Università degli Studi di Modena e Reggio Emilia
Facoltà di Ingegneria – Sede di Modena
Corso di Laurea in Ingegneria Informatica

ONTOLOGIE LESSICALI MULTILINGUA: MULTIWORDNET ED EUROWORDNET

Relatore:
Chiar.mo Prof. Sonia Bergamaschi

Correlatore:
Ing. Daniele Miselli

Elaborato di Laurea di:
Roberto Rasi

Anno Accademico 2002-2003

Parole chiave:

Ontologie lessicali
WordNet
Multilinguismo
EuroWordNet
MultiWordNet

RINGRAZIAMENTI

Desidero ringraziare la prof. Sonia Bergamaschi, l'ing. Domenico Beneventano e l'ing. Daniele Miselli per l'aiuto fornito durante la realizzazione del presente elaborato e per la costante disponibilità dimostrata.

Un sentito ringraziamento va ai miei genitori, che, con il loro continuo sostegno morale ed economico, mi hanno permesso di raggiungere questo traguardo.

Un ringraziamento particolare, infine, alla mia fidanzata, Alisia, che mi è stata molto vicino durante i miei studi universitari.

Indice

Indice delle figure

Introduzione

Spinta dalla necessità di condividere conoscenza e informazioni con applicazioni già costruite, l'attività di ricerca legata alle ontologie è, al momento, molto forte in diversi campi dell'intelligenza artificiale e, più in generale, delle scienze dell'informazione [10]. Le ontologie giocano un ruolo fondamentale anche in uno dei più recenti ambiti di interesse, quello del Semantic Web.

Nel contesto della ricerca sul web semantico si colloca il progetto **SEWASIE** [5] (**SE**mantic **W**ebs and **A**gent**S** in **I**ntegrated **E**conomies), finanziato dall'Unione Europea e guidato dall'Università degli studi di Modena e Reggio Emilia. Tale progetto, della durata di 36 mesi, è partito nel maggio del 2002 con l'obiettivo di sviluppare un motore di ricerca intelligente basato sulla semantica. Il sistema che si intende sviluppare in SEWASIE basa il suo funzionamento su un processo di integrazione e arricchimento semantico di sorgenti di dati eterogenee; questo processo viene realizzato dal componente denominato Ontology Builder, che è preposto alla costruzione e al mantenimento di un'ontologia delle suddette sorgenti. L'integrazione delle sorgenti viene fatta sfruttando il sistema sviluppato nell'ambito del progetto **MOMIS** (**M**ediator Envir**O**nment for **M**ultiple **I**nformation **S**ources), al quale hanno partecipato il Dipartimento di Scienze dell'Ingegneria di Modena e il Dipartimento di Scienze dell'Informazione di Milano.

Il progetto MOMIS si poneva l'obiettivo di integrare sorgenti di dati eterogenee, distribuite ed eventualmente semistrutturate, e di proporne all'utilizzatore una vista globale, aggregata e virtuale, che permesse all'utente di formulare interrogazioni in modo trasparente. Il sistema, sviluppato nell'ambito del progetto, costruisce la vista globale partendo da un insieme di viste parziali, ciascuna delle quali è uno schema rappresentante, nel linguaggio ODL_I^3, una delle sorgenti da integrare, e da un common thesaurus di relazioni, sia intra-schema che inter-schema, tra gli elementi che compongono tali viste. La costruzione del common thesaurus implica l'annotazione delle sorgenti rispetto ad un'ontologia lessicale comune: questa operazione consiste nell'associare ad ogni attributo del suddetto schema ODL_I^3 un termine ed uno o più significati; in questo modo si può sfruttare l'ontologia lessicale data per estrarre nuove relazioni, in base alle le relazioni semantiche che sono in essa definite.

Da quanto detto finora emerge la centralità delle ontologie lessicali nella fase di annotazione, e quindi la loro importanza nel processo di integrazione di sorgenti di dati eterogenee. Per questa operazione MOMIS utilizza il database lessicale WordNet, implementando anche un metodo per estenderlo con l'inserimento di nuovi termini, significati e relazioni. Per quanto riguarda SEWASIE, la prima release del prototipo di Ontology Builder realizza l'annotazione allo stesso modo di MOMIS; dato però il suo contesto europeo, l'obiettivo è quello di realizzare l'annotazione rispetto ad un'ontologia lessicale multilingua, in particolare si desidera utilizzare EuroWordNet.

Argomento del presente elaborato saranno proprio le ontologie lessicali multilingua; in particolare questo lavoro consiste in un'analisi dei due modelli fondamentali (*merge-model*, *expand-model*) e delle rispettive implementazioni (*EuroWordNet*, *MultiWordNet*).

Lo scopo del nostro lavoro è, dunque, di confrontare le architetture dei suddetti database lessicali, al fine di evidenziarne affinità e differenze, e di valutare quindi la compatibilità, a livello di schema concettuale, delle due ontologie. Volendo giudicare le possibilità espressive delle due architetture, e non le loro prestazioni, non saranno considerati nel confronto gli aspetti riguardanti l'implementazione fisica dei due database.

Non è nell'ambito di questo lavoro il progetto completo per l'integrazione dei due database; la nostra intenzione è quella di studiare la fattibilità di tale progetto. Perciò, l'obiettivo che si intende raggiungere con il confronto è di delineare una struttura comune in grado di ospitare e integrare le suddette ontologie: nel fare ciò ci si concentrerà sugli aspetti fondamentali delle due architetture, tralasciando alcuni particolari di secondaria importanza. Saranno espressamente indicati i dettagli che non vengono considerati.

Il presente elaborato è organizzato nei seguenti cinque capitoli:

1. **Ontologie lessicali:** una breve introduzione all'argomento della tesi; sono descritte a livello generale le ontologie lessicali; in seguito viene presentato WordNet, il database lessicale della lingua inglese; infine vengono introdotte le ontologie multilingua e descritti i due modelli fondamentali.

2. **Il database lessicale MultiWordNet:** si descriverà il modello adottato e l'implementazione; verrà descritto poi il processo di reverse engineering ed infine presentato, come risultato dell'analisi, lo schema E/R del database.

3. **Il database lessicale EuroWordNet:** procedendo parallelamente a prima, saranno presentati dapprima il modello, poi l'architettura generale e le strutture dati utilizzate; infine si presenterà lo schema E/R del database.

4. **Confronto tra gli schemi E/R di MultiWordNet ed EuroWordNet:** nel quarto capitolo saranno messi a confronto i due schemi E/R rappresentanti le architetture dei due database e si cercherà di evidenziarne affinità e differenze.

5. **Conclusioni e lavoro futuro:** alcune riflessioni sul lavoro svolto e indicazioni per un suo possibile proseguimento.

Capitolo 1

Le Ontologie Lessicali

Il termine *ontologia* è usato, generalmente, per indicare quella branca della filosofia che studia le modalità fondamentali dell'essere. In altre parole l'ontologia è una sistematica rappresentazione dell'essere e dell'esistenza.

Una delle definizioni di ontologia più largamente accettate è quella proposta da Tom Gruber in [12, pag. 199], che qui riportiamo:

"an ontology is an explicit specification of a conceptualisation"

Un'altra definizione molto quotata è quella data da Nicola Guarino [10]:

"an ontology is a set of logical axioms designed to account for the intended meaning of a vocabulary"

Questo termine oggi è stato preso in prestito in molti ambiti di ricerca delle scienze dell'informazione e dell'intelligenza artificiale; a seconda dei contesti e delle applicazioni essa assume significati diversi.

In seguito al loro crescente utilizzo, alcuni ricercatori, oltre a darne una loro definizione, hanno proposto una classificazione delle ontologie in base a determinate caratteristiche. Ad esempio, lo stesso Guarino [10] propone una classificazione, basata sul grado di generalità dei concetti rappresentati, delle ontologie in quattro categorie:

- **Top level ontology:** descrive concetti molto generici, indipendenti da un particolare problema o dominio di conoscenza (come i concetti di tempo, spazio, materia, oggetto, evento, azione…);
- **Domain ontology:** descrive il vocabolario relativo ad un dominio generico (come la medicina, la fisica…);
- **Task ontology:** descrive il vocabolario relativo ad una generica attività o processo (ad esempio la diagnosi, la vendita);
- **Application ontology:** descrive concetti che dipendono sia da un particolare dominio che da un particolare processo.

Per una panoramica più completa sulle definizioni e sulle classificazioni delle ontologie si rimanda a [11].

Nell'ambito di questo elaborato si farà riferimento alle cosiddette *ontologie lessicali*, in altre parole a quelle ontologie che rappresentano un linguaggio, o una sua parte (o più linguaggi, come si vedrà in seguito). In generale la conoscenza che si vuole esprimere con le ontologie lessicali è composta di due parti:

- una lessicale, formata da una collezione di parole (intese separatamente dal loro significato, o più semplicemente, come stringhe di caratteri);

- una semantica, che raccoglie tutti i significati associati alle parole e le relazioni che intercorrono tra di loro.

Questo tipo di rappresentazione permette non solo di recuperare il significato associato ad una parola, come in un dizionario, ma anche e soprattutto di estrarre un insieme di relazioni semantiche (cioè basate sul significato) a partire da un insieme di parole, e quindi trova applicazione nell'annotazione ed integrazione di sorgenti di dati eterogenee.

Per comprendere meglio questa rappresentazione, nel seguito di questo capitolo si descriverà il database lessicale WordNet, che rappresenta il punto di riferimento per i database lessicali che saranno oggetto della trattazione seguente.

1.1 L'ontologia lessicale WordNet

WordNet è sviluppato dal Cognitive science Laboratory sotto la direzione del professor Gorge A. Miller, [3] presso l'università di Princeton. E' disponibile on-line al sito internet http://www.cogsci.princetn.edu/~wn/, assieme al materiale e alla documentazione relativa.

WordNet è un sistema di gestione di un dizionario lessicale della lingua inglese basato sulle attuali teorie psicolinguistiche della memoria lessicale umana. Le categorie sintattiche (nomi, verbi, aggettivi e avverbi) sono organizzate in insiemi di sinonimi che rappresentano un inerente concetto lessicale. Gli insiemi di sinonimi sono poi collegati tra loro da diversi tipi di relazioni.

A ben vedere, WordNet non è semplicemente un dizionario on-line, ma non può neanche essere considerato un'ontologia, come sottolineato in [2]. Ciononostante, il numero di applicazioni che usano WordNet più come un'ontologia che come una risorsa lessicale è in crescita. Nel presente elaborato ci si riferirà ad esso come ad un'ontologia.

L'organizzazione interna di WordNet è diversa da quella dei dizionari cartacei. Questi ultimi, infatti, adottano un ordinamento alfabetico, cosa che permette ad un lettore di trovare le parole cercate senza dover sfogliare tutto il dizionario; d'altra parte però tutti i significati associati ad una parola sono presentati insieme, anche se spesso non hanno nessun legame tra loro. In WordNet, invece, i termini sono organizzati in concetti e non in ordine alfabetico, in base ai risultati delle ricerche psicolinguistiche. Inoltre le parole sono divise in categorie sintattiche: questo introduce una certa ridondanza perché una parola può avere significati in più categorie sintattiche; d'altra parte, permette di rappresentare e sfruttare le differenze strutturali che esistono nell'organizzazione semantica di tali categorie.

WordNet non può essere considerato quindi come un dizionario tradizionale, e nemmeno ha la pretesa di esserlo, poiché non contiene informazioni quali quelle etimologiche, fonetiche, gli esempi di utilizzo, ecc, ed il numero di termini presenti non è paragonabile. Tuttavia, il modello adottato nella sua costruzione è innovativo perché focalizza sulla differenza tra lemmi e significati, e rappresenta le relazioni tra i significati, le quali permettono navigabilità alle informazioni.

Il punto focale della struttura di WordNet è l'associazione tra *lemma* (la forma scritta o il suono) e *significato* (il concetto associato), rappresentata in figura con una notazione E/R. Esiste infatti un'associazione convenzionale tra la forma delle parole e il concetto che esse esprimono; tale associazione risulta di tipo molti-a-molti, e quindi dà luogo alle proprietà di:

- *Sinonimia*: proprietà di un significato di avere due o più lemmi in grado di esprimerlo. Un gruppo di lemmi tra loro sinonimi è chiamato *synset*.

- *Polisemia*: proprietà di un lemma di poter esprimere uno o più significati.

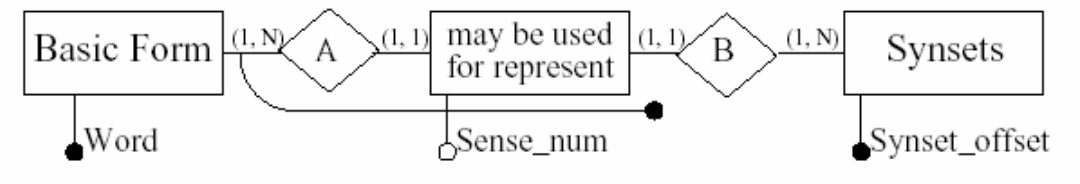


Figura 1 : Relazione tra lemmi e significati in WordNet

La corrispondenza tra la forma delle parole e il significato che esprimono viene sintetizzata nella cosiddetta *Matrice Lessicale*, nella quale le righe riportano il significato delle parole (ogni riga rappresenta un synset) e le colonne rappresentano la forma delle parole (ogni colonna corrisponde ad un lemma). Una $entry_{i,j}$ non nulla nella matrice lessicale significa che la word form *j* può essere usata per esprimere il concetto *i*: in questo caso il numero rappresenta il contatore del significato. I significati associati ad un lemma sono infatti numerati con un *sense number*; in questo modo un synset è identificato univocamente all'interno del database non solo dal suo *synset identifier* (un identificatore numerico univoco) ma anche dalla coppia (lemma, sense number), come è rappresentato nello schema E/R precedente. Inoltre un synset è associato ad una *glossa*, una spiegazione del significato che esso rappresenta.

