Tactical Linguistics: Language Analysis in Cyber Threat Intelligence

Jason Spataro

Tactical Linguistics: Language Analysis in Cyber Threat Intelligence

Author: Jason Spataro

Abstract

The capability to effectively collect and analyze data in strategic foreign languages when intelligence requirements are supported by it is a defining characteristic in a mature Cyber Threat Intelligence (CTI) program. Far beyond its use in attribution, language analysis can be leveraged to approach collection sources from a new perspective. This research seeks to provide a blueprint of those perspectives, as well as a set of critical considerations for those seeking to add or advance language analysis capabilities within their own CTI environments.

Introduction

Imagine a suspicious email is forwarded to a security operations center (SOC). The email and its attached Microsoft Word document are filled with unfamiliar, conjoined letters, and dots are sprinkled across portions of the text. Incident responders tentatively state that the language is Farsi. Later, corresponding Cyber Threat Intelligence (CTI) analysts do not have foreign language proficiency, but use Google Translate as a triage tool and discover that the language is Arabic. The resulting English language translation is muddled, but comprehensible. Nothing about the English-translated text is noteworthy, but the attached Word file is confirmed to be malicious. The attack is ultimately prevented, and leadership considers the situation to be resolved, but a set of deep, underlying questions still remain unanswered.

What did the Arabic in the email say? CTI analysts were provided an English translation by Google Translate, but that does not answer the question. Indistinguishable by the Google Translate application, the original Arabic email body contains contextual details that paint a much more vivid picture of the adversary and their operation.

Was the email written in a formal or informal voice? Foreign languages may separate specific vocabulary and even grammar for exclusive use with different persons, determined by cultural norms or an overarching social hierarchy. Depending on the language, formality may be reserved for people in positions of power, and this can indicate varying degrees of familiarity, respect, and even intention [11].

Does the text in the email contain any homonyms; words with more than one distinct meaning? Similarly, do any of the words carry different meanings in Arabic slang? One key indicator leading to the attribution of APT1 in by Mandiant Intelligence Center was the use of an obscure Chinese slang word referring to infected computers, which had been referenced in malware samples retrieved from compromised systems [8].

What dialect was the Arabic written in? This may be determined by only a few select words belonging to a certain region, or may not contain traces of a specific dialect at all.

Perhaps most importantly, does it appear that the Arabic was written by a native speaker? If it was, it could serve as a key indicator of attribution in future analysis. If it was not, that would call into question why the email was written in Arabic in the first place, who wrote it, who the intended target was, and what indicators within the text could be used to identify other writing by the same author.

The answer to that final question would be incredibly informative, but is entirely dependent on the previous answers to the questions before it. It also could be a complicated answer – Arabic is spoken by certain non-Arab populations across the Middle East. In Iran, the Arabic language is spoken by certain Arab minorities, and also taught across the Iranian public school system, despite not being the native or even secondary language of most Iranians. Discrepancies involving single words could indicate that the Arabic text in the email may not have been written by a native speaker. Even further, the presence of singular letters could suggest that it was written by a native Farsi, Urdu, or Kurdish speaker.

Without the capabilities of a linguist to bridge these gaps, however, the Arabic writing in the email remains wholly unremarkable.

Translation Tools Are Triage Tools

It may be tempting to think of an automated translation tool, such as Google Translate, as a solution to the language analysis needs of a CTI team. But language analysis is not the same as translation, and language itself is too complex to make accurate assessments of via automated tools. When analyzing a threat, translation tools may provide value in short bursts during triage – often to allow

analysts to understand what they are looking at – but fail to properly capture ambiguities or discern underlying meaning.

One example of this particular to CTI is the adopted use of the Latin alphabet by certain languages to create encoded or transliterated versions of themselves. Such is the case of Franco-Arabic, also known as the Arabic 'chat alphabet'. Franco-Arabic is a separate writing system frequently used by Arabic speakers for online communication. In it, the Arabic alphabet is substituted for a combination of Latin letters and Arabic numerals, creating an entirely new alphabet in which the language is written. For instance, Franco-Arabic would use the number three (3) to represent the Arabic letter *'ayn* (ع), and the letter *l* to represent the Arabic letter *lam* (ل). In context, this means that the response common Arabic greeting 'May peace be upon you', typically transliterated *wa alaykumu s-salam* (وَعَلَيْكُمُ ٱلسَّلَامُ), would be written in Franco-Arabic as *w 3laikom essalaam*. To further complicate this, different regional dialects of Arabic have carried over into Franco-Arabic, producing separate versions of the Franco-Arabic writing system based on the exact pronunciation of the speaker, each of which are separately applicable to the slang and colloquialisms found in different major dialects of Arabic [5].

