Dictionary for Computer-Assisted Text Analysis of Cyber Security (TACS)

Ana Levordashka
University of Bath

Adam Joinson
University of Bath

Simon Jones
University of Bath

Abstract

Being able to elicit and analyse people's beliefs about security in a systematic, scalable manner, is critical for the advancement of research on human dimensions of cybersecurity. Computational methods can facilitate and augment traditional forms of language analysis, especially when combined with specialised domain knowledge and qualitative approaches. In this paper we describe the development and evaluation of a dictionary for automated, computer-supported content analysis of cybersecurity texts. The open-source dictionary, or codebook, consists of approximately 700 empirically-derived words and phrases related to cybersecurity, organised into a framework of concepts (e.g., "Hacking", "User") and relatively simplistic categories based on security protections, threats, and contexts. The dictionary is accompanied with a set of functions to tag, annotate, and extract information from textual data. To demonstrate the dictionary and illustrate its core functions, we use the dictionary to contribute to ongoing discussions about people's understanding of cybersecurity and the utility and scope of cybersecurity information provided in the mass media. We report how the portrayal of cybersecurity in the mass media varies, and compares to the narratives of a diverse sample of individuals. Implications for how cybersecurity can be studied using computational approaches to natural language are discussed.

1. Introduction

While considerable progress has been made in understanding, and addressing – the role of human factors in cybersecurity (e.g., [1, 17, 37]), as a research field we still struggle to elicit and analyse people's beliefs about security in a systematic, scalable manner. This matters because while there are tools to measure people's cybersecurity awareness, knowledge and skills, work on people's folk beliefs and mental models tends to be (research) labour intensive, relying heavily on qualitative methods and skilled researchers. Similarly, studies of cyber security materials have traditionally relied on trained researchers to analyse large bodies of textual (and other) material [2].

One method to address both the resource and scale problem with textual materials is the use of computational methods. Within social science, the use of computational methods to analyse large bodies of text has become relatively well-established [15, 25]. These approaches to "computer-assisted content analysis" typically range from simple word counts to topic modelling and the analysis of subject-object-verb triads [25], and can incorporate bespoke discipline-specific tools (e.g. [28]) intended to reveal latent variables for analysis.

Such techniques are not intended to replace qualitative analysis. Indeed, Nelson [25] argues that computational techniques can support qualitative researchers by enabling pattern recognition, hypothesis refinement and pattern confirmation. It has not gone un-noticed that there are considerable similarities between computational methods and grounded theory, with a number of researchers proposing that machine learning can be used to support qualitative methods, for instance by identifying instances of ambiguity ([6]), identifying patterns in the material ([23]) or to confirm patterns identified using traditional methods ([25]).

Much of the power of the tools and techniques used in computational social science is increased when combined with specialised domain-based knowledge. For instance, a corpus of news articles about cybersecurity can be queried using existing tools such as topic modelling without recourse to specialised tools. Such models require domain knowledge to interpret, in much the same way that qualitative researchers might interpret themes using traditional methods. However, if we wanted to automate further, we might want to develop a specialist dictionary that allocated the terms used. At its simplest, this might simply allocate a word or phrase to 'security' or 'not security' categories. This would then enable the researcher to dive deeper into a large dataset – e.g. by identifying trends across time in the proportion of mainstream news articles that mention information security, or to identify passages for further qualitative analysis. An additional stage would be to further divide security-related terms into discrete categories – for instance, those that dealt with protective mechanisms (e.g. anti-virus) and threats (e.g.

USENIX Symposium on Usable Privacy and Security (SOUPS) 2020. August 9 -- 11, 2020, Boston, MA, USA.

malware). This would not only allow a more detailed analysis in the earlier case (e.g. by tracking coverage of threats over time in the mainstream press), but also to combine with additional dictionaries (e.g. LIWC [28]) to examine the association of, for instance, emotion words with threats vs protective mechanisms. A further step would be to associate each term with its grammatical function by allocating each term to either 'subject' 'verb' or 'object' (SVO). SVO tagging allows us to identify the role of each element within an utterance. For instance, the sentence "Hackers attack Sony" can be understood in terms of the subject [hackers], verb [attack] and object [Sony]". If the sentence were reversed, the meaning would be substantially different (i.e. "Sony attacks hackers").

Computational methods can facilitate research on the human dimensions of cyber security. The present paper builds on dictionary-based approaches and computational grounded theory to develop a Text Analysis tool for the Cyber Security domain (TACS). TACS consists of an open-source dictionary, or codebook, of ca. 700 words and phrases related to cyber security, organised into *concepts* (e.g., Hacking, Passwords, User) and grouped into overarching *categories* (e.g., Threat Mechanisms, Security Mechanisms, Cyber Entities), alongside a set of functions to apply the dictionary codes to textual data and transforming it into numeric representations, and extracting key terms, excerpts of text, and informative phrases (e.g., subject-verb-object triples).

2. Background & Related Work

In this section, we provide a brief overview of relevant computational methods for language analysis and how they can be relevant to research on the human dimensions of cyber security. A comprehensive review of these literatures computational language literature is beyond the scope of this paper, instead we focus specifically on a body of work where the integration of qualitative and quantitative methods is discussed in greater depth.

2.1. Computational Language Analysis

Attempts to study human language through computation originate in computational linguistics, or natural language processing (NLP), a branch of computer science which, due to its ground-breaking advances and applications [15], has gained significant attention in various domains, including the social sciences. Some social science fields have a tradition of successfully applying computational approaches (e.g., [13] in political science; [35] in psychology); others have published papers to increase awareness of such approaches and their relevance (e.g., [3] in Journalism; [25] in Sociology, [14] in Communication).

