TERMINOLOGY SYSTEMATIZATION FOR CYBERSECURITY DOMAIN IN ITALIAN LANGUAGE

CLAUDIA LANZA - PHD STUDENT, UNIVERSITY OF CALABRIA AND VISITING AT UNIVERSITÉ DE NANTES C.LANZA@DIMES.UNICAL.IT
BÉATRICE DAILLE – FULL PROFESSOR UNIVERSITÉ DE NANTES

TIA 2019
Terminology and Artificial Intelligence TALN-RECITAL and IC-PFIA 2019
July 1st, 2019. Toulouse. France
OVERVIEW

- Introduction
- Objectives
- Methodology outline
- Domain
- Corpus
- First phase:
  - Mapping with standards
  - Experts of the domain
  - First **Italian** thesaurus draft for Cybersecurity
- Second phase:
  - Term extraction software comparison
  - Candidate terms observation
- Future works
# INTRODUCTION

| Starting elements                      | Problems                                                                 |
|----------------------------------------|--------------------------------------------------------------------------|
| Language of the thesaurus: Italian     | Standardization of the terminology into Italian language (few standards in Italian for Cybersecurity) |
| Terminology extraction tools in Italian| Weak level of granularity given by the terms extraction tools for hierarchy and synonymy detection |
| Support of the experts of the domain   | Contrasting perspective in the adjustment of the terminological assets    |
Some terms are given by the English Standards and by the Italian official glossaries in the English form and they are not translatable in Italian or they have become common use terminology.

- Phishing
- Spam
- Smishing
- Cyber trolling

**Translation from Standards EN-ITA**

**CYBER-DEFENCE**

**CYBER INTELLIGENCE**
OBJECTIVES

- Realization of the Italian thesaurus of Cybersecurity;
- Portray a comparison between the software employed to configure a system of NLP rules to proceed with the enhancement of the candidate terms selection;
- Set out a methodology that could automatize, starting from the source corpus, the semantic updating for this field of knowledge by adjusting the terminological asset of the thesaurus in order to achieve an adequate coverage of Cybersecurity domain.
METHODOLOGY OUTLINE

Document selection and analysis

Terminological extraction from documents (T2K)

First terminological list datasets filtered by frequency (TF/IDF) and accuracy

Mapping with the ICT Security vocabularies contained in the standards

Establishment of the semantic relationships between terms

Collaboration with the experts of the Cybersecurity domain

Creation of the thesaurus as a means of semantic control

Software comparison: T2K, TermSuite, Pke

Terms observation

Claudia Lanza - PhD student, University of Calabria and visiting at Université de Nantes
c.lanza@dimes.unical.it
- Multisciplinarity: ICT and sub-areas (Audiovisual techniques; Computer software; Electronics; etc)
- Specificity: technicisms and standardized terms;
- Cross-field: computer science field, legislative systems, regulations.
CORPUS

The overall 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

Criteria of selection:

— **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.
FIRST (AND SECOND) PHASE MAPPING WITH STANDARDS

Table 2 – Matches between the terms contained in the thesaurus and the ones contained in the Standards for Information Security.
Gestione del rischio cyber

Broader Terms
- Cybersecurity

More specific terms
- Monitoraggio del rischio cyber
- Piani di emergenza in caso di incidenti informatici
- SCAP
- Piano di risposta al rischio cyber
- Mitigazione del rischio cyber
- Valutazione del rischio cyber

Related terms
- Meccanismi di sicurezza informatica
- Rischio residuo

Not a of ambito (SN)

Processo di gestione dei rischi per le operazioni organizzative (compresa missione, funzioni, immagine, reputazione), asset organizzativi, individui, altre organizzazioni e il Paese, che è risultato dall'attività di un sistema informativo, e include: (i) la valutazione della gestione del rischio; (ii) l'implementazione della strategia di mitigazione del rischio; (iii) l'utilizzo di tecniche e procedure per controllare in maniera continuativa lo stato di sicurezza del sistema informativo.

SOURCE: SP 800-53; SP 800-53A; SP 800-37

Processo di gestione dei rischi per le operazioni organizzative (comprese missione, funzioni, immagine, reputazione), asset organizzativi, individui, altre organizzazioni e il Paese, che è risultato dall'attività di un sistema informativo. Esso include la valutazione del rischio, analisi dei costi e benefici; la selezione, l'implementazione e la valutazione dei controlli di sicurezza; e l'autorizzazione formale a utilizzare il sistema. Il processo prende in considerazione l'efficacia, l'efficienza, e i limiti dovuti alle leggi, normative, politiche e regolamenti.

