UNIVERSITÀ DI PISA

Corso di Laurea in Informatica Umanistica

RELAZIONE

Analisi, estrazione e conversione in OntoLex-Lemon dei tratti semantici del lessico computazionale della lingua italiana PSC

Candidato: *Isabel Santucci*

Relatore: *Alessandro Lenci*

Correlatore: *Felice Dell'Orletta*

Correlatore: *Emiliano Giovannetti*

Anno Accademico 2021-2022

Indice

Elenco delle figure

Elenco delle tabelle

1. Introduzione

La presente relazione documenta il lavoro di tesi svolto per l'estrazione di un insieme di tratti semantici contenuti nel lessico computazionale della lingua italiana "Parole-Simple-Clips" (PSC) e la loro conversione sotto forma di triple RDF secondo il modello *OntoLex-Lemon*.

PSC costituisce una risorsa digitale unica per la lingua italiana, la cui ricchezza e varietà di contenuti linguistici è stata sottolineata nella relazione attraverso il confronto della sua componente semantica con alcune caratteristiche delle risorse basate sul modello di "Wordnet".

Le risorse confrontate, modellate con Wordnet, prevedono la rappresentazione di un concetto mediante dei *synset*, ovvero attraverso l'insieme delle parole sinonime che lo denotano e mediante alcune relazioni semantiche presenti tra *synset*, tra cui sinonimia, meronimia e antonimia. La struttura di PSC si basa sulla teoria del Lessico Generativo di Pustejovsky, secondo la quale il significato è determinato secondo i ruoli (*qualia*) che possono essere di di quattro tipi: ***formal role***, ***constitutive role***, ***agentive role*** e ***telic role***. Dopo aver analizzato le principali differenze tra i due modelli, la relazione prosegue con una breve rassegna delle tecnologie del Web Semantico e dei *Linked Data*, funzionali al lavoro successivo di conversione dei dati di PSC.

I *Linked Data* costituiscono un paradigma concepito per collegare dati strutturati sul Web e attraverso il quale è stato possibile dare vita al cosiddetto Web Semantico. Il paradigma alla base della strutturazione dei dati nel Web Semantico è RDF (Resource Description Framework), standard W3C che è brevemente descritto nella relazione. Come modello lessicale di riferimento per la versione *Linked Data* di PSC è stato adottato *OntoLex-Lemon*, sviluppato in seno al W3C Ontology Lexicon Community Group.

Una volta analizzata la risorsa oggetto di studio (attualmente disponibile come database relazionale) e le tecnologie del Web Semantico necessarie alla sua rappresentazione nel nuovo formato, il lavoro di tesi è proseguito con l'estrazione dal database di PSC di una parte dei tratti semantici mediante query SQL, seguita dalla conversione in triple RDF-lemon dei dati estratti attraverso un programma Python sviluppato appositamente.

Il set di triple RDF ottenuto è stato infine integrato nella risorsa PSC già parzialmente convertita nel formato *OntoLex-Lemon* su database GraphDB presso l'Istituto di Linguistica Computazionale del CNR, dove verrà utilizzata a supporto di task di ricerca *full-text* sui testi del Talmud babilonese tradotti in italiano, come descritto nell'ultima sezione dell'elaborato.

2. Il Lessico Computazionale PSC

Il lessico computazionale della lingua italiana "Parole-Simple-Clips", abbreviato in PSC, è stato sviluppato tra il 1996 e il 2003 presso l'Istituto di Linguistica Computazionale "Antonio Zampolli" del Consiglio Nazionale delle Ricerche (ILC-CNR) (Ruimy et al., 2002).

Questa risorsa, che in virtù della particolare articolazione multilivello dei dati linguistici che veicola costituisce un *unicum* tra le risorse lessicali digitali per la lingua italiana, è stata sviluppata nell'ambito di progetti distinti, tra cui si segnalano: "LE-PAROLE" (parte morfologica e sintattica), "LE-SIMPLE" (modello linguistico e parte semantica); CLIPS (aggiunta del *layer* fonologico e aumento delle entrate per la parte semantica e sintattica con la collaborazione della società privata Thamus).

PSC fa parte dei "lessici SIMPLE" (*Semantic Information for Multipurpose Plurilingual Lexica*) sviluppati per 12 lingue nell'ambito di progetti europei. I lessici SIMPLE si caratterizzano per:

- l'attenzione riservata alla modellazione del significato, secondo una teoria e un'architettura comuni;
- l'applicazione di un modello condiviso a più lingue;
- la metodologia comune adottata per la costruzione di lessici combinando strategie *top down* e *bottom up*.

Attualmente la risorsa è disponibile come database MySQL[1]. L'informazione linguistica di PSC è strutturata in quattro livelli distinti ma fortemente interconnessi tra loro: fonologico, morfologico, sintattico e semantico. Il livello semantico è basato sul modello del Lessico Generativo di James Pustejovsky (1995) fondato sulla "Qualia Structure" (QS), composta da quattro ruoli sulla base dei quali ogni "unità semantica" (che in PSC denota il senso di una parola) può essere descritto attraverso un approccio composizionale e per mezzo di numerose relazioni semantiche.

[1] https://dspace-clarin-it.ilc.cnr.it/repository/xmlui/

I ruoli Qualia possono essere di quattro tipi:

- il ***formal role*** che identifica un'entità in mezzo ad altre entità e ne indica la posizione all'interno dell'ontologia dei tipi. La relazione semantica di riferimento è quella iperonimica di "is_a" per i nomi e le entità che connotano eventi.
- il ***constitutive role*** che esprime la costituzione dell'entità analizzata e quella dei suoi elementi costitutivi attraverso le relazioni: "is_a_member_of", "is_a_part_of", "resulting_state", "has_a_property".
- l'***agentive role***, che fornisce le informazioni sull'origine e la direzione dell'entità considerata. Alcune delle relazioni tipicamente usate nel lessico sono: "created_by", "result_of", "caused_by", "agentive_cause", "agentive_experience".
- il ***telic role*** che specifica la funzione dell'entità. Alcune tra le relazioni teliche più utilizzate sono: "used_as", "used_for", "object_of_the_activity", "indirect_telic".

Le relazioni Qualia possono essere combinate tra loro per descrivere *tipi semantici*, strutturati in un sistema definito con il nome di SIMPLE Ontology (Lenci et al. 2000) e costituito da 157 concetti, alcuni dei quali sono stati mutuati dalla risorsa EuroWordNet.

Il livello semantico di PSC è descritto anche attraverso i cosiddetti *template*, schemi predefiniti che veicolano informazione di un dato tipo semantico, e quindi delle unità semantiche che ad esso riferiscono.

Prendiamo l'esempio della parola "gatto" che può essere intesa sia come animale sia come un tipo di frusta, presentandosi perciò, in PSC, sotto forma di due distinte unità semantiche legate alla medesima parola. La parola "gatto", oltre a essere composta da una serie di fonemi, sarà associata a due unità semantiche, una delle quali sarà a sua volta collegata ad un template di tipo "ANIMAL". Questa formalizzazione attraverso i template può rivelarsi estremamente utile sia per ricerche mirate all'interno della risorsa, sia per ricerche sul testo che sfruttino PSC (come si vedrà nella sezione 7) in quanto lessico di supporto.

Numerosi sono stati i lavori di ricerca che, negli anni, hanno interessato PSC, tra cui si citano:

- il popolamento dello strato fonologico (Monachini et al. 2004) con dati appartenenti al DMI (Dizionario Macchina Italiano, Calzolari et al. 1983);
- il collegamento con i concetti di ItalWordNet (Roventini et al. 2007) (cfr. sezione 3.1);
- l'arricchimento della rappresentazione semantica con informazioni relazionali tra eventi e partecipanti (Ruimy, 2010);
- la rappresentazione dell'ontologia SIMPLE in OWL (Monachini e Toral 2007; Piccini et al. 2014);
- una prima conversione del PSC secondo il modello Lemon (Del Gratta et al. 2014);
- uno studio di PSC dedicato al fenomeno della polisemia (Frontini et al. 2014, Khan e Frontini 2015)
- una applicazione che usa PSC a supporto di ricerche *full-text* (Giovannetti et al. 2021)

3. PSC e WordNet

La risorsa lessicale digitale maggiormente utilizzata in ambito linguistico computazionale è WordNet. Per questo motivo, dopo una breve descrizione di tale risorsa, saranno forniti alcuni elementi di confronto tra PSC e WordNet utili a evidenziarne le differenze sostanziali, soprattutto in relazione all'uso che si intende fare di PSC come risorsa di supporto per la ricerca linguistico-semantica sul testo (cfr. sezione 7).

3.1 WordNet

WordNet (WN), nella sua accezione originaria di risorsa sviluppata presso l'Università di Princeton, è un database semantico-lessicale per la lingua inglese basato sul concetto di *synset*, definito come un insieme di sinonimi che denotano uno stesso concetto (Fellbaum, 1998).

Le origini di WordNet possono essere fatte risalire alle cosiddette "reti semantiche", sviluppate a cavallo tra gli anni '50 e '60, che prevedevano una rappresentazione formale della conoscenza mediante grafi i cui nodi rappresentavano concetti e gli archi relazioni semantiche tra di essi.

