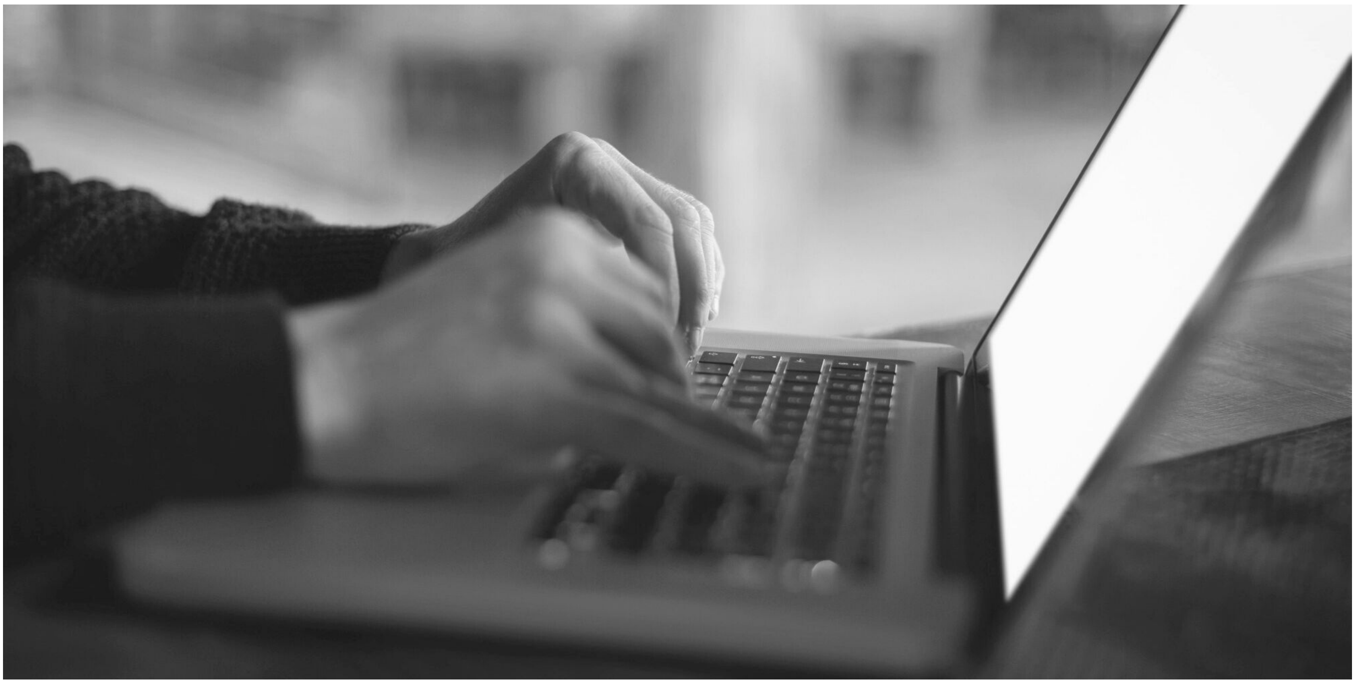
Beratung

Why can’t you simply search for French tea? Language challenges in eDiscovery

27. März 2023, von Irène Wilson



Why searching isn’t as simple as selecting the right search terms

On eDiscovery projects a lot of thought goes into selecting the right search terms, but what about all the other factors that need to be considered to get accurate results? In this article we will discuss some of these factors. With topics ranging from foreign accents to noise words, let’s dive into the conversation.

Dialling in the settings for each eDiscovery step

Data preparation

Most eDiscovery tools can handle common multilingual data sets without any major issues. Languages with different alphabets or reading directions however can present greater challenges.

Optical character recognition (OCR) is usually the biggest language challenge with data preparation. Having the right language selected for OCR increases the quality of the text recognised in your images, as the system knows which characters and accents to expect.

> “Having the right language selected for OCR increases the quality of the text recognised in your images, as the system knows which characters and accents to expect.”

This causes an automation issue, as a program can’t be used to identify a document’s language before OCR is performed due to it requiring the OCR text as an input. This Catch-22 situation can only be resolved by performing an initial manual review for language identification. Selecting too many languages for the OCR process, especially ones with different alphabets, will strongly decrease the quality of the results. Furthermore, we regularly see conversations where there are mixed languages, especially when messages are forwarded and new participants join conversations.

Once text is extracted from documents or identified in processed images via OCR, most tools provide a language identification feature. This is typically based on a statistical approach, where alphabet and frequency of groups of letters are used to make a decision. Depending on the tool, the decision might be binary, assigning one language to each document, or more nuanced, providing a list of likely languages and related percentages. That second option allows to manage the impact of forms or disclaimers, and comes in useful in multilingual conversations as mentioned before. Quite often, conversations remains primarily in the most common language of your data set, with just a couple of messages in a foreign language. Knowing the finer details of the potential languages within such conversations will allow you to assign them to someone who is more likely to understand the full content. Accurate identification of a document’s language is key for a swift and efficient review.

Data reduction

But before getting into a review phase, the data is usually reduced further via a variety of techniques:

Email threading

Repetitive content is usually excluded via technical approaches such as deduplication. While languages don’t impact a file’s hash value (and therefore hash-based deduplication), email threading is often used to bypass limitations of hashes and only keep the most inclusive email in a conversation. This feature sounds intuitive but is not language-agnostic.

