White Paper:

Cybersecurity - What's Language got to do with it?

Judith L. Klavans, Ph.D.[1]
Computational Linguistics and Information Processing (CLIP) Laboratory
University of Maryland Institute for Advanced Computer Studies (UMIACS)
jklavans@umd.edu

Executive Summary

A new opportunity to explore and leverage the power of language analysis in ensuring effective Cybersecurity is presented. Cybersecurity as a discipline includes the protection of computer systems, and the detection of intrusion. In general, cyber experts hail from computer systems analysis, and perform the difficult task of out-thinking the clever hacker community to keep systems safe. However, the role of language in achieving this difficult goal has been under-explored. The purpose of this White Paper is to propose methods to exploit critical information that can only be derived from language to assist in the Cybersecurity effort. For example, determining the language of a user, extracting the content from email or other communication, analyzing descriptions along with photos or videos can all contribute to keeping valuable information safe and thus improving Cybersecurity.

A. Cybersecurity – a 21st Century Phenomenon

The term "Cyber-security" is as meaningful as it is meaningless. "Cyber" means somewhere in Internet space; but, since the Internet has only been around for 50 years, and since accessibility to the internet only took off in the last 20 years, the field of "cyber" as a defined discipline is as new as nano-technology. However, there is one big difference. Internet use has soared, whereas "nano" still lies in the esoteric. According to the Internet World Statistics Group, there are over 2.5 Billion people using the Internet, as of 2013. In contrast to the newness of "cyber", the security part of cyber-security is as old as humankind. The great leader of Carthage, Hannibal, in the 3rd century B.C.E used spies to collect intelligence from Roman troops in order to find out about intentions and capabilities, and thus succeeded in crippling many major Roman attacks. These human spies are analogous to "bots" going into systems and capturing information. Even though the Romans had poor intrusion detection techniques, ironically, they still learned new methods from Hannibal's strategies, then turned these strategies around, and used them to defeat Hannibal's Carthage.

The juxtaposition of these two parts of "cyber" and "security" has given rise to a new, but also old, field. This is the major challenge facing cyber-security, namely, how to apply long-standing principles of safety and defensive awareness from the field of security to the nascent area of cyber. What are the challenges specific to cyber and which are just more of the same, applied to this new area? What is novel about the *cyber* part of cybersecurity? What are some of the specific challenges in cyber that did not exist in prior security challenges? How does language contribute to solutions to the networking challenges of Cybersecurity?

B. Cybersecurity and Government Policy

Cybersecurity has been prominent for government policymakers more and more in recent years. For example, the DHS's 2013 budget request asks for $769 million for cybersecurity efforts – 74 percent higher than 2012's budget request[2]. President Barack Obama said during his State of the Union address that he had signed an executive order aimed at protecting government and businesses from what he called "the rapidly growing threat from cyber-attacks." Two Senate bills have proposed different approaches to the problem; Sen. Joe Lieberman (I-Conn.), along with Sen. Susan Collins (R-Maine) and Sen. Jay Rockefeller (D-W.Va.), introduced the Cybersecurity Act of 2012. The bill gives the Department of Homeland Security regulatory authority over the private companies that control designated critical infrastructure systems — such as telecommunications networks and electric grids — and would require owners and operators of critical infrastructure to meet security standards established by the National Institute of Standards and Technology, the National Security Agency and other designated entities, or face unspecified civil penalties. A second bill introduced by Sen. John McCain (R-Arizona) focuses on information sharing to secure systems, rather than regulation.
The Republican proposal would update the criminal code to reflect the threat cyber-criminals pose, reform the Federal Information Security Management Act (FISMA), and focus federal investments in cybersecurity.