Le ricerche di Quillian sulla memoria associativa possono essere considerate come il punto di partenza per la concettualizzazione delle reti semantiche (Quillian, 1963). Quillian intendeva realizzare un modello che riproducesse un'organizzazione della memoria semantica di un essere umano.

La risorsa WordNet originaria ha dato origine a numerose altre risorse, tutte basate sul medesimo modello di Quillian. Per quanto riguarda la lingua italiana si cita ItalWordNet (IWN), un database semantico-lessicale, anch'esso sviluppato presso l'ILC-CNR, nell'ambito dei progetti di ricerca EuroWordNet e Sistema Integrato per il Trattamento Automatico del Linguaggio (SI-TAL).

Il modello WN struttura i propri dati lessicali attraverso alcune relazioni semantiche fondamentali:

- **Sinonimia**: si ha quando una parola (o espressione) esprime un significato simile a quello di un'altra parola. In WN la sinonimia è fondamentale poiché se due o più parole presentano questo tipo di relazione allora appartengono allo stesso *synset*;

- **Antonimia**: è una relazione di opposizione semantica tra due parole;
- **Iponimia**/**Iperonimia**: un iponimo è una parola di significato più specifico rispetto ad un'altra più generale (es. in italiano "armadio" è iponimo di "mobile"); un iperonimo, per contro, è una parola che ha un significato più generale rispetto ad altre più specifiche (es. "animale" è iperonimo di "gatto" e di "cane");
- **Meronimia**/**Olonimia**: un meronimo è una parola che indica una parte rispetto a un tutto (es. "pagina" è meronimo di "libro"); per contro, l'olonimia indica la relazione tra una parola che indica l'intero e una parola che ne rappresenta una parte (es. "albero" è olonimo di "corteccia" o "tronco").

3.2 Differenze tra PSC e WordNet

Per fornire un'idea diretta della sostanziale differenza nella rappresentazione dell'informazione lessicale tra PSC e WordNet si consideri l'esempio della parola "*lancet"* (che figura con la parola italiana "lancetta" nella risorsa PSC) (Bel et al., 2000, *Simple*, 4).
Si prenda in considerazione la descrizione di uno dei sensi della parola *lancet*:

lancet, lance
=> surgical knife
=> knife
=> edge tool
=> cutter, cutlery, cutting tool
=> cutting implement
=> tool
=> implement
=> instrumentality, instrumentation
=> artifact, artefact
=> object, physical object
=> entity, something
=> surgical instrument
=> medical instrument
=> instrument
=> device

=> instrumentality, instrumentation
=> artifact, artefact
=> object, physical object
=> entity, something

Una delle caratteristiche più conosciute riguardo allo stile di rappresentazione è che i nodi della gerarchia *isa* si riferiscono alle varie ed eterogenee tipologie di informazioni. Per esempio si trova per *lancet* ("*a surgical knife with a pointed double-edge blade; used for punctures and small incisions*": un coltello chirurgico con una lama appuntita e a doppio taglio; usato per punture e piccole incisioni), vi è l'informazione riferita agli aspetti essenziali di *lancet* ("*edge tool*": utensile a taglio); due righe dopo, si trova invece l'informazione che si riferisce allo scopo solitamente associato a *lancet* ("*cutting implement*": strumento da taglio). Proseguendo si trova l'informazione sull'origine di *lancet* ("*artifact*": artefatto). Infine si hanno altre informazioni rilevanti, per esempio che *lancet* appartiene all'ambito della chirurgia, ampliando così la tassonomia. Di conseguenza, anche se l'entrata di Wordnet contiene le informazioni maggiormente importanti per caratterizzare il senso di *lancet,* questo tipo di informazione non è totalmente esplicita, e perciò non è direttamente e facilmente accessibile alle applicazioni. Inoltre differenti tipi di informazione non hanno una posizione fissa all'interno della gerarchia *isa*, così quello stesso tipo di informazione (per esempio l'informazione che concerne il tipico scopo dell'artefatto, o il materiale di cui è fatto) potrebbe essere collocata a livelli differenti della gerarchia per le differenti entrate. Questo fatto rappresenta sicuramente un'altra causa di una potenziale difficoltà per quelle applicazioni che hanno bisogno o che vogliono ottenere specifiche informazioni semantiche.
Diversamente SIMPLE, seguendo i principi del Generative Lexicon, organizza i vari tipi di informazione inserendosi dentro la caratterizzazione di un senso di una determinata parola, come se fosse al di sopra del template *instrument*[2].
Inoltre, ogni informazione semantica è anche scritta e inserita dentro gerarchie strutturate, ognuna caratterizza esplicitamente un certo aspetto del contenuto dei nomi, verbi e aggettivi. In questo modo, l'informazione semantica identifica il senso della parola in maniera esplicita, e può essere direttamente e selettivamente

[2] Nella risorsa il template è indicato come Temp535.

indirizzato dalle applicazioni NLP. Infine, diversamente dallo stile organizzativo di WordNet, l'informazione lessicale in SIMPLE è strutturata in termini di rete semantica piccola, locale, che opera in combinazione con le informazioni principali, una ricca descrizione della struttura dell'argomento e preferenze selettive delle entrate predicative.

La tabella sottostante è relativa all'unità semantica (nel seguito *SemU*) della parola "lancetta" per il senso di *lancet* appena accennato, esemplificando il template *instrument*.

Usem:	Lancet
BC number:	
Template_Type:	[Instrument]
Unification_path:	[Concrete_entity| $Artifact_{Agentive}$ | Telic]
Domain:	*Medicine*
Semantic Class:	*Instrument*
Gloss:	a surgical knife with a pointed double-edge blade; used for punctures and small incisions
Pred_Rep:	<Nil>
Selectional Restr.:	<Nil>
Derivation:	<Nil>
Formal:	*isa*(<lancet>,<knife>: [Instrument])
Agentive:	*created_by* (<lancet>, <metal>: [Substance]) *has_as_part*(<lancet>, <edge>: [Part])
Constitutive:	*made_of* (<lancet>, <metal>: [Substance]) *has_as_part* (<lancet>, <edge>: [Part])
Telic:	*used_for* (<lancet>, <cut>: [Constitutive_change]) *used_by* (<lancet>, <doctor>)
Synonimy:	*Collocates* (<Usem1>,..,>Usemn>)
Complex:	<Nil>

Tabella 1. Esempio di codifica di un senso della parola "lancetta" (lancet) in PSC.

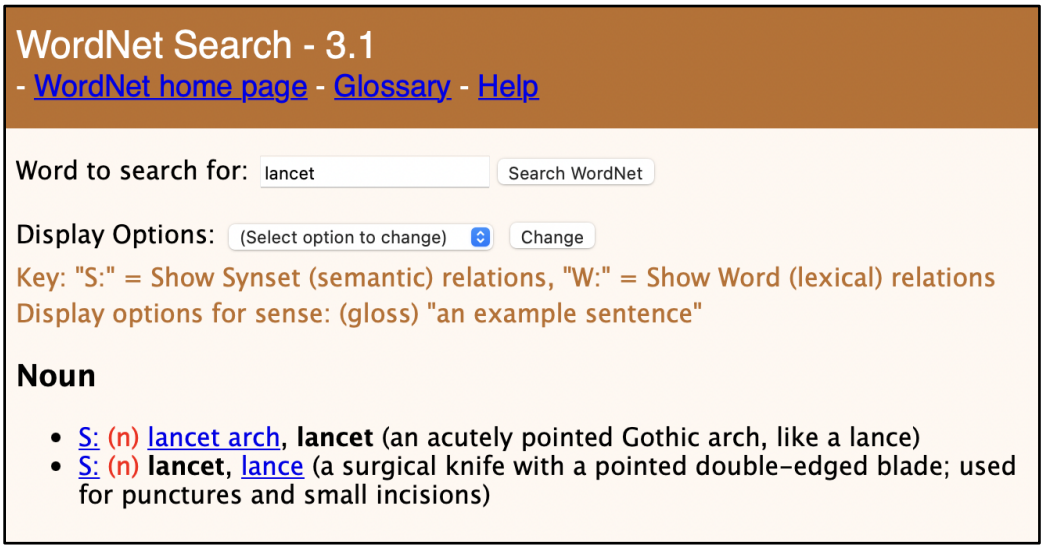

Figura 1. L'entrata della parola "lancet" nella WordNet inglese.

Nella figura 1 si riporta un esempio della parola "*lancet*" così come appare codificata nella WN inglese e visualizzata nell'interfaccia di ricerca disponibile online[3].
Nella tabella 2 sono schematizzate le principali differenze tra WN e PSC: sebbene entrambe le risorse descrivano e strutturino un lessico, i modelli di rappresentazione del dato lessicale sono molto diversi e, soprattutto, presentano sostanziali differenze nella tipologia di informazione linguistica veicolata.

WordNet	**PAROLE-SIMPLE-CLIPS**
La descrizione del significato delle parole è affidata ai *synsets*.	Oltre alla semantica, vi si rappresentano informazioni linguistiche di tipo sintattico, morfologico e fonologico;
Le relazioni descritte comprendono l'iponimia, la sinonimia, l'antonimia e la meronimia.	Appaiono descritte numerose relazioni semantiche. Si segnala inoltre la modellazione della polisemia.
Non è presente descrizione morfologica delle singole forme.	Sono descritte tutte le forme flesse di un dato lemma.

