Terminology systematization for Cybersecurity domain in Italian Language

Claudia Lanza[1, 2] Béatrice Daille[3]
(1) University of Calabria (UNICAL), DIMES, via Pietro Bucci 87036 Arcavacata di Rende, Cs, ITALY
(2) PhD visiting student at LS2N Laboratory, Université de Nantes, FRANCE
(3) Université de Nantes - LINA CNRS UMR 6241, 2, Rue de la Houssiniere, BP 92208, F44322 Nantes Cedex 3, FRANCE
c.lanza@dimes.unical.it, beatrice.daille@univ-nantes.fr

ABSTRACT

This paper aims at presenting the first steps to improve the quality of the first draft of an Italian thesaurus for Cybersecurity terminology that has been realized for a specific project activity in collaboration with CybersecurityLab at Informatics and Telematics Institute (IIT) of the National Council of Research (CNR) in Italy. In particular, the paper will focus, first, on the terminological knowledge base built to retrieve the most representative candidate terms of Cybersecurity domain in Italian language, giving examples of the main gold standard repositories that have been used to build this semantic tool. Attention will be then given to the methodology and software employed to configure a system of NLP rules to get the desired semantic results and to proceed with the enhancement of the candidate terms selection which are meant to be inserted in the controlled vocabulary.

KEYWORDS : alignment, cybersecurity, textual variants, thesaurus, term extraction

1 Introduction

In order to develop a means for terminological knowledge base, TBK (Condamines, 2018), the construction of the corpus represents the first phase, this is followed by the extraction of candidate terms and then by the construction of a relational system (Barrière, 2006). For what concerns the systems on how to extract terminology, the webpage of the European Parliaments gives a detailed overview of the available resources [1]. Among other studies that have given a state-of-the-art on the different ways to extract information there is the analysis given by Bernier et al. (2012) for the gold standard terms construction within a corpus-based methodology, or by Nazarenko et al. (2009) for the evaluation of different terms extractors tools, or still other studies that proposed different methodology to execute terminological extraction, like Aubin and Hamon (2006) or Loginova (2012) . Two of the features to take into account in constructing a terminological knowledge base are accuracy and precision with respect to the contextual requirements and target genre (Condamines, 2007). Another key element for knowledge databases reliability is their representativeness with respect to a specific field of knowledge (Condamines, 2018). According to Leech (1992), a semantic resource should be considered representative if it contains an adequate level of information related to the domain of reference.
Though a definite measurement or shared standardised calculation of the quantitative level of terms

1. http://termcoord.eu/discover_trashed/free-term-extractors/term-extraction-tools/

a thesaurus should have (Caruso *et al.*, 2014) doesn't officially exist, the idea is to have as much information as possible about the domain of study. A useful method to improve the lexical richness could be that of Pastor and Seghiri (2010) according to which "new words will always be included in the corpus, a point is reached when the addition of more documents will not in practice bring anything new to the collection. In other words, it will not make it more representative because all the different situations, as well as the normal range of terminology in this particular field, have already been covered. ". Still, not always the larger size is a signal of a good baseline for a reliable corpus (Leech, 1991).

2 Resources for the Cyber domain

The main purpose of the research activity related to the construction of a thesaurus for Cybersecurity domain in Italian language was that of providing a web resource able to structure the technical information proper to this area of study from a semantic point of view. In order to achieve this goal, an accurate study of the existent repositories in Cybersecurity domain has been carried out. Cybersecurity is a multidisciplinary field of knowledge. Its terminology is highly linked to the world of Information and Communication Technology (ICT) and its sub-areas [2], but it's also related to the legislative systems and regulations of the reference countries. For its being a combination of different fields of study, the corpus creation followed this orientation and it has been made up of documents both deriving from legal world and from sector-oriented magazines. To build up a source corpus that could include this convergence the gold standards have been gathered in order to start with a systematization of specific information. Indeed, these sources, the gold standards (Terryn *et al.*, 2018), give a reliable content that has to be examined as to offer a comprehensive terminological database that has to be processed to create an accurate knowledge representation.