Although national security, including cybersecurity, should be a bipartisan concern, in actual fact, crippling partisanship has stalled legislation. Republicans and business lobbyists have opposed efforts to force companies to adhere to minimum security standards, saying it is unfair for the government to require them to make costly security improvements. Experts say companies should be required to meet security benchmarks or they won't do them. In mid-March 2013, a coalition of Internet advocacy organizations and individuals launched a week of action to combat the CISPA, the Cyber Intelligence Sharing and Protection Act, viewing this as undermining existing privacy laws by giving overly broad legal immunity to companies who share users' private information, including the content of

communications, with the government. Their position is that legislation intended to enhance our computer and network security must not sacrifice long-standing civil liberties and protection. At the same time, the president's executive order says minimum-security standards will be voluntary, not mandatory, and companies will receive incentives to follow them.

C. Defining Cybersecurity

For the purpose of this white paper, we consider first the larger scope of the field, to include networks, computers, programs and data. Strategies for defending cyber threats include two major components:

- monitoring of the security status of systems to detect threats to sensitive data,
- developing methods to respond to threats in real-time.

As for intrusion techniques, the general public is familiar with the Trojan horse technique, where malware, i.e. malicious code, is embedded in a seeming innocuous attachment, in order to achieve a (usually mal-intentioned) purpose, such as collecting passwords, spying on the individual via a web browser, or industry espionage for competitive advantage. Such information, once obtained, has been used to purposefully bring down large systems and to compromise large financial, government and military networks. Firewalls are also familiar to the general public; they look outward for intrusions in order to stop them from happening. Intrusion detection systems (IDS) differ from firewalls because firewalls limit access between networks to prevent intrusion but do not signal an attack from inside the network. In contrast, an IDS evaluates a suspected intrusion once it has taken place and signals an alarm.

In general, intrusion detection systems, firewalls and other Internet security devices can stop the average hacker, but new threats use stealth techniques that these defenses cannot detect on their own. In the past, most of these systems rely purely on software tools and statistical methods to help discover and remediate potential security vulnerabilities but more and more, language also adds a critical component in ensuring Internet security.

D. The International and Multilingual Nature of the Internet

Language plays a fundamental role in many areas of Cybersecurity. First, the international nature of Internet use is shown in Table 1.[3] According to Internet World Stats, the premier tracking site for reliable Internet use statistics, as of June 2012, nearly half (44.8%) of Internet users originate in Asia. Surprisingly,

approximately half of that 44.8 percentage (21.6%) comes from Europe, and then just over half of that 21.6 percentage (11.4%) are North American. Although this shows international penetration of the Internet by continent, the true numbers of users by language is given in Table 2, which is even more germane to the question of language and Cybersecurity.

