A Linguistic Analysis Metric in Detecting Ransomware Cyber-attacks

Diana Florea[1]
Lucian Blaga University of Sibiu
10 Victoriei Bvd
Sibiu, 550024, Romania

Wayne Patterson[2]
Patterson and Associates
201 Massachusetts Ave NE, Suite 316
Washington, DC 20002 USA

***Abstract*—Originating and striking from anywhere, cyber-attacks have become ever more sophisticated in our modern society and users are forced to adopt increasingly good and vigilant practices to protect from them. Among these, ransomware remains a major cyber-attack whose major threat to end users (disrupted operations, restricted files, scrambled sensitive data, financial demands, etc.) does not particularly lie in number but in severity. In this study we explore the possibility of real-time detection of ransomware source through a linguistic analysis that examines machine translation relative to the Levenshtein Distance and may thereby provide important indications as to attacker's language of origin. Specifically, the aim of our research is to advance a metric to assist in determining whether an external ransom text is an indicator of either a human- or a machine-generated cyber-attack. Our proposed method works its argument on a set of Eastern European languages but is applicable to a large(r) range of languages and/or probabilistic patterns, being characterized by usage of limited resources and scalability properties.**

***Keywords*—Cyber-security; cyber-attacks; machine translation; language; Levenshtein distance**

I. Introduction

The recent COVID-19 pandemic has determined an upsurge of remote work that has increased both companies' and end users' exposure to various cyber-attacks. This has complicated an already existing landscape of risks associated with hacking and cyberattacks that the exponential advancement in technologies has brought about. Cyberattacks may be motivated by ideological, financial, or personal reasons and are directed at governments and institutions, businesses and private individuals engendering geopolitical, security, reputational and privacy concerns. While there is a wide literature on the typology, counter measures, policies and security information sharing across state and private sectors [1-8], for our practical purposes, we shall briefly refer in this section to ransomware and ways to address cybersecurity by means of linguistic approaches and instruments.

Ransomware represents a subset of cryptovirology malware that threatens to release and expose the victim's personal information or to permanently disable access to that data until a certain ransom is paid. Whereas some ransomware is designed to lock the system in such a way that it is easily reversible, more advanced malware employs techniques such as cryptoviral blackmail that encrypts the victim's data, rendering them unusable, and demands payment for their decryption [9]. Recovery of data without a decryption key is an uncontrollable problem in a cryptoviral extortion attack, one all the more difficult to trace as crypto currencies, such as Bitcoin, and Dark Web environments are used for completion of ransom payments.

What actually happens in the space between the human brain's complexity and the keyboard strokes on the computers' starred-out password field has been an object of constant inquiry for researchers coming from fields of cognitive sciences, including psychology, philosophy, logic, computer science, neuroscience, etc. In the world of cybersecurity, linguistics has also provided a wide array of approaches, methods and instruments to expose in particular the vulnerability of password creation by exploring various password strength metrics and creation strategies. Such approaches concern lexical patterns (word choices), structural preferences (in composition rules) and syntactic and semantic patterns (such as preference for semantic categories and/or their sequences). Thus, while areas concerning grammar and grammatical rules to crack passphrases [10], or general linguistic patterns in multi-word passphrase selection [11] have been investigated, other practical models and approaches, such as semantic segmentation frameworks of passwords based on Natural Language Processing algorithms [12], phrase generators for cracking pass-phrases [13], probabilistic context-free grammars [14] or predicting technologies [15, 16], represent as many functional models devised to assist in understanding password creation processes and ensuring users' protection in the cyberspace.