Word Meanings	Word Forms				
	F_1	F_2	F_3	. . . . .	F_n
M_1	$E_{1,1}$	$E_{1,2}$			
M_2		$E_{2,2}$			
M_3			$E_{3,3}$		
.				. . . .	
.					
.					
M_m					$E_{m,n}$

Figura 2 : La Matrice Lessicale di WordNet

L'altro aspetto fondamentale di WordNet, come già annunciato, è costituito dalle relazioni che sono rappresentate nel database. Esse si dividono in due categorie:

- *Relazioni semantiche*: coinvolgono due synset, e sono valide per tutti i lemmi ad essi collegati (ad esempio specializzazione/generalizzazione)

- *Relazioni lessicali*: stabiliscono un nesso tra due singoli lemmi (ad esempio un contrario non è detto che sia valido per tutti i termini di un synset, ma solo per uno in particolare)

Non volendo appesantire troppo la trattazione, non entriamo nel merito delle specifiche relazioni. Per una trattazione più ampia dell'argomento si rimanda ai testi [1], [2] e [3] della bibliografia. Questa breve introduzione non ha pretese di completezza, ma è resa necessaria per la comprensione degli argomenti seguenti. Come si avrà modo di vedere, infatti, la struttura innovativa di WordNet è il punto di partenza per la creazione di nuove ontologie linguistiche e delle loro estensioni (anche quelle che introducono il multilinguismo). Le nuove ontologie mantengono un grado di compatibilità con WordNet integrandone in modi diversi il database lessicale. Questo permette di sfruttare il lavoro che già è stato fatto sia per quanto riguarda le ontologie stesse, sia per quanto riguarda le applicazioni che le utilizzano, riflettendosi in entrambi i casi in minori tempi di sviluppo.

1.2 Estensione delle ontologie lessicali e multilinguismo

Un problema che ci si trova ad affrontare quando si utilizzano le ontologie linguistiche è quello della loro estensione. Con questa espressione ci si riferisce alla possibilità di inserire nuovi concetti non presenti nell'ontologia. E' impensabile infatti sviluppare un'ontologia contenente tutti i concetti di tutti i domini di conoscenza: esse contengono infatti solo termini generali, di uso più comune, cosa che le rende flessibili a diversi tipi di applicazioni. Il problema di rappresentare termini molti specifici di alcuni domini deve così essere risolto nelle applicazioni che utilizzano tali ontologie, appunto tramite le estensioni.

Al riguardo citiamo il riferimento [2] della bibliografia, in cui si propone un sistema, WNEditor, sviluppato nel contesto del progetto MOMIS, per risolvere il problema dell'annotazione di termini non presenti nel database lessicale (sull'annotazione e su MOMIS vedi [1] e [4]): questo sistema integra WordNet in un apposito database lessicale denominato momiswn, e consente di estendere l'ontologia con l'inserimento di nuovi lemmi, synset e relazioni, ed inoltre di mantenere traccia dei nuovi elementi aggiunti.

Un altro genere di estensioni sono quelle che portano alla costruzione di ontologie lessicali multilingua. Rimanendo come esempio nell'ambito, sopra citato, dell'annotazione delle sorgenti, questa necessità si pone nel momento in cui si voglia integrare sorgenti in lingue diverse (ad esempio una in italiano e una in inglese); in questo caso, le soluzioni possibili sono:

- Utilizzare un'ontologia monolingua, scegliendo quindi arbitrariamente una lingua rispetto alla quale creare le annotazioni; ciò implica la necessità per l'annotatore di tradurre manualmente gli attributi delle sorgenti scritte in lingue diverse da quella scelta per l'ontologia;

- Creare un'ontologia lessicale multilingua, che permetta l'annotazione delle sorgenti in lingue diverse e l'estrazione di relazione semantiche interlinguistiche.

La scelta seguita inizialmente nelle applicazioni è stata la prima, non essendo disponibili ontologie multilingua, come risulta anche abbastanza chiaro da quanto detto prima; la seconda scelta rappresenta l'evoluzione naturale, nel momento in cui si può disporre di tali strumenti. Per quanto riguarda il sistema MOMIS, esso sfrutta WordNet, integrando anche il meccanismo di estensione sopra citato.

Ma, in che modo è possibile estendere le attuali ontologie lessicali per costruire ontologie multilingua?

Ci sono due possibili approcci, diametralmente opposti, alla soluzione di questo problema ([6], [7]). Entrambi utilizzano come riferimento WordNet, ma lo integrano in modi differenti. Il primo, che prende il nome di *expand-model*, punta ad espandere il database Princeton WordNet collegando i synset con synset di altre lingue; in questo modo le relazioni semantiche rimangono le stesse per tutte la lingue. Il secondo invece, denominato *merge-model*, prevede la creazione di un *wordnet* per ogni lingua che deve essere rappresentata, in maniera indipendente una dall'altra; in un secondo momento i wordnet vengono messi insieme e collegati ai synset di WordNet 1.5 che più si avvicinano come significato.

1.2.1 Expand-model

L'expand-model si basa sul presupposto che tra gli stessi concetti, in linguaggi diversi, intercorrono le stesse relazioni; parlando di ontologie, se due synset in un wordnet sono legati da una relazione, i due synset equivalenti in un altro wordnet saranno legati dalla stessa relazione.

Da un punto di vista tecnico, questo modello è più semplice da implementare, ed inoltre garantisce un maggiore grado di compatibilità tra le strutture semantiche delle lingue rappresentate. Infatti viene presa come base comune WordNet (cui ci si riferisce anche con l'espressione Princeton WordNet, abbreviata in PWN), dal quale vengono importate tutte le relazioni semantiche. I synset degli specifici linguaggi, una volta inseriti, sono collegati, tramite relazioni di equivalenza, ai synset di WordNet. Così facendo si minimizzano le discrepanze strutturali tra i wordnet dovute alle decisioni soggettive che necessariamente intervengono durante la loro costruzione. D'altra parte, l'ipotesi fatta introduce il rischio di forzare eccessivamente la struttura dei wordnet (e quindi delle lingue rappresentate) in base alla struttura semantica della lingua inglese (e americana), come evidenziato in [6]. Inoltre, il modello presentato non consente di utilizzare risorse già esistenti, ma i wordnet devono essere riscritti in toto basandosi sulla struttura di WordNet: questo implica lunghi tempi di sviluppo (ma il problema può essere risolto utilizzando procedure semi-automatiche, vedi capitolo 2).

Questo modello viene implementato nel database lessicale MultiWordNet (MWN), la cui struttura sarà descritta nel capitolo 2. E' bene ricordare a questo proposito, come viene riportato anche in [7] che la procedura per la determinazione dei lexical-gap (concetti che non sono esprimibili in una lingua con un singolo termine, ma solo tramite una parafrasi o un termine con significato vicino), applicata alla coppia italiano - inglese ha rilevato che circa 1% dei synset in entrambe le lingue è un lexical gap. Questo risultato viene presentato come un sostegno empirico al modello per quanto riguarda la struttura lessicale; non dice nulla però su quella semantica, la vera discriminante tra i due approcci.

1.2.2 Merge-model

Il merge-model, a differenza del precedente, è pensato con l'intenzione di mantenere separate le strutture semantiche dei diversi linguaggi. I wordnet sono sviluppati autonomamente (vengono quindi utilizzati anche database e risorse già esistenti). In seguito viene sviluppata la parte di collegamento interlinguistico, mettendo in relazione i synset dei diversi wordnet con quelli di WordNet 1.5 che più si avvicinano nel significato.

In questo modo si mantiene l'indipendenza strutturale delle lingue e non si perdono, come avviene nel caso precedente, le differenze che dipendono dai singoli linguaggi. Inoltre vengono impiegate risorse già esistenti, cosa che permette di risparmiare parecchio lavoro. D'altra parte anche questo modo di procedere ha i suoi difetti: in primo luogo, è necessario assicurare una sovrapposizione sufficiente tra i vari wordnet, mantenendo lo stesso le proprietà specifiche dei linguaggi; in secondo luogo, si deve stabilire come devono essere interpretate le differenze che si possono incontrare tra diversi wordnet. (ad esempio due coppie di synset equivalenti in lingue diverse possono essere collegati da relazioni diverse; inoltre altre relazioni si possono incontrare nella parte comune).

Il merge-model è stato implementato nel database lessicale EuroWordNet (EWN), di cui si discuterà al capitolo 3 del presente elaborato.

Capitolo 2

Il database lessicale MultiWordNet

Il progetto MultiWordNet [7], sviluppato presso l'istituto ITC-irst di Trento, mira a creare un Italian WordNet strettamente allineato con Princeton WordNet. Il sito internet del progetto è http://multiwordnet.itc.it: a questo indirizzo sono disponibili tutti gli articoli riguardanti il progetto, oltre ad un'interfaccia web del database, tramite la quale si può provare il suo funzionamento. Sul sito internet è disponibile anche un depliant, in formato pdf, di presentazione del progetto. Come si legge nell'opuscolo, il progetto è ancora attivo, e al momento è disponibile la release 1.39, contenente 58.000 word meaning italiani organizzati in 32.700 synset, oltre alle corrispondenze con i synset equivalenti di WordNet 1.6. Si legge inoltre che al momento, il database è fornito come dump di tabelle MySQL, assieme con un paccheto di API scritte in java se richieste, mentre con la prossima release sarà disponibile un'interfaccia html-based che utilizzerà Apache/php/MySQL.

MultiWordNet implementa il modello expand-model descritto al capitolo precedente. L'obiettivo che intende perseguire questo progetto è quello di costruire un database lessicale italiano strettamente allineato con quello inglese (Princeton WordNet). L'architettura del database però è stata pensata in modo da renderla flessibile all'inserimento di nuove lingue. Inoltre, sono stati sviluppati in MultiWordNet anche dei meccanismi atti a risolvere i problemi del modello implementato che sono stati descritti al capitolo precedente. Come si vedrà in dettaglio, infatti, nella progettazione del database si è tenuto in considerazione, la necessità di rappresentare le differenze semantiche specifiche di una lingua, permettendo quindi ai nuovi wordnet di divergere dalla struttura di WordNet quando necessario; è stato inoltre inserito un meccanismo per la modellazione dei *lexical gap* (letteralmente "lacune lessicali", concetti di una lingua che non sono esprimibili con un singolo termine in un'altra lingua).

Uno degli obiettivi del progetto è anche la sperimentazione di tecniche di acquisizione (semi)automatiche. Nella costruzione dei nuovi wordnet, infatti, vengono sfruttate due procedure semi-automatiche, il cui funzionamento è reso possibile proprio dal modello adottato; in questo modo è possibile ovviare ai lunghi tempi di riscrittura dei singoli wordnet di cui si era parlato in precedenza (paragrafo 1.2.1).

Nel seguito sarà presentato più in dettaglio la struttura del database e le tecniche di costruzione. Infine si cercherà, con un processo di reverse engineering, di disegnare lo schema E/R del database, che verrà utilizzato in seguito per i confronti con le altre architetture.

2.1 Descrizione del modello

Da un punto di vista architetturale, MultiWordNet rappresenta un'estensione della matrice lessicale di WordNet in una *matrice lessicale multilingua*. Come è rappresentato nella figura che segue, viene aggiunta una terza dimensione che rappresenta la lingua. Con riferimento alla figura quindi, in larghezza si trovano i lemmi specifici di una lingua, in altezza i significati e in profondità si scorrono le lingue. Se prendiamo come base di riferimento WordNet, la costruzione della matrice lessicale multilingua con l'aggiunta dell'italiano consiste nel ri-mappare i lemmi italiani secondo i synset della lingua inglese. Il risultato è una completa ridefinizione delle relazioni lessicali (che collegano un lemma a un synset) della lingua italiana; per quanto riguarda invece le relazioni semantiche, invece, vengono mantenute il più possibile quelle definite per l'inglese. Si noti che questa architettura è facilmente estensibile ad altre lingue.