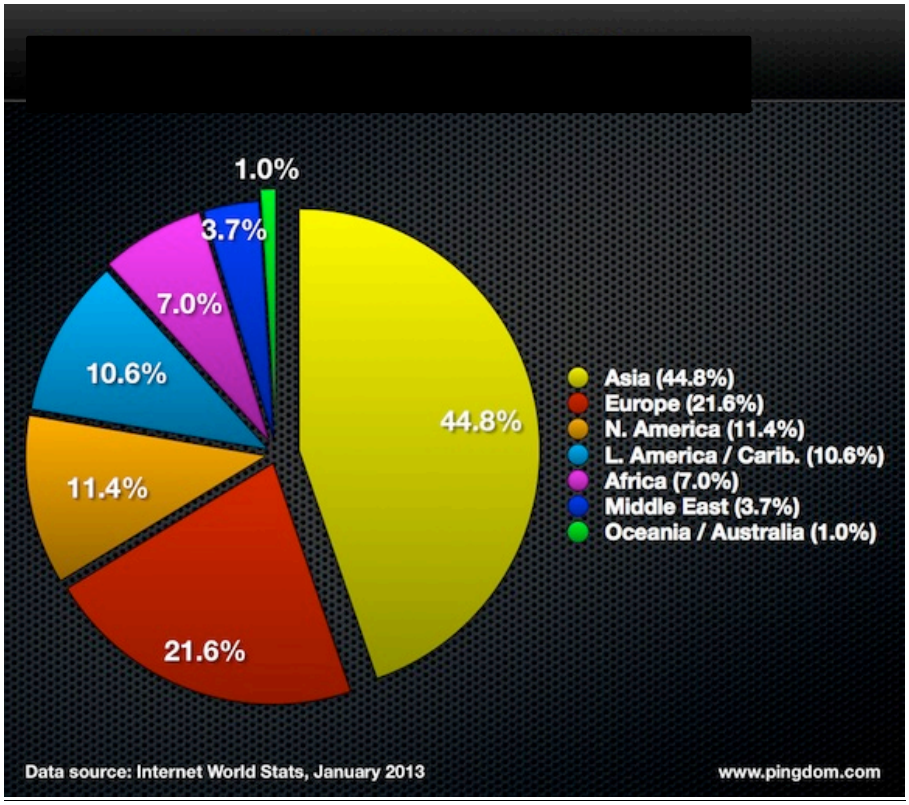

Table 1: Internet users by Region as of June 2012

The data in Table 2 reflect the top languages used on the Internet, since the regional percentages in Table 1 do not fully reflect the language used. For example, a significant percentage of European and Asian users interact primarily in English on the Internet; at the same time, in many communities two or more languages are part of the geographical region. In these cases, mixed usage is common, i.e. a French person might communicate in both French and English, and even use both languages in the same document, SMS or spoken audio. This phenomenon, called "code-switching" is very common in informal language where a group might purposefully use code-switching as a mechanism to identify group membership. For example, so-

called “Spanglish” is often used in the United States by the Hispanic Mexican community to indicate common background.

Mixed language data is not reflected in Table 2, since that would mean double (or triple) counting people; therefore, these numbers add up to the total world population since each person is categorized into a single language count, which is intended to represent his or her primary language, known as L1 in the language literature. This is in contrast to L2, which is usually not the maternal language but a language that could be used in schools vs. home (such as English in schools in Nigeria but Yoruba, Hausa, or Igbo might be the only language used at home), or learned in the community (such as Spanish in Basque-speaking regions of Spain where the languages are in close contact so Spanish is the primary language of commerce and business whereas Basque is used at home and in early education) or second language learning at school (such as is typical in the United States, where children typically start L2 language learning at the late age of 12 or 13.)

Top 10 Languages	Internet Users by Language	World Population for this Language (2011 Estimate)	Internet Penetration	Internet Users % of Total	Growth in Internet (2000 - 2011)
English	**565,004,126**	1,302,275,670	43.40%	26.80%	301.40%
Chinese	**509,965,013**	1,372,226,042	37.20%	24.20%	1478.70%
Spanish	**164,968,742**	423,085,806	39.00%	7.80%	807.40%
Japanese	**99,182,000**	126,475,664	78.40%	4.70%	110.70%
Portuguese	**82,586,600**	253,947,594	32.50%	3.90%	990.10%
German	**75,422,674**	94,842,656	79.50%	3.60%	174.10%
Arabic	**65,365,400**	347,002,991	18.80%	3.30%	2501.20%
French	**59,779,525**	347,932,305	17.20%	3.00%	398.20%
Russian	**59,700,000**	139,390,205	42.80%	3.00%	1825.80%
Korean	**39,440,000**	71,393,343	55.20%	2.00%	107.10%
Total: Top Ten	**1,615,957,333**	**4,442,056,069**	**36.40%**	**82.20%**	**421.20%**
Other Languages	350,557,483	2,403,553,891	14.60%	17.80%	588.50%
World Total	**2,099,926,965**	**6,930,055,154**	**30.30%**	**100.00%**	**481.70%**

Table 2: Languages of the Internet

In order to fully comprehend the data presented in Table 2, the following figures are provided to visualize the number of users by language, how much the Internet has penetrated the overall population, and trends in growth of Internet use by language. The top 10 languages, reflecting columns one and two of Table 2, are shown below in Figure 1:

