Trotter, David/Bozzi, Andrea/Fairon, Cédrick (éd.) (2016). *Actes du XXVII^e Congrès international de linguistique et de philologie romanes (Nancy, 15-20 juillet 2013). Section 16 : Projets en cours ; ressources et outils nouveaux*. Nancy, ATILF :
http://www.atilf.fr/cilpr2013/actes/section-16.html.

Un'ontologia per il DiTMAO (*Dictionnaire des Termes Médico-botaniques de l'Ancien Occitan*)

1. Premessa

Il Progetto DiTMAO pone, fra i propri obiettivi principali, quello di rendere accessibili sul Web, in formato digitale, la terminologia contenuta nei testi di argomento medico-farmaceutico in antico occitano, organizzata in forma di dizionario[1]. Come capita di frequente nella lessicografia storica, anche in questo progetto si è affrontato il problema della modalità con cui realizzare la corrispondenza fra i valori concettuali che le parole, appartenenti ai campi semantici considerati, esprimono nei testi del corpus e quelli equivalenti nella terminologia medico-farmaceutica attuale, ove, naturalmente, tale persistenza esista[2]. Il lavoro lessicografico che è stato avviato consente di indagare non solo fenomeni linguistici, ma anche, tramite essi, di descrivere un universo concettuale sincronico, ovvero un dominio di conoscenze relativo alla medicina e alla farmacopea occitana di epoca medievale. Esso, infatti, se analizzato solo sulla base di fenomeni storici e culturali, risulterebbe incompleto, o impreciso.

Se la redazione delle voci, inoltre, viene realizzata tenendo presente anche la prospettiva diacronica che ha inevitabilmente inciso sull'evoluzione dei valori concettuali di termini che sono sopravvissuti fino ai nostri giorni, si pongono tutte le condizioni favorevoli:

---

1 Si veda la comunicazione di Corradini («La realizzazione del Dictionnaire des Termes Médico-botaniques de l'Ancien Occitan (DiTMAO): problemi di organizzazione della conoscenza medico-farmaceutica attestata dai manoscritti in occitano antico»), presentata in questa stessa sessione, dove sono descritti in dettaglio gli scopi del progetto e i testi sui quali lo spoglio lessicale viene condotto. Il progetto è coordinato da Guido Mensching, Università di Gottinga, finanziato dalla DFG e sviluppato in collaborazione con l'Università di Pisa (Corradini) e l'Università di Colonia (Bos).

2 Tale fenomeno si verifica in particolare riguardo a testi antichi di argomento tecnico per i quali i dizionari generali di una lingua, che non abbiano utilizzato spogli specifici, forniscono informazioni lacunose o non puntuali. Osservazioni a questo proposito, concernenti termini tecnici del *Corpus Hippocraticum* e relative definizioni presenti negli strumenti lessicografici disponibili per il greco antico, si leggono in Bozzi 1982 (in particolare nelle note introduttive alle pp. 1-3). In questo genere di opere, inoltre, si può verificare il fenomeno della specializzazione semantica di parole di uso comune le quali, in contesti particolari, o sull'esempio di modelli greci o latini ove ciò accade molto di frequente, si caricano di accezioni tecniche specifiche.

- per avere un quadro dettagliato del settore specifico sul quale la ricerca viene condotta;
- per mettere a disposizione di molte comunità di filologi e linguisti una mole di dati organizzata in forma ottimale ai fini della consultazione sia nella tradizionale veste di un dizionario cartaceo, sia mediante sistemi informatici di ultima generazione semanticamente orientati.

Per raggiungere gli obiettivi indicati, le voci devono essere redatte seguendo un duplice sistema di classificazione logico-semantica: quello aderente al periodo medievale nel quale esse furono impiegate e quello coerente con l'uso che esse hanno assunto attualmente.