With ransomware, linguistic approaches have been either in the form of actual text-analyses (see [8]) or of ransomware detection devices and apps based on linguistic parsing [17]. Two types of linguistic analysis can be distinguished: one that examines the way the source code was written and another that examines the text that was used. While the former examines the code's style and compares it to other pieces of code discovered in malware samples, the latter is more concerned with the word choices made in user dialogues, code comments, input screens, and other user-visible displays. All ransomware includes ransom notes, however, unlike spam and phishing messages, where attackers must impersonate legitimate entities, ransom notes can conceal clues about the writer's proficiency in that language as well as his/her geographic location.Within efforts of linguists who have struggled with the question of attribution (2014 Sony breach, Coin Vault, Shadow Brokers and Guccifer 2.0), has been the infamous *WannaCry*

and *Petya* ransomware attacks, part of which a thorough linguistic and cultural review of ransom notes was conducted so as to determine the native tongue of the authors (Flashpoint 2017). The research found that almost all ransom notes were translated using Google Translate, three of which being likely to have been written by a human rather than machine translated. Further discovered within the same examination was the fact that machine translation into the other languages was performed by using the English note as the source text. Despite such spectacular results, the main attribute of such linguistic analyses as the above is that they try to shed light on the varying levels of language proficiency of attackers, which, in practice, can often obscure the very origin of a ransom attack. Thus, in order to mislead analysts, attackers frequently use red herrings, manipulate time stamps and/or deliberately implant false language clues and insert cultural references and phrases to instill confusion about either their backgrounds or their locations. This makes a linguist's effort a very complex yet an equally critical task. Nonetheless, linguistic examinations of ransom notes are all the more successful if they can be further combined with additional computer science evidence that points the way toward attribution.

This study is structured as follows: Section I includes some preliminary considerations concerning ransomware and a few linguistic approaches and instruments by means of which cybersecurity can be addressed. Section II describes the reason for using the English language (II-A) scope of our research relative to corpus (II-B) and the Levenshtein Distance metric (II-C) whereas Section III presents the research methodology, operations and emerging results. The final Section IV presents the conclusions and the implications of our study for further research in the field.

II. Scope of Research

Internet access facilitates global outreach which is why a cyber-attack, launched to potentially target any location on Earth, is unconstrained by geography and/or distance. A variable vulnerability may represent the language of the attacker (and/or that) in the ransom notes however recent cases of malicious ransomware indicate that cyber-attackers have been able to develop additional language functionalities, such as the ability to issue ransom demands in as many as 30 languages[1], to enable them to more easily target their cyber-victims worldwide.

In previous studies [7, 8], we have analyzed ransomware external messaging via a mechanism that has involved six extensively used languages (French, Spanish, German, Russian, Chinese, and Hindi) and several round-trip translation (RTT) operations from target language into English using the Google Translate (GT) functionality. Our analysis was then conducted on a number of random quotations from popular culture and English literature and that finally allowed us to devise a procedure which could establish whether a perceived attack was initiated by a human writer with some knowledge of English, or alternatively, by a machine translation. An index was further advanced to assist the cyber-defender in the profiling of potential human or machine cyber-attacks in which the attack message might have been originally written in a different language than English.

Starting from these assumptions and results, and looking to extend our analysis to other regional zones of interest, the scope of this research is to provide a methodology and a systematic metric to assist in detecting the possible origins (language/location-wise) of a remote cyber-attack potentially originating from the Southeast region of Europe. To conduct this analysis, we will use the GT functionality on a number of sampled texts analyzed in six Southeast European languages (Macedonian, Romanian, Albanian, Bulgarian, Greek and Serbian) through a RTT process (a back and forth translation hereafter referred to as ABA). Additionally, while the effectiveness of GT is under scrutiny as well, several comparisons between the earlier data [7] and the languages that now form the basis of our approach will afford a better grasp of the method we are proposing relative to its general use and scalability properties.

A. The Medium: The English Language Rationale

There are a variety of reasons why a cyber-attack launched from any location on the planet may include text or instructions to the target written in the English language. For one reason, cyber-attackers might perceive that potentially more lucrative targets might be easier to find in the English-speaking world. A second reason may well be that there is a far greater usage of the English language throughout the Internet, which from its very inception tended to be far easier to use than compared to any other language based for example on logograms (Chinese) or the Cyrillic writing system (Bulgarian, Macedonian, Serbian, Russian, etc.). Thirdly, as experience has often shown, non-English language speakers are very likely to use English in their Internet communication and resort to GT whenever cyber-attackers are not proficient in English.