The main difficulty in designing the source corpus at first was that English sources in this technical domain seemed more numerous than the Italian ones, therefore a process of alignment between these two languages appeared necessary (Daille, 2012). The main gold standards that have been exploited as the starting points to evaluate the results of terms extraction are in English : NIST 7298 (Kisserl, 2013) and ISO 27000 (2016) , and they have been manually translated to Italian. The goal was to have a Italian semantic resource, as well as to enhance the list of terms included in these latter standards related to Cybersecurity since the collaboration with experts made evident that there was a a level of abstractness in the terminology contained in them. Stating that a level of abstractness was present in the lists of terms the standards provide, means that sometimes words like "Activity", "Analysis" or "Objective" didn't, in accordance with experts of Cybersecurity domain, give an high informative additional value to the terminological structure, then not needed for guiding the domain terminology comprehension that is characterized by a marked level of technicisms. The purpose was, indeed, to provide a precise semantic source that could organize in a piloted way the terminology of this technical domain.

Among the Italian resources used to populate the source corpus there are the glossary edited by the Presidency of Ministers "Glossario Intelligence" [3] and the guideline book about Cybersecurity (Baldoni *et al.*, 2018) which has been launched in 2015 during ITASEC [4] conference. If the latter provides a very well-comprehensive book that explains in details the most relevant elements related to Cybersecurity domain giving also examples of the key episodes occurred in years, the first is conceived as a tool that gives a list of the most important administrative and politics-oriented words

2. http://uis.unesco.org/en/glossary-term/ict-related-fields

3. https://www.sicurezzanazionale.gov.it/sisr.nsf/quaderni-di-intelligence/glossario-intelligence.html

4. https://www.itasec.it/2018.html

linked to Cybersecurity field of knowledge. Looking at the English language, instead, there is a large number of databases that deal with Cybersecurity world. Starting from the standards ISO 27000 :2016 and NIST 7298 (Kisserl, 2013), and moving to the list of vulnerabilities and cyberattacks provided by MITRE [5] [6]. It's worth mentioning also NICCS glossary [7] and the vocabulary offered by ENISA [8] . All of them have represented sources that have guided the process of creating a thesaurus, because it's only by knowing which can be the specialized contexts in which the tool will be used that a semantic resource like a thesaurus can check if the terminology obtained by the lexical extraction is accurate (Condamines, 2018).
The main objective of this activity is to realize an image of the domain realizing a semantic tool that could include the knowledge of the field. In order to achieve this first objective, the collaboration with experts proved to be fundamental. It's been thanks to their supervision that the first draft of the thesaurus could have been structured in main categories. Currently this semantic resource is present among the services on the website developed by the CybersecurityLab in Pisa (Italy) [9] ; this online platform was created to orientate the comprehension the Cybersecurity domain, and the thesaurus contains 245 terms, almost all of them with their definition extracted by the corpus documents.
This paper is focused on the way to improve the term extraction accuracy and to extend the first version of the Italian thesaurus for Cybersecurity by having different rules beyond the lexical engineering process.

3 Cyber corpus

Following the words of Bowker (2002), "A corpus is not only a random collection of texts [...]. Rather, the texts in a corpus are selected according to explicit criteria in order to be used as representative sample of a particular language or subset of that language".
In order to have a set of documents that should characterize the source corpus, choosing specific parameters for selecting them it's a very important phase. Linguistics of corpora (Pearson, 1998) suggests to take in consideration documents that are related to present age, better if they share the same time range and are all written in the same language, including produced in a same national area.
The domain of Cybersecurity is a field that undergoes towards continuous changes over time (Condamines, 2018), and its lexicology gets involved too. Having a semantic source that can monitor the terminology of a technical domain can be helpful both for common users and for experts as to follow the lexical internal system and be obliquely updated with terms modifications in time.
For the realization of the Italian thesaurus for Cybersecurity, the hierarchy of sources in law has been followed (Zagrebelsky, 1984). According to this upper level classification of reference sources from which a construction of corpora should start, the first documents to take in consideration related to the legal world, are, firstly, national, then regional and finally local. Subsequently, the corpus has included divulgative documents both in digital and paper-native forms (i.e., sector oriented magazines).
Therefore, for the development of the source corpus for the Italian thesaurus of Cybersecurity, legislative datasets have firstly been taken into account. The main portals on which this kind of information has been retrieved were the European official websites where it's possible to freely download the laws of interest, i.e., EUR-lex, Altalex. Each of these online platforms allows the selection of many filters related to the classification of legislative sources (Penal, Civil, Administrative Code), the time range, the keywords used to search inside the database. Among these documents that have made up the