The case of Franco-Arabic exemplifies two distinct problems with reliance on translation tools in the context of CTI operations. First, language evolves at a rate that translation tools may not be able to reliably keep up with, particularly in a digitally industrialized world where the evolution of a language takes place across multiple mediums at once. Even tools addressing the translation of Franco-Arabic may struggle to keep up with its evolution as a separate medium of the Arabic language, and that is to say nothing of the numerous other Romanized writing systems similarly adopted by other foreign languages for digital communication.

But perhaps more importantly, Romanized versions of foreign languages may be difficult to recognize as foreign languages at all. For example, in the 2013 Mandiant APT1 report, the malware string containing the Chinese word *ròujī* (肉鸡) was written within a larger Windows directory path as simply 'rouji', likely indistinguishable to those unfamiliar with the Chinese language [8].

This is a prime example of the Johari Window at work [6], more specifically the concept of 'unknown unknowns', where an analyst is met with threats that they neither know, nor are even aware of. By contrast, if an analyst were to see Chinese characters in a data set, they could at least

identify that a foreign language was present, even if they did not know what it meant – this would be a ‘known unknown’. But Romanized versions of foreign languages require linguistic expertise to identify, especially because they first have to be recognized as foreign languages, and then correctly identified as a specific language without the advantage of native characters or symbols.

All this in mind, the language analysis needs of CTI differ significantly from the translation services provided by automated tools. Foreign language encountered by CTI operations may not be immediately discernable as foreign language, and is frequently blurred by contextual ambiguities such as slang, style, or dialect. This problem transcends the capability of translation tools to solve, and requires hands-on analysis by an experienced linguist to accurately assess.

Incorporating Language Analysis into a Cyber Threat Intelligence Program

On strategic and operational levels, the development of language analysis capabilities within a CTI program will hinge upon the defining of overarching objectives, familiarization with proficiency frameworks, and the establishment of organizational standards for language proficiency. This may or may not be a straightforward process, and largely relies on strategic and operational levels of leadership agreeing upon similar language analysis end goals for a CTI program.

The following section outlines several key steps and best practices for the effectively developing these capabilities. These steps are enumerated by subsection, and are written in recommended order to carry out, but are not an exhaustive list of requirements. It should be underscored at all parts of the development process that language analysis objectives are highly unique to each organization, and must remain flexible and subjectable to change in congruence with a shifting threat landscape.

Determining Strategic Languages

The United States Department of Defense, as well as component military branches such as the United States Air Force, maintain their own Strategic Language Lists (SLLs) to document organizational needs for proficiency in key foreign languages. Each SLL is unique, and contains different languages at varying prioritization levels – classified as Immediate, Emerging, and Enduring – based on current and projected language analysis requirements. These requirements include the potential for surge, a rapid growth in capability from otherwise dormant assets, in response to an unexpected or short notice need [9].

What constitutes a strategic language is entirely dependent on intelligence requirements, and must be defined by an organization in accordance with their own unique threat landscape. They likely include languages used by adversary groups who have historically targeted an organization or its industry sector, and that have been confidently attributed to work in the interests of key nation states where the language is spoken. However, strategic languages should not be strictly defined as adversary languages. If an organization has overseas assets, previous intrusion attempts may have involved indicators or artifacts in the native languages of different host countries, which may require analysis despite not being considered adversary communications.

Documentation on strategic languages should prioritize immediate, potential, and long-term language analysis needs, estimate the minimum required and maximum desired number of linguists for each language, and, if possible, quantify a monetary value for strategic language proficiency. Maintaining an updated SLL at an operational level better ensures that language analysis plays an active role in the hiring process, allows organizations to more effectively identify and utilize existing language proficiency amongst employees, and narrows the otherwise overwhelming scope of potential languages by known priority.