Boumans and Thrilling [3] offer a helpful categorisation of computational techniques relevant to the social sciences into: counting-based methods, supervised and unsupervised machine learning. Counting based methods rely on identifying the number of times given words and phrases occur in a text. Supervised machine learning approaches attempt to classify text into pre-determined categories, based on a body of manually classified (annotated) data. Unsupervised approaches attempt to inductively identify and extract themes in text, based, roughly speaking, on the occurrence and distribution of words. Additionally, there are NLP techniques to extract key terms from text [20], infer words' part of speech (noun verb, conjunction), and syntactic dependencies (subject, verb, object, adverbial modifier, determinant), as well as to infer, or "learn", the meaning of words from the context in which they appear [21].

A substantial part of computational linguistics techniques rely or greatly benefit from manually constructed resources, such as machine-readable lexicons, dictionaries, and ontologies, which can be thought of as codebooks, mapping the relationships between words and concepts. Dictionaries can be constructed around specific topics or theoretical constructs [12, 10] or general linguistic relationships [22]. The typical use of dictionaries involves looking for orthographic matches in a body text.

Dictionaries are elegant, and due to their mechanistic operation, extremely reliable [3]; their major disadvantages being the manual, laborious development and linguistic ambiguity. Ambiguity can stem from homonymy (orthographically identical words with different meaning) or use (e.g., sarcasm, metaphors). Although ambiguity remains a major challenge in NLP, there are ways improve accuracy. Classically, ambiguous terms can be embedded in relatively unambiguous phrases. Another approach is developing a rule-based model that take into account information such as neighbouring terms, concordance, or syntactic dependencies. One example of a rule-base model is the sentiment classifier VADER [16], which outperformed LIWC [r] and other benchmark lexicons in accuracy.

Computational language analysis can be accomplished through a number of proprietary (WordStat) and open-sourced applications [4]. Scripting language such as Python or R and specialised libraries (e.g., spaCy [33], gensim [31]) are currently the standard for advanced techniques and are the most transparent, flexible and reproducible [25].

There is currently no dictionary or related lexical resource for the human dimensions of cybersecurity. One paper proposes a framework for developing an ontology [27]. Existing ontologies and taxonomies, such as STIX (Structured Threat Information Expression; [34]) and CAPEC (Common Attack Pattern Enumeration and Classification of Threats; [5]), are designed for technical purposes and incident reporting. There are a number of non-exhaustive glossar-

ies [26, 24]; detailed, scoping definitions [8] and a knowledge organisation framework (CyBoK [30]). Some categorisation of cybersecurity terminology can be found in general lexical resources such as WordNet [22] and ConceptNet [19] but those are fairly limited in scope. We acknowledge and build on these resources, but no single one of them fulfils the purpose intended by the dictionary.

2.2. Computational Grounded Theory

Traditionally programmatic, computational approaches are often adopted for the purposes of quantitative research, and seen as too simplistic for in-depth, interpretative research. While such views are not unfounded and still widely prevalent, there are notable non-trivial discussions on how to best utilise computational language techniques from the social sciences.

Nelson [25] proposed a three-step methodological framework for computer-assisted grounded theory analysis. As a first step, large and messy bodies of text are reduced into simpler, interpretable lists or networks of words, with the help of key-term extraction algorithms and general linguistic resources such as WordNet. These representations are then interpreted along with deep reading of representative text excerpts to produce refined hypotheses. In the final stage, relevant techniques (e.g., supervised machine learning) or resources (e.g., dictionaries) are used to test the prevalence of hypothesised patterns in a body of data. Similar workflows have been proposed by others [6, 23]

2.3. The Language of Cyber Security

Considerable progress has been made in understanding the role individuals play in cybersecurity. For example, that people's security decisions are greatly influenced by their beliefs and the information they receive, and that these are not always optimal [1, 17, 37]. The tools to measure people's cybersecurity awareness, knowledge and skills, beliefs and mental models tends to be labour intensive, relying heavily on qualitative methods and skilled researchers. Computational approaches can greatly support this work. Known or hypothesised phenomena (e.g., good user advice; mental models) can be linked to linguistic markers (e.g., 'use antivirus software' and 'hacker targets') and reliably identified in a body of natural language. Rader & Wash [29] use topic modelling to identify patterns in informal stories that are relevant to individuals' medal models.

Theoretical advances notwithstanding, it should be taken into account that cyber security is a rapidly changing field [32]—not only are new threats and countermeasure constantly emerging from a technological arms race, but so are new actors and domains [9, 18, 11]. Computational language analysis can be used to reveal differences and patterns in previously unexamined domains and generate hypotheses, as was done, for example, for age differences [18].

These few examples illustrate the potential that computational techniques can offer the field of usable security. To support these endeavours, we set out to design and develop a tool that is adaptable and agile, deployable quickly and at scale, and non-prescriptive in terms of definitions or approaches to cybersecurity.

3. Dictionary Development

3.1. Language Data

For the purposes of development and validation, we compiled a collection of language data, including lexical data, comprised of individual words and phrases, and natural-language documents, consisting of multiple full sentences. The aim was to cover a broad range of contexts, topics, author and audience profiles (e.g., age, geographical location, technical expertise). The lexical data consisted of: (1) large-scale survey and (2) online glossaries. The natural language-data included: (1) open-ended survey responses, (2) interview transcripts, (3) articles in periodicals, (4) semi-structured database entires, (5) an online discussion board; on a broad range of topics, including personal accounts of cybersecurity-related experiences, challenges, and behaviours at home and in the workplace by experts and non-experts, advice seeking, expert description of incidents, consequences and countermeasures. We additionally collected a control sample of data matched by a number of characteristics but on topics other than cyber security. Details on the corpus data are provided in the Appendix.

