



PARTNER: SECONDA UNIVERSITÀ DI NAPOLI E UNIVERSITÀ DI ROMA LA SAPIENZA

RESPONSABILE: PROF. BENIAMINO DI MARTINO

Technical Report: **2.2.9**

*LC3 – Laboratorio pubblico-privato di ricerca
sul tema della Comunicazione delle Conoscenze Culturali*

PAG 1 DI 5

Technical Report

TR2.2.9

Costruzione automatica di una ontologia in formato OWL sul dominio “Il rapimento di Aldo Moro”.

DRAFT

| | | | | |
|---------------------|---------------------|----------|------------|--------------|
| <i>PROGETTO LC3</i> | <i>Revisione n*</i> | <i>0</i> | <i>Del</i> | <i>-----</i> |
|---------------------|---------------------|----------|------------|--------------|



Abstract

In questo TR viene descritta brevemente la procedura semi-automatica per produrre un'ontologia sul dominio "Aldo Moro". La metodologia viene descritta ad alto livello e una sua descrizione più dettagliata verrà ripresa nei prossimi TR.

.1 METODOLOGIA.

Un obiettivo di questo OR consiste nella creazione di una ontologia relativa al dominio prescelto per il progetto, denominato "caso Aldo Moro", mediante metodi semi-automatici ed opportune applicazioni web di supporto alla popolazione di ontologie.

Brevemente descriviamo la procedura adottata, rimandando per dettagli al prossimo Deliverable di questo OR.

- Il primo passo del lavoro per la costruzione di tale ontologia di dominio è stata la costruzione del *corpus* documentale, ottenuto tramite acquisizione OCR di un centinaio di articoli di giornale, forniti dall'Istituto Luigi Sturzo di Roma, che appartengono all'anno del caso Moro (1978) e che parlano di questo argomento. Si è ottenuto in questo modo un corpus per un totale di 4 megabyte di documenti in formato testo.
- Il secondo passo è consistito nel trattamento del corpus con l'utilizzo del software open source *Treetagger*¹, che consente di ottenere l'analisi grammaticale (identificazione del Part of Speech) e la lemmatizzazione delle singole parole. Ciò che si è ottenuto da questo spoglio del corpus è stata una lista di circa 4000 parole indicante: frequenza, forma lessicale (così come trovata nel corpus), parte del discorso e lemma. Ecco un estratto di questa lista che mostra le parole più frequenti:

¹ www.ims.uni-stuttgart.de/projekte/complex/TreeTagger/

| | | | | |
|--------------|--------------|---|-----|-------|
| PROGETTO LC3 | Revisione n* | 0 | Del | ----- |
|--------------|--------------|---|-----|-------|



368 via NOUN via
322 presidente NOUN presidente
285 partito NOUN partito
271 messaggio NOUN messaggio
259 terroristi NOUN terrorista
213 polizia NOUN polizia
209 brigatisti NOUN brigatista

3. Il terzo passo è consistito nell'estrarre, attraverso il sistema TermExtractor² (Velardi and Sclano, 2007) espressioni polirematiche e nomi propri. Ad esempio:

caso Moro

Brigate Rosse

Presidente del Consiglio

stella a cinque punte

decreto legge

Amintore Fanfani

- A questo punto, utilizzando il sistema TAV³ (un editor di ontologie sviluppato dalla Sapienza, sul quale si forniranno dettagli nel Deliverable 2.2), si è proceduto come segue:

² accessibile come web application su <http://lcl2.uniroma1.it/termextractor>

| | | | | |
|--------------|--------------|---|-----|-------|
| PROGETTO LC3 | Revisione n* | 0 | Del | ----- |
|--------------|--------------|---|-----|-------|



