# Intelligence Linguistica: L'analisi (meta)cognitiva del linguaggio naturale per il rilevamento di cyber threats

Alice Felli

Società Italiana di Intelligence – SOCINT Press

Società Italiana di Intelligence
SOCINT Press
c/o Università della Calabria
Cubo 18-b, 7° piano
Via Pietro Bucci
87036 Arcavacata di Rende (CS) – Italia
https://www.socint.org
ISBN 979-12-80111-48-7

UNIVERSITÀ DEGLI STUDI DI ROMA
"TOR VERGATA"
MACROAREA DI LETTERE E FILOSOFIA

CORSO DI LAUREA IN LINGUA E CULTURA ITALIANA
A STRANIERI PER L'ACCOGLIENZA E
L'INTERNAZIONALIZZAZIONE

**TESI DI LAUREA MAGISTRALE**
IN LINGUISTICA GENERALE

# *Intelligence linguistica: l'analisi (meta)cognitiva del linguaggio naturale per il rilevamento di cyber threats*

**Relatrice**
Prof.ssa Francesca Dragotto

**Correlatore**
Dott. Emanuele Galtieri

**Laureanda**
Dott.ssa Alice Felli
mat. n° 0297497

Anno Accademico
2021/2022

* * *

*«Non Scholae, sed Vitae discimus»*
*– Non impariamo per la scuola, ma per la Vita.*
*(Seneca)*

* * *

# INDICE

# INTRODUZIONE

Il presente lavoro di tesi di Laurea Magistrale mira a descrivere l'importanza dell'analisi linguistica nelle attività di *Cyber Intelligence:* la raccolta di dati[1] linguistici, la loro selezione e l'estrazione di informazioni rappresentano operazioni strategiche da svolgere con la massima accuratezza, riservatezza e attenzione ai dettagli.

Il mio interesse in materia nasce dalla mia grande passione per il mondo dell'Intelligence e si realizza nella mia posizione come analista linguistica presso una delle società-pilastro operante nel settore della *Cyber Security* e della *Cyber Intelligence.*

Mi occupo di Intelligence linguistica, attraverso la realizzazione di un servizio multilingue di ricerca e analisi delle informazioni, recepite tramite fonti aperte[2] o *database* locali, sotto forma di dati strutturati o non strutturati, in ambito *Corporate* e Governativo.

L'Intelligence linguistica ambisce a districare le reti della (meta)cognizione umana per elaborare, attraverso sofisticati algoritmi, quello che ad oggi rimane il sistema più complesso da decriptare: il pensiero dell'essere umano.

Informatica e linguistica confluiscono in maniera strategica e funzionale grazie a tecnologie all'avanguardia nell'ambito della comprensione e dell'elaborazione del linguaggio umano *(Natural Language Processing and Understanding).*

In questo elaborato, sottolineerò l'importanza della matrice cognitiva nei processi di classificazione delle informazioni: è nella necessità di fruire di *types* e nell'impossibilità di vedersi garantiti dei *tokens* immutabili nel tempo, che ogni persona comprende e concettualizza la propria realtà. Per questo, qualsiasi tipo di filtro cognitivo concepito dall'essere umano è da ritenersi valido al fine di poter realizzare una cyber analisi quanto più esaustiva e puntuale.

I *corpora* testuali (provenienti da fonti OSINT, SOCMINT e/o documenti locali) rappresentano, per gli operatori linguistici, il punto di partenza per la redazione ed

[1] Si definisce "dato" un qualsiasi *«elemento conosciuto, una descrizione elementare di un'entità, di un fenomeno, di una transazione, di un avvenimento».* – A. Tofalo, *Intelligence Collettiva: i dati, l'informazione e il linguaggio*, 2017, Cfr. www.angelotofalo.com/intelligence-collettiva-dati-linformazione-linguaggio/

[2] *«Con il termine "fonte" si indica l'origine della notizia, cioè chi la produce; si definisce "aperta" quella che è accessibile a tutti»* – A. Piras, *Fake news e nuove tecnologie,* ANDIG, 2020, Cfr. https://www.andig.it/saggi/29-fake-news-e-nuove-tecnologie-la-blockchain-puo-realmente-essere-la-nuova-frontiera-della-lotta-alla-disinformazione-in-rete

implementazione di condizioni formali di estrazione e di categorizzazione delle informazioni, ambendo al raggiungimento di *target* predefiniti. La tecnologia OSINT permette di setacciare tutto il web alla ricerca di dati necessari all'analisi.

Le regole atte all'estrazione delle informazioni vengono create mediante linguaggi formali, tra cui le *RegEx*[3] *(Regular Expressions*) e vengono implementate in un *framework* semantico; nello specifico, il *tool* di riferimento per questo lavoro sarà la piattaforma *software* COGITO STUDIO® dell'azienda *Expert.ai* (precedentemente conosciuta come *Expert System)*, di cui dettaglierò le caratteristiche a partire dal terzo capitolo della presente tesi, elencandone i punti di forza, nonché ragionando su migliorie ed eventuali soluzioni a limiti riscontrati. Le condizioni linguistiche realizzate dall'analista devono essere meticolosamente e costantemente validate attraverso operazioni di verifica e raffinamento, ovvero di *fine tuning*, della loro risposta in *output*, al fine di testare la loro efficace applicazione su ampi e diversificati *corpora* testuali. In base ai risultati ottenuti dal processo di elaborazione delle regole, l'operatore valuterà la possibilità di implementare in esse ulteriori condizioni, al fine di potenziare l'addestramento linguistico della macchina.

Il lavoro di analisi linguistica si basa sul processo di riconoscimento, comprensione ed elaborazione del linguaggio naturale da parte di un motore semantico che sfrutta – nel caso specifico di questa tesi magistrale – la sofisticata tecnologia *rule-based* del *Cognitive Computing*, ovvero della computazione cognitiva, la quale necessita della *knowledge* di un operatore linguistico in fase di *input processing*. Ciò consente di combinare i punti di forza dell'A.I. con l'intelligenza umana.

Le tecnologie di analisi linguistica sono in continua evoluzione e i risultati raggiunti, ad oggi, sono sorprendenti: i motori semantici basati sul modello di *Cognitive Computing* sono in grado di ragionare come un vero e proprio "cervello linguistico", presentando capacità di elaborazione dell'informazione mediante reti neurali artificiali equiparabili, o comunque estremamente assimilabili, a quelle umane. Dette reti neurali sono alla base del processo di apprendimento profondo *(Deep Learning)* e consentono alla macchina di processare i dati in entrata seguendo un protocollo di azione investigativa su base cognitiva.

Infine – ma non è un elemento da considerare alla fine del processo linguistico – riveste fondamentale rilievo la componente culturale, intesa sia come *forma mentis* del Cliente che

[3] Nello specifico, utilizzando la sintassi PERL delle espressioni regolari. Cfr. https://www.boost.org/doc/libs/1_53_0/libs/regex/doc/html/boost_regex/syntax/perl_syntax.html

richiede il servizio di analisi, nonché come piano culturale in materia di geopolitica – componente fondamentale per la progettazione di un albero tassonomico, ovvero di una categorizzazione concettuale gerarchizzata del contesto da elaborare. L'operatore dovrà esser in grado di intessere questi fili invisibili per lo sviluppo di un progetto linguistico.

A seguire, nel quarto capitolo, verrà presentato un caso di studio, consistente in progetto linguistico nel quale – grazie all'applicazione di linguaggi formali e alla tecnologia del *Cognitive Computing* – si ricercheranno ed estrarranno informazioni in materia di *Cyber Threat Intelligence* (CTI). La CTI è una disciplina della sicurezza informatica che si può tradurre come *«servizio di informazione strategica sulle minacce informatiche»*[4], nonché diramazione delle *Cyber Operations*, condividendo con esse la stessa matrice tassonomica della *Cyber Intelligence* e della *Cyber Warfare.*

Verrà fornita una rappresentazione tassonomica degli strumenti e delle tecniche per l'individuazione di *cyber threats,* conducendo in prima istanza una raccolta pianificata dei dati grezzi e delle informazioni recuperate dal web mediante le tecnologie OSINT e SOCMINT, ai fini del rilevamento, dell'analisi e del monitoraggio degli attacchi.

A seguito della fase di raccolta dati, verrà sviluppato un modulo di estrazione delle entità interessate seguendo il modello linguistico STIX[5]: il *target* di tale modulo è quello di poter individuare *cyber threats* nel minor tempo e con la maggior accuratezza possibili, ai fini dell'esecuzione di un'analisi predittiva delle minacce, nonché dell'implementazione di misure di mitigazione e/o di un potenziale contro attacco.

Linguisticamente, l'obiettivo da raggiungere è la ricognizione delle (inter)dipendenze sintattiche e semantiche tra elementi, che rendono possibili la classificazione e l'estrazione sia di singole entità (sotto forma di *keywords*), che delle loro interrelazioni.

[4] https://www.pandasecurity.com/it/mediacenter/sicurezza/cose-la-cyber-threat-intelligence/
[5] Acronimo di *Structured Threat Information eXpressions*. Cfr. https://stixproject.github.io/

# 1. SCIENZE *del* LINGUAGGIO *e* CYBER INTELLIGENCE

## *1.1 Il linguaggio della Cyber Intelligence*

Con *Cyber Intelligence* si intende la raccolta e l'analisi di dati di varia natura, soprattutto a carattere linguistico, con l'obiettivo di fornire un solido supporto all'azione strategica di *decision making* degli operatori di Intelligence. Questa tecnologia interessa diversi settori, tra cui, *in primis*, quello militare e governativo. *«La Cyber Intelligence utilizza il linguaggio tipico dello spionaggio e della Cyber Warfare»*[6]*:* partendo da questa citazione, il presente lavoro di tesi ambirà a fornire una panoramica sui modelli di analisi linguistico-computazionale aventi come obiettivo l'individuazione delle minacce, dei loro attori e degli strumenti utilizzati per realizzarle.

L'Intelligence, definito da sempre come *«il secondo mestiere più antico del mondo»*[7]*,* rappresenta, nell'odierna era digitale, la chiave di volta per *«l'acquisizione delle informazioni e per l'interpretazione di una società sempre più trasformata ed influenzata dall'avvento di internet e della tecnologia».*[8] Possiamo definire la *Cyber Intelligence* come quel «*complesso di attività programmate e applicate per identificare, seguire, misurare e monitorare informazioni sulle minacce digitali, nonché dati sulle intenzioni e attività di entità avversarie».*[9] È un'attività, quindi, che tratta di servizi di controspionaggio, di reti di informazioni condivise e di alleanze tra i possibili *target* degli attacchi.

La diffusione di internet e la costante crescita della digitalizzazione in ogni aspetto della vita degli individui hanno generato, con gli anni, un settimo[10] continente: il

[6] https://www.pandasecurity.com/it/mediacenter/sicurezza/cose-la-cyber-threat-intelligence/
[7] A. Mantici, *Intelligence, il secondo mestiere più antico del mondo,* Babilon Magazine – Terapie geopolitiche, 2021. Cfr. www.babilonmagazine.it/spy-games-alfredo-mantici-libro-paesi-edizioni/#:~:text=Il%20%C2%ABsecondo%20mestiere%20pi%C3%B9%20antico,notizie%20segrete%20connesse%20alla%20sicurezza.
[8] N. Nalesso, *Il ruolo della Cyber Intelligence nella tutela della Sicurezza Nazionale,* 2019. Cfr. https://www.cyberlaws.it/2019/il-ruolo-della-cyber-intelligence-nella-tutela-della-sicurezza-nazionale/
[9] U. Gori, L. S. Germani, *Information Warfare. La sfida della cyber-intelligence al sistema Italia: dalla sicurezza delle imprese alla sicurezza nazionale,* Franco Angeli Editore, 2011, pp. 16-17.
[10] In riferimento al "modello a sei continenti", ovvero al criterio storico-etimologico di classificazione dei continenti più diffuso in Italia, in Europa occidentale (ad eccezione delle Isole Britanniche) e in America latina. Tale modello include i seguenti continenti: Africa, Asia, Europa, America, Oceania, Antartide. Tra gli altri sistemi di suddivisione delle terre emerse, si annoverano anche il "modello a sette continenti" e il "modello a cinque continenti".

*cyberspazio.* Con questa definizione ci si riferisce ad eventi – guerre incluse – che non si verificano nei luoghi in cui si trovano fisicamente gli attori dell'azione, bensì in un luogo a sé stante, invisibile, senza patria, spazio o confini. Il *cyberspazio* sta diventando sempre più un *«settore militare operativo»*[11], definito anche come la *«quinta dimensione del conflitto».*[12]

Come si definisce sulla rivista di Intelligence "Gnosis": *«Linguaggio ed Intelligence hanno entrambi a che fare con le informazioni: l'uno è lo strumento attraverso cui si comunicano informazioni, l'altro è l'attività che per definizione si sostanzia nella ricerca ed elaborazione di informazioni».*[13] Proseguendo nella lettura: *«L'Intelligence ha sviluppato con il tempo un proprio linguaggio ed un proprio gergo,* [in virtù] *delle* [proprie] *condizioni operative* [...] *e dell'attività di informazione per la Sicurezza Nazionale».*[14]

Il vocabolo *Intelligence* rappresenta un esempio di ciò che in linguistica viene definito "prestito linguistico" di tipo "bidirezionale" o "mediato": è di conio anglosassone, ma ha evidente etimologia latina (il termine proviene dal verbo *intelligere,* ovvero "legare insieme" scil. i concetti, le idee e pertanto "comprendere").

Questo termine è destinato alla Presidenza del Consiglio dei Ministri in una «*duplice accezione che gli è propria: l'una, soggettiva, che rimanda al complesso delle strutture e delle attività volte a raccogliere notizie utili ai fini della tutela della Sicurezza Nazionale; l'altra, oggettiva, che si riferisce al prodotto di tale attività, funzionale a sostenere le decisioni in materia di protezione degli interessi del Paese».*[15]

Al fine di migliorare sempre più la capacità e le strategie di raccolta e filtraggio delle informazioni, l'Intelligence abbraccia nuove dimensioni tecnologiche.

La *Cyber Intelligence*, non diversamente dall'attività di Intelligence tradizionale, mira ad acquisire e valorizzare le informazioni attraverso l'utilizzo di sistemi informatici e – al contrario dell'assetto informativo dell'era analogica – oggi fa fronte ad una sovrabbondanza

[11]A. Flores D'Arcais, *Usa: la nuova guerra è il cyberspazio*, 2015. Cfr. https://inchieste.repubblica.it/it/repubblica/rep-it/2015/06/15/news/cosi_mi_arruolo_tra_gli_007-115409472/
[12] M. Caligiuri, (a prefazione di) *Cyber Espionage e Cyber Counterintelligence*, A. Teti, Rubbettino Editore, 2018, p. 10.
[13] Presidenza del Consiglio dei Ministri, Dipartimento delle Informazioni per la Sicurezza, *Glossario Gnosis – – Il linguaggio degli organismi informativi*, Rivista Italiana di Intelligence, 2013.
[14] Ibidem.
[15] Ibidem.

di informazioni: il filtraggio delle informazioni – azione preliminare e inderogabile del *ciclo di Intelligence* – rappresenta una delle fasi decisive nell'azione intelligente cibernetica.

Le principali modalità di raccolta delle informazioni sono le seguenti:

- *HUMan INTelligence,* o HUMINT: attività che consistente nella raccolta di informazioni per mezzo di contatti interpersonali. È stata, fino all'introduzione di strumenti tecnologici durante la prima guerra mondiale, l'unica fonte di approvvigionamento di dati. Tutt'ora oggi, permette l'accesso alle notizie più sensibili, non accertabili in altro modo, e in contesti in cui non si può fare totalmente affidamento alle attrezzature elettroniche. Sviluppare abilità di *Human Intelligence* richiede un addestramento specifico;
- *Virtual HUMINT,* o VHUMINT, attività *«che si basa essenzialmente sul rapporto tra "comunicatore" ed "interlocutore"»*[16] all'interno di ambienti virtuali, in cui le interazioni corrispondono a relazioni sociali tra due o più individui (sotto forma di profili); ulteriori obiettivi di tale intelligenza (oltre alla raccolta di informazioni), sono il condizionamento delle masse, la propaganda, nonché azioni di guerra psicologica[17];
- *IMagery INTelligence,* o IMINT: attività di raccolta di informazioni mediante l'analisi di immagini e/o fotografie aeree o satellitari;
- *SIGnals INTelligence,* o SIGINT: ovvero, l'attività di raccolta di informazioni tramite l'intercettazione e analisi di segnali;
- *ACoustical INTelligence,* o ACINT: disciplina di raccolta di informazioni tramite i fenomeni acustici;
- *Social Media Intelligence,* o SOCMINT: si riferisce agli strumenti che consentono di monitorare reti e canali sociali per identificare profili, relazioni, organizzazioni e tracciare reti di conoscenze. I *social media* rappresentano oggi le più grandi armi di "distrazione" di massa;
- *Open Source Intelligence,* o OSINT: disciplina che si occupa della ricerca, raccolta ed analisi di dati e notizie d'interesse tratte da fonti accessibili in rete; una volta definita la loro affidabilità, le fonti aperte vengono integrate con le fonti chiuse – non accessibili al pubblico poiché ottenute mediante specifiche operazioni sul campo, nonché mediante HUMINT – determinanti per indirizzare le scelte dei decisori di Intelligence.

[16] A. Teti, *Virtual Humint – La nuova frontiera dell'Intelligence,* Rubbettino Editore, 2019, p. 24.
[17] Si rimanda al paragrafo 2.1.1 *"(Cyber)Bellum omnium contra omnes – operazioni (psico)cognitive"* del presente lavoro di tesi.

Nel riconoscere il grande valore della rete per la realizzazione dell'attività investigativa cibernetica, negli anni si è affermata sempre più l'importanza delle attività di OSINT e SOCMINT. L'OSINT, come già indicato in precedenza, mira alla raccolta di informazioni online dalle fonti aperte: l'individuo è diventato una fonte inesauribile di produzione di dati, i quali, una volta rilasciati in rete, rimangono a disposizione di chiunque in una dimensione panottica. Bisogna, però, fare una precisazione: quando si parla di rete, non ci riferisce solamente a quel 4-5% di essa che tutti noi utilizziamo, definita *surface web* – la rete di superficie – ma anche alla rete sommersa, definita *deep web*, non indicizzata dai motori di ricerca. Non si può, inoltre, eludere l'esistenza di una rete ancor più profonda, definita oscura: il *dark web*. In quest'ultima, l'illegalità fa da padrona.

Nell'ambito della *Cyber Threat Intelligence* (branca della *Cyber Security)*, effettuare delle operazioni di ricerca strategica sul *dark web* permette l'accesso a dati sensibili sulle intenzioni avverse di *cyber* criminali, nonché contribuisce alla creazione di un processo decisionale efficiente per valutare e mitigare rischi e pericoli. I dettagli e le informazioni raccolte in materia di *cyber threats* (soprattutto nel *deep web* e nel *dark web)* – opportunamente condivisi e analizzati dagli operatori di Intelligence – rendono possibile l'applicazione di strategie preventive, tattiche di intervento e sistemi di monitoraggio.

Inoltre, in queste profondità del web, i servizi di Intelligence *«attraverso determinate tecniche di manipolazione riescono ad intercettare e veicolare messaggi più funzionali ai propri obiettivi strategici.* […] *I compiti assegnati all'Intelligence ricoprono un ampio spettro di attività, che vanno dall'analisi dei pericoli al trattamento delle informazioni* [mediante la *Cyber Intelligence* e la *Knowledge Management], dal contrasto al cyber crime, allo sviluppo di capacità operative».*[18]

Tornando "in superficie", nell'ambito di ricerca delle informazioni su persone e organizzazioni, i *social networks* (da cui la definizione di SOCMINT) hanno assunto una grande rilevanza nell'ambito della *Cyber Intelligence*, in quanto "l'individuo medio" ha di fatto trasferito la propria esistenza in rete, condividendo dati e informazioni sensibili, facendo diventare rapidamente raggiungibili[19] ed esaminabili dall'analista di Intelligence per

[18] M. Caligiuri, *Cyber Intelligence – Tra Libertà e Sicurezza*, Donzelli Editore, 2016, p. 7. Cfr. https://www.sicurezzanazionale.gov.it/sisr.nsf/wp-content/uploads/2016/06/cyber-intelligence-sfida-data-scientist-Caligiuri.pdf

[19] Prima dello scandalo *Facebook-Cambridge Analytica* avvenuto nel 2018 (in cui vennero raccolti i dati personali di circa 50 milioni di utenti *Facebook* senza previo consenso) era possibile estrarre dati dai profili *Facebook* mediante l'utilizzo di un profilo di tipo *developer* e una chiave di accesso (definita *token*, in questo

scopi prefissati. Lo studio dei profili social permette di comprendere desideri e necessità delle masse, nonché dedurre lo stato mentale dei singoli cittadini e/o gli orientamenti socio-politici[20]. Si pensi all'importanza che un'attività di analisi critica del linguaggio possa rivestire anche in chiave anti-terroristica.

Secondo Luciano Romito, docente presso l'Università degli Studi della Calabria, nonché coordinatore dell'OLF – Osservatorio sulla Linguistica Forense: *«Il linguista è utile all'Intelligence per prevenire la radicalizzazione».* Questo perché l'informazione è potere, ma per comprenderla bene occorre avere competenza linguistica e conoscere i modelli di comunicazione. Citando, inoltre, il Prof. Antonio Teti, docente di *IT Governance e Big Data* all'Università "G. D'Annunzio" Chieti-Pescara, nonché Responsabile per la Transizione Digitale e del Settore Sistemi Informativi di Ateneo: *«Vincerà chi sarà in grado di distinguere l'informazione dalla disinformazione e chi saprà avvalersi di specifici algoritmi basati su piattaforme di intelligenza artificiale in grado di elaborare continuamente il flusso dati generati dal web e dai social».*[21]

Paradossalmente, nonostante l'introduzione ed implementazione di tecnologie sempre più raffinate e l'ingente mole di informazioni reperibili in rete in incessante crescita, il fattore umano risulta ancor oggi essere un elemento imprescindibile per svolgere un'analisi intelligente realmente efficiente: l'operatore di Intelligence deve essere in grado di applicare un *«metodo per trasformare le informazioni in conoscenza»*[22], nonché discernere l'informazione dalla disinformazione. Inoltre, è nell'umanità e nello scibile sociolinguistico e culturale dell'analista che risiede quella (meta)cognizione necessaria per la comprensione di un linguaggio che, graficamente o foneticamente, potrebbe non rispondere allo standard.

La lingua – o codice linguistico – infatti, si configura *«come il concretizzarsi storico-sociale della facoltà del linguaggio posseduta da ogni individuo, ovvero come capacità di unire significati a simboli per poter dare espressione ai primi».*[23]

caso). Non si trattò di un caso di *data breach,* né di un *attacco hacker*: era possibile estrapolare dati personali mediante apposite *API (Application Programming Interfaces)* facilmente accessibili.

[20] «*L'Intelligence non è prodotta da una serie di documenti, ma dalle persone»*. – M. Caligiuri, (a prefazione di) *Cyber Espionage e Cyber Counterintelligence*, A. Teti, Rubbettino Editore, 2018, p.13.

[21] Cfr. http://www.osservatorioanalitico.com/?p=9995

[22] M. Caligiuri, *La scienza dell'Intelligence nell'era dell'incertezza,* Formiche, 2015. Cfr. https://formiche.net/2015/02/la-scienza-dellintelligence-nellera-dellincertezza/

[23] M. Martignong, *Riflessione sulla lingua – Il codice linguistico*, 2004, p. 1. Cfr. http://www.insegnareitaliano.it/documenti/Laboratorio%20docenti/italiano/Martignon/riflessione_sulla_lingua/Il%20codice%20linguistico_2004.PDF

In semiotica, con "codice" s'intende il sistema di segni che permette la trasmissione di informazioni (dati linguistici) tra un mittente e un ricevente di un messaggio; l'atto di formazione di un messaggio viene definito codifica; mentre, la sua interpretazione è detta decodifica. Quindi, allo stesso modo dei dati criptati *(encrypted data),* che possono essere letti o elaborati solo dopo essere stati decriptati (a seguito di un'operazione di *decrypting),* parlare una lingua significa definire chiavi di decodifica del (meta)codice linguistico; quindi, definire e concordare dei significati – intenzionali o letterali – trasmessi in prima istanza da un mittente, e negoziati poi con un ricevente.

Al contrario del processo di elaborazione dei dati forniti in *input* in un sistema computazionale, l'atto di produzione e di (effettiva) ricezione (con conseguente *feedback)* di un messaggio, è tutt'altro che binaria: l'interazione (e comunicazione) umana si configura come un sistema complesso – adattativo/adattivo – strutturato da relazioni reciproche tra *«parlanti e ascoltatori* [che] *si muovono frequentemente* [...] *tra strutture diverse, e spesso le fondono».*[24]

Secondo la teoria della rilevanza[25] – dell'ambito comunicazione e alla cognizione, postulata dall'antropologo cognitivo Dan Sperber e dalla linguista Deirdre Wilson – la comunicazione umana intenzionale, nella sua totalità, è di tipo ostensiva-inferenziale. Vale a dire che l'uso di uno stimolo ostensivo è in grado di generare aspettative di rilevanza cognitiva, diversamente da altri tipi di stimoli esperiti; ciò è a sostegno dell'idea che il carattere essenziale della comunicazione sia rappresentata dall'espressione, nonché dalla ricognizione degli intenti.

Un codice linguistico, inoltre, si delinea come (dia)sistema[26] in continua trasformazione, essendo esso fortemente legato agli assetti ed ai contesti socioculturali di una data comunità di parlanti, nonché strutturato sulla base di esigenze comunicative in

[24] L. Renzi, G. Salvi, *Nuova introduzione alla filologia romanza,* 2ª ed., Bologna, Il Mulino, 1994, p. 158.
[25] La teoria della rilevanza (o teoria della pertinenza) viene considerata come evoluzione di una delle più importanti idee del filosofo e pragmatico inglese Paul Grice, uno dei massimi esponenti, insieme a John Austin, della filosofia del linguaggio ordinario.
[26] Un (dia)sistema è l'insieme delle variazioni linguistiche presentanti delle strutture in comune, tali da poter essere descritte e considerate in un sistema collettivo di corrispondenze. Il modello diasistematico – che il linguista Leiv Flydal definisce come *architettura dell'insieme* – si compone di quattro dimensioni: la dimensione diacronica, diamesica, diatopica e diastratica. Gaetano Berruto include anche la diacronia nella lista delle variazioni diasistematiche perché tale variazione va «*peraltro, a confluire con il grande fenomeno del cambiamento o mutamento linguistico, e che in quanto tale fuoriesce dall'ambito di azione della sociolinguistica»* – G. Berruto, *Prima lezione di sociolinguistica,* Roma/Bari, Laterza, 2005, p.85.

continua evoluzione – le quali, a loro volta, portano alla creazione di nuovi canali, strumenti e mezzi per la loro realizzazione.

Secondo Roman Jakobson, infatti – uno dei maggiori esponenti nell'ambito linguistico del XX secolo, nonché uno dei principali iniziatori della scuola del formalismo e dello strutturalismo – il canale orale (mediante cui trova espressione e realizzazione il linguaggio verbale) rappresenta la più importante modalità di comunicazione tra gli esseri umani. Tra le variabili sociolinguistiche (ovvero, tra le forme di una data lingua usate da una comunità di parlanti) la molteplicità delle alternative non marcate definisce lo stato iniziale nell'acquisizione del linguaggio; il suo sviluppo progredisce attraverso il cambio di opzione (e parallelamente a ciò, in relazione al modello dell'ambiente) poiché ogni (dia)sistema linguistico presenta varianti marcate.

L'esigenza comunicativa risulta, quindi, nell'azione mirata al soddisfacimento un determinato fabbisogno informativo, inteso come identificazione ed analisi di tutte le informazioni necessarie all'individuo in un dato contesto sociale/situazionale; il riuscito conseguimento di tale processo è dovuto a strategiche ed intelligenti modalità di lettura di un determinato messaggio, in grado di condurre all'instaurazione, a livello pragmatico, di un corretto rapporto segni-utenti.

La *Cyber Intelligence* linguistica – avvalendosi della molteplicità delle discipline che compongono le scienze del linguaggio e che conducono ad una riflessione (ed azione) metacognitiva, nonché alla realizzazione dell'intenzionalità comunicativa – ha uno scopo predittivo: attraverso un approccio inclusivo dei processi implicati nella conoscenza (servendosi, quindi, della *knowledge* dello specialista linguistico in grado di utilizzare in maniera strategica le informazioni raccolte) è possibile intendere funzionalmente come "guida" i comportamenti (mediamente noti) che spingono gli avversari (o *competitor*) all'azione; tale approccio metacognitivo conduce al tentativo di decifrazione e di previsione delle loro espressioni, intenzioni ed abilità, permettendo, così, la pianificazione e la realizzazione delle migliori strategie di spionaggio, controspionaggio, difesa e/o attacco, anche mediante l'utilizzo di *tool* tattici.

Come definisce il documento *Strategia Nazionale di Cybersicurezza*[27] redatto e pubblicato il 25 maggio 2022 dall'Agenzia per la Cybersicurezza Nazionale (ACN): *«Le ultime tensioni internazionali hanno messo in evidenza l'importanza di un meccanismo di gestione delle crisi cibernetiche, che consenta* […] *di graduare le attività sulla base di scenari predefiniti della minaccia cyber* [...], *al ricorrere dei quali viene innescata l'immediata applicazione di strumenti, procedure e norme di linguaggio comuni».*

In maniera molto sintetica, quindi, l'accesso e la condivisione di uno stesso linguaggio (inteso come codice linguistico, oltreché formale) – ai fini della realizzazione di atti comunicativi ed azioni performative in ambito digitale – riveste un ruolo fondamentale all'interno delle organizzazioni e dei sistemi sociali, operando sulla riduzione della complessità, sulla gestione dell'incertezza e del rischio, e sulla mediazione del conflitto.

L'analista di Intelligence, oggi, viene inserito in una condizione operativa molto simile a quella di un *Data Scientist*, figura con spiccate competenze tecnologiche e capacità di interpretazione ed analisi, indispensabili per garantire una efficiente e funzionale gestione di algoritmi e di statistiche avanzate per l'estrazione di conoscenza. Così come un *Data Scientist*, compito dell'operatore di Intelligence è quello di saper individuare, selezionare, esaminare e disseminare (ai Vertici decisionali) l'informazione – linguistica, numerica o statistica – contenuta nei *Big Data*[28].

L'attività di Intelligence è indispensabile per comprendere la realtà; la *Cyber Intelligence* lo è ancora di più, perché permette di orientarsi sia nella realtà umana – tangibile – che in quella digitale – immateriale ed invisibile.

[27] Agenzia per la Cybersicurezza Nazionale, *Strategia Nazionale di Cybersicurezza (2022-2026)*, 2022. Cfr. www.cybersecitalia.it/wp-content/uploads/2022/05/Strategia-nazionale-di-cybersicurezza_documento.pdf

[28] Con il termine *Big Data* si intende il fenomeno atto ad immagazzinare, gestire e analizzare grandi quantità di dati. Appositi algoritmi sono deputati all'individuazione delle informazioni e alla creazione di modelli in grado di attuare una previsione delle attività umane.

### *1.1.1 Le fasi del ciclo di Intelligence*

Dopo aver chiarito che cosa si intenda per *Cyber Intelligence,* in questo paragrafo si passerà a descrivere come questa operi concretamente.

Nello specifico, si evidenzieranno le procedure che consentono di selezionare in modo strategico le informazioni linguistiche utili all'azione intelligente, e di come individuare i connotati riferibili alla rilevanza ed affidabilità dei dati linguistici raccolti al fine di massimizzare i risultati dell'analisi.

L'aspetto procedurale – che realizza l'azione intelligente sopracitata – è definita "ciclo di Intelligence"[29]; questo rappresenta un percorso temporale-funzionale che, secondo una procedura ben determinata, consente di far arrivare al *decision maker* la giusta informazione "servibile" e contestualizzata al fine di metterlo nella condizione di assumere una determinata decisione[30].

In sintesi, il ciclo di Intelligence è *«l'insieme delle fasi in cui si articola l'attività d'informazione per la sicurezza»*[31]*;* oltre a costituire un processo operativo, tendenzialmente a vocazione scientifica, il processo trasforma una notizia in un'informazione[32]. Tale ciclo mette in relazione diverse fasi del processo di elaborazione di dati reperiti online, nonché della *knowledge* degli analisti per rispondere alle nuove richieste di informazione.

Un programma di Intelligence efficace è quello che mira all'iterazione dei processi da esso realizzati. Inoltre, attraverso il continuo raffinamento delle tecniche utilizzate e mediante funzionalità operative sempre più all'avanguardia, i *software* di cyber analisi e di analisi investigativa risultano essere rispondenti per lo sviluppo di soluzioni[33] a rischi cibernetici tangibili ed altamente organizzati, di cui il *cyberspazio* ne è il teatro operativo.

---

[29]Sistema di informazione per la Sicurezza della Repubblica, *Lezione sull'intelligence,* Presidenza del Consiglio dei Ministri, p. 16. www.sicurezzanazionale.gov.it/sisr.nsf/wp-content/uploads/2014/05/lezione-intelligence.pdf

[30] D. Antiseri, A. Soi, *Intelligence e metodo scientifico*, Rubbettino Editore, 2018, p. 98.

[31]A. Sperini, *Implementazione del ciclo d'Intelligence tramite l'utilizzo della Social Media Intelligence (SOCMINT)*, Ministero della Difesa, 2017. Cfr. www.difesa.it/SMD_/CASD/IM/CeMiSS/Pubblicazioni/ricerche/Pagine/ricerca_Sperini.aspx

[32] «*L'informazione è l'assemblaggio di una serie di dati* [significativi in un contesto sociale] *che analizziamo e confrontiamo e a cui diamo un certo valore per effettuare le nostre scelte»*. – A. Tofalo, *Intelligence Collettiva: i dati, l'informazione e il linguaggio*, 2017. Cfr. www.angelotofalo.com/intelligence-collettiva-dati-linformazione-linguaggio/

[33] Tra cui, soluzioni SIEM per l'individuazione le potenziali minacce, vulnerabilità di sicurezza e anomalie del comportamento degli utenti. Utilizza moduli di A.I. ed è diventata una soluzione imperetibile nei moderni centri operativi di sicurezza (SOC).

Come si legge dal documento pubblicato nel 2014 dal Sistema di Informazione per la Sicurezza della Repubblica, destinato alla loro Scuola di Formazione: *«La legge 124/2007 ha codificato diversi strumenti operativi di cui possono avvalersi i servizi di informazione per l'attività* […] *info-operativa, a sottolineare che il fine primario dell'azione dell'Intelligence è la ricerca informativa,* […] *di notizie utili a prevenire, rilevare, contenere e contrastare le minacce alla Sicurezza Nazionale».*[34]

Il ciclo di Intelligence si compone delle seguenti fasi:

- pianificazione: definizione degli obiettivi e della strategia di analisi;
- raccolta dati: acquisizione, ricerca informativa e gestione dell'informazione;
- analisi ed interpretazione dei dati raccolti, con conseguente sviluppo di scenari;
- eventuale produzione di un report sull'investigazione attuata;
- disseminazione dei risultati[35] e conseguente valutazione dei *feedbacks.*

Secondo tale schema, l'attività inizia su *input* del decisore (o del cliente) che formula ed avanza le proprie richieste in veste di desiderata alle società di Intelligence: il processo inizia quindi con la pianificazione, ovvero, con l'identificazione del fabbisogno informativo, in cui si pianifica una domanda in base a prestabilite esigenze strategiche.

La seconda fase sancisce il momento in cui gli operatori di Intelligence intraprendono la raccolta delle informazioni mediante gli strumenti a loro disposizione: si sfruttano, quindi, fonti OSINT – tra cui le indicizzazioni di *Google News*, i *feed* RSS, le *press releases*, i blog di interesse – SOCMINT, VHUMINT (qualora autorizzati), *database* locali, e molto altro. La capacità di accedere, comprendere e utilizzare efficacemente i dati e le informazioni non strutturate – contenute in e-mail, contratti, messaggi di chat e altri tipi di documenti – è cruciale per il successo del ciclo di Intelligence.

Una volta raccolte le informazioni, inizia la fase più importante e più delicata dell'intero processo di Intelligence: l'analisi delle informazioni, ovvero, la loro valutazione. Questa fase è un passaggio chiave: trasforma i dati e le notizie raccolte in un prodotto finito

[34] Presidenza del Consiglio dei Ministri, *Lezioni sull'Intelligence – Scuola di Formazione,* 2014, p. 18. Cfr. https://www.sicurezzanazionale.gov.it/sisr.nsf/wp-content/uploads/2014/05/lezione-intelligence.pdf

[35] Nel contesto della Sicurezza Nazionale, la disseminazione – intesa come sistema di informazione e comunicazione – è diretta esclusivamente ai Vertici decisionali.

impiegabile, ed è anche il primo momento in cui è possibile riorientare e riconsiderare i parametri di ricerca delle informazioni.

È qui che subentra il prodotto dell'azione linguistica realizzata dagli specialisti linguistici mediante un processore semantico: ai fini della gestione di testi e dialoghi complessi, i *software* di comprensione e analisi del linguaggio – come la tecnologia del *Cognitive Computing*, utilizzata nel seguito di questo lavoro – applicano sistemi adattivi, interattivi, iterativi e contestuali per riconoscere e adattarsi nei significati, nella sintassi, nella morfologia e nei domini appropriati nei confronti del *target* linguistico di interesse.

Ne consegue che *«le competenze della figura professionale di un/a linguista rivestono un ruolo determinante, sia nella collazione di dati dal web, nella definizione degli input a supporto dell'addestramento (training) degli algoritmi, nonché nella comprensione e validazione dell'output ottenuto».*[36]

Attraverso il processore linguistico si creano ed implementano regole di estrazione, categorizzazione e normalizzazione per il riconoscimento e discernimento di informazioni e significati. Inoltre, mediante il processo di disambiguazione semantica e sintattica dei dati linguistici raccolti, gli operatori di Intelligence linguistica sono in grado di centrare l'obiettivo della propria analisi investigativa.

Infine, i risultati delle *RegEx* realizzate sul motore semantico vengono implementati (sotto forma di un pacchetto linguistico, generalmente in formato GSL) in un *software* di *Decision Intelligence* che automatizza e ottimizza tutte le attività del ciclo.

Attraverso il *software* di Intelligence deputato, vengono utilizzati diversi strumenti e tecniche per la visualizzazione degli *output* dell'analisi linguistica intelligente, come ad esempio: *la Link Analysis*, i diagrammi causa-effetto, analisi statistiche percentuali e l'analisi dei sentimenti (ovvero, la dimensione emotiva nascosta del linguaggio online), nonché sezioni che ricordano la struttura dei più comuni *browser* sul web, in cui è possibile creare delle *query* di ricerca, applicare dei filtri semantici e consultare tassonomie linguistiche (standard o realizzate *ad hoc* per il Cliente).

[36] F. Bertolino, *Cy4Gate crea una divisione dedicata all'intelligenza artificiale,* intervista ad Emanuele Galtieri, *Milano Finanza*, 2022. Cfr. www.milanofinanza.it/news/cy4gate-crea-una-divisione-dedicata-all-intelligenza-artificiale-20220125

*«Occorre, quindi, fornire all'analista un quadro documentale sufficientemente ampio in modo che egli ne tragga uno studio comparato e ne valuti i passaggi critici. Allo stesso modo, attraverso la raccolta dei comuni articoli dei media, è possibile ricostruire il messaggio occulto o implicito».*[37]

Al termine della fase analitica vi è la produzione di un report che viene inoltrato (disseminato) ai decisori[38], i quali forniranno un *feedback* al riguardo, per poi eventualmente formulare ulteriori richieste e/o identificare nuovi *target*. Quest'ultima fase è *«volta a definire in che misura i prodotti di Intelligence abbiano soddisfatto le esigenze conoscitive dell'Autorità di Governo o di altri interlocutori istituzionali».*[39]

Con l'identificazione di un nuovo *target*, la ricerca (cyber) investigativa riparte dall'inizio. Seguendo lo stesso *modus operandi*, infatti, anche la disciplina della *Cyber Threat Intelligence* si configura come ciclo – ovvero come processo iterativo – e si suddivide nelle seguenti fasi:

- pianificazione e direzione del progetto, per la realizzazione del processo di indagine – mediante azioni e *tasks* – avente come *target* l'individuazione di *cyber threats*;
- raccolta di dati grezzi *(raw data)*, in grado di soddisfare *tasks* e requisiti definiti nella prima fase. Mediante la tecnologia OSINT (già definita come *Open Source Intelligence),* si setaccia il web per la ricerca di informazioni e dati trafugati. Le fonti OSINT si configurano come caotiche e destrutturate. I dati relativi alle *cyber threats* sono spesso condivisi nelle aree più sommerse del web: *il deep* ed il *dark web;*
- elaborazione ed organizzazione delle informazioni grezze in metadati, attraverso il filtraggio dei dati ridondanti e di quelli definiti "falsi positivi" e/o "falsi negativi";[40]
- fase di analisi, in cui si attua la ricerca di potenziali problemi di sicurezza;
- distribuzione (disseminazione) dei risultati agli operatori di Intelligence;

[37] P. Costantini, *Language Intelligence*, 2019. Cfr. www.paolocostantini.com/language-intelligence/
[38] O trasmesso ai Vertici decisionali.
[39] *Lezioni sull'Intelligence – Scuola di Formazione*, 2014. Cfr. www.sicurezzanazionale.gov.it/sisr.nsf/wp-content/uploads/2014/05/lezione-intelligence.pdf
[40] «*Falso positivo (false positive): Si tratta di un'affermazione che non viene confermata dalla realtà. Ad esempio, l'e-mail viene classificata spam (positive) ma non è spam (false); Falso negativo (false negative): sono le previsioni negative sbagliate. Si tratta di una negazione che non è confermata dalla realtà. Ad esempio, l'e-mail viene classificata no-spam (negative) ma è spam (false)*» – A. Minini, 2020. Cfr. www.andreaminini.com/ai/machine-learning/la-differenza-tra-falsi-negativi-e-falsi-positivi

- restituzione di un *feedback*, che decreta o meno l'avvenuto conseguimento dell'obiettivo (e risoluzione al problema) definito in fase di pianificazione.

Questo processo viene definito "ciclo" perché, nello sviluppo delle operazioni di Intelligence, vengono identificate nuove ulteriori domande – ovvero, nuovi problemi, rischi, potenziali attacchi – che richiedono una rapida risposta, portando così alla definizione di nuovi requisiti per la raccolta delle informazioni per il rilevamento dei profili delle minacce ed attacchi informatici.

Un altro strumento per la messa in atto del pensiero intelligente – ai fini di un adattamento costante all'interno di teatri militari incerti, caotici ed imprevedibili, con l'obiettivo di mitigare potenziali *cyber threats* e/o di saper rispondere in maniera proattiva a *cyber attacks* – è il cosiddetto "ciclo OODA", di cui si delineeranno caratteristiche e strategie nel paragrafo 2.1.1 *"(Cyber)Bellum omnium contra omnes – operazioni psico-cognitive"* del presente lavoro.

"OODA" è un acronimo che rimanda alle quattro fasi del processo decisionale che lo contraddistinguono: osservazione, orientamento, decisione ed azione.

Non si tratta di una metodologia – ovvero, di un *modus operandi et cogitandi* – applicabile esclusivamente a scenari militari, bensì, anche alle situazioni più complesse della vita di ogni individuo, a cui viene richiesta una presa di decisioni più o meno rapida per attenuare conseguenze e raggiungere obiettivi strategici.

Questo approccio favorisce l'agilità di ragionamento nella risoluzione dei problemi, al fine, altresì, di acquisire un vantaggio in un determinato contesto, fornendo una chiave per districare reti di significato più confuse – dall'interpretazione ambigua – garantendo l'attivazione di processi cognitivi per la formazione di nuovi contenuti di conoscenza.

## *1.2 Computazione linguistica: l'incontro tra linguaggio naturale e formale*

Per poter comprendere il connubio tra linguistica ed il mondo dell'Intelligence è conveniente iniziare ad intendere l'atto della comunicazione come un'attività strategica: attuare una funzionale interazione linguistica significa saper analizzare i bisogni comunicativi degli interlocutori nei molteplici contesti, nonché essere in grado di realizzare prodotti comunicativi come risultato di un processo di analisi cognitiva.

Inoltre, proprio come nell'Intelligence, si può guardare alla lingua come potenziale strumento per la realizzazione di strategie di persuasione: gli esseri umani realizzano (più o meno manifestamente) messaggi dai contenuti impliciti, nascondendo logiche e meccanismi psicologici e cognitivi che presuppongono un altro messaggio rispetto a quello che viene effettivamente esplicitato.

Guardando alla lingua come un processo cognitivo (e a fronte delle strategie realizzabili dagli esseri umani attraverso un codice linguistico condiviso), nell'osservare la struttura dei suoni e delle frasi da porre in analisi, la prospettiva degli analisti linguistici si sposta da una visione "etic" (oggettiva, superficiale, scientifica), ad una dimensione "emic"[41] (profonda, ermeneutica e che considera i punti di vista e il trascorso esperienziale dei parlanti). Quest'ultima dimensione riconduce all'"osservazione partecipante" di Bronisław Malinowski[42] e permette la realizzazione dell'analisi delle stesse strutture di suoni e frasi considerando il punto di vista dei soggetti partecipanti all'azione – ovvero, analizzando il modo in cui ogni parola, segno o suono vengono da loro prodotti, compresi e interpretati – che i metodi quantitativi dapprima utilizzati, di tipo "etic", non erano in grado di rilevare.

Nell'ambito della rielaborazione del linguaggio naturale – e ragionando in una dimensione scientifica, oggettiva, "etic" – le prime applicazioni digitali mai tentate prima di allora furono l'invenzione della traduzione automatica e lo spoglio automatico di testi. *«La traduzione automatica nasce nel 1946 come attività di decriptazione dei messaggi, un obiettivo fondamentale dei servizi segreti durante la guerra».*[43]

[41] I concetti di *etic* (etico) ed *emic* (emico) sono stati introdotti per la prima volta dal linguista americano Kenneth Pike per riferirsi al modo in cui il comportamento sociale si manifesta e viene compreso all'interno della società. Tali termini sono stati mutuati dalle desinenze dei lemmi inglesi *phonemics* (fonologia) e *phonetics* (fonetica).
[42] Fondatore della scienza dell'etno-antropologia e pioniere nel campo della ricerca etnografica.
[43] https://www.dir.uniupo.it/pluginfile.php/138267/mod_resource/content/0/domande.pdf

Lo spoglio automatico dei testi, invece, venne progettato nel 1943 dal Gesuita padre Roberto Busa, che, con il supporto della IBM[44], produsse, nell'arco di quasi 40 anni, l'elaborazione completa delle opere di San Tommaso D'Aquino.

Gli studi di padre Roberto Busa hanno costruito le fondamenta sulle quali oggi poggia la linguistica computazionale, conducendo all'utilizzo del computer come strumento di immagazzinamento ed analisi del testo, con conseguente produzione di *corpora* elettronici.

*«La linguistica computazionale si concentra sullo sviluppo di formalismi descrittivi del funzionamento di una lingua naturale»* – tenendo conto sia della *langue,* ovvero, standard sociale linguistico – sia della *parole*[45]*,* ovvero, dell'espressione soggettiva del parlante in lingua, quindi del modello basato sull'uso che se ne fa in contesti reali «[…] *tali da poter essere trasformati in programmi eseguibili da computer».* [46]

I problemi che affronta la linguistica computazionale – come intuibile dal nome stesso della disciplina – consistono nel trovare una mediazione fra il linguaggio umano (oggetto di studio in costante evoluzione) e le capacità di comprensione della macchina, limitate a/da quanto può essere descritto ed implementato mediante regole formali.

Per comprendere le difficoltà che sistematicamente sussistono in fase di ricerca e di (conseguente) implementazione di adeguati meccanismi computazionali per il riconoscimento e l'elaborazione del linguaggio naturale, è necessario conoscere *in primis* la distinzione in livelli del linguaggio stesso – rispettivamente:

- livello sintattico: ovvero, la struttura grammaticale di una data lingua;
- livello semantico: la semantica è quella parte della linguistica che studia il significato delle parole, degli insiemi delle singole lettere, delle frasi e dei testi;
- livello pragmatico: ovvero, lo scambio di interazioni e messaggi in una lingua di comune conoscenza tra due o più interlocutori. La comunicazione, però, va al di là del semplice significato delle frasi (e, quindi, dal solo livello semantico del linguaggio): per poter

[44] «*L'International Business Machines Corporation (comunemente nota come IBM, e soprannominata Big Blue), è un'azienda statunitense, la più anziana e tra le maggiori al mondo nel settore informatico. Produce e commercializza hardware, software per computer, middleware e servizi informatici*» – A. Galilo, *Una data incredibile: 40 anni fa presentato il primo PC IBM*, per *Computermagazine.it,* 2021. Cfr. www.computermagazine.it/2021/08/13/40-anni-fa-presentato-il-primo-pc-ibm/

[45] F. De Saussure, *Cours de linguistique générale, 1916,* trad. it. T. De Mauro, *Corso di linguistica generale,* Laterza, Bari-Roma, 2018.

[46] S. Fidacaro, *Linguistica computazionale la tradizione dell'NLP,* per *Journalpost.info,* 2021. Cfr. https://journalpost.info/linguistica-computazionale-la-tradizione-nel-nlp/

comprendere e analizzare compiutamente una frase, infatti, occorre anche considerare il contesto in cui la frase viene pronunciata, nonché la relazione in essere tra i vari interlocutori. Questi problemi investono il livello pragmatico del linguaggio e vanno risolti al fine di una decodifica del messaggio il più esauriente possibile.

Poiché la comprensione del significato di un enunciato è strettamente legata al contesto di rappresentazione degli stati mentali e della realtà linguistico-culturale e sociale degli interlocutori, non si può non considerare che le tecnologie odierne *«stanno modificando radicalmente non solo l'organizzazione sociale, ma anche l'umanità».*[47]

*«La convergenza al digitale, intesa come il processo di trasferimento progressivo verso il formato digitale di tipologie diverse di informazione tradizionalmente collegate a media diversi, rende possibile un'integrazione inedita tra codici e linguaggi lontani che eravamo abituati a considerare lontani».*[48]

La rivoluzione digitale è il frutto del mero progresso tecnologico (seppur con azione non automatica) e ha ovvi riscontri anche sotto il profilo sociale. Tale riconfigurazione porta alla riformulazione del modo in cui gli individui raffigurano, organizzano e si scambiano informazioni: sono stati creati nuovi canali e strumenti di comunicazione, nonché è un linguaggio nuovo (dei *bit*). Nonostante quest'ultimo non possa essere in grado di sostituire le forme di linguaggio preesistenti, tali nuove opzioni per la realizzazione della comunicazione conducono inevitabilmente a quanto espresso nella teoria riduttiva del "determinismo tecnologico" che *«individua nella tecnologia l'unica causa delle trasformazioni della nostra società».*[49]

Sono trascorsi più di quarant'anni da quando venne sviluppato il primo calcolatore digitale, programmato per compiere delle attività di comprensione in cui si richiedeva una prima forma di intelligenza "non umana". Seguendo questa visione – allora futuristica– vennero sviluppati i primi tentativi digitali per la decodificazione del linguaggio naturale e – imprescindibilmente ad esso – anche della conoscenza umana.

[47] M. Caligiuri, (a prefazione di) *Cyber Espionage e Cyber Counterintelligence* di A. Teti, 2018, p. 9.
[48] E. Pucci, *La Convergenza tecnologica: aspetti tecnici, mercati interessati e tentativi di regolazione*, Università di Pisa, 2013, p. 23.
[49] G. Pecchinenda, *Il determinismo tecnologico*, 2007. Cfr. www.federica.unina.it/sociologia/comunicazione-e-processi-culturali/il-determinismo-tecnologico/

Per far ciò, si è dovuto *in primis* riflettere sul come far memorizzare, processare ed eseguire da una macchina una rielaborazione di informazioni linguistiche e matematiche, nonché come insegnare, far comprendere – sotto forma di apprendimento continuo – e, in qualche modo, far riprodurre alla macchina un'azione interattiva in maniera automatica o semi-automatica. Questi primi esperimenti per la realizzazione di uno strumento digitale intelligente condussero gli scienziati ad ambire, qualche anno più tardi, all'ideazione di una macchina in grado di "generare" comprensione cognitiva artificiale.

I primi studi sull'interazione essere umano-macchina mediante linguaggio naturale, ovvero, sulla possibilità di estrarre contenuti e significati da dati testuali, nonché sulla classificazione di documenti, vennero sperimentati nel corso degli anni '60.

L'applicazione di metodi formali permise una prima interazione tra la linguistica ed il settore dell'intelligenza artificiale.

L'allora neonata tecnologia dell'elaborazione del linguaggio naturale (NLP), fu profondamente influenzata dai metodi deduttivi utilizzati dalla grammatica generativa di Noam Avram Chomsky[50]: tale teoria linguistica aveva come obiettivo quello di individuare delle regole astratte in grado di descrivere la competenza della lingua posseduta da un determinato parlante. Nella dimensione computazionale della lingua, venne, quindi, resa necessaria l'individuazione delle modalità di produzione di norme linguistiche da parte dei parlanti, al fine di rendere possibile la definizione, da parte degli analisti di (cyber) Intelligence, di *patterns* linguistici atti all'estrazione di informazioni mediante un processore semantico.

Al fine di poter comunicare alla macchina il pensiero linguistico umano mediante un approccio cognitivo, il *modus operandi* ideato da scienziati e linguisti prevedeva la raccolta di ingenti quantità di testi – definiti *corpora* testuali – da fornire in *input* al calcolatore elettronico. Attraverso la comprensione della funzionalità nella fruizione di testi autentici – fondamentali per lo studio della struttura del linguaggio, poiché contenenti regolarità linguistiche stabili e reiterate – si poté realizzare una prima validazione dell'azione di classificazione e di estrazione di informazioni effettuate dal processore.

[50] Classe 1928, Noam Avram Chomsky è un linguista, accademico, scienziato cognitivista, teorico della comunicazione, attivista politico, nonché ideatore della grammatica generativo-trasformazionale. Si rimanda al paragrafo 2.2 *"Fondamenti teorici della linguistica cognitiva"*.

La validazione del processo di analisi ha esito positivo se il *corpus* testuale (ovvero, l'*input* linguistico inviato al processore) riesce a riprodurre l'intero ambito di variabilità di tratti e proprietà di una lingua; in tal caso, il *corpus* verrà considerato un campione affidabile.