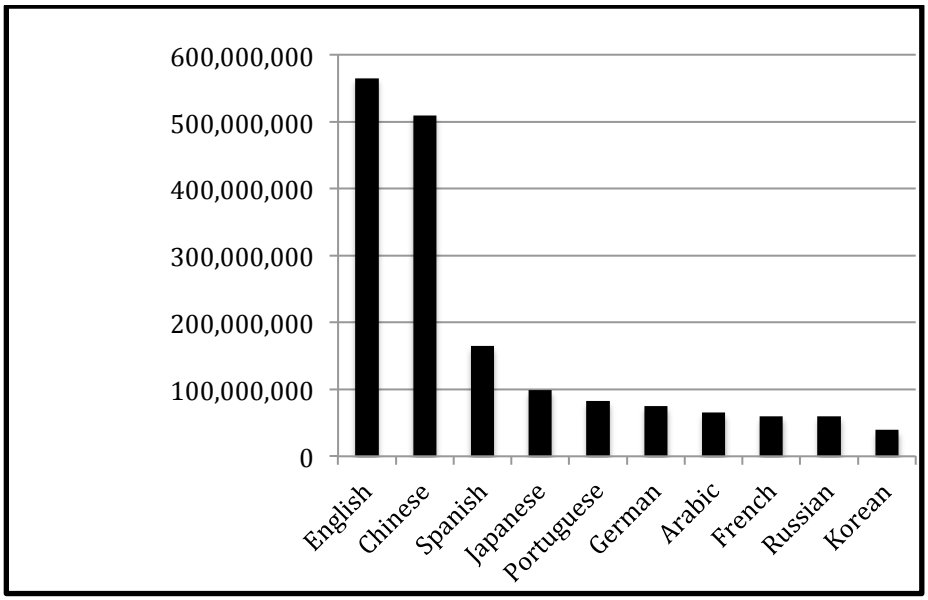

Figure 1: Internet Users by Languages

What Figure 1 demonstrates is that, not surprisingly, the predominant language used on the Internet is English. However, Chinese is a close second, with a rapid drop-off to Spanish and the other top ten languages. This data is important since any Internet analysis of content clearly must prioritize these languages, but the data in Figure 1 represents just one piece of the Languages and Cybersecurity challenge. More important is Figure 2 which demonstrates the rapid change in Internet use by language:

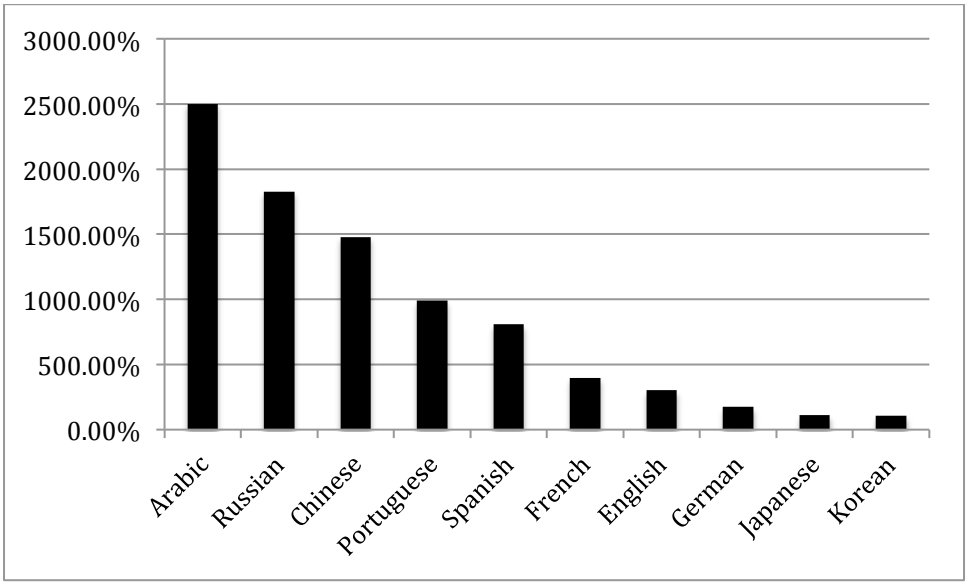

Figure 2: Percentage Increase in Internet Use from 2000-2011

Figure 2 reflects the fact that, although English is the most used language, Arabic, Russian, Chinese users have increased over 1000% in just over ten years. These three language groups are significant in that the vast majority of cyber-crime incidents have originated in Russia and China, whereas Arabic has been more in use for other purposes. This alarming increase sends a warning, to be further discussed in the next section.