[3] http://wordnetweb.princeton.edu/perl/webwn

L'entrata è descritta anche attraverso le sue relazioni lessico-semantiche e il suo impiego nella creazione di sintagmi specifici.	Raggruppamento dei sensi in insiemi definiti attraverso la SIMPLE Ontology.

Tabella 2. Principali differenze tra WN e PSC.

3.3 Da Wordnet e PSC ai Linked Data

Date queste differenze tra Wordnet e PSC, si comprende il vantaggio nell'utilizzare questo tipo di risorsa. Tuttavia la presentazione in un database MySQL risulta particolarmente complessa e perciò è stata necessaria una conversione secondo il paradigma dei *linked data* e del web semantico, trattati più in dettaglio nella seguente sezione.

4. Linked Data e Web Semantico

Al fine di estrarre i tratti semantici da PSC (descritti nella sezione 5) e convertirli secondo le specifiche già chiarite è stato innanzitutto necessario studiare i fondamenti dei *Linked Data*, del Web Semantico e del modello *OntoLex-Lemon*.

4.1 Linked Data

I *Linked Data* costituiscono un paradigma per la pubblicazione sul Web di dati strutturati, interconnessi tra loro e collegabili con altri dati. L'insieme di dati connessi gli uni agli altri, quando opportunamente descritti secondo i principi esposti nella prossima sezione, consente la definizione di una rete di dati elaborabili della macchina, detta *Semantic Web*.
Quando i *Linked Data* collegano dati aperti (*open data*) si hanno i ***Linked Open Data*** **(LOD)**.

4.2 Il Web Semantico

Il Web Semantico è costituito dall'insieme dei dati connessi secondo il paradigma dei *Linked Data* codificati in modo tale da poterne descrivere la semantica dei dati e facilitare la loro interrogabilità da parte della macchina.

Tim Berners-Lee presentò per la prima volta i *Linked Data*[4] nel 2009 e definì il *Semantic Web* come un'estensione del Web attuale; il *framework* RDF consente di descrivere e analizzare risorse concepite per il Web Semantico.

[4] https://fontistoriche.org/linked-data/

4.2.1 RDF

RDF (*Resource Description Framework*) è uno standard W3C utilizzato per lo scambio di dati sul Web. Secondo tale modello, una risorsa viene identificata tramite un URI[5] e descritta utilizzando il *data model* RDF che si basa sui concetti di:

- ***Resource*** (risorsa): ciò che viene descritto mediante RDF e può essere una risorsa Web (una pagina HTML, un documento XML) o anche una risorsa esterna al Web (un libro).
- ***Property*** (proprietà): una proprietà, un attributo o una relazione utilizzata per descrivere una risorsa.
- ***Statement*** (espressione): l'elemento che descrive la risorsa ed è costituito da un soggetto (rappresentante la *Resource*), un predicato (che indica la *Property*) e da un oggetto (*Value*) che indica il valore della proprietà.

Una struttura dati RDF può essere rappresentata tramite un grafo dove gli archi indicano il collegamento nominale tra le risorse che costituiscono i nodi del grafo.

4.2.2 OntoLex-Lemon

OntoLex-Lemon è un modello per la rappresentazione di lessici computazionali sviluppato dal *W3C* Ontology-Lexica Community Group[6].
Questo modello è stato inizialmente concepito per associare informazione linguistica alle ontologie. Oltre al modulo *Core* del modello *OntoLex-Lemon*, schematizzato in figura 2, sono disponibili ulteriori moduli atti a rappresentare aspetti linguistici specifici (per esempio il modulo *synsem* per l'interfaccia sintassi-semantica). In questa sede considereremo principalmente il vocabolario definito dal modulo *Core* di *OntoLex-Lemon*.

[5] URI (*Uniform Resource Identifier*) è una sequenza di caratteri (senza spazi all'interno) che viene utilizzata per identificare una risorsa disponibile in rete.
[6] https://www.w3.org/community/ontolex/

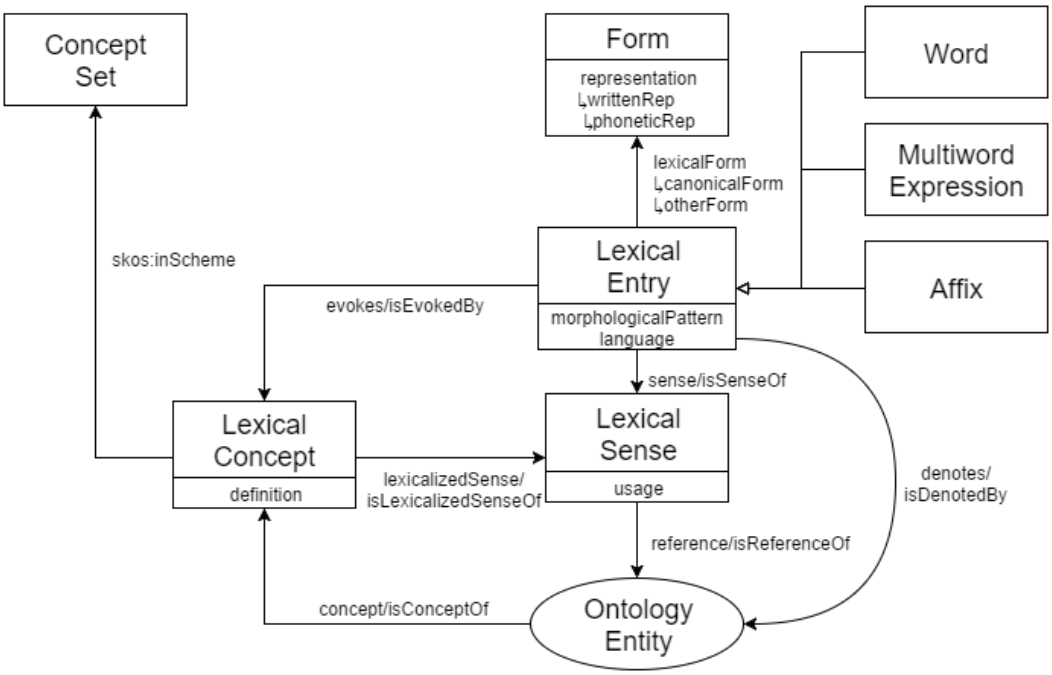

Figura 2. Il modulo *Core* di *OntoLex-Lemon*.

Nella parte centrale dello schema di figura 2 vi sono tre entità: ***Form***, ***Lexical Entry*** e ***Lexical Sense***.

L'entrata lessicale (***Lexical Sense***) è l'astrazione di un lessema le cui forme ortografiche sono rappresentate nella classe ***Form*** (per esempio, "andare" e "vado" sono forme). La forma "andare" è legata alla *Lexical Entry* attraverso la relazione "*canonical form*", che in questo caso definisce il lemma, il quale costituisce la forma dell'entrata da dizionario *Lexical Entry*. Inoltre, quest'ultima si lega al *Lexical Sense*, ovvero la classe che rappresenta i sensi, tramite la relazione "*sense*". La classe indicata come *Ontology Entity*, invece, è esterna al modello *OntoLex-Lemon* e indica i concetti riferiti dai sensi attraverso la relazione "*reference*".

La ***Lexical Entry*** può essere compresa nelle classi: ***Word***, ***Multiword Expression*** e ***Affix***. ***Word*** si riferisce a una parola singola, per esempio "cane". ***Multiword expression*** comprende espressioni composte da due o più elementi, come nel caso di "Presidente della Repubblica"[7] e ***Affix***, infine, rappresenta gli affissi, ovvero un elemento che aggiunto a una parola permette di formarne un'altra di significato differente (ad esempio prefissi, infissi, suffissi, ecc.), come il prefisso "anti".

4.2.2.1 Esempio di codifica di "andare"

Si fornisce di seguito un esempio di codifica in *OntoLex-Lemon,* relativa al verbo italiano "andare".

Innanzitutto, è necessario definire un'istanza della classe "LexicalEntry" (in particolare della sottoclasse "*Word*") dotata di un proprio URI, come ad esempio "lex:andareLexEntr"[8].

Nella classe "***Form***" si definiscono le seguenti istanze relative ad altrettante forme grammaticalmente ammissibili per l'entrata lessicale in questione: "andareForm", "vadoForm", "andreiForm" (per semplicità si omette, nel resto dell'esempio, di indicare i *namespace*).

Nella classe "***Lexical Sense***" si definiscono le seguenti istanze: "andareSense1" (inteso come verbo di movimento) e "andareSense2" (con il significato di "andare fuori di testa").

Si collegano le istanze inserite nelle tre classi attraverso le relazioni definite dal modello:

canonicalForm

andareLexEntr ----------> "andareForm"

otherForm

andareLexEntr ----------> "vadoForm"

otherForm

[7] La definizione fornita in questa sede di "Multiword Expression" è puramente funzionale alla descrizione del modello, che si dichiara agnostico rispetto alle teorie linguistiche, e non mira a descrivere esaurientemente i diversi punti di vista teorici che hanno preso in esame le ME. In linea generale, una multiword expression si caratterizza come insieme di parole percepite come unità, ma il cui significato unitario non è direttamente inferito dal significato dei medesimi termini e con regole specifiche di ordine e coesione.