A tale scopo e considerando soprattutto che il risultato finale della ricerca dovrà essere consultabile sul Web, oltre che, come detto, su supporto cartaceo, si è ritenuto opportuno adottare appropriate tecnologie del cosiddetto Web semantico e una serie di applicazioni basate su strumenti software per il trattamento di testi, immagini e lessici, sviluppati all'ILC-CNR di Pisa[3].

In questo contributo si prende in esame l'approccio metodologico tipico dei modelli concettuali utilizzati per le ontologie di dominio grazie al quale è stato possibile dare una descrizione formale esplicita del particolare dominio di interesse, cioè la medicina e la farmacopea medievale in antico occitano[4].

La riflessione preliminare si è basata sull'analisi delle informazioni e dei materiali disponibili, alcuni dei quali in formato cartaceo, altri in formato elettronico.

Sono stati utilizzati:

- i manoscritti, ovvero le fonti primarie, editi o in corso di edizione[5];
- i termini medici in antico occitano scritti in alfabeto ebraico, con relative corrispondenze in ebraico e in arabo[6];
- indici lemmatizzati di testi editi disponibili in formato elettronico.

I materiali, dunque, sono eterogeni, redatti in lingue e alfabeti diversi ed è indispensabile che il sistema da adottare consenta, come già detto, di stabilire un rete di relazioni semantiche diacroniche fra il significato posseduto dai termini in epoca medievale e quello assunto nell'età contemporanea.

---

[3] L'Istituto di Linguistica Computazionale "A. Zampolli" del CNR di Pisa (‹www.ilc.cnr.it›) ha una pluridecennale esperienza nel trattamento dei testi mediante sistemi informatici; fra questi, è in fase molto avanzata di sviluppo un'applicazione Web di filologia computazionale per la produzione di edizioni elettroniche di testi e per la comparazione fra testi antichi e loro antiche traduzioni. Si veda, per esempio, il sito del progetto ERC Ideas Advanced Grant 249431, "Greek into Arabic. Philosophical Concepts and Linguistic Bridges" (‹www.greekintoarabic.eu›), che ha lo scopo di contribuire, mediante confronto semiautomatico fra alcuni capitoli delle *Enneadi* di Plotino e la traduzione di essi in lingua araba, nota come *Teologia di Aristotele*, all'incremento del *Glossarium Graeco-Arabicum* redatto presso la Ruhr-Universität Bochum.

[4] In questo ambiente il termine "ontologia" significa "specificazione esplicita di una concetto". Si veda Gruber (1993, 199-220). Cfr. anche Staab / Studer (ed.) (2009).

[5] Si veda Corradini (1997).

[6] Si veda Bos / Mensching / Hussein / Savelsberg (2011).

2. Le motivazioni alla base della scelta ontologica per l'annotazione semantica strutturata

Nella prassi redazionale di tipo tradizionale è sempre esistita la possibilità di utilizzare parole chiave in grado di classificare tematicamente le definizioni, come, per esempio, i tecnicismi, gli arcaismi, gli usi metaforici, i sinonimi. Tuttavia, nell'ambito di una più precisa organizzazione semantica dei dati contenuti in un vocabolario riferito ad uno specifico dominio di conoscenza e, per di più, relativo ad un'epoca remota, il fatto di inserire ogni entrata di una eventuale lista di parole-chiave in uno schema precostituito, ove compaiano voci di livello superiore, voci ad esse soggiacenti e relazioni fra voci e sotto-voci, assume senza dubbio notevoli valori aggiunti[7].

Ciò, infatti:

- consente ai redattori di dare una struttura semantica omogenea ed esplicita alle definizioni della terminologia per la composizione del dizionario cartaceo;
- rende più produttive le fasi di consultazione, studio e ricerca da parte, soprattutto, degli specialisti che accederanno alla versione Web del dizionario e all'archivio dei testi;
- rende lo schema concettuale indipendente dalla lingua nel quale vengono descritte classi, sottoclassi e relazioni[8];
- soprattutto, rende lo schema concettuale indipendente dalla lingua del testo studiato. Ciò vale in primo luogo per i testi multilingui (per esempio, le opere che contengono parti in latino, lingue romanze, greco, ebraico, arabo), come nel caso del progetto DiTMAO, sul quale tale ipotesi di organizzazione ontologica dei dati del dominio si sta sperimentando con successo.