3.2. Vocabulary

In this section we detail the approaches applied to construct a vocabulary from the language data outlined in the previous section. At the core of the vocabulary were the key terms, collected from a large, diverse sample of experts and non-experts. The separate key terms were manually grouped into unique concepts. Whenever possible, lexemes were matched to entries in WordNet [22]—a lexical database, where words are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. This allowed us to retrieve relevant related terms (synonyms, antonyms, hypernyms, and hyponyms). The glossary terms were grouped manually. For the natural language data, we extracted key terms from each data set, using the PageRank algorithm [372] implemented via the Python package textacy [7]. We then grouped the terms by semantic similarity, with the help of a vector space model available via the Python library spacy ('en_core_web_lg'), with 685k GloVe vectors trained on Common Crawl, an open repository of petabytes web data.

3.3. Categorisation

The next step in the dictionary development was to organise the vocabulary into a set of helpful categories. The categorisation was performed by the authors and consisted of nu-

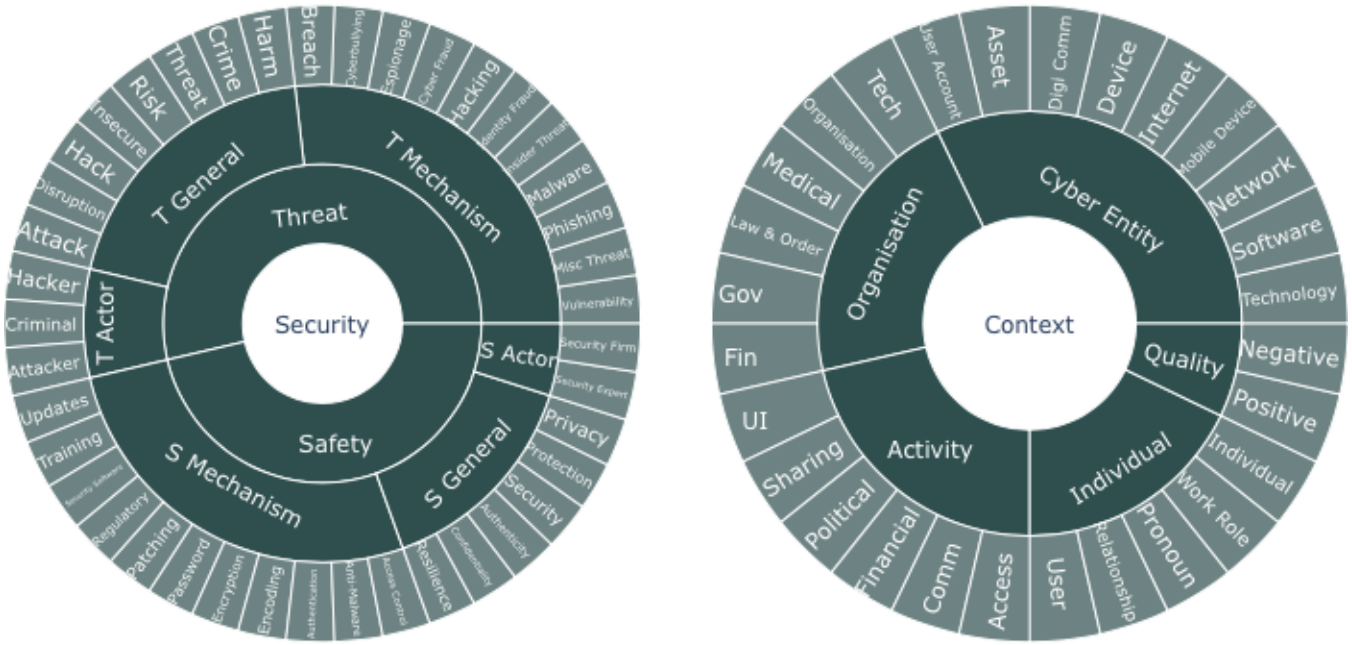

Figure 1: Visual summary of the TACS framework. The visualisation features *categories* (dark tones) and *concepts* (light tones). Not displayed are the *terms* nested within concepts and the orthogonal *domain* categories.

merous iterations. Exploratory card sorting exercises were used to inform our overall strategy and ideate candidate categories with the help of existing resources domain-specific ([27, 34, 5, 8]) and general linguistic resources ([14, 19]). The categorisation framework was then refined through several cycles of (1) manual categorisation of vocabulary lexemes and (2) evaluation through key-word-in-context analyses, whereby a search query was performed for each Concept in the totality of annotated documents and the accuracy of the annotation ("Does the assigned category correspond to the intended meaning?") was judged by the authors for 10 random instances.

The final categorisation we arrived at was comprised of 4 levels of organisation: *Terms* nested within *concepts* nested within in *categories*, with two separate sets of categories—Cyber Security and Context. In addition to the main categorisation, each term was assigned one of the following *domain* categories: Cyber Security (e.g., password, hacker), Security (e.g., threat, safety), Cyber (e.g., computer, user), and General (e.g., person, company). Detailed description of the framework in provided in the following section.

3.4. Disambiguation

Sense ambiguity was handled through the use of ngrams. Firstly, ambiguous terms were embedded in ngrams (e.g., the word 'security' in the phrase 'security flaw' is not counted towards the concept of General Security, but the entire phrase is counted towards Vulnerability).

One major design choice was to include words from domains other than cyber security (e.g., user, government). The inclusion of non-domain words has the advantage of providing rich contextual information. However, it can be problematic in word counting methods, as it would capture these words in the various other contexts in which they often appear. To resolve this, we resorted to a simple rule-based model for context-aware tagging. In the current implementation, the model takes into account neighbouring terms (default window is 30 words). Words belonging to domains other than Cyber Security are only counted if their neighbouring terms include one or more Cyber Security words or a combination of Security and Cyber words.