To push the argument further, for any analysis of cyber-attacks involving exclusively the languages of the South Eastern European countries, it would be also reasonable to ask why any measurements of translation effectiveness should use English at all, since there are currently 24 official languages spoken within the Member States of the EU. Being the world's lingua franca, English is spoken by nearly 360 million native speakers worldwide, with slightly less than 60 million of them residing in Europe. It is the most spoken language in the EU (44%) and the most spoken second language by roughly half European language speakers within the 15- 35 age group who can communicate in English. More recent studies [18] assume that English still remains the EU's most spoken language post-Brexit and that the English figure is in reality much higher as English proficiency has recently increased rapidly among young people across the continent.

B. The Corpus: Southeastern European Languages Medium

The Southeastern European countries represent particular zones of interest for cybersecurity issues due to their increasing reputation on digital skills and education, internet usage (Table I), strengthened national cybersecurity capabilities, IT savviness, and increasing number of successful technology companies. In particular, Romania is home to the European

[1]For more, see: https://www.zdnet.com/article/locky-ransomware-how-this-malware-menace-evolved-in-just-12-months/; https://www.zdnet.com/article/ransomware-an-executive-guide-to-one-of-the-biggest-menaces-on-the-web/

Cybersecurity Competence Center, has a high performance in broadband internet speed and its talent pool in IT is ranked among the best in Europe.

TABLE I. INTERNET USAGE IN THE EASTERN EUROPEAN REGION SOURCE:(INFO COMPILED BY AUTHORS)

Country	World pop. rank	World population	Internet users	Internet user %	World rank of internet users
North Macedonia	124	2,083,160	1,589,659	76.3%	72
Albania	117	2,930,187	2,105,339	71.8%	81
Greece	58	11,159,773	7,923,438	71.0%	83
Serbia	75	8,790,574	6,182,411	70.3%	85
Romania	45	19,679,306	12,545,558	63.8%	105
Bulgaria	88	7,084,571	4,492,326	63.4%	106

The geographical area of the South Eastern European region and the prevalence of languages throughout this region are the very basis for our examination of cybersecurity issues in this study. The six corpus languages are Albanian, Bulgarian, Greek, Macedonian, Romanian, and Serbian which are spoken by a population of approximately 66 million people (Table II). The area of this study has a population of 51,727,571, which, if it were a single country, would rank 28th in the world by population.

TABLE II. PREVALENCE OF LANGUAGES UNDER CONSIDERATION IN THIS STUDY SOURCE: (INFO COMPILED BY AUTHORS)

Language	Approximate number of speakers in Millions (M)
Romanian	24. 3 M
Greek	13. 1 M
Serbian	11 M
Bulgarian	8 M
Albanian	6 M
Macedonian	3.5 M

C. The Metric: The Levenshtein Distance

Levenshtein distance is a string metric used in information theory to quantify the difference between two sequences [19]. The Levenshtein distance established between two units/words is the least possible number of single-character modifications required to convert one to the other. The Levenshtein distance (MLD) was modified to elucidate the fact that certain languages swap the positions of parts of speech when performing an ABA translation. Strings are compared until no characters match. Then proceeding forward, the number of mismatched characters are counted until another match is met. Processing is continued until the example ends. MLD is the sum of mismatched pairs. These three examples demonstrate the process.

1) ERE English-Romanian-English: Romanian: Dintre toate articulațiile de gin din toate orașele din întreaga lume, ea intră în a mea.

Of all the gin joints in / all the towns / in all / the world, she

Of all the gin joints in / every city /around / the world, she

/ store 9 ↑ store 6 ↑

/ walks into / mine.
/ enters / mine.
/ store 9 ↑

Thus MLD = 9 + 6 + 9 = 24.

==========

*2) EBE English-Bulgarian-English:*Bulgarian: Лъжата се разпростира на половината земя, преди истината да има шанса да си сложи гащите.

A /lie /gets / half / way around / the earth before the
The / lie / spreads to / half / / the earth before the
Store 3↑ store 9↑ / store 9↑

truth has a chance to / get / its pants on.
truth has a chance to / put on / its pants.
/ store 6 ↑

Thus MLD = 3 + 9 + 9 +6 = 27.