Measuring Language Proficiency

Foreign language proficiency is a nebulous concept that is traditionally difficult to accurately capture, but nevertheless important to gauge when considering language analysis capability. Literacy is typically measured across the four major modes of communication: Reading, Writing, Listening, and Speaking; but additional context is needed to determine factors such as vocabulary range and sociocultural understanding [3]. Major frameworks for measuring language proficiency include the Interagency Language Roundtable (ILR), the Common European Framework for Reference of Languages (CEFRL), and the American Council on the Teaching of Foreign Languages (ACTFL) Guidelines. Additional language-specific frameworks include the Hanya Shiuping Kaoshi (HSK) and Test of Chinese As A Foreign Language (TOCFL) for the Chinese and Taiwanese Mandarin language, respectively, as well as the Test of Russian As A Foreign Language (TORFL) for the Russian language.

There are disagreements as to how exactly ILR, CEFRL and ACTFL scores correspond with one another, and additional disagreements on how the HSK exam corresponds to the CEFRL. The chart below is a general reference for language proficiency levels across the CEFRL, ACTFL, and ILR frameworks, published by the American University Center of Provence [1], with attached comparisons of the HSK, TOCFL, and TORFL. Other models may contain minor differences:

CEFRL	ACTFL	ILR	HSK	TOCFL	TORFL
A1	Novice (Low/Medium/High)	0 / 0+	1	1	TEL
A2	Intermediate (Low/Medium)	1	2	2	TBL
B1	Intermediate (High)	1+	3	3	TORFL-I
			4		
B2	Advanced (Low/Mid/High)	2	5	4	TORFL-II
		2+	6		
C1	Superior	3 / 3+	N/A	5	TORFL-III
C2	Distinguished	4 / 4+		6	TORFL-IV

Table 1: Comparison of CEFRL, ACTFL, and ILR frameworks alongside HSK, TOCFL and TORFL exams

In certain cases, frameworks may be used as references by significantly different exams, such as ILR scale being administrable by custom exams from both ALTA Language Services and the Defense Language Institute Foreign Language Center (DFIFLC).

Exam providers and academic bodies may deem scores from a language proficiency exam expired after a given period of time. It is advisable that CTI teams consider the validity of scores on a case-by-case basis, and keep in mind the length of time that an individual has maintained documented proficiency in a language, as well as the period and context in which they have leveraged their language skills in a professional environment.

The aforementioned frameworks all produce acceptable means of measuring language proficiency. However, CTI teams should carefully consider how to appropriately gauge proficiency as it pertains their operational and broader strategic goals. A CTI team concerned with activity occurring on social media or underground communities may be seeking colloquial language proficiency that standardized exams – the likes of which almost always measure proficiency based on formal modes of the language – do not accurately measure. In fact, some CTI teams may only really require that a linguist be capable of reading, and not concern themselves with listening, writing, or speaking proficiency at all!

In addition, CTI teams need to be cognizant of what challenges their strategic languages bring to language analysis. Certain languages, such as Arabic, differ so significantly in their spoken varieties that it is debated whether or not major dialects should be considered separate languages in their own right [12]. Languages such as Arabic and Farsi also have distinct differences in their written language from the colloquial, spoken language used in everyday life [10]. These differences may affect the capability of a linguist to conduct language analysis in the context of specific intelligence requirements.

Defining 'Linguist'

What constitutes a linguist? The answer is will vary by organization, but is necessary to standardize in order to establish clear-cut language analysis capabilities. When determining a standard for linguists, organizations should be mindful of using subjective wording like 'fluent' or 'near-native', and should instead rely on a combination of measured proficiency levels and contextual qualifications, such as the ability to conduct technical translation, knowledge of geopolitics, or familiarity with social media slang. Keep in mind that even a native speaker will not understand specialized terminology if they are not familiar with the underlying subject.

Organizations should always seek to establish the lowest standard required to satisfy language analysis needs. The lower a minimum language proficiency is, the more realistic it will be for an organization to find applicable linguists, and for those linguists to maintain the proficiency over time.

To do this, decision-makers need to be familiar with the descriptions of different proficiency levels and understanding what they functionally represent. The CEFRL contains multiple, detailed descriptions for each mode of language proficiency, according to setting. For instance, the CEFRL describes a C1 in Overall Reading Comprehension as follows: "Can understand in detail lengthy, complex texts, whether or not they relate to his/her own area of specialty, provided he/she can reread difficult sections" [3].