4. TACS

4.1. Framework

The final Framework consists of 4 levels of categorisation: *Terms* nested in *concepts* nested in *categories*, with two separate sets of categories—Cyber Security and Context, which can be thought of as separate dictionaries. Figure 1 provides an overview of the main 3 levels, which are displayed in full in Table 1. The lowest-level, *terms*, is simply a grouping of lexical variants and close synonyms, akin to stemming or lemmatisation. Cyber Security terms are grouped along the dimensions of Threat and Safety/Protection; and further into the following categories: General-language terms describing threat- and security-related states and activities, Actors, and specific Mechanisms, (similar to

the Tools Techniques Procedures dimension of STIX [34]). Context terms are grouped into Individuals, Organisations, Activities, Quality/State, and Cyber Entities.

The full list of TACS *categories* and *concepts*, their prevalence in our corpus and a sample of frequent *terms*, and can be seen in Table 1. That none of the *concepts*, including the general-language ones, are over-represented in the control corpus, points to the utility of the rule-based tagging. Pronouns are typically far more prominent, in fact in default preprocessing they are considered 'stop words' and excluded from analyses along with non-function words like propositions.

4.2. Implementation and Functionality

Analysing text with TACS begins with supplying input data, which can be a single document or a group of documents in text or pdf format, or texts within a spreadsheet. Each document is processed sequentially.

First, a document is split into individual tokens. The tokens are matched with dictionary lexemes, starting with multi-word lexemes (ngrams). Matched tokens are assigned (tagged with) the corresponding dictionary categories; ngrams are merged into single tokens (**tagging** function). This first transformation results in tokenised tagged documents serves as the basis of subsequent operations. By default, tokens are only tagged if their 20 neighbouring tokens contain at least one Cyber Security term or a at least one Cyber and one Security term, as determined by the Domain categorisation in TACS. This **rule-based tagging** can be disabled or parametrised (e.g., specifying custom window size and threshold). On a home computer, TACS processes 10,000 documents in 4 minutes with rule-based tagging and under 1 minute without.

The tagged documents can then be transformed into numeric representations by counting the frequency of occurrences of dictionary categories in all documents, each document, or groups of documents (**vectorisation**); TACS also provides the option to normalise the raw frequencies (e.g., per-10k; tf-idf). To contextualise, the *each* vectorisation of TACS would produce the document term matrix commonly used in document classification and topic modelling. TF-IDF is a technique that assesses the relative importance of terms within documents, which TACS implements via a basic formula or the Python library textacy [7].

A third class of TACS functions—**extraction**—return excerpts of full text based on queries. This could be a span of words or sentences containing a certain dictionary category or a combination of categories. A word span around a given or concept is a format commonly referred to as Key-Word-In-Context (KWIC); multi-sentence paragraphs will provide the enough context for some forms of qualitative analysis. Full text can be returned with annotation whereby

Category	Concept	NRF*	Top 3 Terms
Threat Mechanism	Malware	14.7	Virus_2013,Malware_595,Spyware_298
	Vulnerability	7	Vulnerability_1058,Bug_279,Exploit_203
	Breach	4.7	Breach_1030
	Hacking	3.7	Hacking_810
	Cyber Fraud	2.7	Spam_408,Scam_285,Spoofing_32
	Phishing	1.6	Phishing_334
	Misc Threat	1.1	Botnet_154,Zombie_36,Terrorism_19
	Cyberbullying	0.2	Catfishing_28,Bullying_14,Stalking_9
	Espionage	0.2	Eavesdropping_39
Threat General	Threat	3.4	Threat_648,Fear_69,Danger_64
	Risk	3.1	Risk_558,Liability_156,Hazard_20
	Insecure	1.8	Vulnerable_203,Exploitable_178,Insecure_29
	Hack	1.3	Hack_293
Security	Harm	1.1	Damage_199,Loss_62
	Crime	0.5	Crime_128
	Disruption	0.4	Failure_80,Disruption_31
Threat Actor	Hacker	11.1	Hacker_2273,CybCriminal_64
	Attacker	0.9	Attacker_178
	Criminal	0.8	Criminal_178
Security Mechanism	Access Control	11.7	Firewall_1110,Authentication_957,Signature_197
	Password	8	Password_1727
	Encryption	4.7	Encryption_902,Cryptography_90,Hashing_11
	Updates	2.9	Updates_905
	Regulatory	3.3	Safeguard_268,Insurance_154,Compliance_139
	Patching	3.8	Patching_805
	Anti-Malware	3.5	Anti-Virus_614,Anti-Spyware_97,Anti-Malware_20
	Encoding	2.6	Encrypting_545,Encoding_28
	Training	0.3	Training_82
Security General	Protection	17.3	Protection_1209,Data Protection_1209,Secure_736
	Privacy	1.6	Privacy_592
	Security	1.1	Safety_272,Security_6
	Confidentiality	0.8	Confidentiality_189
	Resilience	0.1	Resilience_19
	Authenticity	0	Authenticity_7
	Cybersecurity	8.6	Cyber Security_907,Cybersecurity_907
Quality	Negative	4.2	Uncertain_903,Suspicious_207,Confusing_122
	Positive	2.2	Important_457,Certain_204,Vigilance_67
Organisation	Organisation	14.4	Company_5333,Industry_421,Institution_55
	Tech	1.5	Microsoft_440,Windows_316,Google_119
	Gov	2.3	Government_646,Authorities_43
Context	Fin	1.9	Bank_550
	Law & Order	0.2	Police_53,Watchdog_6,FDA_5
	Medical	0	NHS_9
Individual	Pronoun	34.5	2nd You/Your_2880,1st I/My_2106,1st pl We/Our_1892
	User	5.7	User_879,Customer_422,Consumer_178
	Work Role	4.3	Employee_1261,Contract Manager_51
	Individual	4.6	Person_716,Expert_147,Researcher_125
	Relationship	0.1	Friend_28,Relative_11,Acquaintance_1
Cyber Entity	Internet	7.2	Internet_1311,Website_740,Online_510
	Device	11.9	Computer_2014,Device_702,Home Device_2
	Asset	8.9	Data_2053,Infrastructure_204,Money_161
	Software	8.6	Software_1203,App_704,Code_344
	Technology	7	Tech_1255,Cyber_293,Electronic_247
	Network	8.1	Network_1730,Server_426,Workstation_34
	Mobile Device	2.2	Phone_319,Laptop_174,Tablet_74
	Digi Comm	0.3	Attachment_66
	User Account	0.1	Email_22
Activity	Comm	2.9	Emailing_538,Messaging_229
Activity	UI	1.6	Download_178,Click_125,Delete_61
	Access	0.2	Logging in_41
	Financial	0	Shopping_18