Con il passare degli anni, si è assistito ad una crescita esponenziale del numero dei *corpora* facilmente fruibili dagli utenti di una lingua, grazie alla produzione copiosa e quotidiana di materiale pubblicato sul web, caratterizzato oggi da infinite risorse testuali.

Come già anticipato nell'introduzione di questo lavoro di tesi, considerato l'incessante flusso e l'inarrestabile aumento della mole di informazioni sul web, è emersa la necessità, da parte degli analisti, di ideare una tecnologia in grado di realizzare un'azione di filtraggio linguistico, preliminarmente all'azione di analisi e classificazione dei contenuti. Questa prima operazione di vaglio delle informazioni conduce a una prima catalogazione (ed eventuale archiviazione) sia delle informazioni utili ed impiegabili nell'attività investigativa, che di quelle "fuorvianti", poiché lontane dal *target* di interesse.

Gli unici modelli computazionali realmente rispondenti ai desiderata, quindi funzionali al raggiungimento degli obiettivi, sono di stampo cognitivo. Per questo, nell'operare studi e ricerche in materia di linguistica computazionale non si può prescindere dal considerare ed includere varie discipline, come la logica, la filosofia del linguaggio, la psicologia, l'antropologia e le scienze cognitive.

Nella realizzazione di un atto comunicativo di qualsiasi natura, il messaggio recepito viene automaticamente processato ed interpretato dai riceventi attraverso una complessa «*concettualizzazione del mondo* [in una prospettiva cognitivo-comportamentale altamente soggettiva], *non solo riguardo alla scelta dei modelli, ma anche e soprattutto alla loro espressione linguistica*»[51].

La parola è l'unità minima specie-specifica della comunicazione umana, nonché di riferimento per la trasmissione di concetti. Ciò avviene attraverso il richiamo automatico da parte dell'individuo a convenzioni precedentemente accettate, radicate e condivise dai parlanti – rappresentanti, quindi, un significato socialmente riconoscibile e comprensibile.

[51] P. Petricca, *Semantica: forme, modelli e problemi*, LED Edizioni Universitarie, 2019, p. 93.

*«Quando si vuole avviare l'analisi computazionale del testo, il problema principale è stabilire dei criteri di identificazione per quella che è la sua unità di base: la parola».*[52]

Per poter analizzare un qualsiasi prodotto linguistico realizzato dall'essere umano, quindi, è stato necessario stabilire delle unità minime di significato e di riferimento, in grado di riflettere la natura e la funzione della parola sotto una veste computazionale: i *tokens*.

Gli elaborati meccanismi di identificazione dei *tokens* permettono la classificazione e il raggruppamento in categorie concettuali secondo un principio di somiglianza reciproca, nel significato e nel loro dominio di appartenenza.

Il metodo più diffuso per elaborare tale classificazione è il *Word Embedding*, conosciuto anche come "semantica distribuita" [53], ovvero una rappresentazione "distribuita" delle parole, la quale permette la memorizzazione – *ex post* addestramento del processore elettronico – di informazioni a carattere semantico.

Ciò si realizza attraverso la conversione delle parole fornite in *input* in vettori numerici – la cui geometria cattura e rivela le relazioni che intercorrono tra esse – per poter procedere, così, al loro inserimento negli algoritmi di *Machine Learning.*

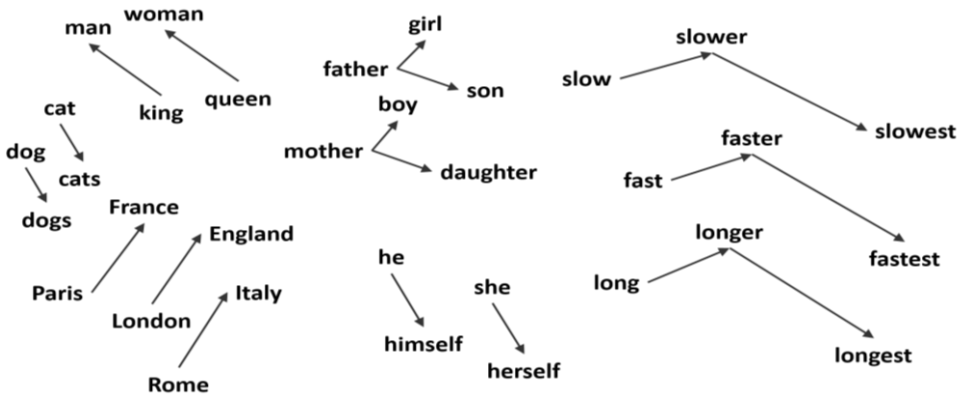



***Figura 1 – Un esempio di Word Embedding***[54]

Ogni modello di *Data Analysis* può essere attivato solo se i dati forniti in *input* risultano quantitativamente e qualitativamente opportuni; detti elementi determinano il successo o il fallimento di qualsiasi processo di elaborazione: dati fuorvianti produrranno risultati approssimativi, che dovranno essere sottoposti – come nel caso del *Word Embedding*

[52] S. Fidacaro, *Linguistica computazionale. La tradizione dell'NLP,* 2021. Cfr. https://journalpost.info/linguistica-computazionale-la-tradizione-nel-nlp/

[53] P. Caressa, *Processare il linguaggio naturale*, per *deeplearningitalia.com*, 2019, pp. 60-63. Cfr. http://www.caressa.it/dli/processare_linguaggio_naturale.pdf

[54] Fonte immagine: https://samyzaf.com/ML/nlp/nlp.html

– a un processo di *debiasing*[55]. Di norma, questo processo è atto alla risoluzione delle asimmetrie sociali (ad esempio, di genere e di ruolo) presenti nel linguaggio[56].

La *tokenizzazione,* invece, è l'operazione mediante la quale si suddivide il testo in *tokens*, ovvero in parole riconosciute dal sistema come tali poiché delimitate da spazi. Si tratta di un processo molto complesso nelle lingue a sistema ortografico continuo *(scriptio continua* o "scrittura continua"), come ad esempio la lingua thailandese o sindhi. In questo caso, l'operazione richiederebbe algoritmi estremamente elaborati, in grado di riconoscere non un sistema linguistico di tipo alfabetico, bensì sillabico, poiché ogni parola – generalmente – è composta da una sola sillaba. Nel sistema sillabico, gli spazi si utilizzano per delimitare la fine di una frase/di un periodo.

In un sistema linguistico di tipo alfabetico trova applicazione il principio di *token, come è noto,* nella linguistica computazionale; ovvero, esso è definibile come *«una qualunque sequenza di caratteri delimitata dagli spazi»*[57], di norma, intendendo ogni sequenza (o stringa) di caratteri come una parola.

Tuttavia, vi sono molteplici eccezioni, come nel caso della presenza dei segni di punteggiatura che sono parte della parola stessa: dall'apostrofo, al punto, ad un carattere speciale come la sempre più diffusa "schwa" [ə][58] nella de-classificazione di genere[59].

---

[55] Ovvero, di "de-polarizzazione". La mancanza di imparzialità nei risultati ricevuti dalle intelligenze computazionali sono dovuti al fatto che: *«Chi è, secondo una lunga serie di parametri, normale, non solo ha modo di sfuggire alle etichette, ma ha anche il potere di decidere quali etichette debbano portare gli altri, gli anormali»*–V. Gheno, *La lingua non dev'essere un museo,* Il Libraio.it, Cfr. https://www.illibraio.it/news/saggistica/vera-gheno-linguaggio-inclusivo-1418943/.

[56] Esempi molto noti di *bias* linguistico-culturali presenti negli *output* di *Machine Learning*: *«L'uomo sta al medico come la donna sta all'infermiera»; «L'uomo sta al Re, come la donna sta alla principessa»; «L'uomo sta al dirigente, come la donna sta alla segretaria»*. Il *Word Embedding* riflette i pregiudizi radicati della/nella società. Si dovrebbe, *in primis*, cercare di eliminare detti pregiudizi tra gli individui, per poter ottenere/aspettarci *output* computazionali imparziali ed equi. *«Le parole sono un importante atto identitario, perché tramite il riconoscimento linguistico reciproco individuiamo i confini delle nostre "tribù" di appartenenza. Dunque, le parole non sono mai "solo parole"»* –V. Gheno, *Verso l'inclusività linguistica e oltre,* Zanichelli Editore, 2021, p. 2. Cfr. www.zanichelli.it/download/media/bq5r/10inparita_Gheno_agg.pdf

[57] S. Fidacaro, *Linguistica computazionale la tradizione dell'NLP,* per *journalpost.info,* 2021. Cfr. https://journalpost.info/linguistica-computazionale-la-tradizione-nel-nlp/

[58] La "schwa" (ə) – detta anche "scevà" – è un simbolo dell'Alfabeto Fonetico Internazionale *(International Phonetic Alphabet – IPA)* che indica una vocale intermedia; vale a dire, essa può denotare sia una vocale debole, nonché l'assenza totale della stessa. Il suono che ne corrisponde è presente in molte lingue, nonché in vari dialetti italiani, come ad esempio, in quello napoletano.

[59] Proprio grazie alla posizione intermedia della "schwa", si sta diffondendo sempre di più il suo uso nella dimensione scritta dell'italiano, con l'obiettivo di eliminare la predominanza maschile nel linguaggio – altresì, per marcare le forme non binarie di genere – puntando, così, ad una maggiore inclusione di genere nel linguaggio (e, conseguentemente, anche nel sociale – e viceversa). Una delle fautrici della diffusione di tale dibattito è la sociolinguista Vera Gheno, sostenitrice dell'uso del sopra citato simbolo [ə] nella lingua italiana scritta e orale, in luogo del maschile sovraesteso. Per un approfondimento: V. Gheno, *Le ragioni del dubbio: l'arte di usare le parole,* Einaudi, 2021.

L'ambiguità della punteggiatura costituisce un problema non irrisorio nel momento in cui si deve identificare le unità linguistiche superiori alla parola, ovvero il sintagma e, in successione, il periodo e la frase.

Le frasi, più comunemente, possono essere definite come sequenze di parole separate da un punto fermo ed uno spazio, e la cui prima parola è rappresentata con la lettera maiuscola; ma sono molte le eccezioni, come, ad esempio, nel caso delle abbreviazioni nei titoli (ad esempio: On. Dep., Mar. Magg., Sig.ra, Dott.ssa) che, secondo questa euristica, verrebbero scisse in frasi distinte. Come già definito, tale definizione non trova applicazione in quelle lingue caratterizzate da *scriptio continua* o "scrittura continua", le quali non distinguono tra lettere maiuscole e minuscole.

La *tokenizzazione* deve, quindi, basarsi su criteri a volte complessi per tenere conto delle possibili eccezioni. Per questo, si è resa necessaria la creazione di espressioni regolari, ovvero notazioni algebriche/alfanumeriche che descrivono formalmente dei *patterns* e che permettono, quindi, la comprensione di stringhe di testo da parte del processore elettronico.

Nella realizzazione di regole per l'identificazione di una stringa di testo, gli operatori booleani, logici e di sequenza rappresentano uno strumento ed ausilio fondamentale.

Dette espressioni regolari vengono definite *RegEx,* ovvero, *Regular Expressions,* che definiscono una funzione che prende in ingresso *(input*) una stringa, e restituisce un valore in uscita *(output)* del tipo sì/no, a seconda che la stringa rispetti o meno un certo *pattern*.

A titolo di semplice esempio dell'interazione tra linguistica e informatica si consideri la seguente *RegEx*[60] per l'individuazione di un indirizzo e-mail non-anonimizzato:

**^[a-ZA-Z0-9_.+-]+@[a-zA-Z0-9-]+\.[a-zA-Z0-9-.]+$**

La presente stringa definisce: un qualsiasi carattere alfanumerico, maiuscolo o minuscolo, eventualmente seguito dal segno ".", ed eventualmente seguito da altri caratteri alfanumerici; a ciò, segue, come condizione necessaria, il carattere "@", seguito da caratteri alfanumerici indefinibili; successivamente, vi sarà sicuramente un punto "\.", seguito da un'ultima stringa di caratteri alfanumerici che rispetta la stessa variabilità ed imprevedibilità delle condizioni espresse nelle porzioni precedenti della stringa.

[60] Sintassi *RegEx* di tipo PERL (acronimo di *Practical Extraction and Reporting Language).*

## *1.3 Intelligenza artificiale e ricognizione linguistica*

*«L'intelligenza artificiale è l'abilità*
*di una macchina di mostrare capacità umane,*
*quali il ragionamento, l'apprendimento, la*
*pianificazione e la creatività».*[61]

L'intelligenza artificiale (definita anche come A.I., o *Artificial Intelligence)* è una tecnologia informatica che sta rivoluzionando il modo con cui l'essere umano interagisce con la macchina, e le macchine tra di loro, attraverso algoritmi di apprendimento in costante aggiornamento. Tali algoritmi permettono alle macchine di essere "addestrate" al fine di essere in grado di compiere azioni e ragionamenti complessi, nonché di poter imparare dai propri errori, sulla scia della teoria del monitor[62] del linguista Stephen D. Krashen, fino a poco tempo fa teorizzata e applicata solamente su apprendenti umani.

Le tecniche di intelligenza artificiale stanno ampliando in modo sorprendente le capacità di comprensione ed elaborazione automatica del linguaggio naturale (rispettivamente, NLU e NLP), da sempre l'obiettivo più ambizioso dell'informatica.

La nostra vita quotidiana vivrà in futuro una notevole semplificazione grazie ad interfacce controllate attraverso digitazione/scrittura e comando vocale. Recentemente, la tecnologia ha permesso una naturale interazione (a tratti, definibile "umana") tra individui e vari *device*, basti pensare ai sistemi *Amazon Alexa*® o *Google Assistant*®.

Vi sono principalmente due forme di intelligenza artificiale:

- A.I. di tipo *software*: assistenti virtuali, *software* di analisi di immagini e di dati linguistici, motori semantici e di ricerca, sistemi di riconoscimento facciale e vocale;

[61] Fonte: Sito Istituzionale del Parlamento Europeo. Cfr. www.europarl.europa.eu/news/it/headlines/society/20200827STO85804/che-cos-e-l-intelligenza-artificiale-e-come-viene-usata

[62] Per un approfondimento: P. Begotti, *L'acquisizione linguistica e la Glottodidattica umanistico-affettiva e funzionale*, Università Ca' Foscari Venezia, 2013.

- A.I. intesa come "intelligenza incorporata"[63]*:* robot, veicoli automatizzati, droni e tutto ciò individuato nell'Internet delle cose *(The Internet of Thing)*, che nasce dall'idea di portare nel mondo digitale gli oggetti della nostra esperienza quotidiana.

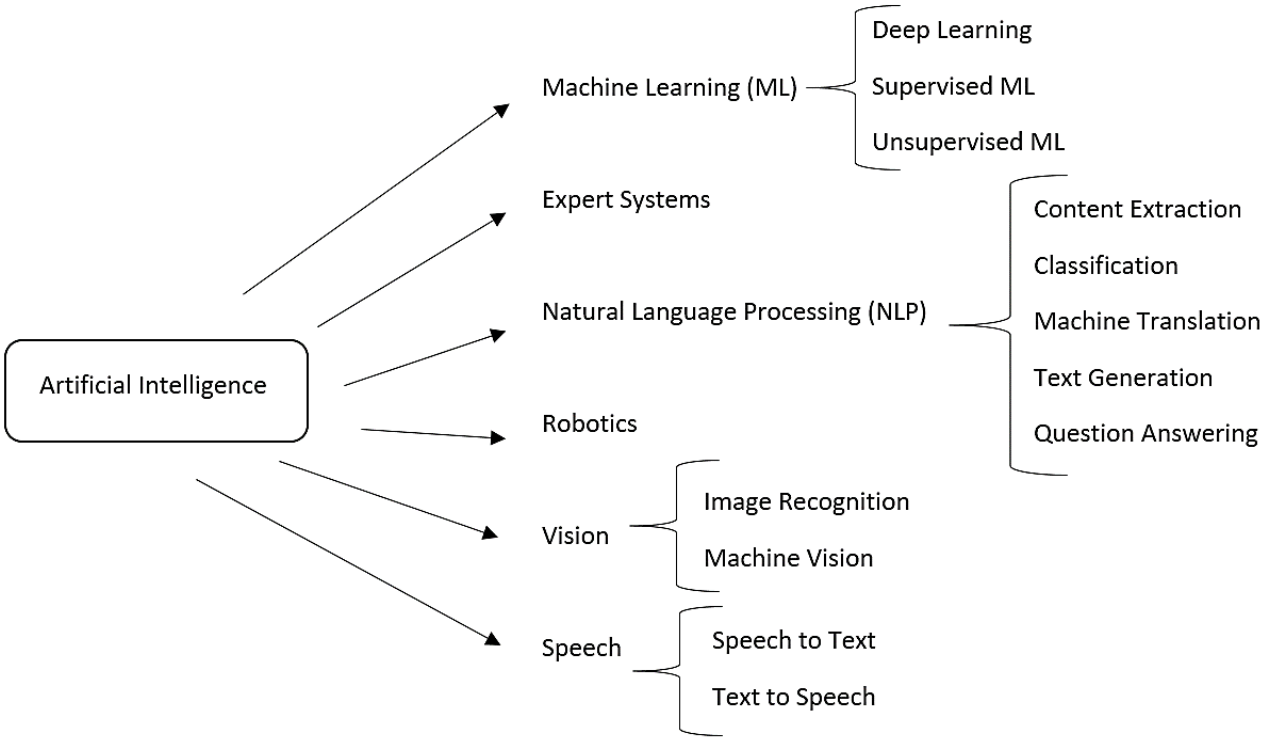



***Figura 2 – Intelligenza artificiale e ricognizione facciale, delle immagini e linguistica***[64]

*«I sistemi di Intelligenza Artificiale possono aiutare a riconoscere e combattere gli attacchi e le minacce informatiche. Lo fanno imparando dal continuo flusso di informazioni* [tra cui flussi di dati linguistici] *e riconoscendo tendenze e ricostruendo come sono avvenuti gli attacchi precedenti».*[65]

Ciò è reso possibile dalla realizzazione di *ANNs – Artificial Neural Networks*, ovvero reti neurali artificiali (modello computazionale composto di neuroni artificiali, ispirato dalla semplificazione di una rete neurale biologica) e dai sistemi di apprendimento delle informazioni *(Machine Learning, Deep Learning)* e di elaborazione e comprensione del

[63] Si rimanda al paragrafo 2.2 *"Fondamenti teorici della linguistica cognitiva",* in cui si delinea il concetto di *corporeità* del linguaggio e della cognizione umana.
[64] Fonte: grafico autoprodotto.
[65] Sito Istituzionale del Parlamento Europeo. Fonte: https://www.europarl.europa.eu/news/it/headlines/society/20200827STO85804/che-cos-e-l-intelligenza-artificiale-e-come-viene-usata

linguaggio naturale *(Natural Language Processing, Natural Language Understanding, Cognitive Computing).*

*«In termini tecnici, l'Intelligenza Artificiale è un ramo dell'informatica che permette la programmazione e progettazione di sistemi sia hardware che software che, a loro volta, permettono di dotare le macchine di determinate caratteristiche che vengono considerate tipicamente umane quali; ad esempio: le percezioni visive, spazio-temporali e decisionali».*[66]

Questo tipo di intelligenza non va intesa solamente come capacità di calcolo o di conoscenza di dati astratti, bensì anche (e soprattutto) di tutte quelle forme di intelligenza riconosciute dallo psicologo Howard Gardner nella sua teoria delle intelligenze multiple[67], e che vanno dalla intelligenza corporeo-cinestetica a quella inter-/intra-personale, da quella musicale a quella logico-matematica, nonché individuando una intelligenza linguistica[68].

*«Un sistema intelligente viene realizzato cercando di ricreare una o più di queste forme di intelligenza che, anche se spesso ancora definite come semplicemente umane, possono essere ormai ricondotte a comportamenti riproducibili anche dalle macchine».*[69]

L'utilizzo delle reti neurali e la creazione di complessi algoritmi in grado di riprodurre ragionamenti prima solo adducibili agli esseri umani hanno permesso ai sistemi intelligenti di migliorare sempre di più le proprie capacità di *performance*.

Tali algoritmi ambiscono ad imitare comportamenti, azioni procedurali e cognitive a carattere umano a seconda degli stimoli ambientali (l'ambiente è un contesto imprescindibile per l'attuazione di processi cognitivi nell'essere umano). Le macchine stanno così diventando sempre più in grado, in questo modo, di auto-apprendere e di prendere loro stesse delle decisioni, ovvero di effettuare scelte a seconda dei contesti in cui sono inserite, dei quali imparano ad avere ricognizione, e nei quali si adattano.

---

[66] S. Rossi, *L'Intelligenza Artificiale,* Rai Cultura, 2020. Cfr. www.raicultura.it/raicultura/articoli/2020/03/Intelligenza-artificiale.html

[67] Si rimanda al paragrafo 2.1 *"Intelligere – processi e stili cognitivi"*.

[68] Intesa come *«abilità che si esprime nell'uso del linguaggio e delle parole, nella padronanza dei termini linguistici e nella capacità di adattarli alla natura del compito»*. Cfr. https://www.lumsa.it/sites/default/files/Lezione%20del%2027-03-2020%20GARDNER.pdf

[69] S. Rossi, *L'Intelligenza Artificiale,* Rai Cultura, 2020. Cfr. www.raicultura.it/raicultura/articoli/2020/03/Intelligenza-artificiale.html

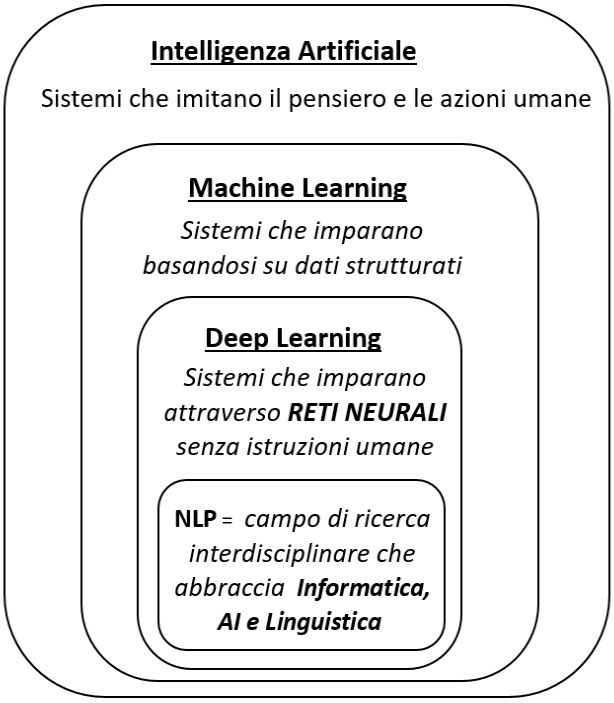



***Figura 3 – Ricostruzione tassonomica dell'applicazione dell'AI nella linguistica***[70]

Inoltre, la ricerca nel campo dell'A.I. ha iniziato ad indagare una delle dimensioni più profonde della mente umana: la metacognizione, ovvero la capacità di imparare ad apprendere, nonché il *«saper astrarre da un dominio specifico di conoscenza delle strategie per risolvere una classe di problemi e applicarle in contesti nuovi e differenti».*[71]

Per rispondere all'esigenza di creare algoritmi sempre più precisi e complessi per poter permettere alla macchina di compiere dei ragionamenti automatici, è stato da poco creato una nuova branca dell'intelligenza artificiale, denominata "rappresentazione della conoscenza", che consiste nella definizione dei simbolismi o dei linguaggi che permettono di formalizzare la conoscenza *(knowledge, o know-how)* in modo astratto.

La rappresentazione della conoscenza studia *«tutte le possibilità di ragionamento di un individuo e, soprattutto, tutte le possibilità di rendere tale conoscenza comprensibile alle macchine tramite un linguaggio e dei comandi sempre più precisi e dettagliati».*[72]

[70] Fonte: grafico autoprodotto.
[71] M. Catalano, *L'intelligenza artificiale e il potere cognitivo delle metafore: spunti di riflessione per una didattica innovativa*, per *ictedmagazine.com*, 2021. Cfr. https://www.ictedmagazine.com/index.php/ricerca-e-innovazione/243-l-intelligenza-artificiale-e-il-potere-cognitivo-delle-metafore-spunti-di-riflessione-per-una-didattica-innovativa.html
[72] A. Minini, *Personal Knowledge Base,* 2021. Cfr. https://www.andreaminini.com/ai/

Le informazioni rappresentanti la conoscenza dell'essere umano vengono trasferite alla macchina tramite diverse modalità; le più importanti delle quali sono *«quelle che si basano sulla teoria dei linguaggi formali e sulla teoria delle decisioni».*[73]

Nella teoria dei linguaggi formali, si sceglie l'approccio migliore per mettere a punto il trasferimento. Gli approcci più noti sono quello generativo, riconoscitivo, denotazionale, algebrico e trasformazionale. Ad oggi, si fa riferimento alle teorie delle stringhe e ai loro utilizzi. *«Le stringhe rappresentano dei veri e propri linguaggi formali le cui proprietà variano proprio a seconda dell'approccio utilizzato».*[74]

La teoria delle decisioni, invece, si basa su modelli predittivi a partire da una serie di dati di partenza. Tali dati vengono disposti in una "struttura ad albero" – che richiama alla mente l'albero sintattico di una stringa di testo, realizzato in accordo a determinate forme grammaticali – definito "albero di decisione". Un albero di decisione permette di valutare per ogni azione e/o decisione tutte le possibili conseguenze; per questo, l'analista può vagliare quale sia il percorso più conveniente da intraprendere.

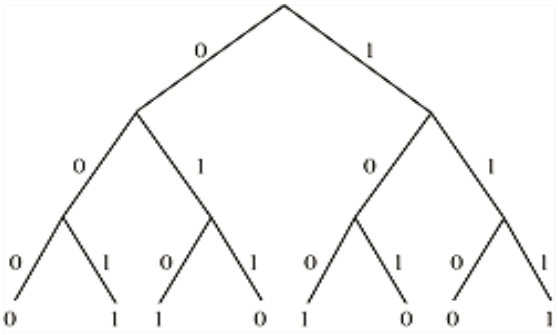


***Figura 4 – Albero di decisione nell'intelligenza artificiale***[75]

Sono i dati forniti di partenza, quindi, a generare una determinata struttura tassonomica ad albero*;* altresì, oltre a rivestire un carattere predittivo, questa struttura permette di verificare l'accuratezza delle previsioni. È l'accuratezza stessa dei dati a generare sistemi intelligenti. Questi ultimi si differenziano tra di loro a seconda delle risposte che sono in grado di fornire in *output*.

[73] Dal portale dedicato all'intelligenza artificiale: https://www.intelligenzaartificiale.it/
[74] G. L. Cascio Rizzo, *Sentiment Analysis: algoritmi di clustering e metodi di classificazione supervisionata,* Università LUISS Guido Carli, p. 28, in riferimento a C. Triberti, M. Castellani, *L'intelligenza artificiale oltre le quattro leggi della robotica. Riflessioni anche alla luce della pandemia da COVID-19*, Ed. GoWare, 2020.
[75] Fonte dell'immagine: www.intelligenzaartificiale.it/intelligenza-artificiale-cose-come-funziona-e-a-cosa-serve/attachment/albero-decisione-intelligenza-artificiale/

Bisogna sottolineare nuovamente che la mole e la qualità dei dati a disposizione ai fini dell'elaborazione di strutture intelligenti può interferire con la precisione del modello utilizzato, qualora non opportunamente filtrati e vagliati in prima istanza.

Per questo motivo, *«i modelli più accurati presentano un numero di informazioni di partenza spesso inferiore a quello che si può immaginare»*[76]*:* la "bontà del modello" – proprio come la "bontà dell'esemplare" in linguistica, afferente alla teoria del prototipo di Eleanor Rosch[77] – viene assicurata dalla tipologia di dati di partenza, dalla loro selezione, nonché, come precedentemente definito, dall'accuratezza degli stessi.

Secondo il modello del prototipo *«dai singoli esemplari di una categoria vengono estratte delle proprietà, ciascuna delle quali possiede dei valori* [e] *di cui si è avuta esperienza»*[78]: ovvero, i concetti vengono classificati, accettati e compresi da una comunità linguistica sulla base delle esperienze similari condivise. Questo tipo di modello di ricognizione linguistico-cognitiva viene individuato anche da Charles J. Fillmore (1929 – 2014) con il nome di *Frame Semantics.*

Ne parlerò nel paragrafo 2.4 *"Frames e modelli cognitivi idealizzati nell'A.I."* del presente lavoro di tesi.

[76] https://www.intelligenzaartificiale.it/ Per un approfondimento: E. Taufer, Alberi *di decisione*, Università di Trento, 2017. Cfr. http://www.cs.unitn.it/~taufer/Labs/L8-Alberi_di_decisione.pdf

[77] Nata a New York nel 1938, E. Rosch è una psicologa statunitense specializzata in psicologia cognitiva. È nota principalmente per i suoi studi sulla categorizzazione cognitiva e per la sua teoria dei prototipi.

[78]A. M. Borghi, *L'organizzazione della conoscenza, aspetti e problemi*, Pitagora Editore, Bologna, 1997, p. 12

## *1.4 Algoritmi e linguistica: proprietà condivise*

*«Algoritmi e scenari: il mondo è basato sempre di più sulle previsioni»*[79]

L'algoritmo[80] è un concetto fondamentale nel mondo dell'informatica, e costituisce la base della nozione teorica di calcolabilità; è, inoltre, un concetto cardine anche nella fase di programmazione e di sviluppo di un *software*. Nella sua veste generale, è una semplice procedura che ha come obiettivo quello di risolvere un determinato problema applicando un certo numero di passi elementari, ovvero una sequenza finita di operazioni, definite anche "istruzioni". L'algoritmo, quindi, opera su dati variabili e grezzi *(raw data)* in *input* per poterli rendere calcolabili, quantificabili, misurabili secondo il linguaggio dei numeri. Per poter raggiungere tale obiettivo, un algoritmo deve soddisfare tutte le sue condizioni di validità, ovvero, deve possedere tutte le proprietà fondamentali che lo definiscono come tale. Dette proprietà sono:[81] la finitezza, l'atomicità, la terminazione, l'effettività, il determinismo e la non ambiguità.

Linguisticamente potremmo associare il concetto di algoritmo alla realizzazione di un atto linguistico: entrambe le condizioni richiedono dei passi ben definiti (e finiti, in termini di tempo) al fine di rendere calcolabile, comprensibile e riproducibile una esecuzione (linguistica o matematica) in uscita *(in output),* dopo aver processato in maniera funzionale e accurata i dati forniti in entrata *(in input).*

Un atto linguistico – inteso come tentativo di comunicazione tra due o più individui – così come altre manifestazioni di trasferimento linguistico (tra cui il processo di traduzione), condivide la stessa natura degli algoritmi: sono attività di risoluzione di problemi. Ma, come precedentemente definito nel presente lavoro di tesi: *«Al contrario del processo di elaborazione dei dati forniti in input in un sistema computazionale, l'atto di produzione e di ricezione di un messaggio, è tutt'altro che binaria: l'interazione (e comunicazione) umana si configura come un sistema complesso – adattativo/adattivo – strutturato da relazioni*

[79] M. Caligiuri, *Cyber Intelligence, la sfida dei data scientist*, 22 giugno 2016. Cfr. www.sicurezzanazionale.gov.it/sisr.nsf/approfondimenti/cyber-intelligence-la-sfida-dei-data-scientist.html
[80] Il termine "algoritmo" deriva dalla trascrizione latina del nome del matematico persiano al-Khwarizmi, considerato uno dei primi autori ad aver fatto riferimento a questo concetto nel suo libro "Regole di ripristino e riduzione" – Università Niccolò Cusano, *Cos'è un algoritmo in matematica e informatica: quello che devi sapere*, 2018. Cfr. https://www.unicusano.it/blog/didattica/corsi/cose-un-algoritmo/
[81] Ibidem.

*reciproche tra parlanti e ascoltatori che si muovono frequentemente tra strutture diverse, e spesso le fondono».*[82]

La comunicazione, infatti, è quell'attività che ci permette di mandare segnali che, a loro volta, generano risposte significative in un determinato ambiente, società e/o contesto. Quest'ultimo dev'essere condiviso da tutti gli attori che prendono parte all'azione linguistica. Si attua, così, un processo interazionale e finito di azioni e retroazioni *(feedbacks),* nonché un processo di *problem solving* più o meno strategico. Infatti, come anche definito da Mario Caligiuri[83] nel suo libro "Cyber Intelligence, la sfida dei data scientist"*: «La rapidità dei mutamenti sociali richiede di saper anticipare gli eventi, perché chi conosce prima ha un vantaggio* [...] *rispetto agli altri».*[84]

Riprendendo la lista delle proprietà fondamentali degli algoritmi delineata nella pagina precedente, ho definito le stesse istruzioni di esecuzione di dati e di informazioni fornite in *input* in una matrice linguistica – al fine di ottenere dei risultati validi e funzionali in un contesto comunicativo – analizzando, così, uno dei problemi primordiali dell'essere umano: come comunicare in maniera efficace.

Tenendo in considerazione, ancora una volta, l'assenza di forze convenzionali nella matrice comunicativa umana, che non restituisce (né garantisce) un *output* binario (0-1) in risposta alla formulazione di un enunciato, in linguistica, detto enunciato si considera potenzialmente funzionale allo scopo, esclusivamente se esso rispetta, nelle modalità di seguito descritte, le stesse caratteristiche che validano gli algoritmi; ovvero, i principi di:

- atomicità: i passi dell'algoritmo devono essere elementari, ovvero non possono essere ulteriormente divisibili; ciò si riflette nella natura della parola (l'unità più piccola, indivisibile, portatrice di significato) e del *token* (unità definibile come sequenza di caratteri delimitata da spazi; è inteso come sinonimo di "parola" nei motori semantici);
- non ambiguità: così come i passi dell'algoritmo non possono essere interpretati in altri modi se non in quelli formalmente definiti, anche le parole (o *tokens*, in linguistica computazionale) non devono essere soggette ad ambiguità. La disambiguazione[85] è il

[82] Nel testo, p. 13
[83] Professore Ordinario di Pedagogia all'Università della Calabria, è considerato uno dei massimi studiosi europei di intelligence a livello accademico. È Presidente della Società Italiana di Intelligence (SOCINT).
[84] M. Caligiuri, Cyber *Intelligence, la sfida dei data scientist,* 2016. Cfr. www.sicurezzanazionale.gov.it/sisr.nsf/approfondimenti/cyber-intelligence-la-sfida-dei-data-scientist.html
[85] Definita in computazione anche come *Word Sense Disambiguation.*

processo con il quale si precisa il significato di una parola o di un insieme di parole. Una parola può denotare significati diversi, predicibili a seconda del contesto. Uno dei problemi più rilevanti del processo di disambiguazione riguarda infatti la polisemia: ovvero, quando una parola assume più significati;

- finitezza: così come un algoritmo deve essere svolto in un numero specifico di passi, *«una grammatica generativa per un linguaggio <L> è un insieme finito di istruzioni e di regole che permettono di generare e riconoscere tutte e sole le frasi appartenenti a <L> e di assegnare a queste frasi un'adeguata descrizione strutturale».*[86]*;*

- terminazione: per quanto concerne gli algoritmi, l'esecuzione del loro schema deve terminare entro un certo periodo di tempo. Anche in linguistica, la realizzazione di un qualsiasi atto linguistico (che sia rappresentativo, direttivo, assertivo, espressivo o dichiarativo), è reso possibile solamente nel momento in cui si stabilisce una fine (attraverso segni di interpunzione o, se oralmente, attraverso l'intonazione) al proprio enunciato. Solo così, si può funzionalmente provocare uno stimolo sul mondo circostante, generando un'interazione con un interlocutore. Altresì, la condizione di finitezza si può applicare in linguistica anche in un'altra dimensione: la produzione di ogni singolo enunciato è unico nel proprio genere, forma e realizzazione, poiché è fortemente caratterizzato ed influenzato da specifiche coordinate spaziali e temporali. Quindi, un'interazione linguistica ha generalmente luogo in un contesto temporale e spaziale che conferiscono identità, forma e struttura all'atto comunicativo stesso;

- effettività: l'esecuzione dello schema algoritmico deve condurre ad un unico *output*. In riferimento alla teoria degli atti linguistici di Austin e Searle, questa proprietà trova un riscontro anche nell'ambito linguistico: si definisce "atto perlocutivo"[87] l'effetto dell'atto linguistico[88] prodotto sull'interlocutore. La perlocuzione è costituita dalle conseguenze provocate, dai risultati ottenuti tramite l'atto illocutorio[89].

---

[86] C. Chesi, *Linguistica Computazionale: gli strumenti linguistico-formali & informatici*, 2009, p. 7.

[87] La perlocuzione è uno dei livelli nell'uso del linguaggio espressi dal linguista J. L. Austin nelle sue lezioni del 1955, successivamente raccolte nell'opera "How to do things with words" ("Come fare cose con le parole"). L'atto perlocutivo (o "perlocutorio") è il terzo dei livelli indicati, dopo quello locutivo ed illocutivo. È incentrato sul destinatario e si può definire come la conseguenza oggettiva provocata dall'atto linguistico.

[88] Pur essendo inteso come "unitario", in un atto linguistico si possono distinguere tre livelli principali: 1) la locuzione, che comprende tutti quegli *«aspetti per cui parlare è un dir qualcosa»* – L. Treccani, *Teorie e analisi degli atti linguistici*, Università Ca' Foscari Venezia, 2016, p. 7; 2) l'illocuzione, che consiste nell'effetto cui il parlante tende con il suo messaggio; 3) la perlocuzione, che riguarda la produzione delle conseguenze sulla situazione in cui avviene l'atto linguistico stesso.

[89] L'atto illocutorio (o "illocuzione") può essere definito come atto di accertamento della ricezione; ovvero, è una forza convenzionale con la quale si compie l'atto locutivo (l'emissione linguistica) per raggiungere un

## *1.5 NLP: elaborazione del linguaggio naturale*

La tecnologia di elaborazione del linguaggio naturale (NLP o *Natural Language Processing)* è *«un campo di ricerca interdisciplinare che abbraccia informatica, intelligenza artificiale e linguistica, con lo scopo di sviluppare algoritmi in grado di analizzare, rappresentare e quindi "comprendere" il linguaggio naturale, scritto o parlato»*[90], realizzando questa abilità in maniera simile a quella propria degli esseri umani.

Il processo di trattamento (semi)automatico delle informazioni e delle risorse testuali redatte in una lingua naturale è eseguito mediante specifici *software*.

L'intelligenza linguistica indotta mediante NLP è un'attività (semi)automatica molto complessa, e *«a rendere particolarmente difficoltosa la comprensione del linguaggio umano da parte di un algoritmo informatico, contribuiscono le sue intrinseche caratteristiche di ambiguità».*[91] Infatti, per poter comprendere il prodotto di una qualsiasi interazione linguistica, anche nel caso di comunicazioni monologiche e messaggi unidirezionali, è necessario non solo che i fruitori della lingua condividano lo stesso codice linguistico, ma anche che la loro comprensione della realtà e conoscenza del mondo confluiscano. Conoscere il significato di ogni singola parola non è, quindi, sufficiente per interpretare correttamente né il messaggio di una frase semplice, né il contenuto di un intero documento. Infatti, nell'azione di analisi semantica (semi)automatizzata, una delle procedure necessarie alla corretta attribuzione di un significato ad una data espressione linguistica viene definita "disambiguazione".

Come si descriverà nel terzo capitolo di questo elaborato magistrale, relativo al *tool* di *Cognitive Computing* COGITO STUDIO®, l'attivazione del processo di disambiguazione (in inglese, *Word Sense Disambiguation)* è un'azione fondamentale per la risoluzione di polisemie o di categorizzazioni morfologiche imprecise delle parole; permette, quindi, la realizzazione di un modello lessicale in grado di descrivere la natura di verbi, sostantivi e aggettivi, discriminando significati opachi.

---

determinato obiettivo o per rispondere a una particolare esigenza. Si configura come secondo livello nell'uso del linguaggio sopra definito.

[90]M. Esposito, *Linguaggio naturale e intelligenza artificiale: a che punto siamo*, 2019. Cfr. www.agendadigitale.eu/cultura-digitale/linguaggio-naturale-e-intelligenza-artificiale-a-che-punto-siamo/

[91]G. Altobello, *Natural Language Processing, cos'è, come funziona e applicazioni*, 2021. Cfr. www.ai4business.it/intelligenza-artificiale/natural-language-processing-tutto-quello-che-ce-da-sapere/

Al fine di ridurre il più possibile gli errori di comprensione del linguaggio naturale da parte dell'elaboratore elettronico, al fine di realizzare una esaustiva *Information Extraction,* è stata resa necessaria la suddivisione del processo di analisi in diverse fasi di azione[92], riconducibili anche ai linguaggi di programmazione:

- fase di analisi lessicale (*Tokenization* o *tokenizzazione):* si realizza la scomposizione di un'espressione linguistica in *tokens*[93]. Come precedentemente menzionato, tale procedura è eseguibile avendo in *input* una lingua a sistema alfabetico[94];
- fase di analisi morfologica *(Part-of-Speech* o POS): le parti del discorso (articolo, nome, aggettivo, pronome, verbo, avverbio, congiunzione, preposizione, interiezione) vengono associate a ogni parola/*token* presente nel testo;
- fase di analisi sintattica (attività di *Parsing):* viene eseguita una disposizione dei *token* mediante una struttura definita "ad albero" (o *parse tree).* Tale analisi assegna al periodo fornito in *input* una struttura osservante le singole unità del discorso e le relazioni che intercorrono tra loro, individuando, così, sintagmi e proposizioni.
- fase del riconoscimento delle entità nominate *(Named Entity Recognition* o NER): ovvero, l'identificazione, l'estrazione e la classificazione automatica in categorie delle entità presenti nel *corpus*. Alcune entità sono generalmente categorizzate in: persone *(people),* luoghi *(places)* e organizzazioni *(organizations);*
- fase dell'analisi semantica: vengono assegnati ed associati dei significati alle singole parole componenti il periodo, sfruttando l'interazione con il vocabolario implementato a sistema. Riveste fondamentale importanza il processo disambiguazione semantica.
- fase dell'analisi dei sentimenti *(Sentiment Analysis):* questa fase ha lo scopo identificare opinioni da fonti aperte (es. pagine web e *social media)* e mira a determinare l'atteggiamento di un autore rispetto a un determinato argomento. Obiettivo di questa attività è l'individuazione della polarità contestuale complessiva (positiva, negativa o neutra) di documenti e/o singoli dati testuali. A volte, fattori culturali, sfumature linguistiche e/o contesti opachi rendono estremamente ardua l'attribuzione di un *sentiment* in maniera completamente affidabile ed imparziale.

[92] Nel testo, p. 78 e p. 108.
[93] In linguistica computazionale *token* viene inteso come sinonimo di "parola".
[94] Sarà diversamente eseguibile nelle lingue a sistema ortografico continuo. Nel testo, p. 27.

In sintesi, quindi, l'NLP fornisce soluzioni per *«analizzare la struttura sintattica di un testo, associando alle singole parole le rispettive categorie morfologiche, identificando entità e classificandole in categorie predefinite, estraendo dipendenze sintattiche e relazioni semantiche».*[95] Inoltre, tale induzione linguistica consente di comprendere la semantica del testo persino quando relazionato al contesto e alle modalità di utilizzo, nonché di classificare le informazioni presenti in un documento in categorie concettuali.

Negli ultimi anni si è assistito alla creazione di nuovi approcci in grado di integrare l'elaborazione del linguaggio naturale con algoritmi di *Deep Learning* (apprendimento profondo), sottocategoria del *Machine Learning* (apprendimento automatico), producendo risultati estremamente efficaci in molteplici scenari e domini di applicazione.

Il processo di creazione di algoritmi di *Deep Learning* è ispirato alla struttura e al funzionamento delle reti neurali cerebrali (come precedentemente indicato, detto processo dà vita alle *Artificial Neural Networks:* reti neurali artificiali), assimilabili a modelli di calcolo matematico-informatici costituiti da interconnessioni di informazioni. Si tratta di un *«sistema "adattivo" in grado di modificare la sua struttura basandosi sia su dati esterni sia su informazioni interne».*[96]

Un primo approccio basato sul *Deep Learning* venne applicato nel 2011 al fine di risolvere alcuni problemi di interpretazione circoscritti all'ambito dell'elaborazione del linguaggio naturale, come ad esempio una identificazione più profonda delle entità e un raffinamento della relazione semantica che intercorre tra esse. Questa esigenza nacque dalla comprensione dell'estrema singolarizzazione di alcuni contesti di cyber investigazione, caratterizzati da elementi semantici altamente specifici e/o settoriali tra loro interconnessi. Ciò condusse gli analisti all'acquisizione di una strategia che potesse realizzare una più puntuale personalizzazione e un costante re-indirizzamento di azione.

È dalla tecnologia dell'NLP, insieme ai sistemi di *Deep/Machine Learning* e di riconoscimento di *patterns*, che nasce il *Cognitive Computing*, scienza alla base del motore semantico utilizzato per la creazione del caso di studio oggetto del quarto capitolo.

---

[95] T. Buonocore, *NLP in Medicina*, 2021. Cfr. https://www.biomeris.it/nlp-in-medicina/

[96] N. Boldrini, *Reti Neurali: cosa sono e a cosa servono,* AI4BUSINESS, Network Digital 360, 2022. Cfr. https://www.ai4business.it/intelligenza-artificiale/deep-learning/reti-neurali. Per un approfondimento: N. Boldrini, *A.I. – Artificial Intelligence: Come è nata, come funziona e come l'Intelligenza Artificiale sta per cambiare il mondo, la vostra vita e il vostro lavoro*, Class Editori, 2018.

### *1.5.1 Types e Tokens*

Ho introdotto in questa tesi di Laurea l'obiettivo di voler sottolineare *«l'importanza della matrice cognitiva nei processi di classificazione delle informazioni: è nella necessità di fruire di types e nell'impossibilità di vedersi garantiti dei tokens immutabili nel tempo, che ogni persona comprende e concettualizza la propria realtà».*[97]

L'elaborazione computazionale di una qualsiasi risorsa, come già esplicitato, ha inizio con la scomposizione dell'informazione testuale in *tokens*, ovvero in occorrenze linguistiche – intese come sinonimi di parole, poiché delimitate, come queste ultime, da spazi o segni di interpunzione. Ma conoscere cos'è un *token* non è sufficiente per eseguire una esauriente decodifica di un segno linguistico[98].

Secondo la semiotica (disciplina che studia i segni e il modo in cui questi abbiano una significazione), per ogni segno vige la distinzione tra *type* e *token*[99]. Attuare tale classificazione permette di comprendere la natura della relazione che intercorre tra questi elementi: per *types* si intendono tutte le entità astratte che fanno da modello cognitivo per la realizzazione di *tokens*, ovvero la realizzazione nel concreto di ogni singola occorrenza linguistica. In altre parole, questa distinzione è necessaria al fine di poter *«distinguere i predicati che sono attribuiti in abstracto ad un tipo di oggetto* [type], *dai predicati che sono attribuiti in concreto ad un singolo oggetto individuale* [token]».[100]

La relazione tra "tipi" e "occorrenze" rimanda alla dicotomia tra *langue* e *parole*[101] di Ferdinand De Saussure, ovvero, tra l'aspetto sociale e oggettivo del linguaggio inteso come codice linguistico – il sistema di segni che formano il codice di un idioma, vale a dire, l'insieme delle convenzioni adottate dai membri di una comunità per comunicare tra loro – e la realizzazione individuale e soggettiva di ogni singola esecuzione linguistica[102].

---

[97] Nel testo, p. 5

[98] Un segno è, secondo Ferdinand de Saussure, l'unione di significante e significato. Ovvero, il prodotto di un processo logico-cognitivo che, a partire dalla registrazione di un evento da parte di un interprete, viene stimolata la sua attività inferenziale per approdare alla formulazione di un'ipotesi esplicativa del fenomeno in questione.

[99] La coppia *type/token* è stata introdotta in semiotica dal matematico, filosofo, semiologo, scienziato e accademico statunitense Charles Sanders Peirce.

[100] L. Glazer Passerini, *Impossibilità di Tokens, necessità di Types*, 2013, p. 83. Cfr. https://www.ledonline.it/ledonline/761-impossibilita-normativa/761-impossibilita-normativa-passerini.pdf

[101] Tale dicotomia ha origine nella seguente opera: F. De Saussure, *Cours de linguistique générale*, 1916.

[102] Il linguista Eugenio Coseriu, negli anni '70, introdusse un termine intermedio tra *langue et parole*: il concetto di *norma*, da non intendersi come insieme di regole da rispettare, bensì come condizione singola da rispettare affinché un enunciato risultasse *corretto*.

La *langue*, così come i *types*, rimandano al concetto di modellizzazione astratta e discreta della realtà; mentre, la *parole*, così come i *tokens*, sono la rappresentazione individuale ed unica del primo fenomeno nel concreto.

*«Un type è una forma significante che determina cose che esistono; esso, tuttavia, non esiste in sé come cosa singola* […], *bensì, per poter essere usato, esso deve necessariamente essere istanziato in un token. La relazione tra un type e i suoi tokens è, dunque, una relazione di istanziazione».*[103]

La soggettività della produzione di ogni esecuzione linguistica, e l'individuazione di determinate occorrenze per poterla realizzare, è dovuta ai modelli cognitivi di cui ogni individuo dispone. Per "modello" si intende una rappresentazione mentale *«semplificata della realtà, ma contestualizzata in situazioni specifiche. Un individuo, anche inconsciamente, tende a fare esperienza di ciò che osserva e lo classifica in un suo personale schema».*[104]

La modellizzazione è un processo cognitivo che conduce alla creazione di modelli, ovvero alla rappresentazione di categorie teoriche o mentali, che sono indispensabili all'essere umano al fine di poter interpretare (mediante la formulazione di inferenze) i segni iconici *in abstracto.*

La formulazione di inferenze è una procedura (a carattere induttivo o deduttivo) di attivazione dell'informazione che non prevede alcun emittente *volontario,* bensì solamente la presenza di oggetti o entità che vengono interpretati come messaggi. Tale processo permette la comprensione del significato delle informazioni non esplicite, rintracciabile unicamente attraverso il recupero delle informazioni correlate a tale oggetto o entità.

A metà degli anni '50, l'economista Herbert A. Simon postulò la teoria delle decisioni a razionalità limitata, secondo cui gli individui non sarebbero veramente in grado di operare delle scelte a livello logico, poiché ininterrottamente condizionati da tre circostanze che si influenzano a vicenda: *«L'impossibilità di possedere tutte le informazioni, i tempi ridotti per assumere le decisioni e i limiti cognitivi individuali».*[105]

---

[103] L. Glazel Passerini, *Impossibilità di Tokens, necessità di Types,* 2013, p. 85. Cfr. https://www.ledonline.it/ledonline/761-impossibilita-normativa/761-impossibilita-normativa-passerini.pdf

[104] G. Negri, *Mental Framing: le conseguenze di un'errata valutazione del modello mentale,* Università Ca' Foscari di Venezia, 2018, p. 4.

[105] M. Caligiuri, *Cyber Intelligence – Tra libertà e sicurezza*, Donzelli Editore, 2016, p. 3.

Modellizzare la realtà che ci circonda è un'attività indispensabile, naturalizzata nell'essere umano e altamente soggettiva, poiché profondamente influenzata dall'acquisizione (e dalla successiva applicazione) di abitudini culturalmente acquisite e di esperienze vissute, in maniera individuale e/o sociale.

In quanto categorie astratte ed altamente personali, i modelli non comprendono tutti gli aspetti della realtà. Inoltre, data la loro natura dinamica – la cui plasticità si riflette nuovamente nel "concetto-paradosso" di *langue*, ovvero di *Système ou tout se tient*[106], la cui stabilità *«è dovuta al costante mutamento esercitato dal tempo e dalla continua* rinegoziazione [di forme e significati] *ad ogni atto di parole»*[107] – ogni modello cognitivo è soggetto a cambiamenti, riscontrabili e valutabili sia in chiave sincronica che diacronica, nonché attraverso parametri dettati dalle varietà linguistiche e da fattori sociali ed extra-linguistici. È per questo che l'essere umano vive nella necessità di fruire di *types*, ovvero di categorie concettuali personali, e nell'impossibilità di vedersi garantiti dei *tokens* immutabili nel tempo, perché la loro stabilità è dovuta (paradossalmente) alle continue negoziazioni realizzate in ogni atto comunicativo[108].

Sono, quindi, infinite le rappresentazioni di un modello in fase di interiorizzazione, poiché esso si configura come il prodotto di un processo inferenziale (ovvero, interpretativo) dei dati sensoriali percepiti e compresi dall'individuo nell'ambiente a lui circostante; si tratta di un'operazione di concettualizzazione molto complessa, formulata e resa possibile solo attraverso giudizi percettivi individuali[109]. L'esperienza è un fattore determinante nel processo cognitivo di percezione.

*«Questo riporta alla ribalta la necessità dell'integrazione dei saperi, richiedendo la fusione di competenze umanistiche e scientifiche. La Cyber Intelligence è un terreno di analisi, di impegno primario e urgente interesse. Un elemento fondamentale per la cultura della sicurezza»*[110]

[106] «*Un sistema in cui tutto si tiene insieme*» – Cfr. F. de Saussure, *Cours de linguistique générale,* 1916.
[107] M. Barbera, *Introduzione alla linguistica Generale,* 2009. Cfr. www.bmanuel.org/corling/corling_idx.html
[108] Nel testo, p. 5
[109] Si rimanda alla teoria della *Gestalt* (o della "forma") che si impegna a comprendere come la percezione umana selezioni in maniera soggettiva gli stimoli, categorizzandoli in "figura" (ovvero, ciò che ci interessa e/o che riusciamo a comprendere) e "sfondo".
[110] M. Caligiuri, *Cyber Intelligence, la sfida dei Data Scientist,* 22 giugno 2016. Cfr. www.sicurezzanazionale.gov.it/sisr.nsf/approfondimenti/cyber-intelligence-la-sfida-dei-data-scientist.html

Il processo di modellizzazione viene applicato nell'ambito della sicurezza (nazionale, internazionale o dei sistemi), con lo scopo di identificare, classificare e analizzare potenziali minacce, attraverso l'adozione di un determinato approccio – e conseguente metodo – per il rilevamento delle stesse, la valutazione di rischi e vulnerabilità[111], e per la previsione di contromisure. Si parla, in questo caso, di *Threat Modeling.*

È da questa categorizzazione e analisi delle minacce che nasce, tra le varie strategie adottate, un programma di *Intelligence* nel campo della *Cyber Security* dedicato al rilevamento delle minacce informatiche: la *Cyber Threat Intelligence* (CTI), come verrà discusso nel quarto capitolo.