But this is then further broken into categories such as Reading Correspondence, Reading for Orientation, and Reading Instructions. So, an organization especially interested in linguists able to read foreign technical documentation may decide to standardize a minimum reading proficiency based on the descriptions provided in a specific category. The B2 proficiency level of Reading

Instructions is explained as "Can understand lengthy complex instructions in his/her field, including details on conditions and warnings, provided he/she can reread difficult sections" [3].

If language analysis skills fitting this description would satisfy an organization's CTI mission, then the organization would benefit from keeping the mandatory minimum reading proficiency to the B2 level, where more individuals will be able to meet and maintain the standard.

Organizations should also consider how experience may be substituted for documented proficiency scores. In certain cases, language proficiency may have been acquired in a household setting, studied in an academic environment, or studied independently, without ever having been formally documented or assessed. If documented language proficiency scores are imperative, an organization may consider sponsoring proficiency exams. Otherwise, additional methods may have to be established to determine the validity and acceptability of a candidate's stated language proficiency.

Understanding Proficiency vs. Capability

What would happen if a group of native English speakers – with no experience in information technology (IT) or security – were asked to explain how Domain Name System (DNS) hijacking subverts the resolution of DNS queries? Would they understand the question? If the answer were explained to them, would they 'get it' after hearing it for the first, second, or even third time?

This situation is important to think about when developing requirements for linguists within a CTI program. One notable mistake is to standardize the highest language proficiency level available, typically ILR 4 (CEFRL C2), and deem it a requirement to any linguist position. The native English speakers in the question set above would be considered an ILR 5 on that proficiency scale, just by virtue of English being their first language. But no ILR or CEFRL level could prepare them for that question, because language proficiency levels cannot describe the intricacies of language in information security any more than they could for other specialized subjects.

Proficiency scales, while important, do not speak to the full needs of a CTI program. They are based on standardized exams that may not gauge any of the specialized vocabulary or underlying context that a CTI team encounters. They are also frequently based on the standardized language, which does not necessarily resemble the way that the language is used in spoken or online contexts.

Although proficiency levels clearly distinguish capability in a foreign language, there is a certain point in which a higher level does not indicate that a linguist will produce greater language analysis capabilities for a CTI team. Organizations should determine what tactical-level situations they expect linguists to encounter based on projected intelligence requirements, and screen linguists to determine whether or not they are familiar enough with these situations to properly comprehend and analyze them.

Identifying Linguists

The larger an organization is, the more likely it already has employees proficient in a given language. Dependent on the positions of these employees, the language analysis needs of the organization, and the communicative abilities of CTI team, it may be advantageous for a language analysis 'asset list' to be established. In this list, any employees who meet the set standard for foreign language proficiency are recorded as 'assets' or 'linguists', and can be called upon for language analysis tasks if needed.

This approach will fail if linguists are unable to be called upon when needed, or if the list is not properly maintained to account for incoming and outgoing employees. For this reason, a successful asset list requires buy-in from the highest levels of organizational leadership, as well as internal coordination to keep updated.

When leveraged successfully, a language analysis asset list can allow organizations to circumvent resource constraints in hiring and retaining a language-enabled CTI team by utilizing a pre-existing group of employees. With this improved resource utilization, it may be possible for an organization to broaden its range of strategic languages. A language analysis asset list should be attached to its corresponding SLL, so that both documents can more readily be referenced, compared, and maintained.

Depending on the size, structure, and culture of the organization, coordinated asset management for language analysis taskings may not be feasible across all sections or offices. At the very least, management can consider implementing a language analysis asset list within the information security or information technology department, thereby better ensuring that employees deemed 'assets' will be in-scope for future language analysis taskings.

One inherent disadvantage of language analysis asset lists is that they require incident responders or CTI analysts to be aware of a foreign language, and potentially identify it, before calling for help from a linguist. Previously, this paper described the issue of 'unknown unknowns', where an analyst may encounter a foreign language and not recognize it as such, especially if the language was written using the Latin alphabet. This problem is not addressed by the use of asset lists, and instead requires that a linguist view the data set themselves to determine the presence of a foreign language. For this reason, it is particularly advantageous to have linguists in positions that respond to incidents directly, whether that be within a security operations center (SOC) as tiered analysts, or compartmentalized Digital Forensics & Incident Response (DFIR) functions such as CTI, threat hunting, and malware analysis.