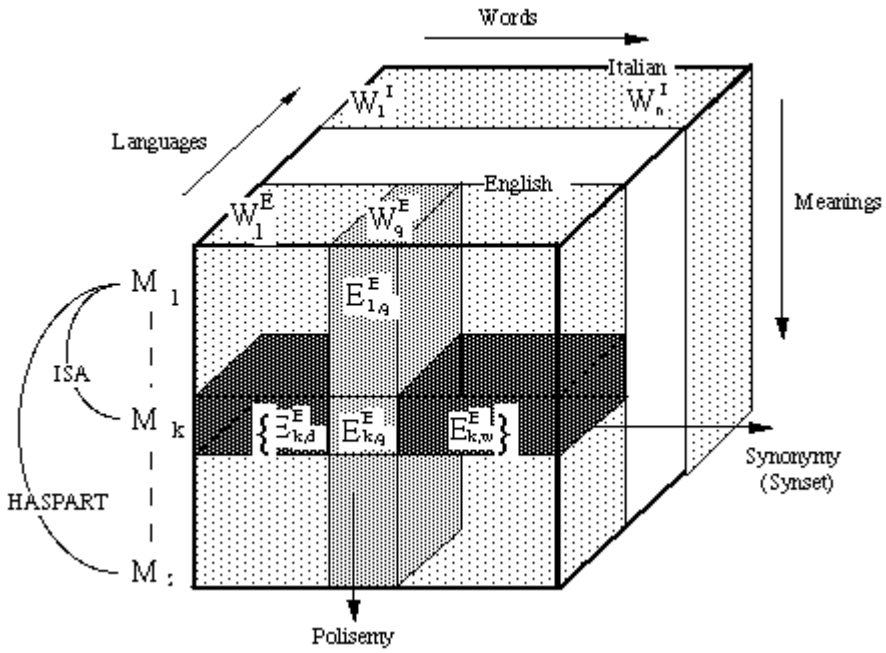


Figura 3 : La matrice lessicale multilingua di MultiWordNet

Nella matrice lessicale multilingua il concetto di synset si evolve, dunque, nel *multisynset*: un significato comune alle due lingue che viene rappresentato e individuato univocamente nel database (e quindi nell'ontologia). Il multisynset non riguarda più le relazioni di sinonimia tra lemmi di una stessa lingua; individua, invece, una relazione di sinonimia in senso lato tra synset equivalenti (e quindi tra i termini che li compongono) in lingue diverse. I multisynset in pratica, essendo indipendenti dalla lingua, rappresentano la dimensione verticale della matrice; in questo modo, le relazioni semantiche (rappresentate come archi che congiungono due synset) vengono estese in validità a tutte le lingue.

2.2 Architettura del database

Dopo aver introdotto la matrice lessicale multilingua, vogliamo ora descrivere l'architettura del database e quindi il modello dei dati di MultiWordNet ad un alto livello di analisi, come si può ricavare da [7]. Sarà poi fatto un cenno alle tecniche semiautomatiche utilizzate per popolare il database stesso.

Come risulta chiaro da quanto detto prima il database deve contenere una parte che è comune a tutti i wordnet e una specifica per ogni lingua che è rappresentata. Più precisamente, con riferimento alla figura, la struttura è organizzata in un modulo centrale, denominato *Common-db*, e due moduli specifici, *Italian-db* e *English-db*. Il Common-db contiene tutte le relazioni semantiche presenti in Princeton WordNet (quindi non i lemmi). A livello concettuale, il Common-db contiene anche i multisynset: in realtà questi non esistono nel database, ma vengono implementati utilizzando lo stesso identificatore per i synset dei moduli linguistici. Nei moduli Italian-db e English-db invece sono contenute le informazioni lessicali, vale a dire i legami tra lemmi e synset, specifiche delle lingue.

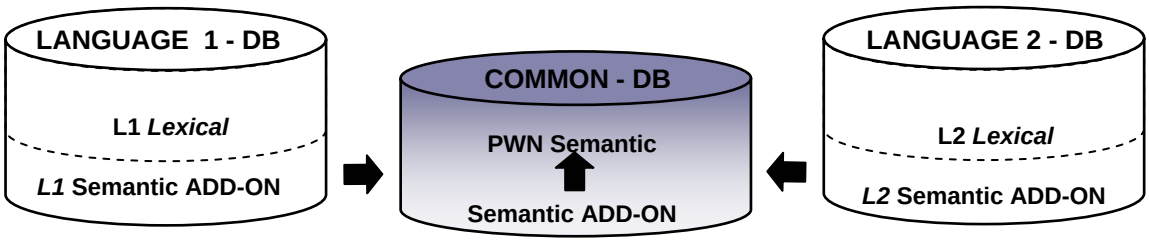


Figura 4 : Architettura software di MultiWordNet

Lo schema descritto finora è adatto a rappresentare le informazioni condivise tra le due lingue: da solo, però, non è in grado di rappresentare le loro differenze strutturali, che invece si vogliono rappresentare; inoltre si vuole poter modificare le relazioni semantiche comuni (cioè quelle di WordNet) e aggiungerne di nuove. La soluzione che è stata adottata consiste in una struttura a add-on: come si vede in figura questi sovrascrivono (nel verso delle frecce) in vari strati il database WordNet, senza però modificarlo fisicamente.

Il primo di questi *semantic add-on* è incapsulato nel Common-db e si pone direttamente sopra a WordNet; esso permette di modificare l'insieme delle relazioni semantiche comuni. Questo modulo è a sua volta sovrascritto dai semantic add-on specifici per ciascuna lingua, che sono incapsulati nei *language-db*: il loro ruolo è di rappresentare le relazioni semantiche incompatibili con quella lingua. Le informazioni contenute in questi ultimi add-on sono essenzialmente di due tipi:

- Relazioni semantiche specifiche di una lingua;
- Lexical gap.

I lexical gap si hanno quando un determinato concetto, che in una lingua è associato ad un synset non vuoto, e quindi può essere espresso con singoli lemmi, in un'altra lingua non trova un synset corrispondente ma può essere solo tradotto con una frase oppure espresso con un termine di significato più generico o più specifico; in questo secondo caso si parla più propriamente di *denotation difference*.

Ad esempio, il synset inglese associato alla glossa *"the pointed head or striking tip of an arrow"* e contenente il solo termine *"arrowhead"*, non trova in italiano un termine corrispondente, ma viene tradotto con l'espressione *"Punta di freccia"*. In questo caso ci troviamo di fronte ad un lexical gap, che viene rappresentato nel Italian-db con un synset italiano che ha associati solo il particolare lemma *"GAP!"* e la glossa *"punta_di_freccia"*. Si noti che "GAP!" viene implementato nel database come se fosse un lemma, per semplicità, anche se, in realtà, non è un vero e proprio termine.

Quando ci si trova in uno di questi casi, il concetto mancante viene collegato, nel language-db, ad un particolare empty-node presente nella parte lessicale; a questo punto si possono seguire due strategie differenti:

- nel caso di lexical gap, nella glossa del nodo viene riportata una traduzione del concetto in quella lingua;

- nel caso di denotation difference, il nodo viene collegato da una o più relazioni *nearest* a synset che esprimono concetti più generali/specifici.

Per quanto riguarda le relazioni, infine, sono importate da WordNet tutte quelle semantiche, che risultano perciò disponibili in MultiWordNet; a queste si aggiungono le relazioni di tipo nearest. L'unica relazione lessicale al momento istanziata è la sinonimia. La tabella seguente, che è stata ricavata dal file di accompagnamento al database, riassume tutti i tipi di relazioni presenti in MultiWordNet.

Il database fin qui descritto è stato implementato con tabelle MySQL, in base ad uno schema, non ben precisato nella documentazione, che si tenterà di ricostruire nel prossimo paragrafo. Per il popolamento del database è poi avvenuto nel modo seguente: dapprima è stato integrato WordNet, separando la parte semantica, che ha riempito il Common-db, da quella lessicale, inclusa nel English-db. In seguito, si sono utilizzate due procedure semiautomatiche per la costruzione del Italian-db che sfruttano entrambe un dizionario bilingue italiano-inglese:

- La prima è denominata *Assign-procedure* e serve per individuare, dato un termine italiano ed un suo significato, una lista dei synset inglesi più simili; i synset vengono pesati in base ad una serie di regole di match per i cui dettagli si rimanda a [7];

- La seconda è la già citata LG-procedure, che aiuta a trovare i lexical gap; i lexical gap English-to-Italian possono quindi essere automaticamente esclusi dai risultati della routine precedente, mentre quelli Italian-to-English rappresentano synset italiani per i quali la Assign-procedure non può restituire alcun risultato, e che perciò devono essere aggiunti manualmente.

Categoria sintattica	Codice	Nome della Relazione	Note
Nome (Noun)	!	Antonym	Lexical
	@	Hypernym	
	~	Hyponym	
	#m	Member-of	
	#s	Substance-of	
	#p	Part-of	
	%m	Has-member	
	%s	Has-substance	
	%p	Has-part	
	=	Attribute	
	|	Nearest	New IRST
Verbo (Verb)	!	Antonym	Lexical
	@	Hypernym	
	~	Hyponym	
	*	Entailment	
	>	Causes	
	^	Also-see	
	$	Verb-group	New 1.6
	|	Nearest	New IRST
Aggettivo (Adj)	!	Antonym	Lexical
	&	Similar-to	
	<	Participle	Lexical
	\	Pertains-to	Lexical
	=	Is-value-of	
	^	Also-see	
	|	Nearest	New IRST
Avverbio (Adv)	!	Antonym	Lexical
	\	Derived-from	
	|	Nearest	New IRST

Tabella 1 : Tipi di relazioni in MultiWordNet

2.3 Schema E/R del database

Quello che si voleva ottenere dallo studio condotto era uno schema E/R equivalente del database, da poter utilizzare nel futuro confronto con EuroWordNet. L'analisi strutturale di MultiWordNet è iniziata con la lettura di [7]. In seguito però è sorta la necessità di chiarire alcuni dettagli dell'implementazione, riguardo i quali la documentazione era vaga e lacunosa. L'unica alternativa possibile è stata quella di operare direttamente sul database una tecnica di riverse engineering.

MultiWordNet viene distribuito come dump di un database MySQL (sul sito del progetto è ora disponibile anche un depliant che illustra le licenze d'uso e i costi). Nei documenti e nel sito sono presenti dei riferimenti più o meno espliciti sia ad una serie di API per l'interfacciamento al database, che ad una vera e propria interfaccia per la navigazione e l'aggiornamento dello stesso. Una versione web di tale interfaccia è presente sul sito del progetto: essa permette di provare via internet la versione del database installata sul server dell'IRST senza bisogno di scaricare e installare alcun software. Sfortunatamente, la versione di MultiWordNet a nostra disposizione non è corredata di tali strumenti, quindi, l'unico modo possibile per interrogare il database sono le query SQL.

Presentiamo ora i risultati della nostra analisi, ossia lo schema E/R del database. Nel diagramma seguente si farà riferimento a titolo d'esempio ai soli Italian-db e Common-db, essendo English-db speculare al primo. Per semplificare il diagramma non sono rappresentati gli identificatori esterni; per completezza, saranno illustrati a parole nel commento che accompagna il diagramma. Per aiutare la lettura sono stati inoltre evidenziati i macro-moduli descritti in precedenza, mentre per uniformità con la presentazione precedente si è cercato di mantenere il più possibile la nomenclatura di MultiWordNet.

Iniziamo il commento partendo dal Common-db. Si vede, al centro del modulo, l'entità *MultiSynset*: la specializzazione totale ed esclusiva, suddivide poi i multisynset tra i *WordNet_Synset* (per i quali viene mantenuto l'identificatore originale di WordNet, una stringa del tipo pos#offset, dove pos è un carattere che identifica la categoria sintattica, offset è un numero), e i *New_Synset*, i nuovi synset aggiunti (per questi l'identificatore è del tipo pos#Noffset). L'indicazione circa le categorie sintattiche (*part of speech*) è dunque contenuta nell'identificatore del synset, e quindi è propria del MultiSynset. Considerando quest'attributo, a ben vedere, di tipo numerabile, l'associazione è stata modellata attraverso l'entità *Part_Of_Speech*, con un procedimento analogo al tipo delle associazioni, *Rel_Type*.

Ristrettamente al Common-db, i MultiSynset partecipano all'associazione con *Common_Relation*, che incapsula le relazioni semantiche comuni: come si vede quest'entità è associata obbligatoriamente ad un *Rel_Type* (i tipi delle relazioni riassunti nella tabella precedente) e a due MultiSynset (un source e un target). L'attributo *status* serve a riconoscere le relazioni importate da WordNet (per le quali vale null) e quelle aggiunte successivamente (attualmente viene utilizzata solo lo stato "new"). Per l'entità Common_Relation esiste un identificatore esterno, costituito dall'unione dei due synset più il tipo della relazione (C_Rel_ID = (Source_MSyn_ID, Target_MSyn_ID, Rel_Type_Code)).