[8] Specifichiamo che sia namespace ("lex") che l'URI nella sua interezza sono puramente funzionali all'esempio.

andareLexEntr ----------> andreiForm

sense

andareLexEntr ----------> andareSense1

sense

andareLexEntr ----------> andareSense2

Supponiamo di avere un'ontologia di riferimento che comprende, tra gli altri, i due seguenti concetti:

"ToGo" e "ToFreakOut" rappresentati da istanze di "***Ontology Entity***"

Il collegamento tra i due sensi del verbo andare e questi due concetti viene realizzato mediante la relazione "*reference*" di *OntoLex-Lemon*:

reference

andareSense1 ----------> ToGo

reference

andareSense2 ----------> ToFreakOut

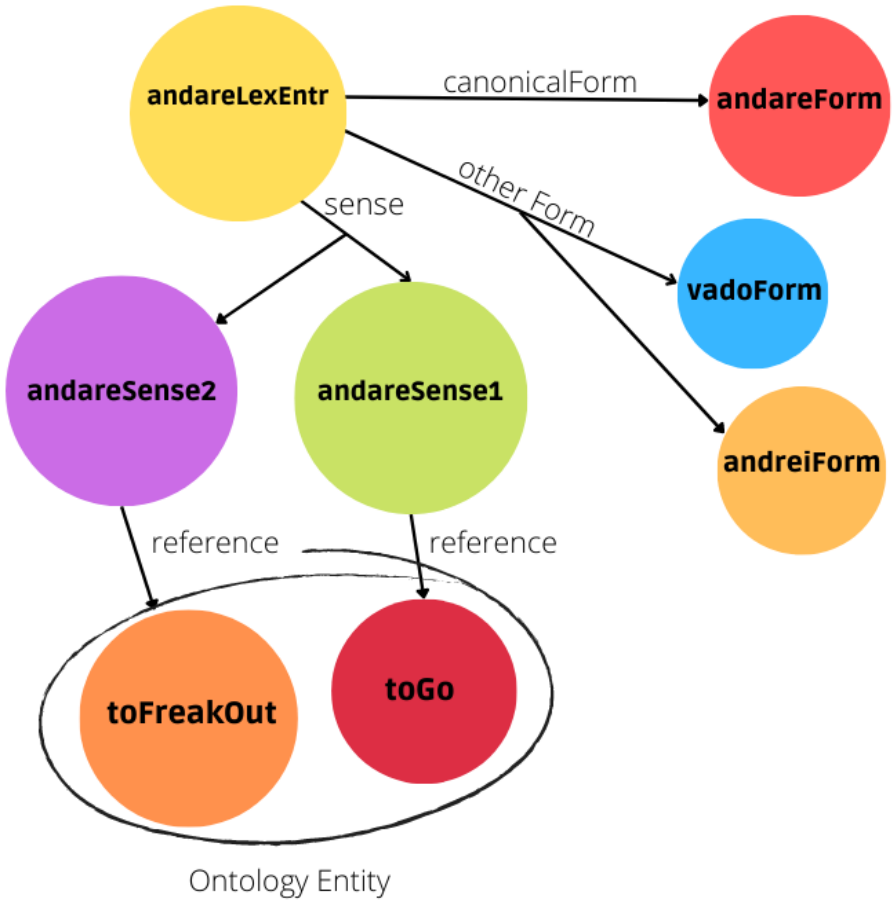

Figura 3. Rappresentazione grafica dell'esempio riportato.

5. I tratti semantici di PSC

In questa sezione si fornisce un breve approfondimento dei tratti semantici presenti nella risorsa PSC. I tratti costituiscono una fonte di dati particolarmente interessante e distintiva di tale risorsa, la cui utilità in un contesto applicativo sarà illustrata nella sezione 7.

Un "tratto semantico" costituisce una formalizzazione all'interno di PSC di un'informazione semantica attribuita ad un'unità semantica.

I tratti semantici (che sul database di origine sono codificati nella tabella "traits", cfr. sezione 6, figura 4) comprendono diversi tipi di informazione semantica. Risulta arduo fornire una classificazione precisa e coerente di tutti i tratti semantici. I tratti possono infatti comprendere informazioni riferite ai template collegati a una *SemU* (per esempio nel caso dei *supertype*, che chiariscono la classe di appartenenza di un dato template; data ad esempio a *SemU* per "tigre" come "tipo di felino", si avrà l'associazione al template "*Animal*" nella tabella "usemtemplates" e al tratto di supertype "Living Entity") oppure essi possono precisare l'appartenenza di una *SemU* a una data classe semantica o a un dominio, secondo la classificazione fornita dall'ontologia monodimensionale integrata in PSC, LexiQuest[9]

Per questo lavoro di tesi si è scelto di considerare e sottoporre ad analisi solo i tratti semantici legati alla descrizione del significato degli aggettivi.

La SIMPLE ontology contiene dei template specifici per gli aggettivi (descritti in dettaglio nelle due prossime sezioni) secondo la classificazione seguente:

- A1 Intensionale
  - Modal
  - Temporal
  - Emotive
  - Manner
  - ObjectRelated
  - Emphasizer

[9] LexiQuest è un'ontologia integrata in PSC concepita per consentire la modellazione di informazione anche attraverso un'ontologia di tipo tradizionale.

- OrdinalNumeral
- A2 Estensionale
- Physical Property
- Psychological Property
- Social Property
- Temporal Property
- Intensifying Property
- Relational Property
- Numeral Property

5.1 Aggettivi intensionali

Gli aggettivi intensionali manipolano il parametro temporale o modale che è rilevante per l'interpretazione dei nomi con cui sono combinati. Di seguito si descrivono le classi dei template per gli aggettivi intensionali, a cui sono collegati i tratti oggetto dell'estrazione.

Gli aggettivi temporali (***temporal adjectives***) indicano una connotazione temporale riferita al nome, come nel caso della frase "il *vecchio* contatore dell'acqua", la quale denota che l'oggetto in questione forniva una determinata funzione in un tempo passato.

Gli aggettivi modali (***modal adjectives***) esprimono una modalità riferita al giudizio di chi parla riguardo alla probabilità della proposizione di essere vera, come nel caso dell'esempio "*pretendente*", cioè qualcuno che potrebbe avere la possibilità di ottenere quel ruolo. In generale riflette il giudizio del parlante sul fatto che la proposizione possa essere vera.

Gli aggettivi ***emotive*** esprimono un forte atteggiamento emotivo riguardo l'oggetto denotato dal nome, come per "uomo *povero*".

Gli aggettivi appartenenti alla classe ***manner*** modificano l'evento associato al nome. Infatti nel caso di "ballerino *bello*" si introduce un'informazione che modifica o aggiunge indicazioni all'evento in questione.

Gli aggettivi ***object-related*** sono in molti casi derivanti a livello morfologico da nomi, come nell'esempio di "avvocato *criminale*".

Gli aggettivi ***emphasize*** hanno un effetto generale di intensificazione che evidenzia l'importanza di appartenenza alla categoria descritta dal sostantivo, come "*grande* vittoria".

5.2 Aggettivi estensionali

Gli aggettivi estensionali si dividono in sei classi associate a un template e conseguentemente a un insieme di tratti, come accade per gli intensionali. Di seguito si descrivono i sei template che li raggruppano.

Gli aggettivi ***psychological property*** riguardano qualità di tipo psicologico, morale o legate all'interiorità, per esempio: triste, meschino, freddo, debole etc. Al contrario gli aggettivi ***physical property*** fanno riferimento a uno stato esteriore come: magro, grasso, pulito, sudicio etc.

Gli aggettivi ***social property*** sono legati al rapporto con un altro individuo: inglese, ligure, canadese, tedesco etc.

Gli aggettivi denominati ***temporal property*** fanno riferimento a questioni temporali: perenne, provvisorio, transitorio, definitivo etc.

Gli aggettivi ***intensifying*** portano ad un'intensificazione come: grande, forte, acuto, violento, profondo etc.

Gli aggettivi ***relational*** sono relativi al legame: uguale, identico, compatibile, contrario etc.

6. Estrazione e conversione dei tratti semantici

In questa sezione è descritta la procedura che ha portato ad estrarre i tratti semantici scelti e, successivamente, a convertirli in triple RDF secondo il modello *OntoLex-Lemon*.