Maintaining Linguists

Language proficiency requires attention to properly maintain, and an exceptional amount of attention to advance, particularly as proficiency rises. In a scale with only six categories to describe all degrees of language proficiency, the difference in capability between the B1, B2, and C1 proficiency levels is noteworthy. Organizational standards may clearly require one of these proficiency levels over another, and CTI teams may find that a B1 or even a B2 is insufficient to address specific language analysis needs, or to do so in a timely or accurate manner.

Because of this, it is imperative that an individual hired to advancing a language analysis capability be given sufficient, continuous language analysis tasks, in order to maintain language proficiency.

Language Analysis Tactics

Within a CTI environment, the objective of language analysis is to provide a more thorough understanding of indicators and to better satisfy new intelligence requirements. Exactly how linguists reach these objectives is determined by what language analysis tactics they are able to develop and effectively execute. Collection sources are constantly changing, and the availability of sources is typically expanding. For this reason, there will inherently be a some creativity involved in devising language analysis tactics as solutions to emerging requirements.

This question of tactics should be periodically revisited to address new intelligence requirements and situational developments. Determinations on what tactics will satisfy an organization's intelligence requirements should occur at the operational level. Linguists should be incorporated into this decision-making process, as they are likely to be very familiar with the ecosystem of available collection sources in their target language, and may be able to provide insight on the feasibility of certain tactics.

This section contains several example language analysis tactics. It is not an exhaustive list.

Vetting Threat Intelligence Sources

A commercial threat intelligence provider may share a tip on an actionable threat based on information gathered from a foreign language communication. This could mean an underground forum post revealing a vulnerability to public-facing web infrastructure, or discussing sensitive documents evident of data exfiltration. The victim organization may need to take immediate action, but how do the appropriately judge the source material without being able to analyze it themselves?

Organizational needs are complex, and typically involve many moving parts. Outside entities such as Information Systems Audit and Control Associations (ISACAs) and commercial threat intelligence providers may share threat intelligence based on their own analysis of foreign language communications, but it is the responsibility of each organization to analyze, assess, and validate what that intelligence means to them. That means the foreign language communications need to be analyzed by in-house linguists, professionals who are highly familiar with the needs of the organization as well as the target language.

It may turn out that the communication is a fragment of a much bigger picture – for example, usernames associated with the forum post could match persona artifacts found during an previous intrusions, unknown to any outside of the organization targeted. Pivoting on these personas could require reading additional parts of the forum, and only linguists with knowledge of the prior intrusion and the capability to access and analyze the course of the communication could have made the discovery.

Alternatively, the foreign language communication itself may contain information that the threat intelligence provider did not recognize was relevant, either due to lack of familiarity with the organization or ineffective language analysis. Were the organization unable to perform their own analysis with in-house linguists, the foreign language communication would have to be assessed at whatever level of detail it was reported as, and any unreported intelligence – perhaps quite valuable to the organization – would have been lost.

Vetting Media Reports

Geopolitical developments may directly or indirectly impact the interests of an organization, and would therefore be of interest to CTI teams. But how does an analyst conduct thorough assessments of an event without access to outside perspectives, other than those of English-language media sources? Similar to threat intelligence sources, media reports on any publicly accessible, foreign language data should be vetted by in-house linguists whenever possible.

Historically, mistranslations propagated by media reporting have had large-scale political consequences. In 1956, Soviet First Secretary Nikita Krushchev gave an address at the Polish embassy in Moscow, in which he stated the Russian phrase *My vas pokhoronim* (Мы вас похороним), translated by his interpreter as “We will bury you!” Time Magazine published this account, specifically using the translated phrase and noting that Khrushchev had threatened the West. The perceived threat would go on to be formally acknowledged and responded to by multiple US politicians, who took it as threat of annihilation [4].

The problem is that the phrase *My vas pokhoronim* (Мы вас похороним) was erroneously translated by Krushchev’s interpreter, and subsequently published by the Times to a national audience. Krushchev’s Russian idiom more closely resembles the phrase “We shall outlast you”

[7]. The difference between the two terms is dire; one denotes and impending threat, while the other can be interpreted as the boast of a rival nation state.