5. https://capec.mitre.org/
6. https://cve.mitre.org
7. https://niccs.us-cert.gov/about-niccs/
8. https://www.enisa.europa.eu/topics/threat-risk-management/risk-management/current-risk/risk-management-inventory/glossary
9. https://www.cybersecurityosservatorio.it/it/Services/thesaurus.jsp

source corpus there are also technical documents, such as reports or guidelines published by official institutions' portals that are monitoring the Cybersecurity sphere in the Italian field, i.e. CERT [10]. Papers or scientific articles have been excluded at the moment because the terminological texture might have been too discursive and the tone of the sentences too adverbial. Cybersecurity is a domain that owns specific technicism, in this way the first selection of documents has been structured, and gradually a list of the main legislative Italian documents related to Cybersecurity over the latest years has been set up. The selection of the texts to be inserted in the collection of document that would have made up the source corpus has been mainly based on the retrieval of the most important laws, norms, legislations, regulations about the domain of Cybersecurity. The importance of certain documents instead of others has been given by some filters related to :

— time range ; documents taken into account were ranged around the latest years, minimum the latest seven ;
— language : only Italian documents have been considered for the analysis ;
— contexts : national, european and regional laws have been analysed.

Nevertheless, given that the scope of the thesaurus for Cybersercurity was to include as much information as possible on this domain, legislative documents, which constituted a relevant part of the corpus, seemed not enough powerful to cover the knowledge specificity domain's spectrum. It is for this latter reason that also sector-oriented magazines have been employed for the data population in the corpus to be processed. Table 1 summarizes the size of the Italian Cybersecurity corpus by the number of documents and the number of words that have been used for the building of the Italian thesaurus of Cybersecurity. The overal size of the corpus is **563** documents with **11 806 558** terms contained in them.

	Number of documents	**Number of terms**
Laws	249	8 918 209
Magazines	315	7 640 732

TABLE 1 – Features of the Italian corpus of Cybersecurity

4 Methodology

For the term extraction process the starting reference schematization has been that contained in the thesaurus whose terms and semantic relations have been evaluated and validated during several months of collaboration with the experts of this domain. The first draft of the Italian thesaurus for Cybersecurity has as its peculiarity that of having been written in Italian official language, and this trait could be considered as a innovative investigation on the development of a semantic resource for organizing the information on Cybersecurity field of knowledge. Terms which are currently present in the Italian thesaurus are partly represented by the most frequent terms given by T2K [11] term extraction and partly structured according to the information progressively given by experts during the important phase of direct collaboration. Up to now, the first comparative structure to check the granularity obtained with TermSuite [12] is the first draft of the Italian thesaurus for Cybersecurity which had been already evaluated from the experts. The terminological process started at the beginning by selecting the most frequent terms, according to TF/IDF formula, among the list of terms given by the semi-automatic software extractor T2K. These terms have gone through a careful validation process with the experts of the domain with whom a selection of the preferred terms has been settled down.

10. https://www.certnazionale.it/
11. http://www.italianlp.it/demo/t2k-text-to-knowledge/
12. http://termsuite.github.io/

This paper aims at describing the further steps that anticipate the improvement of the current semantic structure of the thesaurus. The ongoing goal is, hence, to enhance the current thesaurus structure, made up of 245 terms, through the variations detected by TermSuite in the source corpus and with Pke library (Boudin, 2016) for clustering the keyphrases. In the seek of aligning the information included in the thesaurus for Cybersecurity, the passages hereafter described will present :

1. The mapping procedure executed between the terms of the thesaurus and the standards of Cybersecurity, i.e. Nist and ISO 27000 :2016, showing how many of them appear in these reference resources and why up to now several of them have not been taken into account ;
2. The terminological extraction by exploiting the semi-automatic tools :
   — **T2K** will be the first to be described as the first that has been used to extract the first terminological dataset ;
   — **TermSuite** will be presented as a software that helped in enhancing the structuring of the semantic relations by using variants ;
   — **Pke** as the library that with its models, TopicRank and Multipartite Rank, supported the definition of topical information about the domain. In detail, for what concerns Multipartite Rank, the first executions have been tested on the main four macro-categories included in the Italian thesaurus for Cybersecurity that have been validated by the group of experts : *Cybersecurity*, *Cybercriminality*, *Cyberbullism*, *Cyber Defence*.