==========

3) EGE English-Greek-English: Greek: Ένα ψέμα φτάνειστα μισάτηςγης πριν η αλήθεια έχειτηνευκαιρία να φορέσειτο παντελόνιτης.

A lie / gets / half / way / around the earth before the truth
A lie /reaches / half / / around the earth before the truth
/ store7 ↑ / store 3 ↑

has a chance to / get its/ pants on.
has a chance to / put on/ her pants.
/ store 6 ↑ store 8↑

Thus MLD = 7 + 3 + 6 + 8 = 24.

III. RESEARCH METHODOLOGY

The goal of this research is to construct a metric that may be used to ascertain the probability that a text retrieved from an external source is representative of a cyberattack, whether human- or machine-initiated. We can capture the text in order to establish a profile against which an unrecognized text can be checked and subject it to the ABA test mentioned above in order to ascertain the original language of the probable cyberattack. We created a sequence of twenty English quotations, half of which are quotations from English literature (Q), Table III and the other half from English popular culture, specifically movies (F). Each text sample was exposed to the ABA procedure in each of the six languages mentioned above.

One can reasonably inquire why familiar English language phrases and film or television dialogue should be employed as a test bed. The rationale is that recognized quotations are more likely to adhere to proper English grammar and syntax, whereas cinema dialogue is frequently intended to emulate actual English conversation, being thus more likely to adhere to the conventions of everyday speech.

TABLE III. TEST QUOTATIONS FROM ENGLISH LITERATURE AND FILM SCRIPTS

No.	Category	Quotation	Length (no. of chars)
T1	F	"I'm as mad as hell, and I'm not going to take this anymore!	60
T2	Q	When a person suffers from delirium, we speak of madness. When many people are delirious, we talk about religion.	113
T3	F	Of all the gin joints in all the towns in all the world, she walks into mine.	77
T4	F	Open the pod bay doors, please, HAL.	36
T5	F	Mrs. Robinson, you're trying to seduce me. Aren't you?	54
T6	F	Keep your friends close, but your enemies closer.	49
T7	F	If you build it, he will come.	30
T8	Q	A lie gets halfway around the earth before the truth has a chance to get its pants on.	86
T9	F	I have always depended on the kindness of strangers.	52
T10	Q	Sex and divinity are closer to each other than either might prefer.	67
T11	Q	Political correctness is despotism with manners.	48
T12	Q	The only way to get rid of a desire is to yield to it.	54
T13	Q	Whether you think that you can, or that you can't, you are usually right.	73
T14	Q	There are no facts, only connotations.	38
T15	Q	I'm living so far beyond my income that we may almost be said to be living apart.	81
T16	Q	People demand freedom of speech to make up for the freedom of conviction which they avoid.	90
T17	F	Tell'em to go out there with all they got and win just one for the Gipper.	75
T18	F	Round up the usual suspects.	28
T19	F	Love means never having to say you're sorry.	44
T20	Q	The greatest glory in living lies not in never falling but in rising every time we fall."	89
		TOTAL CHARACTERS	1244

The average MLD values, emerging from the translation into one of the available languages and then back to English for each of the test sentences, are provided in Table IV. The sources for the quotations can be found at [8].

This examination is intended to ascertain the cyberattack's original source language. If the other language could be limited to the six instances, the attacked party would be able to condense the spectrum range of probable assault sources. Additional analyses to ascertain the translated material's validity are used.

A. Comparing Literary Quotes (Q) and Film Dialog (F)

Half of the text quotes were taken from Q examples and half from F examples, on the presumption that the majority of writers quoted in literature observe strict grammatical rules and that screenwriters may be more prone to deviate from grammatical norms for achievement of dramatic effect. As a result, we compared the two subsets of Q and F in order to determine whether translation systems were more accurate in terms of MLD (Table V).