---

[7] Lo schema logico e concettuale potrebbe essere definito 'tassonomia', adottando un termine più comunemente usato nel settore degli studi umanistici rispetto a quello di 'ontologia' che, nato nell'ambito degli studi logici e filosofici, è stato oggi riqualificato dagli ingegneri dell'informazione che ad esso hanno attribuito una connotazione tecnologica e strumentale. La differenza fra le due espressioni consiste soprattutto nel fatto che l'ontologia consente di dare una rappresentazione esplicita di uno specifico dominio di conoscenza, mettendo in relazione fra loro le componenti concettuali le quali non sono quasi mai isolate le une rispetto alle altre, e alle quali è possibile applicare una indefinibile quantità di attributi. Il potenziale descrittivo delle ontologie, pertanto, è molto potente soprattutto nel caso in cui lo si voglia applicare a un dominio specifico e delimitato come, appunto, la medicina e la farmacopea di epoca medievale. L'utilizzo delle ontologie, come elemento descrittivo del lessico, si è molto diffuso anche in conseguenza dell'affermazione della teoria del Lessico Generativo (Pustejovsky, 1995). Un'interessante applicazione di questo metodo al lessico tecnico presente nei lavori di Ferdinand de Saussure si legge in Ruimy / Piccini / Giovannetti (2012, 1043-1056).

[8] Resta inteso, comunque, che le etichette concettuali attribuite a ciascun elemento dello schema devono essere espresse in forma linguistica e, nella fattispecie, l'idioma scelto dal gruppo di ricerca è il francese. Ciò significa che per interrogare il lessico e ritrovare tutti i termini che condividono valori concettuali o partecipano delle medesime relazioni fra concetti diversi, è necessario operare la selezione su liste predisposte in francese. Il vantaggio dell'approccio ontologico consiste anche nel fatto che tali etichette, in numero non rilevante, possono essere facilmente tradotte per un accesso multilingue ai dati del dizionario, sempre che il termine usato nella traduzione (per es., l'inglese) ricopra il concetto di quello francese originario.

La conoscenza dello specifico dominio che i testi, sulla base dei quali viene redatto il dizionario, veicolano, viene pertanto rappresentata in forma ontologicamente strutturata, esplicita. La valutazione puntuale della terminologia medico-farmaceutica medievale in antico occitano è un caso-studio particolarmente interessante poiché necessita di strumenti di analisi più fini di quelli offerti da semplici indici di parole-forma o di lemmi presenti nelle fonti, corredati di eventuali concordanze.

Oltre ad essi, appare oggi sempre più funzionale interrogare la base dei dati terminologica o gli stessi testi utilizzando, come chiave di accesso, un concetto o un tema generico (es. "unguento"), oppure una relazione fra due concetti (es. "unguento" per "ferita"), oppure ancora una relazione fra più elementi (es. "unguento" per "ferita" in una determinata parte del corpo, come la "testa")[9].

I risultati che si ottengono con programmi di tipo tradizionale al fine di produrre vocabolari terminologici di specifici settori, inoltre, potrebbero non essere esaustivi perché è accertato, per esempio:

- che una delle fonti (o chi effettua una ricerca) denoti uno stesso tema con parole diverse da quelle utilizzate da un'altra fonte;
- oppure che in fase di recupero delle informazioni ('information retrieval') venga usata una chiave di accesso (es. "unguento", "ferita", "testa") diversa da quella, semanticamente identica, che è attestata. Ciò provoca una evidente impossibilità di recuperare le informazioni che invece sono presenti, sia pure in una veste diversa.