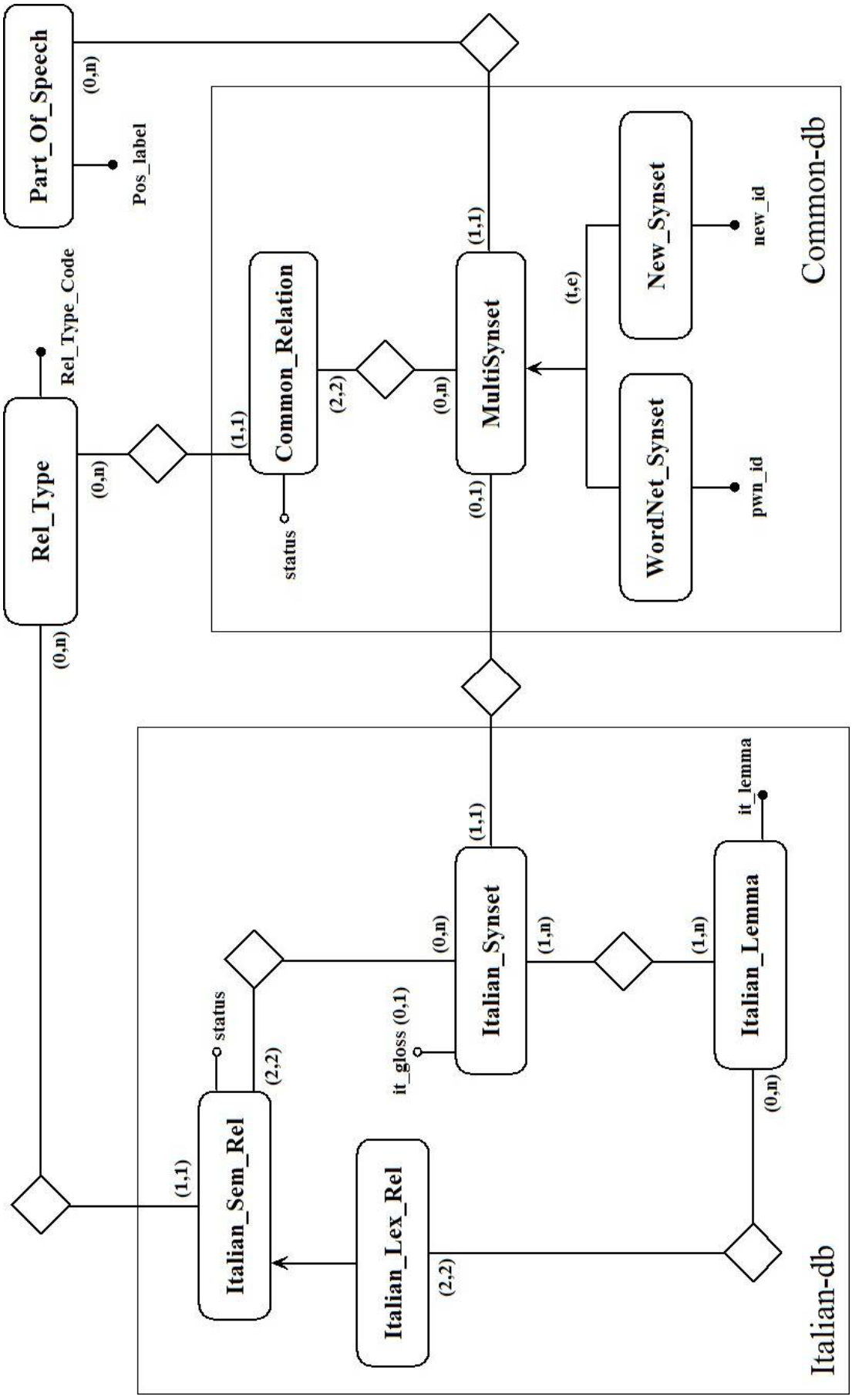


Figura 5 : Schema E/R di MultiWordNet

Si noti come, in pratica, WordNet sia integrato totalmente in MultiWordNet in questo modo: la parte semantica, quella cioè relativa ai synset e alle loro relazioni, viene incapsulata nel common-db, nelle entità WordNet_Synset e Common_Relation; la parte lessicale, riguardante i lemmi e il loro collegamento ai synset, viene invece rappresentata nel English-db (non presente nello schema E/R), tramite le entità English_Synset e English_Lemma.

Il legame tra Common_db e Italian_db è dato dall'associazione tra MultiSynset e *Italian_Synset*. Come si vede un Italian_Synset è necessariamente legato ad uno ed un solo MultiSynset: questo legame dipende proprio dalla rappresentazione dei multisynset tramite la condivisione dell'identificatore per synset di diverse lingue. Perciò l'identificatore non può essere specifico del linguaggio, ma è necessariamente comune: il MultiSynset costituisce dunque un identificatore esterno per l'entità Italian_Synset. L'attributo it_gloss (la glossa del synset) è opzionale, quindi viene indicato con cardinalità (0,1).

Italian_Synset è poi collegato all'entità *Italian_Lemma*, che rappresenta la parte lessicale del database, cioè contiene tutti i termini italiani. L'entità *Italian_Sem_Rel*, permette di rappresentare relazioni semantiche valide solo per la lingua italiana: come Common_Relation, è associata ad un tipo e a due synset italiani, e l'unione dei tre attributi costituisce un identificatore esterno per quest'entità (It_Sem_Rel_ID = (Source_ISyn_ID, Target_ISyn_ID, Rel_Type_Code)). L'entità che la specializza, *Italian_Lex_Rel*, rappresenta invece le relazioni lessicali (cioè associate a due singole parole) specifiche per l'italiano; per quest'ultima entità, l'identificatore esterno comprende anche i due lemmi (It_Lex_Rel_ID = (Source_ISyn_ID, Target_ISyn_ID, Rel_Type_Code, Target_Lemma, Source_Lemma)).

Un'ultima nota riguarda i lexical gap: essi vengono rappresentati come synset associati ad un particolare lemma, la cui parola è "GAP!"; le eventuali relazioni nearest fanno parte delle Italian_Sem_Rel.

2.4 Traduzione dello schema E/R nello schema relazionale

Lo schema relazionale che segue traduce lo schema E/R precedente. Come si vede, la relazione MultiSynset traduce l'omonima entità, nella quale è stata collassata verso l'alto la gerarchia di specializzazione. L'attributo *Type* che è stato aggiunto serve appunto per rappresentare le due categorie con la stessa relazione (ad esempio con i due valori "WordNet" e "New"). Il subset di Italian_Sem_Rel è rappresentato con due relazioni distinte (*Italian_Sem_Rel* e *Italian_Lex_Rel*). Infine sono state inserite le relazioni per il English-db, identiche a quelle del database italiano.

1) **Tipi comuni:**

`Rel_Type (Rel_Type_Code)`

`Part_Of_Speech (Pos_label)`

2) Common-db:

```
MultiSynset (MSyn_Id, Type, Pos_label)
      FK: Pos_label References Part_Of_Speech

Common_Relation (Rel_Type_Code, Source_MSyn_ID, Target_MSyn_ID, status)
      FK: Source_MSyn_ID References MultiSynset
      FK: Target_MSyn_ID References MultiSynset
      FK: Rel_Type_Code  References Rel_Type
```

3) Italian-db:

```
Italian_Synset (ItSyn_ID, It_gloss)
      FK: ItSyn_ID References MultiSynset
Italian_Lemma (It_Lemma)

It_Lemma_Syn (It_Lemma, ItSyn_ID)
      FK: ItLemma References Italian_Lemma
      FK: ItSyn_ID References Italian_Synset

Italian_Sem_Rel (Rel_Type_Code, Source_ItSyn_ID, Target_ItSyn_ID, status)
      FK: Source_ItSyn_ID References Italian_Synset
      FK: Target_ItSyn_ID References Italian_Synset
      FK: Rel_Type_Code   References Rel_Type

Italian_Lex_Rel (Rel_Type_Code, Source_ItSyn_ID, Target_ItSyn_ID,
                     Source_ItLemma, Target_ItLemma, status)
      FK: Source_ItSyn_ID References Italian_Synset
      FK: Target_ItSyn_ID References Italian_Synset
      FK: Rel_Type_Code   References Rel_Type
      FK: Source_ItLemma References Italian_Lemma
      FK: Target_ItLemma References Italian_Lemma
```

4) English-db:

```
English_Synset (EnSyn_ID, En_gloss)
      FK: EnSyn_ID References MultiSynset

English_Lemma (En_Lemma)
```

En_Lemma_Syn (En_Lemma, EnSyn_ID)

FK: EnLemma **References** English_Lemma

FK: EnSyn_ID **References** English_Synset

English_Sem_Rel (Rel_Type_Code, Source_EnSyn_ID, Target_EnSyn_ID, status)

English_Lex_Rel (Rel_Type_Code, Source_EnSyn_ID, Target_EnSyn_ID,
Source_EnLemma, Target_EnLemma, status)

FK: Source_EnSyn_ID **References** English_Synset

FK: Target_EnSyn_ID **References** English_Synset

FK: Rel_Type_Code **References** Rel_Type

FK: Source_EnLemma **References** English_Lemma

FK: Target_EnLemma **References** English_Lemma

Capitolo 3

Il database lessicale EuroWordNet

EuroWordNet è un progetto, finanziato dalla Comunità Europea, iniziato nel marzo del 1996 e terminato nel mese di luglio del 1999, al quale hanno partecipano Istituti e Università di alcuni stati membri. Il sito internet di riferimento, dal quale è possibile scaricare una versione dimostrativa del database e del suo browser Periscope, si trova all'indirizzo http://www.illc.uva.nl/EuroWordNet/.

Come detto il progetto è stato completato nell'estate del 1999, e, al momento, lo schema del database, le relazioni definite, la top-ontology e l'Interlingual Index non vengono sviluppati. Ciononostante altri istituti stanno sviluppando wordnet in altre lingue sulla base delle specifiche di EuroWordNet, in modo che siano compatibili con il modello e che possano, quindi, essere integrati nell'ontologia già esistente.

L'obiettivo del progetto EuroWordNet era la costruzione di un database lessicale multilingua consistente e affidabile, e al tempo stesso in grado di conservare la ricchezza e la diversità delle diverse lingue. Riguardo a EuroWordNet si vedano i documenti [7] e [8] per il modello, e [9] per l'architettura e i dettagli implementativi del database.

Si è deciso di implementare in EuroWordNet il modello merge-model discusso al paragrafo 1.2.2. Per fare ciò, dapprima è stato creato un wordnet per ognuna delle lingue rappresentate, a partire da risorse preesistenti, in un modo simile alla costruzione di Princeton WordNet 1.5 e in maniera indipendente dagli altri wordnet. Il progetto prevedeva inizialmente l'inserimento di tre lingue, italiano, olandese e spagnolo, oltre all'inglese importato da WordNet; in una successiva estensione sono state aggiunte altre quattro lingue, francese, tedesco, ceco ed estone. In un secondo momento, sono stati collegati i synset di ogni wordnet al più vicino significato di WordNet 1.5. Si è cercato infine di risolvere i problemi del modello, vale a dire di assicurare sufficiente sovrapposizione tra i concetti presenti nei singoli wordnet, e di interpretare le differenze strutturali che si incontrano in diversi wordnet.

In questo capitolo verranno approfonditi meglio alcuni aspetti riguardanti il modello e l'implementazione. Sarà illustrata la struttura a moduli del database e, in seguito, si descriveranno i singoli moduli e i record in essi contenuti. Infine si cercherà di riassumere le informazioni raccolte in uno schema E/R equivalente al database presentato.

3.1 Descrizione del modello

Come già detto, il modello punta alla costruzione di un database, o meglio un'ontologia, multilingua, rappresentante alcune lingue europee. I wordnet delle singole lingue vengono costruiti indipendentemente uno dall'altro, sul modello di WordNet (la cui struttura è stata descritta nel capitolo 1) e sfruttando il più possibile risorse preesistenti. I wordnet così costruiti vengono collegati a WordNet 1.5 con relazioni di equivalenza tra synset. In questo modo si è creata un'unica ontologia contenente sia informazioni intralinguistiche che interlinguistiche. Il lavoro è stato poi completato raccogliendo i concetti e i termini più importanti delle diverse lingue in un'unica ontologia centrale, indipendente dalle singole lingue. Questo garantisce la compatibilità e permette di controllare maggiormente la consistenza dei dati tra diversi wordnet; le differenze dipendenti dalla lingua possono comunque essere mantenute nei singoli wordnet.

Il processo di costruzione del database è schematizzato nella figura seguente. Partendo da risorse già esistenti (dizionari elettronici, altri wordnet…), viene selezionato un sottoinsieme (*subset*) di significati, per i quali vengono estratte tutte le parole che li definiscono, vale a dire i lemmi ad essi collegati. Vengono individuate all'interno del subset le relazioni semantiche specifiche delle lingue (*language internal relations*); si ottiene così una classificazione dei significati in synset, che vengono poi collegati, tramite *equivalence relation*, ai synset di WordNet più vicini. Queste operazioni vengono poi ripetute, in modo da costruire incrementalmente i wordnet che saranno poi caricati in EuroWordNet. In seguito vengono effettuati controlli di consistenza interlinguistici: gli errori che emergono in questo passaggio sono corretti ristrutturando le gerarchie dei synset e le loro relazioni. Si costruisce infine una *top-ontology* comune, raccogliendo, dai singoli wordnet, i vertici delle gerarchie e i nodi più frequentemente coinvolti nelle relazioni.

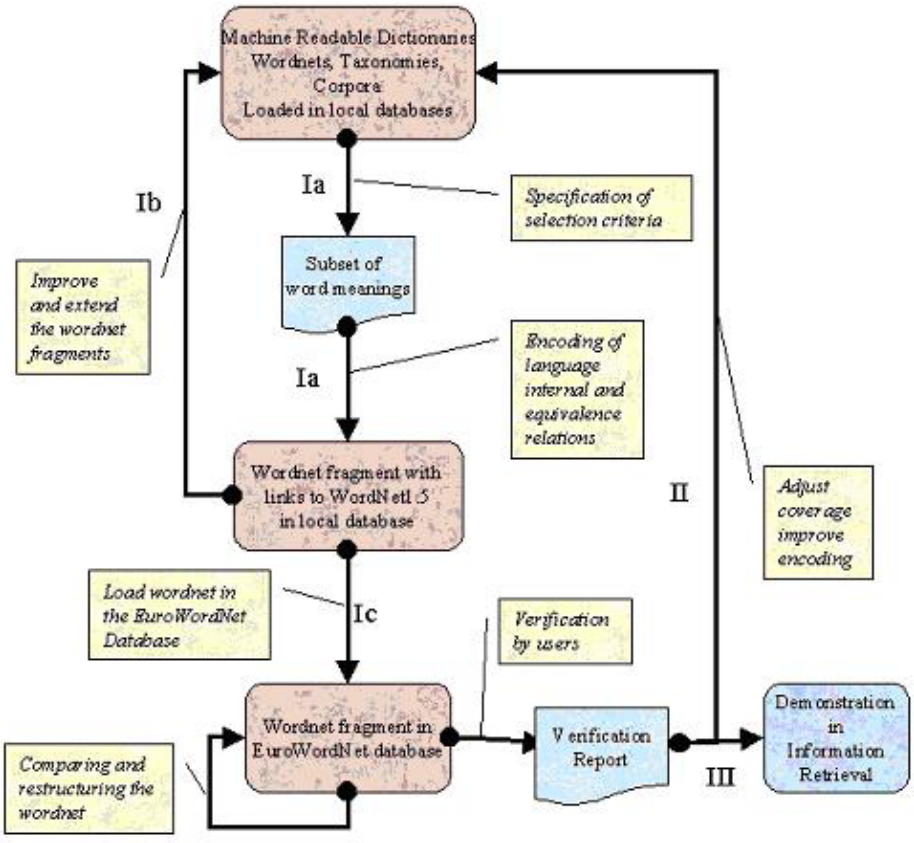


Figura 6 : Il processo di costruzione di EuroWordNet: schema a blocchi

Utilizzando specifici criteri di scelta (come i controlli di consistenza tra diversi wordnet), si mira a codificare in EuroWordNet un subset delle risorse di partenza che contenga tutte le parole più comuni delle lingue rappresentate e i concetti "padre" necessari per definire concetti più specifici; deve inoltre garantire la massima sovrapposizione dei concetti contenuti nei singoli wordnet.