TABLE IV. VALUE OF MODIFIED LEVENSHTEIN DISTANCE (MLD) FOR LANGUAGE PAIRS

ABA Example	Code for Translation	MLD Value Averaged Over All Test Entries
English-Romanian-English	ERE	29.0%
English-Bulgarian-English	EBE	25.1%
English-Macedonian-English	EME	30.2%
English-Greek-English	EGE	30.5%
English-Serbian-English	ESE	26.5%
English Albanian-English	EAE	28.5%
English-French-English	EFE	15.5%
English Spanish-English	ESpE	17.0%
English-German-English	EGeE	20.7%
English-Russian-English	ERuE	32.6%
English-Chinese-English	ECE	35.5%
English-Hindi-English	EHE	30.0%

These findings appear to imply that the translation program works similarly across all sets of cases, regardless of the translation type.

B. Comparing the Most and Least Accurate MLD Measures

The MLD was compared for each language's test bed of twenty items, T1-T20. In terms of accuracy, the six languages under examination fell into two categories (of three languages each): Bulgarian, Serbian, and Albanian, and Greek, Macedonian, and Romanian. Across the entire range of quotes, those for which the MLDs were minimal were found in Table VI.

The identification of specific test items and the MLD values aid in the refinement of the type of test bed that will provide a more precise characterization of the test item. For instance, the T7 test item translation is 94 percent accurate across all ABA language translations chosen. Thus, the T7 is unlikely to be a strong option for determining the source language of a potential hacker. Additionally, Table VII demonstrates the usefulness of the translation type and the test quote. For example, the MLD is zero in six of the test instances T1-T20, indicating that these items are useless for identifying malicious cyberattacks. This is also true for T11 and T14 which correspond to flawless translations in four of our six language pairs.

TABLE V. MLD SCORES ON FILM DIALOGUE (F) AND LITERARY QUOTES (Q)

Language Pair	MLD F Score	MLD Q Score	Total MLD Score	% Difference
English – Romanian	25.54%	20.97%	22.83%	4.57%
English – Bulgarian	17.23%	21.38%	19.69%	4.15%
English – Macedonian	23.17%	24.09%	23.71%	0.92%
English – Greek	22.57%	24.90%	23.95%	2.33%
English – Serbian	21.39%	20.43%	20.82%	0.96%
English – Albanian	21.39%	23.55%	22.67%	2.16%

TABLE VI. QUOTATIONS AMONG T1-T20 WITH MINIMAL AND MAXIMAL MLD VALUES

Minimal MLD Case	Film (F) or Quotation (Q)	MLD Value	Max. MLD Case	Film (F) or Quotation (Q)	MLD Value
T9	F	0.7	T15	Q	30.8
T7	F	2.0	T3	F	27.0
T14	Q	3.3	T8	Q	24.3
T6	F	3.7	T2	Q	23.8
T11	Q	5.3	T16	Q	23.7
T19	F	6.0	T17	F	21.3
T4	F	9.5	T20	Q	21.2
T18	F	9.5	T1	F	19.0
T12	Q	10.8	T10	Q	12.0
T13	Q	11.3	T5	F	11.8

C. Alternative Test to Further Distinguish Eastern European Region Languages

Comparisons are made in this study among the six chosen Eastern European region nations, but the linguistic differences among these nations and their natural languages are somewhat obscured when analyzed in the context of these languages compared to other world languages where the linguistic structures are so different. For example, in an earlier paper [8] we considered more closely related European-based languages compared to Hindi, Russian and Chinese. In order to draw greater distinctions between the set of languages in this study, we used the same test bed (T1-T20) of well-known English language text, and sought to find a subset of the test items that would provide a greater discrimination between the Eastern European languages in the study. One approach to doing this could be to consider subsets of the test items to see if the differences in results applied only to the Eastern European languages would be more pronounced when only a subset of the test items is considered.

To measure the difference in the results for only our six Eastern European countries in consideration, we compute the following. Consider the differences of the results for each pair of languages. In order to ensure that we can eliminate positive and negative differences, we square each comparison of values. For these six languages, there are (6 x 5)/2 = 15 pairs for comparison (Table VIII).

Calculate the squares of the differences for each of the 15 pairs. For example: Albanian – Bulgarian: $(282 – 245)^2 – 37^2 = 1369$; Macedonia – Romanian: $(295 – 284)^2 = 11^2 = 121$.