Il superamento di questi limiti e la certezza di ottenere risultati esaustivi si ottiene mediante un sistema di 'data modeling' grazie al quale l'utente viene assistito nell'organizzazione e nella scrittura dello schema concettuale tipico del dominio di sua particolare competenza ed interesse[10]. La progettazione di questo innovativo sistema è stata effettuata anche per il progetto di edizione elettronica di un corpus di manoscritti di F. de Saussure[11].

Prima di entrare nel dettaglio dello schema ontologico con cui è stata organizzato il dominio della terminologia medico-farmaceutica dell'antico occitano, desideriamo fornire un esempio concreto che contribuisca subito a chiarire almeno la più evidente delle ragioni che hanno spinto ad adottare questo impianto metodologico e a realizzare le conseguenti componenti del sistema informatico.

Si tratta del caso in cui termini differenti sussumono il medesimo valore semantico inserito nello schema: per es., il concetto di "unguento", in antico occitano *oignement* (e varianti *oinhement*, *oinnement*, *onhement*, *hongemen*) e *onguent*, ma anche

[9] Le etichette o denominazioni dei concetti, per comodità espositiva, vengono qui espresse in italiano, mentre, come detto, nel sistema esse compaiono in francese.

[10] In rete si trovano una grande quantità di ottimi siti accademici ove ricavare informazioni, anche di tipo divulgativo, su questi aspetti. Fra i sistemi di scrittura ('editor') più noti e gratuiti per la creazione di ontologie si indica qui *Protégé* (‹http://protege.stanford.edu/overview›).

[11] Si veda il contributo a questo convegno di Pesini / Del Grosso / Bozzi: «Ferdinand de Saussure e la linguistica romanza. Un'applicazione web per l'edizione elettronica dei manoscritti».

dura confeccio e *apostolico*, che sono tipi particolari di unguento. Il problema sarebbe parzialmente risolvibile anche con altri metodi, come, per esempio, la preparazione di tabelle di corrispondenza con cui la macchina sarebbe in grado di equiparare almeno le varianti (grafiche, fonetiche, morfologiche) di una medesima parola. Le caratteristiche linguistiche del corpus, tuttavia, sono tali da indurre la procedura a generare errori: l'utilizzo meccanico delle corrispondenze da parte di un programma di calcolatore potrebbe, infatti, far considerare varianti di uno stesso termine forme che in realtà appartengono a parole diverse.

Con l'attribuzione del valore concettuale "unguento" si superano tutte queste difficoltà e i due lemmi *oignement* e *onguent*, con tutti gli altri eventuali sinonimi oltre che naturalmente tutte le varianti, sono unificati su base semantica e concettuale. L'insieme dei contesti nei quali i lemmi ricorrono si possono ottenere selezionando la voce "unguento" nello schema ontologico predisposto.

Nelle fasi iniziali del progetto si era valutata la possibilità di risolvere il problema della equivalenza semantica fra termini differenti in una stessa lingua, e/o appartenenti a lingue diverse, e/o appartenenti a tranche diacroniche molto distanti, mediante l'utilizzo di un modello, noto nel settore del trattamento automatico del linguaggio col nome di WordNet (*EuroWordNet*, per le lingue europee)[12]. Esso si basa sulla descrizione del valore semantico grazie alla concatenazione di liste di sinonimi ('synset'): in pratica, una parola viene definita, come capita spesso anche nei vocabolari cartacei redatti con metodi descrittivi tradizionali, mediante altre parole che denotano il medesimo concetto o si riferiscono al medesimo oggetto della realtà. Abbiamo, tuttavia, constatato che le versioni del sistema sono state principalmente adottate per lingue parlate contemporanee come, per esempio, l'italiano (*ItalWordNet*)[13], anche allo scopo di avere a disposizione uno strumento di controllo per i programmi di traduzione automatica. L'impianto metodologico centrato sulla sinonimia, che è alla base di *WordNet*, di fronte alla necessità dei redattori di DiTMAO di tenere sotto controllo tecnicismi del lessico medico-farmaceutico e botanico, correlati anche ad aspetti di semantica diacronica (tutti elementi che è impensabile trattare con l'utilizzo di 'synset'), si è dimostrato non pertinente alle finalità.