[111] «*Un exploit sfrutta i bug e/o le vulnerabilità di un sistema operativo o di un'applicazione software per ottenere l'accesso non autorizzato delle informazioni e dei dati al loro interno*». – A. Raffaelli, *Data Breach: una sfida tecnologica, legale ed etica per il futuro*, Università Campus Bio-Medico di Roma, 2022.

## *1.6 Text Data Mining: estrazione di dati dal testo*

La tecnica di analisi del *Text Mining* (o *Text Data Mining* ovvero, "estrazione di dati dal testo") pertiene all'ambito della *Cyber Intelligence* e utilizza l'NLP per trasformare testi non strutturati (ovvero, documenti non aventi un *template* specifico – come, ad esempio, un documento in formato word, un'e-mail, un *post* sui social media) in testi strutturati e normalizzati (ad esempio, in formato tabulare, di tipo non continuo).

Il *Text Mining* unisce la tecnologia della lingua con gli algoritmi del *Data Mining* – ovvero, del processo di individuazione ed estrazione delle informazioni da grandi banche dati. Sono molteplici gli strumenti di analisi di testi ad oggi disponibili: da quelli completamente automatizzati (*full A.I.)* a quelli ibridi, in cui la componente umana è ancora necessaria nel processo di interpretazione del *Big Data* testuale.

Tutte le attività che produciamo nel *cyberspazio* vengono memorizzate ed archiviate sotto forma di *Big Data*[112], ovvero sotto forma di un enorme e complesso flusso di informazioni che generiamo con il solo navigare in rete.

Inconsapevolmente, ad ogni *click* lasciamo tracce di tutte le ricerche e scelte operate online, e permettiamo alle compagnie che ci offrono servizi *(Google, Twitter, Facebook,* per menzionarne alcuni) di documentare la nostra vita quotidiana e di risalire al nostro comportamento personale e sociale. Secondo Viktor Mayer-Schönberger e Kenneth Cukier: «[Grazie ai Big Data] *la società dovrà abbandonare la sua ossessione per la casualità*».[113]

Ricerche scientifiche, infatti, hanno dimostrato una sorprendente capacità dei *Big Data* nella predizione delle attività umane, sia reali che digitali. Ad esempio, analizzando determinate azioni svolte in passato da un utente, è possibile prevedere *«in modo sorprendentemente accurato gli spostamenti degli individui e le attività che questi svolgono nella vita di tutti i giorni»*[114]*:* non solo, è possibile individuare persino lo scopo del suo movimento attraverso la creazione di un modello matematico, nonché tracciare le relazioni intessute dall'utente sui *social networks* o al dispositivo cellulare, mediante un'accurata analisi linguistica di ogni conversazione effettuata.

[112] I *Big Data* sono caratterizzati dalle "cinque V": volume, varietà, veridicità, velocità e valore.
[113] V. Mayer-Schönberger, K. Cukier, *Big Data. Una rivoluzione che trasformerà il nostro modo di vivere e già minaccia la nostra libertà*, Garzanti, Milano, 2013, pp. 15-16.
[114] L., Pappalardo, F. Giannotti, *Capire la mobilità attraverso i big data*, Donzelli Editore, 2015, p. 25.

Compito della *Cyber Intelligence* è anche quello di comparare le azioni illogiche ed imprevedibili degli esseri umani con le informazioni contenute nei *Big Data*, in funzione degli interessi nazionali.

I principali problemi di *Text Data Mining* sono la selezione dei dati, la loro trasformazione, analisi e interpretazione. Un *Text Analytics* applica le tecnologie afferenti al *Natural Language Processing* e/o al *Machine Learning* al fine di identificare ed estrarre informazioni dai dati e testi non strutturati. Un *Data Scientist* è in grado di comprendere come analizzare un insieme di dati complessi per derivarne informazioni strategiche e affrontare decisioni definite *data driven.*

Alla base del *Text Mining* vi è la tecnica dell'*Information Extraction*, ossia l'estrazione automatica (detta anche "localizzazione") di informazioni specifiche sotto forma di dati non strutturati espressi in linguaggio naturale che vengono trasformati in dati strutturati.

Gli obiettivi principali del *Text Mining* sono:

- l'individuazione di parole chiave, entità e significati, nonché delle loro eventuali associazioni e/o co-occorrenze implicite;
- la classificazione degli argomenti ad essi associati;
- l'assegnazione di una polarità alle entità estratte, ovvero l'attività di *Sentiment Analysis:* attraverso specifici algoritmi di linguistica computazionale viene rilevato il sentimento (positivo, negativo o neutro) socialmente condiviso della parola astratta; in questo caso, quindi, sfruttando la tecnologia della social media intelligence (SOCMINT);
- i dati estratti, inoltre, permetteranno l'addestramento di motori semantici e di ricerca, come descritto di seguito.

Il processo di *Text Mining* si articola generalmente nelle seguenti fasi:

1. fase dell'*Information Retrieval:* ovvero, del recupero delle informazioni (raccolta dati) che saranno oggetto di analisi;
2. fase dell'*Information Extraction*: prevede l'estrazione di informazioni dai documenti precedentemente raccolti e selezionati. Si procede all'indicizzazione, nonché alla suddivisione di ogni elemento testuale in frasi e in *token*s. Questi ultimi vengono ricercati all'interno di un dizionario al fine di poterne individuare le relazioni grammaticali, e di realizzare un'analisi lessicale e sintattica profonda del documento. A seguito

dell'identificazione delle parole chiave, oggetto di analisi, si effettuerà l'attività di *pruning,* ovvero il filtraggio completo delle informazioni non necessarie;

3. ottenuto in questo modo un *corpus* di dati, informalmente definiti *puliti,* si passa alla fase successiva dell'*Information Mining,* in cui viene applicato un determinato algoritmo di *Data Mining*. Vi sono varie classi di algoritmi a seconda dell'obiettivo di analisi da perseguire, tra cui gli algoritmi di *Machine Learning:* ovvero, di apprendimento supervisionato, non supervisionato o rinforzato per la classificazione automatica delle informazioni;
4. l'ultima fase, invece, è dedicata alla valutazione e all'interpretazione dei risultati ottenuti.

L'utilizzo delle tecnologie semantiche per l'analisi dei dati testuali *(Text Analysis)* e per l'estrazione di informazioni *(Text Mining),* porta grandi vantaggi:

- riduzione dei costi del servizio: soprattutto nel caso in cui si deve gestire una ingente mole di *set* di dati e in cui l'analisi presenta una ridondante reiterazione. Una volta creato e convalidato il *framework* di analisi, e una volta definite le regole di classificazione che permetteranno di individuare i *topics* più rilevanti all'interno di un documento, il processo viene automatizzato. L'unica fase in cui viene richiesto l'apporto di un operatore umano, riguarda la validazione *(testing)* dei risultati restituiti dal processo di analisi automatizzata;

- riduzione dei tempi di esecuzione: grazie all'automazione del processo di analisi e di classificazione delle informazioni, l'operazione di trasformazione ed estrazione delle informazioni richiede tempi esigui, i quali si riducono ulteriormente qualora si presentassero al sistema dei modelli di analisi precedentemente processati;

- coerenza: un'accurata automatizzazione del processo di analisi conduce ad una coerenza superiore in termini di classificazione, se messa a confronto con una individuazione dei *topics* eseguita manualmente; quest'ultima porta con sé il rischio di parzialità e soggettività delle azioni eseguite dall'operatore umano;
- adattabilità e riutilizzo dei modelli: una volta creato un *framework* di analisi per rispondere a un particolare *task*, *target* o desiderata del Cliente, questo può essere duplicato, adattato, applicato e riutilizzato a progetti o ad indagini di estrazione futuri;

- correlazioni tra fonti diverse e co-occorrenze tra entità: richiamando il punto precedente, il *framework* adottato per l'analisi del testo può essere applicato a molteplici fonti di dati, conducendo così ad un addestramento (con conseguente apprendimento) del sistema al fine di sviluppare (ed in seguito, automatizzare) l'abilità di correlare argomenti, nonché di definire le co-occorrenze tra entità per un'analisi predittiva e strategica;
- *Better Knowledge* – una conoscenza più profonda: l'analisi del testo tramite tecnologie semantiche non è solo un'alternativa più economica all'approccio manuale: guidando la macchina verso una memorizzazione e ricognizione dei dati forniti in *input* si aumenta in maniera consistente la capacità di individuazione ed estrazione di reti di significato, discernendone confini e domini di appartenenza da un'elevata mole e tipologia di testi.

Il rischio più grande nell'utilizzare l'intelligenza artificiale per lo sviluppo e l'applicazione di modelli e tecniche di analisi del linguaggio è la potenziale parzialità degli algoritmi applicati: questi potrebbero essere non del tutto *unbiased,* ovvero potrebbero risultare consapevolmente o inconsapevolmente "schierati". Questo perché dietro lo schermo – ovvero, dietro l'immensità del mondo artificiale – vi è la mente umana. Ogni individuo possiede una propria *Weltanschauung,* ovvero una propria concezione del mondo e della vita[115], che ne determina visioni, credenze e priorità nella presa di decisioni nelle più diverse situazioni.

Il bagaglio esperienziale ed enciclopedico di ognuno, così come ogni elemento relegato alla dimensione del culturale connaturata nell'essere umano (di cui la lingua è parte) rischiano di riflettersi negli algoritmi che egli stesso genera, e che applicherà nei sistemi intelligenti. Uno dei compiti degli analisti è proprio quello di validare l'imparzialità degli *output* restituiti dalla macchina, al fine eventuale di emendare la polarità e la simmetria (a carattere sociale, di genere, di ruolo) dell'elemento semantico estratto.

*«La maggior parte delle persone che programmano queste macchine sono maschi, bianchi, etero e questo porta a squilibri anche gravi nelle capacità delle intelligenze Artificiale di capire il mondo che le circonda. Rischiando di renderle inutili, se non addirittura pericolose».*[116]

[115] Termine definito nel dizionario Treccani come *«modo in cui singoli individui o gruppi sociali considerano l'esistenza e i fini del mondo e la posizione dell'essere umano in esso»*. Cfr. https://www.treccani.it/enciclopedia/weltanschauung_%28Dizionario-di-filosofia%29/

[116] E. Capone, *Il futuro visto da Annalisa Barla, la prof che insegna machine learning,* intervista ad Annalisa Barla per *Italian.Tech,* 2021. Cfr.

Per cercare di mitigare questi potenziali danni, alcuni esperti hanno individuato due elementi da considerarsi imprescindibili alla fase di sviluppo degli algoritmi: il primo elemento è la necessità di avere un *team* che sia variegato in termini di pensiero, genere, cultura ed esperienza di vario tipo; il secondo elemento richiede il soddisfacimento del fattore della spiegabilità[117]: *«Gli algoritmi ed i dati su cui vengono addestrati devono essere entrambi trasparenti»*[118], ergo, non ambigui.

Le tecniche di *Text Data Mining* sono applicabili a qualsiasi ambito di analisi e indagine; stanno attualmente vivendo grande crescita e diffusione grazie ai progressi della sua formalizzazione matematica nel campo della NLP e grazie alla sua conseguente implementazione nei motori di ricerca, negli *Application Service Providers* (ASP), nonché in specifici *software* di Intelligence.

Un *software* di Intelligence – nell'operare una ricerca mirata di informazioni e dati specifici all'interno di uno o più documenti – è, generalmente, in grado di processare un'ingente mole documentale con dei tempi di inferenza davvero esigui, eseguendo preliminarmente una lettura di tipo *scanning*, ovvero selettiva, eseguendo preliminarmente una lettura di tipo *scanning*, ovvero selettiva, mirata all'individuazione di *keywords* all'interno dei dati testuali forniti in *input*.

A seconda delle condizioni linguistiche definite in fase di investigazione e interrogazione semantica, l'operatore di Intelligence validerà la collezione eseguita dalla macchina di determinati dati afferenti ai campi semantici o a classi lessicali di interesse, come ad esempio: sostantivi (nomi) indicanti persone, luoghi, organizzazioni, date e altre micro-categorie semantiche previamente definite e variabilmente lontane dallo standard.

In questo caso si parla di vero e proprio *mining* delle informazioni, definito anche come "estrazione semantica" o "estrazione delle entità" *(Named Entity Recognition)*, reso possibile dall'implementazione per mano dell'analista linguistico/a di condizioni afferenti al linguaggio "E", ovvero al linguaggio "di estrazione/*extraction"*[119].

www.italian.tech/2021/12/27/news/il_futuro_visto_da_annalisa_barla_la_prof_che_insegna_machine_learning-331506239/

[117] Si rimanda al paragrafo 1.4 *"Algoritmi e linguistica: proprietà condivise"*.

[118] E. Galtieri, A. Melegari, *Il lato oscuro degli algoritmi* per *Panorama*, 2021. Cfr. https://www.panorama.it/Tecnologia/cyber-security/algoritmo-intelligenza-artificiale-computer

[119] Si rimanda al paragrafo 3.6 *"Linguaggio 'E' – Estrazione".*

Solitamente, gli stessi *software* di azione linguistica sono in grado di approcciarsi anche in maniera orientativa all'informazione ricevuta e da processare: è questo il tipo di lettura che prende il nome di *skimming*, consistente nell'individuazione di macro-argomenti all'interno dei testi, nella conseguente categorizzazione degli stessi e nell'assegnazione (manuale, automatica o ibrida) di un punteggio *(score)* ad ognuno di questi – qualora, come nella maggior parte dei casi, si individuassero più *topics* all'interno dello stesso testo – al fine di conferire maggiore rilevanza solo ad alcuni di questi.

In questo caso, si parla di "categorizzazione" o – come definito in precedenza – "classificazione" delle informazioni e – nel *software* da me individuato come *tool* per la realizzazione dell'analisi linguistica del caso di studio presentato nel quarto capitolo – viene resa possibile attraverso condizioni che esplicitano il linguaggio "C", appunto, di *categorization*, di cui ne illustrerò applicazioni e funzionalità nel paragrafo 3.5.

# 2. COGNITIVISMO *ed* ELABORAZIONE *delle* INFORMAZIONI

## *2.1 Intelligere – processi e stili cognitivi*

*«L'attività di Intelligence è intimamente legata alle modalità con cui lavora il cervello umano: essa, infatti, opera da una parte grazie a processi logici rigorosamente codificati, e dall'altra si affida anche ad intuizioni e processi estremamente creativi».*[120]

Le scienze cognitive sono alla base delle scoperte e delle innovazioni in tema di intelligenza artificiale; tra esse rientrano, citandone alcune: la filosofia della mente, la neuroscienza cognitiva, la psicologia cognitiva e la linguistica cognitiva.

Il cognitivismo è un indirizzo di studio afferente alla scienza della psicologia che individua analogie tra la mente umana e gli elaboratori di informazioni, entrambi dotati di una modalità di organizzazione delle stesse di tipo sequenziale.

A partire dagli anni '50, il filone della psicologia cognitiva – che rappresenterà una vera e propria rivoluzione nell'ambito delle scienze psicologiche – incentrerà i suoi studi e la sua attenzione nei processi cognitivi messi a punto dall'essere umano: *focus* dell'ambiziosa ricerca sarà quello di comprendere pensieri, memorie ed immagini coinvolti nel processo di elaborazione dell'informazione, nonché di verificare il loro conseguente impatto sul comportamento umano.

Per "processo cognitivo" s'intende la sequenziazione di tutti quegli eventi ritenuti necessari alla realizzazione di una qualsiasi forma di conoscenza attraverso l'attività della mente. Dapprima, l'oggetto di studio si configurava solo nella mente umana, oggi anche in quella digitale ed artificiale.

Steven A. Pinker, professore dell'Università di Harvard, sostenitore della psicologia evoluzionista e della teoria computazionale della mente, postulò cinque assunti che

[120] Gruppo di lavoro 71ª sessione di Studio dell'Istituto Alti Studi per la Difesa, *L'impatto dell'Intelligenza Artificiale sul ciclo di Intelligence e sugli strumenti a disposizione per i pianificatori militari e le forze dell'ordine,* Centro Militare di Studi Strategici, Ministero della Difesa, 2020. Cfr. www.difesa.it/SMD_/CASD/IM/CeMiSS/DocumentiVis/Rcerche_da_pubblicare/Pubblicate_nel_2020/AP_CC_01.pdf

portarono ad una seconda "rivoluzione della cognizione"[121]. Tali assunti possono essere espressi come segue[122]:

1. il mondo della mente è ancorato al mondo fisico per mezzo dei concetti di informazione, calcolo e stimolo-risposta[123] *(input* e *feedback)*;
2. la mente non nasce – e non può esser, quindi, intesa, come *tabula rasa*[124], poiché essa non ha cognizione e non agisce; sono necessari, quindi, necessari dei meccanismi e dei processi innati nella mente dell'essere umano;
3. una gamma infinita di comportamenti può essere generata dalla combinazione di programmi nella mente. Come definito da J.R. Searle nel 1980: *«I programmi non sono semplici strumenti che ci permettono di verificare le spiegazioni psicologiche. I programmi sono essi stessi quelle spiegazioni»*[125]*;*
4. i meccanismi universali della mente sono alla base della varietà antropologico-culturale;
5. la mente è un sistema complesso composto da diversi moduli che interagiscono tra loro.

Il cognitivismo vede l'essere umano come ideatore e realizzatore attivo della propria realtà (unica e contraddistinta da tutte le altre realtà dei suoi simili), in grado di costruire, interrogare e raffinare il proprio scibile in maniera reiterativa, modellando incessantemente le proprie rappresentazioni del mondo, *intelligendo* (riuscendo a legare tra loro) informazioni e dati sensoriali raccolti. Questo perché il canale percettivo-sensoriale influenza in maniera notevole il modo in cui raccogliamo, elaboriamo e comprendiamo informazioni le nuove (o rivalutiamo quelle già note).

Il filosofo Maurice Merleau-Ponty, esponente della fenomenologia francese del '900, sostenne che la percezione e l'interazione con il mondo – e non solo la coscienza, come fino ad allora asserito – rappresentassero delle componenti essenziali per la comprensione della realtà circostante. In questa prospettiva, quindi, il corpo ed i descrittori sensoriali non potevano in alcun modo essere scissi dai processi di cognizione umana.

[121] In riferimento a: H. Gardner, *Nuova scienza della mente. Storia della rivoluzione cognitiva*, 1985.
[122] Per un approfondimento: S. A. PINKER, *How the mind works*, 1997, W. W. Norton & Company, trad. it. M. Parizzi, *Come funziona la mente*, Mondadori, 2002.
[123] Sulla base della teoria comportamentista e del *riflesso condizionato*, di Ivan Pavlov, 1927.
[124] In riferimento a: S. A. Pinker, *Tabula rasa. Perché non è vero che gli uomini nascono tutti uguali*, Mondadori, 2006.
[125] *John R. Searle, Menti, cervelli e programmi,* in D. Dennett, D. Hofstadter, *L'io della mente,* Adelphi, Milano, 1985, pag. 341.

Tutti i tipi di (ri)cognizione possono essere visti come un tipo di elaborazione dell'informazione, per cui è stata mantenuta una stretta relazione con la computazione e l'intelligenza artificiale, sia per la costruzione di sistemi cognitivi artificiali che per la comprensione di quelli naturali.

*«L'Intelligenza Artificiale, che ha come obiettivo quello di replicare su un elaboratore elettronico funzioni di pertinenza esclusiva dell'intelligenza umana, impersonando il tentativo perpetuo di replicare processi mentali di tipo cognitivo, trova in questo ambito un ambiente di applicazione naturalmente affine».*[126]

I processi e modelli di tipo cognitivo della mente umana sono fortemente strutturati e radicati nella profondità della psiche. Replicare tali dinamiche in una mente digitale, che – in assenza di *input* esterni – non ambisce ad interrogarsi, modificarsi ed aggiornarsi, condurrebbe a una realizzazione incompleta del processo: quindi, in assenza delle componenti di apprendimento e di (conseguente) crescita, fondamentali per una rigenerazione continua dei saperi, si arresterebbe il ciclo iterativo ed integrativo che è alla base della conoscenza.

Infatti, l'apprendimento può essere definito come il processo di revisione, modifica e ristrutturazione dei dati raccolti nella memoria a lungo termine. L'acquisizione di nuova conoscenza procede mediante molteplici percorsi culturali, sociali ed intellettuali fra loro interconnessi, sotto forma di schemi rappresentativi e significativi per l'individuo.

Altri settori che conducono l'essere umano, data la sua natura neotenica, ad un apprendimento continuo, integrativo e di tipo *long-life* (ovvero, in grado di essere assimilato lungo tutto l'arco della propria vita) sono: l'attivazione dell'attenzione, la formulazione del pensiero (che si riconduce, in ultima analisi, a puro calcolo; questa concezione dette vita alla macchina di Turing come *«modello di calcolo per dare risposta al problema di decisione»*[127]*)* e la creatività – intesa come espressione artistica.

---

[126] Gruppo di lavoro 71ª sessione di Studio dell'Istituto Alti Studi per la Difesa, *L'impatto dell'Intelligenza Artificiale sul ciclo di Intelligence e sugli strumenti a disposizione per i pianificatori militari e le forze dell'ordine,* Centro Militare di Studi Strategici, Ministero della Difesa, 2020. Cfr. www.difesa.it/SMD_/CASD/IM/CeMiSS/DocumentiVis/Rcerche_da_pubblicare/Pubblicate_nel_2020/AP_CC_01.pdf

[127] P. Forte, *Pensiero Computazionale e la Macchina di Turing per il problem solving*, 2020, p .8. Cfr. www.sbai.uniroma1.it/terza_missione/fascino_matematica_e_sue_applicazioni/slides-forte-macchina_turing.pdf

Il cervello umano, con le sue reti neurali, è analogo ad un processore di dati dotato di abilità di lettura e di scrittura; è in grado, quindi, di decodificare (dapprima strumentalmente) segnali in *input*, per attivare una comprensione degli stessi in una dimensione più profonda, a livello funzionale; come risultato finale, genererà una risposta in *output*. Ogni dato indirizzato alla macchina da parte dell'operatore (che darà l'avvio a un processo di *ingestion* delle informazioni), dovrà essere elaborato mediante regole e condizioni a carattere dicotomicamente binario, al fine di potersi tramutare in *output*.

La risoluzione del problema di comprensione dei dati e della, conseguente, generazione di una risposta, viene resa possibile da algoritmi e codici che realizzano un linguaggio formale che, come quello naturale, mette in comunicazione due o più attori nel teatro dell'azione intelligente.

L'intelligenza artificiale si inserisce tra le discipline delle scienze cognitive, al fianco (citandone solo alcune) della psicologia, della neuroscienza, della linguistica cognitiva, dell'antropologia, dell'etologia, della filosofia della mente, dell'arte, nonché, della matematica e dell'informatica, coinvolte soprattutto nella formazione (e formalizzazione) degli strumenti di modellazione: le reti neurali. Nell'ambito delle scienze cognitive, quindi, confluiscono discipline apparentemente diverse tra di loro. Si riporta di seguito una esemplificazione grafica, sotto forma di esagramma, della complessità di tale interrelazione.

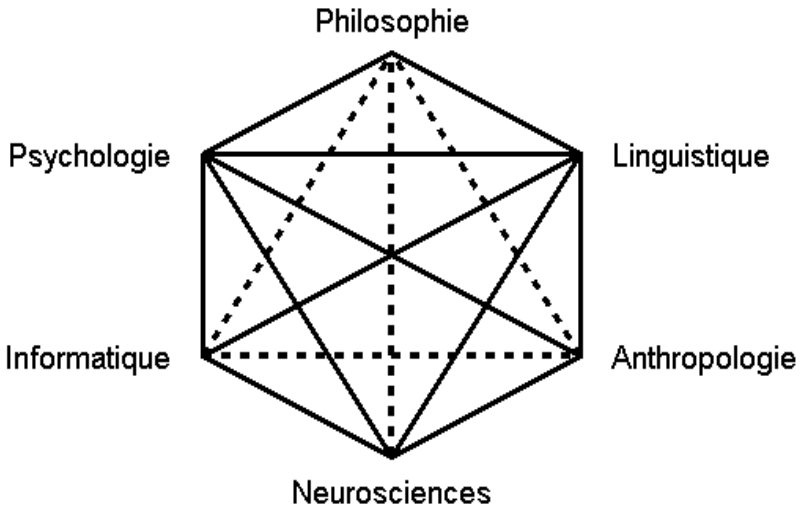



***Figura 5 – Rete delle Scienze cognitive[128]***

[128] Fonte dell'immagine: P. Tabossi, *Intelligenza naturale e intelligenza artificiale,* Il Mulino, 1988.

La base della conoscenza umana (in inglese e in ambito informatico definita come *Knowledge Base)* è, come sostiene oggi Avram Noam Chomsky, formata da regole e, in senso più moderno, da istruzioni informatizzate.

Come definisce Mario Caligiuri nel suo libro "Intelligence e Scienze Umane": «*L'Intelligence potrebbe candidarsi come materia di incontro delle scienze umane, in uno scambio interdisciplinare con altri ambiti,* [perché] *non vi una sostanziale differenza tra i problemi che affronta uno scienziato nel ricostruire il passato astronomico* […] *e i problemi che affronta lo storico nel ricostruire il passato degli uomini»*.

Discorso speculare avviene per il futuro dell'umanità, poiché l'Intelligence si configura come bussola per orientarsi nell'oceano dei *Big Data.*

Il ruolo e le responsabilità assunte dal *Data Scientist* risultano strategicamente nella fusione tra le competenze scientifiche e quelle umanistiche «*perché il sapere è globale e la sua suddivisione specialistica ha comportato nello stesso tempo progressi e perdite. Vanno sviluppati approfondimenti nei settori delle tecnologie, delle scienze cognitive e perfino dei saperi considerati non scientifici. L'efficace e immediato utilizzo dell'informazione che potrà conferire alle élites pubbliche* [come, ad esempio, la Sicurezza Nazionale] *gli strumenti adatti per contrastare minacce criminali e terroristiche».*[129]

George Armitage Miller (1920 – 2012), uno dei fondatori e massimi esponenti storici della psicologia cognitiva, si occupò dello studio della psicologia dei processi mentali applicando tecniche sperimentali da lui ideate, individuando un collegamento tra le scienze cognitive, l'informatica e la linguistica.

Dopo aver suddiviso l'attività cognitiva in tre processi principali (la percezione, la memoria e il pensiero), Miller identificò dei sotto-processi, in seguito definiti come "stili cognitivi". Il termine "stile cognitivo", introdotto dallo psicologo statunitense Gordon Willard Allport nel 1937, si può definire come «*la propensione ad elaborare le informazioni attraverso specifiche modalità*».[130]

[129]M. Caligiuri, *Cyber intelligence, la sfida dei data scientist*, 2016. Cfr. www.sicurezzanazionale.gov.it/sisr.nsf/wp-content/uploads/2016/06/cyber-intelligence-sfida-data-scientist-C. aligiuri.pdf

[130]E. Morelli, D. Palamà, C. Meneghetti, *Il ruolo degli stili cognitivi e degli aspetti emotivo-motivazionali nella buona riuscita nello studio,* Edizioni Centro Studi Erickson, 2015, p. 479.

Ogni stile cognitivo si fonda su alcune predisposizioni cognitive e comportamentali di base di un individuo, ma può modificarsi nel corso della vita in base alle circostanze sociali e ambientali, al tipo di educazione ricevuta e al bagaglio esperienziale soggetto a costanti evolutive. Viene, inoltre, definito come il modo più *«efficace* [per] *godere di una certa autonomia* [nell'] *affrontare compiti stimolanti.* [Tale] *competenza si riferisce al voler essere efficaci, agire nel proprio ambiente e raggiungere i risultati voluti*»[131]; in altre parole, come la personale propensione nell'interpretare, analizzare ed intendere la realtà mediante quei criteri ritenuti da un individuo come i più economici, efficaci e semioticamente inequivocabili.

Gli stili cognitivi non sono da intendersi come abilità, in senso di proficue capacità per lo svolgimento di un'attività o di un compito; bensì, tale concetto esprime la metodologia adottata da un individuo, in maniera spontanea e naturale, al fine di poter sfruttare al meglio le capacità di cui dispone.

Possiamo definire gli stili cognitivi come:

- dipendente o indipendente dal contesto: le persone che elaborano le informazioni tenendo conto del contesto, riescono più difficilmente ad isolare i dati in esso inclusi; diversamente, chi possiede uno stile cognitivo non dipendente dal contesto, è in grado di estrapolare più facilmente i concetti dal contesto stesso, per poi riutilizzarli;
- globale o analitico: gli individui che possiedono uno stile cognitivo di tipo analitico scompongono l'attività o il compito da svolgere in unità discrete; invece, chi possiede uno stile globale tende ad approcciare l'attività o il contesto intesi come *insieme;*
- verbale o visivo: alcuni riescono ad elaborare meglio le informazioni nuove e ad acquisirle mediante il codice linguistico; altri, prediligono il codice visivo-spaziale;
- impulsivo o riflessivo: il primo stile induce a fornire risposte in maniera immediata; nel secondo, si procede all'elaborazione di una risposta solamente a seguito di un'attenta analisi della situazione;
- sistematico o intuitivo: in uno stile di tipo sistematico, si adotta un approccio alle diverse situazioni procedendo a piccoli passi e considerando tutte le variabili in gioco; procedendo intuitivamente, invece, si lavora preferibilmente su ipotesi da confermare o confutare;

[131] P. Boscolo, *La motivazione ad apprendere tra ricerca psicologica e senso comune,* Edizioni Scuola e Città, 2002, pp. 88-89.

- convergente o divergente: con il primo, si procede all'azione in maniera logica e sulla base di informazioni disponibili; il secondo, invece, procede creativamente, immaginando nuove risposte.

Ogni essere umano non è dotato di una sola abilità di *intelligere* il mondo, bensì di molteplici capacità e forme di intelligenza che, come definito da Howard Gardner nel 1983 nella teoria delle intelligenze multiple[132] (come menzionato nel paragrafo 1.3 *"Intelligenza artificiale e ricognizione linguistica")*, hanno grande influenza sugli stili cognitivi.

Altresì, nessun individuo possiede un solo stile cognitivo mediante il quale realizzare attività di presa di decisione o di elaborazione di nuove conoscenze; bensì – a seconda del contesto e delle esigenze in una data situazione – la mente umana, a carattere adattivo, può attingere, inconsciamente, a molteplici stili di pensiero.

[132] Teoria pubblicata in H. Gardner, *Frames of Mind*, 1983.

### 2.1.1 *(Cyber)Bellum omnium contra omnes*[133] *– operazioni psico-cognitive*

*«Le guerre cognitive domineranno lo scenario geopolitico del futuro, incidendo sulla competitività dei sistemi-paese e sulla stessa loro coesione sociale e politica».*[134]

Gli stili cognitivi influenzano in maniera notevole, come già definito, l'apprendimento, lo svolgimento di operazioni ed attività, nonché la presa strategica di decisioni anche in ambito militare. Infatti, in materia di cognizione e ricognizione *«alcune nazioni della NATO hanno già riconosciuto che le tecniche e le tecnologie neuroscientifiche hanno un alto potenziale per l'uso operativo in una estremamente ampia varietà di imprese di sicurezza, difesa e Intelligence».*[135]

Nell'ambito della presa di decisioni in ambito strategico-militare, non si può prescindere dal menzionare il ciclo decisionale definito come *"*OODA *Loop"*[136], acronimo che sta a definire le fasi che compongono i passi per la comprensione, l'apprendimento e l'attuazione dell'azione migliore da compiere in un determinato contesto: *Observe* (osservare), *Orient* (orientare/orientarsi), *Decide* (decidere), *Act* (agire).

Molti *software* di *Decision Intelligence* si basano ed utilizzano tale *mindset* strategico per processare nella maniera più intelligente i dati (strutturati e/o non strutturati, multimediali o multiformato) raccolti da molteplici fonti.

Alla domanda «*Can A.I. help military organizations in their Cyber Security decision-making process?»*[137], avanzata dal Dott. Emanuele Galtieri, CEO di Cy4Gate S.p.A., nel suo omonimo articolo pubblicato sul *magazine* online "Everywhere Rapidly della NATO Rapid Deployable Corps", lo stesso Dott. Galtieri definisce la scelta del metodo di implementazione di algoritmi A.I. al "ciclo OODA" non univoca nella sua applicazione a

---

[133] Il titolo del presente paragrafo è un omaggio all'opera del Prof. R. Trinchero *"(Cyber)Bellum omnium contra omnes. Strategie educative di prevenzione alla guerra cognitiva in Rete"*, Edizioni Erickson, 2018.
[134] G. Gagliano, Gen. C. Jean (a prefazione di), *Guerra psicologica. Saggio sulle moderne tecniche militari cognitive e di disinformazione*, Fuoco Edizioni, 2012, p. 9.
[135] F. Valli, *Cos'è la Cognitive Warfare?*, 2021. Cfr. https://www.mittdolcino.com/2021/10/23/cose-la-cognitive-warfare-la-nato-ed-il-progetto-di-guerra-tramite-controllo-del-cervello-umano-condotta-dai-militari-occidentali-seconda-parte/
[136] Nel testo, p. 20.
[137] The Magazine of the NATO Rapid Deployable Corps – Italy, *Deep Watching on Cyber Threat, Everywhere Rapidly*, 2021. Cfr. www.nrdc-ita.nato.int/db_object/www_nrdc-ita_nato_int/usr/file/ER-Magazine-Everywhere-Rapidly-July-2021-NATO-NRDCITA.pdf

scenari differenti, bensì dipendente da quanto rapidamente e facilmente si dovrà e potrà risolvere una data situazione.

Nello stesso articolo, inoltre, si sostiene che, in un contesto militare, l'uso dell'intelligenza artificiale crei ancora profonde incertezze e preoccupazioni, soprattutto nei riguardi dell'impatto che potrebbe generare un completo conferimento e trasferimento dei processi decisionali a intelligenze non umane; di conseguenza, molte organizzazioni evidenziano l'importanza del ruolo dell'A.I. in veste di supporto agli operatori deputati alla presa di decisioni, ma non come strumento a totale sostituzione di questi ultimi.

La NATO è attualmente attiva nello sviluppo di tattiche di guerra che mirano al condizionamento del comportamento umano per ottenere una determinata risposta/reazione come risultato finale. È questo il concetto cardine della *Cognitive Warfare,* che combina le capacità di combattimento non cinetico della *cyber ingegneria psicologica sociale* e di informazione per vincere sull'avversario senza scontro fisico. Si tratta di una guerra di informazione che ha come obiettivo quello di influenzare i pensieri e/o le azioni dei cittadini, al fine di destabilizzare una Nazione [138].

Comprendendo il potenziale di tale attività di manipolazione, obiettivo della NATO, ad oggi, è quello di far convergere tutte le discipline erogate nelle accademie militari nel contesto della guerra elettronica cognitiva. All'interno dell'esercito, tra le altre aree, le competenze in antropologia sono più richieste che mai: «[Tra le] *direzioni possibili di impegno dell'antropologia nei confronti del mondo militare* [vi è] *quella che potrebbe definirsi l'antropologia "nel" mondo militare; in questo caso il ricercatore è spinto a, e in sostanza deve, negoziare con i suoi interlocutori i modi, i tempi, le caratteristiche, i fini e i limiti della sua azione»*[139]. Lo stesso rapporto del 2020 sulla guerra elettronica cognitiva indica che su quest'ultima si stia sempre più sovrapponendo l'*Artificial Intelligence: «La rapida evoluzione delle neuroscienze come strumento di guerra* [è] *potenziata dalla marcia inarrestabile di una troika trionfante fatta di Intelligenza Artificiale e di Big Data».*

[138] *«La guerra cognitiva è la forma più avanzata di manipolazione mai vista finora»,* cit. della Tenente Colonnello Marie-Pierre Raymond, esercito canadese, oggi scienziata della Difesa e manager del portafoglio di innovazione.
[139] A. Colajanni, *Una difficile sfida per l'antropologia applicativa,* Università degli Studi di Roma "La Sapienza", 2015. Cfr. https://riviste-clueb.online/index.php/anpub/article/view/16/24

Alcune personalità del mondo dell'Intelligence ritengono che l'intelligenza artificiale potrebbe comportare *«un salto di specie, proprio come quello che segnò il passaggio dall'uomo di Neanderthal all'Homo sapiens»*[140], e da quest'ultimo, si sta già arrivando ad una nuova specie umana, caratterizzata dall'ibridazione tra l'individuo e le macchine.

La guerra cognitiva mira a costruire e a far radicare determinate idee e rappresentazioni mentali nei riguardi dell'opinione pubblica per orientare le emozioni, gli atteggiamenti, i ragionamenti, le scelte, le reazioni e i comportamenti degli individui.

*«La cyber war è una nuova modalità di condurre un conflitto, a volte complementare ad esso, con le stesse finalità di una guerra tradizionale* […], *seppur con il ricorso ad alcune peculiarità tattiche e strategiche proprie dello specifico dominio».*[141]

Questo tipo di conflitto non si realizza attraverso strumenti ed armi tangibili, bensì attraverso la capacità di utilizzo più intelligente della conoscenza.

Non prevede, quindi, alcuna violenza fisica, ma mira alla manipolazione sistematica dell'informazione per trarre un vantaggio sull'avversario[142]; questo perché *«il conflitto cibernetico fa leva a una escalation di tecniche e tattiche che* [incidono] *sulla sfera emotiva di una comunità»*[143], al fine di indebolirla e agevolare le condizioni per la resa.

Per questo, la guerra cognitiva si serve dell'apporto di numerose discipline che studiano i processi conoscitivi e comunicativi, quali, ad esempio: la pedagogia, la psicologia, la sociologia, l'antropologia, la semiotica e le discipline STEM.

È in questo contesto che si parla, in ambito militare, di operazioni psicologiche – definite *Psychological Operations,* o *PSYOP,* ovvero, *«l'insieme di prodotti e/o azioni che condizionano o rafforzano attitudini, opinioni ed emozioni di specifici target quali governi di Paesi stranieri, organizzazioni, gruppi o singoli individui al fine di indurli a comportarsi in modo tale da supportare gli obiettivi di politica nazionale»*[144].

---

[140]M. Caligiuri, *La mente come campo di battaglia,* per *Formiche,* 2022. Cfr. https://formiche.net/2022/03/campo-battaglia-definitivo-mente-persone/

[141]E. Galtieri, *Da Anonymous ai malware, l'arte della guerra cyber,* per *Formiche*, 2022. Cfr. https://formiche.net/2022/02/arte-della-guerra-era-cyber-galtieri-cy4gate/

[142] Tale concetto è assimilabile semanticamente a quello di guerra dell'informazione *(Information Warfare).*

[143] E. Galtieri, *Da Anonymous ai malware, l'arte della guerra cyber,* per *Formiche*, 2022. Cfr. https://formiche.net/2022/02/arte-della-guerra-era-cyber-galtieri-cy4gate/

[144] Ten. Col. FONTANA L., SMD II Rep. – Uff. Materiali di Armamento e Alta Precisione, *Le Operazioni Psicologiche Militari (PSYOP), La 'conquista' delle menti,* Ministero della Difesa, 2016. Cfr. https://www.difesa.it/InformazioniDellaDifesa/periodico/IlPeriodico_AnniPrecedenti/Documents/Le_Operazioni_Psicologiche_militar_620menti.pdf

Le PSYOP possono essere definite come «*il complesso delle attività psicologiche pianificate in tempo di pace, crisi o guerra, dirette verso gruppi obiettivo nemici, amici o neutrali, al fine di influenzarne gli atteggiamenti ed i comportamenti che incidono sul conseguimento di obiettivi prefissati di natura politica e militare*» [145].

Uno scandalo connesso alla manipolazione del pensiero attraverso la gestione illecita (e conseguente sottrazione) di dati personali, fu quello che ha interessato il colosso dei *social networks Facebook* e la società di *data analytics* denominata *Cambridge Analytica*. *Cambridge Analytica* venne accusata di aver manipolato sia la campagna elettorale di Donald Trump del 2016, che il referendum inglese *pro-Brexit*, influenzando le scelte elettorali dei cittadini attraverso strategie e tecniche di profilazione degli utenti e di manipolazione del consenso dei dati personali. L'obiettivo era tentare di prevedere le scelte politiche dei cittadini e re-indirizzarle. Vennero acquisiti senza previo consenso dati personali di circa 50 milioni di utenti *Facebook*[146], che vennero elaborate da modelli e complessi algoritmi per individuare (e, attraverso l'attività strategica di dissuasione o persuasione, rigenerare) i profili degli utenti, con un approccio simile a quello adottato nella psicometria – al fine di misurare le abilità, il pensiero, i comportamenti e le caratteristiche della personalità di ogni persona.

La manipolazione delle personalità avviene all'interno di un ambiente comunicativo digitale, al fine di esplorare vulnerabilità cognitive e comportamentali delle persone[147]. Ciò avviene, ad esempio, mediante la pubblicazione di un *post* puntando al suo *re-post* da parte di una ragnatela di contatti, configurandosi come nuova espressione del concetto di socializzazione. Il messaggio condiviso inizialmente da un utente può generarne altri in grado di modificare o contraddire il significato di quello prodotto in origine, influenzando, così, gli aspetti emozionali e comportamentali dei lettori.[148]

«*Con l'avvento di Internet e dei social network* […], *tutti possono diventare pubblicatori* [e] *dichiarare guerre cognitive a tutti: una vera e propria "(Cyber)Bellum*

[145] F. Angius, *PSYOPS – operazioni psicologiche 2. Struttura e modalità di decisione e pianificazione*, Archivio Disarmo, Istituto di Ricerche Internazionali, 2008, p. 2.
[146] Inclusi tutti i dati linguistici prodotti da ogni utente mediante la pubblicazione di *post*, commenti a questi ultimi e a foto, e – in alcuni casi – persino dati testuali estratti dalla sezione di messaggistica privata.
[147] Per un approfondimento: E. Santagata, A. Melegari, *Social media: il preoccupante rovescio della medaglia*, per *Analisi Difesa*, 2020. Cfr. www.analisidifesa.it/2020/10/social-media-il-preoccupante-rovescio-della-medaglia/
[148] In riferimento a: A. Teti, *Virtual Humint – La nuova frontiera dell'Intelligence,* Rubbettino Editore, 2019.

*omnium contra omnes".*[149] *La diffamazione sui social networks, lo stalking e il cyberbullismo possono assumere la forma di tante, piccole guerre cognitive».*[150]

La guerra cognitiva, quindi, si serve di numerose strategie, utilizzate in modo strategico e coordinato, come ad esempio:

1. la pubblicità: il diffondere messaggi il cui scopo esplicito non è quello di informare le masse, bensì di influenzarle;
2. la *deception*, ossia l'inganno: occultare i fatti realmente accaduti attraverso depistaggi sistematici;
3. la disinformazione[151], ovvero la diffusione di notizie infondate con l'obiettivo di danneggiare l'immagine pubblica di un avversario;
4. l'intossicazione, ossia il fornire all'avversario informazioni errate per influenzare negativamente le sue decisioni;
5. la propaganda, ovvero l'attività di disseminazione di idee e informazioni con lo scopo di indurre a specifici atteggiamenti e azioni.

Mario Caligiuri, nel suo articolo "La mente come campo di battaglia", ritiene «*opportuno cominciare a delineare una "geopolitica della mente", intesa come il campo di battaglia dove si sta svolgendo la lotta per il potere, in modo da esercitare il dominio sulle persone e sulle nazioni, poiché oltre il controllo della mente non può esserci altro*»[152]. L'obiettivo della guerra cognitiva è rendere ogni essere umano una potenziale arma: la qualità e la quantità delle informazioni in nostro possesso determinano la nostra capacità di previsione ed azione.

[149] Traduzione: "(Cyber)Guerra di tutti contro tutti". Omaggio all'omonima opera del Prof. R. Trinchero, 2018.
[150] R. Trinchero, *Against the cognitive war. Promoting active skepticism*, Riviste Erickson, 2018. Cfr. https://www.unipa.it/persone/docenti/c/gianna.cappello/.content/documenti/TRINCHERO.pdf
[151] *"La capacità di un algoritmo di creare testi credibili è infatti già ampiamente utilizzata per sviluppare chatbot, riassumere testi, scrivere sceneggiature di film e, purtroppo, anche per robotizzare la produzione di fake news veritiere"* – E. Santagata, A. Melegari, *Ecco quanto (poco) costa un robot capace di disinformare (tanto),* per *Analisi Difesa,* 2019. Cfr. www.analisidifesa.it/2019/06/ecco-quanto-poco-costa-un-robot-capace-di-disinformare-tanto/
[152] M. Caligiuri, *La mente come campo di battaglia,* per *Formiche.net,* 2022. Cfr. https://formiche.net/2022/03/campo-battaglia-definitivo-mente-persone/

## *2.2 Fondamenti teorici della linguistica cognitiva*

La linguistica cognitiva, insieme degli studi sviluppati tra gli anni '70 e gli anni '80, è una branca della linguistica tradizionale che esamina la relazione tra il linguaggio e la mente umana, analizzando le produzioni linguistiche realizzate all'interno di un contesto specifico, al fine di ricostruirne i processi cognitivi sottesi.

Inoltre, essa *«estende la sua analisi ai meccanismi cognitivi che stanno alla base della struttura della lingua e del comportamento linguistico, evidenziando i processi di acquisizione, elaborazione, produzione e comprensione della conoscenza, tramite pensiero, esperienza e sensi»*.[153]

L'incontro tra cognitivismo e linguistica si realizzò (ma solo inizialmente) in contrapposizione alla grammatica generativo-trasformazionale teorizzata dal linguista Avram Noam Chomsky, dal momento che la nuova corrente linguistica evidenziava la rilevanza del ruolo della semantica nel comportamento linguistico – intesa come ponte fra facoltà cognitiva umana e capacità linguistica, costruito in funzione del significato e dell'uso che se ne fa[154]– a differenza della teoria di Chomsky, che assegnava questo primato alla sintassi.

Inoltre, Chomsky definiva il linguaggio come capacità innata nell'essere umano, inscritta in un modulo deputato della mente di ogni individuo – a differenza del pensiero linguistico-cognitivo che sostiene, invece, che le varie componenti del linguaggio vengano create e strutturate in base alle molteplici situazioni con cui l'individuo si interfaccia.

La linguistica cognitiva richiama il pensiero di Giambattista Vico – filosofo italiano dell'età dei Lumi – secondo il quale il sapere (*scire)* non risiederebbe esclusivamente nella pura *cogitatio*, bensì nella capacità umana di produrre simboli, nel ricomporre gli elementi della conoscenza (definita da Vico come *divina intelligentia*[155]*)* attraverso questi ultimi e nella trasformazione dei segni prodotti in linguaggio.

---

[153]C. Bazzanella, *Linguistica cognitiva. Un'introduzione*, Roma-Bari, Edizioni Laterza, 2014, p. 182.
[154] «*Il significato di una parola risiede interamente nel suo uso, e viene dato in una spiegazione*». – S. Oliva, *Dal non-senso al gesto: Wittgenstein e il giudizio di valore,* 2017. Cfr.
https://journals.openedition.org/estetica/2165
Fonte: https://journals.openedition.org/estetica/2165
[155] De Luise, Farinetti, *Lezioni di storia della filosofia,* Zanichelli Editore, 2010, p. 41. Cfr.
https://studylibit.com/doc/4505782/lettura-1---giambattista-vico--vero-e-fatto-si-convertono

Nella linguistica cognitiva è, quindi, il significato a permettere la generazione di una categorizzazione della realtà – attraverso processi concettuali – nonché l'assunzione di un approccio induttivo e tassonomico. Questo approccio induttivo e profondo decreta il linguaggio come il deposito della conoscenza del mondo.

Il pensiero di Giambattista Vico si incentra sull'importanza di realizzare e far propri degli schemi concettuali individuali (ma al contempo, condivisi con la società, ergo predicibili da parte dell'individuo) rappresentanti gli strumenti ed i mezzi cognitivi attraverso i quali si rende possibile l'incontro tra individuo, mondo e linguaggio.

Si introduce, così, il concetto di "dominio linguistico" *(frame):* ovvero, scatole *(buckets)* contenenti concetti, realizzate sulla base dell'esperienza e delle conoscenze enciclopediche dell'utilizzatore della lingua. Il concetto di dominio implica un grado di interdipendenza cognitiva e di gerarchizzazione tassonomica dei segni, rivalutabili nel corso della vita di ognuno. L'interfaccia tra sintassi e semantica viene, quindi, profondamente esplorata ed esaminata per indurre e condurre a un'indagine delle capacità cognitive umane, e di come esse siano in grado di utilizzare il linguaggio come strumento per l'organizzazione e l'elaborazione di pensieri, opinioni, memorie e conoscenza.

Il linguista statunitense Charles J. Fillmore (1929 – 2014) ha incentrato i suoi studi sulla semantica della comprensione: ogni parola susciterebbe una determinata lettura ed interpretazione mediante il richiamo a uno o più *frame(s)* in cui "inscatolare" concettualmente quanto recepito. Ogni parlante associa, quindi, i propri significati a domini linguistici specifici. Quando questo *format* di associazioni semantico-concettuali viene condiviso tra due o più interagenti[156], si pongono le basi per intentare un atto comunicativo funzionale. L'efficacia di tale atto comunicativo dipenderà unicamente dalle dinamiche interpretative (altamente soggettive e difficilmente prevedibili), dai modelli cognitivi in uso degli *interattanti* e dalla volontà di riconsiderare i significati previamente acquisiti e riconosciuti. La sola somma degli *input* recepiti/inviati dai parlanti non garantisce il successo della comunicazione.

[156] Degno di menzione è l'effetto psicologico del *priming*, per cui l'esposizione ad uno stimolo influenza la risposta a stimoli successivi. Ciò può esercitarsi a livello percettivo, semantico o concettuale. *«Alla base del priming c'è il fenomeno per il quale riconoscere una parola appare più semplice se questa viene seguita da un'altra parola correlata ad essa»* – I. Zambri, *Modelli di memoria semantica e lessicale,* Università di Bologna, 2015, p. 20. (Per un approfondimento: A. Laudanna, C. Burani, *Il lessico: processi e rappresentazioni,* Firenze, La Nuova Italia Scientifica, 1993)

Il linguaggio naturale pertinente all'essere umano, quindi, si sviluppa a seconda delle esperienze linguistiche condivise. Infatti, anche i linguisti e docenti dell'*Universität* Koblenz-Landauof (Mainz, Germania) René Dirven e Frank Polzenhagen hanno sottolineato che *«i modelli culturali sono schemi cognitivi condivisi dai gruppi sociali».*

Questo assunto si collega alla teoria dello schema culturale[157], postulato nel 1999 da Kiwamu Nishida, professore dell'Università di Tokyo. Gli schemi culturali sono strutture cognitive che contengono conoscenza, a seguito di interazioni sociali nell'ambiente culturale di un individuo. Ne risulta che, ogni volta che un individuo interagisce con dei membri della sua stessa cultura in situazioni e contesti diversi, vengano creati e fissati nella memoria degli schemi culturali, che vengono reiterati nel corso della sua vita in società.

Nella linguistica cognitiva è ipotizzato il diretto collegamento tra mente e interpretazione del mondo tramite la propria fisicità, contrapponendosi, in questo caso, alla semiotica linguistica, la quale affermava la centralità del segno nell'azione di decodifica della realtà.

È in questa formulazione che risiede il concetto formulato da George Lakoff di "cognizione incorporata" *(embodied cognition)* o "filosofia del corpo", secondo la quale: *«I processi di alto livello* [ovvero, il pensiero ed il ragionamento] *siano radicati nelle emozioni e nei cosiddetti processi sensorimotori di basso livello* [ovvero, nel cosiddetto "pensiero primitivo", relegato ad istinti e pulsioni]. *Tutti gli aspetti della cognizione (idee, concetti e categorie) sarebbero plasmati da aspetti del corpo».*[158]

Il concetto di *embodiment*, collocandosi nei processi cognitivi di acquisizione delle informazioni, si estende automaticamente a tutte le capacità e abilità naturali (o, meglio, *naturalizzate*[159]) nell'essere umano; si parla, quindi, anche di *embodied language*[160] e, di conseguenza, di *embodied language processing*, ovvero di *semantica incorporata* – l'esistenza di determinate sezioni cerebrali che permettono l'elaborazione del significato di una data parola solo attraverso il suo richiamo nella sfera sensoriale motoria ad essa

[157] Teoria dello schema culturale. Cfr. https://it.knowledgr.com/06172276/TeoriaDiSchemaCulturale

[158] I. Adornetti, A. Chiera, F. Ferretti, *Embodied Cognition e Origine Del linguaggio: il ruolo cruciale del gesto*, Università degli Studi di "Roma Tre", 2019, p. 44.

[159] In antropologia, così come in tutte le discipline nate grazie all'interconnessione con essa, si fa distinzione tra ciò che è "naturale" – ovvero, tutto ciò che è "natura"*:* elementi tramandati geneticamente e inscritti nel DNA dell'essere umano (capacità innate) – e ciò che è "naturalizzato" – ovvero, qualcosa non inscritto nei geni, bensì appreso dopo la nascita, ma radicato così profondamente nella coscienza e nel corpo di ognuno, tale da poter esser perseguito ed agito con inconscia naturalezza (automatizzazione); inoltre, senza detti elementi/capacità, un individuo non riuscirebbe a vivere contestualmente in una società.

[160] Per un approfondimento: R. Barthes, *La grana della voce, Interviste 1962 - 1980*, Edizioni Einaudi, 1981.

associata. Pertanto, l'*embodiment* si configura come rifiuto del dualismo cartesiano "mente *versus* materia" – quindi, tra pensiero di "livello superiore" e di "livello inferiore".

La costruzione della realtà è, quindi, mediata ed interferita dalla natura dei nostri corpi; e, il corpo, mutando in maniera inafferrabile nel corso della vita, costringe l'individuo ad una costante rivalutazione del proprio sistema linguistico.

Ma il concetto di "corporeità" non è l'unico ad essere elaborato dal movimento cognitivo contemporaneo, tra i cui esponenti si menzionano i filosofi Andy Clark e David Chalmers; l'*embodiment* è infatti parte di un gruppo di teorie denotato come *4E Cognition.* Le altre ipotesi considerano la mente come *embedded* (situata), *extended* (estesa) ed *enacted* (liberamente intesa come "interattiva"), definite come di seguito[161]:

- ipotesi della cognizione situata*:* è l'idea che i sistemi intelligenti richiedano un costante adattamento e puntuale allineamento con l'ambiente circostante *(input);*
- ipotesi della cognizione estesa*:* attraverso l'utilizzo di quaderni, lavagne e/o dispositivi elettronici si appuntano (e trasmettono) informazioni dapprima contenute solamente nel nostro cervello. Questa teoria sostiene che i mezzi utilizzati per la presa di cognizione, rappresentino un'estensione della cognizione umana stessa;
- ipotesi della cognizione interattiva*:* la maggior parte delle funzioni cognitive dipende da un legame interattivo tra esseri umani ed ambiente. Ogni essere umano "mette in scena" il mondo (ed il modo) in cui vive; l'interazione tra esso e l'ambiente ha influenza sugli aspetti della percezione e modella il processo di cognizione stesso.