Infine, altre tecniche vengono utilizzate per individuare e interpretare le differenze semantico-lessicali tra diversi wordnet. Ad esempio si possono istanziare diversi tipi di relazioni di equivalenza, per rappresentare le denotation difference (vedi paragrafo 2.2): non soltanto la sinonimia interlinguistica, quindi, ma anche l'ipernimia e l'iponimia.

3.2 Architettura funzionale del database

La figura seguente mostra la struttura del database EuroWordNet, come viene presentato in [8] e in [9].

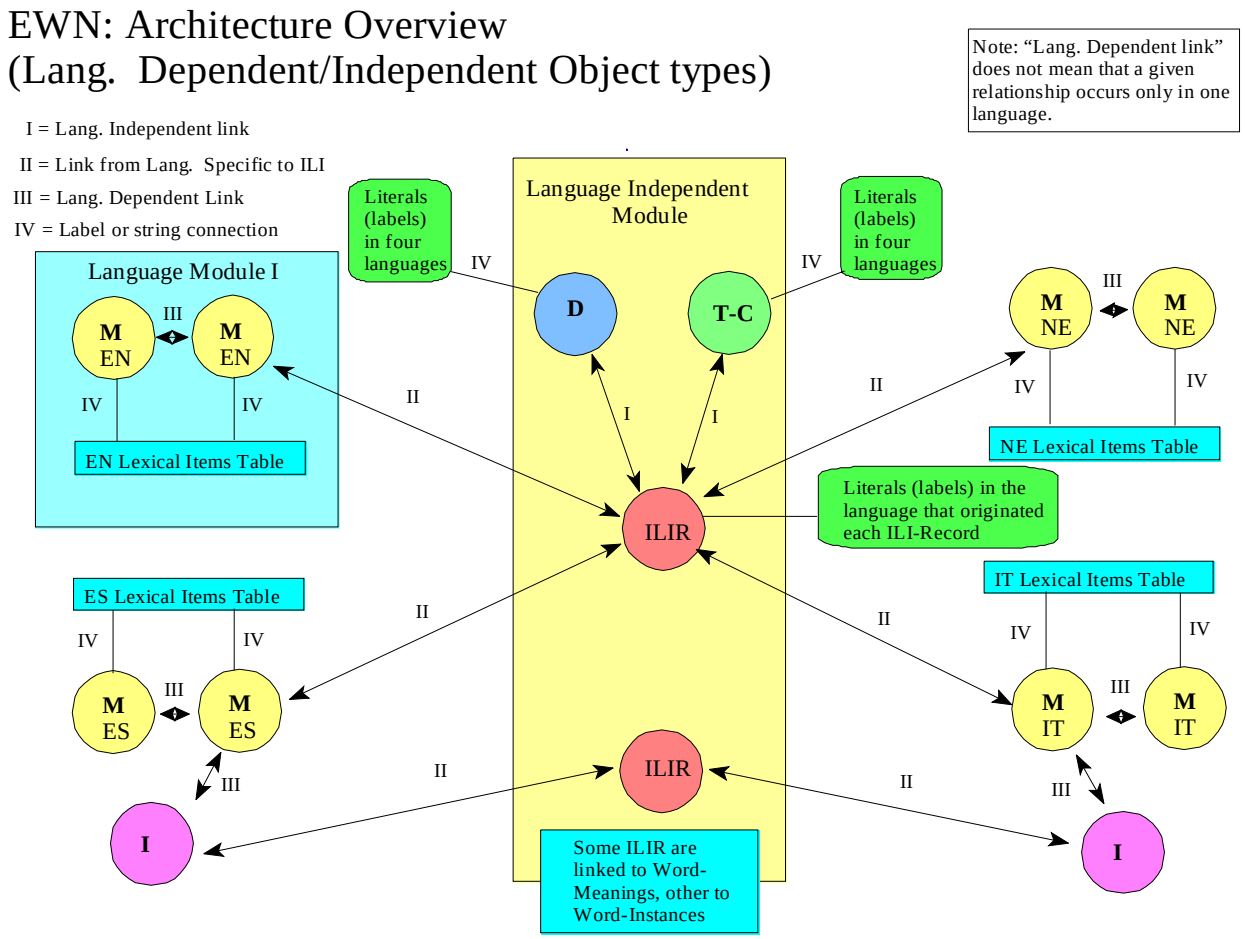


Figura 7 : Architettura funzionale del software EuroWordNet

Come si vede, ad un alto livello di dettaglio, sono visibili due elementi costitutivi fondamentali:

1. Un insieme di *Language Module*, rappresentanti ciascuno un singolo wordnet; ne è perciò presente uno per ogni lingua rappresentata;

2. Un *Language Indipendent Module*, che incapsula la top-ontology descritta sopra e che funge da ponte interlinguistico tra i singoli wordnet.

Scendendo nel dettaglio, i moduli sono composto ciascuno da specifici record, collegati tra loro da relazioni sia interne che esterne ai moduli stessi. Nella figura vengono indicate con i numeri romani le quattro categorie in cui si suddividono tali relazioni. Nel seguito sarà presentata la struttura interna dei moduli, mentre nel paragrafo 3.3 si descriveranno uno ad uno i singoli record. Si è scelto in tutta la trattazione seguente di mantenere la nomenclatura originale inglese utilizzata in [8].

3.2.1 Language Module

Ogni Language Module contiene due tipi di record, I *Word-Meaning* record (WM) e i *Word-Instance* record (I). Essi corrispondono ai synset di WordNet, come si vedrà meglio in seguito; la differenza non è ben chiara, avendo essi struttura identica: l'unica differenza è che i primi possono essere collegati a tutte le categorie sintattiche, mentre i secondi solo a quella dei nomi propri (*Proper_Noun*). Contiene inoltre l'insieme dei lemmi della lingua cui si riferisce, indicato nel disegno come XY Lexical Items Table, dove XY rappresenta il codice della lingua (IT = italiano, ES = spagnolo, EN = inglese, NE = olandese). I lemmi sono incapsulati in oggetti chiamati *literal*, la cui struttura sarà illustrata nel seguito.

I suddetti record partecipano a due tipi di relazioni, chiamate *Language Internal Relation* e *Interlingual Relation*. Le prime, indicate in figura con l'indice III, coinvolgono due record appartenenti allo stesso Language Module e sono le relazioni semantiche proprie di una lingua. Le seconde, invece, collegano tra loro un record di un Language Module ed un ILI-Record, e rappresentano perciò le relazioni di equivalenza interlinguistiche; sono individuate nella figura dal numero II.

3.2.2 Language Indipendent Module

Si suddivide nei seguenti tre sotto-moduli:

1. **InterLingual Index (ILI):** Contiene gli *ILI-Record* (ILIR). Essi partecipano a due tipi di relazioni, Interlingual Relation (vedi paragrafo precedente) e Language Indipendent Relation (collegano un ILIR ad un Top-concept record o ad un Domain record).

2. **Domain Ontology:** Contiene i *Domain* record (DOM). Essi partecipano a due tipi di relazioni: Module Internal Relation (tra due Domain record) e Module External Relation (tra un Domain record ed un ILIR), entrambe classificate come Language Indipendent Relation.

3. **Top-concept Ontology:** Contiene i *Top-concept* Record (TC). Questi partecipano a 2 tipi di relazioni: Module Internal Relation (tra due Top-concept record) e Module External Relation (tra un Top-concept record ed un ILIR), entrambe classificate come Language Indipendent Relation.

Le relazioni Language Indipendent Relation sono indicate nella figura con il numero I.

Da ricordare infine che tutte le tre categorie di record partecipano ad associazioni con i literal (ovvero i lemmi), delle singole lingue. Tali associazioni, indicate con il numero IV, saranno precisate meglio nel prossimo capitolo.

3.3 Struttura dei record

In questo capitolo saranno descritti in dettaglio i record introdotti al capitolo precedente, per meglio chiarire la successiva costruzione dello schema E/R. Per maggiori informazioni al riguardo si veda [9].

3.3.1 Literal e Part of Speech

Nel database EuroWordNet i lemmi sono rappresentati attraverso oggetti denominati *literal*. In generale, la struttura di un literal è la seguente:

- String: la parola vera e propria
- ID number: identificatore (numerico) univoco
- Per ogni significato a cui è associato, viene riportato un Part Of Speech ed un sense number
- Campi opzionali per i lemmi composti da più parole (multiword), ed un language label non sempre presente

Nel database di EuroWordNet, i significati associati ad una singola parola sono raggruppati in base alla categoria sintattica. Rispetto a WordNet, ne viene introdotta una nuova, quella dei nomi propri. Ci si riferisce quindi ad un tipo enumerabile part-of-speech (POS) che può assumere i seguenti cinque valori:

- Noun (N)
- Proper Noun (PN)
- Verb (V)
- Adjective (Adj)
- Adverb (Adv)

Come in Wordnet, ad ognuno dei significati associati ad un lemma viene associato un numero (sense number) univoco all'interno della stessa categoria sintattica. In questo modo ogni significato viene identificato univocamente in due modi:

- Da un identificatore (numerico) univoco
- Dalla tripla (lemma, POS, sense_number)

Il campo relativo alla lingua, infine, è indicato come opzionale: i moduli delle diverse lingue infatti sono memorizzati separatamente, per cui l'indicazione sulla lingua è implicita nella scelta della tabella per quanto riguarda i literal relativi a Word-Meaning e Word-Instance. Alcuni literal però sono associati al Language Indipendent Module, e possono appartenere a ciascuna delle lingue di EuroWordNet.

3.3.2 Word-Meaning record e Word-Instance record

Questi due record hanno la stessa identica struttura e contengono i seguenti campi obbligatori:

- Synset ID: identificatore numerico univoco del record
- Synset type: una label che assume i valori "Word-Meaning" e "Word-Instances" a seconda dei casi
- POS: un identificatore di tipo part-of-speech. Nel caso di Word-Instance record, può assumere solo il valore PN (proper noun)
- Un array di coppie (Relation ID, Synset ID) che rappresentano le relazioni con synset appartenenti alla stessa lingua (indipendentemente di tipo WM o I), le cosiddette Language Internal Relation
- Un array di coppie (Relation ID, ILI ID) che rappresentano le relazioni con ILI-record (Interlingual relation)
- Un array di literal, per ciascuno dei quali è necessariamente indicato:
    - String
    - Sense number
    - Status (new, revise…)

Nel caso di Word-Meaning e Word-Instance, dunque, possono esserci più literal riferiti allo stesso record; in questo modo si rappresentano i synset specifici di una determinata lingua. Come si vede per questi synset non è possibile specificare una glossa. Da quanto detto in precedenza, discende che:

- I synset hanno un identificatore numerico univoco all'interno del database
- I synset collegati ad un singolo literal sono raggruppati per categoria sintattica e identificati univocamente, all'interno di ciascuna categoria, da un sense number.

Infine, sono possibili molti campi opzionali che qui non riportiamo e che riguardano essenzialmente literal (per rappresentare variazioni dialettali, particolari traduzioni…). Per maggiori dettagli vedere [9].

3.3.3 Interlingual Index record

La struttura degli ILI-record è formata dai seguenti campi obbligatori:

- ILI Record ID: identificatore numerico univoco del ILI-record
- POS: un'etichetta di tipo part-of-speech (vedi paragrafo precedente)
- GLOSS: la glossa (in inglese) che descrive il record
- Un array di coppie (Relation ID, Synset ID) che rappresentano le relazioni con record di tipo Word-Meaning e Word-Instance (Interlingual relation)

- Un array di coppie (LI Relation ID, ILI/TC/DOM Record ID) rappresentanti le relazioni con record di tipo ILI-record, Top-concept record, Domain record (le cosiddette Language Indipendent Relation)
- Un array di Origin ID: identificatore che indica il synset originario per l'ILI-record.
- Un array di literal, per ciascuno dei quali è necessariamente indicato:
    - String
    - Sense number
    - Status (new, revise…)
    - Language

Gli unici campi opzionali riguardano i literal e servono per specificarne l'origine.

Dal fatto che le relazioni interlinguistiche non abbiano un attributo che specifichi la lingua di appartenenza del synset, si deduce (ma non è specificato nei documenti) che il Synset ID è univoco non solo all'interno di un Language Module, ma in tutto il database.