4.1 Mapping

The process of checking the candidate terms has been firstly executed by mapping the term records obtained with T2K with the words contained in the list of NIST and ISO 27000 :2016 vocabularies that can best suite the needs of the group of experts who played a relevant role for the construction of the Italian thesaurus. The Italian thesaurus for Cybersecurity also underwent through a process of validation in comparison with the main gold standards, i.e. NIST and ISO 27000 :2016, the terminological data set inside it can be reasonably conceived as a reliable structure with which juxtapose the results from the terms candidates extraction software. The lists of terms contained in the standards NIST and ISO 27000 :2016 have been manually translated with the support of terms databases in IATE system [13] in order to retrieve the information using the same language in input as we worked for the construction of the thesaurus on Cybersecurity in Italian. Through the mapping with these standards, the check was aimed to verify the presence or not of the selected terms given by T2K extraction. The further passage dealt with coverage of the terms contained in the corpus used to build the semantic resource with respect to the standards, as to demonstrate how the terminology given by the texts gathered in the corpus could be overlapped with the list of terms in the standards.Table 2 summarises the scores given by the comparison. As stated in section 2, several general words like "Analysis" have been put aside for the moment because too vague for the purposes of providing a semantic resource to structure the technicisms of the domain.

	Nist Match	**ISO match**
Thesaurus terms	75/1299	17/88

TABLE 2 – Matches between the terms contained in the thesaurus and the ones contained in the Standards for Information Security.

13. https://iate.europa.eu/home

4.2 Term Extraction software

Once defined the documents that should have made up the source corpus, the candidate term extraction begun. For this phase the decision was to use several tools that presented different features. For instance, two of the principal software that have been exploited to retrieve the terminology are Text To Knowledge (T2K) (Dell'Orletta *et al.*, 2014) and TermSuite (Cram & Daille, 2016).
This activity refers to an ongoing process of selection of the more adequate tool from which terminology on this domain has to be taken into consideration. For the first development of the thesaurus in Italian language for Cybersecurity, T2K has been used, but, as it will be better specified in the nexts paragraphs, some of its internal restrictions limited the visualization of different types of variants. Nonetheless, for the purposes of the first draft of the Thesaurus, this term extractor software proved to give reliable results that have been approved by the consensus of the expert on this specified field. The selection of terms amongst the ones given in output by T2K has been based both on the first scores sorted according to TF/IDF formula, and on the schemas owned by the experts to frame the main Cybersecurity concepts, i.e. the head term Cyber accompanied by several key concepts of this field of study, such as *"hygiene", "espionage", "threat"*. In order to have a better systematization of the Italian thesaurus for Cybersecurity to be finalized also in agreement with the community of experts, another resource that is giving interesting results is Pke (Boudin, 2016). This library dedicated to keyphrases extraction gave as output a good grouping structuring of the information contained in the source corpus that can guide the orientation of certain macro-categories inserted or to be inserted in the thesaurus with other concepts with which they are related.
While there are some fundamental differences between term extraction and keyword extraction approaches (Lossio-Ventura *et al.*, 2014), we wanted to test their faculty to propose other types of terms selected according to their keyness properties at the document level.

4.2.1 T2K

T2K (Dell'Orletta *et al.*, 2014) is native Italian extractor, has been created by the Italian Computational Linguistics Group at CNR in Pisa. In addition to extract domain-specific terminology and phrases, it organizes and structures the set of extracted terms and phrases into taxonomical chains, provides PoS tagging for the corpus as well as a dependency parsing overview.
Its main advantages are pretty much related to the fact that is a software developed by Italian spoken experts, so the rules beyond the system result to be specifically orientated to the source language of the thesaurus for Cybersecurity. Another benefit is that it can apply a contrast function with another reference vocabulary or a list already processed, as to make two terminilogical sets comparable for statistical analysis.
For what concerns one of its cons, even though it's a corpus oriented extraction tool, it doesn't present a very wide customization intervention in the grammar rules. However, from the beginning it's possible to decide which kind of lexical chains is the extraction meant to provide, as well as the threshold frequency range and the length of terms.
The candidate terms extraction resulted to be quite different in using the two term extractor software. The results are presented in Table 3. In detail, given the limited range of modifications T2K allows to do, the total numbers of entries (single terms and MTW) was 446 325, a very high number that reveals how many noise can be present among the records. That is probably linked to the fact that T2K doesn't show the possibility to insert a personalised stopword list or a set of variation rules like TermSuite. The outnumbered results in T2K have been obtained by applying the following configuration structure :

1. the frequency has been set up to level 1 ;
2. the semantic chains have been defined as Common Nouns (S) followed by an Adjective (A),

	T2K	TermSuite
number of candidates	446 325	16 641

TABLE 3 – Terminology extraction : number of candidate terms extracted by each tool.

or/with a a Preposition (E), Articulated Preposition (EA), Common Noun (S), and ending with a Common Noun (S) or an Adjective (A) ;
3. Maximum length terms has been fixed at 8.