Queste tre ipotesi sono *«accomunate da un "esternismo" di fondo, per cui la mente sarebbe diffusa nell'ambiente esterno, in tutto e per tutto parte della realtà che ci circonda».*[162]

[161] A. Clark, D. Chalmers, *The Extended Mind*, Oxford University Press on behalf of The Analysis Committee, Vol. 58, No. 1, 1998, pp. 7-19.
[162]D. Versari, *Pensare con lo spazio che ci circonda: Il ruolo del linguaggio nell'ipotesi della mente estesa,* Magazine Treccani, 2021. Cfr. treccani.it/magazine/chiasmo/storia_e_filosofia/Spazio/IUSS_pensare_con_lo_spazio_Versari_seconda_parte

## *2.3 Reti semantiche – interpretazione e confini del significato*

*«You see, it's like a portmanteau – there are two meanings packed up into one word»*
*– È come un baule[163], capisci, ci sono due significati imballati dentro un'unica parola.[164]*

L'*Homo sapiens* è una creatura in grado di produrre significati. Lo fa attraverso l'esperienza, la contemplazione, l'interpretazione ma, soprattutto, mediante l'immaginazione. Queste attività vengono definite "naturalizzate"[165] nella coscienza dell'essere umano, ovvero "apprese naturalmente"[166]. Gli esseri umani sono gli unici animali in grado di conversare, effettuando ragionamenti e scatenando guerre per cose *«che esistono solo nella loro immaginazione: come divinità, nazioni, leggi e soldi».[167]*

L'importanza della produzione di senso si riflette nella necessità, da parte dell'individuo, di creare dei macro-campi concettuali per l'ampliamento, il delineamento e la discriminazione dei confini di significato di ogni entità – concreta o astratta – individuata, compresa e, in seguito, radicata nella propria mente. Tale processo di potenziamento e di continuo raffinamento della conoscenza agisce e consegue anche sulle intelligenze e sugli stili cognitivi propri di ogni individuo, che garantiscono un apprendimento plastico e duraturo nel corso della propria esistenza *(long-life learning),* nonché un'incessante rivalutazione di quanto appreso fino a quel determinato momento.

Nell'avviare, in maniera del tutto spontanea e naturale, tale processo di ricerca, sviluppo e discrimine di significati, si avverte come necessaria la creazione – del tutto personale e performativa[168], seppur necessariamente condivisa (almeno) dai membri della

[163] Un *portmanteau*, in linguistica, viene anche definito come "parola macedonia", "neologismo sincratico" o "mashup". Ovvero, si tratta di un neologismo nato dall'unione di due parole, talvolta fondendo una o più lettere che le compongono. Può essere considerato una estensione dell'acronimo. Alcuni *portmanteaux* di conio anglosassone, ma radicati nell'uso linguistico a livello globale, sono i termini inglesi *smog* e *motel*.
[164] L. Carroll, *Through the Looking-Glass and What Alice Found There*, 1871, p. 225, trad. it., *Attraverso lo specchio e quel che Alice vi trovò*, Umberto Notari Editore, 1914, p. 221.
[165] Come già menzionato nel paragrafo 2.2 *"Fondamenti teorici della linguistica cognitiva".*
[166] Si sottolinea, nuovamente, la dicotomia "acquisizione spontanea e naturale" Vs. "apprendimento guidato", rovesciata nel paradosso semantico insito nel concetto di "apprendimento naturale" nei riguardi delle manifestazioni delle azioni dell'essere umano relegate, o persino inscindibili, dal proprio assetto culturale.
[167] Y. N. Harari, *Da Animali a Dèi. Breve storia dell'umanità,* 2011, trad. it, G. Bernardi, Bompiani, 2017, p. 1
[168] In riferimento al concetto di *performance* (o prestazione) di Naom Chomsky, ovvero, il modo in cui il sistema linguistico è usato nella comunicazione in maniera del tutto personale; è in opposizione al concetto di

propria comunità culturale-linguistica di appartenenza – di specifiche categorie[169] in cui "inscatolare" ogni nuovo concetto, idea o entità in fase di elaborazione, per riuscire a contestualizzare tali nuovi elementi nel modo più coerente ed idoneo possibile, nonché a renderlo immediatamente fruibile nel momento e nella situazione più opportuni.

Quando proviamo ad interpretare un comportamento o uno scenario, riveste fondamentale importanza l'aver consapevolezza e cognizione dei molteplici processi (logici, cognitivi, semantici e sociali) ed implicazioni che questa operazione comporta.

La funzionalità di ogni enunciato è resa possibile dalle relazioni logico-sintattiche degli argomenti che lo compongono, nonché dalle connessioni che intercorrono tra essi, le quali accrescono la complessità e la profondità della lettura della frase in termini di interpretazione – di tipo "emic".[170]

Le parole, che costituiscono un prodotto sociale, possono essere contestualizzate e decodificate solamente se si trovano all'interno di frasi, di testi e/o di situazioni comunicative concrete – in cui, peraltro, si manifesta la lingua dell'uso.[171] Di contro, qualora si isolassero e decontestualizzassero le parole nel tentativo di ricercare un'interpretazione semantica più lineare e/o una dimensione semantica meno articolata, la decodifica del messaggio risulterebbe incompleta, se non fallimentare.

Inoltre, non si può prescindere dell'esecuzione di una puntuale valutazione sul valore informativo contenuto in ogni enunciato, ovvero, il saper riconoscere quale sia il rapporto tra parola/significante *(token)* – ovvero la rappresentazione grafica/visiva del concetto – e il referente – la rappresentazione del concetto nella realtà esterna – per tentare di esaudire la realizzazione del suo significato.

*competence*, ovvero nella grammatica: il linguaggio "ideale" che rende possibile la produzione e la comprensione standard di un numero infinito di frasi in un determinato codice linguistico.

[169] Come già menzionato nel paragrafo 2.2 *"Fondamenti teorici della linguistica cognitiva".*

[170] Nel testo, p. 21.

[171] Degno di nota è il *Grande dizionario italiano dell'uso – GRADIT*, di Tullio de Mauro.

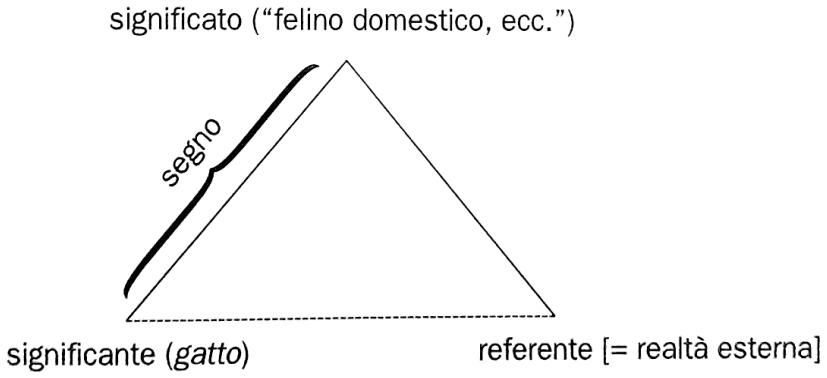



***Figura 6 – Il triangolo semiotico***[172]

Considerando la molteplicità degli studi linguistici, il campo della semantica è il settore scientificamente più complesso e meno strutturato fra tutti, e ciò è dovuto a una serie di problemi epistemologici che mettono a dura prova i più efficaci metodi esplorativi della conoscenza umana. Al contrario degli studi attinenti alla morfologia, alla fonologia e alla sintassi, ambiti nei quali si può procedere con metodologie e strumenti validati, reiterati e consolidati negli anni in termini di rappresentazione, spiegazione ed insegnamento, *«il campo della semantica offre delle trattazioni sistematiche basate su* […] *postulati o restrizioni, non considerabili "canonici"* […], *la cui maggiore condivisibilità scientifica è fortemente dovuta alle restrizioni di dominio».*[173]

A seconda della natura e del grado di analisi del significato su cui bisogna agire in fase di investigazione linguistica, lo studio della semantica presenta al suo interno una suddivisione disciplinare e paradigmatica che è propria della linguistica sincronica[174], ovvero, risultante in una triplice possibilità di valutazione del dato testuale: a livello lessicale, frasale e/o del discorso. A ciò, si aggiungono molteplici influenze multidisciplinari che, interconnettendosi con il piano linguistico, integrano la propria epistemologia – fatta di conoscenze e metodi per raggiungere tale sapere – nel tentativo di aggiungersi, o subentrare parzialmente – al ruolo centrale della linguistica. Alcune di queste discipline pertengono, ad esempio, all'ambito della filosofia, della logica, della psicologia, dell'antropologia, delle scienze cognitive e della statistica.

[172] *Triangolo di Ogden-Richards.* Il triangolo semiotico è una rappresentazione geometrica dei componenti essenziali della significazione. Fonte dell'immagine: https://it.quora.com/Cos%C3%A8-il-triangolo-semiotico
[173] P. Petricca, *Semantica, forme, modelli e problemi*, LED, Milano, 2019, p. 7.
[174] Concettualmente opposta alla linguistica diacronica, la linguistica sincronica studia gli elementi costitutivi e i principi fondamentali del sistema di una lingua in un determinato momento storico, linguistico e culturale.

La rappresentazione della conoscenza da parte dell'essere umano avviene mediante reti semantiche: una rete semantica si basa sull'utilizzo su "nodi" ed "archi" *(links)* afferenti a una determinata entità e/o concetto. Nello specifico, i nodi identificano le entità/gli oggetti, mentre gli archi stabiliscono le relazioni semantiche che intercorrono tra due o più nodi concettuali. Le reti semantiche agevolano la comprensione della realtà di un dato dominio.

Quando un elemento/entità è in relazione di appartenenza con una categoria, ne acquisisce le proprietà (concetto di ereditarietà). Ogni nodo/entità può avere molteplici relazioni di appartenenza *(links)* con categorie/domini differenti.

Una categoria può essere definita come un cerchio concentrico (o una cellula), al cui centro (o nucleo) si trovano gli elementi più rappresentativi e radicati nella lingua dell'uso, tra cui l'entità che fornisce il modello – definito anche come stimolo – che permette la ricognizione e la collocazione di tutti gli altri argomenti afferenti a quella data categoria, nella mente del parlante.

In una rete semantica la mera distinzione tra entità e categorie, però, non è sufficiente a rappresentare in maniera univoca e fedele la realtà. Ad esempio, se si indicasse come nodo-oggetto una persona (di nome, ad esempio, *Andrea*) e si creasse una relazione semantica *(link)* che collegasse tale nome al dominio/categoria *(field)* definito come insieme di "persone generiche", ciò potrebbe causare una subordinazione meronimica troppo ampia – nonché un'ambiguità di genere, essendo Andrea, in questo caso, anche nella lingua italiana, un nome sia femminile, che maschile.

Sarebbe, quindi, preferibile creare tanti nodi quante sono le caratteristiche che differenziano e caratterizzano cada entità, nonché, altrettante categorie di appartenenza quanti sono i domini individuabili; qualora la molteplicità di sfumature di dominio venisse inglobata in un'unica categoria semantica, ciò potrebbe condurre ad una interpretazione troppo opaca dei confini di significato sia del concetto stesso, e sia dell'entità da inglobare, o meno, a sé.

I confini delle categorie sono estremamente plastici e variabili, e gli elementi inclusi al loro interno sono suscettibili al passare del tempo; inoltre, l'evolversi delle circostanze politico-sociali generano dinamicità al (dia)sistema linguistico: nuove parole vengono coniate, alcune cadono in disuso, altre ancora vengono adottate ed applicate a nuovi domini di uso, oppure subiscono un cambiamento, più o meno radicale, nella loro polarità semantica.

Vi è una correlazione psico-linguistica tra l'individuazione della posizione dei confini del significato di un'entità (nodo-oggetto) e la velocità di risposta *(buffering)*[175] ad una richiesta di categorizzazione/collocazione a livello concettuale dell'entità stessa. Tanto più il confine risulta saldo nella mente del parlante, quanto più sarà semplice per lui inquadrare tale entità in un dominio. Allo stesso modo, tanto più il significato risulta opaco, quanto più lenta sarà la risposta nell'individuazione della categoria di appartenenza.

Esiste uno stato – ed una funzione – marginale di alcuni elementi che esperiscono una molteplice ereditarietà concettuale: è in questi casi che si parla di "vaghezza dei confini di significato". Alcune persone possono, quindi, delimitare in maniera diversa la stessa entità, oppure possono collocare in altre categorie concettuali lo stesso oggetto – soprattutto nei casi di polisemia semantica o di dubbio nei riguardi del rapporto di iponimia o iperonimia che intercorre tra vari elementi.

Questo perché l'interpretazione della realtà – da cui ne consegue, realizzandola, la propria conoscenza del mondo – è altamente soggettiva, suscettibile agli *input* sensoriali, influenzata dal bagaglio esperienziale, culturale ed enciclopedico (auspicabilmente, in continua evoluzione e rivalutazione) di ogni individuo[176].

Jean-Claude Coquet, linguista e semiotico francese, ha dedicato i suoi studi ai rapporti che legano natura e linguaggio, percezione e comunicazione. Egli afferma che *«vi* [sia] *un continuum tra il linguaggio, il mondo nel quale il soggetto si trova e sul quale agisce, e l'essere».*[177] La realtà empirica in cui viviamo *«è la risultante di un processo di semiosi illimitata»*[178] in cui ogni entità – astratta o concreta che sia – viene etichettata; ogni etichetta definita ed allocata è carica di significato e si collega semanticamente ad altri elementi, fino a costituire un reticolo di significati che attiva e rinvia continui processi interpretativi, anche in assenza di interlocutori.

[175] La memoria ed il richiamo sono aspetti molto importanti della ricerca linguistica cognitiva, nonché nei modelli computazionali con alla base l'intelligenza artificiale.
[176] *«Gli esseri umani sono quegli strani animali intrappolati nelle reti di significato che essi stessi hanno tessuto»*. cit. Max Weber, sociologo e filosofo tedesco, 1864-1920.
[177] P. Fabbri (a cura di) J. C. Coquet, *Le istanze enuncianti. Fenomenologia e semiotica*, Bruno Mondadori, 2008. Cfr. https://www.paolofabbri.it/libri/le-istanze-enuncianti/
[178] P. M. Dominici, *La funzione del linguaggio tra arbitrarietà e ambiguità*, Il Sole 24 Ore, 2014. Cfr. https://pierodominici.nova100.ilsole24ore.com/2014/05/05/per-una-comunicazione-responsabile-la-funzione-del-linguaggio-tra-arbitrarieta-e-ambiguita/

Una caratteristica tra le più significative del segno linguistico è l*'arbitraire du signe,* in contrapposizione al concetto di iconicità: tale convenzionalità si realizza quando alcuni elementi del segno linguistico non vengono motivati, né etichettati secondo una logica semantica. *«La scrittura è dunque basata sulla sistematicità conferita dall'ordine che le è proprio, dall'arbitrarietà nel rapporto fra lettera e suono, dal valore acquisito per negatività e differenza, tra la pienezza del grafo e il vuoto della spaziatura»*[179].

Detti elementi sono, quindi, il frutto di una convenzione tra i parlanti di una lingua. L'arbitrarietà semantica si realizza tanto sul piano dell'espressione (il significante), quanto su quello del contenuto (il significato). *«La parola arbitraire* [...] *non deve dare l'idea che il significante dipenda dalla libera scelta del soggetto parlante* [...]; *noi vogliamo dire che è "immotivato", vale a dire arbitrario in rapporto al significato, col quale non ha nella realtà alcun aggancio naturale»*[180].

Le reti semantiche – ed altri formalismi ad esse ispirati – vengono usate nei *software* di elaborazione del linguaggio naturale (NLP), con l'obiettivo di rappresentare il contenuto semantico dei singoli elementi di una frase, nonché le relazioni tra le parole *(tokens)* e i concetti *(fields,* o domini) corrispondenti. Graficamente, sia i nodi che gli archi sono contrassegnati da un'etichetta, definita *label,* che serve ad attribuire loro sia un significato, che una relazione.

Il complesso delle entità etichettate (in linguaggio tecnico, *labelizzate,* oppure "annotate") e delle relazioni logico-sintattiche-semantiche indicate, si configura in un "grafo dei sensi"[181], che permette una visione d'insieme della natura del sintagma, del periodo o della frase complessa analizzati, e con il quale si può interagire al fine di analizzare la profondità dei collegamenti e/o discriminando soglie di significato.

[179] D. Poli, *Il mito dell'interpretazione in Ferdinand de Saussure,* Università di Macerata, 2018, p. 73.
[180] F. De Saussure, *Cours de linguistique générale,*, 1916, trad. it. T. De Mauro, *Corso di linguistica generale*, Laterza, Roma-Bari, 2005, pp. 86-87.
[181] Nel *tool* di riferimento di *Expert.ai*, utilizzato ai fini della realizzazione del lavoro di analisi linguistica presentato in questa tesi di Laurea Magistrale, il *grafo dei sensi* è denominato *Sensigrafo*®.

Alcune relazioni semantiche rilevanti:

- meronimia: A è parte di B, quindi B ha A come sua parte[182];
- olonimia: B è parte di A, quindi A ha B come sua parte[183]:
- iponimia, o troponimia: A è subordinata a B; A è un tipo di B[184];
- iperonimia: B è sovraordinata ad A[185];
- sinonimia: A denota la stessa cosa di B;
- antonimia: A denota il concetto opposto a B;
- relazioni di causa-effetto;
- relazioni di proprietà: A è caratterizzata dal possedere B come proprietà o abilità[186].

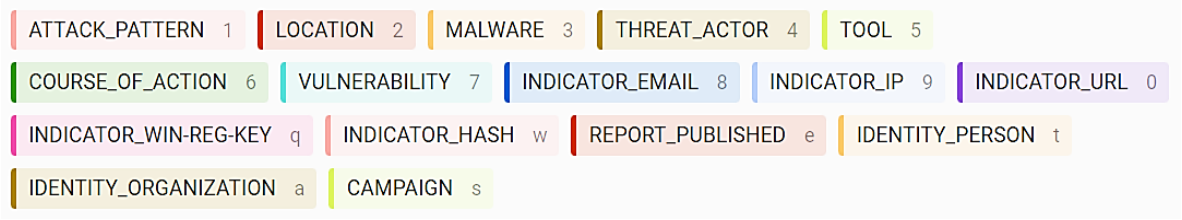



*Figura 7 – Annotazioni semantiche realizzate mediante il software LabelStudio®*[187]

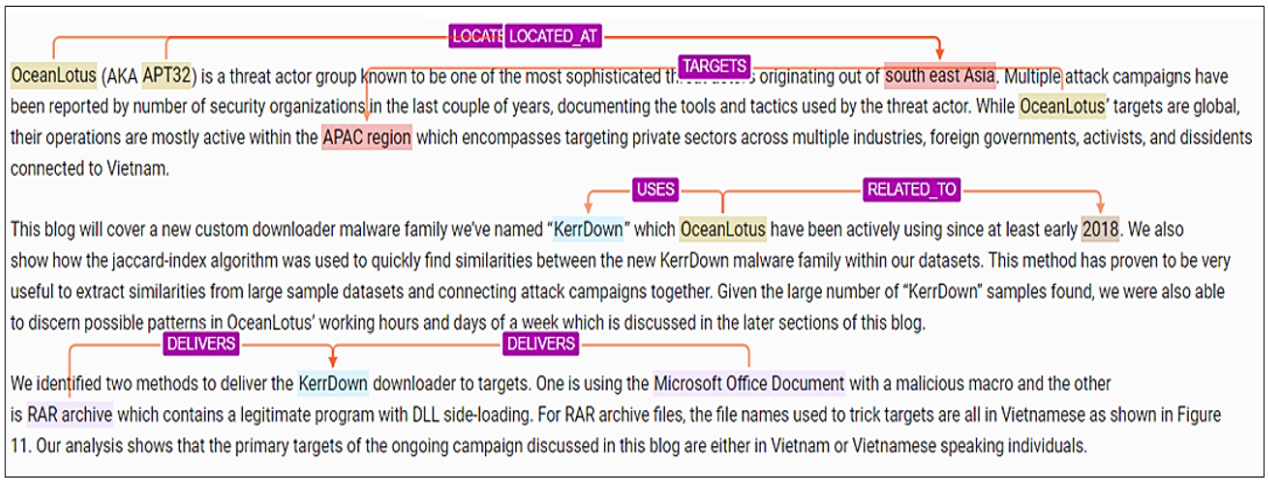



*Figura 8 – Esempio di labelizzazione semantica e delle relazioni mediante il software LabelStudio®*[188]

[182] Relazione definita anche "IS-A", dall'inglese *"is a":* "è un/a". Può essere integrata la *label* "IS NOT".
[183] Relazione definita anche come "PART-OF", traduzione dall'inglese: "è parte di".
[184] Relazione definita "SUBSET", ovvero "sottoinsieme".
[185] Relazione definita come "SUPERSET", ovvero "sovra-insieme".
[186] Relazione definita anche come "HAS-CAN", ovvero: "è – può/è in grado di fare". Possono essere integrate le *labels* "HAS NOT-CANNOT"
[187] Fonte immagine: grafico autoprodotto mediante *software LabelStudio®*
[188] Ibidem.

L'annotazione ai fini della realizzazione di reti semantiche viene sviluppata negli anni '50-'60 del '900, basandosi sugli studi del linguista cognitivo Charles Peirce di inizio secolo e sul suo grafo (da egli definito "esistenziale"), già caratterizzato da nodi ed archi.

La realizzazione di grafi dei sensi aveva come scopo quello di costruire un dizionario elettronico in grado di tradurre il linguaggio naturale nel linguaggio logico-matematico degli elaboratori elettronici.

Il sistema *Nude* ideato da Richard Hook Richens nel 1956, in collaborazione con il *Cambridge Language Research Unit,* prevedeva, ad esempio, una rete semantica come interlingua per la traduzione automatica, ossia come linguaggio intermedio da utilizzare nella traduzione di un testo da una lingua naturale ad un'altra[189].

Nel corso degli anni '70 e '80 del '900, le reti semantiche vengono, poi, sviluppate sotto forma di ipertesto, contribuendo – in questo modo, anche se indirettamente – alla nascita dei linguaggi informatici ipertestuali, come l'HTML *(HyperText Markup Language)*, atto alla formattazione e impaginazione di documenti, appunto, ipertestuali.

All'inizio degli anni '90, le reti semantiche iniziano ad essere utilizzate anche negli studi sull'intelligenza artificiale e per la realizzazione dei primi sistemi esperti.

Un altro esempio degno di nota di rete semantica è il *database* lessicale inglese *WordNet®*, con cui è possibile rappresentare descrizioni logiche usando grafi concettuali – ovvero, notazioni logiche basate sui "grafi esistenziali" di Charles Sanders Peirce[190] – con potere un espressivo pari allo standard della "logica dei predicati del primo ordine" (o "calcolo dei predicati").

La teoria della logica dei predicati del primo ordine è un particolare sistema formale con cui è possibile esprimere enunciati e/o dedurre le loro conseguenze logiche. Viene chiamato anche "logica del primo ordine", ed è spesso utilizzato per la formulazione di un linguaggio atto a rappresentare la conoscenza.

[189] R. Richens Hook, *Preprogramming for Mechanical Translation,* 1956, Vol. 3, n°1. Cfr. https://aclanthology.org/www.mt-archive.info/50/MT-1956-Richens.pdf

[190] John F. Sowa, nel 1984, usò i grafi concettuali per rappresentare gli schemi utilizzati nei *database*, fornendo applicazioni ad un'ampia gamma di argomenti nei riguardi dell'intelligenza artificiale, dell'informatica e delle scienze cognitive.

Si presenta, infatti, *«come il sostituto formalizzato del linguaggio naturale e propone un meccanismo inferenziale* […] *per la manipolazione della conoscenza: l'attività di deduzione genera la conoscenza implicitamente contenuta nella conoscenza iniziale, che compare sotto forma di assiomi. In particolare, consente di costruire database deduttivi»*[191].

La logica del primo ordine introduce quantificatori esistenziali e universali, predicati, funzioni, variabili e costanti che apportano una maggiore potenza espressiva al calcolo dei predicati – ovvero, al calcolo delle espressioni linguistiche collegate a uno o più elementi del dominio per poter formare una frase – estendendo, così, il concetto di logica proposizionale.

Con *WordNet*®, ogni elemento sintattico viene associato ad un insieme sinonimi cognitivi denominati *synset*[192], ognuno dei quali esprime un concetto puntuale e distinto. I *synset* (che, in questo caso, corrispondono ai singoli *token)* sono interconnessi tra loro per mezzo di relazioni semantico-lessicali e di dominio/concettuali.

La maggior parte delle relazioni di *WordNet*® collega le parole della stessa parte del discorso (POS – *part-of-speech*), nonché tra parti del discorso differenti (relazioni *cross-POS)*, le quali includono *links* morfo-semantici tra parole che condividono la stessa radice.

Per mezzo di questa interconnessione, *WordNet*® è in grado di collegare tra loro non solo le singole stringhe di parole, bensì anche i significati intrinsechi di ogni unità minima. Di conseguenza, le parole che si trovano in prossimità l'una dall'altra vengono semanticamente e sintatticamente disambiguate, creando reti semantiche concettualmente coerenti.

[191] https://ugochirico.com/page/rappresentazione-della-conoscenza.aspx

[192] Nel *tool* di analisi linguistica COGITO STUDIO® di *Expert.ai*, i *synset* vengono indicati come *syncon*, e sono associati ad un numero identificativo (ID) contenuto nel *Knowledge Graph* del sistema – denominato *Sensigrafo*®.

### 2.3.1 *Il processo di disambiguazione*

La disambiguazione *(o Word Sense Disambiguation)* è il processo con il quale si precisa il significato di una parola (disambiguazione semantica) o di un insieme di parole (disambiguazione sintattica), che rimanda a significati diversi a seconda del contesto cui si fa riferimento, al fine di risolvere le ambiguità che sono proprie di ogni lingua.

Tale processo viene eseguito da *software* o *Application Programming Interface* (API) di elaborazione del linguaggio naturale che, mediante appositi algoritmi di intelligenza artificiale – più o meno complessi, a seconda della struttura del codice linguistico preso in analisi – hanno capacità di ricognizione ed interpretazione non solo delle unità minime di significato, ma della totalità dei concetti espressi in un determinato dato testuale.

L'essere umano sviluppa e raffina la capacità di disambiguare parole e frasi mano a mano che espande le proprie conoscenze linguistico-culturali – nonché il proprio bagaglio esperienziale – tanto da riuscire a processare cognitivamente il significato di parole e frasi associandolo ad un contesto di appartenenza, in un tempo molto breve se la posizione dei confini di tale significato non è vaga[193].

Il problema principale del processo di disambiguazione da parte di tecnologie atte all'interpretazione automatica del linguaggio riguarda appunto l'identificazione dei molteplici significati che una parola (intesa come "inventario di senso") può esprimere. Questa proprietà è detta *polisemia* (dal greco *polysemos*, "dai molti significati").

Alcuni *framework* di analisi linguistica sono in grado di risolvere casi di ambiguità lessicale[194], grammaticale, semantica e sintattica grazie al "grafo della conoscenza" (o *Knowledge Graph*): questa fondamentale risorsa in grado di applicare una lettura funzionale all'*input* testuale, identificandone il contesto corretto, e risolvendo i conflitti di significato, *token* dopo *token*. Tale grafo fornisce descrizioni grafiche sull'ontologia di COGITO STUDIO® sul funzionamento dei sistemi cognitivi: qualsiasi testo non strutturato viene compreso ed analizzato dal programma mediante sistemi di intelligenza artificiale basati sul *Natural Language Understanding* (NLU) e sul *Natural Language Processing* (NLP).

[193] In riferimento al concetto di *buffering* espresso nel paragrafo precedente.

[194] Qualora ci si imbattesse in un caso di polisemia, l'analista linguistico/a potrà forzare il disambiguatore centrale del sistema affinché la macchina possa ignorare – e, quindi, non considerare in quel dato progetto linguistico – il *token* di disturbo.

Prendendo come riferimento il *tool* di analisi linguistica COGITO STUDIO®, presento, di seguito, un esempio di disambiguazione fornendo come *input* il seguente testo:

***«Gli attacchi informatici all'Ucraina fanno parte di una strategia che serve ad agevolare le azioni militari russe sul campo. Oggi le guerre si combattono (e si vincono) anche dal cyberspazio».*** [195]

| **SENTENCE** | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **PRINCIPALE** | | | | | | | | | | |
| **NP** | | | **PP** | | | **VP** | | **PP** | | |
| **ART** | **NOU** | | **PRE** | **N/A** | **NPR** | **VER** | | **PRE** | **ART** | **NOU** |
| **ART** | **NOU** | **ADJ** | **PRE** | **PNT** | **NPR** | **VER** | **NOU** | **PRE** | **ART** | **NOU** |
| Gli | attacchi | informatici | all | ’ | Ucraina | fanno | parte | di | una | strategia |

| **SENTENCE** | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **RELATIVA** | | **IMPLICITA** | | | | | | | | **GEN** |
| **RP** | **VP** | **VP** | | | **NP** | | **DP** | | | **N/A** |
| **PRO** | **VER** | **PRE** | **VER** | **ART** | **NOU** | | **ADJ** | **ADV** | | **PNT** |
| **PRO** | **VER** | **PRE** | **VER** | **ART** | **NOU** | **ADJ** | **ADJ** | **PRE** | **NOU** | **PNT** |
| che | serve | ad | agevolare | le | azioni | militari | russe | sul | campo | . |

| **SENTENCE** | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **PRINCIPALE** | | | | | **PRINCIPALE** | | | | | **PRINCIPALE** | | | **GEN** |
| **DP** | **NP** | | **VP** | | **N/A** | **CP** | **VP** | | **N/A** | **PP** | | | **N/A** |
| **ADV** | **ART** | **NOU** | **PRO** | **VER** | **PNT** | **CON** | **PRO** | **VER** | **PNT** | **ADV** | **PRE** | **NOU** | **PNT** |
| **ADV** | **ART** | **NOU** | **PRO** | **VER** | **PNT** | **CON** | **PRO** | **VER** | **PNT** | **ADV** | **PRE** | **NOU** | **PNT** |
| Oggi | le | guerre | si | combattono | ( | e | si | vincono | ) | anche | dal | cyberspazio | . |

***Figura 9 – Processo di disambiguazione con COGITO STUDIO® di Expert.ai*** [196]

[195] M. Reina, *Ucraina, come funzionano gli attacchi malware dei russi*, 2022. Cfr. https://www.webnews.it/2022/03/05/ucraina-come-funzionano-gli-attacchi-malware-dei-russi/

[196] Fonte: grafico autoprodotto.

| TABELLA DEGLI ACRONIMI | | | |
|---|---|---|---|
| CP > Complementizer phrase | NP > Noun phrase | PP > Prepositional phrase | VP > Verb phrase |
| DP > Determiner phrase | RP > Relative phrase | VER > Verbo | PRO > Pronoun |
| NPR > Nome proprio | CON > Congiunzione | ART > Articolo | NOU > Noun |
| ADV > Adverb | ADJ > Adjective | PRE > Preposition | PNT > Punteggiatura |

Come si può notare dagli esempi grafici riportati, viene eseguita l'attività di disambiguazione procedendo per *layers:* stabilendo, quindi, una gerarchia tra le tipologie di analisi linguistica da affrontare.

L'analisi del testo svolta dal *software* è, quindi, costituita da diverse fasi consecutive[197], che comprendono:

- l'analisi sintattica atta ad individuare un primo livello di gruppi di parole rappresentati da locuzioni[198] (nominali, verbali, avverbiali, preposizionali, oggettivali, ecc.), seguìto da un secondo livello rappresentato da proposizioni, ed infine un terzo livello rappresentato da periodi[199];
- l'analisi lessicale e grammaticale: il testo viene suddiviso in singoli *tokens* prendendo come riferimento i caratteri separatori presenti tra le unità minime di significato, in genere lo spazio *blank* e i segni di punteggiatura, al fine di identificare le parti variabili del discorso;
- l'analisi semantica: questo livello di analisi permette di associare le parole ai significati sfruttando l'interazione tra il motore semantico per la comprensione del linguaggio naturale ed il *Sensigrafo*®, ovvero il "grafo dei sensi/della conoscenza". Il disambiguatore procede filtrando l'elenco delle opzioni possibili per ciascuna parola, dopodiché considerando il contesto a cui ogni parola viene associata, al fine di restituire all'utente il significato corretto.

[197] https://www.expert.ai/it/the-platform/tecnologia/natural-language-understanding/
[198] *«Le locuzioni possono essere assimilate a varie classi di parole, di cui condividono distribuzione e funzioni»*. Cfr. https://www.treccani.it/enciclopedia/locuzioni_%28Enciclopedia-dell%27Italiano%29/
[199] In grammatica, il periodo è un'unità complessa del discorso, composta da frasi semplici, o proposizioni, combinate in una sola struttura dal senso compiuto.

### 2.3.2 La metafora: deviazione e trasposizione di senso nella computazione

*«L'intelligenza artificiale potrebbe presto capire le metafore linguistiche»*, esordisce così l'articolo omonimo del *CORDIS – servizio Comunitario di Informazione in materia di Ricerca e Sviluppo* dell'Unione Europea[200]– che si interroga sulla serie di vantaggi che porterebbe un'intelligenza artificiale in grado comprendere le metafore.

In generale, la ricerca – sempre più all'avanguardia – ha fatto enormi passi avanti nel campo dell'elaborazione profonda[201] del linguaggio naturale, in relazione alle diverse attività quotidiane dell'essere umano nel prossimo futuro.

L'obiettivo dei *team* di sviluppo e di ricerca in ambito A.I. è quello di raggiungere una comprensione metaforico-concettuale del linguaggio naturale che riesca ad andare oltre le capacità dei sistemi attuali[202]. Si punta, quindi, alla restituzione da parte della macchina di *output* linguistici che riescano a rimandare al significato sorgente della metafora, persino nei casi in cui l'*input* consista in *post* e commenti raccolti online – manifestazione per antonomasia della lingua dell'uso di una data comunità linguistica, in un determinato periodo di tempo.

Non bisogna dimenticare che le molteplicità e diversità linguistiche sono il prodotto delle caratteristiche proprie di ogni lingua, influenzate (nonché, generate) da fattori ambientali, storici, sociali, nonché individuali – ovvero, propri di ogni individuo, creatore di *parole* a partire da grammatica *(langue)* socialmente accettata e condivisa[203].

L'ipotesi di Sapir-Whorf, conosciuta anche come teoria della relatività linguistica, afferma che *«lo sviluppo cognitivo di ciascun* [individuo sia] *influenzato dalla lingua che parla».*[204] Pertanto, il modo di esprimersi (nonché di pensare) di ognuno porta con sé le caratteristiche della propria lingua materna e/o della lingua appresa dopo il periodo critico e sviluppata fino al raggiungimento di un bilinguismo consecutivo tardivo[205].

---

[200]Cfr. cordis.europa.eu/article/id/182697-artificial-intelligence-may-soon-understand-language-metaphors/it

[201] In riferimento alla scienza del *Deep Learning* – dell'apprendimento *profondo.*

[202] Il nuovo sistema di *Cognitive Computing* COGITO STUDIO® di *Expert.ai* si configura oggi come una *hybrid NL Platform*, prevedendo la combinazione di sistemi di *Machine Learning* con gli approcci basati sulla conoscenza *(rule-based system).*

[203] In riferimento alla dicotomia *Langue/parole* di Ferdinand de Saussure, ampiamente elaborata nel suo *Cours de linguistique générale* del 1916.

[204]P. Diadori, *Insegnare italiano a stranieri*, Milano, Le Monnier, 2011, p. 9.

[205] *«Si parla di bilinguismo consecutivo tardivo quando il bambino inizia ad apprendere una seconda lingua (L2) dopo aver raggiunto buone competenze nella lingua madre, che verrà quindi definita L1»* – FLI:

Come anche sostenuto da Avram Naom Chomsky: *«Ragionare non significa elaborare simboli senza significato»*. L'essere umano è l'unico animale dotato di immaginazione, ed in grado di realizzare collettivamente[206] ciò che idea. Tale capacità cognitiva di concretizzare idee, figure e pensieri, si riflette nel modo in cui l'individuo concettualizza il mondo: lo fa attraverso delle strutture immaginative che lui stesso ha creato, e che si riflettono nelle sfaccettature della sua cultura e nelle proprietà della sua lingua; tra queste, le figure retoriche, di cui la metafora ne è un iponimo.

Gli studi sulla metafora hanno rappresentato uno dei capitoli principali della storia della linguistica cognitiva, perché *«la comprensione e l'interpretazione di nuove figure generano conoscenza.* […] *La figura non nasce aggiungendo qualcosa alla parola, ma nasce per mezzo di intersezioni, antitesi, inclusioni, soppressione di aree concettuali»*.[207]

Infatti, secondo Lakoff e Johnson: *«Le metafore concettuali rappresentano il fulcro stesso del pensiero e della ragione, in quanto si basano su un sistema concettuale culturale, in parte convenzionale (dato dalla lingua), in parte risultato di un'attività cognitiva che conduce a metafore di nuova creazione»*.[208]

La metafora è un tropo (o traslato), ovvero è *«l'utilizzo retorico di una deviazione e trasposizione di significato»*[209]*;* si manifesta quando l'uso di un'espressione normalmente legata ad un determinato campo semantico viene attribuito – mediante un processo di estensione – ad altri oggetti o campi semantici. In altre parole, si realizza *«quando un vocabolo o una locuzione sono usati per esprimere un concetto diverso da quello che normalmente significano»*[210].

Le metafore sono rappresentazioni di mappature tra *frames*; la struttura del *frame* del dominio di origine viene mappata sulla struttura del frame del dominio di destinazione.

---

Federazione Logopedisti Italiani, *Conoscere il bilinguismo,* 2014. Cfr. https://aler.fli.it/files/2014/03/BILINGUISMO-DEFINIZIONI.pdf

[206] In riferimento, nuovamente, al libro di Y. N. Harari, *Da Animali a Dèi. Breve storia dell'umanità,* 2011, trad. it. G. Bernardi, Bompiani, 2017.

[207] C. Diana, *La linguistica Cognitiva*, 2015. Cfr. https://filosofiafammiunthe.wordpress.com/2015/07/08/linguistica-cognitiva/amp/

[208] C. Dalla Libera, *Le metafore concettuali in un approccio comunicativo nell'apprendimento delle lingue straniere,* Università Ca' Foscari Venezia, Vol. 6, N. 1, EL.LE, 2017, p. 28. In riferimento a: G. Lakoff, L. Johnson, *Metaphors We Live by,* Chicago, Chicago University Press, 1980.

[209] B. M. Garavelli, *Il parlar figurato. Manualetto di figure retoriche,* Laterza, 2010, p. 9.

[210] https://www.treccani.it/enciclopedia/metafora/

La lingua – come ogni altro sistema semiotico – si costituisce nella forma di una macchina in grado di realizzare previsioni e di risolvere problemi, ovvero, in grado di determinare – come menzionato all'inizio di questo lavoro di tesi – sia l'elemento definibile come "bontà dell'esemplare" di una data categoria concettuale, che soluzioni rispetto al meccanismo di generazione di frasi semplici o complesse. Rispetto a questa meccanicità, *«la metafora rappresenta il guasto, e al tempo stesso il motore di rinnovamento della lingua».*[211]

Umberto Eco suggerisce che il problema dell'interpretazione delle metafore si possa risolvere da punto di vista semiotico realizzando un processo di condensazione semantica, in cui *«le unità di contenuto* [o sememi] *coinvolte acquistano proprietà e ne perdono delle altre».*[212]

Lakoff e Johnson (1980) propongono, inoltre, far distinzione tra due tipi di espressioni simbolico-evocative: la metafora concettuale (es. *«il tempo è denaro», «la discussione è una guerra»),* e quella linguistica (es. *«perdere tempo», «risparmiare tempo»).*

Il riconoscimento dell'ubiquità della metafora nel linguaggio ha portato a un grande interesse per l'identificazione automatica di espressioni simboliche. La difficoltà di definire un'unica visione teorica nei riguardi della metafora è la causa della varietà di sistemi di NLP che si pongono come obiettivo quello di riuscire a distinguere automaticamente i significati di una parola o di un periodo in termini metaforici, nonché di individuarne l'ampliamento dei confini di significato partendo dagli usi letterali della parola (o dei concetti) interessati.

Studi sperimentali, basati su modelli di analisi del linguaggio neurale, si mostrano in grado di rilevare il contesto corretto di ciascun termine, identificandone funzioni ed usi senza dover impiegare esplicitamente connessioni e rilevanze semantiche tra le parole interessate nel tropo, nonché i concetti ad essi correlati.

Sfruttando i vantaggi della semantica distribuzionale – che comprende una serie di teorie e metodi di linguistica computazionale per lo studio della distribuzione semantica delle parole nel linguaggio naturale[213]– lo studio *"Deep Learning based, end-to-end metaphor detection in Greek language with Recurrent and Convolutional Neural Networks"*, attuato da alcuni ricercatori dell'Università di Atene, dimostra come si possa

[211] G. Altamura, *Umberto Eco, il giocoliere dell'Intelligenza*, Università di Bari Aldo Moro, 2019, p. 46.
[212] Ibidem.
[213] *«Tali modelli derivano da una prospettiva empiristica e assumono che una distribuzione statistica dei termini sia preponderante nel delinearne il comportamento semantico».* Cfr. https://biblio.toscana.it/argomento/Semantica%20distribuzionale

supplire al confronto semantico tra le varie entità/parole, per *«sostituirlo con un confronto a livello numerico della loro rappresentazione distribuzionale nello spazio vettoriale».* [214]

L'identificazione e l'interpretazione computazionale delle metafore, ad oggi, si sono basate su una varietà di strumenti e *features*, come la modellazione statistica[215], la categorizzazione tassonomica, il *clustering*[216], o la regressione logistica[217].

Per eseguire la *Metaphor Analysis* si può adottare sia un approccio introspettivo ed induttivo di tipo *top-down*, che un approccio deduttivo, di tipo *bottom-up* nei confronti dei *corpora* testuali selezionati*;* quest'ultimo pone l'enfasi sul modo in cui le metafore vengano utilizzate partendo da un testo nel suo complesso, con l'obiettivo di mostrare un quadro più chiaro dell'inventario delle espressioni metaforiche individuate, oltre a fornire una percentuale più accurata della loro frequenza d'uso.

L'analisi linguistico-investigativa ideale sarebbe quella in grado di rilevare ed estrarre dette figure retoriche all'interno di un testo fornendo informazioni semantiche su ciascuna di esse, fornendo una puntuale identificazione dei domini *(semantic frames)* di origine e di destinazione, delle eventuali dipendenze gerarchico-tassonomiche tra questi ultimi, nonché dei collegamenti tra domini e le metafore concettuali soggiacenti.

Come mostrato nello schema seguente, ad esempio, se considerassimo gli assunti: *«la povertà infetta la società»* e *«il crimine sta affliggendo la nazione»*, e se, a titolo esemplificativo, li traslassimo nelle seguenti metafore: *«la povertà è una malattia»* e *«il crimine è una malattia»*, possiamo notare, a livello di analisi computazionale, come tali assunti ricerchino, a livello semantico, eredità da metafore di più ampio spettro (ad esempio: *«I problemi sociali sono afflizioni fisiche»).*

---

[214] K. Perifanos, E. Florou, D. Goutsos, *Deep Learning-based, end-to-end metaphor detection in Greek language with Recurrent and Convolutional Neural Networks*, Department of Linguistics, National and Kapodistrian – University of Athens, Greece, 2007, p. 1. Cfr. https://arxiv.org/pdf/2007.11949.pdf

[215] Metodologia semplificata e matematicamente formalizzata per approssimare la realtà e fare previsioni a partire da queste approssimazioni – GMSL Srl, *Cos'è la modellazione statistica?* 2018. Cfr. https://www.gmsl.it/wp-content/uploads/2018/09/Articolo-XLSTAT-Modellazione-statistica.pdf

[216] *«Il clustering consiste in un insieme di metodi per raggruppare oggetti in classi omogenee. Un cluster è un insieme di oggetti che presentano tra loro delle similarità, ma che, per contro, presentano dissimilarità con oggetti in altri clusters».* Cfr. https://www.dataskills.it/tecniche-di-clustering/#gref

[217] *«In statistica, il modello logit, noto anche regressione logistica, è un modello di regressione non lineare utilizzato quando la variabile dipendente è di tipo dicotomico».* – IBM.com, *Cos'è la regressione logistica?* 2020. Cfr. https://www.ibm.com/it-it/topics/logistic-regression

Seguendo, in questo caso, una dinamica *waterfall* inversa (di tipo *bottom-up*) – possiamo stabilire che, in computazione, i concetti espressi tramite metafore, procedano logicamente nella ricerca di matrici semantiche contenenti anch'esse, a loro volta, metafore, e che queste ultime vengano ereditate man a mano che si procede in una indagine (in questo caso) per iperonimia concettuale, fino a giungere alla ipotetica macro-metafora di partenza:

**«Le condizioni sociali valutate negativamente sono fisicamente nocive».**

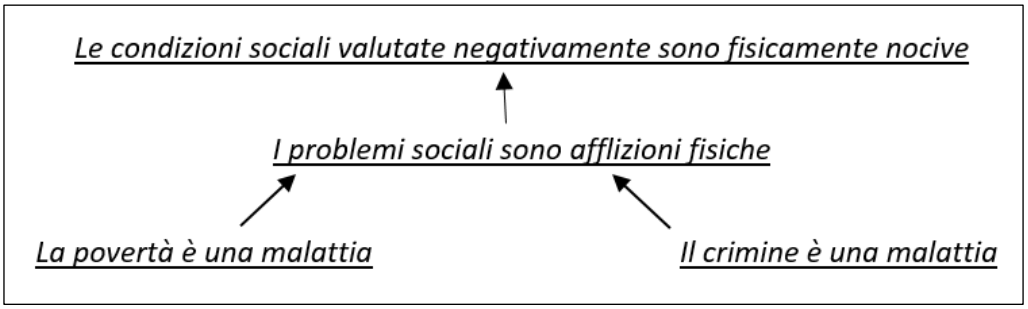



***Figura 10 – Processo logico-metaforico***[218]

Il processo di rilevamento delle metafore in un testo è seguito dall'attività di annotazione di *tokens*; tali annotazioni possono essere utilizzate sia per validare e perfezionare il sistema utilizzato, che per procedere ad ulteriori tipi di analisi basate sulla complessità del *corpus* (analisi di tipo *bottom-up).* Le informazioni fornite dalle annotazioni includono il tipo di lemma, la categoria grammaticale (POS), il dominio/*frame,* e la struttura tassonomica dei *fram*es di entrambi i termini di origine e di destinazione.

Per rendere possibile tale analisi, è necessario disporre di strumenti di NLP come:

- lemmatizzatori, che procedono alla riduzione di una forma flessa di una parola alla sua forma canonica;
- *POS tagger,* che realizzano un'analisi grammaticale identificando la categoria lessicale (e relative sottocategorie) di ogni parola nel contesto nel quale viene utilizzata;
- *parser* di dipendenze sintattiche[219], per eseguire un'analisi del flusso continuo di dati in *input,* al fine di determinare la correttezza della struttura del flusso stesso, mediante l'utilizzo di una determinata grammatica formale.

[218] Fonte: grafico autoprodotto.
[219] L'analisi delle dipendenze sintattiche è la fase più dispendiosa in termini di *effort* dal punto di vista computazionale.

## 2.4 *Frames e modelli cognitivi idealizzati nell'A.I.*

I domini (o *frames)* si configurano come rappresentazioni concettuali della conoscenza enciclopedica dell'essere umano. Per questo, il *frame* di ogni concetto viene inteso come infinitamente complesso, mai oggettivo, nonché legato a quella che il linguista John Searle postula come "sfondo"[220] *(Background of Meaning)*[221], ovvero legato al significato non puramente semantico di parole ed espressioni linguistiche del singolo individuo, bensì con cognizione di sfondo, che struttura la coscienza e permette alla percezione di avere luogo.

La struttura semantica del concetto ha origine in quella cognitiva: le unità minime di significato, simboliche, rivelano quindi le strutture cognitive del linguaggio.

La duplicità del termine "concetto" (nel senso semantico e cognitivo) può creare sia delle difficoltà interpretative e metodologiche, che costituire un paradigma interdisciplinare della linguistica cognitiva.

La teoria del *frame* (in inglese, *Frame Theory)*, infatti, presenta come principale oggetto di analisi il modo in cui il linguaggio mette in prospettiva una determinata concettualizzazione del mondo attraverso l'espressione linguistica: l'associazione di una determinata espressione ad un dato modello concettuale permette la creazione di nuovi *layers* di modellizzazione[222], legati a loro volta cognitivamente a una data visione e considerazione del mondo, evocando e/o producendo *bias*[223], limiti o bisogni cognitivi.

Al contempo, la scelta di una data espressione permette anche la riconfigurazione, l'ampliamento, oppure la restrizione del campo di cognizione di un dominio, definito anche come "cornici interpretative" con le quali si organizza il proprio pensiero.

È da questo presupposto che, alla fine del 1960, si diffonde un nuovo approccio alla semantica, denominato *Frame Semantics*[224]: uno studio sulla semantica linguistica che si

[220] Concettualmente legato alla psicologia della *Gestalt,* movimento incentrato sui temi della percezione e dell'esperienza. In questo caso, facendo riferimento all'allineamento *figura-sfondo* nell'organizzazione della mente, sul modo in cui il nostro cervello interpreta ed organizza i dati della percezione.
[221] In riferimento all'opera: J. R. Searle, *The Background of Meaning*, Cambridge University Press, 1980.
[222] Si rimanda al paragrafo 1.5.1 *"Types e Tokens".*
[223] Vengono identificate quattro categorie che gli analisti di Intelligence devono considerare: *«personali, culturali, organizzativi, a cui si aggiungono le distorsioni cognitive»*. – M. Caligiuri (a prefazione di), A. Teti, Cyber Espionage e Cyber Counterintelligence, Rubbettino Editore, 2018, p. 17.
[224] In riferimento all'opera: J. C. Fillmore, *Frame semantics – Cognitive Linguistics: Basic Readings*, Hanshin Publishing, 1982.

interroga sul sapere necessario al fine di accedere alla comprensione di un enunciato linguistico, sia esso una parola, un sintagma, una frase o un intero testo.

Il significato di parole o di espressioni linguistiche viene inteso nei termini di *frame* semantici, strutture e categorizzazioni delle conoscenze, contenenti elementi *culture-specific.* Infatti, Charles J. Fillmore indica che ogni qual volta si è impegnati in attività di percezione, pensiero e comunicazione, ha luogo il processo di *framing* ovvero: *«The appeal* […] *to structured ways of interpreting experiences»*[225], ovvero modi strutturati di interpretare l'esperienza, perché *«the literal meaning of an ordinary sentence is often consistent with unacceptable interpretations»*[226].

In ogni situazione, attraverso il richiamo alla propria conoscenza *«la memoria attiverebbe un pacchetto di dati rilevanti a una data situazione tale da consentire di comprendere, cioè riconoscere e poi strutturare, quella data circostanza»*[227].

Tale "pacchetto" – definito in precedenza in questo lavoro di tesi anche come *bucket*[228] – risulta in un *Idealized Cognitive Model* (ICM)[229], ovvero in un modello cognitivo idealizzato, che equivale, a sua volta, alla definizione di dominio cognitivo. L'ICM equivale, quindi, *«ad una simbolica convenzione semantica contenente un insieme di tratti cognitivi acquisiti mediante l'esperienza»*.[230]

Sia gli ICM, che i domini cognitivi, sono caratterizzati da alcune caratteristiche che li rendono facilmente richiamabili alla memoria *(buffering)* di ogni persona, come, ad esempio, la massima rappresentatività – ovvero se un dato elemento un dato elemento possiede i tratti caratteristici che ne permettono l'associazione ad un dato *frame* – ed il minimo sforzo cognitivo nell'identificazione delle categorie, definito come tipicità, mediante la loro associazione a proprietà diagnostiche[231].

---

[225] J.C. Fillmore, *Frame semantics and the nature of language*, University of California-Berkeley, 1976. Cfr. https://www1.icsi.berkeley.edu/pubs/ai/framesemantics76.pdf
[226] D. Sosa, *Checking Searle's Background*, Luis Manuel Valdés-Villanueva Publishing, 1999, p. 109.
[227] P. Scaruffi, *Thinking about Thought,* iUniverse Publisher, 2014. Cfr. https://www.scaruffi.com/univ/tat1.pdf
[228] Si rimanda al paragrafo 2.2 *"Fondamenti teorici della linguistica cognitiva"*.
[229] Concetto postulato e proposto dai linguisti George Lakoff e Gilles Fauconnier.
[230] A. Kosz, *La rappresentazione delle conoscenze – diversi modelli delle strutture concettuali nell'ambito della linguistica cognitiva*, Università di Slesia Katowice, 2020, p. 15.
[231] Il linguaggio diagnostico è una modalità comunicativa di *prestigio*, una tendenza a descrivere l'intera realtà in termini investigativi; è uno stile di comunicazione socialmente distribuito che si sviluppa sulla base del duplice significato della parola diagnosi: quello *di processo investigativo* (tecnica diagnostica) e quello di *prodotto linguistico* (formulazione diagnostica).

Per modello cognitivo – come indicato in precedenza – si intende la rappresentazione mentale contenente la struttura essenziale degli oggetti o degli eventi nel mondo reale, sulla base di quanto esperito e conosciuto dal singolo individuo, altresì necessariamente condiviso con la propria comunità linguistico-culturale ai fini dell'interazione sociale.

Il risultato dei processi conoscitivi e percettivi basati sull'esperienza e sulla conoscenza acquisita prende la forma di "costrutto mentale": idee e concetti si strutturano in maniera complessa, sotto forma di "costrutti concettuali" – definiti, appunto, ICM – modelli cognitivi idealizzati (G. Lakoff, 1987), oppure *frame* (Charles J. Fillmore, 1975), od anche *script* o scenari (Abelson, 1977), nonché – come trattato precedentemente – reti semantiche (R. Quillian, 1968).

Lo scopo principale dell'analisi linguistica – in particolare semantica, anche in ambito computazionale – è comprendere cosa i concetti riescono a rappresentare, ed in che modo essi riescano a trovare una significazione condivisa; inoltre, mediante la struttura lessicale-semantica del concetto prestabilita, si può risalire alla struttura cognitiva rappresentata sia nella lingua, nonché attraverso essa, mediante un approccio metacognitivo e metalinguistico.