Express each term as a fraction of the square of the overall number of comparisons, N = 1244 (see page 4). Thus $N^2 = 1547536$. Thus, the Albanian – Bulgarian ratio is 1369/1547536 = 0.00088463. Averaging all the difference for the 15 comparisons gives a separation factor related to the specific quotation differences, in this case 0.0043163.

In order to be able to distinguish GT for the translation, we look for a subset of the test items where the separation factor is greatest. In that way, we can more easily distinguish which languages were used through the average magnitude of the separation factor. To give a more distinct separation, we choose the following subset of test items, {T1, T3, T13, T15, T17, T20}. Constructing the same table as before, one obtains (Table IX).

Calculate the squares of the differences for each of the 15 pairs. For example: Albanian – Bulgarian: $(120 – 108)^2 – 12^2 = 144$; Macedonian – Romanian: $(155 – 139)^2 = 16^2 = 256$. Express each term as a fraction of the square of the overall number of comparisons, N = 455. Thus $N^2 = 207025$. Thus, the Albanian – Bulgarian ratio is 144/207025 = 0.00069556. Averaging all the difference for the 15 comparisons gives a separation factor related to the specific quotation differences, in this case 0.043163. Thus, the separation ratio is multiplied by approximately 5 between the values calculated (0.043163 vs. 0.00847) for the selected test items { T1, T3, T13, T15, T17, T20 } as compared to all 20 test items T1-T20.

TABLE VII. FONT SIZES OF HEADINGS

Translation Type	No. of Exact Translations (of 20)	Test Quotation (T or Q)	No. of Exact Translations
EBE	5	T9 (F),	5
ERE	4	T11 (Q), T14 (Q)	4
EME, ESE	3	T6 (F)	3
EAE	2	T4 (F)	2
EGE	1		

TABLE VIII. CHARACTERS CHANGED IN TRANSLATION FOR EACH LANGUAGE PAIR USING ALL QUOTES T1-T20

	Albanian	Bulgarian	Greek	Macedonia	Romanian	Serbian
Chars changed in translation	282	245	298	295	284	259
Albanian		1369	256	169	4	529
Bulgarian			2809	2500	1521	196
Greek				9	196	1521
Macedonia					121	1296
Romanian						625
Serbian						
Column sum		1369	3065	2678	1842	4167
% of total chars squared		8936/ 207025 =	.0043163 (see below)			8936

TABLE IX. CHARACTERS CHANGED IN TRANSLATION FOR EACH LANGUAGE PAIR USING 6 INDICATED QUOTES

	Albanian	**Bulgarian**	**Greek**	**Macedonia**	**Romanian**	**Serbian**
Chars changed in translation of subset	120	108	141	155	139	121
Albanian		144	441	1225	361	1
Bulgarian			1089	2209	961	169
Greek				196	4	400
Macedonia					256	1156
Romanian						324
Serbian						
Column sum		144	1530	3630	1582	2050
% of total chars squared			.00847864			8936

IV. CONCLUSION

Despite the fact that the information gathered from the given data set of quotations and language translation methods is to a certain extent limited, the study may be a valuable strategy in real-time detection of ransomware and other cyberattacks because it can narrow the field of suspects of an attack. For instance, if the language used in a suspected attack is exposed to the methods given in this paper, the approaches presented herein may yield critical information on the language of origin used by the attacker. The metric we have presented above is applicable to a large range of languages and is characterized by sustainability, usage of limited resources and scalability properties.

ACKNOWLEDGMENT

The authors are grateful for the contributions and assistance provided by Professor Silvia Florea of the Lucian Blaga University of Sibiu, Romania.

REFERENCES

[1] Ablon, Lillian, Martin C. Libicki and Andrea A. Golay. Markets for Cybercrime Tools and Stolen Data. RAND Corporation. 2014.

[2] Carr, Madeline. Public–private Partnerships in National Cyber-security Strategies. Chatham House, January 2016.