This particular incident carries three key takeaways. First, media reports of statements by foreign officials may not have been translated by in-house linguists. Second, organizations should be wary of trusting the standards of linguists outside of their organization, particularly when confronted with complex subjects such as the interpretation of political dialogue. Finally, the implications of media reporting are sometimes more noteworthy than initial subject of the reporting itself.

Meta-Analysis of Media Reports

Because media reporting itself is impactful, an analysis of foreign language media reporting may result in valuable conclusions. For example, in December 2019, Mohammad-Javad Azari Jahromi, serving as the Minister of Information and Communications Technology for Iranian President Hassan Rouhani's administration, released a tweet announcing that an intrusion attempt by APT27 against Iranian government electronic systems had been successfully defended against by 'Dejfa', Iran's national cyber defense system.

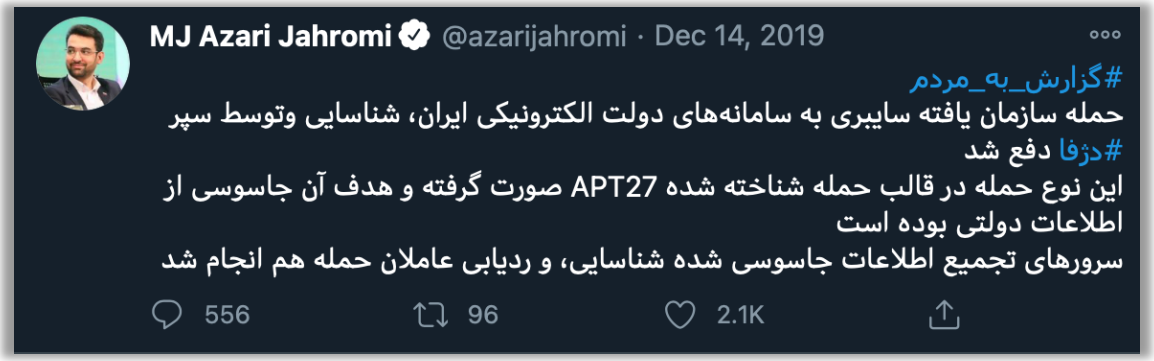

Figure 1: Tweet by Mohammad-Javad Azari Jahromi on Attack by APT27

In the final line, Jahromi states, among other things, that the perpetrators of the attack were tracked. Without any indicators of compromise (IOCs) or additional data sets, the purported intrusion is impossible to validate. But CTI analysts responding to intrusions by APT27 may be interested in the impact of the statement itself.

By publicly alleging to have detected an APT27 intrusion and initiated monitoring efforts against it, Jahromi could potentially scare APT27 operators into abandoning pieces of their active Command & Control (C2) infrastructure. This is a possible outcome whether or not the initial intrusion claims by Jahromi are actually true, as even false claims could potentially alert APT27 to recalibrate their operations. If APT27 were to abandon pieces of its C2 infrastructure as a response to this statement, it could lead to an intelligence loss for CTI teams responding to ongoing APT27 intrusions, who may have chosen to surreptitiously track corresponding C2 infrastructure in sight of long-term intelligence gain.

While likely not a reportable detail in and of itself, CTI analysts may also find it noteworthy that a key Iranian political figure, in a native-language communication, used a FireEye naming convention to refer to a major threat actor.

Multilingual Open-Source Intelligence (OSINT)

When a CTI team leverages OSINT against its own organization, it allows them to see through the lens of adversary reconnaissance, and what that reconnaissance could discover from publicly available information. This is already an advisable activity for CTI teams, but a multilingual team can perform a vastly different kind of OSINT collection. Linguists proficient in strategic or adversary languages can search for OSINT about their own organization in those languages.

This is a creative process as much as an analytical one, as there may be a multitude of words capable of referring an organization outside of its formal title. For example, a military unit could be referred to as simply ‘military’, by their service branch, or by any number of component organizations within its structure. Similarly, CTI teams working in the private sector may be interested in specifying the names of their company, products, and executives, as well as any key terminology relatable to their broader industry and the technology that it depends on.

CTI teams may be similarly interested in using multilingual OSINT gathering against known or suspected adversaries, by researching publicly available channels such as underground forums, chat platforms, and ‘dark web’ content. Here again, success often relies on specialized search terminology in the target language. Early research on the use of open-source communication channels by jihadist groups discovered that successful data collection relied on targeting clusters of specific Arabic words most associated with jihadist rhetoric [2]

Operational management should note that even linguists with near-native proficiency may require time to assemble and regularly update specialized vocabulary and research key sources before being capable of effectively performing multilingual OSINT operations.