I modelli cognitivi, in quanto strutture conoscitive (o "modelli conoscitivi"[232]), permettono una comprensione e una interpretazione dello scenario nel processo cognitivo.

Come per le reti semantiche[233], è difficile dare una definizione di *frame* che sia al contempo puntuale e generale al fine di comprendere le realizzazioni che trovano riscontro nei vari sistemi e tecnologie di intelligenza artificiale, tra cui il *Cognitive Computing* con alla base il *Natural Language Processing.*

I *frame*, proprio come le suddette reti di significato, sono strutture rappresentanti dati di conoscenza: è quindi di fondamentale rilevanza riuscire ad estrarre (nonché utilizzare) mediante appositi algoritmi e processi di categorizzazione e/o estrazione – l'informazione in esse contenuta. Queste procedure eseguono i cosiddetti "processi inferenziali".

*«L'inferenza è un ragionamento logico mediante il quale si esercita il processo di conoscenza,* [e] *consiste nel produrre una conclusione* [output] *a partire da una serie di premesse* [in input]».[234] In questo processo vengono impegnate una moltitudine di funzioni

[232] In riferimento alla teoria degli spazi mentali, postulata dal linguista Gilles Fauconnier nel 1985.
[233] Come descritto nel paragrafo 2.3 *"Reti Semantiche – Interpretazione e confini del significato"*.
[234] F. Mattarella, *Inferenze e ragionamento causale*, 2022. Cfr. www.pensierocritico.eu/logica--cosa-sono-le-inferenze

cognitive complesse, tra cui la modellizzazione, la categorizzazione e l'elaborazione (meta)cognitiva.

Secondo alcuni esperti cognitivi, tra cui il linguista R. Jakobson, il contenuto informativo di ogni enunciato viene codificato all'interno di stringhe di elementi sintattici che, a loro volta, costituiscono una frase; i processi di codifica e di decodifica delle informazioni, quindi, sono legati ai procedimenti inferenziali di comprensione, intesi come sistemi di segni indispensabili alla comunicazione[235].

All'interno del contesto dell'A.I., i *frame* si configurano come "strutture dati" per la rappresentazione della conoscenza umana[236] che permettono la suddivisione di tale conoscenza in sottostrutture rappresentanti delle situazioni stereotipate[237].

- Esempi di lemmi associati a *frame* nell'ontologia di COGITO STUDIO®.

Come precedentemente menzionato, ogni lemma può avere molteplici relazioni di appartenenza *(links)* con categorie/domini differenti. Per dimostrare questo assunto, presento un caso di polisemia processato dal *tool* di applicazione di analisi linguistica sopra citato.

| | Linked from | Definition | Domain 1 | Domain 2 | ID |
|---|---|---|---|---|---|
| supernomen | N parasite | An organism that grows, feeds, and is... | biology 30% | | 5676 |
| | N animal microorganism | An organism of microscopic or ultramicroscopic size... | microbiology ... | | 78269 |
| | Syncon | Definition | Domain 1 | Domain 2 | ID |
| syncons | N virus, v, V | A minute particle that lives as a... | microbiology ... | medicine ... | 5472 |
| | N computer virus, virus | An intrusive program that infects computer files... | software 40% | crime 40% | 127969 |
| | N virus | Anything that has a corrupting or poisonous... | | | 160140 |
| | Links to | Definition | Domain 1 | Domain 2 | ID |
| subnomen | N adenovirus | Any of a group of viruses including... | medicine 100% | | 5475 |
| | N alphavirus | A virus transmitted by mosquitoes that can... | microbiology ... | medicine ... | 203088 |
| | N animal virus | An animal pathogen that is a virus... | microbiology ... | | 5484 |
| | N arborvirus, arbovirus | A virus transmitted by bloodsucking arthropods, for... | biology 100% | | 161678 |
| | N arenavirus, arenaviridae | A virus of the family that causes... | microbiology ... | medicine ... | 203882 |
| | N bacterial virus, bacteriophage, bacteriophagy, phage | Virus parasitic in bacteria; multiplies within its... | microbiology ... | | 5482 |
| | N bunyavirus | A virus spread by mosquito bites that... | microbiology ... | medicine ... | 208057 |
| | N coronavirus | Any of a group of viruses that... | microbiology ... | medicine ... | 213275 |

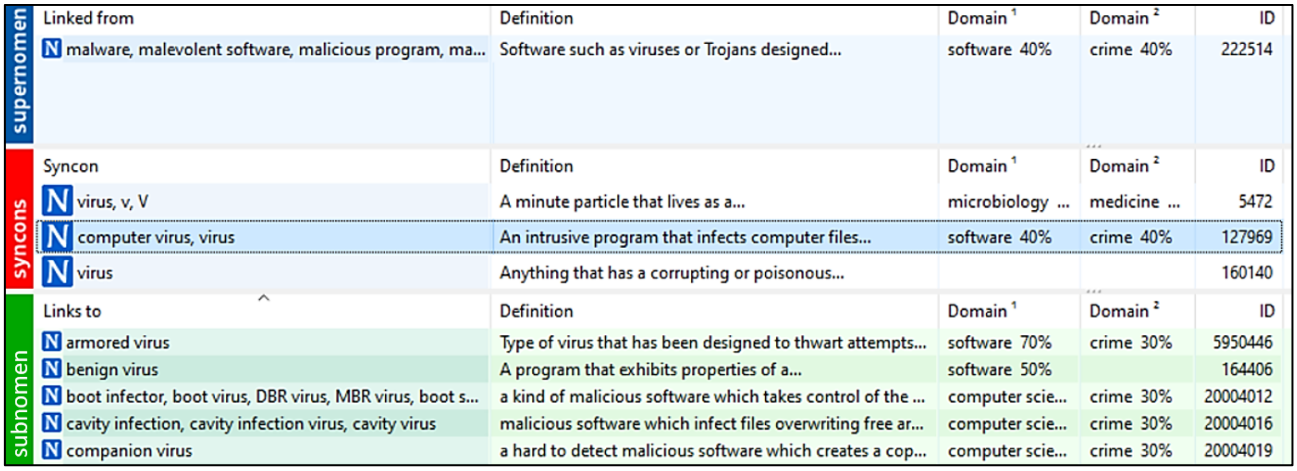
***Figura 11 – COGITO STUDIO®, lemma "virus" nell'ambito della medicina***[238]

[235] F. Bonacci, *Processi inferenziali Vs. processi di codifica/decodifica nei modelli di trasmissione dell'informazione fra individui*, Università della Calabria, 2009. Cfr. http://www.rifl.unical.it/index.php/rifl/article/view/130/120

[236] Il concetto è stato introdotto per la prima volta nel 1974 dal linguista M. Minsky, all'interno del suo articolo *"A Framework for Representing Knowledge"*. Cfr. https://courses.media.mit.edu/2004spring/mas966/Minsky%201974%20Framework%20for%20knowledge.pdf

[237] Un'altra terminologia indicante la rappresentazione della conoscenza nell'A.I. è "linguaggio *frame"*.

[238] Fonte: *feature Sensigrafo*® del *software* COGITO STUDIO®



| | Linked from | Definition | Domain [1] | Domain [2] | ID |
|---|---|---|---|---|---|
| supernomen | N malware, malevolent software, malicious program, ma... | Software such as viruses or Trojans designed... | software 40% | crime 40% | 222514 |
| | **Syncon** | **Definition** | **Domain [1]** | **Domain [2]** | **ID** |
| syncons | N virus, v, V | A minute particle that lives as a... | microbiology ... | medicine ... | 5472 |
| | N computer virus, virus | An intrusive program that infects computer files... | software 40% | crime 40% | 127969 |
| | N virus | Anything that has a corrupting or poisonous... | | | 160140 |
| | **Links to** | **Definition** | **Domain [1]** | **Domain [2]** | **ID** |
| subnomen | N armored virus | Type of virus that has been designed to thwart attempts... | software 70% | crime 30% | 5950446 |
| | N benign virus | A program that exhibits properties of a... | software 50% | | 164406 |
| | N boot infector, boot virus, DBR virus, MBR virus, boot s... | a kind of malicious software which takes control of the ... | computer scie... | crime 30% | 20004012 |
| | N cavity infection, cavity infection virus, cavity virus | malicious software which infect files overwriting free ar... | computer scie... | crime 30% | 20004016 |
| | N companion virus | a hard to detect malicious software which creates a cop... | computer scie... | crime 30% | 20004019 |

*Figura 12 – COGITO STUDIO®, lemma "virus" nell'ambito dell'informatica*[239]

Come si può notare delle immagini precedenti, il sistema riesce a processare il lemma "virus", fornendo una molteplice lettura nell'interpretazione dei suoi significati: nel primo caso, si propongono come *syncon* di partenza il lemma associato ai domini *microbiology e medicine,* indicando puntualmente pertinenti iperonimi *(supernomen)* e iponimi *(subnomen),* i cui corrispettivi *syncons* (in questo caso, sostantivi nella loro totalità) appartengono anch'essi alle reti semantiche correlate all'ambito della medicina, della biologia, della microbiologia e della virologia.

Nella seconda immagine presentata, invece, la scelta del lemma di partenza è stata indirizzata verso il *syncon* "virus", relegato, in questo caso, al dominio/frame dell'informatica. Anche questa volta, il sistema COGITO STUDIO® presenta una gerarchizzazione meronimica dei lemmi associati al *syncon* selezionato, e tutti i sostantivi *(nouns)* ad esso associati risultano pertenere al campo della *computer science,* o inscriversi nel dominio *software*[240].

Questo perché – a sostegno del pensiero del linguista R. Langacker – le strutture semantiche ed il significato delle espressioni lessicali sono determinati dai rispettivi domini cognitivi di appartenenza, ovvero dal risultato della rielaborazione dell'informazione da parte di un essere umano, a partire dall'attivazione di processi concettuali atti all'interpretazione e alla conoscenza della realtà.

[239] Ibidem.

[240] In COGITO STUDIO®, gli operatori linguistici possono inserire manualmente nuovi lemmi tramite la funzionalità *Sensigrafo Editor*®. Ogni lemma può essere associato ad un massimo di due domini, per i quali è obbligatorio applicare una percentuale di rilevanza/di appartenenza al dominio stesso.

Interpretazione e conoscenza che nascono come processi individuali e soggettivi, ma che – grazie alla funzione comunicativa della lingua, che tende alla comprensione reciproca – vengono sottoposti a processi di oggettivazione (di standardizzazione) nell'associazione del significato dell'espressione lessicale a determinati modelli cognitivi idealizzati (ICM), risultanti nel prodotto di come la memoria organizza le informazioni semantiche ai fini comunicativi.

## 2.5 *Categorizzazione tassonomica*

La categorizzazione è la funzione principale del processo di concettualizzazione, prospettiva idealizzata e semplificata[241] della conoscenza che ogni individuo possiede del mondo, del quale percepisce la necessità di fornire una rappresentazione mediante *types* – definiti anche come categorie, classi, ordini, tipi, *o buckets* – al fine di riuscire ad elaborare, interpretare e comprendere la realtà in cui vive.

Per concettualizzazione, pertanto, si intende il processo cognitivo che permette *«lo sviluppo o la costruzione di idee astratte dall'esperienza: è la nostra comprensione cosciente, non necessariamente "vera", del mondo».*[242]

Il processo cognitivo di categorizzazione – imprescindibilmente da quello di percezione – permette la lettura funzionale dei caratteri essenziali e (semi)costanti della propria esistenza, senza eccessive opacità della determinazione dei confini di significato nella decodifica di pensieri, comportamenti e concetti; inoltre, tramite detta operazione, è possibile formulare nuovi contenuti della conoscenza, nonché rievocare una sequenziazione dei concetti già radicati.

Secondo Ronald Langacker, le costruzioni tassonomiche[243] costituiscono *«un inventario strutturato della conoscenza che un parlante ha della propria lingua».*[244]

Tale inventario viene rappresentato sotto forma di rete tassonomica, composta da nodi che definiscono e regolano, seguendo una relazione di meronimia, la disposizione del modello/ prototipo di una determinata categoria, nonché le sue specifiche concettuali man mano che ci si allontana dal suo nucleo centrale.

Ergo, il metodo di ragionamento tassonomico è di tipo induttivo (gergalmente definita *funneling,* rimandando all'idea di un imbuto): si procede partendo dalle caratteristiche generale di una classe di entità, terminando con l'analisi di quelle più nel particolare.

Secondo il linguista Anderson (1991), le persone categorizzano il mondo raggruppando oggetti ed entità che hanno funzioni e/o caratteristiche simili, formando

[241] In riferimento al principio di semplificazione della modellazione statistica, descritto nel paragrafo 2.3.2 *"La metafora: deviazione e trasposizione di senso nella computazione"* del presente lavoro di tesi.
[242] http://conceptmaps.it/KM-Conceptualization-it.html
[243] Il termine "tassonomia" proviene dal greco: *τάξις, tàxis* = ordinamento; e *νόμος, nòmos* = norma.
[244]W. Croft, A. D. Cruse, *Cognitive Linguistics,* Cambridge University Press, 2004, pp. 262-263.

macrocategorie concettuali. Questo *habitus* cognitivo-comportamentale realizza quello che viene considerato come uno dei vantaggi della categorizzazione: attiva il cosiddetto processo di economia cognitiva, che conduce alla massimizzazione concettuale di un'informazione, minimizzando le risorse cognitive per richiamarla alla memoria.

Le categorie concettuali vengono definite come classi flessibili, poiché si adattano alle circostanze mutevole della vita sociale dell'individuo, ed evolvono sia in base alla sua necessità/all'utilità (ai fini della sua sopravvivenza), nonché mediante la sua curiosità intellettuale. Questo continuo stimolo cognitivo garantisce una stabile flessibilità dell'attività cerebrale. L'esperienza, l'uso e la cultura, quindi, sono veicoli – nonché costanti – conducenti a una concettualizzazione personale di una stessa realtà condivisa.

Nell'ambito dell'NLP, la categorizzazione permette la comprensione dei macro-argomenti trattati in un determinato dato testuale. Ai macro-argomenti e domini individuati (o da individuare) viene assegnato uno *score* (un punteggio) di estrazione tassonomica al fine di conferire maggiore rilievo al *topic* – oggetto della cyber investigazione.

Definire un albero tassonomico per la categorizzazione/estrazione di dati da documenti specifici o da fonti aperte selezionate non è un'operazione facilmente automatizzabile, soprattutto nel caso in cui operi in domini specifici (come, ad esempio, quello militare, *cyber* o riguardante la *warfare*).

Altrimenti, in altri contesti d'uso – come, ad esempio nell'ambito della *business Intelligence* – si può procedere alla categorizzazione di articoli, blog, o documenti trattanti *topics* di più ampia azione comunicativa, utilizzando tassonomie predefinite, riscontrabili online e facilmente integrabili nel proprio *tool* di azione linguistica.

La tassonomia (o libreria tassonomica) di tipo standard maggiormente fruita online, nonché maggiormente utilizzata per l'analisi dei contenuti testuali, è denominata *IPTC Media Topics,* acronimo di *International Press Telecommunications Council – Media Topics* (Consiglio internazionale per le telecomunicazioni della stampa), noto per redigere norme tecniche per la definizione di uno standard *business-to-business* ai fini dello scambio e della classificazione di notizie, testi, foto e altri media, condivisi dalle più importanti agenzie di informazione del mondo.

La tassonomia *IPTC Media Topics* è stata creata nel 2010 e viene aggiornata ogni anno con termini e ramificazioni nuove. È utilizzata in tutto il mondo, poiché disponibile in 11 lingue: arabo, inglese (UK), cinese mandarino, danese, francese, tedesco, norvegese, portoghese (per Portogallo e Brasile), spagnolo e svedese. L'italiano non è tra queste.

Di seguito, riporto i primi nodi della tassonomia *IPTC Media Topics*, mostrando, per motivi di spazio, solo il primo livello di ramificazione tassonomica. Tale *screenshot* è stato effettuato nella *repository* personale di COGITO STUDIO®, che implementa ed esporta tale classificazione – definita come *domaintree* – in formato *.xml.

Dalla foto si può notare che ogni nodo di partenza di tale matrice venga inteso nel file di configurazione del sistema come *domain*, ovvero come dominio concettuale.

```
<DOMAINTREE>
    <DOMAIN DESCRIPTION="Intelligence Taxonomy" NAME="INTELLIGENCE_TAX">
    + <DOMAIN DESCRIPTION="Arts, Culture and Entertainment" NAME="010">
    + <DOMAIN DESCRIPTION="Crime, Law and Justice" NAME="020">
    + <DOMAIN DESCRIPTION="Disaster and Accident" NAME="030">
    + <DOMAIN DESCRIPTION="Economy, Business and Finance" NAME="040">
    + <DOMAIN DESCRIPTION="Education" NAME="050">
    + <DOMAIN DESCRIPTION="Environmental Issue" NAME="060">
    + <DOMAIN DESCRIPTION="Health" NAME="070">
    + <DOMAIN DESCRIPTION="Human Interest" NAME="080">
    + <DOMAIN DESCRIPTION="Labour" NAME="090">
    + <DOMAIN DESCRIPTION="Lifestyle and Leisure" NAME="100">
    + <DOMAIN DESCRIPTION="Politics" NAME="110">
    + <DOMAIN DESCRIPTION="Religion and Belief" NAME="120">
    + <DOMAIN DESCRIPTION="Science and Technology" NAME="130">
    + <DOMAIN DESCRIPTION="Social Issue" NAME="140">
    + <DOMAIN DESCRIPTION="Sport" NAME="150">
    + <DOMAIN DESCRIPTION="Unrest, Conflicts and War" NAME="160">
    + <DOMAIN DESCRIPTION="Weather" NAME="170">
    + <DOMAIN DESCRIPTION="Public Companies" NAME="180">
    + <DOMAIN DESCRIPTION="Critical Infrastructure" NAME="190">
    </DOMAIN>
</DOMAINTREE>
```

*Figura 13 – Esempio di nodi tassonomici relativi alla IPTC Taxonomy (in formato *.xml)*[245]

[245] Fonte: grafico autoprodotto mediante *software* COGITO STUDIO®

È possibile scovare online persino librerie tassonomiche relative all'ambito del *cyber crime* (come raffigurato nella seguente immagine), ma al fine di garantire un'estrazione dei risultati il meno "rumorosa" e il più puntuale possibile, viene richiesta una profonda *knowledge* – nonché azione intelligente – da parte dell'operatore, al fine di creare mirate condizioni semantico-grammaticali atte al rilevamento *(mining)* dell'informazione e/o del macro-argomento di indagine linguistica.

COGITO STUDIO® di *Expert.ai* permette la modellazione dei nodi tassonomici, sia nei riguardi della gerarchizzazione dei contenuti implementati in formato *.xml, che di ramificazioni concettuali da definirsi *ex novo;* inoltre, permette di stabilire quale sia il *macro-topic* di estrazione concettuale da decretare come modello di categoria – definito anche "bontà dell'esemplare"[246] da Eleanor Rosch e Ludwig Wittgenstein – ovvero, l'elemento considerato centrale all'interno di una categoria e/o di un macro-argomento.

Come si vedrà nel caso di studio presentato nel quarto capitolo del presente lavoro, la tassonomia *cyber crime,* ivi presentata, è stata modellata sulla base delle esigenze di tale tesi di Laurea, e per congiungersi coerentemente alle categorie concettuali, nonché al linguaggio standardizzato STIX[247], pertinenti l'ambito della *Cyber Threat Intelligence* (CTI) per il rilevamento di minacce informatiche.

[246] Nel testo, p. 34 e p. 81

[247] *«Acronimo di Structured Threat Information eXpression, linguaggio standardizzato sviluppato da MITRE e dall'OASIS Cyber Threat Intelligence (CTI) Technical Committee per descrivere i dati relativi alle minacce informatiche».* https://www.anomali.com/it/resources/what-are-stix-taxii

```
<DOMAIN TREE="Cyber illegal Taxonomy" NAME="CYB_TAX">
    <DOMAIN DESCRIPTION="Cyber Attack" NAME="28001000">
        <DOMAIN DESCRIPTION="DoS attack" NAME="28001010"/>
        <DOMAIN DESCRIPTION="Intrusion (computer or network)" NAME="28001020">
            <DOMAIN DESCRIPTION="Defacing" NAME="28001021"/>
            <DOMAIN DESCRIPTION="Account compromised" NAME="28001022"/>
            <DOMAIN DESCRIPTION="Data dump/Data loss" NAME="28001023"/>
        </DOMAIN>
        <DOMAIN DESCRIPTION="Interception attack" NAME="28001030"/>
        <DOMAIN DESCRIPTION="Injection" NAME="28001040"/>
        <DOMAIN DESCRIPTION="Cross-site scripting" NAME="28001050"/>
        <DOMAIN DESCRIPTION="Cross-site request forgery" NAME="28001060"/>
        <DOMAIN DESCRIPTION="Broken authentication and session management" NAME="28001070"/>
        <DOMAIN DESCRIPTION="Man-in-the-middle" NAME="28001080"/>
    </DOMAIN>
    <DOMAIN DESCRIPTION="Cyber deception" NAME="28002000">
        <DOMAIN DESCRIPTION="Information gathering" NAME="28002010">
            <DOMAIN DESCRIPTION="Identity theft" NAME="28002011"/>
            <DOMAIN DESCRIPTION="Phishing" NAME="28002012"/>
        </DOMAIN>
        <DOMAIN DESCRIPTION="Credit card fraud" NAME="28002020">
            <DOMAIN DESCRIPTION="Skimming" NAME="28002021"/>
        </DOMAIN>
        <DOMAIN DESCRIPTION="Email spamming" NAME="28002030"/>
        <DOMAIN DESCRIPTION="Scam" NAME="28002040"/>
    </DOMAIN>
    <DOMAIN DESCRIPTION="Cyber Violence" NAME="28003000"/>
    <DOMAIN DESCRIPTION="Content-related Computer Crime" NAME="28004000">
        <DOMAIN DESCRIPTION="Cyber piracy" NAME="28004020"/>
        </DOMAIN>
    <DOMAIN DESCRIPTION="Cyber criminals" NAME="28005000"/>
    <DOMAIN DESCRIPTION="Cyber Security" NAME="28006000">
        <DOMAIN DESCRIPTION="Vulnerabilities" NAME="28006010">
            <DOMAIN DESCRIPTION="Mobile vulnerability" NAME="28006011"/>
            <DOMAIN DESCRIPTION="Application and software vulnerability" NAME="28006012"/>
            <DOMAIN DESCRIPTION="Firmware vulnerability" NAME="28006013"/>
            <DOMAIN DESCRIPTION="Zero-day" NAME="28006014"/>
        </DOMAIN>
        <DOMAIN DESCRIPTION="Threat and vectors" NAME="28006020">
            <DOMAIN DESCRIPTION="Malware and virus" NAME="28006021"/>
            <DOMAIN DESCRIPTION="Botnet" NAME="28006022"/>
            <DOMAIN DESCRIPTION="Advanced Persistent Threat" NAME="28006023"/>
            <DOMAIN DESCRIPTION="Ransomware" NAME="28006024"/>
        </DOMAIN><DOMAIN DESCRIPTION="Security software and devices" NAME="28006030"/>
    </DOMAIN>
    <DOMAIN DESCRIPTION="Cyber Terrorism" NAME="28007000"/>
    <DOMAIN DESCRIPTION="Cyber Espionage" NAME="28008000"/>
    <DOMAIN DESCRIPTION="Hacktivism" NAME="28009000"/>
    <DOMAIN DESCRIPTION="Malicious devices" NAME="28010000"/>
</DOMAIN>
```

*Figura 14 – Categorizzazione tassonomica "Cyber Crime" (in formato *.xml)*[248]

[248] Fonte: grafico autoprodotto mediante *software* COGITO STUDIO®

# 3. STRATEGIE di INTELLIGENZA LINGUISTICA

*«Intelligenza è la capacità di risolvere problemi, o di creare prodotti, che siano apprezzati all'interno di uno o più contesti culturali.* […] *L'intelligenza linguistica permette agli individui di comunicare e di costruire il significato del mondo attraverso il linguaggio».*

È così che Howard Gardner, neuropsicologo americano, definisce l'intelligenza linguistica, una delle (almeno) nove intelligenze da lui individuate e descritte nel suo libro "Frames of Mind", in cui espone la sua teoria delle intelligenze multiple.

Quella linguistica è il tipo di intelligenza maggiormente analizzata e studiata in ambito didattico e cognitivo-psicologico, nonché condivisa da tutti gli individui: si configura come capacità di acquisire/apprendere ed utilizzare un codice linguistico (una lingua materna e/o straniera), nella sua matrice scritta e/o orale, al fine di realizzare atti comunicativi attraverso segni – richiamando, quindi, dei significanti (canali e strumenti) per poter esprimere una moltitudine di significati socialmente condivisi ed accettati, in grado di creare, a loro volta, intricatissime reti semantiche.

È nell'infinità di segni (non sempre oculatamente) scelti, prodotti e dimenticati dagli individui sul web, che risiede il *target* principale della *Cyber Intelligence*: raccogliere quante più informazioni, per estrapolarne parole e concetti semantici chiave.

Questi verranno filtrati in maniera intelligente dagli analisti linguistici mediante tecnologie create sulla base di modelli cognitivi che realizzano intelligenza artificiale.

Detti modelli – come descritto nel paragrafo 2.5 *"Categorizzazione tassonomica"* – attivano processi di ragionamento di tipo *top-down*, ricercando, quindi, informazioni e dati partendo dal macro-argomento in generale, sino ad arrivare al dettaglio puntuale.

Tra le strategie applicabili, si possono progettare tassonomie *ad hoc* ad impostazione induttiva, realizzando ed implementando – nel proprio *tool* di azione linguistica – regole e condizioni linguistiche concettuali (di categorizzazione), nonché di estrazione.

Queste ultime permetteranno la cernita e il *mining* di quelle singole componenti linguistiche ricche di dettagli e di informazioni, ad esempio, sull'autore del messaggio, poiché la *parole* Saussuriana che gli esseri umani utilizzano per esprimersi, è la manifestazione più atomizzata e personalizzata della nostra identità.

Per questo, è importante che l'analista linguistico/a sappia orientarsi nelle diverse varietà dell'italiano contemporaneo/neo-standard – in continua ed incessante evoluzione – tenendo conto sia della grammatica italiana *(langue),* nonché dell'italiano dell'uso, ovvero di quella varietà linguistica realizzata nei contesti reali e quotidiani, tra cui il mondo di internet, attraverso l'espressione di sé stessi tramite blog e *social network.*

Anche il linguaggio è, infatti, una forma di dati, ed alimenta l'odierna realtà del business: ogni volta che digitiamo qualcosa online, lasciamo una traccia delle nostre scelte e preferenze. Commenti e *post* pubblicati sul web, ad oggi, rappresentano il principale strumento di interazione sociale tra individui e si configurano come impronte digitali del loro *modus cogitandi et loquendi*.

A designati *data scientists* viene richiesto di possedere competenze, conoscenze ed abilità di *critical thinking «per identificare dati strategici e relazioni più rilevanti tra le informazioni a loro disposizione, al fine di poterle impiegare a supporto dei processi decisionali».*[249] Per tale motivo, vi è la necessità di sviluppare – o di raffinare in maniera costante – un'intelligenza linguistica sempre più in grado di processare ed interpretare le informazioni mediante la riproduzione della (ri)cognizione umana: l'operatore di Intelligence linguistica dovrà aver cura di riuscire a leggere e interpretare le strutture dinamiche della comunicazione – proprie di ogni individuo – a livello metalinguistico e metacognitivo, rievocando costantemente la propria *knowledge* in materia socio-culturale e geopolitico, al fine di realizzare un'azione di *disguise* a carattere psico-linguistico per riuscire a ragionare ed esprimersi come farebbe l'autore delle minacce alla (cyber)sicurezza nazionale ed internazionale – obiettivo *target* dell'indagine linguistica in ambito computazionale.

[249] *Decision Making,* 2022. Cfr. https://www.stateofmind.it/tag/decision-making/

## *3.1 Potenzialità del motore semantico*

Un motore semantico contiene e processa algoritmi in grado, non solo di dedurre il macro-argomento trattato nel dato testuale, ma anche la prossimità semantica – sfruttando le reti semantiche che vi sono alla base – degli elementi linguistici a un determinato macro-concetto, seguendo un processo di comprensione ed interpretazione associativa.

Si sente sempre più frequentemente parlare di semantica in ambito informatico-computazionale, e di quanto essa incida sull'utilizzo della piattaforma *internet*.

Il concetto stesso di semantica non è, però, di immediata (ed univoca) comprensione; si riscontrano, infatti, due chiavi distinte di lettura:

1. "semantica" può riferirsi alla scienza dei significati, dedicata all'analisi dei segni linguistici d'uso comune del linguaggio naturale proprio di ogni individuo; scienza che, in computazione, viene realizzata e resa possibile mediante *frameworks* o motori semantici, atti all'individuazione dei rapporti sintattici e lessicali tra tutti i costituenti di una frase (o di una porzione di testo appositamente specificata) ai fini della concettualizzazione di macro-argomenti e/o dell'estrazione di parole chiave;
2. altresì, "semantica" può indicare il *semantic web*, ovvero – nella visione dell'informatico Sir Timothy John Berners-Lee – inventore, assieme a Robert Cailliau, del *World Wide Web*, spazio informativo globale – di un *«web of data that can be processed by machines»*[250], ovvero una realtà web – che sfrutta algoritmi di *information retrieval* – in cui i *link* di collegamento a parole chiave o ad interi documenti presentano un significato dicotomico/binario (0-1) processabile da un *software*[251].

Queste due interpretazioni sono (mediante le odierne tecnologie disponibili) comunicanti tra di loro (il *semantic web* non esisterebbe senza, appunto, la componente computazionale di processazione semantica che vi è alla base) ma "motore semantico" e "motore di ricerca semantica" sono da considerarsi come due realizzazioni dell'analisi del dato differenti: nel primo caso, infatti, ci si riferisce al contesto di realizzazione del *Natural Language Processing* (NLP), ovvero dell'elaborazione del linguaggio naturale; mentre, nel

[250] T. Petkova, *A Web of People and Machines: W3C Semantic Web Standards*, 2014. Cfr. www.ontotext.com/blog/a-web-of-people-and-machines-w3c-semantic-web-standards/
[251] In riferimento all'articolo di A. Bolioli, *Motore di Ricerca Semantico: che cos'è e a cosa serve?*, 2016. Cfr. https://www.celi.it/blog/2016/04/motore-di-ricerca-semantico/

secondo caso, si rimanda al *Semantic Web* e ai *Linked Data* – afferenti ai più comuni motori di ricerca online[252].

Poiché la semantica della mente viene intesa come non completamente accessibile – essendo legata ad una interpretazione umana del mondo altamente soggettiva ed intima, nonché non rintracciabile meramente nei singoli significati di ogni parola – la semantica computazionale – che si realizza a partire dalla emulazione del contenuto (anche in termini di strutture neurologiche) della mente dell'essere umano da parte di una macchina, mediante l'applicazione di un linguaggio formale a carattere binario – per poter accedere alla creazione della conoscenza (ma non dell'autocoscienza) deve vedersi garantiti quattro elementi principali:

a) i dati: sono il punto di partenza per minimizzare la distorsione analitica ed informativa del linguaggio. Si devono basare su fatti osservati e puntuali, e devono essere in prima istanza filtrati e processati dell'operatore di *Intelligence* o da un *layer* di analisi linguistica; solo dopo potranno esser forniti come *input* al *software*. Ad esempio, volendo presentare al motore semantico una struttura sintattica non complessa, come una frase dichiarativa composta da [soggetto + verbo + oggetto], il *framework* linguistico, mediante questo *input*, sarà in grado sia di associare ogni sintagma ad ogni funzione, e sia di determinare la relazione lessicale-sintattica tra loro instaurata;
b) l'interpretazione dei dati: mediante una descrizione formale del dato processato, il motore sarà in grado di fornire una categorizzazione logico-concettuale dei domini correlati alle entità;
c) i sistemi per la gestione dell'informazione: consente di attribuire un significato univoco e condiviso al dato informativo inviato in *input*, nonché di reperire le informazioni in un dato testuale in maniera veloce ed efficiente, nonché di applicare delle condizioni linguistiche deputate al discernimento o alla ricognizione di dettagli legati (o meno) a un'informazione;
d) i sistemi di elaborazione della conoscenza/*knowledge*: per poter generare *conoscenza*, è necessario che nel *framework* semantico sia presente una componente a carattere cognitivo, la cosiddetta *Knowledge Base*[253] ("base di conoscenza"), in grado di organizzare ed elaborare l'insieme di formazioni rilevanti memorizzate dal sistema, al

[252] Di cui, il più famoso è senza dubbio il motore di ricerca *Google.* Altri esempi di motori di ricerca: *Bing, Yahoo, AltaVista, Lycos, Virgilio*.
[253] Nel testo, p. 55.

fine di diffondere una comprensione ed una soluzione informativa relativamente al problema di partenza[254].

I modelli di *Machine Learning* vengono addestrati per analizzare il linguaggio naturale ma non possiedono il quarto elemento sopra indicato: la conoscenza. I dati, da soli, non sono sufficienti per vedersi garantito un processo di analisi linguistica esaustivo; è necessario, invece, che la macchina sia in grado di leggerli in maniera funzionale, comprendendo i significati di determinate entità o macro-contesti, proponendo chiavi di lettura differenti.

È per questo motivo che l'elemento fondamentale per poter indirizzare nella maniera più corretta un *software* nell'elaborazione e nell'interpretazione del linguaggio naturale è il *Knowledge Graph* – come già visto in precedenza in questo lavoro di tesi[255] – ovvero, una "mappa cognitiva", che permette di carpire le (innumerevoli) sfumature di una lingua, semplificando la gestione della conoscenza non strutturata, decodificando significati, associando questi ultimi a domini e contesti d'uso, nonché individuando una polarità contestuale complessiva (positiva, negativa o neutra) del documento processato.

A un sistema dotato di capacità di apprendimento, quindi, non sarà sufficiente la sola ingestione e *processing* non supervisionato di dati, bensì, necessiterà di dette mappe cognitive per poter riuscire ad individuare concetti e risolvere espressioni problematiche.

Il processo di *cyber analysis* è un processo di risoluzione di problemi, di cui non ci si può permettere la reiterazione. Per questo, al meccanismo di autoapprendimento, si affianca spesso la presenza dei tecnici umani, conducenti indispensabili al timone del motore semantico, in grado di risolvere ambiguità del *software*, integrandone la conoscenza.

Attraverso l'implementazione e l'utilizzo di mappe cognitive-concettuali, quindi, la ricerca semantica tenta di avvicinarsi progressivamente al meccanismo di apprendimento umano e di analisi del testo, ambendo ad un approccio sempre più "artificialmente intelligente" nell'elaborazione linguistica.

[254] Si parla anche di *knowledge management* ("conoscenza organizzativa"), intesa come *«l'insieme di strategie e metodi per identificare, raccogliere, sviluppare, conservare e rendere accessibile la conoscenza* […], *avvalendosi in genere di strumenti delle tecnologie dell'informazione»*. – E. Bonati, *La conoscenza è nulla se non è organizzata,* 2021. Cfr. www.forme.online/2021/06/03/la-conoscenza-e-nulla-se-non-e-organizzata/

[255] Il *Knowledge Graph* – denominato *Sensigrafo*® in COGITO STUDIO® – interpreta l'*input* testuale identificando contesti di appartenenza e risolvendo conflitti di significato.

## *3.2 Cognitive Computing: COGITO STUDIO® by Expert.ai*

Premessa:

*Il presente capitolo, interamente dedicato al motore semantico COGITO STUDIO® realizzato dalla società Expert.ai, nonché ogni menzione effettuata nel presente lavoro di tesi,* ***NON sono stati realizzati a fini promozionali*** *– né nei riguardi dello strumento, né della sua casa madre – né tanto meno si tratta di un'attività realizzata a scopo di lucro. Ogni menzione e la testimonianza di utilizzo di tale strumento si configurano unicamente come* ***libera attività di ricerca universitaria****. Ritengo, inoltre, che strumenti diversi da COGITO STUDIO® abbiano le dovute potenzialità per realizzare l'analisi linguistica presentata nel seguente caso di studio; infatti, il mio percorso di analisi e ricerca personale in ambito linguistico-/computazionale si realizza e si amplia anche attraverso tool differenti, con alla base tecnologie altrettanto innovative.*

***

COGITO STUDIO® è un sistema *software* realizzato dall'azienda *Expert.ai* per l'analisi linguistica di testi e contenuti di varia natura e struttura, sfruttando la tecnologia del *Cognitive Computing*, con alla base algoritmi di intelligenza artificiale e la capacità di elaborazione dei segnali *(Signal Processing Intelligence).*

Sono molteplici i modelli operati dal *Cognitive Computing*, il quale include, ad esempio, algoritmi di apprendimento automatico *(Machine Learning)*, di elaborazione del linguaggio naturale (NLP) e dello *Speech Recognition and Vision (Object Recognition).*

Intelligenza artificiale e *Machine Learning*, inoltre, hanno un obiettivo comune: replicare le funzioni e le attività cognitive del cervello umano[256].

La prima tecnologia si configura come un'architettura computazionale composta da reti neurali artificiali (ANNs), ispirata ad una semplificazione delle reti neurali biologiche; mentre, la seconda tecnologia prevede algoritmi per la risoluzione di problemi, che consentono, quindi, ai sistemi di apprendere e di auto-migliorarsi in fase di restituzione *(output)* dei dati processati.

Il *Cognitive Computing* si inscrive tra l'A.I. ed il *Machine Learning:* si tratta di un sistema altamente avanzato che attiva funzionalità cerebrali nelle macchine, con l'obiettivo di far risultare più "naturale" l'interazione essere umano-computer; il successo di tale attività si deve ad una comprensione ed elaborazione più profonda dei dati forniti in *input*.

[256] Cfr. https://tecnologia.libero.it/che-cose-il-cognitive-computing-14479

Il modello di apprendimento automatico che compone parte di questa tecnologia, infatti, permette agli algoritmi di *Cognitive Computing* di imparare una lingua allo stesso modo dei bambini (al di sotto degli 8 anni[257]) nella loro prima fase di apprendimento (ma non di acquisizione) linguistica.

Per le "macchine cognitive" comprendere una sequenza di parole[258] è obiettivo minimale: esse, infatti, possono contare su un livello di conoscenza (una *Knowledge Base*) del linguaggio umano tale da riuscire a capire il significato profondo delle informazioni.

Le migliori piattaforme con alla base il *Cognitive Computing*, quindi, sono in grado di realizzare una scelta "consapevole" del significato più corretto da conferire alle parole che sentono o leggono, riconoscendo il loro contesto di riferimento, nonché distinguendo volti – a partire da una o più immagini che ritraggono un umano – e voci delle persone[259].

Così come la conoscenza umana può essere migliorata imparando cose nuove, anche la *knowledge* di COGITO STUDIO® può essere raffinata ed ampliata attraverso l'acquisizione di nuove informazioni. Un dato degno di nota nell'ambito dell'informatica, è che questo sistema non è una *black box*[260], bensì un sistema aperto, dotato di contenuto e struttura comprensibili all'essere umano, nonché quindi facilmente validabili ed adattabili per ottimizzare ed automatizzare le molteplici prestazioni richieste nei vari settori.

I modelli di *Cognitive Computing* restituiscono risultati esemplari nell'elaborazione di grandi quantità di dati, e nell'analisi di informazioni aggregate, in formato non omogeneo (strutturato e/o non strutturato), che, contrariamente, risulterebbero difficilmente assimilabili dalle applicazioni informatiche tradizionali.[261]

Un sistema di computazione cognitiva si può definire come:

- adattivo: il modello alla base è in grado di riconoscere quando le informazioni fornite in *input* subiscono variazioni (è in grado di apprendere sulla base di quelle precedentemente

[257] Ovviamente, il processo di lateralizzazione della dominanza emisferica non viene realizzato/trasferito nei sistemi di intelligenza artificiale.
[258] In riferimento ai motori semantici in grado di riconoscere ed elaborare codici linguistici a sistema alfabetico, non a sistema sillabico e/o di scrittura continua.
[259] Circa la ricognizione delle voci si richiede una tecnologia all'avanguardia per l'identificazione e/o la verifica dell'identità di una persona, essendo ogni voce considerata tanto unica, quanto le impronte digitali.
[260] Un modello black box è un sistema che *«non consente una chiara interpretazione del risultato ottenuto»* – AA. VV., *L'intelligenza artificiale per lo sviluppo sostenibile,* Università degli Studi di Bari "Aldo Moro", 2014, p. 75.
[261] Cfr. https://www.focus.it/tecnologia/innovazione/che-cos-e-il-cognitive-computing

fornite) e quando i requisiti imposti dall'operatore evolvono. Il sistema sa risolvere ambiguità linguistiche ed alfanumeriche, e rimane flessibile all'imprevedibilità;

- interattivo: il sistema è in grado di interagire con gli utenti umani, fornendo *feedbacks* e restituendo risultati attesi. Se configurato opportunamente, può anche interagire con altri processori, dispositivi e servizi *cloud* – nonché, con le persone;
- iterativo e con stato: il sistema può aiutare a definire la natura di un problema *«ponendo domande o trovando ulteriori fonti di input se un'affermazione del problema è ambigua o incompleta»*[262]*;* ricorda le eventuali interazioni avvenute in un processo precedente, e restituisce le informazioni atte alla risoluzione del problema incontrato;
- contestuale: il sistema può *«comprendere, identificare ed estrarre elementi contestuali come: significato, sintassi, ora, posizione, dominio appropriato, normative, profilo dell'utente, processo, attività e obiettivo».*[263] Può attingere a varie fonti informative, tra cui dati testuali e numerici strutturati e non strutturati – realizzando l'attività di *Text Analytics* – nonché *input* a carattere sensoriale.

Le innovazioni introdotte con COGITO STUDIO® presentano le caratteristiche e le strategie di computazione linguistica sopra delineate, e si configura come piattaforma *rule-based* ibrida di intelligenza artificiale che permette a *data scientists* di personalizzare la *Knowledge Base* contenuta al suo interno *«in ogni fase del processo: dall'acquisizione e dalla pulizia dei dati, fino alla messa a punto del modello linguistico»*[264] ovvero, fino all'implementazione del pacchetto linguistico *renderizzato* su *software* di Intelligence.

Il funzionamento di tale processore linguistico è basato sul processo di disambiguazione che permette di realizzare tre tipologie di analisi: morfologico-lessicale, sintattica *(parsing)* e semantica.

Un'altra innovazione del sistema è la definizione manuale di una serie infinita di regole (o "condizioni linguistiche") che consentono di analizzare ed elaborare il testo fornito realizzando sia un processo di categorizzazione tassonomica delle informazioni, che di estrazione (ad esempio, di *keywords*, entità standard e *patterns),* per un'individuazione efficace e puntuale del *target* dell'investigazione linguistica.

[262] *Dizionario Informatico*, 2021. Cfr. www.dizionarioinformatico.blogspot.com/2021/07/definizione-di-cognitive-computing-cc.html
[263] Ibidem.
[264] https://www.expert.ai/it/the-platform/nl-ops/

All'interno della piattaforma *rule-based*, le condizioni linguistiche possono essere redatte ed implementate dall'operatore in tre linguaggi differenti, ognuno deputato alla realizzazione di una data funzionalità ed ognuno proprietario di *Expert.ai*[265]*:*

- linguaggio "C"[266] (file *.efo): per la realizzazione del processo di categorizzazione. È possibile procedere alla compilazione di condizioni linguistiche atte alla ricerca di categorie e concetti all'interno del documento[267], solo a seguito dell'implementazione (o creazione *ex novo)* dell'albero tassonomico di progetto (file *.xml). Il linguaggio "C", inoltre, permette la trasformazione del documento in blocchi logici, delimitando la ricerca dell'informazione nel: titolo, paragrafo, frase, corpo del testo, o in altro tipo di segmentazione (in tal caso, *customizzata)*;
- linguaggio "D" (file *.efd): effettua una manipolazione sui domini presenti all'interno del dizionario del sistema (il *Sensigrafo*®) o su un *set* di domini editati dall'operatore;
- linguaggio "E" (file *.efe): per la realizzazione del processo di estrazione delle informazioni contenute nei documenti; sono molti gli attributi, le funzionalità e le condizioni linguistiche compilabili per garantire un *mining* puntuale ed immediato del dato da ricercare.

Oltre all'operazione di categorizzazione e di estrazione delle informazioni da documenti, tale *framework* semantico fornisce ulteriori funzionalità, come:

- l'analisi dei sentimenti (*Sentiment Analysis):* è possibile analizzare il sentimento espresso in un documento individuando polarità semantiche di parole e/o domini;
- il processo di normalizzazione[268]: mediante la creazione di file *.txt implementabile a sistema, il testo (o il *corpus* di dati testuali redatto sulla base della propria *knowledge*) viene normalizzato dall'operatore in modo da stabilire una variante linguistica di riferimento per una data entità [ad esempio: (Isaac Wiper = ISAAC WIPER|isaac wiper|IsaacWiper)]; altresì, in modo da eliminare segni di interpunzione e/o spazi superflui o ripetuti, o anche trasformando i caratteri maiuscoli in minuscoli (e viceversa) in caso di gestione di liste di dati di natura *case sensitive.*

[265] Ad eccezione dell'implementazione di RegEx PERL standard, mediante la compilazione di condizioni linguistiche contenenti l'attributo PATTERN nel linguaggio "E" di estrazione.
[266] Non vi è alcuna connessione tra il linguaggio "C" di COGITO STUDIO® e il linguaggio "C/C++" di programmazione.
[267] L'operatore dovrà assegnare dei punteggi *(scores)* di rilevanza ai domini concettuali da individuare.
[268] Si rimanda al paragrafo 4.4.1 *"Regole di normalizzazione"*.

## 3.3 *Il Sensigrafo®: reti e nodi della conoscenza*

Il *Sensigrafo®* (o "grafo dei sensi") è una *core feature* presente nel sistema di analisi semantica di COGITO STUDIO®: non può essere definito come un dizionario, poiché si configura come un insieme di reti semantiche interconnesse tra loro tramite una struttura a grafo in cui ogni nodo – come già mostrato in precedenza – rappresenta un concetto.

In tale processore linguistico, i lemmi non sono disposti in ordine alfabetico, a differenza di un qualsiasi altro dizionario, bensì i concetti espressi all'interno del *Sensigrafo®* sono rappresentati in maniera univoca da un numero identificativo (ID) definito come *syncon*[269], ed ognuno di essi è associato a uno o più domini, che ne determinano la significazione. Ciò è di essenziale rilevanza in caso di polisemia.

Il *Knowledge Graph* (o "grafo della conoscenza"[270]) permette, quindi, di risolvere casi di ambiguità lessicale, grammaticale, semantica e sintattica, attraverso una rappresentazione "intelligentemente artificiale" della conoscenza del mondo reale, in cui ogni concetto/dominio viene definito ed interconnesso con altri *types/frames*.

Ciò si realizza mediante la ricognizione e, successivamente, l'instaurazione di relazioni semantiche, rese possibili da complessi algoritmi di *Deep Learning*, con protagoniste le reti neurali artificiali. Detti algoritmi permettono l'accesso alla comprensione e all'elaborazione del linguaggio naturale[271]. Ciò non è reso possibile nei sistemi puri di *Machine Learning*.

Le reti presenti nel grafo della conoscenza di COGITO STUDIO® si concretizzano, da parte di *Expert.ai*, grazie all'adozione di un approccio (meta)cognitivo del linguaggio.: Il *Sensigrafo®*, infatti, è stato configurato come una struttura aperta, modificabile dall'analista linguistico senza alcuna limitazione, al fine di garantire all'operatore una stabile ed immediata comprensibilità e comprensione dei suoi contenuti linguistici, nonché la possibilità di modellare ed ampliare entità e relazioni tra *syncon* sulla base della propria conoscenza, umanamente in continua evoluzione.

[269] Definito *synset* nella maggior parte dei programmi di analisi linguistica.
[270] Come già definito nel paragrafo 2.3.1 *"Il processo di disambiguazione"*.
[271] Per un approfondimento: E. Santagata, A. Melegari, *Gli USA sognano un computer capace di ragionare come un bambino*, per *Analisi* Difesa, 2019. Cfr. https://www.analisidifesa.it/2019/04/gli-usa-sognano-un-computer-capace-di-ragionare-come-un-bambino/

Tramite il *Sensigrafo*®, quindi, si può apprendere circa: l'interconnessione tra concetti, le specifiche riguardanti l'appartenenza o meno di un certo elemento ad un certo dominio lessicale, nonché informazioni sulla frequenza di utilizzo di un certo concetto.

La frequenza di utilizzo di una determinata entità (ereditando in essa uno o più concetti), è strettamente collegata al concetto di radicamento linguistico[272] che, a sua volta, trova riscontro ed espressione nel fenomeno della grammaticalizzazione di una forma e/o variante linguistica, nonché nella reificazione[273] dei domini a cui essa è collegata.

Inoltre, ogni elemento presente nella conoscenza del motore semantico contiene un insieme di attributi – rappresentanti il ruolo grammaticale, la relazione semantica, la definizione di significati, domini e frequenza di ogni *syncon* – che stabiliscono le caratteristiche di lemmi e concetti, e attraverso cui è possibile effettuare un'immediata analisi del contenuto testuale.

In questa rappresentazione della conoscenza, ogni *syncon* (o nodo) è collegato ad altri *syncon* mediante una gerarchizzazione delle relazioni semantiche, basata su una struttura meronimica. In questo modo, ogni *syncon* viene "arricchito" (in inglese, *enriched)* da tutte le *features* e significati dei nodi (o *links)* ad essi correlati.

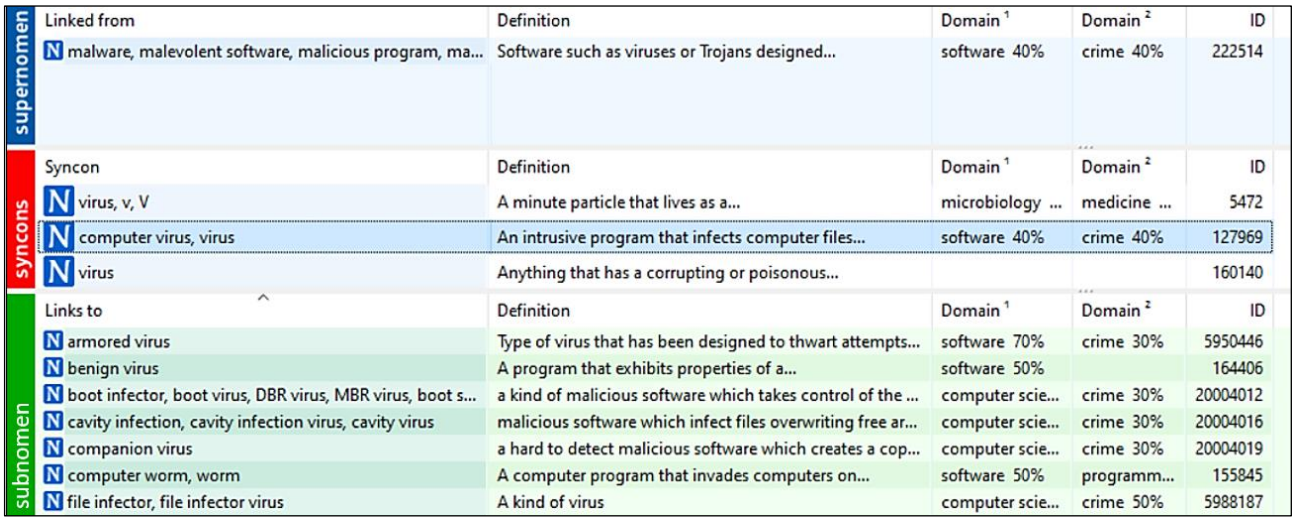
Presentando l'esempio della parola "virus" che, nel *Sensigrafo*® implementato a sistema, si dirama nei contesti della medicina e dell'informatica, vediamo che la struttura meronimica di ogni lemma viene rappresentata attraverso un listato di *links* collocati mediante condizione di iperonimia (nella terminologia di COGITO STUDIO®: *supernomen – syncon linked from)* e iponimia (nella terminologia di COGITO STUDIO®: *subnomen – syncon links to).*

È possibile cliccare su ogni entità al fine di visualizzare e conoscere i collegamenti più profondi da cui hanno origine e/o che originano i *syncon.*

[272] «*La grammaticalizzazione non è un fenomeno interno alla lingua,* [bensì è] *il risultato finale del consenso linguistico,* […] *anch'esso,* [risultato] *di un processo essenzialmente sociolinguistico. Dire che un'innovazione viene grammaticalizzata, equivale a dire che essa è stata accettata dalla comunità linguistica come norma».* – M. Alinei, *Linguistica storica e reificazione del linguaggio*, Università di Utrecht, 2011, p. 204.
[273] Con "reificazione" s'intende un *«processo mentale per cui si converte in qualcosa di concreto, o si viene a considerare tale, ciò che ha soltanto esistenza astratta».* Cfr. Treccani: https://www.treccani.it/vocabolario/reificazione/

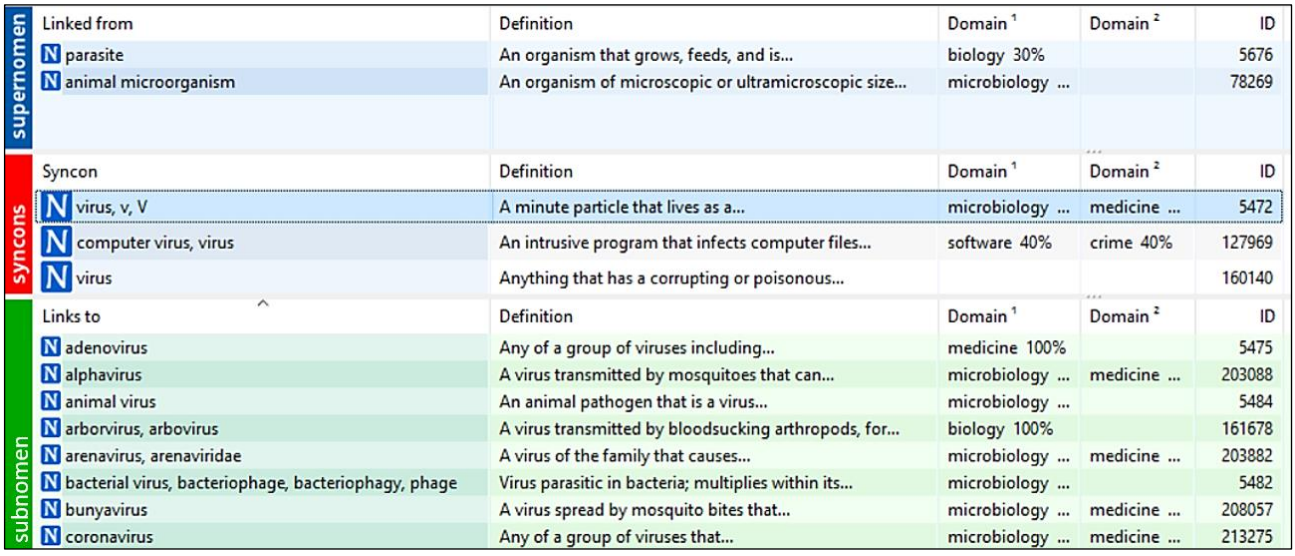
Ad esempio, partendo dal lemma “virus” annesso all’ambito informatico, si clicchi sul suo *supernomen* “*software* malevolo” – come evidenziato nella figura successiva.