---

[12] Si vedano il sito ufficiale di *WordNet* (‹www.illc.uva.nl/EuroWordNet›) e Rodrìguez / Climent / Vossen et alii (1998, 117-152).

[13] A questo proposito si veda Marinelli / Roventini / Enea (2004, 465-468); il manuale d'uso di *ItalWordNet* è accessibile sul sito dell'ILC-CNR (‹www.ilc.cnr.it›). Solo di recente, nell'ambito di un progetto che l'ILC sta realizzando in collaborazione con il *Perseus Project* (Tufts University, Boston), si sta effettuando un utilizzo di *WordNet* in lessicografia greca classica (*AncientGreekWordNet*) con lo scopo, però, di favorire lo sviluppo di nuove modalità di interrogazione, semanticamente orientate, di archivi testuali, a fini principalmente didattici. Al momento attuale non esiste ancora una pubblicazione scientifica dedicata all'argomento.

3. Lo schema ontologico

Il primo passo per la definizione di classi, proprietà e regole dello schema ontologico è consistito nell'analisi dei materiali e nella successiva verifica dell'esistenza di standard, riconosciuti a livello internazionale, idonei allo scopo. Alcune prove preliminari di popolazione[14] dello schema ontologico, eseguite al fine di verificare la validità e l'applicabilità di tali standard a questo specifico progetto, hanno evidenziato come esse non fossero sufficienti a rappresentare il dominio di conoscenza relativo a tale ambito nei materiali disponibili[15]. È stato necessario, quindi, estendere lo schema ontologico, pur mantenendo la compatibilità con gli standard di partenza. Si è trattato di un processo iterativo che, grazie a successive fasi di popolamento dello schema di volta in volta affinato, ha consentito di progettarne uno pienamente rispondente ai requisiti. Senza scendere in eccessivi dettagli non necessari in questa occasione, descriviamo qui i punti principali dell'attività che si è svolta su due fronti diversi, ma correlabili fra loro.

Sono stati predisposti due schemi.

Il primo, descritto con linguaggio UML[16], è in grado di rappresentare le classi principali e le relazioni per quanto riguarda gli aspetti grammaticali (grafico-fonetici e morfologici), che sono spiegate in dettaglio nel contributo di Maria Sofia Corradini citato sopra alla nota n. 1. Grazie a tale linguaggio, è possibile catalogare lemmi, sottolemmi, sinonimi e varianti (morfologiche, grafico fonetiche, ecc.), indicando l'alfabeto, la categoria grammaticale, il numero, il significato, la lingua, il nome scientifico, l'eventuale corrispettivo in un'altra lingua antica, il periodo nel quale ogni singola voce (lemma, sottolemma e/o variante) era in uso e in quali documenti è attestata.

Lo schema è stato popolato con i reali lemmi estratti dalle fonti che, entrando nella classificazione ontologica, ne hanno assunto le relative proprietà. Questo processo è avvenuto grazie ad una specifica interfaccia grafica, semplice ed intuitiva.

---

[14] 'Popolare/popolazione' sono termini tecnici adoperati dagli informatici che si occupano di sistemi ontologici e si riferiscono alla "attività di attribuzione di istanze ai singoli elementi dello schema".

[15] Fra gli standard esistenti, sono stati considerati quelli più adatti al tema di questo progetto: si tratta di FRBR-oo (*Functional Requirements for Bibliographic Records-object oriented*) adottato dalla 'International Federation of Library Associations' (IFLA), per il quale si veda ‹http://archive.ifla.org/VII/s13/wgfrbr/FRBR-CRMdialogue_wg.htm›. Esso incorpora elementi dello schema concettuale e terminologico presente nel CIDOC-CRM (*Conceptual Reference Model*), validato come standard ISO-21127 nel 2006, per il quale si veda ‹www.cidoc-crm.org›. Abbiamo, inoltre, preso in considerazione il *Lexical Markup Framework* (LMF), ISO-24613 validato come standard nel 2008, per il quale si veda ‹www.lexical-markupframework.org›.