Prioritizing Language Analysis Collection Sources

At its core, language analysis allows CTI teams to leverage intelligence sources that would otherwise be difficult or impossible to effectively utilize. The scope of these sources, however, is virtually infinite, and CTI teams have to prioritize which sources best satisfy their intelligence requirements, with the understanding that new intelligence requirements may require re-prioritization. The following is a non-exhaustive overview of several major types of collection sources.

Official State Communications

CTI teams interested in geopolitical developments may require official, attributable communications from a nation-state to produce high confidence assessments. Typically, these communications will be broadcast from formal channels, including major foreign media organizations. Communications may also be available via official social media channels, such as Twitter. Both official and unofficial sources may include dedicated channels for officials within different branches of government, which can help CTI teams focus on collection sources related to their topics of interest. Additionally, there may be officially, publicly available documentation on projected developments of a government, military, or economy, known as a National Strategic Plans.

Official Information Security Communications

Multilingual CTI teams also have access to publicly available threat intelligence and information security research produced by foreign entities. This includes reports and white papers by threat intelligence providers, which may provide secondary perspectives on emerging threats. Dependent on context, reports by vendors providing offensive security services could hint at tactics, techniques, and procedures (TTPs) being leveraged by adversaries. Similarly, security-focused technological research published by foreign universities might contain findings and context unavailable in English-language publications. Releases by foreign Computer Emergency Response Teams (CERTs) can be used to track activity groups internationally.

Parallel to the previous point on official state communications, announcements by politically appointed technology officials may provide insight on the security concerns being faced in strategic locations.

It is important to underscore that these examples are applicable not only to nations attributable to adversary activity, but in any region encompassing an organization's strategic interests.

Social Media, Chat Platforms, and Forums

Valuable threat intelligence can also be found on social media, chat platforms, and blogging platforms by foreign information security professionals and enthusiasts, who may write in their native language to report discoveries on malware or exploit activity. In certain circumstances, these digital platforms may host 'dissident' groups, who seek to expose the cyberwarfare activities of nation states through leaked documentation and insider reporting.

One of the more notable examples of this is Lab Dookhtegan (لب دوختگان), a purported Iranian dissident group operating on Telegram, whose activity has historically involved leaking data related to the active campaigns, tool samples, and actor identities of prominent Iranian threat groups.

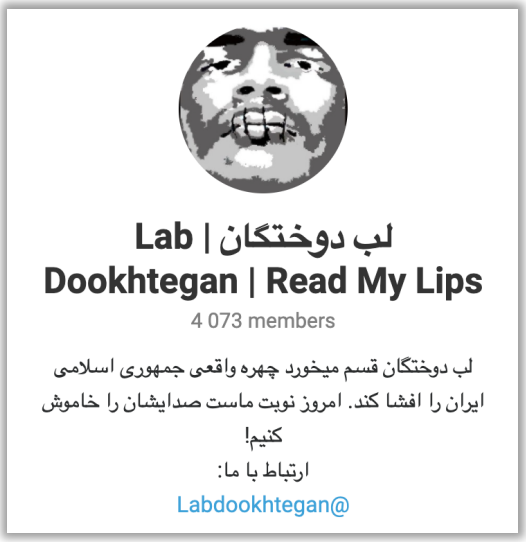

Figure 2: Telegram Channel for Purported Iranian Dissident Group Lab Dookhtegan

These sources may also be leveraged by threat actors to organize attacks, dump or sell successfully exfiltrated data, and collaborate building custom attack tools. Strategic and operational levels of leadership may be tempted to have CTI teams explore these channels to directly reach high fidelity discoveries, but should be aware that identifying valuable collection sources in this category is often a laborious process, and requires an intimate understanding of the ecology of an organization's threat landscape.

Identifying high-value collection sources across social media, chat platforms, and forums is challenging, because their locations are scattered, and their names are frequently unintuitive. For this reason, linguists seeking to enumerate collection sources across these categories should conduct searches based on the content they consider high-value. This can be as simple as a list of foreign language search terms for key collection targets, such as '*leaked documents*', or may involve a more dynamic list of search templates that can be configured to accommodate emerging intelligence requirements. Source enumeration should be leveraged on individual platforms as well as Google, which can be used to filter search results via Google Dorking techniques.