In TermSuite, the overall results on the same corpus, made up of almost 563 documents, resulted to be 16 641, with a discrepancy of 429 684 units. Even though the results have been lower than T2K, the application of Italian rules beyond the system, the customization of stopword lists, the configuration of prefix banks and derivational datasets, allowed to obtain more sophisticated results.

4.2.2 TermSuite

TermSuite (Cram & Daille, 2016) is not an Italian native extractor but, as it is multilingually designed, it could be extended or adapted to another language. TermSuite computes the termhood and the unithood of a term candidate as defined by Kageura and Umino (1996). It adopts the two core steps of the terminology extraction process :

1. Identification and collection of term-like units in the texts, mostly a subset of nominal phrases ;
2. Filtering of the extracted term-like units that may not be terms, syntactically or terminologically ; sorting of the candidate terms according to their unithood, their terminological degree and their most usefulness for the target application.

Among its main features, the variants recognition, that could help for relation detection, proved to be very helpful in terms of enhancing the first structure of the thesaurus, because acting on the linguistics rules the output can be more accurate and customized. Other features beside the patterns recognition, are for example the definition of different typologies of frequency (i.e., Document Frequency, TF-IDF, General Frequency Norm).

4.2.3 TermSuite customization for Italian

One of the main advantages of TermSuite has been that of creating customized patterns that enabled the extraction of desired lexical outputs.
The morphology package was the one that has been mostly edited, starting from the prefixes and suffixes, acting on the derivation bank in order to adjust the rules to the Italian ones. The number of grammar rules in Italian does not differ from the French ones. For the purposes of the ongoing enhancing systematization for the Cybersecuirity Italian thesaurus, the variation rules made in French language appeared strictly similar to what might have been the Italian ones. A core grammar of variants is instantiated for five languages including French and Spanish languages which gather morphological and syntagmatic rules and three main categories of variants : denominative, conceptual and linguistic variants (Daille, 2017). Table 4 summarises the number of rules by language of the default TermSuite Grammar.
To enhance the terminological extraction for the Italian thesaurus for Cybersecurity, a structuring of Italian variation rules has been arranged. Therefore, in the extraction given by TermSuite, to obtain different form of variations, syntactic as well as morphological, an alignment with the Italian grammar rules proved to be necessary. Restructuring the patterns for Italian meant the specification of the most common semantic chains the Italian language owns, the base-terms (Daille, 2003). Indeed, in Italian language the main basic syntactic structures are made up of : **Noun Adjective** *sicurezza informatica*

		Spanish	French
Denominative variants	Morphological	2	2
	Syntagmatic	4	5
Conceptual variants	Morphological	2	4
	Syntagmatic	17	13
Linguistic variants	Morphological	0	2
	Syntagmatic	15	11
Total	Morphological	4	8
	Syntagmatic	36	29
Number of rules		40	37

TABLE 4 – TermSuite grammar : number of rules for Spanish and French, by structure type, and by morphological or syntagmatic nature.

(cybersecurity), **Noun(Prep(Det))Noun** *titolare del trattamento* (treatment manager)
Each of these basic patterns of base-terms can be modified by using the head part of these latter and clustering the occurrences found in the source texts.
To give some examples of how the variations related to Cybersecurity domain in Italian language appear after having activated the variation rules in TermSuite, we observed the three types of variants :

1. **Denominative variants**
   — Composition : npn : hacker del telefono (phone hacker) → na : hacker telefonico
   — Lexical reduction : npna : indirizzo di posta elettronico (mail address) → na : indirizzo elettronico (electronic address)
2. **Conceptual variants**
   — Derivation : na : contenuto legale (legal content) → na : contenuto illegale (illegal content)
   — Expansion : na : fascicolo sanitario (personal health record) → naa : fascicolo sanitario elettronico (electronic personal health record)
3. **Linguistic variants**
   — Graphical segmentation : cyber-security ↔ cybersecurity
   — Coordination : npn : sicurezza della rete (network security) → npncpn : sicurezza della rete e del sistema (network and system security)

Denominative variants in the thesaurus shall be considered as units to be incorporated as synonymous variants of terms which have been approved by the experts of the domain. Having a list of terms accompanied by the synonymy form is a very helpful base from which begin to structure the network system of the semantic relationships in the thesaurus, as well as knowing the conceptual variants that, by expanding the terms output through the insertion of words with which they appear in the source corpus, can provide a better frame of the information inside the documents of the corpus as to decide which can be the best to be inserted in the thesaurus. For what regards the linguistic variants, such as coordination or the graphical segmentation, all represent a valid decision-making system to recognize which can be, among the entries given by a candidate term extraction, the preferred ones approved by the experts communities.