Il sistema mostrerà i collegamenti profondi del lemma individuato, generando, anche per esso, *supernomen* da cui ha origine, e *subnomen* che originerà.



| | Linked from | Definition | Domain 1 | Domain 2 | ID |
|---|---|---|---|---|---|
| supernomen | N malware, malevolent software, malicious program, ma... | Software such as viruses or Trojans designed... | software 40% | crime 40% | 222514 |
| | **Syncon** | **Definition** | **Domain 1** | **Domain 2** | **ID** |
| syncons | N virus, v, V | A minute particle that lives as a... | microbiology ... | medicine ... | 5472 |
| | N computer virus, virus | An intrusive program that infects computer files... | software 40% | crime 40% | 127969 |
| | N virus | Anything that has a corrupting or poisonous... | | | 160140 |
| | **Links to** | **Definition** | **Domain 1** | **Domain 2** | **ID** |
| subnomen | N armored virus | Type of virus that has been designed to thwart attempts... | software 70% | crime 30% | 5950446 |
| | N benign virus | A program that exhibits properties of a... | software 50% | | 164406 |
| | N boot infector, boot virus, DBR virus, MBR virus, boot s... | a kind of malicious software which takes control of the ... | computer scie... | crime 30% | 20004012 |
| | N cavity infection, cavity infection virus, cavity virus | malicious software which infect files overwriting free ar... | computer scie... | crime 30% | 20004016 |
| | N companion virus | a hard to detect malicious software which creates a cop... | computer scie... | crime 30% | 20004019 |
| | N computer worm, worm | A computer program that invades computers on... | software 50% | programm... | 155845 |
| | N file infector, file infector virus | A kind of virus | computer scie... | crime 50% | 5988187 |

***Figura 15a – Esempio di gerarchia semantica del syncon relativo al lemma “virus” > dominio computer science***[274]



| | Linked from | Definition | Domain 1 | Domain 2 | ID |
|---|---|---|---|---|---|
| supernomen | N parasite | An organism that grows, feeds, and is... | biology 30% | | 5676 |
| | N animal microorganism | An organism of microscopic or ultramicroscopic size... | microbiology ... | | 78269 |
| | **Syncon** | **Definition** | **Domain 1** | **Domain 2** | **ID** |
| syncons | N virus, v, V | A minute particle that lives as a... | microbiology ... | medicine ... | 5472 |
| | N computer virus, virus | An intrusive program that infects computer files... | software 40% | crime 40% | 127969 |
| | N virus | Anything that has a corrupting or poisonous... | | | 160140 |
| | **Links to** | **Definition** | **Domain 1** | **Domain 2** | **ID** |
| subnomen | N adenovirus | Any of a group of viruses including... | medicine 100% | | 5475 |
| | N alphavirus | A virus transmitted by mosquitoes that can... | microbiology ... | medicine ... | 203088 |
| | N animal virus | An animal pathogen that is a virus... | microbiology ... | | 5484 |
| | N arborvirus, arbovirus | A virus transmitted by bloodsucking arthropods, for... | biology 100% | | 161678 |
| | N arenavirus, arenaviridae | A virus of the family that causes... | microbiology ... | medicine ... | 203882 |
| | N bacterial virus, bacteriophage, bacteriophagy, phage | Virus parasitic in bacteria; multiplies within its... | microbiology ... | | 5482 |
| | N bunyavirus | A virus spread by mosquito bites that... | microbiology ... | medicine ... | 208057 |
| | N coronavirus | Any of a group of viruses that... | microbiology ... | medicine ... | 213275 |

***Figura 15b – Esempio di gerarchia semantica del syncon relativo al lemma “virus” > dominio microbiology***[275]

[274] Fonte: grafico autoprodotto mediante *feature Sensigrafo*® del *software* COGITO STUDIO®
[275] Ibidem.

## *3.4 Il processo di analisi linguistica in COGITO STUDIO®*

L'investigazione linguistica si configura come analisi scientifica di un dato testuale. Essa può interessare uno o più livelli linguistici – fonologico, morfologico, sintattico, semantico e pragmatico – come delineato di seguito:

- livello fonologico: si riferisce, essenzialmente, allo studio dei suoni di una lingua; l'analisi fonologica di una lingua si fonda su regole basate sui principi della commutazione tra i foni, della loro distribuzione e del loro raggruppamento in "classi naturali", definite da somiglianze fonetiche;
- livello morfologico: atto allo studio della struttura interna delle parole di una lingua; l'analisi fonologica individua le unità minime di significato, i morfemi;
- livello sintattico: riguarda lo studio della struttura di una frase. L'analisi sintattica tenta di definire e descrivere le regole utilizzate dai parlanti di una lingua per combinare funzionalmente le parole tra loro per realizzare frasi significative;
- livello semantico: come visto in precedenza in questo lavoro di tesi, la semantica si occupa dello studio dei significati e dei loro confini concettuali; condurre un'analisi semantica significa assegnare un significato ad una struttura sintattica e, di conseguenza, all'espressione linguistica. Il processo di computazione automatica in grado di realizzare l'associazione tra parole e corrispettivi sensi è definito "disambiguazione";
- livello pragmatico: atto a descrivere l'uso e gli aspetti sociali del campione linguistico da analizzare. L'analisi pragmatica individua il rapporto fra i segni ed i loro utenti, analizzando l'uso[276] che i parlanti fanno di determinati segni in un dato contesto.

L'analisi linguistica può essere utilizzata per descrivere le regole ed i processi inconsci che i parlanti di una lingua utilizzano per realizzare atti comunicativi. Il modo di tale comunità linguistica di utilizzare la *langue/competence*[277] , la quale prende forma nella *parole/performance*[278], fornisce informazioni – seppur confutabili – circa il *background*

[276] Si rimanda alla teoria dei giochi linguistici di Ludwig Wittgenstein, elaborata nel suo lavoro "Ricerche Filosofiche" del 1953, in cui espresse che il significato delle parole dipenda dall'uso che se ne fa.
[277] Definita nel 1957 da Avram Noam Chomsky come *«la conoscenza astratta del proprio linguaggio»*, ovvero il saper padroneggiare quell'insieme di regole e/o principi linguistici che permettono ad un parlante nativo o altamente competente di determinare se una data espressione rispetti la norma, o meno.
[278] Per *performance* linguistica si intende, invece, l'effettivo uso della conoscenza e padronanza della grammatica e della norma; è l'espressione personale e unica della lingua, ed è diretto riflesso della *competence*.

geografico, quindi linguistico-culturale, di una persona e, conseguentemente, circa un potenziale schieramento politico e/o appartenenza ad un Credo religioso.

Infatti, alcune Agenzie Governative elaborano tale processo per poter accogliere, oppure negare, richieste di asilo politico o di cittadinanza ricevute; altresì, molte agenzie private di Intelligence forniscono il servizio di analisi (pluri-)linguistica per la ricerca di informazioni, e/o di strategie di comunicazione complesse, risultanti negli strumenti che si possono sfruttare – in modo integrato oppure selettivo – per la trasmissione e decodifica di un determinato messaggio.

Il problema principale nell'avviare un processo di elaborazione del linguaggio naturale tramite un *tool* di analisi linguistica è la disambiguazione: ovvero, il discernimento del significato più idoneo per ogni singolo *token*.

Con COGITO STUDIO® si realizza la disambiguazione di un dato testuale mediante l'analisi di quattro livelli (su cinque) di elaborazione linguistica precedentemente descritti:

1. analisi lessicale: in questa fase si scompone l'espressione linguistica in singoli *tokens*;

2. analisi morfologica: operazione che associa ciascun *token* ad una parte del discorso (sostantivi, nomi propri, verbi, aggettivi, articoli, ecc.), nonché ad un ruolo all'interno della frase; inoltre, vengono fornite informazioni sulla forma base dei lemmi. Già in questa fase emergono le prime difficoltà legate all'ambiguità lessicale. Per superare il problema, i sistemi NLP usano varie strategie: ad esempio in presenza di una parola ambigua si procede all'analisi delle categorie grammaticali delle parole ad essa prossime, realizzando, così, un'analisi morfo-sintattica;

3. analisi sintattica (processo di segmentazione o *parsing):* si strutturano i *tokens* ottenuti mediante l'analisi lessicale in una struttura ad albero *(parse tree)*. È il cuore dei processi NLP/NLU: un *parser* è in grado di riconoscere le principali relazioni grammaticali tra gli elementi della frase; ne identifica, ad esempio: il soggetto, i complementi, i nuclei nominali complessi. Questa fase rappresenta la componente fondamentale per poter procedere a degli *step* successivi, come l'annotazione *(labeling)* e/o l'estrazione semantica *(mining/extraction);*

4. analisi semantica: ad ogni elemento della struttura sintattica viene associato un significato, a sua volta collegato ad una rete di concetti; le relazioni tra i concetti ed i *syncon (o synset)* consentono di realizzare una struttura reticolare definita "ontologia"[279]. Il sistema realizza l'analisi semantica sfruttando l'interazione tra il motore semantico ed il *Sensigrafo*®. Il disambiguatore filtra l'elenco delle opzioni possibili per ogni *token*, considerando il contesto in cui ognuno di essi viene rilevato, al fine di definire il significato corretto.

[279] Tra le relazioni più comuni ci sono quelle di appartenenza a una categoria «IS-A» o ad un oggetto «PART-OF», come definito nel paragrafo 2.3 *"Reti semantiche, interpretazione e confini del significato"*.

### *3.4.1 Attributi ed operatori booleani, logici e di sequenza*

Le regole atte all'estrazione e alla categorizzazione dell'informazione linguistica da un determinato documento *«descrivono dei prototipi di sequenza di testo* [o stringhe] *che devono coincidere con il testo che si vuole verificare».*[280]

Per poter rilevare informazioni all'interno di una porzione di testo – sulla base di un documento fornito in *input* al sistema – l'analista linguistico/a dovrà creare ed implementare sul motore semantico puntuali condizioni formali.

Come già descritto in precedenza, COGITO STUDIO® presenta due linguaggi proprietari: il linguaggio "C", per la realizzazione del processo di categorizzazione, ed il linguaggio "E", per il processo di estrazione dell'informazione.

Entrambi i linguaggi prevedono la fruizione, all'interno delle espressioni linguistiche, di attributi, operatori booleani, operatori di sequenza ed operatori logici. Tali operatori permettono di stabilire un collegamento e/o una relazione di dipendenza tra più attributi espressi all'interno di una regola.

Gli attributi che permettono l'individuazione di un dato linguistico *target* all'interno di un testo, nonché di verificare – in fase di validazione dell'*output* di *mining* – il *matching* tra il codice realizzato e la porzione di testo indicata nella condizione, sono i seguenti[281]:

[280] D. Bedogni, *Progettazione e Sviluppo di un Sistema di Risposta Automatico per la Richiesta di Informazioni Riguardanti i Servizi Ferroviari,* Università degli Studi di Modena e Reggio Emilia, 2014, p. 16
[281] In riferimento alla guida utente del *tool* reperibile al seguente url: https://docs.expertsystem.com/

| ATTRIBUTO | DESCRIZIONE |
|---|---|
| SYNCON | Identifica come *token* una parola presente all'interno del *Sensigrafo*®, il cui significato le è stato assegnato dal disambiguatore semantico; è associato ad una rete di domini, nonché collegato concettualmente ai suoi iperonimi e iponimi. Ogni *syncon* è associato ad un numero identificativo. Si possono creare *syncons ex novo.* |
| LEMMA | Identifica come *token* una parola presente nel *Sensigrafo*® espressa nella sua forma base. Indicando, ad esempio, il LEMMA *ransomware*, il sistema considererà tutte le entrate *ransomware* presenti all'interno della *Knowledge Base*, indipendentemente dalla categoria grammaticale e/o appartenenza a domini. |
| KEYWORD | Identifica come *token* una sequenza precisa di caratteri espressa dall'operatore. Tale sequenza di caratteri non viene in alcun modo ricondotta al *Sensigrafo*® del sistema, e non rileva differenze tra caratteri numerici e quelli linguistici. L'attributo KEYWORD è *case sensitive.* |
| ANCESTOR | Tale attributo si basa sulla gerarchizzazione meronimica tra parole, caratterizzante il "grafo dei sensi" del sistema. Indicando, ad esempio, il *syncon* "virus", tale attributo permette l'inclusione e l'individuazione in fase di estrazione (concettuale o NER) di tutti gli iponimi a lui relativi all'interno del *Sensigrafo*®. Ogni parola ad esso associato, inoltre, farà puntuale riferimento alla categoria/al dominio del *syncon* di partenza. |
| LIST | Permette il richiamo ad una lista di termini (lette dal sistema come parole chiave) realizzata all'interno del progetto linguistico dall'operatore (in formato *.txt). Se nella porzione di testo da analizzare viene rilevato un termine presente nella lista, allora il *matching* è positivo. Le liste possono contenere dati testuali di ogni tipo (ogni *token* viene definito dal punto e a capo) ed è *case sensitive.* |
| TYPE | Questo attributo permette di rilevare la categoria grammaticale [ADJ\|ART\|AUX\|ADV\|CON\|NOU\|NPH\|PNT\|PRE\|PRO\|VER] di un *token*, nonché di estrarre automaticamente alcune entità predefinite dal sistema [ADR=address\|DAT=date\|ENT=entity\|MON=money] |
| ROLE | Identifica un *token* in base al suo ruolo logico nella frase. Il ruolo si riferisce a una delle unità sintattiche di base comunemente riconosciute durante l'analisi logica di frasi o proposizioni: OBJECT: oggetto diretto; SUBJECT: soggetto; NOMINAL_P: predicato nominale; VERBAL_P: predicato verbale; COPULA: copula; INDIRECT: oggetto indiretto; OTHER: altro ruolo. |

| | |
|---|---|
| POSITION | Identifica un *token* in base alla sua posizione all'interno del documento. Fa riferimento ad un elenco di valori predefiniti che identificano le posizioni chiave per gli elementi testuali all'interno del documento stesso. I valori possono essere: BEGIN SENTENCE: Il *token* si trova all'inizio della frase; END SENTENCE: Il *token* si trova alla fine della frase; BEGIN PARAGRAPH: Il *token* è all'inizio di un paragrafo; END PARAGRAPH: Il *token* è alla fine di un paragrafo; BEGIN SECTION: Il *token* è all'inizio del documento; END SECTION: Il *token* si trova alla fine del documento. |
| PATTERN | Permette l'identificazione di un *token* o di una stringa semantica mediante l'utilizzo di *RegEx – Regular Expressions* (sintassi formale PERL)[282]. |

Gli attributi da soli, però, non sono sufficienti per garantire un'analisi linguistica puntuale: è necessario, quindi, fruire degli operatori (logici, booleani e di sequenza) che permettono la redazione di condizioni e/o dipendenze tra più attributi più complesse[283].

| **OPERATORE di SEQUENZA** | **DESCRIZIONE** |
|---|---|
| >><br><< | Indica una sequenza diretta tra attributi: ovvero, tra di loro non ci devono essere ulteriori elementi/*tokens*. Nella redazione della regola, dovrà essere indicata la posizione del *token* che segue o precede l'altro attributo di interesse = [#1], [#2], [#3], ecc. |
| ><br>< | Indica una sequenza debole tra attributi. Gli attributi possono essere separati da segni *«dal debole valore semantico»*[284] come, ad esempio: aggettivi, avverbi, congiunzioni, articoli. |
| < > | Indica una sequenza flessibile tra attributi. Gli elementi che si trovano prima e dopo tale sequenza devono essere specificati nella regola; qualsiasi *token* viene accettato tra di loro. |

[282] Si rimanda al paragrafo 3.7, dedicato alle espressioni regolari *RegEx.*
[283] In riferimento alla guida utente del *tool* reperibile al seguente url: https://docs.expertsystem.com/
[284] Ibidem.

| OPERATORE LOGICO | DESCRIZIONE |
|---|---|
| &SV<br>&VS | Sequenza soggetto-verbo. |
| &SO<br>&OS | Sequenza soggetto-oggetto. |
| &VO<br>&OV | Sequenza verbo-oggetto. |
| &SS | Sequenza soggetto-soggetto |
| &OO | Sequenza tra due oggetti diretti dipendenti dallo stesso verbo |
| &SS | Relazione tra due soggetti dipendenti dallo stesso verbo |

| OPERATORE BOOLEANO | DESCRIZIONE |
|---|---|
| NOT | Data una porzione di testo, due o più attributi non possono essere in essa contemporaneamente presenti. |
| AND | Data una porzione di testo, due o più attributi devono trovarsi in una condizione di compresenza affinché vi sia un *matching* positivo. |
| OR | Presenza alternativa o contemporanea di più attributi o espressioni valide in una data porzione di testo. |
| XOR | Presenza alternativa e non contemporanea di più attributi o espressioni valide in una data porzione di testo. |

## 3.5 *Linguaggio "C" – Categorizzazione*

COGITO STUDIO® permette due principali moduli di realizzazione dell'analisi delle informazioni presenti all'interno di un documento: il modulo di categorizzazione (realizzato mediante regole in linguaggio "C"[285] e partendo da una tassonomia concettuale – di tipo standard, oppure creata *ex novo* dall'operatore), ed il modulo di estrazione delle entità (realizzato mediante linguaggio proprietario "E").

Le entità, anche in questo caso, possono essere predefinite all'interno del sistema, oppure implementabili sulla base della *knowledge* dell'analista.

La categorizzazione permette la selezione di porzioni di testo (definiti dal/nel sistema come *scopes),* nonché il loro raggruppamento, sulla base di una serie di domini concettuali interconnessi con le entità di riferimento.

Il testo viene, quindi, suddiviso in sezioni, sul fondamento di un'azione predittiva effettuata in origine dal(la) linguista per la ricerca di un'informazione specifica, che può configurarsi come dato testuale e/o numerico, nonché come dato non strutturato/strutturato.

Gli *scopes* maggiormente utilizzati nell'analisi concettuale sono:

- DOCUMENT: l'informazione viene ricercata all'interno dell'intero dato testuale fornito in *input* dall'operatore; il sistema individua il primo *token* del testo come punto iniziale del DOCUMENT, e l'ultimo *token* (seguito o meno da un segno di interpunzione) come suo punto finale;
- SECTION: la categorizzazione viene effettuata all'interno di una sezione specifica del testo; l'operatore definirà nel file di configurazione i riferimenti di inizio e fine sezione (solitamente, si utilizzano delle parole chiave per delimitare tale *scope);*
- PARAGRAPH: il sistema definisce (e riconosce) la porzione di testo definita come paragrafo quella che inizia e termina con un punto e a capo;
- SENTENCE*:* l'inizio di tale porzione di testo viene identificato con il primo *token* del documento, oppure con il primo *token* preceduto da un punto, punto interrogativo o punto esclamativo; altresì, la fine di una porzione di testo di tipo SENTENCE viene definita alla presenza di un punto, un punto interrogativo o di punto esclamativo.

[285] Il linguaggio "C" di COGITO STUDIO® non rimanda in alcun modo al linguaggio di programmazione "C/C++".

Il linguaggio "C" prevede, inoltre, l'assegnazione (facoltativa) di un punteggio *(score)* ad ogni regola di categorizzazione, al fine di far emergere un determinato dominio su tutte le altre categorie potenzialmente associabili ai *syncon* e/o lemmi di riferimento alla regola stessa; la lettura e l'interpretazione delle singole parole è fortemente influenzata dall'informazione da esse ereditata dal concetto (o dai concetti) di appartenenza.

Il punteggio può osservare i seguenti parametri: STANDARD, LOW e HIGH.

Una regola redatta nel linguaggio "C" presenta la seguente forma:

```
SCOPE [tipologia di porzione del documento]
{ DOMAIN [indice del nodo tassonomico da associare ( + score level) ]
{      [regola di mining concettuale]      }
}
```

*Figura 16 – Esempio di regola di categorizzazione nel linguaggio "C"*[286]

Nella seconda riga della condizione, l'informazione relativa al nodo da associare alla regola – che ne permetterà il richiamo in fase di categorizzazione – fa riferimento all'albero tassonomico implementato nel progetto linguistico, indispensabile per l'avvio dell'investigazione concettuale di un testo.

Come si mostra nel seguente esempio – collegato al caso di studio "CTI" presentato nel quarto capitolo del presente lavoro – il nodo concettuale dell'albero tassonomico associato all'argomento *ransomware* è il DOMAIN numero 1.18 (indice definito dall'analista), ovvero, si configura come 18^ diramazione del primo macro-nodo tassonomico[287] *"Cyber Crime"*.

Facoltativamente, si può assegnare anche un peso alla regola (in questo caso, HIGH), ovvero si stabilisce un punteggio al dominio al verificarsi della regola stessa.

Il processo di categorizzazione avviene, quindi, anche attraverso la somma dei punteggi di tutte le regole implementate; in tal modo, si ha, per ogni dominio, un punteggio

[286] Fonte: grafico autoprodotto mediante *software* COGITO STUDIO®
[287] Si mostrerà l'intera matrice tassonomica del caso di studio nel quarto capitolo del presente elaborato.

in grado di indicare se una data porzione di testo – a cui sono state associate regole di *mining* – appartenga in misura maggiore o minore ad una determinata categoria.

Per quanto concerne lo *scope*, in questo caso, è di tipo SENTENCE: l'informazione verrà ricercata all'interno di ogni singola frase; la definizione di uno *scope*, come già evidenziato, è obbligatoria perché permette la delimitazione dell'ambiente testuale entro il quale la regola verrà inizializzata, e dal quale l'informazione verrà estratta.

```
SCOPE SENTENCE
{
     DOMAIN(1.18: HIGH)
     {  SYNCON (5950444)//#5950444:ransomware, Ransomware,
                cryptotrojan, cryptovirus, extortionware
        AND
        LEMMA ("attack")
     }
}
```

*Figura 17 – Esempio di regola di categorizzazione del progetto "CTI"*[288]

All'interno del corpo del { DOMAIN } l'operatore dovrà creare l'espressione di categorizzazione, ovvero la condizione linguistica che indicherà al sistema che cosa dovrà essere riconosciuto nel testo affinché un dato concetto venga individuato.

L'espressione può essere costituita da uno o più attributi connessi tra di loro mediante operatori booleani e di sequenza[289]. Un attributo permette di identificare un determinato *token* nel testo, cioè una data espressione linguistica, definendone la natura e le caratteristiche specifiche che permetteranno di far "scattare" in maniera puntuale la regola di *mining* concettuale.

Nel caso indicato nella precedente figura, il primo attributo indicato è il SYNCON (ID n° 5950444) del *Sensigrafo*® indicante il NOUN *ransomware*, ed associato ai domini *software e crime.*

[288] Fonte: grafico autoprodotto mediante *software* COGITO STUDIO®
[289] Come descritto nel paragrafo 3.4.1 *"Attributi ed operatori booleani, logici e di sequenza"*.

| Syncon | Definition | | | |
|---|---|---|---|---|
| N ransomware, Ransomware, cryptotrojan, cryptovirus,... | Type of malware used for data kidnapping; an exploit in which the attacker encrypts the victim's data ... | | | |
| ransomware | | | | |
| Ransomware | | | | |
| cryptotrojan | | | | |
| cryptovirus | | Domain [1] | Domain [2] | ID |
| extortionware | | software 70% | crime 50% | 5950444 |

*Figura 18 – SYNCON (ID n° 5950444) ransomware*[290]

Il secondo attributo è il LEMMA *attack:* aver scelto l'attributo LEMMA permette di includere nel processo di analisi concettuale tutte le entrate *attack* del dizionario *Sensigrafo*® di COGITO STUDIO®. Identifica, quindi, tale *token* nella sua forma base, indipendentemente dalla sua categoria grammaticale e/o associazione a uno o più domini. Inoltre, verranno inclusi nella ricerca tutte le varianti grafiche ed estensioni semantiche (tra cui, i sinonimi) associate a ogni lemma *attack* presenti nel *Sensigrafo*®.

| Syncon | Definition | Domain [1] | Domain [2] | ID |
|---|---|---|---|---|
| N assault, attack | A campaign or series of actions that... | | | 27104 |
| N flak, attack, flack, blast, fire | Intense adverse criticism; "Clinton directed his fire... | | | 30881 |
| N attack, offence, offense | The players making up the part of... | sport 100% | | 37572 |
| N attack | An episode or onset of a disease,... | medicine 30% | | 60826 |
| N plan of attack, approach, attack | Method, a way of doing or solving... | | | 178506 |
| N offensive tactic, offence, offense, attack | A sports tactics and offensive actions to... | sport 100% | | 188378 |
| N attack | Commencement of a task | | | 192366 |
| N attack | An attempted rape. | crime 30% | criminal la... | 192367 |
| N attack | The experience or beginning of a feeling,... | | | 199537 |
| N cyber attack, computer attack, computer crime attac... | An attack on, or by means of,... | computer scie... | crime 50% | 302946 |
| N attack | An attack by an animal | zoology 10% | | 326767 |
| V attack | To cause an infection, illness, or damage... | medicine 30% | | 66100 |
| V attack | To begin something such as work with... | | | 67529 |
| V assail, attack, assault, aggress | To try to defeat an enemy or... | military 40% | | 70759 |
| V assault, attack, assail, attempt | To try to harm somebody by using... | | | 70761 |
| V assail, attack, assault | Attack someone psychologically or verbally | | | 124315 |
| V attack | To attempt to defeat, or score against,... | sport 100% | | 147152 |
| A assault, attack | Used in an attack; | military 5% | | 90521 |
| A attack | Designed, planned, or employed for initiating, supportin... | | | 204256 |

*Figura 19 – Lemmi "attack" contenuti nel Sensigrafo*®[291]

Infine, il SYNCON *ransomware* (ID n° 5950444) ed il LEMMA *attack* sono posti in condizione di interdipendenza mediante operatore booleano AND: ovvero, la regola di

[290] Fonte: grafico autoprodotto mediante *feature Sensigrafo*® del *software* COGITO STUDIO®
[291] Ibidem.

categorizzazione “scatterà” solamente se detti attributi fossero contemporaneamente presenti all’interno della porzione di testo previamente definita *(scope* SENTENCE), pena la non estrazione dell’informazione. Per eseguire la validazione della regola di categorizzazione appena creata, l’operatore dovrà fornire in *input* un dato testuale idoneo al raggiungimento dei desiderata concettuali.

A titolo esemplificativo, si è scelta una porzione dell’*abstract* relativo dell’articolo *“Information security breach due to ransomware attacks: a systematic literature review”*.[292]

Il risultato del processo di categorizzazione effettuato dal sistema, sulla base della condizione linguistica precedentemente espressa, è il seguente:

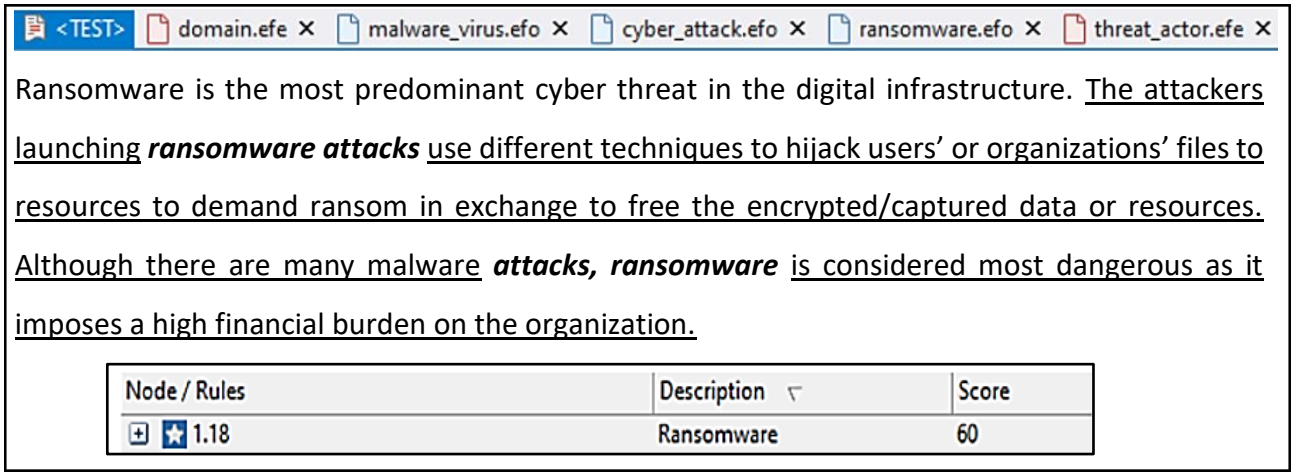



*Figura 20 – Validazione della condizione di categorizzazione*[293]

Come si può notare, ogni volta che la parola *ransomware* si trova in condizione di compresenza con il lemma *attack*, la porzione di testo di tipo SENTENCE di riferimento viene evidenziata in grassetto. Vengono, invece, sottolineati gli attributi/elementi linguistici che hanno fatto “scattare” la regola.

Sulla base della funzionante espressione linguistica implementata, come evidente, non viene evidenziata la parola *ransomware* nella prima SENTENCE (né appare, quindi, sottolineata la frase stessa), poiché non presente anche il lemma *attack* nella porzione di testo definita.

[292] T.R. RESHMI©. Cfr. https://www.sciencedirect.com/science/article/pii/S2667096821000069
[293] Fonte: grafico autoprodotto mediante *software* COGITO STUDIO®

## 3.6 *Linguaggio "E" – Estrazione*

Il processo di estrazione di dati linguistici da un *corpus* testuale, definito anche come operazione di *Text Mining*[294], consente di individuare ed estrapolare informazioni contenute in un testo/documento, nonché di memorizzare tali informazioni in apposite strutture dati.

In questa fase, i dati destinati all'analisi linguistica vengono *in primis* raccolti, selezionati e filtrati da diverse fonti – a seconda del progetto da realizzarsi – e, se necessario, vengono in un secondo momento trasformati e/o destrutturati al fine di soddisfare specifiche esigenze operative.

Il processo di estrazione consta di quattro sotto-processi realizzanti le cosiddette "Operazioni IETL": *Identification, Extraction, Transformation* e *Loading.* Nel dettaglio:

1. *Identification:* nella fase di identificazione (fase I) si individuano eventuali dati strutturati e/o non strutturati presenti nei *corpora* destinati alla cyber analisi linguistico-investigativa;
2. *Extraction:* nella fase di extraction (fase E) si creano puntuali condizioni linguistiche atte alla ricognizione ed estrapolazione di dati;
3. *Transformation:* nel processo di trasformazione (fase T) i dati estratti possono essere eventualmente *renderizzati* e salvati in un formato alternativo a quello predefinito;
4. *Loading:* l'ultima fase, di caricamento (fase L), i dati vengono implementati in un *Client* e memorizzati in strutture apposite.

In questo processo, nel sistema di COGITO STUDIO®, le prime due fasi (di identificazione ed estrazione) sono regolate dal linguaggio proprietario di tipo "E".

In particolare, il processo di estrazione si compone di determinati *steps:*

1. definizione dei *template* e dei *field* ad essi connessi: ovvero, si definiscono delle "strutture dati" designate alla classificazione e alla memorizzazione delle informazioni estratte. I *template* si trovano in una condizione di iperonimia nei confronti dei *field* –

[294] Come precedentemente analizzato nel paragrafo 1.6 *"Text Data Mining – estrazione di dati dal testo".*

adottando un approccio *top-down*. I *template* si possono, quindi, definire come dei contenitori concettuali di singole *keywords* e/o entità *(field)* da estrarre. Ad esempio:

```
TEMPLATE(CYBERTHREAT)
{
   @ATTACK_PATTERN, // Keywords Attack Patterns
   @LOCATION, // Estrazione di luoghi
   @MALWARE, // + RANSOMWARE
   @THREAT_ACTOR, // APT
   @TOOL, // Lista Tools utilizzati
   @COURSE_OF_ACTION, // COA (STIX)
   @VULNERABILITY, //CVE + CWE
   @CAMPAIGN, // Nomi Campagne di Attacchi
   @REPORT_PUBLISHED, // Estrazione di date
   @IDENTITY_ORGANIZATION, // Organizzazioni IT
   @IDENTITY_PERSON, // Identità Threat Actors
   @INFRASTRUCTURE // Botnet
}
```

*Figura 21 – Modulo di estrazione: Template e Fields*[295]

Prendendo come esempio il caso di studio che verrà presentato nel quarto capitolo, il *template "Cyberthreat"* – realizzato dall'analista sulla base della sua *knowledge* – racchiude al proprio interno i *fields* (seguiti dal carattere speciale "@") ad esso inerenti; mediante le *labels* indicanti i *fields*, verranno visualizzate le estrazioni sotto forma di parola chiave o altro dato testuale;

2. redazione delle regole: mediante il linguaggio "E" si definiscono condizioni linguistiche al fine dell'estrazione dei dati di interesse, attraverso l'utilizzo (come visto precedentemente) di attributi, operatori booleani, logici e di sequenza, nonché attraverso il richiamo a liste di *keywords* realizzate dall'analista, nonché ad espressioni regolari *(RegEx)* con sintassi PERL;

3. validazione delle estrazioni *(testing):* come presentato nel processo di categorizzazione, dando in *input* al sistema pertinenti dati testuali, è possibile validare quanto operato, al fine di effettuare un *fine tuning* delle regole implementate e/o la

[295] Fonte: grafico autoprodotto mediante *software* COGITO STUDIO®

creazione di ulteriori *templates/fields* per una classificazione ancor più raffinata delle informazioni.

- Esempio di condizione linguistica estratta dal progetto "CTI" – caso di studio presentato nel quarto capitolo – espressa mediante linguaggio "E":

```
SCOPE PARAGRAPH
{
    IDENTIFY(CYBERTHREAT)
    {
    @THREAT_ACTOR [KEYWORD ("Threat actor", "Threat actors",
                   "threat actor", "threat actors",
                   "Threat Actor", "Threat Actors")]
    <>
    LEMMA("execute")
    AND
    KEYWORD (EXPAND "malware.txt")
    }
}
```

*Figura 22 – Esempio di regola di estrazione dal progetto "CTI"*[296]

La regola presentata differisce dalla condizione di categorizzazione perché non fornisce indicazioni circa l'albero tassonomico del progetto, né (di conseguenza) al punteggio che, in maniera facoltativa, l'operatore potrebbe assegnare ad ogni nodo.

Subito dopo la definizione della porzione di testo *(scope* PARAGRAPH), anziché presentare l'indicatore DOMAIN, le regole di estrazione linguistica presentano l'indicatore IDENTIFY, facente riferimento al *template* definito *ex ante* per l'estrapolazione del *target* linguistico. Il *field,* relazionato al *template,* viene espresso mediante il carattere speciale "@". La condizione espressa nella stringa del *field* rappresenterà il dato da estrarre in fase di analisi.

In riferimento alla figura precedente, l'obiettivo di tale condizione linguistica è quello di estrarre l'informazione *Threat Actors* (espressa graficamente in tutte le varianti linguistiche possibili), a patto che sia contenuta:

[296] Ibidem.

- all’interno di una porzione di testo *(scope),* definita come PARAGRAPH*;*
- all’interno della stessa frase, dev’essere presente il termine *execute.* Non vengono fornite informazioni circa la categoria concettuale di tale lemma, né circa la sua eventuale coniugazione qualora si trattasse di un verbo. Mediante l’operatore di sequenza < >, si indica che la parola *execute* può trovarsi in posizione antecedente o successiva all’elemento *target* di estrazione linguistica *Threat Actors;*
- le precedenti due condizioni sono imprescindibili dalla presenza di una delle parole elencate nella lista “malware.txt”, redatta dall’operatore linguistico:

| **“malware.txt”** | |
|---|---|
| ABK | Hive |
| ransomware | Purple Fox |
| malware | CHEMISTGAMES |
| Dharma | Cherry Picker |
| RPis | China Chopper |
| Sandworm | CHOPSTICK |
| NotPetya | Circles |
| WannaCry | Clop |
| Cyclops Blink | CloudDuke |
| SandWorm | CoinTicker |
| Terra Loader | ArtraDownloader |
| Zeus Gameover | Agent Smith |
| USCYBERCOM | Agent Tesla |
| AgentTesla | Cyclops Blink |
| HawkEye | Agent Tesla |
| Noon | HermeticWiper |

***Figura 23 – Parte della lista “malware.txt”del progetto “CTI”***[297]

Al fine di poter validare l’effettiva operatività di tale espressione linguistica, si prendano come esempio le prime tre frasi tratte dall’articolo: *“From Ransomware to DDoS: Guide to Cyber Threat Actors – How, Why, and Who They Choose to Attack”,* della redazione di “*flashpoint.io”*[298].

***«What do Cyber Threat Actors want? Money, mostly. These actors who execute cyberattacks, such as ransomware, can wreak havoc on organizations across the private and public sectors. These Cyberattacks put their reputation, assets, stakeholders, and customers at stake».***

[297] Fonte: grafico autoprodotto.
[298] https://www.flashpoint-intel.com/blog/guide-to-cyber-threat-actors/

Risultato del processo di estrazione[299]:

| EXTRACTION RESULTS: RECORDS/FIELDS | INPUT TEXT <TEST> |
|---|---|
| CYBERTHREAT<br>MALWARE Cyberattacks | <TEST> domain.efe malware_virus.efo<br>These **Cyberattacks** put their reputation, assets, stakeholders, and customers at stake. |
| CYBERTHREAT<br>MALWARE cyberattacks | <TEST> domain.efe malware_virus.efo<br>These actors who execute **cyberattacks,** such as… |
| CYBERTHREAT<br>MALWARE ransomware | <TEST> domain.efe malware_virus.efo<br>These actors who execute cyberattacks, such as **ransomware,** can wreak havoc on organizations… |
| CYBERTHREAT<br>THREAT_ACTOR Threat Actors | <TEST> domain.efe malware_virus.efo<br>What do **Cyber Threat Actors** want? |

*Figura 24 – Validazione della condizione di estrazione*[300]

La condizione linguistica di estrazione presentata in precedenza trova esito positivo nel 4° risultato di estrazione: come si può notare, l'elemento linguistico (estrazione *target) Threat Actors* viene rilevato in formato grassetto.

Il sistema evidenzia, inoltre, mediante sottolineatura, l'intera frase. Ciò avviene perché all'interno della frase sono ravvisabili i seguenti elementi: il termine *execute* (in questo caso, sotto forma di verbo) e la presenza di (almeno) uno degli elementi presenti nella lista "malware.txt" redatta dell'operatore linguistico ed implementata a sistema (i termini interessati sono *cyberattacks e ransomware).*

[299] Mediante regole di normalizzazione – come verrà descritto nel paragrafo 4.4.1 *"Regole di Normalizzazione"* – le differenti varianti grafiche indicanti lo stesso referente (in questo caso, *cyber attacks*) verranno unite in un'unica condizione, decretando una di esse come "bontà del modello". Conseguentemente, nell'interfaccia del *software* in cui verrà implementato il pacchetto linguistico creato, tutte le varianti grafiche riscontrate risulteranno uniformate in un unico risultato di estrazione.

[300] Fonte: grafico autoprodotto mediante *software* COGITO STUDIO®

## *3.7 RegEx – PERL Regular Expressions Syntax*

In fase di estrazione linguistica, come già menzionato in precedenza, è possibile utilizzare l'attributo PATTERN per poter implementare a sistema delle condizioni contenenti espressioni regolari (definite anche *RegEx*, o *Regular Expressions).*

Le *RegEx* permettono l'estrazione di stringhe contenenti informazioni di tipo testuale, algebrico, alfanumerico, caratterizzate o meno dalla presenza di caratteri speciali. All'interno di COGITO STUDIO® viene riconosciuta e processata la sintassi PERL.

Nel presente paragrafo, non verranno delineati tutte i comandi e/o le funzionalità caratterizzanti le *RegEx* PERL, bensì verrà posta in evidenza la rilevanza di tale linguaggio ai fini della ricerca e dell'estrazione di minacce informatiche *(cyber threats)* e/o di informazioni anonimizzate contenute all'interno di documenti di varia natura ed emissione.

Si ricorre a tale linguaggio formale nel caso in cui il dato da estrarre – stringhe contenenti informazioni a carattere linguistico, alfanumerico, o caratterizzato da caratteri speciali – sia di difficile previsione, ma di cui se ne conosce la struttura.

«*Un'espressione regolare definisce una funzione che prende in ingresso una stringa, e restituisce in uscita un valore del tipo sì/no* [*output dal carattere dicotomico-binario], *a seconda che la stringa segua o meno un certo pattern».*[301]

Di seguito, si riporta una tabella contenente *patterns* di *RegEx* più frequenti, utilizzate anche per l'elaborazione del caso di studio *"CTI: Cyber Threat Intelligence per il rilevamento di minacce informatiche",* presentato nel quarto capitolo di questo lavoro.

[301] https://docs.microsoft.com/it-it/dotnet/standard/base-types/regular-expression-language-quick-reference

| ***RegEx patterns*[302]** | |
|---|---|
| ^ | Corrisponde all'inizio di una stringa. Se viene utilizzata la modalità multilinea /m, verrà applicata anche immediatamente dopo il primo carattere della nuova riga. |
| $ | Corrisponde alla fine di una stringa. Se viene utilizzata la modalità multilinea /m, verrà applicata anche in relazione al primo carattere della nuova riga. |
| [abc] | Individua i caratteri 'a', 'b' o 'c'. |
| [^abc] | Individua tutti i caratteri, ad eccezione di 'a', 'b' o 'c'. |
| \D | Individua tutti i caratteri, ad eccezione di quelli numerici/decimali. |
| \d | Individua tutti i caratteri numerici. È equivalente a [0-9]. |
| [a-z] | Individua tutti i caratteri alfabetici, da 'a' a 'z' ('z' inclusa). |
| [^a-z] | Individua tutti i caratteri, ad eccezione quelli all'interno del *range* 'a-z'. |
| [a-zA-Z] | Individua tutti i caratteri nel *range* 'a-z' o 'A-Z'. |
| \w | Individua tutte le lettere, numeri e *underscores*. È equivalente a [a-zA-Z0-9_]. |
| \W | Individua tutti i caratteri, ad eccezione di lettere, numeri e *underscores*. È equivalente alla funzione [^a-zA-Z0-9_]. |
| a{3} | Individua esattamente 3 caratteri di tipo 'a' consecutivi. |
| a{3,6} | Individua dai 3 ai 6 caratteri consecutivi di tipo 'a'. |
| \s | Individua tutti gli spazi e le nuove righe. |
| \S | Individua tutti i caratteri, ad eccezione degli spazi e delle nuove righe. |

Conoscere tali *patterns* ha rivestito fondamentale importanza per l'elaborazione del caso di studio presentato nel presente lavoro di tesi; nello specifico, ai fini dell'estrazione delle informazioni inerenti gli *IoC – Indicators of Compromise*.

Tali indicatori di compromissione rappresentano una prova a livello forense circa potenziali intrusioni in un sistema *host* o in una rete.

Quando si parla di *RegEx*, solitamente non ci si riferisce a condizioni di (inter)dipendenza sintattica e/o semantica tra gli elementi, bensì, di estrazione puntuale di una determinata unità di significato pertinenti non al linguaggio naturale, ma al linguaggio formale.

[302] È stato preso come riferimento il seguente sito di validazione di *RegEx:* https://regex101.com/

L'estrazione di tali elementi, però, può essere associata a ulteriori unità di senso, dalle quali potranno poi dipendere o con le quali dovranno co-occorrere al fine di acquisire significato per l'operatore.

Nello specifico, tra le informazioni *target* del progetto linguistico trattato nel quarto capitolo – la cui estrazione è stata resa possibile tramite *RegEx* PERL – vi sono:

- funzioni (crittografiche) di *hash:* un *hash* è una stringa di *bit* correlata con i dati in *input*. Consta di 32, 40 o 64 caratteri alfanumerici. Di seguito, alcuni esempi di *RegEx* per l'individuazione di diversi tipi di *hash*:

- tipo di *hash: MD5* [32 caratteri alfanumerici] | *RegEx*: [A-Fa-f0-9]{32}
  Esempio: f614909fbd57ece81d00b01958338ec2
- tipo di *hash: SHA1* [40 caratteri alfanumerici] | *RegEx:* [A-Fa-f0-9]{40}
  Esempio: 5c07e333d381c7a13491cd88f783b4138e32d5db
- tipo di *hash: SHA256* [64 caratteri alfanumerici] | *RegEx:* [A-Fa-f0-9]{64}
  Es.: cafe8f704095b1f5e0a885f75b1b41a7395a1c62fd893ef44348f9702b3a0deb

```
SCOPE PARAGRAPH
{
     IDENTIFY(CYBERTHREAT)
     {
        @INDICATOR_HASH [KEYWORD ("[a-Za-Z0-9] {64}")]
        OR
        @INDICATOR_HASH [KEYWORD ("[a-Za-Z0-9] {32}")]
        OR
        @INDICATOR_HASH [KEYWORD ("[a-Za-Z0-9] {40}")]
        }
}
```

*Figura 25 – Regola di estrazione per gli hash su COGITO STUDIO®* [303]

- strutture di domini, anche anonimizzati; ai fini dell'individuazione di domini anonimizzati, bisogna indicare la presenza di eventuali caratteri speciali, oltre ai caratteri alfanumerici, purché non in posizione iniziale o finale della stringa; *RegEx:* [a-z0-9][-] | [a-Za-z0-9][-]

[303] Fonte: grafico autoprodotto mediante *software* COGITO STUDIO®

- esempi validi: *cocomo1.my-vnc.com* | *coco-mo2.myvnc.com*
- esempi non validi: *-cocomo1.myvnc.com* | *cocomo1-.myvnc.com*

- strutture di url ed e-mail anonimizzati e non: la presenza delle parentesi quadre tra i *dots* della stringa solitamente sono indice di anonimizzazione, così come la sostituzione dell'*Hypertext Transfer Protocol [over Secure Socket Layer]* – ovvero, dei protocolli *http* e/o *https* – con i prefissi *hxxp* e/o *hxxps*, utilizzata per la trasmissione di *malware;*

- esempi di e-mail anonimizzate: *alice*[.]*312@gmail.com* ; rain.5@outlook[.]com

```
SCOPE SENTENCE
{
   IDENTIFY(CYBERTHREAT_IND)
   {
   @INDICATOR_EMAIL[PATTERN("[A-z0-9\.\+_-]+@[A-z0-9\._-]+\.[A-z]{2,6}")]
   OR
   @INDICATOR_EMAIL[PATTERN("[A-z0-9\.\+_-]+@[A-z0-9\._-]+\[punto\]+[A-z]{2,6}")]
   }
}
```

*Figura 26 – Esempio di estrazione di email anonimizzate su COGITO STUDIO®*[304]

- esempi di domini e url anonimizzati: *decent4.myvnc**[.]**com* | *cocomo1.ddns.**]**net* | *cocomo1.ddns**[**.net* | *http[s]://domain**[**.tld/page* | *http://108**[.]**61.189.174* (indirizzo IP mascherato) | ***hxxp[s]****://offlineearth**[.]**com/upload?id=111* (protocollo mascherato) | *http://34.13.42**[.]**35/uploads/2.jpg* (indirizzo IP mascherato + allegato).

- strutture di indirizzi IP, anche anonimizzati: in questo caso, al fine di estrarre un indirizzo IP, la condizione formale deve configurarsi come una stringa a carattere esclusivamente numerico, caratterizzata da quattro *chunks* da uno a tre cifre, nessuna delle quali superiore a 255, ed ognuno dei quali contrassegnato da un punto (e da parentesi quadre attorno ad esso, in caso di IP mascherato);

- esempio di IP non anonimizzato: "186.12.110.32". *RegEx* per individuarlo:

^((25[0-5]|2[0-4][0-9]|[01]?[0-9][0-9]?)\.){3}(25[0-5]|2[0-4][0-9]|[01]?[0-9][0-9]?)$

[304] Ibidem.

- esempio di IP anonimizzato: “179.43.176[.]118”. *RegEx* per raggiungerlo:
^((25[0-5]|2[0-4][0-9]|[01]?[0-9][0-9]?)\[.]){3}(25[0-5]|2[0-4][0-9]|[01]?[0-9][0-9]?)$

Di seguito, la *RegEx* implementata all’interno del progetto linguistico “CTI” per l’estrazione di IP anonimizzati e non, mediante l’utilizzo del linguaggio “E”:

```
SCOPE SENTENCE
{
  IDENTIFY(CYBERTHREAT_IND)
  {
  @INDICATOR_IP[PATTERN("^((25[0-5]|2[0-4][0-9]|[01]?[0-9][0-9]?)\.){3}

                        (25[0-5]|2[0-4][0-9]|[01]?[0-9][0-9]?)$")]

  OR
  @INDICATOR_IP[PATTERN("^((25[0-5]|2[0-4][0-9]|[01]?[0-9][0-9]?)\[.]){3}

                         (25[0-5]|2[0-4][0-9]|[01]?[0-9][0-9]?)$")]

  }
}
```

*Figura 27 – Esempio di estrazione di IP anonimizzati e non, su COGITO STUDIO®* [305]

[305] Ibidem.

# 4. ELABORAZIONE *di* UN CASO *di* STUDIO: *"Cyber Threat Intelligence per il rilevamento di minacce informatiche"*

## *4.1 Introduzione alla Cyber Threat Intelligence*

La *Cyber Threat Intelligence* (CTI) è una componente essenziale nell'architettura della *Cyber Security:* si configura come «*servizio di informazione strategica sulle minacce informatiche*»[306], e fa riferimento alle metodologie e agli strumenti atti all'identificazione, alla prevenzione e alla mitigazione di *cyber threats* e di altri eventi legati alla sicurezza delle reti.

Nello specifico, il processo di analisi alla base della *Cyber Threat Intelligence* si fonda su tre fattori principali: gli attori delle minacce, le loro intenzioni e le loro capacità. *«Si tiene quindi conto di tattiche, tecniche e procedure (TTP), motivazioni e accesso agli obiettivi previsti. Studiando questa triade è spesso possibile effettuare valutazioni strategico-operative».*[307]

Per poter effettuare una qualsiasi azione in questo ambito, è importante avere una conoscenza non opaca del concetto di minaccia informatica, definita dal *National Institute of Standards and Technology* (NIST)[308] come: *«Any event* […] *with the potential to adversely impact organizational assets or individuals through an information system via unauthorized access, destruction, disclosure, modification of information or denial of service».*[309]

La CTI è un processo circolare che segue le stesse fasi del ciclo di Intelligence[310]: l'analisi per il rilevamento delle minacce ha avvio, quindi, con la raccolta dati *(Information*

[306]https://www.pandasecurity.com/it/mediacenter/sicurezza/cose-la-cyber-threat-intelligence/
[307]Laboratorio di Ricerca e Innovazione per la Sicurezza. Cfr. https://rislab.it/cyber-threat-intelligence-cose-e-perche-e-cosi-importante/
[308] U.S. Department of Commerce. Cfr. https://www.nist.gov/
[309] National Institute of Standards and Technology (NIST), Federal Information Processing Standards (FIPS), Minimum Security Requirements for Federal Information and Information Systems, 2006. Cfr. https://csrc.nist.gov/glossary/term/threat [Traduzione: *«Qualsiasi evento [...] con il potenziale di avere un impatto negativo sulle risorse organizzative o sugli individui attraverso un sistema informativo mediante accesso non autorizzato, distruzione, divulgazione, modifica delle informazioni o negazione del servizio»].*
[310] Si rimanda al paragrafo 1.1.1 *"Le fasi del ciclo di Intelligence"*.

*Gathering)* da fonti aperte o *database* locali (ciò si configura come fase di *data input*), proseguendo poi nell'azione di *analysis* e *filtering* dei dati raccolti, fino ad arrivare alla loro valutazione e finale implementazione.

Gli *output* del processo di *threat detection* vengono accuratamente processati al fine di produrre e garantire una solida *knowledge*; la condivisione di tale conoscenza viene attuata mediante una oculata *dissemination* degli *outcomes* a *decision makers*.

L'*output* della *cyber analysis* – al fine di poter diventare *outcome* – viene costantemente validato e/o rivalutato all'emergere di nuove minacce, informazioni e *feedback*; ciò permette di avere un quadro quanto più aggiornato e puntuale possibile in materia di *cyber risk.*

Dopo aver operato tale individuazione, la CTI si pone come obiettivo il tempestivo preavviso di minacce, a supporto del processo decisionale degli analisti di Intelligence. Le *cyber threats* vengono scagliate quotidianamente tanto ad organizzazioni militari – configurandosi come minacce alla Sicurezza Nazionale – quanto a *Corporates*.

Questa (crescente) persistenza ha comportato un cambio di prospettiva nella concezione di sicurezza cibernetica stessa che, se previamente intesa come soluzione per la realizzazione di una sicurezza perimetrale, ora sta cedendo il posto ad una nuova visione – a 360°, interattiva e sincronica – destinata non solo alla difesa, bensì alla modellazione delle minacce, ai fini sia della loro ricognizione, che dell'esecuzione tempestiva di azioni di mitigazione e/o *counter-attacks.*

Negli ultimi mesi, a causa dei conflitti in essere, in particolar modo lo scontro russo-ucraino, si è registrato un incremento del 196% nella disseminazione di *malware e wipers*[311] evoluti verso Enti Governativi, a testimonianza del fatto che, al giorno d'oggi, le armi cibernetiche siano ormai parte integrante delle strategie adottate anche in ambito militare, in un'ottica *cyber warfare.*

Secondo *Gartner S.p.A.* – multinazionale di consulenza strategica, ricerca ed analisi nel campo della tecnologia dell'informazione – operando una stima valutativa e

[311] F. Cofini, *La guerra sul web: in Ucraina cyber attacchi triplicati e servizi fondamentali a rischio*, per *RaiNews.*, 2022. Cfr. www.rainews.it/articoli/2022/03/la-guerra-sul-web-in-russia-e-ucraina-cyber-attacchi-triplicati-e-servizi-fondamentali-a-rischio-73de6ad4-ecc9-40a1-a21a-8d6a6a3c2f52.html

quantificativa delle contromisure adottate negli ultimi anni[312] finalizzate al controllo degli accessi alle infrastrutture e alla messa in protezione dei dati – oltre il 75% degli attacchi realizzati hanno avuto come obiettivo applicativi *software*; hanno mirato, quindi, alle vulnerabilità dei sistemi dell'Ente attaccato, compromettendone il funzionamento ed estraendo illegalmente dati sensibili da essi.