Conclusion

In the context of CTI, the obstacle of foreign language is arguably as much a means of obfuscation as an encoding or encryption algorithm. Intentionally or otherwise, it provides means to conceal communications, mask intent, and create barriers to understanding using underlying cultural context. Language analysis allows a CTI program to overcome this obstacle, and expand intelligence collection and analysis capabilities in support of different intelligence requirements.

At the operational level, proficiency requirements must be tailored to language analysis needs, to ensure that CTI teams have linguists that can accurately perform their analysis duties. It may be tempting to push these requirements to the furthest end of the proficiency scale as a perceived quality guarantee, but a careful reading of language proficiency scales and the addition of context-based qualifications is the only way to ensure that companies are searching for linguists that meet their mission standards. Higher proficiency does not necessarily indicate the ability to function in increasingly technical situations.

At the tactical level, linguists may be aware of use cases for language analysis that satisfy intelligence requirements in a way that others are not privy to. They are likely subject matter experts (SMEs) in the culture and region associated with their target language, and should be incorporated into determinations on how language analysis can satisfy intelligence requirements.

There does not have to be an emerging incident to justify the development of language analysis capabilities. In a digitally industrialized world, many adversaries speak and use foreign languages in the context of their offensive operations. Expanding CTI programs to meet strategic language needs before the capability is critically needed allows organizations more time to prepare the underlying foundations necessary for effective language analysis.

References

[1] "Understanding the TEF Results: Level Breakdown and Equivalencies," American University Center of Provence. 2014. [Online] Available: https://web.archive.org/web/20140116123131/http://www.aucp.org/sous_pages/aix/Students_Aix_docs/TEF_Levels_and_Equivalencies.pdf

[2] Chen, H., Chung, W., Qin, J., Reid, E., Sageman, M., & Weimann, G. "Uncovering the dark Web: A case study of Jihad on the Web. Journal Of The American Society For Information Science And Technology," 2018.

[3] "Common European Framework for Reference of Languages: Learning, Teaching, Assessment," Council of Europe. 2018. [Online] Available: https://rm.coe.int/cefr-companion-volume-with-new-descriptors-2018/1680787989

[4] "Khrushcev's 'We Will Bury you'," Central Intelligence Agency. 1956. [Online] Available: https://www.cia.gov/library/readingroom/docs/CIA-RDP73B00296R000200040087-1.pdf

[5] Danet, B., & Herring, S. "The Multilingual Internet," New York: Oxford University Press. 2007.

[6] Luft, J., & Ingham, H. "The Johari window, a graphic model of interpersonal awareness," Proceedings of the Western Training Laboratory in Group Development. Los Angeles: University of California, Los Angeles. 1955. [Online] Available: https://static1.squarespace.com/static/572d003b40261d2ef97e5b0b/t/5ca20f5d6e9a7f566fc3ef14/1554124637198/The-Johari-Window.pdf

[7] Polizzotti, Mark."Why Mistranslation Matters," New York Times. 2018. [Online] Available: https://www.nytimes.com/2018/07/28/opinion/sunday/why-mistranslation-matters.html

[8] Mandiant. "APT1: Exposing One of China's Cyber Espionage Units," FireEye. 2013. [Online] Available: https://www.fireeye.com/content/dam/fireeye-www/services/pdfs/mandiant-apt1-report.pdf

[9] Murphy, Keith. "Language Capability in the United States Air Force," Marine Corps University. 2009. [Online] Available: https://apps.dtic.mil/dtic/tr/fulltext/u2/a513814.pdf

[10] Negar Mohammadi. "Discourse Markers in Colloquial and Formal Persian" University of Florida. 2018. [Online] Available: https://ufdc.ufl.edu/UFE0052106/00001

[11] Sharifi, S., & Azadmanesh, M. "Persian Back Channel Responses in Formal versus Informal Contexts," Linguistic Discovery. 2012. [Online] Available:

https://journals.dartmouth.edu/cgi-bin/WebObjects/Journals.woa/1/xmlpage/1/article/401?htmlOnce=yes

[12] Zaidan and Callison-Burch. (2007). University of Pennsylvania. [Online] Available: https://www.cis.upenn.edu/~ccb/publications/arabic-dialect-id.pdf