4.2.4 Pke

Another means that has been tested is the library Pke (Boudin, 2016). Pke library is purely statistical, so it's by nature multilingual and is used for keywords extraction. It is not purely term extraction, it's a document oriented extraction. It was used to extract information about the topic terms, single

or multi-word, that characterised the documents contained in the corpus as to use them in order to check the thesaurus structure of semantic relationships. Pke algorithms have two steps : a keyphrase candidate extraction step followed by a refinement step. Sequences of adjacent words, restricted to nouns and adjectives, are considered as keyphrase candidates. Refinement step orders the keyphrase candidates according to various statistical or graph-based methods.Indeed, this means to retrieve keywords from a source corpus proved to be very useful in detecting the main semantic areas.The following are example taken from the output given by executing TopicRank algorithm, where the first are the main topics retrieved in the corpus alongwith their weight size :'sistema informatico' (*informative system*), 'reato'(*crime*),'dati(*data*),'accesso illegittimo'(*illegitimate access*), 'accesso abusivo'(*abusive access*),'rete'(textitnetwork), 'sito'(*website*),'server'(*server*).
We then tried to perform the Multiplartite Rank (Boudin, 2018) and the results proved to be better for the purposes of helping in structuring the categories present in the current thesaurus. The following are examples of the several clusters sorted by relevance given by this library that could depict how the terms can be correlated using the keyphrases extraction logic. These representative cases are the results given by searching the way the main four macro-categories of the current structure of the thesaurus, i.e. *Cybesecurity*, *Cybercriminality*, *Cyberbullism*, *Cyber Defence*, are organized by this model :

1. **Cybersecurity :**
   — From a technical document : sicurezza (*security*), sistema (*system*), attacchi cibernetici (*cyber attacks*), dati personali (*personal data*), cybersecurity (*cybersecurity*)
   — From a divulgative document : here it's intersting to note that the term is written in a separated way, i.e. Cyber security, that could be a signal that the divulgative terminology is different from the technical one :
   cyber security (*cyber security*), sicurezza (*security*), reti (*networks*), informazioni (*information*), macchine intelligenti (*smart cars*),
2. **Cybercriminality :**
   — From a legal document : cibercriminalità (*cybercriminality*), informazione (*information*), lotta (*fight*), reti (*networks*), reato informatico (*cybercrime*)
   — From a divulgative document : informazioni (*information*), sicurezza (*security*), criminalità informatica (*cyber criminality*), comunicazione (*communication*), reti (*networks*)
3. **Cyberbullism :**
   — From a divulgative document : "cyberbullism" gives just one occurrence prevenzione (*prevention*), cyberbullismo (*cyberbullism*), sistema educativo (*educational system*), scuole (*school*), formazione (*education*),
4. **Cyber Defence :**
   — From a divulgative document : attacchi (*attacks*), attacchi cibernetici (*cyber attacks*), minacce (*threats*), difesa cibernetica (*cyber defence*), reti *networks*)
   — From a legal document : informazioni on-line (*on-line information*), sicurezza (*security*), attacchi informatici (*cyber attacks*), difesa informatica (*cyber defence*), difesa (*defence*)

From these examples we can understand that *Cyber defence* is related to *Cyber attacks* which is in turn connected with *Threats*, or that *Cyber bullism* is contextualized with the educational system and schools. This schematization helped in structuring the information.

5 Candidate terms observation

To give an example of how different perform the two software, the following are some of the terms present in the current thesaurus, which, for instance, have already undergone through a validation

process by the experts of Cybersecurity domain, with their source extraction in T2K compared with Termsuite. Sometimes, as the first example, TermSuite proved to provide a more informative result, other times, T2K gave more outputs for one entry even not grouping it in a variation cluster, or, on the contrary, it happened also that TermSuite presented a result that was not present inside the T2K extracted terminology , but was very important for the experts of the domain, as the third example shows. Finally an example of a new term provided by the Pke clustering.