L'immagine mostrata in figura 4 riporta i dati presenti nel database nella tabella "traits" che si riferisce ai tratti semantici.

idTrait Row unique identifier	trait Semantic trait	traitPrefix Trait prefix	hierarchyDom Hierarchy Domain	orderDom Order Domain	idTemplate Semantic template identifier	order Order
M1018	PsychologicalProperty@Positive	WVSFMeaningCompPsychologicalPropertyPositive	1	*NULL*	PsychologicalProperty	1
M1019	PsychologicalProperty@Negative	WVSFMeaningCompPsychologicalPropertyNegative	1	*NULL*	PsychologicalProperty	1
M1020	PsychologicalProperty@Neutral	WVSFMeaningCompPsychologicalPropertyNeutral	1	33	PsychologicalProperty	1
M1021	Experience@Positive	WVSFMeaningCompExperiencePositive	1.1	*NULL*	PsychologicalProperty	1
M1022	Experience@Negative	WVSFMeaningCompExperienceNegative	1.1	11	PsychologicalProperty	1
M1023	Experience@Neutral	WVSFMeaningCompExperienceNeutral	1.1	*NULL*	PsychologicalProperty	1
M1024	PsychState@Positive	WVSFMeaningCompPsychStatePositive	1.2	*NULL*	PsychologicalProperty	1
M1025	PsychState@Negative	WVSFMeaningCompPsychStateNegative	1.2	*NULL*	PsychologicalProperty	1
M1026	PsychState@Neutral	WVSFMeaningCompPsychStateNeutral	1.2	34	PsychologicalProperty	1
M1027	Cognition@Positive	WVSFMeaningCompCognitionPositive	1.3	319	PsychologicalProperty	1
M1028	Cognition@Negative	WVSFMeaningCompCognitionNegative	1.3	315	PsychologicalProperty	1
M1029	Cognition@Neutral	WVSFMeaningCompCognitionNeutral	1.3	317	PsychologicalProperty	1
M1030	AttitudeSalience@Positive	WVSFMeaningCompAttitudeSaliencePositive	1.4	297	PsychologicalProperty	1
M1031	AttitudeSalience@Negative	WVSFMeaningCompAttitudeSalienceNegative	1.4	294	PsychologicalProperty	1
M1032	AttitudeSalience@Neutral	WVSFMeaningCompAttitudeSalienceNeutral	1.4	296	PsychologicalProperty	1
M1033	AttitudeEvaluation@Positive	WVSFMeaningCompAttitudeEvaluationPositive	1.5	306	PsychologicalProperty	1
M1034	AttitudeEvaluation@Negative	WVSFMeaningCompAttitudeEvaluationNegative	1.5	268	PsychologicalProperty	1
M1035	AttitudeEvaluation@Neutral	WVSFMeaningCompAttitudeEvaluationNeutral	1.5	292	PsychologicalProperty	1
M1036	Moral@Positive	WVSFMeaningCompMoralPositive	1.5.1	*NULL*	PsychologicalProperty	1
M1037	Moral@Negative	WVSFMeaningCompMoralNegative	1.5.1	*NULL*	PsychologicalProperty	1

Figura 4. La tabella "traits" nel database MySQL che codifica la risorsa PSC.

La prima colonna denominata "**idTrait**", riporta un codice alfanumerico identificativo per ogni tratto presente nella tabella. Successivamente si ha la colonna "**trait**", che specifica il nome del tratto. Nella terza colonna "**traitPrefix**" si trovano i tratti "WVSF", cioè *Weight Value Semantic Feature* (Bel et al., 2000, 37). "**HierarchyDom**" è la gerarchia dei domini e riguarda la posizione dei tratti nella SIMPLE ontology. L"**orderDom**" è l'ordine dei domini, il quale non viene sempre descritto per ogni tratto. Infine si trova l"**idTemplate**", cioè la relazione con la tabella "Template" (cfr. figura 7 più avanti).

Le colonne della tabella "traits" che sono state utilizzate per l'estrazione dei tratti sono le seguenti:

- **idTrait**
- **trait**
- **idTemplate**

6.1 L’estrazione dei tratti dal database di PSC

Al fine di estrarre dal database MySQL i dati relativi ai tratti semantici richiesti è stato necessario tenere in considerazione sia la tabella dei tratti (“traits”), la tabella relativa alle unità semantiche (“usem”), la tabella di collegamento tra le due (“usemtraits”), e la tabella “templates”.

idUsem Row unique identifier	naming Lemma	pos Part of speech	exemple Example	definition Lemma definition
USem01934maiale	maiale	N	cucinare delle bistecchine di maiale	la carne del maiale ingrassato e macellato
USem068832tassello	tassello	N	manca un tassello per risolvere il caso	pezzo di qualcosa, concreto o astratto
USem069520accantonamento	accantonamento	N	l'accantonamento di denaro da parte di Leo	l'accantonare
USem069521abbozzare	abbozzare	V	abbozzare un progetto, un disegno	delineare un disegno, uno scritto dandogli una pri...
USem069522abbozzo	abbozzo	N	l'abbozzo di un ritratto	prima forma di un'opera d'arte o di un testo
USem069523dispotismo	dispotismo	N	il dispotismo si oppone alla democrazia	governo di un despota
USem069524dissenso	dissenso	N	tra i parlamentari c'era un gran dissenso	divergenza di opinioni, interessi ecc.
USem069525dissesto	dissesto	N	dissesto familiare	condizione di squilibrio, specialmente economico
USem069526parolaccia	parolaccia	N	dire una parolaccia	parola sconcia, detta per offendere
USem069527abbozzare	abbozzare	V	abbozzare un sorriso	accennare
USem069528dissidenza	dissidenza	N	registrare atteggiamenti di dissidenza	contrasto o conflitto di opinioni all'interno di u...
USem069529dissidenza	dissidenza	N	la dissidenza cubana	l'insieme dei dissidenti
USem069530dissidio	dissidio	N	essere in dissidio con i genitori	contrasto divisione tra due o più persone
USem069531dissipazione	dissipazione	N	la dissipazione del patrimonio	il dissipare, dissiparsi, sperpero
USem069532abbreviazione	abbreviazione	N	l'abbreviazione della procedura da parte di Leo	l'abbreviare
USem069533abbreviare	abbreviare	V	abbreviare una parola	troncare con una abbreviazione; nella metrica clas...
USem069535alito	alito	N	*NULL*	fiato; lieve soffio
USem069536parolina	parolina	N	una parolina d'amore	parola benevola e dolce
USem069537dissolvenza	dissolvenza	N	immagini in dissolvenza	graduale apparizione o oscuramento dell'immagine, ...
USem069538allergia	allergia	N	l'allergia verso gli antibiotici	ipersensibilita' verso certe sostanze
USem069539cifrario	cifrario	N	la biblioteca contiene numerosi cifrari	testo contenente la chiave per comprendere una scr...
USem069540distanza	distanza	N	una distanza di un chilometro tra il punto A e il ...	spazio che intercorre tra due luoghi

Figura 5. La tabella “usem” nel database MySQL di PSC.

In figura 5 è mostrato un estratto del contenuto della tabella “usem”. Nella tabella vi sono diverse colonne caratterizzanti le unità semantiche, ma ai fini del presente lavoro è stato considerato solo l’attributo “**idUsem**”, ovvero il codice alfanumerico distintivo per ogni *SemU*.

In figura 6 sono mostrati alcuni dati della tabella “usemtraits” di raccordo tra le *SemU* e i tratti, i primi riferiti dai succitati identificatori, e i secondi dagli identificatori dei tratti introdotti nella sezione precedente.

idUsem Semantic unit identifier	idTrait Semantic trait identifier
USem01934maiale	T11
USem01934maiale	T35
USem01934maiale	T667
USem01934maiale	T675
USem068832tassello	T293
USem068832tassello	T692
USem068832tassello	T750
USem069520accantonamento	T1273
USem069520accantonamento	T274
USem069520accantonamento	T828
USem069520accantonamento	T834
USem069521abbozzare	T1297
USem069521abbozzare	T1305
USem069521abbozzare	T430
USem069521abbozzare	T782
USem069521abbozzare	T822
USem069521abbozzare	T835
USem069522abbozzo	T1297
USem069522abbozzo	T1305
USem069522abbozzo	T430
USem069522abbozzo	T782
USem069522abbozzo	T822
USem069522abbozzo	T835
USem069523dispotismo	T1273
USem069523dispotismo	T276

Figura 6. La tabella “usemtraits” nel database MySQL.

Infine, nella figura 7 è rappresentata la tabella dei template, della quale sono stati utilizzati gli attributi “template” e “selected”.

idTemplate Row unique identifier	template Semantic Template	pos Part of speech	templateLabel Label	type Template type	selected Selected flag
Temp-1	Undefined	*NULL*	*NULL*	*NULL*	*NULL*
Temp0	Top	N	WVSFTemplateTopPROT	1	0-Top
Temp1	Entity	N	WVSFTemplateEntityPROT	1	1-Entity
Temp2	Telic	N	WVSFTemplateTelicPROT	1	2-Telic
Temp3	Agentive	NV	WVSFTemplateAgentivePROT	1	3-Agentive
Temp31	Cause	NV	WVSFTemplateCausePROT	1	31-Cause
Temp4	Constitutive	N	WVSFTemplateConstitutivePROT	1	4-Constitutive
Temp41	Part	N	WVSFTemplatePartPROT	1	41-Part
Temp411	Body_Part	N	WVSFTemplateBodyPartPROT	1	411-Body_part
Temp42	Group	N	WVSFTemplateGroupPROT	1	42-Group
Temp421	Human_Group	N	WVSFTemplateHumangroupPROT	1	421-Human_group
Temp43	Amount	N	WVSFTemplateAmountPROT	1	43-Amount
Temp5	Concrete_Entity	N	WVSFTemplateConcreteEntityPROT	1	5-Concrete_entity
Temp51	Location	N	WVSFTemplateLocationPROT	1	51-Location
Temp511	D_3_Location	N	WVSFTemplate3DLocationPROT	1	511-3_D_location
Temp512	Geopolitical_location	N	WVSFTemplateGeopoliticalLocationPROT	1	512-Geopolitical_location
Temp513	Area	N	WVSFTemplateAreaPROT	1	513-Area
Temp514	Opening	N	WVSFTemplateOpeningPROT	1	514-Opening
Temp515	Building	N	WVSFTemplateBuildingPROT	1	515-Building
Temp516	Artifactual_area	N	WVSFTemplateArtifactualareaPROT	2	516-Artifactual_area**
Temp52	Material	N	WVSFTemplateMaterialPROT	1	52-Material

Figura 7. La tabella “templates” nel database MySQL.