Table 1: Relative prevalence of TACS Concepts in cyber security texts from Periodicals [Data: MPD] and Interviews [Data: I*]), as compared to control texts [Data: CI, CM]). The scores represent the difference between the cyber and control corpus, with higher scores indicating more occurrences in the cyber corpus. The scores are normalised by dividing the number of occurrences by the number total words in the sub-corpus multiplied by 10,000 and can be interpreted as number of occurrences per 10,000 words.

dictionary categories appear next to the text they are assigned to (in html, markdown, or plain text). Parsing of syntactic (via the Python library spacy) can provide data for analyses such as subject-object-verb or object-centred sentiment. Finally TACS provides a number of helpful **visualisations** for its outputs.

TACS is open-source and its full vocabulary is available for users to implement in any number of ways. The current implementation we provide is a python function and a notebook, which can be used with minimum programming experience when following its documentation. Future plans include the development of desktop and online applications and an R package. The project is available at [ANON].

4.3. Use Case: Media Mental Models

To demonstrate TACS and exemplify its core functions, we report a use case related to human-centred security research. The use case contributes to ongoing discussions within the field on the utility and scope of cybersecurity information provided in the mass media.

Scholars have questioned the focus of mass media cybersecurity reporting, noting that end users receive too much information regarding cyber security, not all of which relevant and useful [17, 37]. Exposed to such information, users may develop mental models, which are unhelpful. Ion et al. [17] showed discrepancies between expert and user protection measures, most notably updates. Wash [36] studied user mental models qualitatively, identifying unhelpful believes (e.g., "hackers target only rich, important individuals"), later constructing a survey instrument to study such believes in a representative sample, finding that they indeed relate to insecure behaviour. Rader and Wash [29] studied the informal stories of cyber security that users encounter on the media both qualitatively and with topic modelling. Such research is relevant and timely.

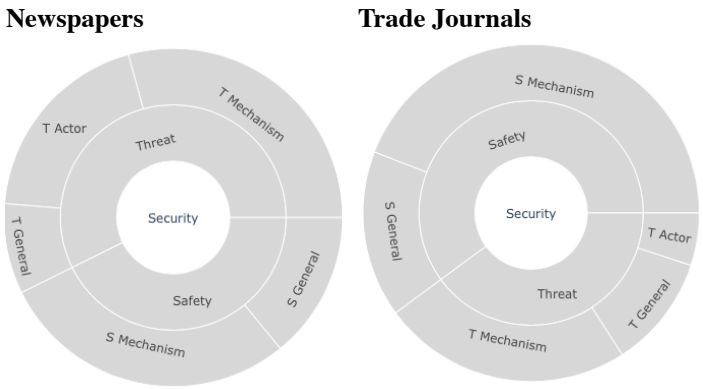

Figure 2: Category prevalence (tf-idf) in articles, published in Newspapers and Trade Journals [Data: MPD]

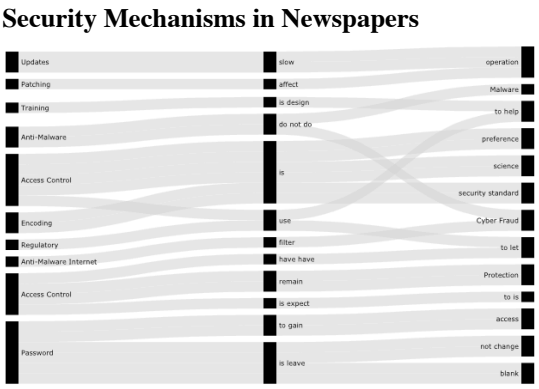

Figure 3: Subject-Verb-Object with Security Mechanisms as Subject in Newspapers [Data: MPD]

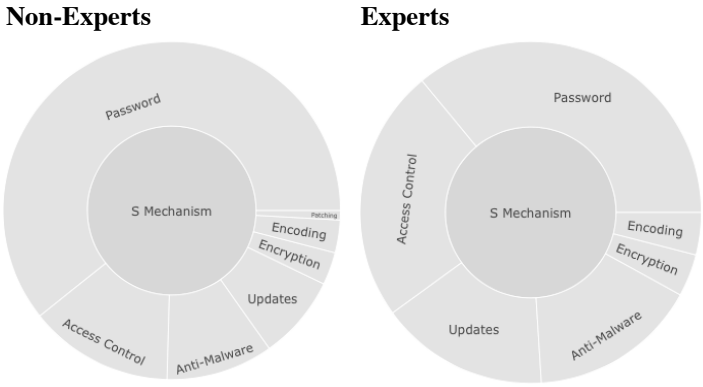

Figure 4: Category prevalence (tf-idf) in narratives of experiences and protection measures by experts and non-experts [Data: SPS]

Non-Experts

account i have norton	**update**	passwords and use 2step authenticity
when using open wifi	**updates**	software change passwords be careful
regularly	**up to date**	security and privacy settings
software security	**update**	never click on links within
family media	**update**	passwords regularly

Experts

sometimes even	**update**	java etc
security news and apply	**updates**	or change settings if
where they exist i	**update**	my systems i use long
oss	**up-to-date**	
regularly	**update**	passwords with password manager

Figure 5. Key-Word-In-Context for Updates in the language of experts and non-experts. [Data: SPS]

To show how TACS can be used to explore the prevalence of security content and tap into the implied processes and mental models, we use it to automatically analyse and compare the cyber-security content of 504 newspapers articles, 577 articles published in trade journals [Data: MPD], and narratives elicited from 22 security experts and 303 non-experts [Data: SPS]. Details can be found in Appendix.