[3] Deibert, Ron. Bounding Cyber Power: Escalation and Restraint in *Global Cyberspace*. *Internet Governance Papers: Paper No. 6*. Center for International Governance Innovation. 2013.

[4] Deibert, Ron. The Geopolitics of Cyberspace After Snowden. *Current History* 114, no. 768. 2015, pp 9–15.

[5] Libicki, Martin C., Lillian Ablon, Timm Webb. The Defender's Dilemma: Charting a Course Toward Cybersecurity. RAND Corporation. 2016.

[6] Meyer, Paul. Outer Space and cyberspace: A Tale of Two Security Realms. In *International Cyber Norms: Legal, Policy & Industry Perspectives* edited by Anna-Maria Osula and Henry Rõigas, NATO CCD COE Publication, Tallinn, 2016.

[7] Patterson, Wayne and Cynthia Winston-Proctor, *Behavioral Cybersecurity*, CRC Press, 2018, pp 178.

[8] Patterson, Wayne, Acklyn Murray and Lorraine Fleming, Distinguishing a Human or Machine Cyberattacker, Proceedings of the 3rd Annual Conference on Intelligent Human Systems Integration, Modena, Italy, February 2020, pp. 335–340.

[9] Young, A. and Moti Yung. *Malicious Cryptography: Exposing Cryptovirology*. Wiley, 2004.

[10] Rao, A., B Jha G Kini. Effect of Grammar on Security of Long Passwords. Proceedings of the 3rd ACM Conference on Data and Application Security and Privacy, Ser. CODASPY '13. New York, NY, USA: ACM, 2013, pp. 317–324.

[11] Bonneau, Joseph, Ekaterina Shutova. Linguistic Properties of Multi-Word Passphrases. Proceedings BS12 of USEC '12: Workshop on Usable Security, March 2012.

[12] Veras, R., C Collins, J Thorpe. On the Semantic Patterns of Passwords and Their Security Impact. NDSS. Internet Society, 2014.

[13] Sparell, Peder, Mikael Simovits. Linguistic Cracking of Passphrases Using Markov Chains. Cryptology ePrint Archive, Report 2016/246.

[14] Weir, M., Aggarwal, S., Medeiros, B. D., and B. Glodek. Password Cracking Using Probabilistic Context-Free Grammars. Proceedings of the IEEE Symppsium on Security and Privacy, 2009, pp 391–405.

[15] Ur, Blasé et al. The Art of Password Creation. *34th IEEE Symposium on Security and Privacy, SP '13, IEEE*, May 2013. http://passwordresearch.com/papers/paper442.html

[16] Komanduri, Saranga et al. Telepathwords: Preventing Weak Passwords by Reading Users' Minds. Proceedings of the 23rd USENIX Security Symposium. August 20–22, 2014. San Diego, CA.

[17] Alzahrani, A., Alshehri, A., Alshahrani, H., Alharthi, R., Fu, H., Liu, A., & Zhu, Y. RanDroid: Structural Similarity Approach for Detecting Ransomware Applications in Android Platform. 2018 IEEE International Conference on Electro/Information Technology (EIT), 2018, pp 0892–0897.

[18] Keating, Dave. "Despite Brexit, English Remains The EU's Most Spoken Language By Far". Forbes. Retrieved 7 February 2020.

[19] Levenshtein, Vladimir I. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*. February 1966, 10 (8): pp. 707–710.

ONLINE SOURCES

https://www.flashpoint-intel.com/blog/linguistic-analysis-wannacry-ransomware/

https://www.theguardian.com/technology/2017/may/17/hackers-shadow-brokers-threatens-issue-more-leaks-hacking-tools-ransomware

https://www.vice.com/en/article/d7ydwy/why-does-dnc-hacker-guccifer-20-talk-like-this

https://www.theregister.com/2015/09/18/coinvault_ransomware_arrests_dutch_netherlands/

https://www.bbc.com/news/business-34589710

https://www.zdnet.com/article/locky-ransomware-how-this-malware-menace-evolved-in-just-12-months/;

https://www.zdnet.com/article/ransomware-an-executive-guide-to-one-of-the-biggest-menaces-on-the-web/