Tale capacità di Intelligence sviluppata in ambito *Cyber Security* dev'essere, pertanto, una componente necessaria in tutte le soluzioni *software;* esse devono, quindi, prevedere *features* di analisi e gestione del rischio in tempo reale, al fine sia di evitare *data breaches*, che di mettere in atto misure di sicurezza in caso di futuri attacchi – anche nel caso di emergenze dalla portata internazionale.

[312] AGID, Agenzia per l'Italia Digitale. Cfr. www.agid.gov.it/sites/default/files/repository_files/allegato_4_linee_guida_per_la_modellazione_delle_minacce-dlt.pdf

## *4.2 Presentazione del progetto*

Come precedentemente definito nella presente tesi di Laurea, coniugare l'azione linguistica alla *Cyber Threat Intelligence* (CTI) per il rilevamento di minacce informatiche è stato – per me e per l'azienda per la quale, con grande passione, opero come linguista dal settembre 2020 – una sfida estremamente stimolante.

La realizzazione del seguente caso di studio ha permesso ad esperti di dominio CTI, specialisti di reti neurali artificiali ed esperti del linguaggio di sedersi alla stessa tavola rotonda, condividendo la propria *knowledge* e ragionando su metodologie, strategie e strumenti applicabili per il raggiungimento dei desiderata del Cliente.

Il risultato dell'analisi di *Intelligence linguistica* è stato reso possibile attraverso l'utilizzo di una complessa tecnologia semantica; elemento imprescindibile allo strumento, ai fini dell'avvio di tale progetto, però, è stato l'attuazione da parte degli analisti di un cosciente processo di immedesimazione – inteso come tentativo di scambio di identità ed acquisizione del *modus operandi et cogitandi* – nei confronti degli attori di *cyber threats.*

Tale iato tra l'operatore di *Intelligence* linguistica e l'identità/l'azione criminale ha permesso la ricognizione (e il successivo *processing)* della minaccia considerata, nonché la supposizione di potenziali contromisure per la sua mitigazione e/o la sua neutralizzazione.

Questo processo si può assimilare, in ambito *cyber*, alle architetture di *reverse engineering* – in tale ambito, per la realizzazione di un *digital twin*, ovvero di un "gemello digitale": una replica virtuale di dispositivi e risorse fisiche, equivalenti a oggetti, processi, entità, sistemi e dispositivi – in questo caso, al fine di effettuare un'analisi predittiva delle strategie di azione che guidano gli *hackers/hacktivists* verso lo scagliamento, ad esempio, di un *ransomware*.

Considerate le dinamiche imprevedibili ed in continua evoluzione degli attacchi informatici, viene, alla data odierna, richiesto agli operatori linguistici e agli esperti del quinto dominio, un costante aggiornamento in materia di *cyber crime* e *cyber attacks*, al fine di acquisire una *knowledge* sempre più stabile in tale ambito, e farsi trovare preparati a nuove minacce e nuovi strumenti in grado sia di inviarle, che di schermarle.

Ad oggi, sempre più compagnie ed Enti Governativi si trovano a fare (letteralmente) i conti con *ransomware* o altro tipo di circuizioni economiche che, se non tempestivamente soddisfatte (generalmente, nell'arco di 72 ore), ne lederebbero in maniera irreversibile il prestigio ed il *business*, per la conseguente fuoriuscita di dati ed informazioni sensibili sia della società stessa, che sui propri clienti (o dei propri cittadini, nel caso di organizzazioni militari e/o *Governance).*

Tale caso di studio non si prefigura come fondamento linguistico per la creazione di una soluzione in stile "anti-virus", bensì come "base di conoscenza strategica" a cui attingere per la navigazione, acquisizione ed estrazione di informazioni da fonti OSINT, SOCMINT e altro tipo di dato testuale proveniente da *database* locali, ai fini del potenziamento della *cyber awareness* dell'Ente.

In particolare, per la raccolta di *raw data* testuali, indispensabili per attivare in prima istanza il processo di elaborazione linguistica, si è preso come riferimento il sito web *Cyber Security Review*[313], la cui missione è quella di promuovere quotidianamente lo scambio di informazioni e la collaborazione tra aziende di tipo *Software House, stakeholders,* tra il mondo accademico e gli esperti di cyber-sicurezza di tutto il mondo.

*Cyber Security Review* è una piattaforma online concepita per attingere alla *knowledge,* alle capacità e all'esperienza combinata dei membri della sua *Cyber Security Community,* al fine di identificare le minacce emergenti e facilitare lo sviluppo di *policy* (azioni e non azioni) pertinenti a carattere pubblico e privato.

Attraverso la pubblicazione giornaliera di *post* ed articoli, nonché attraverso la sua sezione libera "*Resources"* – contenente sia *link* a risorse esterne in ambito CTI, che una *repository* per la condivisione di quanto già pubblicato in precedenza – gli analisti linguistici hanno potuto intendere tale piattaforma come punto di partenza per la collezione di dati, sfruttando così la tecnologia OSINT e SOCMINT nello stesso momento.

In accordo con i desiderata del Cliente, sono stati elaborati tre moduli *(work packages)* di classificazione ed individuazione delle informazioni relativamente agli attori di *cyber threats* (nonché associazioni o gruppi criminosi), alle loro intenzioni, capacità e agli strumenti da loro utilizzati:

[313]https://www.cybersecurity-review.com/

1) modulo di categorizzazione, con annessa tassonomia *customizzata,* in grado di riconoscere ed associare rapidamente domini e macro-argomenti pertinenti alle aree e ai contesti di interesse; questo modulo permette all'utente un accesso rapido all'informazione mediante una lettura da parte del *software* di tipo *skimming*[314];

2) modulo di estrazione *(Text Mining),* per l'individuazione di parole chiave, resa possibile da condizioni formali a carattere sintattico e semantico create dall'operatore linguistico. Il tipo di lettura automatizzata realizzata è di tipo *scanning*[315]*;*

3) modulo di relazione semantica tra entità, per l'individuazione di (inter)dipendenze semantiche tra due o più elementi estratti mediante il modulo di estrazione. Tali relazioni permettono la comprensione delle dinamiche delle azioni criminose, ovvero: chi sono gli attori delle minacce *(Threat Actors),* a quali gruppi di *hackers/hacktivists* appartengono, quali sono le loro intenzioni, chi subisce l'attacco e con quali strumenti esso viene generato.

Per lo sviluppo di tutti e tre i moduli, è stato preso come riferimento il modello STIX: *Structured Threat Information eXpressions*[316]*:* un linguaggio standardizzato per la trasmissione di dati inerenti alle minacce alla sicurezza informatica.

Tale linguaggio viene utilizzato da *threat analysts* per esaminare le minacce informatiche e altre attività illecite online. *«Lo standard STIX permette di identificare modelli in grado di rimandare ad intrusioni, reati o minacce sospette e/o concrete»*[317], nonché permette di realizzare attività di risposta a dette azioni malevole.

Sotto la guida dell'esperto di dominio presente nel *team* di progetto, si è dato avvio alla prima fase del ciclo di *Threat Intelligence:* la raccolta di dati dalle più rilevanti ed attendibili fonti aperte, *press news*, nonché dai forum più consultati nell'ambito informatico; ciò ci ha permesso di partire da una base di dati (un *raw dataset)*, con l'obiettivo di fornire al Cliente un prodotto il più all'avanguardia possibile.

Nel *team* è stato necessario fondere le proprie capacità proattive di analisi e di ricerca, occorrendo ad una riflessione (in modalità *think tank)* riguardante il dominio CTI nella sua

[314] Come precedentemente definito nel paragrafo 1.6 *"Text Data Mining: Estrazione di dati dal testo".*
[315] Ibidem.
[316] *Structured Threat Information eXpression (STIX™) 1.x Archive Website.* https://stixproject.github.io/
[317] L. Zanotti, *Sicurezza e collaboration. Come funziona il framework STIX?,* 2017. Cfr. www.zerounoweb.it/techtarget/searchsecurity/sicurezza-e-collaboration-come-funziona-il-framework-stix/

totalità, per poi applicare tale *knowledge «nel teatro operativo della conoscenza acquisita e condivisa».*[318]

L'impegno di raccolta documentale si è focalizzato sulle minacce costituite da gruppi, micro-gruppi e/o circuiti più ampi e transnazionali, costituiti in particolar modo da attori radicalizzati ed attivi online e in contatto tra loro soprattutto mediante *social networks*. Questo perché *«gli ambienti virtuali hanno* […] *un effetto aggregante rispetto a comunità distinte e/o frammentate, le cui differenze linguistiche, culturali ed etniche* [potrebbero impedire] *qualsiasi commistione».*[319]

A seguito della fase di *Data Gathering* – seguendo lo standard STIX come punto di interconnessione – si è passati alla comprensione delle parole chiave più rilevanti in materia di *Cyber Security and cyber crime*, per la realizzazione dei moduli di estrazione e di categorizzazione.

Circa quest'ultimo, è stato indispensabile creare un albero tassonomico *ad hoc*, in grado di rispondere ai *macro-topics* di dominio, per una classificazione accurata e categorica delle informazioni.

[318] Presidenza del Consiglio dei Ministri, Sistema di Informazione per la Sicurezza della Repubblica, *Connessi con la sicurezza. Il racconto di una Intelligence diffusa*, LeggIntelligence, 2015. Cfr. https://www.sicurezzanazionale.gov.it/sisr.nsf/letture/connessi-con-la-sicurezza-il-racconto-di-una-intelligence-diffusa.html

[319] Presidenza del Consiglio dei Ministri, Sistema di Informazione per la Sicurezza della Repubblica, *Relazione annuale sulla politica dell'informazione per la sicurezza*, 2021, p 84.

## *4.3 Modulo di categorizzazione*

Come precedentemente analizzato nel paragrafo 3.5 *"Linguaggio 'C' – Categorizzazione"*, il linguaggio "C" di COGITO STUDIO® è un linguaggio dichiarativo che definisce un insieme di condizioni che devono verificarsi durante l'analisi linguistica del testo fornito in *input*.

Mediante tali condizioni formali, vengono associati uno o più domini di interesse al dato testuale, con lo scopo di effettuare una affidabile categorizzazione automatica dell'informazione.

Obiettivo del *work package* di categorizzazione del progetto CTI è stato la corretta identificazione da parte del sistema linguistico di macro-argomenti relativi all'ambito del *cyber crime*, risolvendo significati dubbi, elaborando i casi di polisemia – attraverso l'associazione di parole chiave strategiche al dominio *target* – al fine di rilevare puntualmente *topics* e categorie predefiniti[320].

Per realizzare il modulo "C", è necessaria l'implementazione di un albero tassonomico[321] nel progetto. Ogni condizione formale richiamerà un nodo della tassonomia. Progettare una tassonomia è un'azione di precisione, mediante la quale dipenderà il livello di accuratezza nella categorizzazione di concetti nonché nella (successiva) estrazione di significati.

Si presenta di seguito l'albero tassonomico creato dal *team* di progetto ed implementata nel *framework semantico* per la realizzazione del modulo di classificazione delle informazioni:

[320] Se si verificano le condizioni definite nella regola, il sistema assegna un punteggio *(score)* al dominio associato.
[321] In un motore semantico, l'albero tassonomico è lo strumento che realizza la collaborazione tra linguistica, psicologia cognitiva e computazione.

| CYBER_TAXONOMY | Cyber Illegal Taxonomy |
|---|---|
| 1.1 | Cyber Attack |
| 1.2 | Broken Authentication and Session Management |
| 1.3 | Man-in-the-middle |
| 1.4 | Cyber Deception |
| 1.5 | Information Gathering |
| 1.6 | Identity Theft |
| 1.7 | Phishing |
| 1.8 | Mobile Vulnerability |
| 1.9 | Application and Software Vulnerabilities |
| 1.10 | DoS Attack |
| 1.11 | Zero-day |
| 1.12 | Threat and Vectors |
| 1.13 | Malware and Virus |
| 1.14 | Botnet |
| 1.15 | Advanced Persistent Threat (APT) |
| 1.16 | Ransomware |
| 1.17 | Intrusion (Computer and Network) |

***Figura 28a – "Cyber_ Illegal Taxonomy" in formato tabulare***[322]

A differenza degli alberi tassonomici ideati per progetti linguistici sviluppati *ex ante,* la tassonomia realizzata per il progetto "CTI" – denominata "*Cyber Illegal Taxonomy"* non presenta elementi disposti in ordine gerarchico e/o meronimico tra loro. Questo perché non sussiste un vero e proprio rapporto di subordinazione tra i nodi tassonomici indicati, né una dinamica o condizione temporale nel susseguirsi degli eventi cibernetici.

Dal formato *.xml dell'albero, infatti, si evince che tutti gli elementi sono disposti sullo stesso livello.

[322] Fonte: grafico autoprodotto.

```
<DOMAINTREE>
 <DOMAIN DESCRIPTION="Cyber Illegal Taxonomy" NAME="CYBER_TAXONOMY">
          <DOMAIN DESCRIPTION="Cyber Attack" NAME="1.1"/>
          <DOMAIN DESCRIPTION="Broken Authentication and Session Management" NAME="1.2"/>
          <DOMAIN DESCRIPTION="Man-in-the-middle" NAME="1.3"/>
          <DOMAIN DESCRIPTION="Cyber Deception" NAME="1.4"/>
          <DOMAIN DESCRIPTION="Information Gathering" NAME="1.5"/>
          <DOMAIN DESCRIPTION="Identity Theft" NAME="1.6"/>
          <DOMAIN DESCRIPTION="Phishing" NAME="1.7"/>
          <DOMAIN DESCRIPTION="Mobile Vulnerability" NAME="1.8"/>
          <DOMAIN DESCRIPTION="Application and Software Vulnerabilities" NAME="1.9"/>
          <DOMAIN DESCRIPTION="DoS Attack" NAME="1.10"/>
          <DOMAIN DESCRIPTION="Zero-day" NAME="1.11"/>
          <DOMAIN DESCRIPTION="Threat and Vectors" NAME="1.12"/>
          <DOMAIN DESCRIPTION="Malware and Virus" NAME="1.13"/>
          <DOMAIN DESCRIPTION="Botnet" NAME="1.14"/>
          <DOMAIN DESCRIPTION="Advanced Persistent Threat (APT)" NAME="1.15"/>
          <DOMAIN DESCRIPTION="Ransomware" NAME="1.16" />
          <DOMAIN DESCRIPTION="Intrusion (Computer and Network)" NAME="1.17"/>
 </DOMAIN>
</DOMAINTREE>
```

*Figura 28b – "Cyber_Illegal_Taxonomy.xml"*[323]

Partendo dal primo nodo tassonomico *<1.1 – Cyber Attack>* mostrerò la logica applicata nella redazione di regole linguistiche di categorizzazione dell'informazione.

<1.1> In primo luogo, bisogna stabilire la porzione di testo nella/dalla quale si vorrà individuare l'informazione: lo *scope* prescelto sarà di tipo SENTENCE ovvero, le condizioni di categorizzazione si realizzeranno solo nel perimetro sintattico della frase.

A livello semantico, il concetto da rilevare *(Cyber Attack)* viene espresso, in lingua inglese, mediante l'unione di due termini: l'aggettivo *cyber* – posto sempre in posizione antecedente al sostantivo[324] – ed il sostantivo *attack.* In lingua inglese, inoltre, tale termine può essere presentato – senza alcuna variazione di senso – mediante tre varianti grafiche, da cui l'esperto/a linguistico/a non può prescindere nella redazione delle condizioni formali: *cyber attack, cyber-attack* e *cyberattack*[325]. Il concetto di "attacco informatico", poi, può essere espresso anche mediante sinonimi, sia dell'aggettivo "cibernetico", che del sostantivo "attacco".

[323] Fonte: grafico autoprodotto mediante *software* COGITO STUDIO®
[324] A differenza della lingua italiana, in inglese gli aggettivi hanno posizione fissa: sono posti sempre prima del sostantivo e hanno la stessa forma al maschile, femminile, singolare e plurale.
[325] Nelle regole di normalizzazione (come verrà descritto nel paragrafo 4.4.1) aggiungerò condizioni formali in cui sono presenti le varianti grafiche presentanti caratteri in maiuscolo.

Il/la linguista indicherà all'interno del sistema tali concetti sotto forma di SYNCON, legandoli tra loro mediante l'operatore booleano AND. In questo modo, al sistema verrà dettata una precisa condizione per la categorizzazione dei macro-argomenti: l'impreteribile co-presenza – all'interno di una porzione di testo predefinita – di tutte le *keywords* indicate nella regola linguistica.

Indicando il TYPE SYNCON, anziché LEMMA, si dà ordine al motore semantico di richiamare l'informazione, nonché di rielaborarla, attraverso la propria *Knowledge Base*, ovvero mediante il *Sensigrafo*® di COGITO STUDIO®, che – come descritto nel paragrafo 3.3 del presente lavoro di tesi– si configura come insieme di reti semantiche interconnesse tra loro tramite una struttura a grafo, in cui ogni parola è sia collegata ad altri SYNCON, che associata a uno o più concetti/DOMAINS. Mediante l'associazione a domini, il *Sensigrafo*® è in grado di risolvere le ambiguità semantiche.

Infine, l'analista linguistico/a potrà indicare strategicamente l'attributo ANCESTOR ad entrambi i lemmi (sia per *cyber*, che per *attack)* al fine di poter includere tutti i loro iponimi in fase di estrazione concettuale; tali iponimi dovranno esser associati esclusivamente a domini pertinenti, come, ad esempio: *cyber*, *cyber crime*, *computer science*, *software.*

```
1.
SCOPE SENTENCE
{ DOMAIN (1.1: HIGH)
   {  SYNCON (302946,5925431) //#302946: cyber attack, computer attack
   }
}


2.
SCOPE SENTENCE
{ DOMAIN (1.1: HIGH)
   {  SYNCON (302946,5925431) //#302946: cyber attack, computer attack
      AND
      LEMMA ("attack", "crime", "fraud", "malicious")
   }
}
```

*Figura 29a – Regole di categorizzazione del nodo tassonomico <1.1 – Cyber Attack>*[326]

[326] Fonte: grafico autoprodotto mediante *software* COGITO STUDIO®

```
3.
SCOPE SENTENCE
{ DOMAIN (1.1: HIGH)
   { ANCESTOR (302946,5925431) //# 302946: cyber attack, computer attack,
      computer crime attack; 5925431: cyber attack, cyberattack
   }
}

4.
SCOPE SENTENCE
{  DOMAIN (1. 1: HIGH)
   {  ANCESTOR (286933) //# 286933: cyber, virtual
      AND
      ANCESTOR (3133,3950) //# 3133: attack, assailment, attempt; 3950:
      attack, onrush.
   }
}
```

*Figura 29b – Regole di categorizzazione del nodo tassonomico <1.1 – Cyber Attack>*[327]

Dalle precedenti figure, si può riscontrare un altro elemento: al fianco dell'indicazione del nodo tassonomico <1.1>, viene espresso uno *score*, ovvero un punteggio. Ciò permette di conferire minore o maggiore prestigio ad una determinata regola, in relazione ad altre condizioni linguistiche potenzialmente implementate nel motore semantico. In questo caso, viene attribuito un punteggio massimo alle condizioni indicate.

Per la validazione delle condizioni linguistiche di categorizzazione sopra indicate, utilizzerò come *input* testuale la definizione di *Cyber Attacks* fornita da *Wikipedia*© in lingua inglese, aggiungendo, ai fini del presente test, differenti varianti linguistiche, espresse a livello grafico, del concetto *Cyber Attack*, come precedentemente elencate.

[327] Ibidem

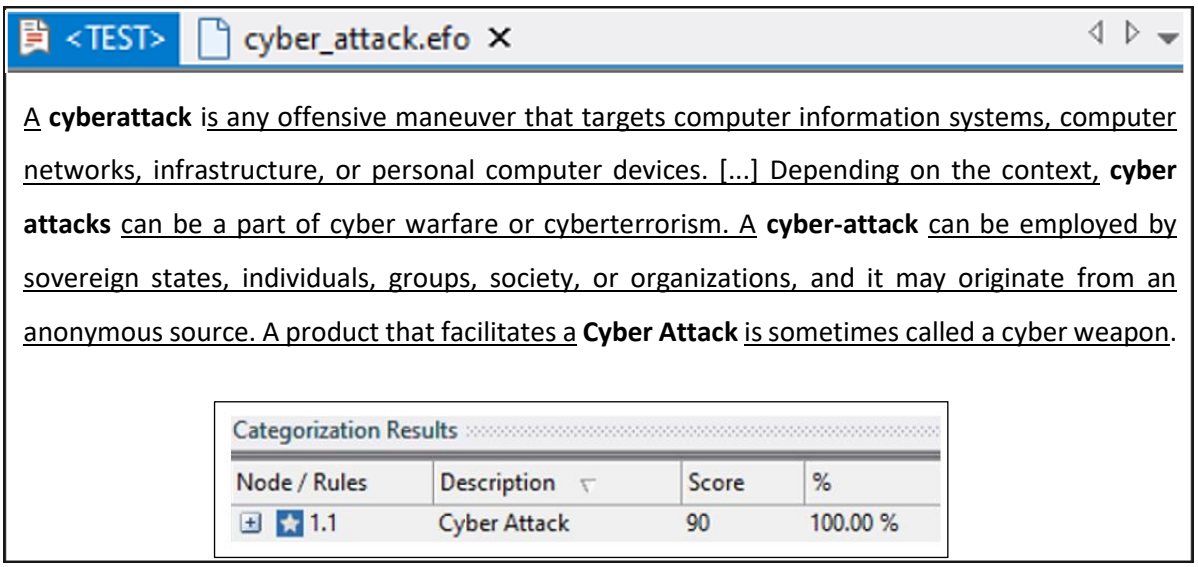



*Figura 30 – Validazione delle regole di categorizzazione*[328]

Dalla figura si può notare che la regola ha restituito un risultato positivo al 100%.

<1.13> | <1.16> Proseguendo con l'analisi della tassonomia *Cyber Illegal Taxonomy* del progetto CTI, ai fini del presente lavoro di tesi di Laurea, presenterò lo sviluppo di altri due nodi tassonomici: *<1.13 – Malware and Virus>* | *<1.16 – Ransomware>*.

La sfida nello sviluppo di questi due nodi concettuale consiste nel far comprendere al sistema che, seppur *ransomware* sia iponimo[329] di *malware*, in questo progetto le due entità devono essere intese come due categorie separate. Dette categorie possono co-esistere e "scattare" all'interno di uno stesso dato testuale – in presenza di *keywords* condivise ai fini del richiamo concettuale – ma è importante che le caratteristiche distintive di ciascuna entità siano sempre ravvisabili. La loro individuazione ed estrazione non deve essere mai ambigua nell'elaborazione di COGITO STUDIO®.

Di seguito, si presentano degli esempi di condizioni formali di categorizzazione per l'individuazione della categoria *<1.13 – Malware and Virus>*.

[328] Fonte dell'*input* testuale: *https://en.wikipedia.org/wiki/Cyberattack;* Fonte dell'immagine: grafico autoprodotto mediante *software* COGITO STUDIO®

[329] Giorgio Sbaraglia, membro del comitato scientifico del CLUSIT, definisce un *ransomware* come segue: *«Un ransomware è un tipo di malware che limita l'accesso del dispositivo che infetta, chiedendo un riscatto (ransom, in inglese) da pagare per rimuovere la limitazione»*. Cfr. https://www.giorgiosbaraglia.it/tag/ransomware/

```
SCOPE SENTENCE
{
      DOMAIN (1.13: HIGH)
      {
      SYNCON (222514) //# 222514: malware, malevolent software, malicious program,
                      malicious software, malfare
      AND
      LEMMA ("cyber", "computer", "computing")
      }
}

SCOPE SENTENCE
{
      DOMAIN (1.13: HIGH)
      {
      SYNCON (127969) // # 127969: computer virus, virus
      AND
      LEMMA ("cyber", "computer", "computing")
      }
}

SCOPE SENTENCE
{
      DOMAIN (1.13: HIGH)
      {
      ANCESTOR (222514) //# 222514: malware, malevolent software, malicious program,
                        malicious software, malfare
      AND NOT
      LEMMA ("ransomware")
      OR
      SYNCON (5950444) //# 5950444: ransomware, Ransomware, cryptotrojan,
                        cryptovirus, extortionware
      }
}

SCOPE SENTENCE
{
      DOMAIN (1.13: NORMAL)
      {
      LEMMA ("malicious")
      AND
      LEMMA ("software")
      }
}
```

*Figura 31 – Regole di categorizzazione del nodo tassonomico <1.13 – Malware and Virus>*[330]

Anche in questo caso, la porzione di testo da cui rilevare l'informazione è circoscritta nel perimetro di una frase (SENTENCE). Ciò significa che, se gli elementi che costituiscono le condizioni linguistiche si trovassero singolarmente su frasi e/o punti differenti del testo, le regole implementate non avranno esito positivo.

[330] Fonte: grafico autoprodotto mediante *software* COGITO STUDIO®

Inoltre, le quattro regole presentate a titolo esemplificativo ravvisano dei punteggi differenti: tre di loro hanno uno *score* HIGH, e l'ultima NORMAL. Questo permetterà al sistema di dar prestigio alle prime, ma non all'ultima, poiché le informazioni contenute nelle prime tre regole sono da considerarsi più rilevanti ai fini della categorizzazione concettuale di un testo *target*.

L'esperto/a linguistico/a stabilisce il livello di prestigio di una condizione anche sulla base del radicamento che una determinata forma linguistica ha subito all'interno di una comunità di parlanti, diventando, per essa, un modello[331].

Seguendo la medesima logica redazionale per la compilazione delle condizioni relative al dominio <1.1>, si stabiliranno, anche per il dominio <1.13>, attributi di tipo SYNCON per richiamare la *Knowledge Base* del *Sensigrafo*® al fine di attivare i collegamenti in essere tra nodi concettuali e le parole ad essi correlate semanticamente.

Si indicheranno, invece, attributi di tipo LEMMA per richiamare tutte le entrate del "grafo dei sensi" presentanti la medesima forma grafica, indipendentemente dal dominio di appartenenza. Si può applicare questo tipo di condizione, qualora la molteplicità dei domini potenzialmente associabili ad una data parola non darà origine ad ambiguità cognitivo-semantiche.

La terza condizione rimanda alla relazione di meronimia sottesa tra le parole *malware* e *ransomware*. L'operatore linguistico potrà applicare l'attributo ANCESTOR al fine di includere nella condizione linguistica – e quindi, indicandoli come *output* eleggibili – tutti gli iponimi di *malware* presenti nel *Sensigrafo*®, ad eccezione sia del LEMMA, che del SYNCON *ransomware*, poiché – ai fini del progetto linguistico CTI, e come definito in precedenza – le due entità devono essere classificate in due categorie differenti.

Viene, così, applicato l'operatore booleano AND NOT seguito dal LEMMA *ransomware* e dal SYNCON relativo a tale entità – quest'ultimo preceduto dall'operatore booleano OR, che stabilisce una presenza alternativa o contemporanea di più espressioni valide nella SENTENCE – e non l'operatore AND, poiché avrebbe stabilito una condizione ineludibile di compresenza tra il LEMMA ed il SYNCON.

[331] Si rimanda al concetto di "bontà dell'esemplare". Nel testo, p. 34, p. 81 e p. 93.

Per quanto concerne il nodo tassonomico *<1.16 – Ransomware>*, essendo un *ransomware* una diramazione di *virus/malware*, il/la linguista dovrà indicare nelle condizioni di categorizzazione l'elemento che lo differenzia dalle entità poste nei suoi confronti in condizione di iperonimia.

Ciò che contraddistingue un *ransomware* da un *malware*, è che il primo rappresenta una circuizione economica: bisogna pagare un riscatto (appunto, un *ransom*) in *bitcoin* o altro tipo di cripto-valuta per poter tornare ad utilizzare il sistema attaccato e/o per evitare una fuga di dati da essi (a seguito di un accesso non autorizzato, con conseguente furto).

È nella terza e nella quarta condizione di categorizzazione, che si stabilisce la necessaria compresenza dell'elemento monetario con l'entità relativa al particolare virus informatico, medianti l'attributo booleano AND e l'attributo di posizione < >, il quale indica una sequenza flessibile tra attributi – all'interno della porzione SENTENCE.

```
SCOPE SENTENCE
{
        DOMAIN (1.16: HIGH)
        {
        LEMMA ("ransomware")
        OR
        SYNCON (5950444) //# 5950444: ransomware, Ransomware, cryptotrojan
                           extortionware
        }
}

SCOPE SENTENCE
{
        DOMAIN (1.16: HIGH)
        {
        SYNCON (5950444) //# 5950444: ransomware, Ransomware, cryptotrojan
                           extortionware
        AND
        LEMMA ("attack")
        }
}

SCOPE SENTENCE
{
        DOMAIN (1.16: NORMAL)
        {
        KEYWORD ("crypto")
        <>
        LEMMA ("malware", "virus", "ransom")
        }
}
```

```
SCOPE SENTENCE
{
        DOMAIN (1.16: NORMAL)
        {
        LEMMA ("ransomware")
        AND
        SYNCON (199508) //# 199508: crypt-, crypto-, krypt-, krypto-
        AND
        LEMMA ("currency")
        }
}
```

*Figura 32 – Regole di categorizzazione del nodo tassonomico <1.16 – Ransomware>*[332]

Ai fini della validazione delle regole create ed implementate a sistema, si uniranno nella stessa pagina di controllo le definizioni di *malware* e di *ransomware* disponibili su *Wikipedia*© in lingua inglese. Le *keywords* che fanno "scattare" la regola di categorizzazione sono messe in rilievo mediante il formato grassetto; la porzione di testo in cui è presente il macro-argomento estratto viene sottolineato dal sistema.

<TEST> malware_virus.efo ×

Ransomware is a type of malware from cryptovirology that threatens to publish the victim's personal data. [...] The **virus** writer can effectively hold all of the money ransom until half of it is given to him. Even if the e-money was previously encrypted by the user, it is of no use to the user if it gets encrypted by a cryptovirus.

***

**Malware** (a portmanteau for **malicious software**) is any software intentionally designed to cause disruption to a computer, server, client, or computer network, leak private information, gain unauthorized access to information or systems, deprive users access to information or which unknowingly interferes with the user's computer security and privacy. By contrast, software that causes harm due to some deficiency is typically described as a software bug.

Categorization Results

| Node / Rules | Description | Score | % |
|---|---|---|---|
| 1.13 | Malware and virus | 150 | 55.56 % |

*Figura 33a – Validazione delle condizioni relative al nodo tassonomico <1.13 – Malware and Virus>*[333]

[332] Fonte: grafico autoprodotto mediante *software* COGITO STUDIO®
[333] Ibidem.

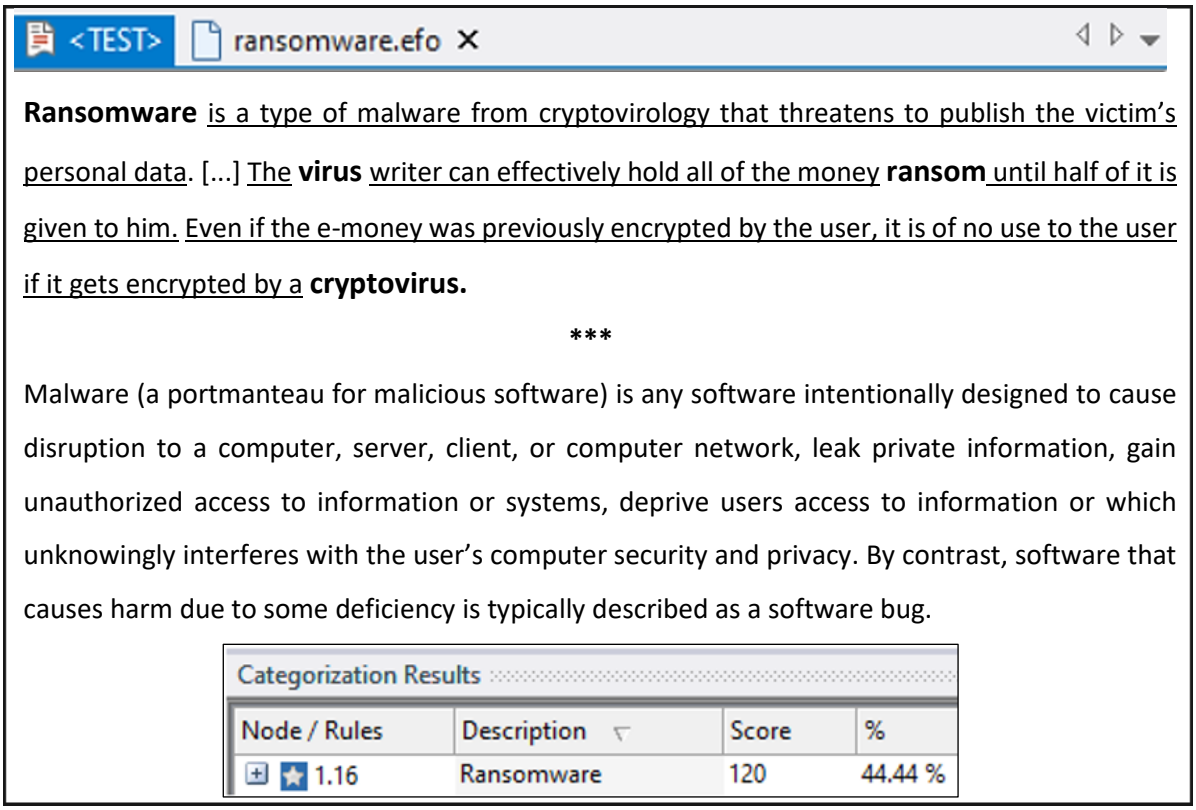



*Figura 33b – Validazione delle condizioni relative al nodo tassonomico <1.16 – Ransomware>*[334]

Come si può notare, il nodo <1.16> relativo ai *ransomware* non fa "scattare" le porzioni di testo in cui è contenuta l'informazione relativa ai *malware*; allo stesso modo, la categoria *<1.13 – Malware and Virus>* non fa "scattare" la prima SENTENCE: seppur sia presente la parola *malware*, essa si trova in una condizione di compresenza nella stessa frase con la parola *ransomware*.

Al fine di far comprendere al sistema che – nonostante il rapporto di gerarchia semantica in essere tra i due nodi concettuali – *malware* e *ransomware* avrebbero dovuto essere intese come due categorie nettamente differenti, ha rivestito una decisiva importanza l'inserimento nelle regole linguistiche la condizione di impossibilità di compresenza tra i due termini nella stessa frase.

[334] Ibidem.

## *4.4 Modulo di estrazione*

Le condizioni linguistiche elaborate su COGITO STUDIO® mireranno al raccoglimento di informazioni dal web e da fonti di varia natura, una volta che il pacchetto linguistico generato dal sistema[335] *(General Scripting Language* – GSL, o *Language Package* – LPK) verrà implementato su di un *software* di *Decision and Continuous Intelligence.* La creazione di dette condizioni dev'essere operata dagli analisti di *Intelligence* con (meta)cognizione e lungimiranza strategica, al fine di non interfacciarsi con risultati devianti – in gergo, definiti "sporchi" – in fase di *Data Filtering*.

L'estrazione linguistica (o "estrazione terminologica" o, ancora, "estrazione della conoscenza"), come già definito in precedenza, permette l'identificazione (semi)automatica di parole chiave rilevanti in un dato *corpus* testuale.

Ogni regola linguistica redatta va intesa, pertanto, come strumento per raggiungere o discriminare rigidamente alcuni significati, dopo aver definito un contesto di azione.

Così come le condizioni di categorizzazione sono indissolubilmente legate alla tassonomia implementata nel progetto linguistico – che sia essa stata creata *ex novo,* seguendo i desiderata del Cliente, o che sia la rappresentazione di modello standard di classificazione di concetti – le regole di estrazione sono rese possibili solo dopo aver definito uno o più *template(s).*

Realizzare un *template* significa *«definire una struttura dati che può contenere uno o più valori»*[336] e può essere inteso come una tabella macro-concettuale al cui interno saranno presenti sottogruppi semantici (denominati *fields* – definiti anche come *labels* in altri *tool* di semantica – ovvero, delle denominazioni con le quali l'informazione estratta viene "etichettata") in cui verranno raggruppate le informazioni da ricercare mediante puntuali *keywords* o *patterns*.

Nella redazione di regole di estrazione, quindi, si riscontra – ancora una volta – un approccio metodologico di analisi di tipo *top-down*: dal macro-concetto, alla singola informazione sotto forma di parola chiave.

[335] Generato o meno con configurazione crittografata (in formato *.EFX).
[336] D. Bedogni, *Progettazione e Sviluppo di un Sistema di Risposta Automatico per la Richiesta di Informazioni Riguardanti i Servizi Ferroviari,* Università degli Studi di Modena e Reggio Emilia, 2014, p. 21

La struttura per la realizzazione di condizioni formali nel linguaggio "E" di COGITO STUDIO®, pertanto, presenta una matrice di stampo gerarchico, all'interno della quale i vari costituenti sono interrelati mediante una relazione di meronimia (PART-OF) – paragonabile ad uno schema in stile *matrioska*, come rappresentato nella seguente figura:

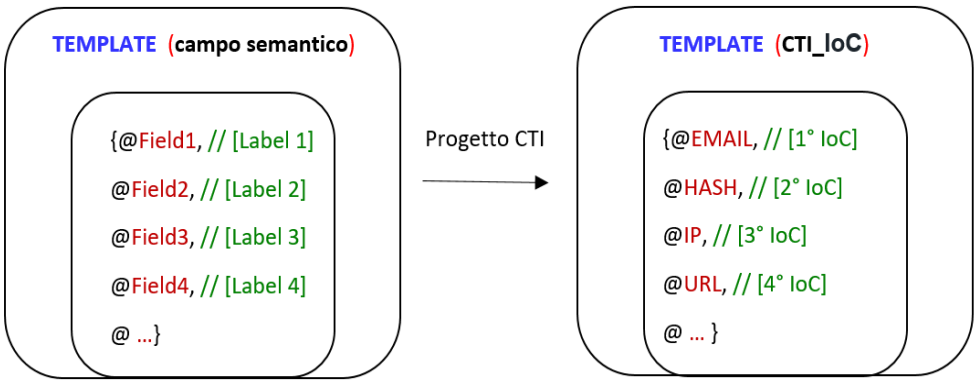



*Figura 34a – Template(s) e Fields*[337]

Nel progetto CTI, l'organizzazione di *templates* e *fields* è stata realizzata come segue:

1.

```
TEMPLATE(CYBERTHREAT)
{
    @ATTACK_PATTERN,
    @LOCATION, // estrazione di luoghi
    @MALWARE, // (include ransomware)
    @THREAT_ACTOR, // APT
    @TOOL,
    @COURSE_OF_ACTION, // COA
    @VULNERABILITY, //CVE + CWE
    @CAMPAIGN, // nomi campagne di attacchi
    @REPORT_PUBLISHED, // estrazione di date
    @IDENTITY_ORGANIZATION, // organizzazioni IT
    @IDENTITY_PERSON, // identità Threat Actors
    @INFRASTRUCTURE // Botnet
}
```

2.

```
TEMPLATE(CYBERTHREAT_IOC) // A tutti è stata applicata anche Anonimizzazione [.]
{
    @INDICATOR_EMAIL,
    @INDICATOR_HASH,
    @INDICATOR_IP, // IPV4 + IPV6
    @INDICATOR_WIN_REG_KEY,
    @INDICATOR_URL,
    @INDICATOR_DOMAIN
}
```

*Figura 34b – Template(s) e Fields del progetto "CTI"*[338]

[337] Fonte: grafico autoprodotto.
[338] Fonte: grafico autoprodotto mediante *software* COGITO STUDIO®

Ogni condizione linguistica, quindi, farà riferimento ad un determinato *field* che, a sua volta, richiamerà un *template*. Non sarà possibile associare più *templates* all'interno della stessa condizione linguistica, bensì, sarà possibile richiamare più *fields* appartenenti allo stesso *template*.

All'interno del progetto CTI sono stati realizzati due *templates*, intesi come contenitori di macro-concetti, e diciotto *fields*. Entrambe le categorie fanno riferimento al linguaggio STIX: *Structured Threat Information eXpressions*[339], che ha fornito degli standard di comunicazione che hanno garantito stabile comprensione tra gli sviluppatori del progetto linguistico e i Clienti (tra loro, anche esperti di dominio), futuri fruitori del servizio.

Analizzando la struttura dei due *templates* realizzati, una prima suddivisione è stata realizzata sulla base della macrocategoria a cui appartengono le singole entità:

1. il primo *template ("Cyberthreat")* è costituito dagli *STIX Domain Objects (SDO)*, ovvero dai modelli di dati e componenti informatiche legati tra loro da articolate relazioni, espresso attraverso lo standard *STIX Relationship Objects (SRO)*[340]. Gli autori di *HackInBio*[341] definiscono gli *SDO* come segue: *«These data models provide the capability to fully express information about their targeted conceptual area»;*
2. il secondo *template ("Cyberthreat_IoC")* racchiude gli indicatori di compromissione *(Indicators of Compromise – IoCs)*, ovvero, gli «*artefatti forensi di un'intrusione che possono essere identificati su di un host o una rete. Mediante l'utilizzo di tali indicatori, tutti i dettagli relativi ad un incidente informatico diventano condivisibili con altre organizzazioni»*[342].

Nella redazione delle regole linguistiche relative a questo secondo *template*, l'estrazione degli *IoCs* è stata resa possibile mediante *RegEx* PERL, implementate nel motore semantico COGITO STUDIO® mediante l'attributo PATTERN.

Essendo un progetto di estrazione mirata di eventi, strumenti e attori delle minacce, nel fruire del modulo *AI-based* di COGITO STUDIO®, ha rivestito un ruolo cruciale la

[339] Tale linguaggio ha come obiettivo, quindi, la standardizzazione delle informazioni relative al dominio della *Cyber Threat Intelligence*.
[340] Come verrà descritto nel paragrafo 4.5 *"Modulo di estrazione delle relazioni tra entità".*
[341] *HackInBo* è la più grande conferenza sulla sicurezza informatica in Italia. https://www.hackinbo.it/index.php
[342] Traduzione della citazione presente in: *Dalla Malware Analysis alla CTI Sharing*, di *HackInBio*, 2017. Cfr. www.hackinbo.it/slides/1508354164_SCHIFILLITI_Slides_HackinBO2017_DraftVer05.pdf

possibilità di consultare il *Sensigrafo*® ed incorporare in esso nuovi dati attraverso la sua *feature* di *editing*.

La raccolta e l'inserimento delle informazioni relative a *malware* e di altro tipo di virus informatico[343], ad *hackivists/hackers* e alla loro appartenenza a gruppi e/o campagne di attacco, hanno infatti permesso al "grafo dei sensi" di elaborare e migliorare la propria *knowledge*, ai fini del raggiungimento del completo riconoscimento automatizzato delle diverse entità sulla base della loro posizione, ruolo semantico e funzione sintattica rivestita all'interno di un testo.

Mediante la funzionalità *Sensigrafo Editor*®, quindi, si è potuto procedere all'inserimento degli *items* a disposizione del *team* – previa classificazione delle loro varianti grafico-linguistiche, anche in vista della successiva (indispensabile) redazione di regole di normalizzazione[344] – fornendo descrizioni puntuali su ogni nuova *entry*.

Ciascuna di esse è stata associata ad una parte del discorso, a uno o più domini di appartenenza, nonché a *parent syncons*, al fine di realizzare una stabile (ma plastica) struttura semiotica del nuovo *corpus* linguistico immesso. Infine, un altro *core element* del *Sensigrafo*® di COGITO STUDIO® è il motore di realizzazione della struttura meronimica tra *syncons*. Ciò permette una gerarchizzazione ed una classificazione meticolosa dell'informazione.

Il *Sensigrafo Editor*® permette, quindi, di potenziare e migliorare la rete semantico-cognitiva del sistema con la *knowledge* personale dell'analista linguistico/a. L'elemento umano rimane, in questo sistema intelligente, un elemento cruciale nell'elaborazione del linguaggio naturale, in grado di regolamentare *output* in cui sono presenti *bias* linguistico-culturali, a loro volta, intrinsechi negli algoritmi che ne realizzano il funzionamento.[345]

[343] Le presenti informazioni sono in continuo aggiornamento, considerate le continue evoluzioni degli attacchi informatici, anche in relazione agli attuali scontri russo-ucraini/globali.
[344] Come descritto nel paragrafo 4.4.1 *"Regole di normalizzazione".*
[345] Nel progetto CTI sono stati ravvisati alcuni *bias* nell'elaborazione dell'aggettivo *positive*, contestualizzato in *positive risk* o *positive* nel senso di *affetto dalla minaccia*, in cui si indicava una polarità ed un *sentiment* – appunto – favorevole all'azione o alla situazione espressa.

| Syncon | ID |
|---|---|
| 3PARA RAT, 3para RAT, 3para rat, 3 PARA RAT, 3 para RAT, 3 para rat | 2006132 |
| 4H RAT, 4h RAT, 4h rat | 2006133 |
| abdupd, ABDUPD, Abdupd | 2006135 |
| ABK, abk | 2006134 |
| AdFind, Adfind, adFind, adfind, ADFIND | 2006136 |
| ADROIDOS_ANSERVER.A, Adroidos_Anserver.A, Adroidos_anserver.A, adroidos_anserver.A, adroidos... | 2006147 |
| Adups, adups, ADUPS | 2006137 |
| ADVSTORESHELL, Advstoreshell, ADVSTORESHELLE | 2006138 |
| Agent Smith, AgentSmith, Agent smith, agent smith, AGENT SMITH | 2006139 |
| Agent Tesla, AgentTesla, Agent tesla, agent tesla, AGENT TESLA | 2006140 |
| Agent.btz, Agent.Btz, Agent.BTZ, agent.btz, agent.Btz, agent.BTZ, AGENT.BTZ Agent btz, Agent Btz, ... | 2006141 |
| Allwinner, allwinner, ALLWINNER | 2006142 |
| Anchor, anchor, ANCHOR | 2006143 |
| AndoRAT, andoRAT, ANDORAT, Andorat, andorat | 2006148 |
| Android/AdDisplay.Ashas, Android/addisplay.Ashas, android/addisplay.ashas, Android/adDisplay.A... | 2006144 |
| Android/Chuli.A, Android/chuli.A, android/chuli.A, android/chuli.a, Android Chuli.A, Android chuli... | 2006145 |
| AndroidOS/MalLocker.B, AndroidOS/malLocker.B, AndroidOS/mallocker.B, androidOS/mallocker.B... | 2006146 |
| Anubis, ANUBIS, anubis | 2006149 |

| ***Parent syncon(s)*** | ***Part of speech*** | ***Domain[1]*** | ***Domain[2]*** | ***Definition*** |
|---|---|---|---|---|
| Malware, malevolent software, malicious program | Noun | Internet 60% | Crime 40% | Malware typology |

***Figura 35 – Elenco di entità di tipo malware inserite manualmente nel Sensigrafo Editor®[346]***

Alcuni degli attributi implementabili nelle condizioni linguistiche di COGITO STUDIO®, richiamando le reti semantiche del *Sensigrafo*®, permettono la (semi)automatizzazione del processo di estrazione di un'informazione all'interno di un testo, nonché la verifica, in seconda istanza, il *matching* tra il codice realizzato e la porzione di testo di interesse – validandone così l'*output* di *mining.*

Ad esempio, nel progetto CTI, nell'elaborazione del *field* @REPORT_PUBLISHED, attinente al *template "Cyberthreat",* si è implementato l'attributo TYPE associandolo alla *entity* [DAT=date], in una porzione di testo di tipo SENTENCE, al fine di riuscire a rilevare automaticamente ogni data e/o riferimento temporale da un testo.

[346] Fonte: grafico autoprodotto mediante *feature Sensigrafo Editor*® di COGITO STUDIO®

```
SCOPE SENTENCE

{ IDENTIFY(CYBERTHREAT)

      { @REPORT_PUBLISHED [TYPE (DAT)]
      }
}
```

*Figura 36 – Regola di estrazione mediante attributo TYPE (DAT)*[347]

Inoltre, in un'altra condizione formale, è stato utilizzato il TYPE PATTERN per l'estrazione di una stringa contenente un'informazione a carattere temporale, risultante, questa volta, in un dato strutturato. La data verrebbe riconosciuta sia qualora essa venisse espressa nel formato [giorno/mese/anno], utilizzato maggiormente in Italia e gran parte di Europa, Africa, Asia e Oceania – oppure mediante il formato [mese/giorno/anno], utilizzato maggiormente in USA e Canada.

```
SCOPE SENTENCE
{
      IDENTIFY(CYBERTHREAT)
      {
            @REPORT_PUBLISHED[PATTERN("^(0?[1-9]|[12][0-9]|3[01])[\/\-
            \.](0?[1-9]|1[012])[\/\-\.]\d{4}$")]
      }
}
```

*Figura 37 – Regola di estrazione mediante attributo PATTERN*[348]

Si fornisce come *input* il seguente testo per la validazione delle regole implementate:

**«The Top Cyber Attacks of November 2021. November's roster of data breaches is an excellent illustration on that point: a mix of surprising methods, unusual motivations, and one old-fashioned data heist on one of the internet's most tempting targets. A major data breach was first identified on November 17. Date of Attack: November 3, 2021 (11/03/2021)».**[349]

[347] Fonte: grafico autoprodotto mediante *software* COGITO STUDIO®
[348] Ibidem.
[349] Rielaborazione dell'articolo di Artic Wolf, *The Top Cyber Attacks of November 2021*, 2021. Cfr. https://arcticwolf.com/resources/blog/top-cyber-attacks-november-2021

COGITO STUDIO® restituisce i seguenti risultati:

| EXTRACTION RESULTS: RECORDS/FIELDS | INPUT TEXT <TEST> |
|---|---|
| CYBERTHREAT<br>MALWARE Cyber Attacks | <TEST> threat_actor.efe ×<br>The Top **Cyber Attacks** of November 2021. |
| CYBERTHREAT<br>REPORT_PUBLISHED 11/03/2021 | <TEST> report_published.efe ×<br>Date of Attack: November 3, 2021 **(11/03/2021)** |
| CYBERTHREAT<br>REPORT_PUBLISHED 2021 | <TEST> report_published.efe ×<br>• The Top Cyber Attacks of November **2021.**<br>• Date of Attack: November 3, **2021** (11/03/**2021)** |
| CYBERTHREAT<br>REPORT_PUBLISHED Nov-17 | <TEST> report_published.efe ×<br>A major data breach was first identified on **November 17.** |
| CYBERTHREAT<br>REPORT_PUBLISHED Nov-2021 | <TEST> report_published.efe ×<br>The Top Cyber Attacks of **November 2021.** |
| CYBERTHREAT<br>REPORT_PUBLISHED Nov-3-2021 | <TEST> report_published.efe ×<br>Date of Attack: **November 3, 2021 (11/03/2021)** |
| CYBERTHREAT<br>REPORT_PUBLISHED November | <TEST> report_published.efe ×<br>• **November**'s roster of data breaches is an excellent illustration…<br>• The Top Cyber Attacks of **November** 2021.<br>• Date of Attack: **November** 3, 2021 (11/03/2021) |

*Figura 38 – Validazione delle condizioni di estrazione nel progetto "CTI"*[350]

Le condizioni linguistiche hanno restituito una risposta positiva.

Proseguendo nella presentazione di modelli di estrazione linguistica implementati su a sistema per la realizzazione del progetto "CTI", mediante l'utilizzo dell'attributo LIST, è

[350] Fonte: grafico autoprodotto mediante *software* COGITO STUDIO®

stato, inoltre, possibile associare ad un qualsiasi *field* una lista di *entities* esterne al *Sensigrafo*® – lista redatta manualmente dall'operatore linguistico – per il richiamo puntuale, a *patterns* e/o codici alfanumerici difficilmente inscrivibili (e non indispensabili) nella rete semantica del sistema.

È stato questo il caso del *field* @ATTACK_PATTERN, attinente al *template "Cyberthreat".* Gli *attack patterns* sono un insieme di metodi per trovare *bug* o errori nel codice nei sistemi. Forniscono, fisicamente o in riferimento, il modello di soluzione comune per prevenire l'attacco. Il sito *Cyber Security 360* definisce gli *attack patterns* come *«una tipologia di TTP (Tactics, Techniques and Procedures) che descrive le modalità con cui gli attori delle minacce tentano la compromissione dei target».*[351]

Grazie all'esperto di dominio presente nel *team*, e facendo come sempre riferimento agli standard STIX, si è creata una lista contenente i più rilevanti cyber *attack patterns*.

Di seguito, se ne presenta un esempio:

| **"attack_pattern.txt"** | |
|---|---|
| Don't Fragment Bit Echoing Probe | Accessing Functionality Not Properly Constrained By Acls |
| Absolute Path | Add Malicious File To Shared Webroot |
| Operation Wizardopium | Accessing Functionality Not Properly Constrained By Acls |
| Internal Reconnaissance | Adversary In The Middle |
| Absolute Path Traversal | Adding A Space To A File Extension |
| Absolute Traversal | Adversary In The Middle |
| Accessing Functionality | Adding A Space To File |
| Accessing Functionality Not Properly Constrained | Adversary In The Browser |
| Aitb | Ajax Footprinting |
| Aitm | Alteration Of A Software Update |
| Altered Component Firmware | Altered Installed Bios |
| Alternative Execution Due To Deceptive Filenames | Accessing/Intercepting/Modifying Http Cookies |
| Account Footprinting | SOCKS Proxy Server |
| Action Spoofing | Active Os Fingerprinting |

***Figura 39 – Lista "attack pattern.txt"***[352]

[351] www.cybersecurity360.it/soluzioni-aziendali/cyber-threat-intelligence-e-condivisione-delle-informazioni-conoscere-le-minacce-per-prevenirle/

[352] Fonte: grafico autoprodotto.

Nel paragrafo 3.7 *"RegEx – PERL Regular Expressions Syntax"* ho presentato il processo di estrazione dei *fields* del *template "Cyberthreat_Ioc"* – nello specifico, lo sviluppo di condizioni formali utilizzando il linguaggio "E" di COGITO STUDIO® per il *mining* di dati (anonimizzati o meno) relativi a: IP, URL, email, domini e *hash*.