La query MySQL finale ha perciò coinvolto quattro tabelle del database che sono state messe in “*join*” tra loro:

```
SELECT traits.idTrait, usem.idUsem,
SUBSTRING_INDEX(Trait,'@',+1) as TraitName,
SUBSTRING_INDEX(Trait,'@',-1) as TraitValue,
templates.selected as Template
FROM usem, usemtraits, traits, templates
WHERE usem.idUsem=usemtraits.idUsem and
traits.idTrait=usemtraits.idTrait and
traits.idTemplate=templates.template and traits.idTrait
LIKE 'M%'
```

Da notare che nella parte di SELECT è stata inserita una funzione per il calcolo di sottostringhe che ha permesso di dividere il nome del tratto in due ulteriori colonne: “TraitName” e “TraitValue”. La prima, che riguarda il nome del tratto, si ottiene

selezionando i dati che precedono il carattere "@", mentre il valore del tratto lo si ricava dai dati che seguono tale carattere. Con "templates.selected", sempre nella parte di SELECT, si estraggono i valori del template aggettivale collegato al tratto in questione. Nella parte di WHERE si realizza una *join* attraverso i diversi "id" comuni tra le varie tabelle ("idUsem", "idTrait") e l'attributo "template" della tabella Templates. Infine, con la clausola "`traits.idTrait LIKE 'M%'`" si richiede di ottenere solo gli "idTrait" che cominciano con la lettera "M" (i.e. i tratti aggettivali).

La query SQL descritta ha permesso di estrarre un totale di 2.702 relazioni tra unità semantiche e tratti aggettivali. La tabella 3 riporta alcuni dati estratti a titolo esemplificativo.

idTrait	idUsem	TraitName	TraitValue	Template
M1018	USemD6217libero	PsychologicalProperty	Positive	A22-PsychologicalProperty
M1018	USemD6782indipendente	PsychologicalProperty	Positive	A22-PsychologicalProperty
M1021	USem62000gradevole	Experience	Positive	A22-PsychologicalProperty
M1021	USem62015amabile	Experience	Positive	A22-PsychologicalProperty
M1021	USem62075acceso	Experience	Positive	A22-PsychologicalProperty
M1021	USem62120bello	Experience	Positive	A22-PsychologicalProperty
M1021	USem62129carino	Experience	Positive	A22-PsychologicalProperty
M1021	USem62139stupendo	Experience	Positive	A22-PsychologicalProperty

Tabella 3. Un esempio di dati risultanti dall'esecuzione sul database di PSC dalla query SQL.

6.2 La conversione dei dati estratti in triple RDF

Al fine di convertire i dati estratti dal database MySQL in triple RDF è stato sviluppato un programma in Python.
Il file di input indicato nel codice è nel formato "csv" (*comma-separated values*), in cui vi sono i dati della tabella, ottenuti precedentemente tramite la query, separati da virgole.

6.2.1 Il codice Python e le triple

Il codice Python prodotto è riportato di seguito:

```
import pandas as pd
df=pd.read_csv('Desktop/traits.csv')

for index, row in df.iterrows():
print('lex:'+row['idUsem'],'simple:trait','simple:'+row['TraitName'
]+row['TraitValue'])
print('simple:'+row['TraitName'],'a', 'owl:Class')
print('simple:'+row['TraitName'],'rdfs:label','"'+row['TraitName']+
'"'+'@en')
print('simple:trait','a','owl:objectProperty')
print('simple:trait','rdfs:domain','ontolex:LexicalSense')
print('simple:trait','rdfs:label','"'+'trait'+'"'+'@en')
print('simple:'+row['TraitName']+row['TraitValue'],'a','owl:NamedIn
dividual')
print('simple:'+row['TraitName']+row['TraitValue'],'a','simple:'+ro
w['TraitName'])

print('simple:'+row['TraitName']+row['TraitValue'],'rdfs:label','"'
+row['TraitValue'].lower()+'"'@en')
print('simple:'+row['TraitName']+row['TraitValue'],'dc:identifier',
'"'+str(row['idTrait'])+'"')
print('simple:'+row['TraitName']+row['TraitValue'],'simple:template
','simple:'+row['selected'])
print('simple:template','a','owl:ObjectProperty')
```

Per prima cosa è stata importata ***Pandas***, una libreria di Python per l'analisi dei dati, ed è stato aperto in lettura il file di input con le associazioni *SemU*-tratti nel formato csv ("traits.csv"). Per l'iterazione sui dati in input è stato realizzato un "for", e, ad ogni iterazione relativa alla lettura di una riga del file sorgente, sono state prodotte le corrispondenti triple RDF, come più avanti dettagliato ed esemplificato nella tabella 4.

lex:USemD6217libero	simple:trait	simple:psychologicalPropertyPositive
simple:PsychologicalProperty	a	owl:Class
simple:PsychologicalProperty	rdfs:label	"Psychological property"@en
simple:trait	a	owl:ObjectProperty
simple:trait	rdfs:domain	ontolex:LexicalSense
simple:trait	rdfs:label	"trait"@en
simple:psychologicalPropertyPositive	a	owl:NamedIndividual
simple:psychologicalPropertyPositive	a	simple:PsychologicalProperty
simple:psychologicalPropertyPositive	rdfs:label	"positive"@en
simple:psychologicalPropertyPositive	dc:identifier	"M1018"
simple:psychologicalPropertyPositive	simple:template	simple:TempA22-PsychologicalProperty
simple:template	a	owl:ObjectProperty

Tabella 4. Le triple RDF prodotte a fronte della lettura di una riga del file di input.

Per ognuna delle associazioni tra una *SemU* e i rispettivi tratti semantici, quindi, è stato prodotto un set di triple RDF. E' necessario specificare che è stato fatto riferimento a diversi tipi di vocabolari *Linked Data*, che nelle triple sono indicati con i rispettivi namespace che precedono i due punti. Alcuni vocabolari sono stati creati ex novo (come "lex" e "simple"), mentre altri erano già esistenti ("owl", "rdfs", "dc").

La tripla della tabella 4.1 fa riferimento a un senso, “USemD6217libero”, la cui definizione è già presente in PSC. Questo senso è associato attraverso un predicato, “simple:trait”, ad un altro oggetto “simple:psychologicalPropertyPositive”, che in questo caso rappresenta un tratto di tipo “PsychologicalProperty” con polarità “positiva”.

lex:USemD6217libero	simple:trait	simple:psychologicalPropertyPositive

Tabella 4.1 Tripla che associa un tratto ad un'unità semantica.

Da notare che ogni elemento deve sempre essere definito: in questo caso le definizioni di “simple:trait” e “simple:psychologicalPropertyPositive” sono fornite nelle due triple della tabella 4.2.

simple:PsychologicalProperty	a	owl:Class
simple:PsychologicalProperty	rdfs:label	“Psychological property”@en

Tabella 4.2 Triple che definiscono la classe di un tratto semantico e la relativa etichetta.

Il tratto semantico “PsychologicalProperty” viene rappresentato come una classe presente nel vocabolario “owl”, che, nella riga successiva, viene dotata di un’etichetta “Psychological property”, il nome del tratto nella lingua inglese, mediante il vocabolario “rdfs”.

simple:trait	a	owl:ObjectProperty
simple:trait	rdfs:domain	ontolex:LexicalSense
simple:trait	rdfs:label	“trait”@en

Tabella 4.3 Triple che definiscono la natura della relazione “trait” che collega sensi e tratti.

Nelle triple della tabella 4.3 viene definita la relazione “simple:trait” che lega le *SemU* a un tratto.

Nella prima riga si afferma che “simple:trait” è una proprietà del vocabolario “owl”. Successivamente si stabilisce che il dominio di tale proprietà (attraverso il predicato “rdfs:domain”) è la classe “LexicalSense” di *OntoLex-Lemon*.