Figure 2 shows two sunburst diagrams comparing the tf-idf scores of the dictionary's Security categories for each publication type. Tf-idf is a form of weighted term frequency, whereby terms that appear in all documents are assigned lower weights and discriminating terms get higher scores. TACS produced interactive diagrams allowing its users to zoom in on a category and examine the underlying concepts and terms.

The figure reveals that newspapers, as compared to trade journals, report more threats and focus on threat actors, specifically hackers. Wash and Rader [37] found in a representative sample that it is common for individuals to believe that hackers do not target home users. Prevalence of hackers/hacking on the media was also observed using topic modelling [29]. Taken together, these findings can be seen as an indication that the majority of individuals, who tend to learn about cybersecurity from the popular media, might not receive the coping-oriented, balanced communication that industries benefit from.

Term prevalence alone might not provide sufficient information with regard to what information is being communicated. Another TACS function is the extraction and optional visualisation of syntactic dependencies, such as subject-verb-object triples. Figure 3 is the output of a query for 'Security Mechanisms' as subject. Such data can reveal suggested processes, such as 'updates slow operation'.

Figure 4 displays category tf-idf scores in the narratives of experts and non-experts. A visual inspection reveals that, out of all protection mechanisms, non-experts spoke primarily of passwords. Experts were more likely than non-experts to refer to updates, which is a well-know discrepancy [17]. One way to get a richer picture of how updates are discussed by each group is by retrieving excerpts of text where updates are mentioned. Depending on the intended analysis, these could be short phrases (KWIC) or entire paragraphs.

These use cases are intended to illustrate how TACS can be utilised to study the language surrounding cyber security in a systematic, scalable manner, rather than provide conclusive evidence. The research design, sampling, and interpretation required for empirical work are beyond the scope of this paper. Instead, our aim was to briefly exemplify how the TACS categories and outputs can be used to address existing research questions in the field.

5. Discussion

The aim of the present research was to design and develop a dictionary suitable for automated, computer-supported content analysis of cybersecurity texts. We used established methods within computational linguistics to develop and validate our lexicon, and allocated the words in the dictionary to categories within a relatively simplistic framework based on security and threat mechanisms, along with additional context.

TACS is designed specifically for the purposes of human-centred research and such that it can be used without extensive technical expertise or resources. It builds on advanced NLP techniques and libraries to offer simple, usable, and computationally efficient approach to language analysis. At the same time it maintains the integrity of these libraries and can provide users with the outputs and parametrisation required for advanced computational research.

The dictionary organisation along lexical variants (*terms*) and unambiguous *concepts* makes it versatile and applicable to a broad range of topics. With its large, empirically-derived vocabulary, it is a unique contribution to human-centred cyber security researcher and a foundation for the development of further, more specialised resources. The framework - while not intended to be exhaustive or complete - allows for research questions to be addressed, as shown in our use case. TACS allows researchers who are not trained in computational methods to conceptually replicate analyses used in published computational research on human-centred security (e.g., [29, 18]).

The decision to include words from other domains, classify them as Context, and handle them with rule-based tagging is a methodological innovation aimed to facilitate an all-round streamlined meaning extraction. This approach is a direct response to efforts to assist in-depth interpretative analysis. As with any innovation, further refinement will be necessary; the principle, however, it a valuable contribution and can be adopted by other domains. The Context portion of TACS is suitable for the broader area of Human-Computer Interaction.

In the design and development of TACS, we considered a range of uses on the spectrum between computational and humanistic interpretive analytical approaches. For the purpose of rapid insights, or "gisting", TACS offers interactive visualisations along with easy inspection of performance (e.g., KWIC, annotated text excerpts). Full-text outputs should serve the needs of a broad range of qualitative approaches. The possibility to easily alternate between full-text, semi-structured (e.g., SVO) and numeric (e.g., term

frequency) representations, as well as to edit the dictionary and use custom queries, makes TACS compatible with iterative approaches, such as computational grounded theory proposed by Nelson [25]. At the same time, TACS can be used to bootstrap advanced modelling, which, too, involve operations with basic linguistic features.

For users choosing not to delve into NLP, TACS offers sensible default parameters (e.g., normalised frequency for corpus-wide summaries, but tf-idf scores for group comparisons). The full functionality of TACS does require setting up Python, but the dictionary component can be used with existing software (e.g., MEH [4]). Users interested in advanced techniques would benefit from standard outputs (e.g., dtm) and good integration with Python and its industry-standard NLP packages (e.g., [33]).

5.1. Limitations & Future Work

A notable limitation of lexicon-based resources, including TACS, is that they categorise words and simple phrases, which are often insufficient for deriving meaning. TACS can reliably reveal how often the term password appears in a body of text, but not whether it is mentioned in the context of advice given to end users or a vulnerability targeted by a threat actor. Although not currently sufficient to identify such topics, TACS would greatly facilitate research seeking to address them. Firstly, TACS can, in a manner of seconds, parse thousands of documents and not only extract relevant excerpts of text but also structure them in helpful ways (e.g., cluster by subject and verb). It is now common to follow up manual coding with attempts to build automated classifiers that would reliably code unseen data. Classifiers rely on lexical features and their accuracy is often improved when dictionaries like TACS are used to extract and group these features. As a lexicon-based resource with information extraction functionality, TACS can support work on various topics, as well as inductive, exploratory work.

