Challenges in Searching Multilingual Data Sets in Global Investigations

by Ben Rusch

Consilio®
Together. Stronger.

Challenges in Searching Multilingual Data Sets in Global Investigations

This note is addressed to in-house legal counsel and private practice lawyers who perform cross-border investigations work. The purpose of the note is to highlight some of the challenges associated with searching multilingual data for eDiscovery purposes and to outline strategies for searching multilingual data. The focus here are those challenges that only arise because of the multilingual nature of the documents under investigation, not challenges of searching data in general.

Introduction

Global businesses speak many languages. While conducting business in multiple languages is necessary and inevitable, it also creates grave business risks as some of the most egregious and concerning compliance issues become much harder to detect. Sound risk management requires an effective approach to searching and reviewing multilingual data sets. Multilingual reviews can be complex and consequently, it is important to get multilingual searches and related workflow choices right.

How Do You Detect Compliance Issues in Your Business When the Employees Are Located Abroad?

Violations of FCPA principles, infringements of antitrust rules or any other conduct amounting to white collar crime can be difficult to detect. Since the advent of systematic compliance training, employees rarely do their compliance departments or outside counsel the favor of helpfully prefacing concerning emails with phrases such as "for your eyes only" or "don't tell Compliance." Disregarding the rules of engagement further, an attempt to establish an anticompetitive cartel is rarely referred to as a "cartel" in work emails. Keyword searches are therefore an inherently imperfect tool to detect concerning behavior.

The poor efficacy of keyword searches is exacerbated by the number of languages spoken in global businesses, and around 95% of the world's population does not speak English natively. It (nearly) goes without saying that the 95% of employees around the world do not suddenly switch to English when plotting the next cartel or FCPA violation. Instead, to the extent that actions that raise compliance issues are recorded in writing at all, the written evidence is much more likely to be indirect, cautious and at pains to avoid keywords, especially English keywords.

Accordingly, when the records of custodians are searched, the considerable thought that goes into English search terms, including the occasionally protracted process of iteratively refining them, will need to be replicated when crafting and developing search terms in other languages.

Translating and Transposing Multilingual Keyword Searches

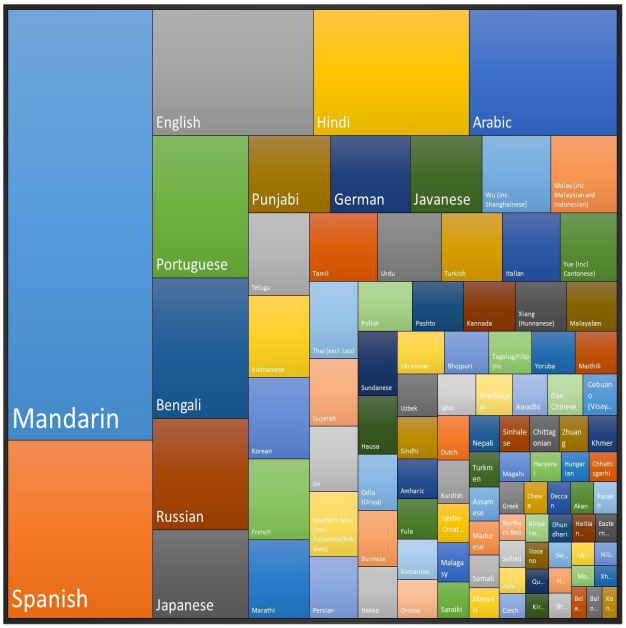
All other things being equal a literal translation of keywords is ineffective. Challenges include the following:

1) Hypocorisms

Native English speakers tend to underestimate the creativity that goes into nicknames elsewhere, and a search for a first

name would need to reflect nicknames that are in common usage.

To illustrate the point, when searching for names, we might reflexively search for "Jon OR Jonathan" or "Tim OR Timothy" when searching English-language correspondence for names. Some languages take the range of nicknames to extremes. In Russian, for instance, the names Dima, Dimka, Dimochka, Dimulia, Dimon, Dimych, Mitya, Miten'ka, Mitënka, Mityusha and Mit'ka are all nicknames for the well-known first name Dmitry (the question of spelling is addressed further below), while in Serbian, Miloš, Mihailo, Mihajlo, Miljan, Milovan, Miodrag, Milivoje, Milorad, Milutin, Milun, Milenko, Milojko, Milisav, Milomir, Miomir, Milić, Miško, Miša, Mišo, Šomi, Mićko, Mića, Mićo, Ćomi, Mile, Milo, Mija and Mijo are all nicknames of the widespread first name Milan.



Languages by number of native speakers[1]

2) Transliteration of first names and place names into English

Speakers of languages with a non-Latin script may transliterate place names and first names in different ways. When searching English documents, lawyers should therefore be conscious of different ways to render, in English, the spelling of names where the original language uses non-Latin script.