Infine, a “simple:trait” viene assegnata un’etichetta (“label”) sempre tramite “rdfs”, che si riferisce a “trait”@en: il nome della proprietà seguito dal codice ISO della lingua inglese.
Riassumendo, con l’ultimo set di triple si è definito che “simple:trait” è una proprietà, che ha come dominio il senso lessicale, e che ha una certa etichetta nella lingua inglese.
Nelle righe mostrate nella tabella 4.4 viene definito il tratto "simple:psychologicalPropertyPositive" attraverso cinque triple. Nella prima e nella seconda esso viene definito come appartenente alla classe degli individui (“NamedIndividual”) di OWL, e alla classe dei tratti “PsychologicalProperty” definita in precedenza[10].

simple:psychologicalPropertyPositive	a	owl:NamedIndividual
simple:psychologicalPropertyPositive	a	simple:PsychologicalProperty
simple:psychologicalPropertyPositive	rdfs:label	“positive”@en
simple:psychologicalPropertyPositive	dc:identifier	“M1018”
simple:psychologicalPropertyPositive	simple:template	simple:TempA22-PsychologicalProperty

Tabella 4.4 Triple che definiscono un tratto semantico con il suo valore.

Inoltre, ad esso viene assegnata anche un’etichetta nella lingua inglese: “positive”. All’individuo “simple:psychologicalPropertyPositive” viene anche associato l'identificatore “M1018” (così come estratto dal database) attraverso il predicato “dc:identifier”, dove “dc” rappresenta il vocabolario “Dublin Core”[11].

Infine, nella tripla in tabella 4.5 viene definita la proprietà OWL “template”, per mezzo della quale ogni tratto viene associato ad uno dei template già presenti nella risorsa.

[10] il token “a” come predicato rappresenta, nella notazione utilizzata, una abbreviazione di “rdf:type”
[11] http://purl.org/dc/terms/identifier

simple:template	a	owl:ObjectProperty

Tabella 4.5 La tripla che definisce la relazione di attribuzione di un template ad un tratto semantico.

In totale sono state prodotte, come output del programma Python, un totale di 32.424 triple RDF.

6.2.2 Verifica e ripulitura dei dati ottenuti

Per analizzare collaborativamente i dati ottenuti è stato necessario importarli in un foglio di calcolo online condiviso. Le componenti "soggetto", "predicato" e "oggetto" delle triple sono state separate su tre colonne utilizzando le funzionalità di importazione del foglio di calcolo dividendo i dati sugli spazi.
Nella tabella 5 sono riportati alcuni dei dati intabellati, tra i quali appare evidente la presenza di triple ripetute generate dal programma, evidenziate in grassetto.

lex:USemD6217libero	simple:trait	simple:PsychologicalProperty Positive
simple:PsychologicalProperty	a	owl:Class
simple:PsychologicalProperty	rdfs:label	"PsychologicalProperty"@en
simple:trait	a	owl:objectProperty
simple:trait	rdfs:domain	ontolex:LexicalSense
simple:trait	rdfs:label	"trait"@en
simple:PsychologicalPropertyPositive	a	owl:NamedIndividual
simple:PsychologicalPropertyPositive	a	simple:PsychologicalProperty
simple:PsychologicalPropertyPositive	rdfs:label	"positive"@en
simple:PsychologicalPropertyPositive	dc:identifier	"M1018"
simple:PsychologicalPropertyPositive	simple:template	simple:A22-PsychologicalProperty
simple:template	a	owl:ObjectProperty

lex:USemD6782indipendente	simple:trait	simple:PsychologicalProperty Positive
simple:PsychologicalProperty	**a**	**owl:Class**
simple:PsychologicalProperty	**rdfs:label**	**"PsychologicalProperty"@en**
simple:trait	**a**	**owl:objectProperty**
simple:trait	**rdfs:domain**	**ontolex:LexicalSense**
simple:trait	**rdfs:label**	**"trait"@en**
simple:PsychologicalPropertyPositive	**a**	**owl:NamedIndividual**
simple:PsychologicalPropertyPositive	**a**	**simple:PsychologicalProperty**
simple:PsychologicalPropertyPositive	**rdfs:label**	**"positive"@en**
simple:PsychologicalPropertyPositive	**dc:identifier**	**"M1018"**
simple:PsychologicalPropertyPositive	**simple:template**	**simple:A22-PsychologicalProperty**
simple:template	**a**	**owl:ObjectProperty**

Tabella 5. I dati intabellati nel foglio di calcolo con le triple ridondanti.

Per eliminare tali ridondanze e quindi rimuovere le triple ripetute è stata utilizzata la funzione di rimozione dei duplicati presente nell'applicazione di *editing* del foglio di calcolo utilizzato. In definitiva, il numero di triple RDF ottenute è risultato esser pari a 3.790.

7. L'uso dei tratti in un contesto applicativo di ricerca sul testo

In questa sezione viene mostrato come i tratti semantici estratti e convertiti siano stati integrati e utilizzati in una applicazione sviluppata presso l'ILC-CNR per la ricerca linguistica sul testo.

In particolare, l'applicazione in questione è stata concepita e realizzata per sperimentare una ricerca *full-text* supportata da un lessico computazionale (Giovannetti et al, 2021). L'esigenza di tale applicazione è emersa nel contesto del progetto di traduzione del Talmud babilonese, in particolare dalla necessità di fornire ai traduttori una modalità avanzata di ricerca testuale su base semantica, come meglio descritto nella prossima sezione.

7.1 PSC a supporto di un task di Full-Text Search

L'applicazione per la ricerca sul testo, come descritto in dettaglio nell'articolo sopra citato, utilizza le seguenti componenti del lessico PSC:

- le unità morfologiche, classificate in base alla loro POS;
- le unità fonologiche che rappresentano le forme flesse;
- le unità semantiche (*SemU*) che descrivono i sensi.

Inoltre, sono state utilizzate le seguenti informazioni contenute nella risorsa:

- i tratti morfologici (per esempio, genere e numero);
- un primo set di relazioni semantiche tra le *SemU*;
- l'associazione tra le *SemU* e i template descritti nella SIMPLE Ontology .

La risorsa utilizzata dall'applicazione è stata convertita in un formato LOD, sulla base del modello *OntoLex-Lemon*, per garantirne l'interoperabilità con altre risorse e favorirne l'utilizzo in diversi *task* nonché per la possibilità di eseguire processi di "*reasoning*" [12] sui dati.

[12] il "reasoning", nell'ambito del settore della rappresentazione della conoscenza, è una tecnica che estrapola dati impliciti dai dati espliciti contenuti in una risorsa, ad esempio sfruttando la transitività di una relazione di "isa" o la simmetria di una relazione di sinonimia.

Tecnicamente, l'applicazione comprende dei servizi di *backend* per l'accesso al lessico PSC e per l'interrogazione del testo. Il testo del Talmud, segmentato in frasi, è memorizzato in un database relazionale, mentre il lessico è codificato in un repository GraphDB[13]. L'accesso ai dati è stato implementato con una serie di servizi REST che possono essere invocati dal *web client*. Sul lato *front-end*, l'interfaccia grafica è stata sviluppata in Angular.

L'applicazione comprende tre modalità di ricerca:

- Parola-chiave (ricerca su forma singola);
- Forma-Lemma (ricerca su tutte le forme di un dato lemma, con possibilità di inserire vincoli morfologici e semantici);
- Template (ricerca "esplorativa" sulla base dei raggruppamenti dati dai template).

Di seguito si riassumono le fasi del processo di ricerca ed espansione della query:

1. L'utente inserisce una prima serie di dati per formulare la query desiderata nell'interfaccia grafica;
2. l'interfaccia interroga i servizi di *backend* del lessico, i quali restituiscono i dati linguistici che corrispondono alla query iniziale;
3. l'utente completa la query, selezionando i dati linguistici desiderati e avvia la ricerca;
4. l'interfaccia esegue la query espansa e invoca, tramite Web API, i servizi di *backend* di accesso al testo che raccolgono, etichettano e restituiscono le porzioni testuali del Talmud nelle quali appaiono le parole cercate;
5. l'interfaccia mostra i risultati all'utente.

7.2 Esempi di ricerca per tratti semantici

In questa sezione si forniscono due esempi di interrogazione al testo del Talmud babilonese finalizzati a mostrare il valore aggiunto fornito dall'inserimento dei tratti in PSC.

Le query mostrate di seguito possono essere riprodotte e testate dal lettore accedendo all'applicazione mediante il *link* in nota[14].

[13] https://www.ontotext.com/products/graphdb/

[14] https://klab.ilc.cnr.it/talmudSearch/

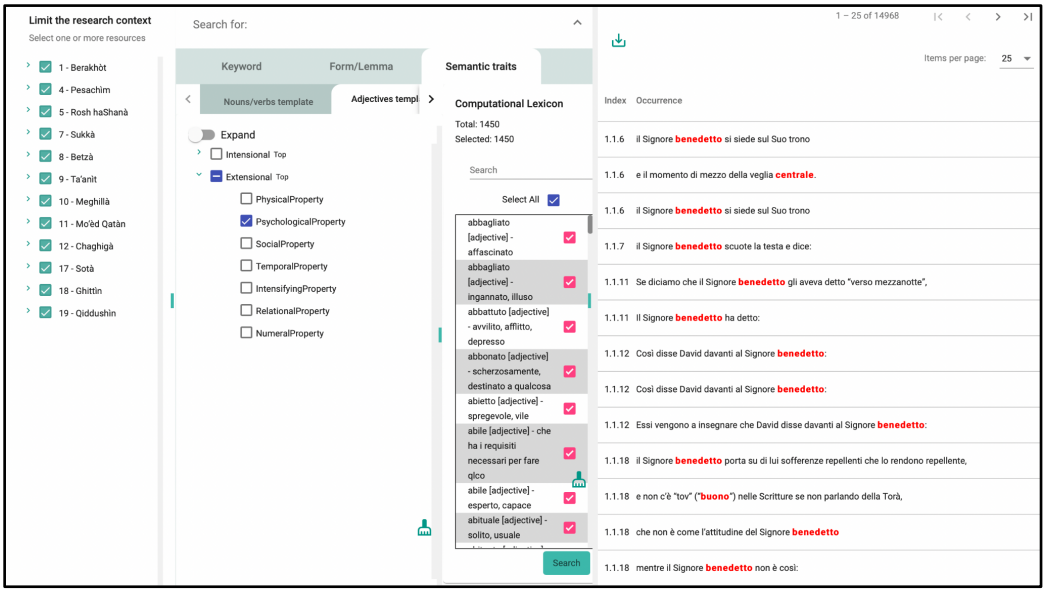

Figura 8. Esempio di ricerca per template.