[16] UML ('Unified Modeling Language'). Nello schema i nomi delle classi e delle relazioni, per convenzione, sono scritti in inglese. Viene utilizzato il francese come lingua del progetto per quanto riguarda la denominazione dei campi presenti nella scheda di immissione dei dati nell'ontologia e per le fasi di consultazione. Si veda anche la nota n. 8.

Essa ha l'aspetto di una scheda per l'immissione di metadati bibliografici. Con questo sistema sono stati inseriti fino ad oggi i dati relativi a più di 1500 lemmi.

Il secondo schema è invece in grado di rappresentare le classi, sottoclassi e relazioni per quanto riguarda gli aspetti semantico-concettuali. Esso è correlato a quello precedente mediante la superclasse 'OldOccitanicMedicalTerminology' e consente di attribuire ai lemmi un valore concettuale (fig. 1).

Anche 'OldOccitanicMedicalTerminology' racchiude classi e relazioni: esse descrivono la classificazione medievale dei termini medici e botanici a partire da tre sottoclassi principali: 'AnimalWorld' "mondo animale", 'VegetalWorld' "mondo vegetale", 'MineralWorld' "mondo minerale".

Il legame tra la 'OldOccitanicMedicalTerminology' e la 'ModernMedicalTerminology' rende esplicita la capacità del sistema di esprimere la relazione diacronica fra i termini.

Si riportano di seguito alcuni esempi di termini relativi al globo oculare attestati nella *Notomia* di Anric de Mondavilla, testo in lingua d'oc che accoglie numerosi latinismi di ambito anatomico. Tali voci mostrano corrispondenze o discrepanze di differente tipologia rispetto alle voci equivalenti nella lingua moderna, secondo un processo parallelo a ciò che avviene nella gran parte dell'area romanza come, per esempio:

- *cornea* "cornea" che, pur nel mutamento di forma, mantiene nella lingua moderna (*cournèio*) il medesimo significato posseduto nel medioevo;
- *uvea* "iride", che è stato sostituito, nella denominazione della medesima entità, da *iris*;
- *aranea*, termine oggi scomparso assieme all'entità a cui si riferiva, che è stata concettualmente assimilata alla retina.

4. Dalla classificazione della terminologia ai testi digitalizzati

Il processo di classificazione della terminologia medico-farmaceutica, secondo i due schemi predisposti dai redattori con l'ausilio del personale informatico che collabora al progetto, non è isolato dai testi nei quali la terminologia è documentata. Per questa ragione, le informazioni inserite nello schema ontologico saranno utilizzate per effettuare una lemmatizzazione semi-automatica su ulteriori testi resi disponibili in formato elettronico. Inoltre, il componente di filologia computazionale che gestisce tutto l'archivio, sarà in grado di associare alle trascrizioni delle opere del corpus anche le corrispondenti immagini digitalizzate delle fonti originali. Un componente aggiuntivo consentirà, infine, di utilizzare la classificazione dello schema ontologico anche per consentire ai redattori di introdurre annotazioni su testi e immagini (o porzioni di immagini).

Pisa, ILC-CNR — Andrea BOZZI
Università di Bologna — Damiana Luzzi

Bibliografia

Bos, Gerrit / Mensching, Guido / Hussein, Martina / Savelsberg, Frank, 2011. *Medical Synonym Lists from Medieval Provence: Shem Tov Ben Isaac of Tortosa, Sefer ha-Shimmush*, Book 29. Part 1: *Edition and Commentary of List 1* (Hebrew-Arabic-Romance/Latin), Leiden, Brill.

Bozzi, Andrea, 1982. *Note di lessicografia ippocratica. Il trattato sulle arie, le acque, i luoghi*, Roma, Edizioni dell'Ateneo.