1. **Thesaurus term : "Attacchi a forza bruta"**(*BruteForce Attack*), which is the more specific term of "Cyber attacks mechanisms" :
   — **TermSuite** a visualization by variation : Term na : attacco bruto → Denominative variants : npna : attacco a forza bruto and npna : attacco di forza bruto ;
   — **T2K :** Prototypical form : attacco a forza bruta, Lemma of term :attacco a forza bruto ;
   — **Pke :** this term is not specifically present, but if we look for *Attack* we have different score according to if its a cluster of ten keyphrases in a legal document or in a sector-oriented issue of the magazines :
      (a) Legal document :Sicurezza (*security*),sistema (*system*), attacchi cibernetici (*cyber attacks*), dati personali (*personal data*), cybersecurity (*cybersecurity*), attacchi (*attacks*), rete (*network*), protezione (*protection*) ;
      (b) Divulgative document : siti(*website*), http (*http*), pagina (*page*), attacco (*attack*), sistema (*system*), posta elettronica (*e-mail*), etica hacker (*hacker ethics*), virus (*virus*) ;
2. **Thesaurus term : "Phishing"**, which is a more specific term of "Spam", is given in two different outputs :
   — **Termsuite :** n : phishing and npn : messaggio di phishing (*phishing message*) ;
   — **T2K :** the terms is under 176 terminological units, among these latter the following are the first five representative results of the extraction. Prototypical forms : tecniche di phishing (*phishing techinques*), mail di phishing (*phishing mail*), spear phishing, attacchi di phishing (*phishing attacks*), siti di phishing (*phishing websites*) ;
   — **Pke :** not present
3. **Thesaurus term : "Cavalli di troia"**(*Trojan Horses*), which is a more specific term of Dipendent Software, that are, in turn, more specific terms of Malicious Software :
   — **T2k :** not present
   — **TermSuite :** TermSuite retrieved both "Trojan" alone and "Trojan-horses", which is more precise and accurate for the purposes of the thesaurus on Cybersecurity ;
   — **Pke :** not present

Pke new term : threat intelligence whose clusters are : threat intelligence (*threat intelligence*), controspionaggio informatico (*cyber counter-espionage*), sistemi informativi (*informative systems*), azioni (*actions*), intelligence (*intelligence*).

6 Conclusion

Enhancing the thesaurus terminological structure is a an ongoing perspective to be progressively achieved. TermSuite extraction helped in obtaining more accurate results in terms of variants : having a list of different ways to represent a terms results to be a valid supporting system to decide, alongside the consensus of experts, the best entries with reference to accuracy and synonyms. Finally, Pke library with Multipartite Rank gives several terminological groups which can help in detecting the semantic interrelations among terms and work in future on the construction of thesaurus basic relationships by employing its collection reasoning. as well as the inclusion of new terms given by the keyphrases.

Références

AUBIN S. & HAMON T. (2006). Improving term extraction with terminological resources. In T. SALAKOSKI, F. GINTER, S. PYYSALO & T. PAHIKKALA, Eds., *Advances in Natural Language Processing (5th International Conference on NLP, FinTAL 2006)*, number 4139 in LNAI, p. 380–387 : Springer.

BALDONI R., DE NICOLA R. & PRINETTO P. (2018). *Il Futuro della Cybersecurity in Italia : Ambiti Progettuali Strategici Progetti e Azioni per difendere al meglio il Paese dagli attacchi informatici*. Laboratorio Nazionale di Cybersecurity (CINI) - Consorzio Interuniversitario Nazionale per l'Informatica.

BARRIÈRE C. (2006). Semi-automatic corpus construction from informative texts. In L. BOWKES, Ed., *Text-Based Studies in honour of Ingrid Meyer*, Lexicography, Terminology and Translation, chapter 5. University of Ottawa Press.

BERNIER-COLBORNE G. . (2012). Defining a gold standard for the evaluation of term extractors. In *in Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, p. 15–18.

BOUDIN F. (2016). pke : an open source python-based keyphrase extraction toolkit. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics : System Demonstrations*, p. 69–73, Osaka, Japan : The COLING 2016 Organizing Committee.

BOUDIN F. (2018). Unsupervised keyphrase extraction with multipartite graphs. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 2 (Short Papers)*, p. 667–672, New Orleans, Louisiana : Association for Computational Linguistics.