SOURCE: SP 800-82; SP 800-34

https://www.cybersecurityosservatorio.it/it/Services/thesaurus.jsp
FIRST PHASE:
EXPERTS OF THE DOMAIN

Informatics and Telematics Institute of CNR in Pisa
(Tuscany)
4 Main categories decided alongside with the experts in a abstraction information process according to their presence in the Cybersecurity taxonomies included in the gold standards and their frequency in the terminological lists:

_Cybersecurity, Cybercriminality, Cyberbullism, Cyber Defence._

- 245 candidate terms;
- Semantic Relationships decided according to the Head-term based derived with T2K and the approval by the experts;
- Scope notes (Definitions of the terms) added in accordance to the co-occurrences in the official sources of the corpus.
Cybersecurity
Non preferred terms

UE: Sicurezza informatica

More specific terms

SC1: Cyber minacce
SC1: Dati
SC1: Gestione del rischio cyber
SC1: Meccanismi di attacchi informatici
SC1: Meccanismi di sicurezza informatica
SC1: Proprietà di sicurezza
SC1: Vulnerabilità

Related terms

RT: Cyber defence
RT: Cyber forensics
RT: Cyber guerra
RT: Cyber laundering
RT: Cyber molestie
RT: Cyber terrorismo
RT: Cyber theft
RT: Cyber Threat Intelligence
RT: Cyber-hygiene
RT: Cyberbullismo
RT: Cybercriminalità
RT: Cyberspionaggio
RT: Framework nazionale della Cybersecurity
RT: Hacker
RT: Hacking
RT: Privacy
RT: Processi di autenticazione
RT: Software

Cybercriminalità
More specific terms

SC1: Cifrativaute
SC1: Cyber crime
SC1: Cyber guerra
SC1: Cyber laundering
SC1: Cyber minacce
SC1: Cyber terrorismo
SC1: Cyberspionaggio

Related terms

RT: Cyber security
RT: Dark web
RT: Hacking
RT: Meccanismi di attacchi informatici

Cyberbullismo
More specific terms

SC1: Cyber gang
SC1: Cyber harrassment
SC1: Cyber molestie
SC1: Cyber stalking
SC1: Cyber trolling

Related terms

RT: Cyber bulli
RT: Cyber security
RT: Dark web

Cyber defence
More specific terms

SC1: Cyber guerra
SC1: Cyber terrorismo
SC1: Cyber-hygiene
SC1: Cyberspionaggio
SC1: Intercettazioni

Related terms

RT: Cyber forensics
RT: Cyber security

https://www.cybersecurityosservatorio.it/it/Services/thesaurus.jsp
Claudia Lanza – PhD student, University of Calabria and visiting at Université de Nantes c.lanza@dimes.unical.it TIA 2019 – 1 July
SECOND PHASE
SOFTWARE COMPARISON

- **T2K** has been used to extract the first terminological dataset (*Italian native tool*);
- **TermSuite** a software that helped in enhancing the structuring of the semantic relations by using variants (*denominative, conceptual, linguistic*);
- **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.

| Source       | Cybersecurity          | Cybercriminality       | Cyberbullism   | Cyber Defence                        |
|--------------|------------------------|------------------------|----------------|-------------------------------------|
| **Legal/technical** | Security; Cyber attacks; personal data; system | Information; fight; networks; cybercrime | / | On-line information; security; cyber attacks; defence |
| **Divulgative** | **Cyber security**; security; networks; smart cars | Information; security; cyber criminality; networks | Prevention; educational system; schools: education | Attacks; cyber attacks; threats; networks |
## CANDIDATE TERMS OBSERVATION

| Term                | T2K                  | TermSuite              | Pke                      |
|---------------------|----------------------|------------------------|--------------------------|
| **Brute Force Attack** | **2 occurrences**   | **1 occurrence**       | Not present              |
| Terms: attacco a forza bruto; forza bruta          | Term: attacco bruto  | But see for Term: attack |
|                     | Denominative variants: | Legal document: security; systems; cyber attacks; cybersecurity ;personal data |
|                     | - npna: attacco a forza bruto | Divulgative document: website; system; e-mail; virus; hacker ethics |
|                     | - npna: attacco di forza bruto |                               |
| **Phishing**        | **176 occurrences** | **2 occurrences**       | Not present              |
| Terms: phishing techniques, phishing mail, spear phishing, phishing attacks, phishing website | Term: phishing         |                               |
|                     | Conceptual variant( expansion): |                               |
|                     | - npn: messaggio di phishing |                               |
| **Trojan Horses**   | Not present          | **1 occurrence**       | Not present              |
|                     | **X**                |                         |                         |

New info from PKE

**Threat Intelligence** clustered as Cyber counter-espionage; informative systems, actions, intelligence
FUTURE WORKS

- Enhancement of categories organization through Pke;
- Detection of semantic relationships starting by a corpora processing;
- Checking of the terminological coverage with respect to terms variation through time
THANK YOU FOR YOUR ATTENTION