Corradini, Maria Sofia, 1997. *Ricettari medico-farmaceutici medievali nella Francia meridionale*, Firenze, Olschki.

Gruber, Thomas Robert, 1993. «Translation Approach to Portable Ontology Specifications», *Knowledge Acquisition*, 5 (2), 199-220.

Marinelli, Rita / Roventini, Adriana / Enea, Alessandro, 2004. «Building a Maritime Domain Lexicon: Few Considerations on the Database Structure and the Semantic Coding», *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)*, Paris, The European Language Resources Association, II, 465-468.

Pustejovsky, James (1995). *The Generativ Lexicon*, MIT Press, Cambridge, MA.

Rodrìguez, Horacio / Climent, Salvador / Vossen, Piek *et al*., 1998. «The Top-Down Strategy for Building EuroWordNet: Vocabulary Coverage, Base Concepts and Top Ontology», *Computers and the Humanities* 32 (2-3), 117-152.

Ruimy, Nilda / Piccini, Silvia / Giovannetti, Emiliano, 2012. «Les Outils Informatiques au Service de la Terminologie Saussurienne», pubblicato in linea negli atti del Congrès Mondial de Linguistique Française (Lyon, 2012), consultabile liberamente all'indirizzo www.ilc.cnr.it/viewpage.php/sez=ricerca/id=917/vers=ita, 1043-1056.

Staab, Steffen / Studer, Rudi (ed.), 2009. *Handbook on Ontologies*, Berlin – Heidelberg, Springer.

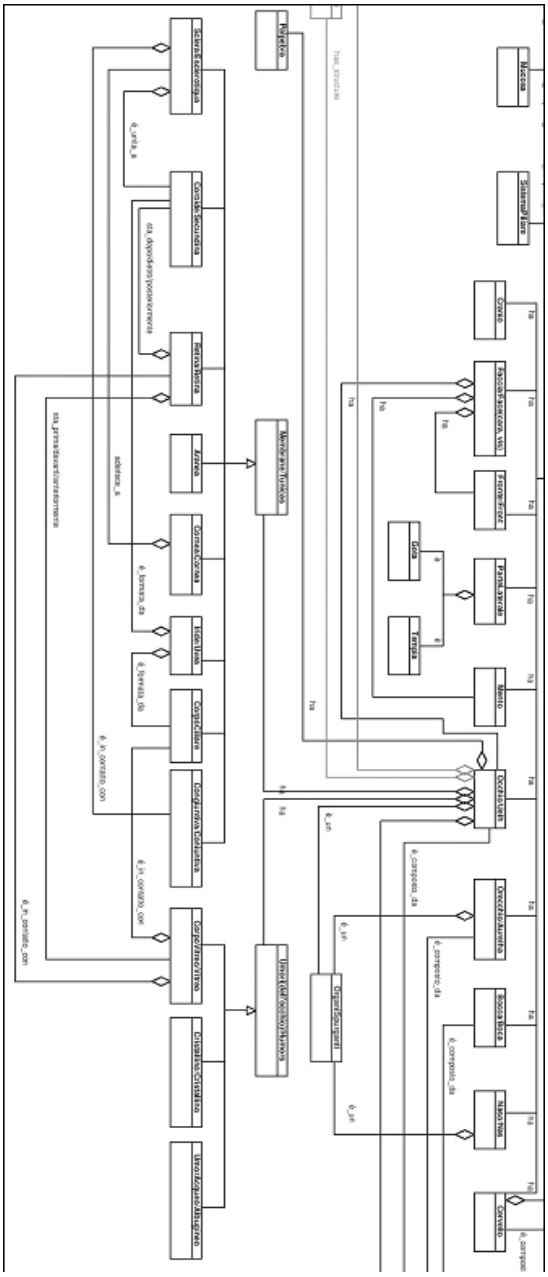

Fig. 1. Schema, descritto in UML, degli aspetti semantico-concettuali relativi al globo oculare