Finally, Figure 3 reflects what is known as "Internet Penetration" by language, i.e. the ratio of those using the Internet in a given language, and the overall number of speakers of that language. Internet penetration reflects how much the population as a whole is using the Internet in their native language.

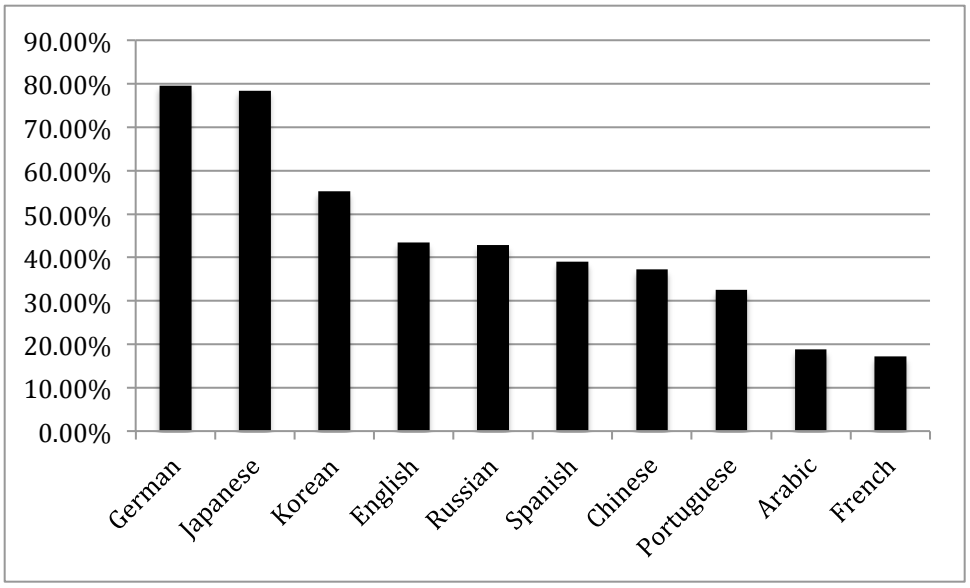

Figure 3: Internet Use Penetration by Language

To clarify the significance of Figure 3 for Cybersecurity, compare English, Russian, Chinese and Arabic. For English, 43.4% of the native English-speaking population is using the Internet, and by now that number has likely risen to 50%. This means that nearly half of the English-speaking world is using the Internet and using it in English. As can be clearly seen in Figure 3, the percentage for Russian is very close at 42.8%, indicating that there is major penetration across the population. Chinese is close behind at 37.2%, just behind Spanish at 39.0%. In contrast, Arabic penetration is only 18.8%, but this lower penetration ratio must be viewed in light of Figure 2. The rapid increase in Arabic users over the past 10 years suggests that Arabic Internet penetration will soon surpass English and Chinese, as well as others. The increased Arabic penetration statistics may well be a factor in the role of cyber-activism in the Arab Spring in Egypt in 2011, and served as an effective cyber-weapon to be discussed in the next section on cyber-crime, cyber-activism, cyber-intelligence, and cyber-weaponry.

E. The Contribution of Language in Cybersecurity

Although the traditional areas of cybersecurity focus on systems protection and on threat neutralization, increasingly content analysis is recognized as a key to both external and internal cybersecurity. The Internet is increasingly multilingual.[4] Some of the areas where language is essential are outlined below:

1. **Language is the medium of communication for cyber-crime and cyber-activism.**
    a. What is the language of the cyber-crime group?
    b. How does this group usually work?
    c. Who else is connected to this group (usually by language)?

2. **Language is used to filter collection**
    a. Should we pay attention to this document, text, audio or video file?
    b. Is this something we should save for the future?
    c. If so, why?