As an example, the common Greek first name Θάλεια can be transliterated as "Thaleia" or "Thalia". Greek employees writing in English might use both spellings to refer to the same individual. Russian-speaking employees may transliterate the Russian first name Dmitry (original Russian spelling: Дми́трий) as Dmitrii, Dimitry or Dimitri.

3) Transliteration of general text into English – informal language and diacritics

Especially in informal communications, speakers of languages with a non-Latin script may use Latin script to communicate. This is because their native language keyboard may not be installed at work, or it might be easier not to switch between scripts. In many languages, conventions have developed as to the most sensible way to render non-Latin letters using Latin script.

Most search algorithms give users the option of including diacritics along with the underlying letter so that a German-language search for "Ohr" will return "Ohr" (ear) as well as "Öhr" (the eye of a needle); a French-language search for "la" will return both "la" (feminine definite article) and "là" (there).

However, there are limits to what a search engine can recognize as interchangeable. For instance, the Russian letter "И" could be rendered in English as "i" or as "y." The Greek letter "θ" could be rendered as "th" or as the number 8 because of its visual similarity to the Greek letter. While the latter rendering is rather informal (also pejoratively referred to as "Greeklish" spelling), keywords will need to reflect different ways to spell foreign language text, even when using Latin script. In the Greek example, the phrase "I will be" would need to be transposed as both "tha eimai" and as "8a eimai" to ensure both transliteration conventions are captured.

[1] Mikael Parkvall, "Världens 100 största språk 2007" (The World's 100 Largest Languages in 2007), in Nationalencyklopedin (https://en.m.wikipedia.org/wiki/Nationalencyklopedin)

4) Transliteration of English words into another language

This section describes the reverse scenario of searching for English keywords that are customarily transposed into another language. Languages using Latin script normally render English expression 1:1 in that the German expression for a "Compliance Department," for instance, is in fact "Compliance Abteilung" (or even Compliance Department as a synonymous expression to search for).

Languages using non-Latin script need to find ways to spell loan words—"imported" words for which there is no equivalent in the foreign language. The katakana script— one of the three Japanese scripts alongside hiragana and kanji—is used to spell words of non-Japanese origin. While there are established conventions in respect of some words, there are multiple permissible spellings in respect of other words. To illustrate, if one was to search for the Japanese translation of the keyword "vendor" one would need to search for the string ベンダー (the letters form the sounds "benda") as well as the different spelling of ヴェンダー (a newer spelling with a "vee"-sound affixed).

It is important to stress, however, that not every English word will even be transliterated by speakers of the investigation language. Common names or place names may be left in their more common Latinized form, especially if the translation is not in common use. For example, it might make sense to search, when investigating Greek-language text, for the place name "Brussels" as well as the official translation "Βρυξέλλες," even though the official translation is actually in common use. The same holds true for other languages in situations where English terms such as company names may not be as well known in their local transliteration. For example, in Japanese text, the company name "Consilio" may be rendered in the official spelling "コンシリオ" or the original Latin script spelling as "Consilio."

5) Grammatical inflections and the limits of stemming and lemmatization – conjugation

One of the things that makes English data sets comparatively easy to search is that English barely inflects through grammatical cases and tenses.

To illustrate what happens on the conjugation side: the search term "conceal*" will catch the first person singular present "I conceal" as well as the third person plural future "they will conceal" and the second person singular past perfect "you had concealed" and so on. This simple, manual stemming-type search would be perfectly serviceable in English but fail even in mainstream European languages.

Contrast this with Romance languages such as Italian. The Italian infinitive for "to conceal" is "celare." The first person singular present would be "io celo" while the third person plural future would be "essi celeranno." A search for "cel*" would be too broad as it would catch entirely different words (e.g., "celebrare", the Italian verb for "celebrate"). Searching for "cela*" would be too narrow as it would miss many inflections— e.g., "io celo" (I conceal) and "noi celiamo" (we conceal).

One of the ways of catching all grammatical inflections of the Italian verb without drawing in too many different words would be the search string "celav* OR celi* OR celat* OR celas* OR celer* OR celo OR cela OR celano." At that point, stemming technology would already have reached its limit, unless we permit a dictionary lookup. Not to mention that we would of course still need to add the synonyms nascondere, occultare, mascherare and velare to match similar synonyms in English (hide, conceal, mask, obscure, disguise, cover up, withhold, etc.).

A quick word on lemmatization. Lemmatization technology will normally be able to recognize, by way of a dictionary lookup, the different inflections of a word where stemming

functionality fails. However, the maturity of both stemming and lemmatization technology is likely much higher for English than it is for exotic languages. Neither technology therefore obviates the need for cautious crafting of search terms to take account of grammatical inflections.

6) Grammatical inflections – declensions

As with conjugations, the declension of nouns and adjectives in languages other than English presents some practical challenges. In English, nouns only inflect to reflect the grammatical number (one garden, two gardens) but not to indicate case. Whether a noun is used in the nominative, genitive, dative or accusative it always looks identical. The same applies to gender in that English has no distinction between masculine, feminine or neuter gender.