Per quanto riguarda poi gli Origin ID, normalmente essi sono synset identifier di WordNet 1.5: come detto in precedenza, infatti, per costruire le relazioni interlinguistiche ci si è appoggiati inizialmente a WordNet. In seguito però sono stati aggiunti nuovi ILI-record, ad esempio per concetti mancanti in WordNet; perciò quindi anche i Synset ID di EuroWordNet possono essere Origin ID.

Infine bisogna sottolineare che anche gli ILI-record possono essere dei synset, in quanto possono essere collegati ad uno o più literal. Il linguaggio di default per i literal collegati agli ILI-record è l'inglese, in quanto derivanti da WordNet; tuttavia ne possono essere aggiunti di nuovi che abbiano lemmi in altre lingue. Nei documenti invece si specifica che la glossa deve essere espressa in inglese.

3.3.4 Top-concept record e Domain record

Un record di tipo Top-concept contiene i seguenti campi obbligatori:

- Top-concept record ID
- GLOSS
- Un array di coppie (LI Relation ID, ILI/TC record ID) rappresentanti le relazioni con record di tipo ILI-record o Top-concept record (Language Indipendent Relation)
- Una coppia (Language ID, string) che identifica un literal in un determinata lingua

Un record di tipo Domain, invece:

- Domain record ID
- GLOSS
- Un array di coppie (LI Relation ID, ILI/Domain record ID) rappresentanti le relazioni con record di tipo ILI-record o Domain record (Language Indipendent Relation)
- Una coppia (Language ID, string) che identifica un literal in un determinata lingua

Per TC e DOM non possono essere costruiti dei synset. Ciascuno di questi record è infatti associato ad un solo literal: per questo motivo non presente il campo sense number. Per quanto riguarda la lingua dei literal e della glossa, valgono le stesse considerazioni fatte per gli ILI-record.

3.4 Schema E/R del database

Il database EuroWordNet è distribuito in due formati:

- Un insieme di file di database, visualizzabili attraverso l'interfaccia Periscope
- Un insieme di file di testo.

I file di database non sono in formato relazionale, come per MultiWordNet, ma sono codificati secondo un modello chiamato Novell concept-net. I file di testo, invece, sono strutturati secondo un formato di import/export che è specificato in [9] e che è utilizzato sia per inserire dati nel database che per estrarli.

L'obiettivo che ci siamo prefissi era di rappresentare EuroWordNet con uno schema E/R. Non avendo a disposizione indicazioni sulla struttura fisica del database, e non potendo analizzarla direttamente, ma solo tramite il suo browser, ci siamo dovuti basare esclusivamente sull'architettura concettuale del database, come descritta nei paragrafi 3.2 e 3.3. D'altra parte, su di essa si basa il formato di import/export dei dati, per cui possiamo dire che rappresenta l'immagine visibile dall'esterno di EuroWordNet. Inoltre, questa struttura a record assomiglia molto al modello relazionale, e, a ben vedere, anche al modello E/R, per cui il nostro lavoro di reverse engineering è partito proprio da queste informazioni. Come risultato, si è ottenuto uno schema E/R che rappresenta non direttamente la struttura fisica di EuroWordNet, ma come esso si presenta ad un utilizzatore esterno.

Lo schema E/R ottenuto, che sarà presentato nel proseguimento di questo paragrafo, data la sua complessità, è stato suddiviso in cinque parti. Si è cercato di mantenere il più possibile la nomenclatura utilizzata al capitolo precedente, per uniformare la trattazione. Si è scelto, inoltre, di non rappresentare i campi opzionali accennati al capitolo precedente. Le scelte fatte, in ogni caso, non limitano la validità della nostra analisi, in quanto non contrastanti con l'obbiettivo del presente elaborato, che consiste nel confrontare le due architetture da un punto di vista strutturale, nelle loro componenti essenziali.

Il primo diagramma mostra la parte centrale dello schema. Al centro dello schema si vede l'entità Interlingual_Rel, che collega tra loro un ILI_record, un EuroWordNetSynset_ID e un Rel_Type. E' stata rappresentata la generalizzazione Synset_ID in funzione dell'associazione *origin* cui partecipa anche ILI_record e che è stata spiegata nel capitolo precedente. E' stato inoltre necessario rappresentare la generalizzazione EWN_Lang_Syn ed associare ad essa il synset id, per poter esprimere l'univocità dell'identificatore in tutto il database; questa generalizzazione si specializza poi nei synset specifici delle singole lingue (per semplicità sono state indicate solo le quattro originarie).

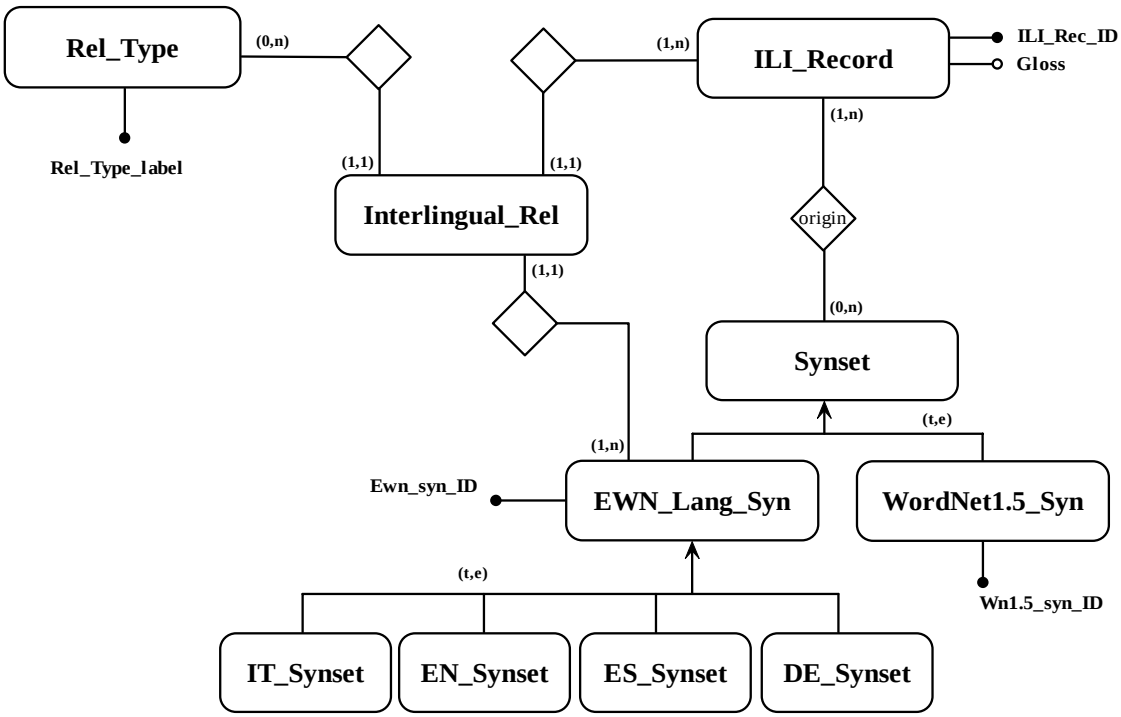


Figura 8 : Schema E/R di EuroWordNet, parte I: Interlingual Relation

Il secondo e il terzo grafico mostrano invece un particolare del Language Module italiano, preso come esempio dei moduli linguistici. Entrambi si collegano al precedente tramite l'entità IT_Synset. Il primo è centrato sui synset e sulle relazioni intralinguistiche; il secondo invece mostra la particolare associazione (reificata) tra synset e literal ed esprime il vincolo per cui la tripla (word, POS, sense number) identifica univocamente il synset.

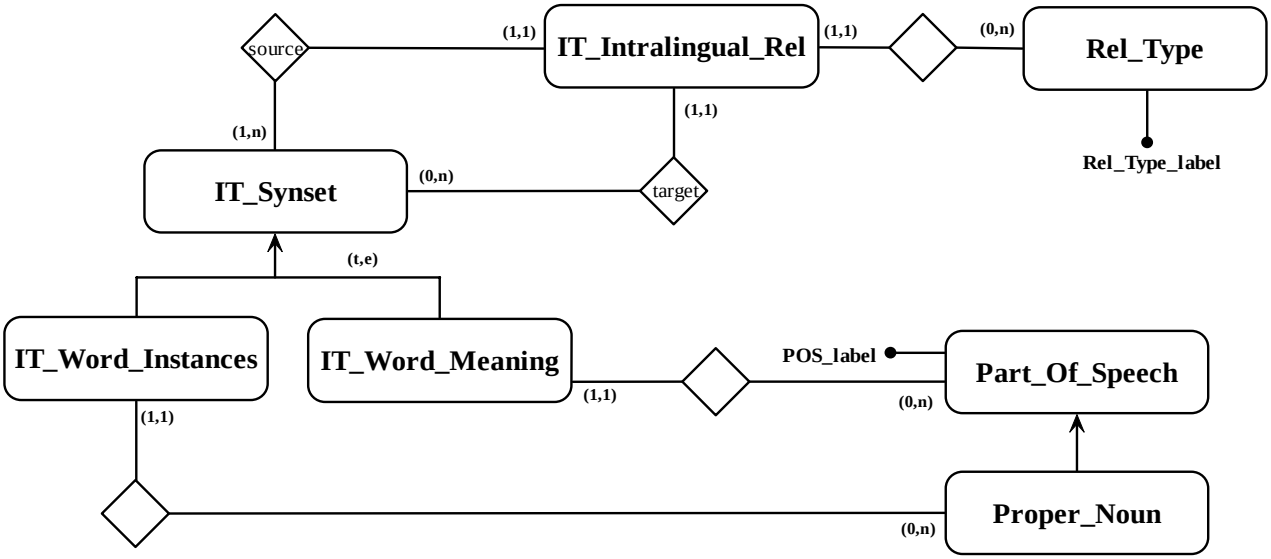


Figura 9 : Schema E/R di EuroWordNet, parte II: Intralingual Relation e Synset

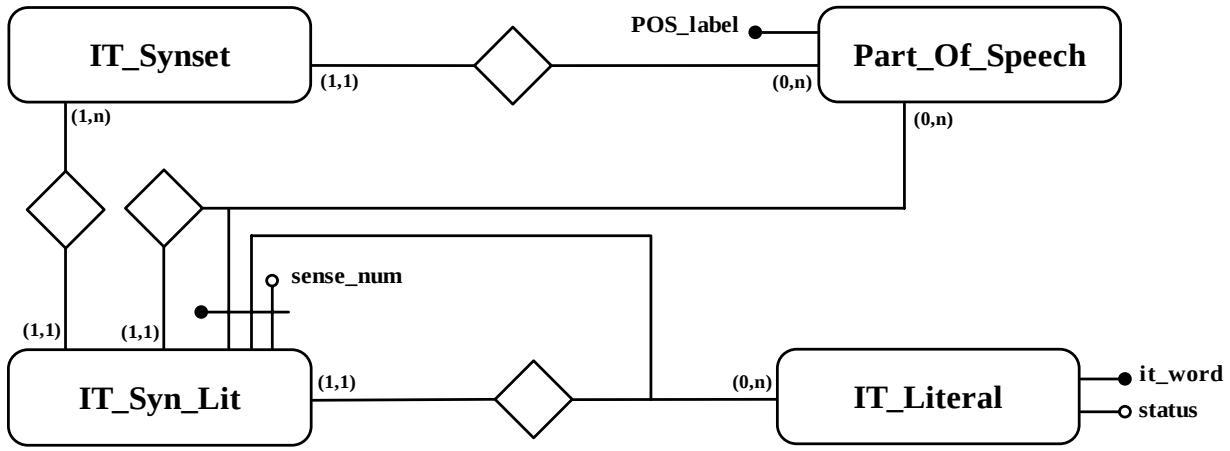


Figura 10 : Schema E/R di EuroWordNet, parte III: Synset e Literal

Gli ultimi tre diagrammi riguardano infine il Language Indipendent Module: i primi due mostrano le relazioni che sono interne al modulo stesso (Language Indipendent), mentre il secondo, le associazioni con i Literal delle lingue (XX è il codice della lingua). Le relazioni che sono presentate nel primo schema come semplici associazioni, dovrebbero essere reificate, come viene indicato nel secondo diagramma. Per quanto riguarda l'ultimo grafico, non è necessario aggiungere nulla rispetto a quanto detto a proposito della figura 10 e, al capitolo precedente, a proposito dei record.