Nella redazione di dette regole linguistiche, è stato applicato a tutti i *fields* sopra menzionati l'attributo PATTERN, seguito da una condizione di tipo *RegEx* PERL.

Allo stesso modo, sono state implementate altrettante espressioni regolari per quanto riguarda i *fields* contrassegnati come: @WIN_REG_KEY (*Windows Registry Key* – Registro di sistema di Windows); @COURSE_OF_ACTION (un insieme di misure che possono essere prese sia in risposta a un attacco che come misura preventiva precedente ad un attacco); @VULNERABILITY (un *bug* nel *software* che permette l'accesso non autorizzato ad un sistema o ad una rete da parte di un *hacker*).

Per quanto riguarda, invece, l'estrazione dei *fields* @MALWARE e @THREAT_ACTOR, si è proceduto con una modalità mista di attributi ed operatori all'interno delle condizioni, includendo così nella metodologia di *mining*, sia l'utilizzo di attributi SYNCON ed ANCESTOR, che il richiamo a liste create *ad hoc* (come si è mostrato nel paragrafo 3.6 *"Linguaggio 'E' – Estrazione"*, la lista "malware.txt").

Si presentano di seguito degli esempi di condizioni formali applicate al progetto "CTI" aventi come *target* l'estrazione delle informazioni concernenti i *fields* @MALWARE, @THREAT_ACTOR e @CAMPAIGN:

```
1.
SCOPE PARAGRAPH
{ IDENTIFY (CYBERTHREAT)
     { @MALWARE [ANCESTOR(222514)]// # 222514: malware, malevolent
     software, malicious program, malicious software, malfare
     AND
       @THREAT_ACTOR [KEYWORD ("threat actor", "threat actors", "Threat
        actors", "Threat actor", "Threat Actor", "Threat Actors")]
     OR
     @CAMPAIGN [LEMMA ("campaign")]
     }
}
```

```
2.
SCOPE SENTENCE
{ IDENTIFY (CYBERTHREAT)
      {      @CAMPAIGN [LEMMA ("campaign")]
             AND
             KEYWORD (EXPAND "campaign.txt")
      }
}

3.
SCOPE SENTENCE
{     IDENTIFY(CYBERTHREAT)
      {      @THREAT_ACTOR [KEYWORD ("threat group", "Threat Group",
                                     "Threat group")]
             AND
             KEYWORD (EXPAND "apt.txt")
      }
}
```

*Figura 40 – Regole di estrazione di tre fields del progetto "CTI"*[353]

Si utilizzerà il seguente testo per la validazione delle regole sopra indicate:

**«Researchers have revealed details of a threat actor that has targeted thousands of organizations globally with over a dozen different commodity malware payloads since at least 2017. The threat group, which researchers with Proofpoint labeled TA2541, has targeted the aviation, aerospace, transportation, manufacturing and defense industries with remote access trojans (RAT) that have the capability to remotely control compromised machines. The threat group has recurring targets in North America, Europe and the Middle East. The activity of the threat group named as TA2541 has previously been publicly reported by various security researchers since 2019.**

**For instance, in September Cisco Talos analyzed a campaign targeting the aviation industry, which it linked to an actor that has been running malware campaigns for more than five years. Proofpoint researchers said this is the first time all of this comprehensive data is being shared under one threat activity cluster».**[354]

[353] Fonte: grafico autoprodotto mediante *software* COGITO STUDIO®

[354] O'Donnell, Welch, *Threat Actor Targets Transportation Firms In Malware Campaigns*, 2022. Cfr. https://duo.com/decipher/threat-actor-targets-transportation-firms-in-malware-campaigns

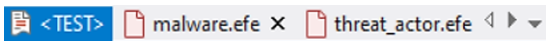
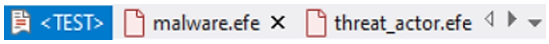
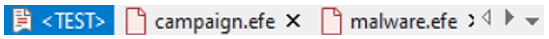
| EXTRACTION<br>RESULTS:RECORDS/FIELDS | INPUT TEXT <TEST> |
|---|---|
|  |  <br>Researchers have revealed details of a **threat actor** that has targeted thousands of organizations globally with over a dozen different commodity **malware** payloads since at least 2017. |
|  |  <br>The **threat group,** which researchers with Proofpoint labeled **TA2541,** has targeted the aviation, aerospace, transportation, manufacturing and defense industries with remote access **trojans (RAT)** that have the capability to remotely control compromised machines. **The threat group** has recurring targets in North America, Europe and the Middle East. The activity of the **threat group** named as **TA2541** has previously been publicly reported by various security researchers since 2019. |
|  |  <br>For instance, in September Cisco Talos analyzed a **campaign** targeting the aviation industry, which it linked to an actor that has been running **malware campaigns** for more than five years. |

*Figura 41 – Validazione delle regole di estrazione del progetto "CTI"*[355]

La validazione ha restituito un *matching* positivo tra le condizioni formali implementate ed il dato testuale fornito in *input.*

[355] Fonte: grafico autoprodotto mediante *software* COGITO STUDIO®

### *4.4.1 Regole di normalizzazione*

In matematica, così come in informatica, con il verbo transitivo "normalizzare" si intende «*il procedimento di dividere tutti i termini di un'espressione per uno stesso fattore, in modo che l'espressione risultante abbia una certa norma*».[356]

La normalizzazione è un procedimento atto al raggiungimento dell'uniformità tra un dato – rappresentante il valore di base (in questo caso, una determinata forma linguistica) – e le sue deviazioni di forma, scelte da un utente. Questa azione punta all'eliminazione della ridondanza informativa, nonché alla preventiva risoluzione di conflitti in caso di incoerenza tra i dati presenti nel *database*; altresì, permette una metodica organizzazione del *dataset* di interesse.

Esistono varie condizioni per normalizzare i dati, nonché per validarne la qualità sulla base allo schema del *database* realizzato. La regolamentazione delle forme linguistiche non costituisce una vera e propria metodologia di progettazione, bensì, permette di conformare un *dataset* e di verificare gli *output* della progettazione eseguita *ex ante*.

All'interno del sistema di COGITO STUDIO® questo processo è reso possibile mediante l'*upload* di file *.txt al cui interno vi sono contenute condizioni di normalizzazione di singole entità o di stringhe di *tokens* – che realizzano significati complessi – espressi mediante una pluralità di varianti linguistiche all'interno del progetto creato.

Pertanto, il motore semantico avvierà la normalizzazione uniformando la rappresentazione grafica di uno o più *token(s)*, ovvero dei vari significanti di uno stesso referente. L'operatore assemblerà in un'unica stringa di testo una determinata forma linguistica che verrà decretata come "bontà dell'esemplare – BdE"[357], seguita da tutte le sue possibili varianti grafico-linguistiche, affinché l'estrazione di quel dato *token* – o insieme di *tokens* – abbia come unica visualizzazione la forma grafica della BdE[358].

Questa modalità di regolamentazione dei dati linguistici non va a "forzare" la rete semantica che è alla base del *framework* semantico, ovvero, non appone modifiche o

[356] https://www.treccani.it/vocabolario/normalizzazione/
[357] Elaborata da Eleanor Rosch e Ludwig Wittgenstein, afferente alla teoria del prototipo.
[358] All'interno del *Sensigrafo*®, la variante linguistica di riferimento di ogni entrata nel dizionario viene formalmente definita e realizzata con la funzione SMARTENTRY.

deviazioni agli algoritmi di intelligenza artificiale che rendono possibile la comprensione (meta)semantica del *Sensigrafo*®. Non vengono, pertanto, intaccati i significati o i domini di appartenenza di un dato *token*.

Le condizioni formali di normalizzazione presentano un uso differente della punteggiatura da quello canonico/informale: ad esempio, il punto non definisce il termine di una stringa, così come la virgola o il trattino non delimitano un *token* da un altro; tutti i caratteri diversi dalla barra verticale (o *pipe:* **|** ) vengono intesi come parte integrante di un *token* – ovvero, come espressione grafica del suo significato.

Quindi, nello schema di configurazione regolativa, al fine di poter comunicare a COGITO STUDIO® l'inizio e la fine di una variante linguistica, si utilizza il *pipe:* |.

Il motore semantico andrà a "fotografare" ogni stringa e decreterà il *token* posto antecedente al simbolo = come BdE, ovvero come modello, unica espressione grafica di significato da considerare in fase di visualizzazione del dato estratto.

Si presentano di seguito alcune delle normalizzazioni contenute all'interno del file *.txt implementato nel progetto linguistico *"CTI – Cyber Threat Intelligence per il rilevamento di minacce informatiche".* Le presenti regole hanno interessato, soprattutto, varie classificazioni di *malware, virus, tool, APT*[359] e *Threat Actors.*

In grassetto, ai fini del presente lavoro di tesi, vengono mostrate le varianti linguistiche individuate come BdE, ovvero, come modello di estrazione.

[359] *«La sigla APT, acronimo di Advanced Persistent Threat, indica una tipologia di attacchi mirati e persistenti portati avanti da avversari dotati di notevole expertise tecnico e grandi risorse»*. Cfr. https://www.cybersecurity360.it/nuove-minacce/minacce-apt-cosa-sono-le-advanced-persistent-threat-come-funzionano-e-come-difendersi/

| **“Normalizzazioni_Cyberthreat.txt”** |
|---|
| **Ajax Security Team=**Ajax security team\|ajax security team\|Ajax Security\|ajax security\|Operation Woolen-Goldfish\|operation woolen-goldfish\|Operation woolen-goldfish\|OPERATION WOOLEN-GOLDFISH\|Operation Woolen Goldfish |
| **APT-C-36=**apt-C-36\|APT C 36\|apt C 36\|APT-c-36\|apt-c-36\|APT c 36\|apt c 36 |
| **Cleaver=**cleaver\|CLEAVER\|Threat Group 2889\|THREAT GROUP 2889\|threat group 2889\|Threat-Group-2889\|THREAT-GROUP-2889\|treat-group-2889\|Threat-group-2889 |
| **Cobalt Group=**Cobalt group\|cobalt group\|COBALT GROUP\|GOLD KINGSWOOD\|Gold Kingswood\|gold kingswood\|Gold kingswood\|Cobalt Gang\|Cobalt gang\|cobalt gang |
| **Cyber Attack=**Cyber Attacks\|cyber attack\|cyber attacks\| CYBERATTACK\|CYBERATTACKS\|CYBER ATTACK\|CYBER ATTACKS\|CYBER-ATTACK\|CYBER-ATTACKS\|cyberattack\|cyberattacks\|Cyber-attack\|Cyber-attacks |
| **Cyber Threat=**Cyber Threats\|cyber threat\|cyber threats\| CYBERTHREAT\| CYBERTHREATS\| CYBER THREAT\|CYBER THREATS\|CYBER-THREAT\|CYBER-THREATS\|cyberthreat\|cyberthreats\|Cyber-threat\|Cyber-threats\|Cyber-Threat\|Cyber-Threats\|cyber-threat\|cyber-threats |
| **Deep Panda=**DeepPanda\|Deep panda\|deep panda\|DEEP PANDA\|Shell Crew\|Shell crew\|shell crew\|SHELL CREW\|WebMasters\|Web Masters\|webmasters\|Webmasters\|web masters\|WEB MASTERS\|WEBMASTERS\|KungFu Kittens\|Kungfu kittens\|kungfu kittens\|KUNGFU KITTENS |
| **Dragonfly=**dragonfly\|DRAGONFLY\|TG-4192\|tg-4192\|TG4192\|tg4192\|TG 419\|tg 4192\|Crouching Yeti\|Crouching yeti\|crouching yeti\|CrouchingYeti\|Crouching Yeti\|CROUNCHINGYETI\|CROUNCHING YETI\|IRON LIBERTY\|Iron liberty\|Iron Liberty\|iron liberty |
| **FIN6=**FIN 6\|fin6\|fin 6\|Magecart Group 6\|Magecart group 6\|magecart group 6\|MAGECART GROUP 6\|Magecart Group6\|Magecart group6\|magecart group6\|MAGECART GROUP6 |
| **Lazarus Group=**LazarusGroup\|Lazarus group\|lazarus group\|LAZARUS GROUP\|HIDDEN COBRA\|Hidden Cobra\|Hidden cobra\|hidden cobra |
| **MuddyWater=**Muddywater\|Muddy Water\|Muddy water\| muddyWater\| MUDDYWATER\| |
| **Patchwork=**patchwork\|PATCHWORK\|Hangover Group\|Hangover group\|hangover group\|HANGOVER GROUP\|Dropping Elephant\|Dropping elephant\|dropping elephant |
| **Threat Actor=**Cyber Threat Actors\|cyber threat actor\|cyber threat actors\|threat actor\|threat actors\|Threat Actor\|Threat Actors\|actor\|actors\|Actor\|Actors\| |
| **Threat Group-1314=**Threat group-1314\|threat group-1314\|THREAT GROUP-1314\|ThreatGroup-1314\|Threatgroup-1314\|threatgroup-1314 |
| **Turla=**turla\|TURLA\|Group 88\|group 88\|GROUP 88\| Group88\| group88\|GROUP88 |
| **Trojan.Karagany=**Trojan.karagany\|trojan.Karagany\|trojan.karagany\|TROJAN.KARAGANY\| Trojan Karagany\|Trojan karagany\|trojan Karagany\|trojan karagany |

Ai fini della validazione *(testing)* delle regole di normalizzazione implementate nel progetto “CTI”, si presenta di seguito un esempio di visualizzazione delle estrazioni di un testo fornito in *input*. La porzione di testo scelta per eseguire il presente *testing* è una rielaborazione dell’articolo: *“From Ransomware to DDoS: Guide to Cyber Threat Actors – How, Why, and Who They Choose to Attack”*, della redazione di “*flashpoint.io”*[360].

[360] https://www.flashpoint-intel.com/blog/guide-to-cyber-threat-actors/

**«What do Cyber Threat Actors want, such as the Lazarus Group? Money, mostly. These actors who execute cyberattacks, can wreak havoc on organizations across the private and public sectors. Hidden cobra's cyber attacks put their reputation and customers at stake».**

| EXTRACTION RESULTS | INPUT TEXT <TEST> | NORMALIZED OUTPUT |
|---|---|---|
| CYBERTHREAT<br>└ MALWARE<br>cyber attacks | <TEST> cyber_attack.ef<br>Hidden cobra's **cyber attacks** put their reputation and customers at stake. | **EXTRACTION TEMPLATE=CYBERTHREAT FIELD NAME= MALWARE OUTPUT=Cyber Attack** |
| CYBERTHREAT<br>└ MALWARE<br>cyber attacks | These actors who execute **cyberattacks**. | |
| CYBERTHREAT<br>└ THREAT_ACTOR<br>Threat Actors | <TEST> threat_actor.efe<br>What do Cyber **Threat Actors** want, such as the Lazarus Group? | **EXTRACTION TEMPLATE=CYBERTHREAT FIELD NAME= THREAT_ACTOR OUTPUT=Threat Actor** |
| CYBERTHREAT<br>└ THREAT_ACTOR<br>Cyber Threat Actors | What do **Cyber Threat Actors** want, such as the Lazarus Group? | |
| CYBERTHREAT<br>└ THREAT_ACTOR<br>Hidden cobra | <TEST> threat_actor.efe<br>**Hidden cobra**'s cyber attacks put their reputation and customers at stake. | **EXTRACTION TEMPLATE=CYBERTHREAT FIELD NAME= THREAT_ACTOR OUTPUT=Lazarus Group** |
| CYBERTHREAT<br>└ THREAT_ACTOR<br>Lazarus | What do Cyber Threat Actors want, such as the **Lazarus** Group? | |
| CYBERTHREAT<br>└ THREAT_ACTOR<br>Lazarus Group | What do Cyber Threat Actors want, such as the **Lazarus Group**? | |

*Figura 42 – Validazione delle regole di normalizzazione del progetto "CTI"*[361]

Il processo di normalizzazione ha restituito una risposta positiva, unificando le varianti linguistiche riscontrate nel testo, rispettivamente, delle seguenti entità: *Cyber Attack, Threat Actor* e *Lazarus Group.*

[361] Fonte: grafico autoprodotto mediante *software* COGITO STUDIO®

## *4.5 Modulo di estrazione delle relazioni tra entità*

Il sistema COGITO STUDIO® non possiede un modulo per l'estrazione delle relazioni semantiche tra entità. Al fine, quindi, di realizzare il terzo requisito dei desiderata del Cliente, si è adoperato un *Open Source Data Labeling Tool* denominato *LabelStudio®*[362].

Tale attività ha avuto come obiettivo l'individuazione di interdipendenze semantiche tra due o più entità rilevate mediante il precedente *work package* di estrazione dati.

Tali relazioni hanno permesso la comprensione delle dinamiche delle *cyber criminal activities* (singole o *clusters* – espresse in *campaigns)* intentate o compiute da *Threat Actors* – o gruppi di *hackers/hacktivists* – nonché degli strumenti da loro utilizzati.

Si è fatto riferimento, anche in questo caso, allo standard STIX, rielaborando la tabella esplicativa delle *Relationship Objects (SROs)*[363], per la comprensione dei collegamenti tra *SDOs (STIX domain Objects)* e le *SCOs (Cyber-Observable Objects)*[364].

Si è proceduto ad una semplificazione della tabella ai fini del progetto CTI, fornendo una identificazione su misura dei ruoli semantici delle entità estratte.

Nello specifico, sono stati messi in evidenza due ruoli: quello di un agente/attore – ovvero, la parte attiva di un evento, su cui ha o meno il controllo e/o di cui può esserne o meno la causa – definito **[SOURCE]**, e di un paziente/esperiente – entità che subisce o che è coinvolta passivamente nell'evento attivato dall'agente/attore – definito **[TARGET]**, interrelati da una specifica **[TYPE OF RELATIONSHIP]**, come mostrato di seguito.

[362] https://labelstud.io/
[363] Come indicato nell'*Appendix B. Relationship Summary Table* del sito web *Introduction to STIX*. Cfr. https://docs.oasis-open.org/cti/stix/v2.1/os/stix-v2.1-os.html#_e2e1szrqfoan
[364] https://oasis-open.github.io/cti-documentation/stix/intro.html

| [SOURCE] | [TYPE OF RELATIONSHIP] | [TARGET] |
|---|---|---|
| Campaign | **Targets** | Identity<br>Vulnerability<br>Location |
| Attack pattern | | Vulnerability<br>Location |
| Malware | | Identity<br>Location |
| Threat Actor | | Location<br>Vulnerability<br>Identity |
| Threat Actor | **Attributed to** | Identity |
| Campaign<br>Malware | **Authored by** | Threat Actor<br>Identity |
| Malware | **Communicates with** | IP<br>Url<br>Domain<br>Hash |
| Identity | **Downloads** | Malware<br>Tool |
| Malware | **Exploits** | Vulnerability |
| Tool | **Has** | Vulnerability |
| Threat Actor | **Impersonates** | Identity |
| Course of action | **Investigate** | IP<br>Url<br>Domain<br>Hash |
| Identity<br>Threat Actor | **Located at** | Location |
| Course of action | **Mitigate** | Vulnerability<br>Domain |
| Malware<br>Campaign | **Originates-from** | Location |
| Report_published | **Related to** | Threat Actor<br>Malware<br>Vulnerability<br>Campaign<br>Tool |
| Attack pattern | **Targets** | Vulnerability<br>Location |
| Malware | | Vulnerability<br>Identity |

| | | |
|---|---|---|
| Threat-actor | | Location<br>Vulnerability<br>Identity |
| Campaign | | Location<br>Vulnerability<br>Identity |
| Attack pattern | **Uses** | Malware<br>Tool |
| Malware | | Attack pattern<br>Malware<br>Tool |
| Threat Actor | | Attack pattern<br>Malware<br>Tool<br>Infrastructure |
| Campaign | | Attack pattern<br>Malware<br>Tool |
| Malware | **Variant of** | Malware |

***Figura 43 – Tabella per la definizione delle relazioni semantiche tra entità***[365]

A seconda del rapporto vigente tra il verbo che definisce il **[TYPE OF RELATIONSHIP]**, ed il soggetto della frase – individuato nella sezione **[SOURCE]**– *LabelStudio*® permette all'operatore di stabilire la diatesi attiva o passiva attraverso le seguenti *features*:

| | |
|---|---|
| → | La forma del verbo fornita in *input* non subisce alcuna variazione e mantiene la diatesi indicata nel **[TYPE OF RELATIONSHIP]**. |
| ← | La diatesi del verbo fornita in *input* viene invertita: da forma attiva si trasforma in forma passiva, e viceversa. |
| ↔ | Partendo da un qualsiasi verbo fornito in *input*, questa funzionalità considera e stabilisce ogni diatesi possibile (attiva, passiva e riflessiva). |

***Figura 44 – Feature di trasformazione della diatesi verbale***[366]

[365] Fonte: grafico autoprodotto mediante *Open Source Data Labeling Tool LabelStudio*®
[366] Ibidem.

A titolo esemplificativo, prendendo come *input* il **[TYPE OF RELATIONSHIP]** EXPLOITS, presentandosi nella tabella delle relazioni con la seguente struttura: *"malware exploits vulnerability":*

1) nel primo quadrante, si stabilirà la struttura sintattica originaria **[SOURCE,** *malware***] [TYPE OF RELATIONSHIP,** *exploits***] [TARGET,** *vulnerability***]**;

2) nel secondo quadrante, la diatesi del verbo verrà invertita, trasformando la relazione in **[TARGET,** *vulnerability***] [TYPE OF RELATIONSHIP,** *is exploited***] [SOURCE,** *by malware***]**;

3) nel terzo quadrante, verranno stabilite entrambe le strutture sintattiche, instaurando una condizione di interrelazione tra **[SOURCE]** e **[TARGET].**

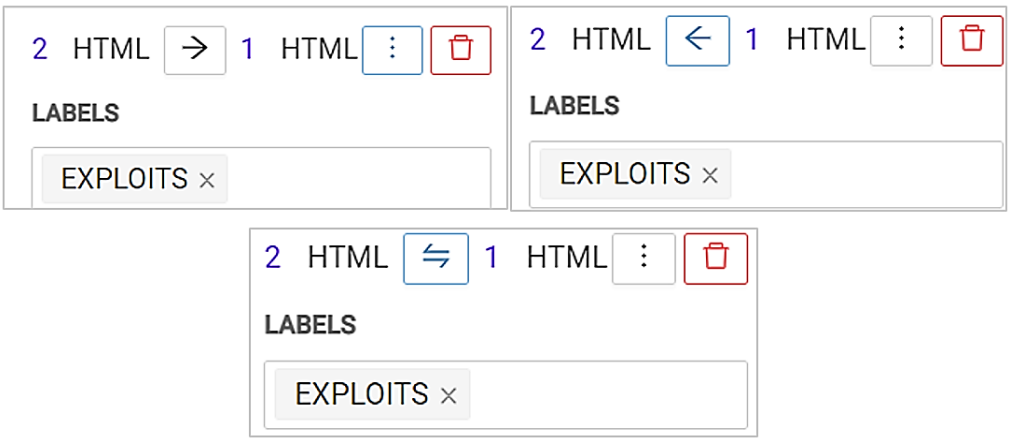



*Figura 45 – Trasformazione diatesi verbale*[367]

In *LabelStudio*®, la definizione delle relazioni tra entità è un processo attivato e perseguito al 100% manualmente; non si configura, pertanto, come modulo (semi)automatizzato, né apprendente. La realizzazione di tale attività è, pertanto, resa possibile unicamente mediante la *knowledge* degli esperti linguistici e di dominio CTI.

Come prima *action* – nel caso specifico del progetto CTI – il *team* ha provveduto al trasferimento degli *output* di estrazione ricevuti da COGITO STUDIO® verso la piattaforma *LabelStudio*®. Tali estrazioni vengono definite *labels* da questo secondo *tool*.

[367] Ibidem.

Ciò ha permesso una pre-annotazione automatica dei testi, rilevando, così, su *LabelStudio*® le entità definite nei moduli di COGITO STUDIO®.

Sono stati riportati, quindi, i diciotto *fields* individuati ed estratti *ex ante*, unificati nei loro *templates*; è stato, inoltre, caricato a sistema un *raw dataset* di circa 800 testi per la ricerca e l'individuazione delle relazioni tra entità.


*Figura 46 – Labels afferenti al progetto "CTI"*[368]

Accanto ad ogni *label*, il sistema indicherà il numero relativo alla quantità di estrazioni rilevate nel testo fornito in *input* afferenti a quello specifico *field*, come mostrato di seguito.

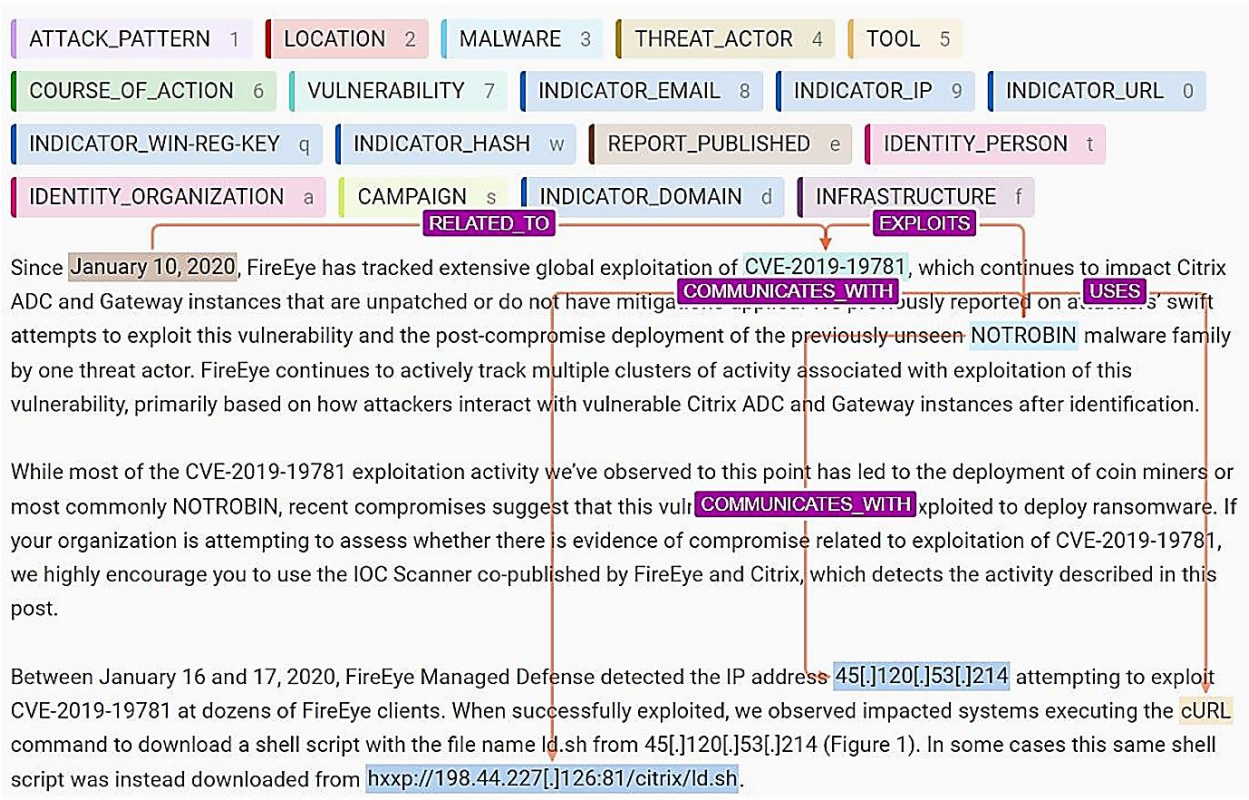



*Figura 47 – Input testuale con relazioni tra entità annotate*[369]

[368] Fonte: grafico autoprodotto mediante *Open Source Data Labeling Tool LabelStudio*®
[369] Ibidem.

Partendo da una porzione *pre-labelizzata* del testo fornito in *input*, si analizzino ora le relazioni tra le varie entità, stabilite dagli operatori:

| [SOURCE] | [TYPE OF RELATIONSHIP] | [TARGET] |
|---|---|---|
| **[Report_published]** "January 10, 2020" | 15 HTML → 14 HTML<br>LABELS<br>RELATED_TO × | **[Vulnerability]** "CVE-2019-19781" **[Malware]** "NOTROBIN" |
| **[Malware]** "NOTROBIN" | 2 HTML → 1 HTML<br>LABELS<br>EXPLOITS × | **[Vulnerability]** "CVE-2019-19781" |
| **[Malware]** "NOTROBIN" | 2 HTML → 5 HTML<br>LABELS<br>USES × | **[Tool]** "cURL *(command)"* |
| **[Malware]** "NOTROBIN" | 51 HTML → 1 HTML<br>LABELS<br>COMMUNICATES_WITH × | **[IP]** "45[.]120[.]53[.]214" |
| **[Malware]** NOTROBIN | 51 HTML → 2 HTML<br>LABELS<br>COMMUNICATES_WITH × | **[Url + IP]** "hxxp://198.44.227[.]12" |

*Figura 48 – Analisi della definizione delle relazioni*[370]

Mediante tale attività di riconoscimento e definizione delle relazioni tra entità è stato possibile eseguire un ulteriore *deep mining* del dato testuale, che ha condotto ad una validazione aggiuntiva dei moduli elaborati *ex ante*, concludendo, così, il progetto CTI.

[370] Ibidem.

# CONSIDERAZIONI FINALI

Obiettivo del presente lavoro di tesi di Laurea Magistrale è stato invitare ad una maggiore consapevolezza (sperando in una conseguente diffusione) in materia di Intelligence linguistica[371], sia essa realizzata mediante applicativi di analisi semantica (semi)automatizzata, che condotta mediante abilità di tipo OSINT, SOCMINT e (V)HUMINT.

Ho considerato fondamentale evidenziare la necessità di conoscere le origini scientifico-cognitiviste del pensiero linguistico, al fine di esplorare appieno le potenzialità del suo funzionamento per concretizzare approcci e metodologie di applicazione ai fini sia della realizzazione di una comunicazione efficiente, nonché di un'investigazione olistica degli elementi (anche para-, extra- e non-) verbali che la compongono.

L'Intelligence linguistica sfrutta competenze in ambito psicologico, etnoantropologico e cognitivo al fine di realizzare un *framework* (e conseguente teatro) di azione funzionale del dominio del linguaggio, il quale, in maniera proattiva e/o reattiva, permette all'essere umano di creare nuove visioni, costruire nuove alleanze, scalfire teorie o distruggere patti e/o unioni.

Il linguaggio rappresenta un vero e proprio potere per chi lo sa destreggiare *(«La parola determina un suono e, con esso, una visione del mondo.* […] *È alla base della nostra capacità di influenzare»*[372]*);* la propensione verso una sensibilità linguistica, però, sembra aver lasciato terreno ad una (quasi)totale analisi computazionale.

Indubbiamente, algoritmi e condizioni formali sono in grado di offrire vantaggi in termini di economicità di tempo e di risorse impiegate nella comprensione ed analisi di testi, ma dette funzionalità, seppur "intelligenti" non godono di abilità di comprensione della *Weltanschauung*[373] del parlante, né sono soggette ad automatico aggiornamento; non seguono, pertanto, il passo della naturale ed incessante evoluzione della lingua.

[371] Il concetto di "intelligenza linguistica" venne postulato da Howard Gardner nel 1983 nella sua teoria delle intelligenze multiple, contenuta nel libro "Frames of Mind".
[372] J. Nabben, *Il potere del linguaggio,* Edizioni LSWR, 2014, Introduzione.
[373] «*Concezione della vita, del mondo; modo in cui singoli individui o gruppi sociali considerano l'esistenza e i fini del mondo e la posizione dell'* [essere umano] *in esso*». Cfr. https://www.treccani.it/enciclopedia/weltanschauung_%28Dizionario-di-filosofia%29/

La comunicazione umana, resa necessaria per soddisfare esigenze primordiali per la sopravvivenza della specie, dà forma da sempre anche ad emozioni, fantasie ed intenzioni: l'essere umano è, infatti, in grado sia di realizzare sensi propri, che figurati.

Tutto ciò, ad oggi, non può esser completamente elaborato o emulato da una macchina. Le macchine, infatti, ereditano una porzione del linguaggio – risultante in un *database* di codici standard, a volte molto lontani dalla struttura linguistica realmente utilizzata in alcuni contesti comunicativi (in maniera particolare, il mondo dei *social)*, e tramite algoritmi – talvolta contenenti *biases*[374] e/o prodotti sulla base di dati incompleti – ed "inscatolano" concettualmente il mondo in maniera formale ed oggettiva, contrariamente a ciò che è realmente, ovvero: naturale e soggettivo.

Anche in questo contesto, l'apporto umano risulta indispensabile per l'indirizzamento degli algoritmi verso la giusta direzione, i quali regolano e danno vita al *mare magnum* di *software* ad oggi in circolazione. È solo partendo dalla piena consapevolezza del proprio potere intelligente come esseri umani (prima ancora di elaboratori, analisti ed operatori), che si possono destreggiare e creare nuove intelligenze artificiali, le quali, però *[rasserenandosi, così, la sottoscritta]*, non potranno mai completamente sostituirsi all'umanità dell'azione e del pensiero. Infatti, in aggiunta ai limiti dell'odierna A.I., una macchina o un robot non sono inoltre (ancora) in grado di leggere quanto viene espresso, sotteso e/o sottinteso generalmente mediante *body language*[375], risultante in movimenti involontari del corpo che realizzano gesti e/o (micro)espressioni facciali.

Oggi, in ambito digitale, si cerca di realizzare tale comunicazione extra-verbale mediante *gif* ed *emoticons* di qualsiasi tipo, ma nessun *software* di ricognizione linguistica è in grado di decifrare tali segni stilizzati, che influenzano in maniera significativa il *sentiment* del testo, ma che sfruttano il fenomeno della pareidolia.

In aggiunta a ciò, gli elaboratori computazionali del linguaggio naturale, se non addestrati, non riescono a riconoscere malapropismi, errori di battitura, "orrori" grammaticali, regionalismi e/o forme dialettali.

---

[374] *Biases* cognitivo-culturali trasmessi dagli sviluppatori stessi. Un operatore informatico deve rimanere sempre imparziale e neutro nelle accezioni, in qualsiasi teatro di azione.

[375] Lo scienziato ed antropologo Albert Mehrabian ritiene che solamente il 7% di tutte le informazioni che ci arrivano da un discorso passa attraverso le parole (elementi verbali); il 38% che ci perviene da elementi paraverbali (tono della voce, volume, ecc.); ed il 55% dal linguaggio del corpo (elementi non verbali).

L'operatore linguistico deve essere in grado, pertanto, di prevedere tali disgrafie: il mondo di internet, in particolare, sta cambiando sempre di più le regole vigenti nella comunicazione scritta. Ciò sta portando alla formulazione di una nuova tipologia di italiano (definita "neo-standard"[376]) il quale considera come accettabili alcuni elementi che una volta erano ad uso esclusivo del parlato, anche nello scritto. La maggior parte di tali peculiarità pertengono al livello morfosintattico[377].

Nei riguardi del *Natural Language Processing*, negli ultimi anni, si è assistito ad enormi progressi nell'accuratezza di estrazione e di analisi linguistica. Di seguito, alcuni dati[378] elaborati da *NLP tools* scritti in *Python* che includono *Natural Language Toolkit* (NLTK)[379] e la libreria *open source SpaCy*:

- riconoscimento della lingua: per testi brevi, soggetti a maggiore ambiguità dei contesti concettuali: risultati positivi all'80%; per testi di lunghezza più ampia, maggiormente esaustivi in termini di assegnazione a domini: accuratezza al 99%;
- *tokenizzazione* delle parole e segmentazione dei periodi: circa 98%;
- analisi morfologica e lessicale: tra il 90% e il 97%;
- analisi sintattica: tra il 70% e oltre il 90%.

Stiamo assistendo a cambiamenti sempre più rapidi non solo nel campo del *Natural Language Processing*, bensì della stessa *Artificial Intelligence:* punto di forza dello sviluppo digitale mondiale. L'intelligenza artificiale è, ad oggi, considerata la più grande forma di assistenza mai creata, in grado di fornire sia soluzioni indirizzate al miglioramento della qualità della vita di ogni individuo, nonché tecnologie in grado di individuare *cyber threats* e/o *malicious activities* nell'ambito non solo della *Cyber Security*, bensì nei teatri di guerra cibernetica dispiegati a livello internazionale.

[376] Per un approfondimento: N. Grandi, *Che tipo, l'italiano Neostandanrd!,* Società di Linguistica Italiana, 2019. Cfr. https://www.societadilinguisticaitaliana.net/wp-content/uploads/2019/08/004_Grandi_Atti_SLI_LII_Berna.pdf

[377] A livello morfologico, si assiste oggi ad una riorganizzazione del sistema pronominale, del sistema dei dimostrativi, nella selezione delle congiunzioni, una semplificazione del sistema verbale (nei modi e nei tempi), uso del ""ci" attualizzante, del "che" polivalente, nonché di forme ridonanti. A livello sintattico, si assiste, invece, a un uso sempre più frequente di dislocazione nell'ordine dei costituenti; tale peculiarità era prima relegata solo al canale orale.

[378] G. Altobello, *Natural Language Processing, cos'è, come funziona e applicazioni*, 2021. Cfr. https://www.ai4business.it/intelligenza-artificiale/natural-language-processing-tutto-quello-che-ce-da-sapere/

[379] Il NLTK è una «*suite di librerie e programmi per l'analisi simbolica e statistica nel campo dell'elaborazione del linguaggio naturale*». – S. Bird, E. Klein, J. Baldridge, *Multidisciplinary instruction with the Natural Language Toolkit*, 2008. Cfr. https://aclanthology.org/W08-0208.pdf

# RIFERIMENTI BIBLIOGRAFICI


- ADORNETTI, I., CHIERA, A., FERRETTI, F. (2019), *Embodied Cognition e Origine Del linguaggio: il ruolo cruciale del gesto*, Università degli Studi di "Roma Tre".
- ALINEI, M. (2011), *Linguistica storica e reificazione del linguaggio in margine a un articolo-recensione di Adiego*, Università di Utrecht.
- ANTISERI, D., SOI, A. (2013), *Intelligence e metodo scientifico*, Rubbettino Editore.
- ANTISERI, D., SOI, A. (2015), *La scienza dell'Intelligence nell'era dell'incertezza,* supplemento a *Formiche*.
- BARTHES, R. (1981), *La grana della voce, Interviste 1962 - 1980,* Edizioni Einaudi.
- BAZZANELLA, C. (2014), *Linguistica cognitiva. Un'introduzione,* Edizioni Laterza.
- BEGOTTI, P. (2013), *L'acquisizione linguistica e la Glottodidattica umanistico-affettiva e funzionale*, Laboratorio ITALS, Università Ca' Foscari di Venezia.
- BOLASCO, S. (2005), *Statistica testuale e Text Mining: alcuni paradigmi applicativi*, Università degli Studi di Roma "La Sapienza".
- BONACCI, F. (2009), *Processi inferenziali Vs. processi di codifica/decodifica nei modelli di trasmissione dell'informazione fra individui,* Università della Calabria.
- BORGHI, A. M. (1997), *L'organizzazione della conoscenza, Aspetti e Problemi,* Pitagora Editore, Bologna.
- CALIGIURI, M. (2016), *Cyber Intelligence – Tra Libertà e Sicurezza*, Donzelli Editore
- CLARCK, A., CHALMERS, J. D. (2010), *The extended mind,* MIT University Press.
- COQUET, J. C. (2008), *Le istanze enuncianti. Fenomenologia e semiotica*, trad. it. NICOLINI, E., Bruno Mondadori, Milano, 2008.
- CROFT, W., CRUSE, D. A. (2004), *Cognitive Linguistics*, Cambridge University Press.
- CUMMINS, J. (1979), *Cognitive/Academic Language Proficiency, Linguistic Interdipendence, the Optimum Age Question and Some Other Matters*, Working Papers on Bilingualism.
- DE LUISE, FARINETTI (2010), *Lezioni di storia della filosofia*, Zanichelli Editore.
- DE SAUSSURE, F. (1916), *Cours de linguistique générale.*
- DIADORI, P. (2015), *Insegnare l'italiano come seconda lingua,* Carocci Editore.
- FILLMORE, C. J. (1976), *Frame Semantics and the Nature of Language*, University of California-Berkeley.
- FILLMORE, C. J. (1982), *Frame semantics. Cognitive Linguistics: Basic Readings,* Hanshin Publishing.
- FORTE, P. (2020), *Pensiero Computazionale e la Macchina di Turing per il problem solving,* Università degli Studi di Roma "La Sapienza".
- GARDNER, H. (1983) *Frames of Mind: The Theory of Multiple Intelligences*, Basic Books, New York.
- GARDNER, H. (1993), *Multiple Intelligences: The Theory in Practice,* Basic Books, New York.
- GIUSTOZZI, C. (2019), *Cos'è il rischio cyber e perché ce ne dobbiamo preoccupare,* ISPI – Istituto per gli studi di politica Internazionale.
- GLAZEL PASSERINI, L. (2013), *Impossibilità di Tokens, necessità di Types.*
- GORI, U., GERMANI, L. S. (2011), *Information Warfare. La sfida della cyber-intelligence al sistema Italia: dalla sicurezza delle imprese alla sicurezza nazionale*, Franco Angeli Editore.
- Gruppo di lavoro 71ª sessione di Studio dell'Istituto Alti Studi per la Difesa (2020), *L'impatto dell'Intelligenza Artificiale (AI-Artificial Intelligence) sul ciclo di Intelligence e sugli strumenti a disposizione per i pianificatori militari e le forze dell'ordine,* Ministero della Difesa.
- HARARI, Y. N. (2011), *Da Animali a Dèi. Breve storia dell'umanità,* trad. it. BERNARDI, G. (2017), Bompiani.

- KOSZ, A. (2020), *La rappresentazione delle conoscenze – diversi modelli delle strutture concettuali nell'ambito della linguistica cognitiva,* University of Silesia in Katowice
- KRASHEN, S. D. (1981), *Second Language Acquisition and Second Language Learning*, Oxford.
- KRASHEN, S. D. (1985), *The Input Hypothesis: Issues and Implications*, Longman, London.
- MCKENZIE, P. W. (2014), *Intelligenze Multiple e Tecnologie per la didattica*, Edizioni Erickson.
- MERLEAU-PONTY, M. (1945), *Phénoménologie de la perception,* Paris: Gallimard, trad. it. BONOMI, A. (1965), *Fenomenologia della percezione*, Milano, Il Saggiatore, 1965.
- MININI, A. (2021), *Personal Knowledge Base.*
- MINSKY, M. (1974), *A Framework for Representing Knowledge,* MIT-AI Laboratory.
- PAPPALARDO, L., GIANNOTTI, F. (2015), *Capire la mobilità attraverso i big data*, Donzelli Editore.
- PAVLOV, I. (1927), *Conditional Reflexes,* Dover Publications
- PERIFANOS, K., FLOROU, E., GOUTSOS, D. (2007), *Deep Learning-based, end-to-end metaphor detection in Greek language with Recurrent and Convolutional Neural Networks,* Department of Linguistics, National and Kapodistrian – University of Athens, Greece.
- PETRICCA, P. (2019), *Semantica: forme, modelli e problemi*, LED, Milano.
- PHYNTION, M. (2013), *Understanding the Intelligence Cycle*, Rutledge, New York.
- PINKER, S. A. (2002), *Come funziona la mente*, trad. it. PARIZZI, M., Mondadori.
- PINKER, S. A. (2002), *The Blank Slate: The Modern Denial of Human Nature*, trad. it. PARIZZI, M. (2007), *Tabula rasa. Perché non è vero che gli uomini nascono tutti uguali*, Mondadori
- PRESIDENZA DEL CONSIGLIO DEI MINISTRI (2013), *Gnosis – Rivista Italiana di Intelligence*, Dipartimento delle Informazioni per la Sicurezza.
- RAFFAELLI, A. (2022), *Data Breach: una sfida tecnologica, legale ed etica per il futuro,* Università Campus Bio-Medico di Roma.
- SEARLE, J. (1976), *Atti linguistici: saggio di filosofia del linguaggio,* trad. it. CARDONA, G. R. (ed. or. 1969), Boringhieri, Torino.
- SEARLE, J. R. (1980), *The Background of Meaning,* Cambridge University Press.
- SEARLE, J. R. (1980), *Menti, cervelli e programmi: un dibattito sull'intelligenza artificiale,* trad. it. TONFONI, G. (1984), CLUP Milano.
- SELINKER, L. (1972), *Interlanguage*, IRAL.
- SOSA, D. (1999), *Checking Searle's Background,* Luis Manuel Valdés-Villanueva Publishing.
- SPERINI, A. (2017), *Implementazione del ciclo d'Intelligence tramite l'utilizzo della Social Media Intelligence (SOCMINT),* Ministero della Difesa.
- TETI, A. (2018), *Cyber Espionage e Cyber Counter Intelligence,* Rubbettino Editore.
- TETI, A. (2019), *Virtual Humint – La nuova frontiera dell'Intelligence,* Rubbettino Editore.
- TRIBERTI, C., CASTELLANI, M. (2020), *L'intelligenza artificiale oltre le quattro leggi della robotica. Riflessioni anche alla luce della pandemia da COVID-19,* GoWare Editore.
- TRINCHERO, R. (2018), *(Cyber)Bellum omnium contra omnes. Strategie educative di prevenzione alla guerra cognitiva in Rete",* Erickson Edizioni.
- VALLI, F. (2021), *Cos'è la Cognitive Warfare?*
- WOLF, M. (2012), *Proust e il calamaro. Storia e scienza del cervello che legge*, trad. it. GALLI, S., Vita e Pensiero Edizioni.

# RISORSE ONLINE


- CALIGIURI, M. (22 giugno 2016), *Cyber Intelligence, la sfida dei data scientist*. www.sicurezzanazionale.gov.it/sisr.nsf/approfondimenti/cyber-intelligence-la-sfida-dei-data-scientist.html
- TOFALO, A. (2017), *Intelligence Collettiva: i dati, l'informazione e il linguaggio.* www.angelotofalo.com/intelligence-collettiva-dati-linformazione-linguaggio/
- GALTIERI, E., MELEGARI, A. (2021), *Il lato oscuro degli algoritmi*, per *Panorama*. www.panorama.it/Tecnologia/cyber-security/algoritmo-intelligenza-artificiale-computer
- Sicurezza Nazionale (2014), *Lezioni sull'Intelligence,* Scuola di Formazione. www.sicurezzanazionale.gov.it/sisr.nsf/wp-content/uploads/2014/05/lezione-intelligence.pdf
- GALTIERI, E. (2022), *Da Anonymous ai malware, l'arte della guerra cyber*, per *Formiche*. https://formiche.net/2022/02/arte-della-guerra-era-cyber-galtieri-cy4gate/
- COSTANTINI, P. (2019), *Language Intelligence.* www.paolocostantini.com/language-intelligence
- MARTIGNON, M. (2004), *Riflessione sulla lingua – Il codice linguistico.* www.insegnareitaliano.it/documenti/Laboratorio%20docenti/italiano/Martignon/riflessione_sulla_lingua/Il%20codice%20linguistico_2004.PDF
- NALESSO, N. (2019), *Il ruolo della cyber intelligence nella tutela della sicurezza nazionale.* www.cyberlaws.it/en/2019/il-ruolo-della-cyber-intelligence-nella-tutela-della-sicurezza-nazionale/
- Sito Istituzionale del Parlamento Europeo (2020), *Che cos'è l'A.I. e come viene usata.* www.europarl.europa.eu/news/it/headlines/society/20200827STO85804/che-cos-e-l-intelligenza-artificiale-e-come-viene-usata
- CATALANO, M. (2021), *L'intelligenza artificiale e il potere cognitivo delle metafore: spunti di riflessione per una didattica innovativa.* www.ictedmagazine.com/index.php/ricerca-e-innovazione/243-l-intelligenza-artificiale-e-il-potere-cognitivo-delle-metafore-spunti-di-riflessione-per-una-didattica-innovativa.html
- ESPOSITO, M. (2019), *Linguaggio naturale e intelligenza artificiale: a che punto siamo*. www.agendadigitale.eu/cultura-digitale/linguaggio-naturale-e-intelligenza-artificiale-a-che-punto-siamo/
- ALTOBELLO, G. (2021), *Natural Language Processing, cos'è, come funziona e applicazioni* www.ai4business.it/intelligenza-artificiale/natural-language-processing-tutto-quello-che-ce-da-sapere/
- SANTAGATA, E., MELEGARI, A. (2017), *Putin esorta gli studenti a dedicarsi all'Intelligenza Artificiale*, per *Analisi Difesa.* www.analisidifesa.it/2017/09/putin-esorta-agli-studenti-a-dedicarsi-allintelligenza-artificiale/
- BUONOCORE, T. (2021), *NLP in Medicina*. https://www.biomeris.it/nlp-in-medicina/
- CAPONE, E. (2021), *Il futuro visto da Annalisa Barla, la prof che insegna Machine Learning*, per *Italian.Tech.*www.italian.tech/2021/12/27/news/il_futuro_visto_da_annalisa_barla_la_prof_che_insegna_machine_learning-331506239/
- NATO RAPID DEPLOYABLE CORPS – ITALY (2021), *Deep Watching on Cyber Threat, Everywhere Rapidly*. www.nrdc-ita.nato.int/db_object/www_nrdc-ita_nato_int/usr/file/ER-Magazine-Everywhere-Rapidly-July-2021-NATO-NRDCITA.pdf
- CALIGIURI, M. (2022), *La mente come campo di battaglia*, per *Formiche.* www.formiche.net/2022/03/campo-battaglia-definitivo-mente-persone/
- TRINCHERO, R. (2018), *Against the cognitive war. Promoting active skepticism.* https://www.unipa.it/persone/docenti/c/gianna.cappello/.content/documenti/TRINCHERO.pdf
- Ten. Col. FONTANA L., SMD II Rep. – Uff. Materiali di Armamento e Alta Precisione (2016), *Le Operazioni Psicologiche Militari (PSYOP), La 'conquista' delle menti,* Ministero della Difesa. www.difesa.it/InformazioniDellaDifesa/periodico/IlPeriodico_AnniPrecedenti/Documents/Le_Operazioni_Psicologiche_militar_620menti.pdf
- HOOK RICHENS, R. (1956, Vol. 3, n°1), *Preprogramming for Mechanical Translation*.

www.aclanthology.org/www.mt-archive.info/50/MT-1956-Richens.pdf
- CORDIS EUROPA (2020), *L'intelligenza artificiale potrebbe presto capire le metafore linguistiche.* www.cordis.europa.eu/article/id/182697-artificial-intelligence-may-soon-understand-language-metaphors/it
- SANTAGATA, E., MELEGARI, A. (2020), *Social media: il preoccupante rovescio della medaglia,* Analisi Difesa. https://www.analisidifesa.it/2020/10/social-media-il-preoccupante-rovescio-della-medaglia/ www.filosofiafammiunthe.wordpress.com/2015/07/08/linguistica-cognitiva/amp/
- ALTAMURA, G. (2019), *Umberto Eco, il giocoliere dell'Intelligenza,* Università di Bari Aldo Moro. www.uniba.it/ateneo/editoria-stampa-e-media/linea-editoriale/fuori-collana/volumeco
- PETKOVA, T. (2014), *A Web of People and Machines: W3C Semantic Web Standards*. https://www.ontotext.com/blog/a-web-of-people-and-machines-w3c-semantic-web-standards/
- BOLIOLI, A. (2016) *Motore di Ricerca Semantico: che cos'è e a cosa serve?* https://www.celi.it/blog/2016/04/motore-di-ricerca-semantico/
- FOCUS (2014), *Che cos'è il Cognitive Computing?* www.focus.it/tecnologia/innovazione/che-cos-e-il-cognitive-computing
- REINA, M. (2022), *Ucraina, come funzionano gli attacchi malware dei russi.* www.webnews.it/2022/03/05/ucraina-come-funzionano-gli-attacchi-malware-dei-russi/
- CALIGIURI, M. (2022), *La mente come campo di battaglia*, per *Formiche.net* www.formiche.net/2022/03/campo-battaglia-definitivo-mente-persone/
- BOLASCO, S. (2005), *Statistica testuale e text mining: alcuni paradigmi applicativi.* www.didattica-2000.archived.uniroma2.it/Statistica_Sociale/deposito/bolasco.pdf
- SANTAGATA, E., MELEGARI, A. (2019), *Gli USA sognano un computer capace di ragionare come un bambino*, per *Analisi Difesa.* https://www.analisidifesa.it/2019/04/gli-usa-sognano-un-computer-capace-di-ragionare-come-un-bambino/
- VERENI, P. (2022) *Fuori tempo massimo. https://pierovereni.blogspot.com/*
- FLORES D'ARCAIS, A. (2015), *Usa: la nuova guerra è il cyberspazio.* www.inchieste.repubblica.it/it/repubblica/rep-it/2015/06/15/news/cosi_mi_arruolo_tra_gli_007-115409472/
- SANTAGATA, E., MELEGARI, A. (2019), *Ecco quanto (poco) costa un robot capace di disinformare (tanto),* per *Analisi Difesa.* https://www.analisidifesa.it/2019/06/ecco-quanto-poco-costa-un-robot-capace-di-disinformare-tanto/
- DE BIASE, L. (2017), *Cogito, ergo capisco come voi umani*, per *Il Sole 24 Ore.* www.ilsole24ore.com/art/cogito-ergo-capisco-come-voi-umani-AEKONoDC
- COFINI, F. (2022), *La guerra sul web: in Ucraina cyber attacchi triplicati e servizi fondamentali a rischio*, per *RaiNews.* www.rainews.it/articoli/2022/03/la-guerra-sul-web-in-russia-e-ucraina-cyber-attacchi-triplicati-e-servizi-fondamentali-a-rischio-73de6ad4-ecc9-40a1-a21a-8d6a6a3c2f52.html
- PRESIDENZA DEL CONSIGLIO DEI MINISTRI, Sistema di Informazione per la Sicurezza della Repubblica (2015), *Connessi con la sicurezza. Il racconto di una Intelligence diffusa*, LeggIntelligence. www.sicurezzanazionale.gov.it/sisr.nsf/letture/connessi-con-la-sicurezza-il-racconto-di-una-intelligence-diffusa.html
- ACERBI, A. (2005), *La mente nella cultura: cognizione ed analisi dei fatti culturali*. www.acerbialberto.com/files/ITA_2005_annali.pdf
- BUONOCORE, T. (2019), *Man is to Doctor as Woman is to Nurse: the Gender Bias of Word Embeddings.* https://towardsdatascience.com/gender-bias-word-embeddings-76d9806a0e17


Pubblicato nel novembre 2023
Società Italiana di Intelligence
*SOCINT Press*
https://press.socint.org/

979-12-80111-48-7