To illustrate, juxtaposing only nominative and genitive:

	NOMINATIVE	GENITIVE
English	...the new house is large...	...the roof of the new house...
German	...das neue Haus ist groß...	...das Dach des neuen Hauses...

In the example above, the article, adjective and noun in "the new house" each inflect through their grammatical case; and an English translation will need to take this into account.

Notice how all three German words in the two tables that follow—articles, adjectives and nouns (translation: "the new house/tree/number")—inflect through the different cases and genders.

	Singular	Plural
NEUTER		
Nominative	Das neue Haus	Die neuen Häuser
Genitive	Des neuen Hauses	Der neuen Häuser
Dative	Dem neuen Haus	Den neuen Häusern
Accusative	Das neue Haus	Die neuen Häuser

	Singular	Plural
MASCULINE		
Nominative	Der neue Baum	Die neuen Bäume
Genitive	Des neuen Baums	Der neuen Bäume
Dative	Dem neuen Baum	Den neuen Bäumen
Accusative	Den neuen Baum	Die neuen Bäume
FEMININE		
Nominative	Die neue Zahl	Die neuen Zahlen
Genitive	Der neuen Zahl	Der neuen Zahlen
Dative	Der neuen Zahl	Den neuen Zahlen
Accusative	Die neue Zahl	Die neuen Zahlen

While one would not ordinarily search for articles ("the"), it is worth bearing in mind that in many languages other than English they normally inflect.

The same applies to adjectives which are more frequently used as search terms. The direct translation of "new," i.e., "neu" would in fact only identify the adverb. To capture the adjective as well, the translation would need to include the words neue, neues and neuen.

Most importantly, consider the inflection of nouns. To capture the German word for "house," one would need to search for the words Haus, Hauses, Häuser and Häusern. The translation will therefore need to reflect that.

7) Synonyms and local culture

Most lawyers investigating data would intuitively use appropriate synonyms for nouns and verbs. For example, if a suggestion that someone received a bribe needs to be investigated, one might look for the keyword "bribe" and synonyms such as "corruption," "enticement," "incentive," "gift," "inducement," "lure," "bait," "sweetener" and "kickback."

When translating search terms, the translator will also need to be instructed to add similar concepts in the target language which may not strictly be synonyms. For example, in several

Southeast Asian countries, Chinese mooncake is given as a gift on formal ceremonies and as corporate hospitalities. The translation of the word "mooncake" should therefore be added to the translation if one was investigating, say, a Singaporean custodian since the mooncake could be a euphemism for an inappropriate gift.

8) Linguistic variety

Especially for translations into languages other than European languages one should consider whether, taking account of regional varieties of the target language, it makes sense to include in the translation additional vocabulary and expressions or alternative spellings. For example, when translating into Arabic, it makes sense to check first if the search terms should be translated into Modern Standard Arabic or whether regional varieties come into play such as Western (Maghrebi), Central (e.g., Egyptian, Sudanese), Northern (e.g,. Levantine, Iraqi), or Southern Arabic (e.g,. Gulf, Hejazi).

Machine Translations: When Is It Better to Think Inside the Box?

1) Machine translations and workflow

Budget pressures and time constraints sometimes make machine translations an appealing option. The thinking is that a cursory glance at the machine translation will enable the monolingual English reviewer to determine whether a given record might contain relevant information and as such merit further review. The further review would then be conducted by a native speaker.

As the quality of machine translations is generally poor, however, this is quite a risky practice when it comes to internal investigations. This is because machine translations are not sufficiently sophisticated to detect the sort of nuance employed by those wishing to disguise egregious conduct in an otherwise innocuous document. In workflow and risk management terms, it is always preferable to have native speakers review the original document.

2) Machine translations and search terms

Because of the quality issues associated with machine translations and because machine translations do not supply synonyms, search terms should always be transposed into the foreign language and run over the original text.

What Next?

To minimize the risk of regulatory sanctions, global businesses need a plan for how to search and review documents in different languages. If the search methodology is flawed, then lawyers will risk missing critically important documents, missing documents that the investigating regulator found and drawing the wrong conclusions from their review exercise. Culturally aware multilingual consultants can help minimize these risks and help iteratively refine keyword searches until they are as good as they can be to keep businesses and clients' businesses safe.

About the Author

Ben Rusch is a Vice President at Consilio and a U.K.-qualified Solicitor. He has set up review projects in Japan, Luxembourg, Germany, Serbia, Austria, Belgium, France and the U.K., involving review teams that natively spoke English, Japanese, Serbo-Croatian, German, Dutch, French, Mandarin, Italian, Russian, Spanish, Catalan, Kazakh, Thai, Uzbek, Turkish, Kyrgyz, etc. Ben speaks German, English, French and basic Greek.