3. **Language is required to achieve the cyber-crime and cyber-activity goal**
    a. What language is this?
    b. What does it say?
    c. What are the intentions of the author or authors?
    d. What purpose does this document, tweet, SMS, audio or video file serve?
    e. What is said 'between the lines'?

Well-known cyber-crime groups have established mini-communities and societies centered around language. Establishing trust in order establish reputation can be inferred from social media.[5] All trusted groups use language cues to establish credibility; this is a well-established socio-linguistic principle which has been documented since the 1950's.[6] For example, the Russian group called "Moonlight Maze" goes back to 1998; Moonlight Maze was the code name for a long-term infiltration of American defense institutions, which lasted from 1998 until its accidental discovery about two years later. To the alarm of computer security specialists, who were surprised by the attack, the computer systems of the Pentagon, NASA, the Department of Energy, and leading research universities were compromised by cyberattackers operating from somewhere inside the former Soviet Union, though Russia denied any involvement. They apparently succeeded in accessing thousands of sensitive files, including maps of American military installations around the world, troop configurations, doctrine, and blueprints of military hardware.

The Moonlight Maze case exemplifies the three aspects of cyber-crime where language is key:

1) Establishing a trust group of criminals through language
2) Cyber-sleuthing to determine which sites are targeted to collect information
3) Cyber-analysis to analyze and extract valuable intelligence from the information collected, including names from maps, people and their actions in reports, and guidelines for military or defense actions.

Trust is a key component of group formation. The group had to have used email and phones to communicate since this was the pre-SMS era. Modern groups often embed in email communication video and audio which are harder to identify and crack, but which still contain valuable information.

A more recent cyber-crime incident occurred in mid-March 2013 on South Korean broadcasters and banks. This was a coordinated attack that hit roughly 32,000 computers on 20 March at 2:00 pm local time and wiped the hard drives and master boot record of at least three banks and two media companies simultaneously. The attacks reportedly put some ATMs out of operation, preventing South Koreans from withdrawing cash from them. Initial reports alleged that the attack was initiated in North Korea, in retaliation for UN sanctions on nuclear testing; then China became the focus of suspicion. Tracing IP addresses is typically complex, but any evidence of communication between the people planning this attack would give clear clues as to the origins of the group and to the individuals with the most influence over the group's intentions, plans and actions.

A case where the object of the cyber-attack was the detection and capture of language information can be found in the incident named Operation Aurora 2009-2010. For several months, hackers operating inside the People's Republic of China engaged in a systematic campaign targeting Google's computer systems, as well as those of a variety of other top-tier American companies. While it is likely that intellectual property and potential state secrets were sought, the primary goal of the attack, according to Google's investigation, was to access the personal gmail accounts of Chinese human rights activists throughout the world. In order for this information to be of use, the dissident's personal communications would need to be captured and analyzed in whatever language the activists were using. The first step would be to identify the language (be it Chinese, English, or another language) and then to extract targeted information.

Recent cyber-activism relies on social media, such as Twitter and Facebook. Mentioned above was the example of the Arab Spring. A chain-reaction of

revolutionary activity initiated in Tunisia in December 2010 resulting in the overthrow of the prime minister, Mohamed Ghannouchi; this activity then rapidly spread through Bahrain, Egypt, Yemen, Libya, Sudan, Jordan, and then Syria. In Iran's capital, students broke into the British embassy, looted and burned British property. Protest has since spread rapidly across the Middle East, communicated primarily through the channels of social media. For the critics who had previously dismissed platforms like Facebook and Twitter as vapid mechanisms for celebrity gossip, puerile activity, and self-aggrandizement, the toppling of entire regimes in Tunisia and Egypt suggested that these tools were as effective for organizing protests and revolutions as they were for organizing birthday parties. Revolutionary movements throughout the Arab world have demonstrated that social media is an effective tool for cyber-activism. As the ongoing tumult throughout the Middle East enters a sort of adolescence, however, the true role of social media in the revolutions is undergoing a necessary closer inspection.