BOWKER L. & PEARSON J. (2002). *Working with Specialized Language : A Practical Guide to Using Corpora*. London/New York : Routledge.

CARUSO A., FOLINO A., PARISI F. & TRUNFIO R. (2014). A statistical method for minimum corpus size determination. In *12es Journées internationales d'Analyse statistique des Données Textuelles (JADT2014)*, Paris, France.

CONDAMINES A. (2007). L'interprétation en sémantique de corpus : le cas de la construction de terminologies. *Revue française de linguistique appliquée*, **Vol. XII**(2007/1), 39–52.

CONDAMINES A. (2018). Terminological knowledge bases from texts to terms, from terms to texts. In *The Routledge Handbook of Lexicography*. Routledge.

CRAM D. & DAILLE B. (2016). Terminology extraction with term variant detection. In *Proceedings of ACL-2016 System Demonstrations*, p. 13–18, Berlin, Germany : Association for Computational Linguistics.

DAILLE B. (2003). Conceptual structuring through term variations. In F. BOND, A. KORHONEN, D. MACCARTHY & A. VILLACICENCIO, Eds., *Proceedings ACL 2003 Workshop on Multiword Expressions : Analysis, Acquisition and Treatment*, p. 9–16 : ACL.

DAILLE B. (2012). Building bilingual terminologies from comparable corpora : The ttc termsuite.

DAILLE B. (2017). *Term Variation in Specialised Corpora : Characterisation, automatic discovery and applications*, volume 19 of *Terminology and Lexicography Research and Practice*. John Benjamins.

DELL'ORLETTA F., VENTURI G., CIMINO A. & MONTEMAGNI S. (2014). T2K : a system for automatically extracting and organizing knowledge from texts. In N. C. C. CHAIR), K. CHOUKRI, T. DECLERCK, H. LOFTSSON, B. MAEGAARD, J. MARIANI, A. MORENO, J. ODIJK & S. PIPERIDIS, Eds., *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland : European Language Resources Association (ELRA).

ISO/IEC 27000 (2016). *Information technology – Security techniques – Information security management systems – Overview and vocabulary*. International Standard.

KAGEURA K. & UMINO B. (1996). Methods of automatic term recognition : a review. *Terminology*, **3**(2), 259–289.

KISSERL R. (2013). *Glossary of Key Information Security Terms*. National Institute of Standards and Technology. NISTIR 7298 Revision 2.

LEECH G. (1991). *The state of the art in corpus linguistics*. London : Longman.

LEECH G. (1992). Corpora and theories of linguistics performance. In J. S. (ED.), Ed., *Directions in Corpus Linguistics : Proceedings of Nobel Symposium*, p. 105–122, Berlin : Mouton de Gruyter.

LOGINOVA CLOUET E., GOJUN A., BLANCAFORT H., GUEGAN M., GORNOSTAY T. & HEID U. (2012). Reference Lists for the Evaluation of Term Extraction Tools. In *Terminology and Knowledge Engineering Conference (TKE)*, Madrid, Spain.

LOSSIO-VENTURA J. A., JONQUET C., ROCHE M. & TEISSEIRE M. (2014). Towards a mixed approach to extract biomedical terms from text corpus. *IJKDB*, **4**(1), 1–15.

NAZARENKO A., ZARGAYOUNA, H. ; HAMON O. & VAN PUYMBROUCK (2009). Evaluation des outils terminologiques : enjeux, difficultés et propositions. *Traitement Automatique de la Langue (TAL)*, **50**(1), 257–281.

PASTOR G. C. & SEGHIRI M. (2010). Size matters : A quantitative approach to corpus representativeness. *Language, translation, reception. To honor Julio César Santoyo*, p. 1–35.

PEARSON J. (1998). *Terms in Context*. John Benjamins, Amsterdam.

TERRYN A. R., HOSTE V. & LEFEVER E. (2018). A Gold Standard for Multilingual Automatic Term Extraction from Comparable Corpora : Term Structure and Translation Equivalents. In N. C. C. CHAIR), K. CHOUKRI, C. CIERI, T. DECLERCK, S. GOGGI, K. HASIDA, H. ISAHARA, B. MAEGAARD, J. MARIANI, H. MAZO, A. MORENO, J. ODIJK, S. PIPERIDIS & T. TOKUNAGA, Eds., *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan : European Language Resources Association (ELRA).

ZAGREBELSKY G. (1984). *Il sistema costituzionale delle fonti del diritto*. Turin : UTET.