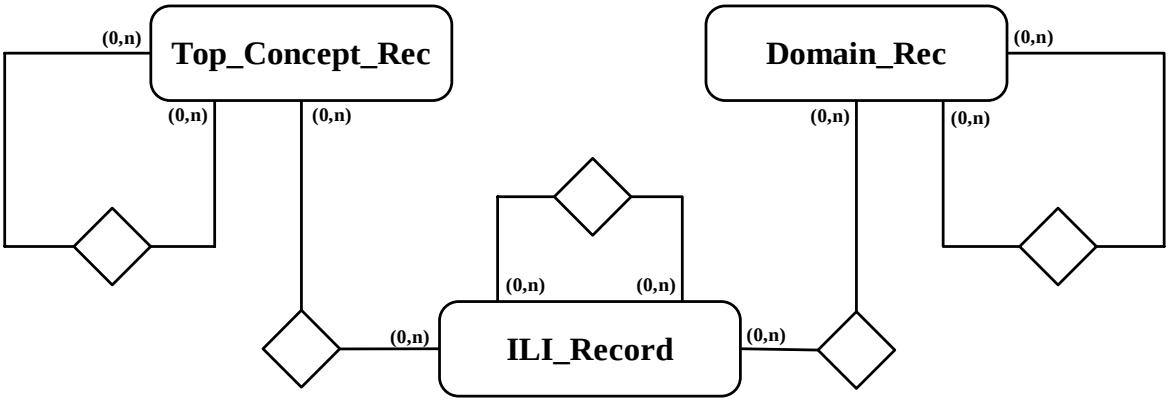
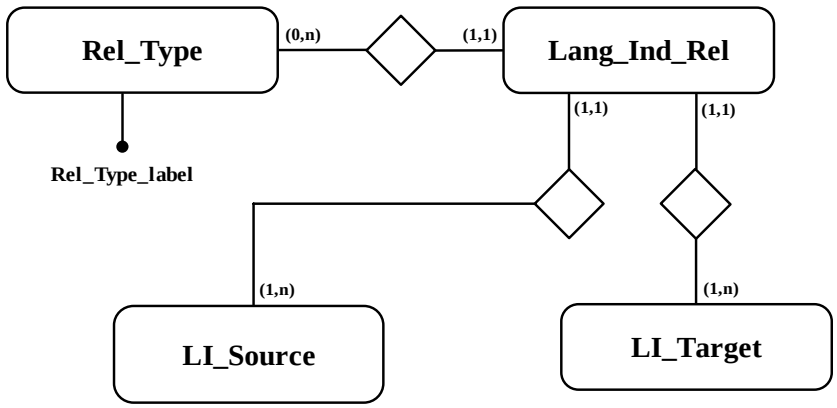


Figura 11 : Schema E/R EuroWordNet, parte IV: Language Indipendent Relation



LI_Source	LI_Target
ILI_Record	ILI_Record
ILI_Record	Top_Concept
ILI_Record	Domain
Top_Concept	Top_Concept
Top_Concept	ILI_Record
Domain	Domain
Domain	ILI_Record

Figura 12 : Schema E/R di EuroWordNet, parte IV: particolare di Language Indipendent Relation

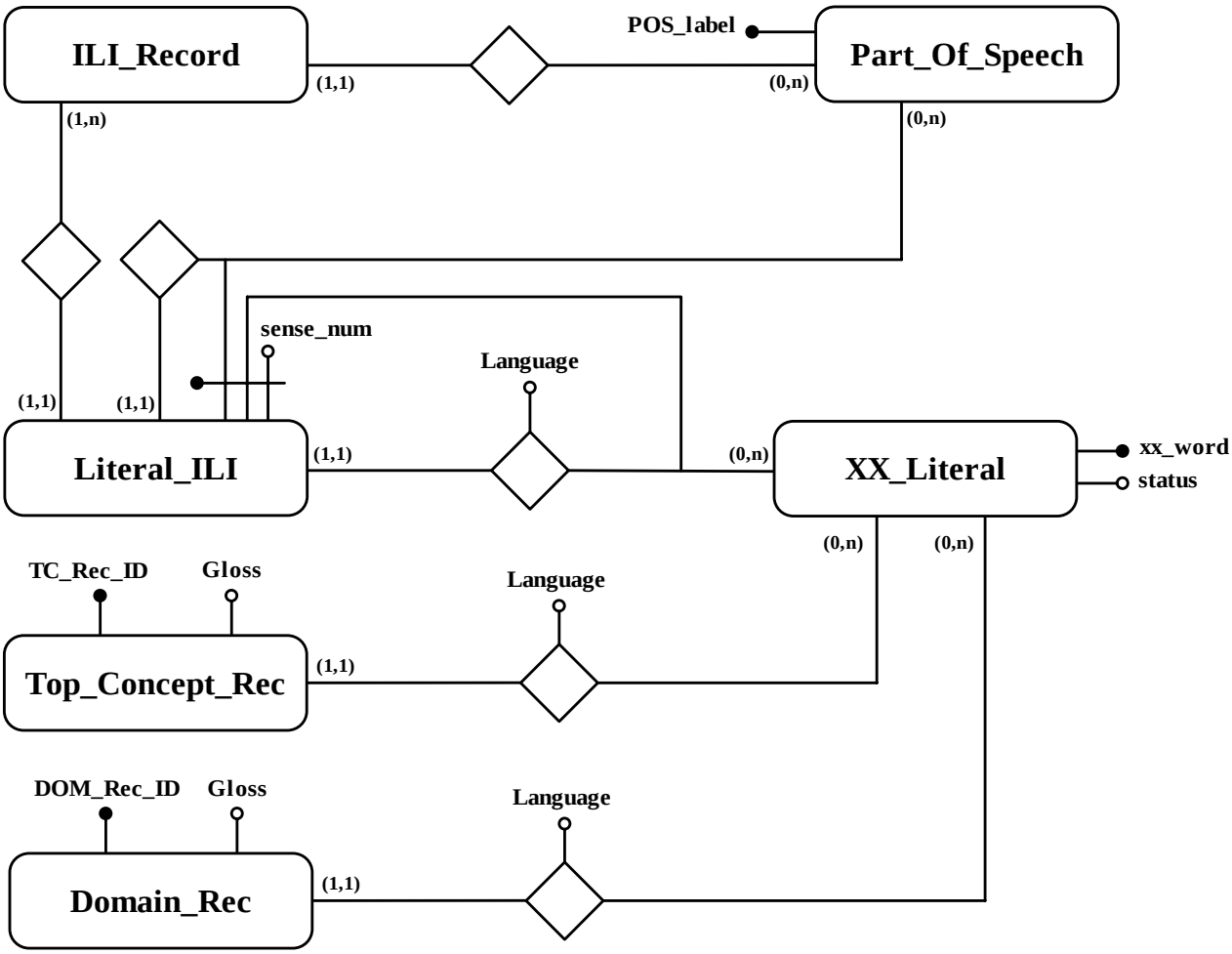


Figura 13 : Schema E/R di EuroWordNet, parte V: Literal e Language Indipendent Module

3.5 Traduzione dello schema E/R nello schema relazionale

Segue la traduzione in schema relazionale del diagramma E/R precedente. Per quanto riguarda i moduli dipendenti dalla lingua, abbiamo riportato solo le relazioni che descrivono il modulo della lingua italiana, essendo gli altri identici ad esso. Le due specializzazioni presenti sono state semplificate nel seguente modo:

- It_Word_Meaning e It_Word_Instances sono state collassate verso l'alto e raggruppate nella relazione It_Synset;
- L'entità Synset è stata separata nelle sue due sotto-entità (collasso verso il basso) e rappresentata con le relazioni Lang_Synset e WN15_Synset, la seconda delle quali potrebbe essere direttamente individuata in WordNet.

1) Tipi comuni:

```
Rel_Type (Rel_Type_label)

Part_Of_Speech (POS_label)

Lang_Synset (Lang_Syn_ID)

Wn15_Synset (Wn15_Syn_ID)
```

2) Language Indipendent Module

```
ILI_record (ILI_rec_ID, Gloss, POS_label)
      FK: Pos_label References Part_Of_Speech

ILI_Origin_Ewn (Lang_Syn_ID, ILI_rec_ID)
      FK: Lang_Syn_ID References Lang_Synset
      FK: ILI_rec_ID References ILI_record

ILI_Origin_Wn15 (Wn15_Syn_ID, ILI_rec_ID)
      FK: ILI_rec_ID References ILI_record
      FK: Wn15_Syn_ID References Wn15_Synset

Interlingual_Rel (Rel_Type_label, ILI_rec_ID, Lang_Syn_ID)
      FK: Rel_Type_label References Rel_Type
      FK: ILI_rec_ID References ILI_record
      FK: Lang_Syn_ID References Lang_Synset
```

```
Literal_ILI (ILI_rec_ID, word, language, POS_label, Sense_num)
      FK: ILI_rec_ID References ILI_record
      FK: Pos_label References Part_Of_Speech
      FK: word References Literal
```

L'attributo word si riferisce ad un Literal in una qualsiasi lingua: tale indicazione viene poi specificata dall'attributo language. Perciò tale attributo è nella chiave primaria, perché word da solo non identifica univocamente il Literal, non specificando la lingua.

Le ultime due note, su word e language valgono anche per le due relazioni seguenti:

```
Top_Concept_Rec (TC_rec_ID, Gloss, word, language)

Domain_Rec (Dom_rec_ID, Gloss, word, language)
```

```
Lang_Ind_Rel (Source_LI_rec_ID, Source_LI_Type, Target_LI_rec_ID,
Target_LI_Type, Rel_Type_label)
      FK: Rel_Type_label References Rel_Type
```

Source_LI_rec_ID e Target_LI_rec_ID si riferiscono a ILI_record, Top_Concept_Rec e Domain_Rec, secondo la tabella in figura 12: tale riferimento viene specificato con gli attributi Suorce_LI_type e Target_LI_Type.

3) Italian Module

```
It_Synset (Synset_Type, POS_label, Lang_Syn_ID)
      FK: Pos_label References Part_Of_Speech
      FK: Lang_Syn_ID References Lang_Synset
      Synset_Type = (Word_Meaning) or (Word_Instance)

It_Intralingual_Rel (Source_ItSyn_ID, Target_ItSyn_ID, Rel_Type_label)
      FK: Rel_Type_label References Rel_Type
      FK: Source_ItSyn_ID References It_Synset
      FK: Target_ItSyn_ID References It_Synset

It_Literal (It_word, status)

It_Syn_Lit (It_Syn_ID, It_word, POS_label, Sense_num)
      FK: It_Syn_ID References It_Synset
      FK: Pos_label References Part_Of_Speech
      FK: It_word References It_Literal
```

Capitolo 4

Confronto tra gli schemi E/R di MultiWordNet ed EuroWordNet

Dopo aver introdotto e descritto i database lessicali MultiWordNet ed EuroWordNet, vogliamo concludere il lavoro confrontando le due architetture. Nel fare ciò ci riferiremo agli schemi E/R ricavati in precedenza, nei capitoli 2 e 3. Nei diagrammi che sono presentati in questo capitolo, si è deciso di riportare, come esempio, solo il wordnet della lingua italiana, oltre ovviamente ai moduli per il collegamento interlinguistico.

4.1 Moduli indipendenti dalla lingua

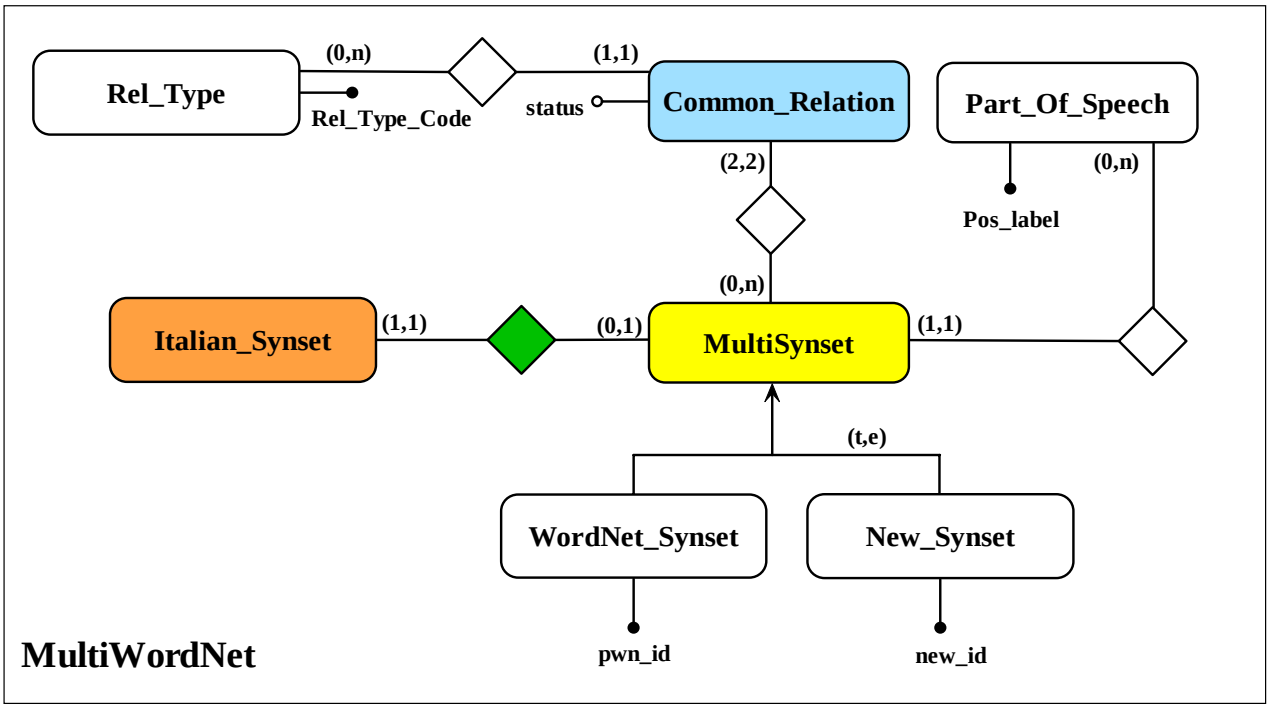
Cominciamo la nostra analisi osservando come entrambe le architetture contengano il concetto di categoria sintattica (part of speech); in EuroWordNet, però, è stata aggiunta la categoria proper noun non presente in MultiWordNet, né in WordNet. Per quanto riguarda, poi, i tipi delle relazioni, sono essenzialmente gli stessi; in EuroWordNet però esistono diversi sottotipi per le relazioni di equivalenza interlinguistica.

Passiamo ora a considerare i multisynset di MultiWordNet. Essi rappresentano due concetti:

- Il collegamento interlinguistico tra i synset dei singoli wordnet
- Il synset identifier

Come sappiamo, infatti, l'equivalenza cross-linguistica in MultiWordNet è realizzata utilizzando lo stesso synset identifier; inoltre vengono utilizzati anche gli identificatori di WordNet, per i synset da esso importati.