The scope of TACS is another deliberate design choice which poses certain limitations. The size of the vocabulary necessitated a relatively shallow, flat structure, which can be refined further with the help of domain experts or empirically through topic modelling and combination with other lexical resources. For example, subject-verb-object analysis can reveal relationship types between the categories and inform the development of an ontology. That TACS and its outputs are conductive for such refinements, is indicative that its current stage of development constitutes a complete, usable tool.

Although we did follow the development and validation process adopted by major lexical resources, additional steps can be taken to further validate accuracy in terms of both precision and recall. The KIWC analyses we performed internally can be outsourced to subject matter experts or a larger sample of non-experts to get precision estimates and inter-rater agreement. The multiple rounds of query expansion (i.e., retrieving semantically related terms via WordNet, ConceptNet and spacy's GloVe vectors) ensure good recall. To formally test this, however, the dictionary transformation would have to be compared to manually annotated text.

Catering to the needs of both qualitative and quantitative research is at the heart of TACS and user studies can be used to develop it further. To achieve good usability, we ensured that a broad range of outputs can be accessed via streamlined functionality and documentation. User studies would help refine the documentation, inform the development of additional functions, and design a suitable graphical user interface.

Labour intensive and notoriously challenging to develop, manually-constructed domain dictionaries are a valuable asset to the research field they were build for [3]. This paper is a major step towards developing a comprehensive resource for computational language analysis in human-oriented cyber security research.

6. Conclusions

TACS is an open-source tool for automated analysis of cyber security language, designed specifically for the purposes of human-centred research and such that it can be used without extensive technical expertise or resources. It is the first manually-constructed domain dictionary for the human dimensions of cyber security. Tailored to meet the requirements of advanced computational and humanistic interpretive analytical approaches, TACS is a useful, practical addition to a growing toolkit of approaches aimed to bridge the gap between qualitative and qualitative methods. It offers simple, usable, and computationally efficient approach to computer-assisted text analysis.

References

[1] Anne Adams and Martina Angela Sasse. 1999. Users are not the enemy. *Commun. ACM* 42, 12 (1999), 41–46.

[2] Maria Bada, Angela M Sasse, and Jason RC Nurse. 2019. Cyber security awareness campaigns: Why do they fail to change behaviour? *arXiv preprint arXiv: 1901.02672* (2019).

[3] Jelle W Boumans and Damian Trilling. 2016. Taking stock of the toolkit: An overview of relevant automated content analysis approaches and techniques for digital journalism scholars. *Digital journalism 4*, 1 (2016), 8–23.

[4] RL Boyd. 2014. MEH: Meaning extraction helper. (2014).

[5] CAPEC. 2019. CAPEC - Common Attack Pattern Enumeration and Classification. https://capec.mitre.org/. (2019).

[6] Nan-Chen Chen, Margaret Drouhard, Rafal Kocielnik, Jina Suh, and Cecilia R Aragon. 2018. Using machine learning to support qualitative coding in social science: Shifting the focus to ambiguity. ACM Transactions on Interactive Intelligent Systems (TiiS) 8, 2 (2018), 1–20.

[7] Burton DeWilde. 2017. textacy Documentation. (2017).

[8] Enisa. 2016. Definition of Cybersecurity - Gaps and overlaps in standardisation. https://www.enisa.europa.eu/publications/definition-of-cybersecurity. (2016).

[9] EPSRC. 2014. Outputs of the human dimensions of cyber security workshop. https://epsrc.ukri.org/files/research/outputs- of- the- human- dimensions- of- cyber- security- workshop/. (2014).

[10] Alastair J Gill, Asimina Vasalou, Chrysanthi Papoutsi, and Adam N Joinson. 2011. Privacy dictionary: a linguistic taxonomy of privacy for content analysis. In Proceedings of the SIGCHI conference on human factors in computing systems. ACM, 3227–3236.

[11] HM Government. 2017. Internet Safety Strategy - Green paper. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/650949/Internet_Safety_Strategy_green_paper.pdf. (2017).

[12] Jesse Graham, Jonathan Haidt, and Brian A Nosek. 2009. Liberals and conservatives rely on different sets of moral foundations. Journal of personality and social psychology 96, 5 (2009), 1029.

[13] Justin Grimmer and Brandon M Stewart. 2013. Text as data: The promise and pitfalls of automatic content analysis methods for political texts. Political analysis 21, 3 (2013), 267–297.

[14] Elisabeth Günther and Thorsten Quandt. 2016. Word counts and topic models: Automated text analysis methods for digital journalism research. Digital Journalism 4, 1 (2016), 75–88.

[15] Julia Hirschberg and Christopher D Manning. 2015. Advances in natural language processing. Science 349, 6245 (2015), 261–266.

[16] Clayton J Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In Eighth international AAAI conference on weblogs and social media.

[17] Iulia Ion, Rob Reeder, and Sunny Consolvo. 2015. "... no one can hack my mind": Comparing Expert and Non-Expert Security Practices. In Eleventh Symposium On Usable Privacy and Security ({SOUPS} 2015). 327–346.

[18] Simon L. Jones, Emily I. M. Collins, Ana Levordashka, Kate Muir, and Adam Joinson. 2019. What is 'Cyber Security'?: Differential Language of Cyber Security Across the Lifespan. In Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems (CHI EA '19). ACM, New York, NY, USA, Article LBW0269, 6 pages. DOI: http://dx.doi.org/10.1145/3290607.3312786

[19] Hugo Liu and Push Singh. 2004. ConceptNet—a practical commonsense reasoning tool-kit. BT technology journal 22, 4 (2004), 211–226.

[20] Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into text. In Proceedings of the 2004 conference on empirical methods in natural language processing. 404–411.

[21] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013).

[22] George A Miller. 1995. WordNet: a lexical database for English. Commun. ACM 38, 11 (1995), 39–41.