Amongst other data, email threading recognises email headers available in the conversation history within a message. The fields within those headers, usually From, To, Cc, Date and Subject, are adjusted according to the user’s mailbox language. eDiscovery tools often support a limited list of languages for email threading, which might provide reduced quality depending on the languages present in your data set. For example, Italian is an official Swiss language that is not currently listed among the supported languages for email threading in Relativity.

Keywords

Most projects use keywords to identify potentially relevant material and decrease the number of documents to review. To do so in a multilingual data set you will need to not only translate your keywords, but also be aware of the different languages’ nuances, and consider synonyms, plurals, conjugated verbs, etc. For example, adjectives don’t vary at all in English. In French, an adjective varies depending on plural and word gender, while in German, even the word position in the sentence impacts the ending of the adjective. Wildcards are usually a good approach to account for all variations, but used incorrectly they can also lead to a massive amount of false positive and drastically affect performance. In an article published last year (i), Robert Wagner shines some light on how search syntax also requires adjustment in languages with tokenisation that differs from English, such as Japanese. It is always recommended to have someone familiar with the language review your keywords, and not just trust translation tools.

> “It is always recommended to have someone familiar with the language review your keywords, and not just trust translation tools.”

Analytical indexes

Several tools on the market offer advanced analytical features as an alternative to keywords, or as a complementary approach to increase speed and accuracy. Textual duplicates, concepts, and technology-assisted review are amongst the most common. These features are usually based on an analytical index, which groups documents based on key words and concepts, and are therefore indirectly language-sensitive.

If you don’t separate your documents per language before creating an analytical index it might misread languages for key words or concepts, and group documents based on recurring language-specific words. To make sure such recurring words (often called noise words) are ignored, they need to be excluded from the analytical index. Noise words are language specific, for example “the” is a noise word in English but means the drink “tea” in French. Excluding “the” for all documents, no matter their language, could seriously decrease the quality of the analytical index.

> “Noise words are language specific, for example “the” is a noise word in English but means the drink “tea” in French. Excluding “the” for all documents, no matter their language, could seriously decrease the quality of the analytical index.”

Ideally, the approach should differentiate between documents grouped by language. Multilingual conversations however will likely offer limited quality with such features.

Review

Once the data set for review is clearly identified, the strategy taken needs to account for the languages required.

Does your review team have all the language skills required? If so, then batching per language with a “Foreign language” option in the coding panel to allow reassigning documents is the most efficient approach. Automatic language identification has its limits. It is always better to account for software inaccuracy and multilingual conversations in your review workflow, and provide a way to reassign documents when they require different language skills than expected.

> “It is always better to account for software inaccuracy and multilingual conversations in your review workflow, and provide a way to reassign documents when they require different language skills than expected.”

If some language skills are lacking in your team, you might want to consider translation. Automated software translation doesn’t guarantee accuracy, but it is usually sufficient to assess the importance of a document.

Human translation is often more expensive and slower than automated approaches, so it might be more efficient to have a first review and decrease the number of documents requiring such accurate translation.

Production

Most production standards already account for foreign languages in their encoding requirements. Some courts or parties will ask to receive translations of foreign documents, in which case a certified translation is required. We recommend to clearly differentiate original documents from translations, and consider carefully which metadata to produce and in what languages.

> “We recommend to clearly differentiate original documents from translations, and consider carefully which metadata to produce and in what languages.”

Does metadata such as email subject also require translation? Should a hash value match the original or the translated version? Metadata such as dates and author can also bring their share of confusion. While you have to conform to what is requested, it is always good practice to be proactive in avoiding any ambiguity.

Conclusion

Multilingual data sets are actually extremely common, and need to be handled accordingly. Not only will you face them in multilingual countries such as Switzerland or Singapore, but any international company owns a multilingual data set. This brings additional challenges throughout the eDiscovery workflow, which are only tackled efficiently if you proactively make the required adjustments.

Discovering during the review that there are other languages involved, and that you haven’t adjusted your keywords accordingly, can be a major setback and lead to missing the court deadline.

> “Discovering during the review that there are other languages involved, and that you haven’t adjusted your keywords accordingly, can be a major setback and lead to missing the court deadline.”

Swiss FTS don’t just provide technical expertise to prepare the data, we also advise our clients on the full eDiscovery process by leveraging our experience and what we’ve learned from all of our international and multilingual projects. eDiscovery projects is our bread and butter, and using a collaborative approach we can help you navigate the complexities your multilingual projects effectively and efficiently. Don’t hesitate to reach out to learn more about our values and services.

References

(i) Robert Wagner, 2022, “Finding And Understanding Multilingual Documents: Where Technical Nuance & Language Converge”, https://www.jdsupra.com

Irène Wilson

Irene Wilson ist auf digitale Forensik und eDiscovery spezialisiert und hat im Laufe ihrer langjährigen Erfahrung für Kunden aus einer Vielzahl von Branchen in ganz Europa gearbeitet. Zu ihren Qualifikationen gehören die renommierten Master-Titel für Nuix Workstation und Nuix Discover.