L'esempio mostrato nella figura 8 riporta l'esito di una ricerca per *template* di aggettivi, eseguita dalla sezione "Adjectives template" dell'interfaccia. Operativamente, è stato innanzitutto selezionato il *template* "PsychologicalProperty" tra gli aggettivi estensionali disponibili. Una volta selezionato il *template*, l'applicazione ha restituito tutti i sensi collegati ad esso, per un totale di 1.450 unità semantiche presenti nel lessico; successivamente, al click del pulsante "Search", è stata eseguita la query sul testo (costituita da tutte le forme flesse degli aggettivi legati ai sensi ottenuti) e sono stati estratti 14.968 contesti dai trattati del Talmud babilonese selezionati. In questi contesti appaiono occorrenze di diversi aggettivi che richiamano qualità o aspetti psicologici o riferiti alla psiche, anche molto differenti tra loro (per esempio con accezione sia positiva che negativa).

Una ricerca che, invece, venga effettuata tenendo conto dei nuovi tratti semantici inseriti, permette di affinare ulteriormente questi risultati.

Come mostrato in figura 9, mediante la neo-aggiunta sezione "Traits", l'utente è in grado di selezionare uno o più tratti semantici tra quelli disponibili, con l'ausilio di due filtri: "Search Template" per mostrare tutti e soli i tratti afferenti a un determinato template aggettivale, e "Search Trait" per affinare la ricerca o filtrare sulla base del valore (e.g. "positive").

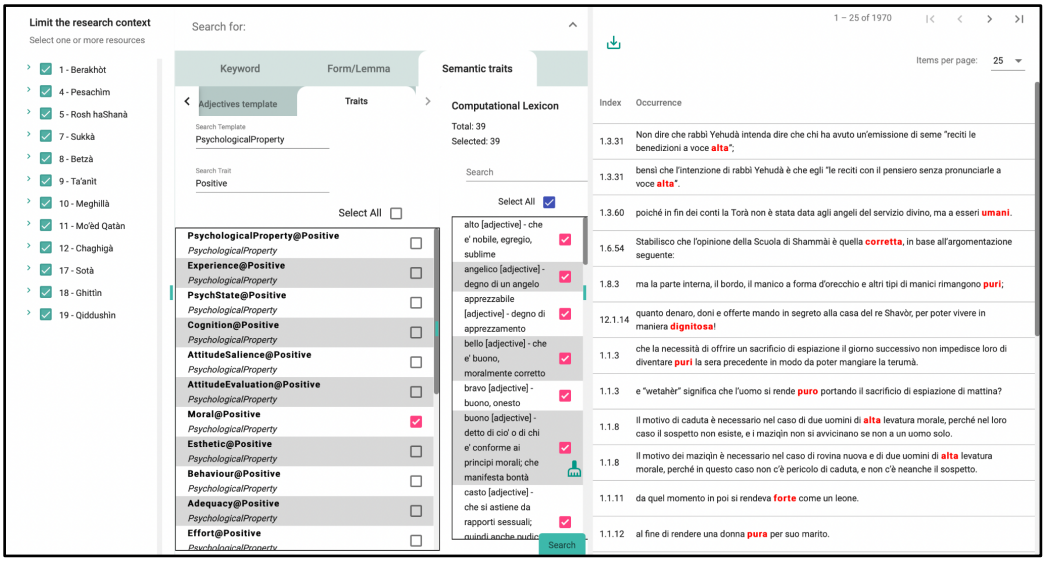

Figura 9. Esempio di ricerca su tratto.

Nell'esempio in figura è stata inserita la stringa "PsychologicalProperty" nel filtro per template ed è stata inserita la stringa "positive" nel campo di filtro sui tratti. Dai tratti risultanti è stato selezionato un tratto specifico: "Moral@Positive".

Il sistema, quindi, recupera dal lessico e mostra in una apposita finestra la lista dei sensi degli aggettivi che sono associati a tale tratto (tra cui "alto", "angelico", "apprezzabile", "bello", ecc.) che l'utente può selezionare. Una volta cliccato su "Search", viene avviata la ricerca all'interno del testo del Talmud di tutte le occorrenze degli aggettivi selezionati, nelle loro diverse forme.

La disponibilità dei tratti semantici all'interno della risorsa PSC (e il contestuale adattamento dell'applicazione alla loro gestione) consente perciò di potere effettuare ricerche semantiche ancora più sofisticate e precise.

8. Conclusioni

Partendo da uno studio teorico sulla risorsa PSC, ampliato attraverso il confronto con il database lessicale WordNet, si è arrivati alla conversione di un suo insieme di dati semantici e alla relativa integrazione in una applicazione per la ricerca semantica sul testo. Dopo un excursus sui fondamenti dei *Linked Data*, del Web Semantico e del modello *OntoLex-Lemon*, è stata descritta l'attività di tesi condotta, che ha previsto l'estrazione dei tratti semantici aggettivali dal database MySQL di PSC e la successiva conversione in RDF, ovvero in un formato allo stato dell'arte conforme al paradigma dei *Linked Open Data*.

L'aggiunta dei tratti semantici alla risorsa PSC, già in parte convertita presso l'Istituto di Linguistica Computazionale del CNR nel formato RDF secondo il modello *OntoLex-Lemon*, ne ha arricchito la componente semantica, costituendo un risultato di per sé.

Inoltre, come documentato nell'ultima parte della relazione di tesi, l'arricchimento di PSC attraverso i tratti ha permesso di incrementare le potenzialità offerte dall'applicazione Web per la ricerca *full-text* che utilizza la risorsa stessa come lessico computazionale di supporto.

9. Bibliografia

Albanesi D., Bellandi A., Giovannetti E., Marchi S., Papini M., Sciolette F. 2021, 2022, *The Role of a Computational Lexicon for Query Expansion in Full-Text Search*. *In Proceedings of CLiC-it Italian Conference on Computational Linguistics, Milan, Italy.. CEUR workshop proceedings, ISSN 1613-0073.* http://ceur-ws.org/Vol-3033/paper33.pdf

Bel N., Busa F., Calzolari N., Gola E., Lenci A., Monachini M., Ogonowski A., Peters I., Ruimy N., Villegas M., Zampolli A. (2000), *Simple: A General Framework for the Development of Multilingual Lexicons*. *International Journal of Lexicography 13(4), 249-263.*

Bianco I., Guazzini E., Molino S., Ulivieri M., (2006), *Some Aspects of the PAROLE-SIMPLE-CLIPS Semantic Layer: Uses and Advantages*.

Busa F., Calzolari N., Gaizauskas R., Gola E., Guimier E., Humphreys L., Lenci A., McCauley C., Monachini M., Ogonowski A., Peters I., Petes W., Rakovsky U.V.,

Calzolari N., Busa F., Gola E., Lenci A., Monachini M., Ruimy N., Zampolli A. (2020), *Simple Work Package 2 Linguistic Specifications Deliverable D2.1*

Calzolari N., Distante R., Guazzini E., Molino S., Monachini M., Ruimy N., Ulivieri M., and Zampolli A.,(2002) , *Clips, a multi-level italian computational lexicon: A glimpse to data. In Proceedings of the Third International Conference on Language Resources and Evaluation (LREC02)*

Fellbaum C., (1998), *Wordnet: An Electornic Lexical Database, Cambridge: MA:MIT Press*

Pustejovsky J., (1995), *The Generative Lexicon. MA: MIT Press*

Quillian R., (1963), *A notation for representing conceptual information: An application to semantics and mechanical English para- phrasing. SP-1395, System Development Corporation, Santa Monica.*

Recourcé G., Ruimy N., Villegas M., Zampolli A., (2000), *SIMPLE Work Package 2 Linguistics Specifications Deliverable D2.1*

Roventini A., Ruimy N., (2005), *Towards the Linking of two Electronic Lexical Databases of Italian*.

10. Sitografia

https://fontistoriche.org/linked-data/ (visitata il 26/07/2022)

https://it.wikipedia.org/wiki/RDF_Schema (visitata il 30/07/2022)

https://www.ontotext.com/products/graphdb/ (visitata il 3/07/2022)

https://en.wikipedia.org/wiki/OntoLex (visitata il 4/08/2022)