[23] Michael Muller, Shion Guha, Eric PS Baumer, David Mimno, and N Sadat Shami. 2016. Machine learning and grounded theory method: convergence, divergence, and combination. In Proceedings of the 19th International Conference on Supporting Group Work. 3–8.

[24] NCSC. 2018. NCSC glossary. https://www.ncsc.gov.uk/information/ncsc- glossary. (2018).

[25] Laura K Nelson. 2017. Computational grounded theory: A methodological framework. Sociological Methods & Research (2017), 0049124117729703.

[26] NICCS. 2018. A Glossary of Common Cybersecurity Terminology. https://niccs.us- cert.gov/about- niccs/glossary. (2018).

[27] Leo Obrst, Penny Chase, and Richard Markeloff. 2012. Developing an Ontology of the Cyber Security Domain.. In STIDS. 49–56.

[28] James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. 2015. The development and psychometric properties of LIWC2015. Technical Report.

[29] Emilee Rader and Rick Wash. 2015. Identifying patterns in informal sources of security information. Journal of Cybersecurity 1, 1 (2015), 121–144.

[30] Awais Rashid, George Danezis, Howard Chivers, Emil Lupu, and Andrew Martin. 2017. Scope for the Cyber Security Body of Knowledge. (2017).

[31] Radim Rehurek. 2019. Gensim - topic modelling for humans. https://radimrehurek.com/gensim/. (2019).

[32] Bruce Schneier. 2000. Secrets & Lies: Digital Security in a Networked World, John Wiley & Sons. Inc. New York, NY, USA (2000).

[33] spaCY. 2019. spaCy. https://spacy.io/. (2019).

[34] STIX. 2019. Structured Threat Information eXpression (STIX). A structured language for cyber threat intelligence. https://stixproject.github.io/. (2019).

[35] Yla R Tausczik and James W Pennebaker. 2010. The psychological meaning of words: LIWC and computerized text analysis methods. Journal of language and social psychology 29, 1 (2010), 24–54.

[36] Rick Wash. 2010. Folk models of home computer security. In Proceedings of the Sixth Symposium on Usable Privacy and Security. ACM, 11.

[37] Rick Wash and Emilee Rader. 2015. Too much knowledge? security beliefs and protective behaviors among united states internet users. In Eleventh Symposium On Usable Privacy and Security ({SOUPS} 2015). 309–325

Appendix

Table 1. Summary of Corpus

Data Type	ID	Dataset description	Dem	N docs (n toks)	Control dataset	N docs (n toks)
List of Words	SPF	Open-ended survey responses from participants asked to list **characteristics or features of 'cyber security'**. Sample: 146 Children; 211 Working-Age adults; 200 Working-Age security professionals; 146 Older Adults [18]	wa-e; wa-n; ch; oa	469 (4,780)	Individuals in the same survey, responding to the same question but about '**security**' rather than 'cyber security	438 (3,584)
	GLO	Online glossaries of **cybersecurity terms**. Sample: [26, 24]	*wa-e*	2 (500)	x	
Survey Open-Ended Response	SPS	Open-ended survey responses from participants asked to describe a **personal experience related to cyber security** and **protection measures** taken. Sample: Same as SPF.	wa-e; wa-n; ch; oa	600 (10,673)	Individuals in the same survey, responding to the same question but about '**security**' rather than 'cyber security	457 (3,970)
Interview Transcript	IFA	Transcripts from qualitative interview studies on **cyber risk** management and **boundary negotiation** within the **family**. Sample: Families with underage children residing in the UK.	wa; ch	16 (59,716)	[CI] Interview transcripts from research projects retrieved from the UK Data Archive, various topics including technology-related topics and children as participants.	84 (503,000)
	IOA	Transcripts from qualitative interview studies on **cyber risk** management and **advice seeking** among older adults. Sample: Older adults residing in the UK.	oa	33 (242,049)		
	IEP	Transcripts from qualitative interview studies on **phishing** and other **cyber risk** in organisations. Sample: Security professionals in organisations.	wa-e	6 (29,578)		
	IEO	Transcripts from qualitative interview studies on **cyber risk** and **organisational culture**. Sample: Security professionals in organisations.	wa-e	30 (183,746)		
Article in Periodical	MPD	Articles in periodicals retrieved from the ABI/Inform database, which have **Computer Security** as their main topic, as indexed by the database. The sample was stratified by publication type (magazines, newspapers, trade journals) and publication decade (90s, 00s, 10s).	*wa*	1,114 (1,542,039)	[CM] Articles in periodicals retrieved from the ABI/Inform database, on topics related to **technology** and **security** but not computer security; stratified by publication type and decade.	3,641 (3,960,834)
Database Entry	DHM	Brief descriptions of **cyber security incidents**. Source: Hackmageddon data. Retrieved 2018 from https://www.hackmageddon.com	*wa*	1198 (32,359)	x	
	DDB	Brief descriptions of **data breaches**. Source: Data Breaches visualisation. Retrieved 2018 from https://www.informationisbeautiful.net/visualizations/worlds-biggest-data-breaches-hacks/	*wa*	330 (9,182)	x	
	DCM	Common Attack Pattern Enumeration and Classification dictionary and classification taxonomy (CAPEC [5]), featuring semi-structured descriptions of **security threats**, their respective **countermeasures** and **consequences**.	*wa-e*	3,624 (152,275)	x	
Online Discussion	FNS	Full threads (Question, Answers, and Comments) from Stack Exchange Information Security, an online **developer discussion** forum on **information security**.	*wa*	1,000 (847,954)	Full threads (Question, Answers, and Comments) from the online discussion forum Stack Exchange Politics, SuperUser, Travel, and Stackoverflow	1,200 (1,154,711)

Note. Topics appear in bold font. Research data with no reference is unpublished and used with permission from primary investigator.