforms. During the MT phase, the content being translated is as secure as the rest of the data, as it's stored and processed in the same data centers. The translated documents also maintain the same confidentiality and privilege as the rest of the documents involved in the dispute. TLS's AI-based MT is integrated directly into our instance of Relativity. The MT technology can translate in bulk or document-by-document, detecting the language as it goes. Once a responsive document or set of documents looks like it requires a thorough human translation, another click of a button within the review portal prompts an expert linguist to translate the document and re-upload it—all behind the scenes without wasted time, cost, or further inaccuracy.

Technology and language expertise combine to limit document review as far as “machinely” possible. When handling multiple languages in the review pool, there are a few strategies to further limit review time. The case study we've touched on above also involved English and Russian. The budget was restrictive, so we were challenged to build an alternative workflow.

We ran our advanced language detection and linguistic OCR on the 40,000 mixed-language documents, and then spent ten hours consulting to build idiomatically and syntactically equivalent search terms in English, Russian, and Arabic to mirror the structure of a Redfern schedule. The responses to the terms reduced the original set by 70%, leaving 12,000 documents for first-level review. English, second-level review included MT and flagged 250 documents for human translation. The adapted workflow reduced the overall review cost from millions to hundreds of thousands of dollars.

There is a whole host of variables presented by multi-language matters, demanding a more sophisticated approach beyond de-duplication, translation, and linear review. It is no longer enough to present a stock technology solution. When our clients are confronted by multi-language e-discovery, we advise them to consider these challenges in conjunction, not in isolation, and partner with experts that have the knowledge and tools to make sure it's managed effectively, efficiently, and accurately.