- Si è generata una versione in linguaggio owl dell'ontologia ItalWordNet, che ha costituito la base di partenza per il presente lavoro;
- ItalWordNet è l'estensione italiana di MultiwordNet⁴. Attualmente è disponibile la versione 1.39 che contiene 58.000 sensi italiani e 32.700 *synsets* ("set di sinonimi", ai quali corrispondono altrettanti *synsets* inglesi).
- ItalWordNet non copre che il 50% dei termini singoli e polirematiche, estratti dal corpus di dominio come delineato nei precedenti punti. Si è dunque proceduto, mediante il sistema TAV e altri supporti informatici, ad arricchire ItalWordNet con i sensi mancanti (il che ha implicato i) riconoscere i vari sensi dei lemmi, selezionare i sensi effettivamente presenti nel contesto mediante opportune funzionalità del sistema TAV, e ii) individuare il corretto posizionamento del nuovo senso nell'ontologia)
- Per ogni senso inserito, è stata anche aggiunta una definizione, come mostrato in figura, ed altre informazioni che verranno descritte nel Deliverable 2.2. Inoltre, sono state inserite eventuali varianti lessicali, ovvero forme diverse che designano la stessa entità. Questo fenomeno è frequente in particolare per gli antroponimi, ad esempio Aldo Moro, Presidente Moro, Moro, ecc. Il sistema TAV fornisce un supporto all'identificazione di queste varianti, tramite una analisi dei contesti nel corpus.

Il task verrà completato entro il prossimo SAL.

³ accessibile come web application su <http://lcl.di.uniroma1.it/tav>

⁴ <http://multiwordnet.itc.it/english/home.php>

| | | | | |
|--------------|--------------|---|-----|-------|
| PROGETTO LC3 | Revisione n* | 0 | Del | ----- |
|--------------|--------------|---|-----|-------|

written by Roberto Navigli
TAXonomy Validator

| WordNet | Paravia | occurrences

Taxonomy

- set
- organized crime
- subculture
- nonalignment
- political system
- moicity
- tribe
- movement
 - artistic movement
 - common front
 - cultural movement
 - political movement
 - Movimento Popolare
 - Movimento Rivoluzionario
 - avanguardia
 - partito armato
 - Brigate Rosse**
 - Settembre Nero
 - Sinn Fein
 - syndicalism
 - reform movement
 - religious movement
 - collection

| | | | |
|---------------|---|----------------------------------|-------------------------------|
| Concept label | Brigate Rosse | | |
| Concept id | 0000000615 | | |
| Subclass of | partito armato | | |
| Definition | partito armato comunista fondato da Alberto Franceschini e Renato Curcio nel 1970, considerato il maggiore gruppo terroristico del secondo dopoguerra in Italia (alicepaesetto@yahoo.it) | | |
| Has-variants | organizzazione, Movimento di resistenza proletario offensivo, br, Partito Comunista Combattente ➤ Add a new instance | | |
| Instance | true | | |
| Changes | Author | Superclass | Date |
| | alicepaesetto@yahoo.it | movimento terroristico (deleted) | Wed Nov 28 12:02:33 CET 2007 |
| | alicepaesetto@yahoo.it | UK movement, UK front | Wed Nov 28 12:02:33 CET 2007 |
| | alicepaesetto@yahoo.it | partito armato | Fri Oct 05 15:33:00 CEST 2007 |
| | alicepaesetto@yahoo.it | movimento terroristico (deleted) | Sat Nov 17 15:43:00 CET 2007 |
| | alicepaesetto@yahoo.it | partito armato | Sat Nov 17 15:48:00 CET 2007 |
| Options | 🔍 Expand subtree 🔍 Collapse subtree 📶 Up one level ➤ Add conceptual relation ➤ Add concept ✖ Delete concept 🔄 Move concept 🔄 Move subtree | | |

Brigate Rosse search • Download OWL • Download TXT • Change ontology • Logout

Figura 1. Visualizzazione di una schermata del sistema TAV. Si notino le funzionalità disponibili nella finestra *Options*.

Riferimenti:

- F. SCLANO AD P. VELARDI “TERMEXTRACTOR: A WEB APPLICATION TO LEARN THE COMMON TERMINOLOGY OF INTEREST GROUPS AND RESEARCH COMMUNITIES “ 9TH CONF. ON TERMINOLOGY AND ARTIFICIAL INTELLIGENCE TIA 2007, SOPHIA ANTINOPOLIS, OCTOBER 2007
- P. Velardi, A. Cucchiarelli and M. Petit “A Taxonomy learning Method and its Application to Characterize a Scientific Web Community “ IEEE Transaction on Data and Knowledge Engineering (TDKE), vol. 19, n. 2, February 2007, pp. 180-191

| | | | | |
|--------------|--------------|---|-----|-------|
| PROGETTO LC3 | Revisione n* | 0 | Del | ----- |
|--------------|--------------|---|-----|-------|