Social media is reshaping human language through the unprecedented mixing of idioms, dialects, and alphabets. Computational linguists and sociolinguists have been monitoring Twitter to track on-the-ground sentiment over the course of the Arab Spring, particularly in Egypt and Libya.[7] Linguist David Beaver and his associates analyzed Arabic-language tweets before and after the death Muammar Qaddafi, the deposed leader of Libya. Beaver et al. (op. cit.) used Twitter's system of geocoding, translated from Arabic to English to perform standard sentiment analysis tests on this data, looking for positive and negative words and other features of this text. A dynamic description of Libya's Twitter traffic emerged which reflected how traffic not only increased, but also terms related to positive sentiment like "good" and "wonderful" rose as well. Religious sentiment was also on display, with a significant increase in the frequency of words like "Allah," "sacrifice" and "gospel."[8] Computational linguistic analysis and translation of twitter, along with blogs and other social media, has become a full-fledged field of study[9], showing how language analysis is key to understanding cyber-activism. Linguistic modeling will reveal multiple viewpoints, influences, centers of power, disputed topics, and overall sentiment. Understanding the attributes of nodes based on content in social media provide much more information that structure alone[10], including topic analysis based on the language and culture of the communication itself and the communicator.

F. Language in Cybersecurity

This brief paper has outline the critical role that Language plays in ensuring a cyber-secure nation and global infrastructure.

[1] This paper was originally written in 2013, and was distributed informally. The Technical Report was submitted to the University of Maryland in 2015. Citations should reflect the 2015 date. Since 2013, funding on this topic has produced additional research and developments.

[2] http://www.dhs.gov/xlibrary/assets/mgmt/dhs-budget-in-brief-fy2013.pdf

[3] http://www.internetworldstats.com/languages.htm. For this table, Internet World Stats notes that the most recent Internet usage information comes Nielsen Online, International Telecommunications Union, GfK, and other reliable sources. World population information comes from the U.S. Census Bureau.

[4] John Rittinghouse and William M. Hancock (2003) Cybersecurity Operations Handbook. Elsevier: Burlington, MA.

[5] Jennifer Golbeck and James Hendler (2004) Accuracy of Metrics for Inferring Trust and Reputation. In Proceedings of the 14th International Conference on Knowledge Engineering and Knowledge Management, Northamptonshire, UK. (pp. 116-131). Springer Berlin / Heidelberg

[6] Elihu Katz and Paul Felix Lazarsfeld (1955). Personal Influence. New York: Free Press, based on earlier work from Paul Felix Lazarsfeld, Bernard Berelson, Hazel Gaudet (1944). The People's Choice: How The Voter Makes Up His Mind In A Presidential Campaign. Columbia University Press: New York.

[7] Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, and Rebecca Passonneau. 2011. Sentiment analysis of Twitter data. In Proceedings of the Workshop on Languages in Social Media (LSM '11). Association for Computational Linguistics, Stroudsburg, PA, USA, 30-38.

[8] http://languagelog.ldc.upenn.edu/nll/?p=3537, as reported in http://www.nytimes.com/2011/10/30/opinion/sunday/twitterology-a-new-science.html.

[9] Other relevant citations include:

(a) Samuel Brody, Nicholas Diakopoulos (2011) Cooooooooooooooollllllllllllll!!!!!!!!!!!!!! Using Word Lengthening to Detect Sentiment in Microblogs. EMNLP 2011: 562-570.

(b) Laura Jehl, Felix Hieber and Stefan Riezler (2012) Twitter Translation using Translation-based Cross-Lingual Retrieval.. In Proceedings of the Seventh Workshop on Statistical Machine Translation (WMT 2012), Montral, Quebec, Canada.

(c) Shane Bergsma, Paul McNamee, Mossaab Bagdouri, Clayton Fink, and Theresa Wilson. 2012. Language identification for creating language-specific Twitter collections. In Proceedings of the Second Workshop on Language in Social Media (LSM '12). Association for Computational Linguistics, Stroudsburg, PA, USA, 65-74.

[10] Jennifer Golbeck (2013) Analyzing the Social Web. Morgan Kauffman: New York.