Questi concetti in EuroWordNet vengono espressi in modo differente. Innanzitutto il synset identifier non è condiviso, ma è univoco in tutto il database, e non è un identificatore di WordNet. Il ponte interlinguistico è rappresentato dagli ILI-record, ai quali synset specifici di ogni linguaggio vengono collegati direttamente, cioè tramite istanze dell'entità Interlingual_Rel, per permettere di rappresentare diversi tipi di relazioni.



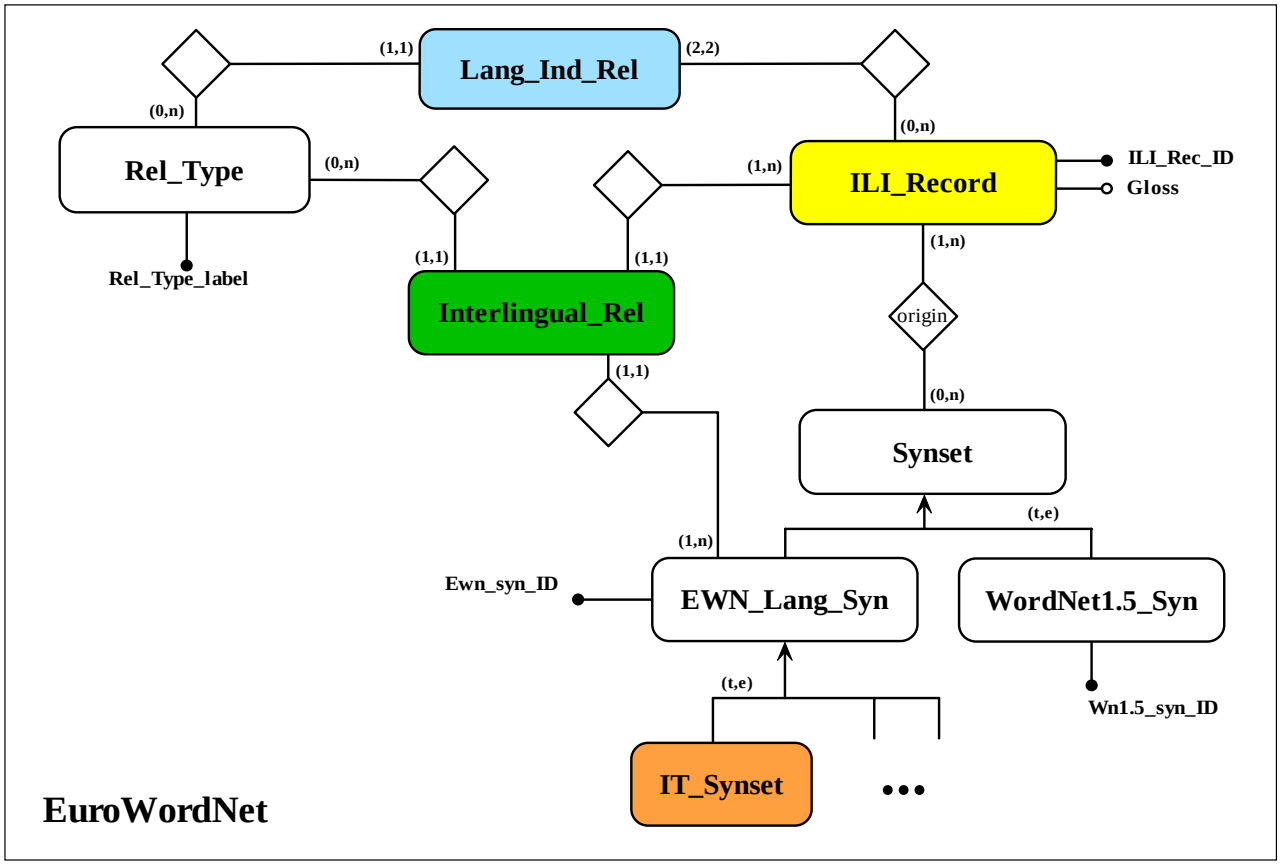


Figura 14 : Confronto degli schemi E/R, parte I: common-db

E’ difficile, perciò, trovare una corrispondenza tra le due strutture per quanto riguardo il synset identifier, data la diversità di rappresentazione. Più facile è invece interpretare il MultiSynset, il quale sicuramente viene messo in corrispondenza con l’ILI_record: nella figura sono state evidenziate con il colore giallo le due entità corrispondenti. Sempre nella figura sono state evidenziate con gli stessi colori altre parti corrispondenti dei due database: con il verde si mostra come l’associazione tra Italian_Synset e MultiSynset sia un corrispondente a Interlingual_Rel, o meglio ad un suo sottoinsieme; con l’arancione si sottolinea la corrispondenza tra i synset specifici delle lingue. Una Common_Relation, infine, corrisponde ad una Lang_Ind_Rel che collega due ILI_record, come mostrato in figura con il colore azzurro. Per quanto riguarda la relazione “origin”, infine, essa non è presente in MultiWordNet: risulta, in parte, implicita nell’uso degli identificatori di WordNet.

Non sono stati riportati nel diagramma i Top-concept e i Domain, né tantomeno le associazioni di questi e degli ILI_record con specifici lemmi, in quanto sono concetti esclusivi di EuroWordNet.

4.2 Moduli dipendenti dalla lingua

La seconda figura mostra a confronto le parti delle architetture di EuroWordNet e MultiWordNet contenenti le informazioni specifiche delle singole lingue: anche in questo caso, come esempio, si è scelto il database linguistico italiano. Come già fatto nella figura precedente, sono state messe in evidenza con i colori le parti comuni ai due diagrammi.

Partiamo considerando le entità che rappresentano i synset: si vede meglio, in questo grafico la corrispondenza tra i synset di MultiWordNet e i synset di EuroWordNet. Si noti inoltre che, a rigore, si dovrebbe limitare la corrispondenza ai soli Word-Meaning: in MultiWordNet, infatti, non è presente la categoria sintattica proper noun.

I lemmi vengono rappresentati allo stesso modo, mentre l’associazione tra lemmi e synset da una parte è una semplice associazione, mentre dall’altra è reificata, per poter esprimere il vincolo dell’identificatore esterno.

Per quanto riguarda le relazioni infine si fa notare che:

- le relazioni semantiche vengono rappresentate nella stessa maniera, tramite le entità Italian_Sem_Rel, e It_Intralingual_Rel;

- le relazioni lessicali, cioè quelle che sono valide per due singoli lemmi, sono espressamente rappresentate in MultiWordNet, mentre in EuroWordNet esse vengono modellate come relazioni intralinguistiche e poi è possibile specificare i lemmi utilizzando appositi campi opzionali.

- mancano in EuroWordNet gli identificatori delle relazioni intralinguistiche: si potrebbe anche ipotizzare che essi siano stati inseriti nella successiva implementazione del database, ma ciò andrebbe verificato.

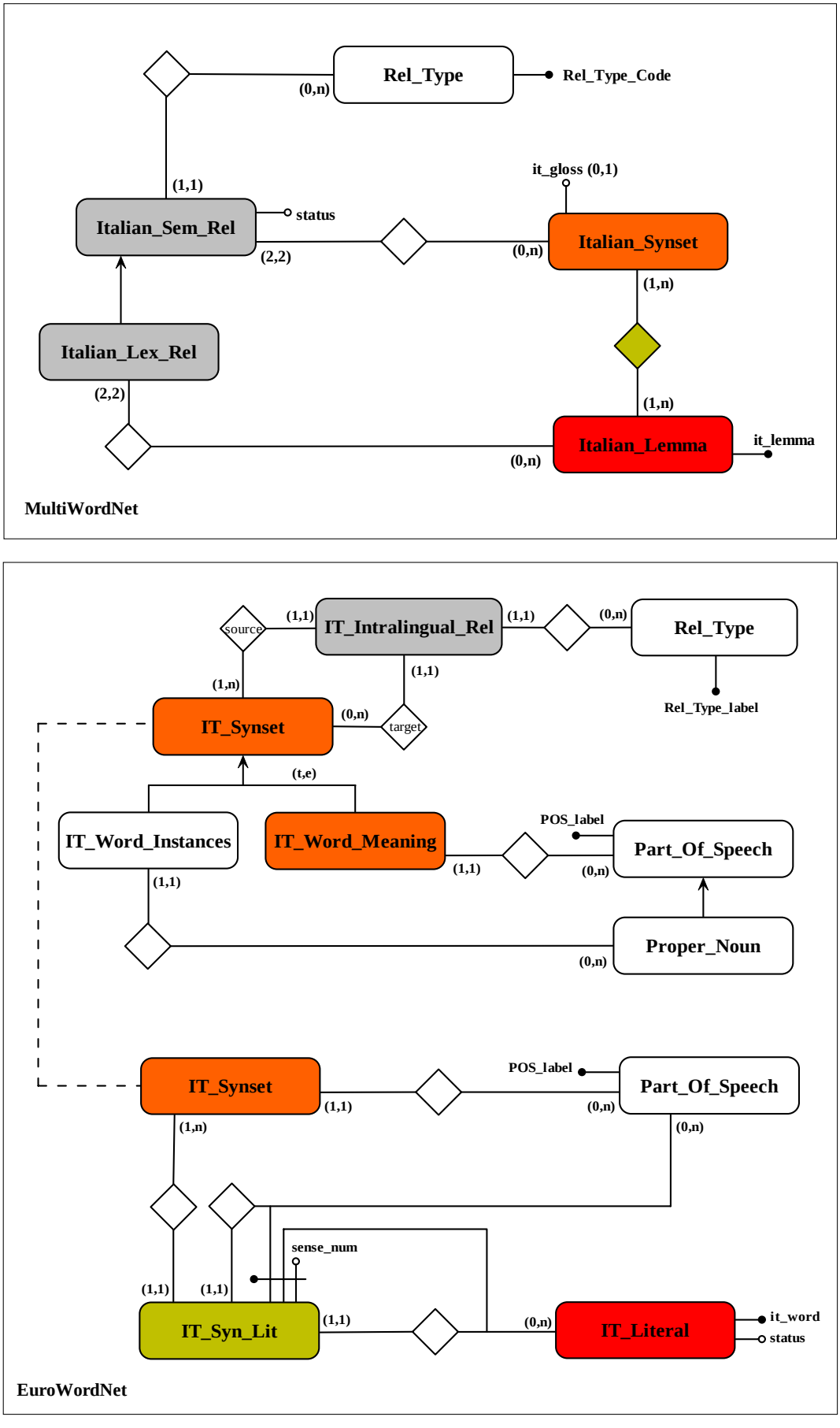


Figura 15 : Confronto degli schemi E/R, parte II: italian-db

4.3 Integrazione delle ontologie in un'architettura comune

Come scritto dell'introduzione, l'obiettivo del presente elaborato consisteva nel delineare uno schema comune in grado di ospitare ed integrare gli schemi concettuali di MultiWordNet ed EuroWordNet.

Il problema di integrazione affrontato non è stato particolarmente complicato da risolvere, poiché è stato possibile sfruttare lo schema di EuroWordNet ricavato in precedenza. Dal confronto, infatti, è emerso come la struttura di EuroWordNet sia più generale rispetto a MultiWordNet, e anzi come essa contenga tutti i concetti espressi in quest'ultimo. Non sembrano esserci vincoli, perciò, al tentativo di importare, con opportune procedure, MultiWordNet in EuroWordNet. Volendo individuare uno schema concettuale comune in grado di ospitare ed integrare le due ontologie, si è pensato di utilizzare quello di EuroWordNet, anche nella versione semplificata presentata al capitolo 3. Di questa ontologia è infatti già disponibile sia lo schema E/R che quello relazionale. Essa inoltre si presta bene all'inserimento di nuovi concetti, come d'altra parte anche quella di MultiWordNet. E' inoltre possibile, e semplice, effettuare alcuni cambiamenti sullo schema dato, ad esempio per inserire i vincoli di integrità referenziale di MultiWordNet non presenti nello schema di EuroWordNet.

Si fa notare inoltre che il processo di integrazione descritto implica un'operazione di mapping delle strutture di MultiWordNet in corrispondenti strutture di EuroWordNet. Tale operazione segue fedelmente le analogie che sono state fatte al capitolo precedente; non riportando in questo momento i dettagli ricordiamo solo che:

- i synset, i lemmi e le relazioni specifiche di ogni lingua vengono tradotte in EuroWordNet con strutture analoghe;
- il legame lessicale tra synset e lemma deve essere reificato ed aggiunto il vincolo dell'identificatore (lemma, sense number, part of speech) per il synset;
- i multisynset vengono tradotti con istanze di ILI-record; i synset di WordNet a cui è collegato il multisynset danno luogo a istanze dell'associazione origin;
- l'associazione tra synset e multisynset diventa una relazione interlinguistica di sinonimia;
- le relazioni tra multisynset in corrispondenti relazioni tra ILI-record.

Capitolo 5

Conclusioni e lavoro futuro

5.1 Note finali sulle architetture presentate

Il confronto, fatto al capitolo precedente, tra i due database mette in luce la maggiore ricchezza del modello di ontologia multilingua di EuroWordNet rispetto a MultiWordNet. Si potrebbe dire, infatti, che solo la prima è una vera e propria ontologia multilingua, che integra al suo interno ontologie monolingua e una parte nuova di relazioni interlinguistiche. Per quanto riguarda MultiWordNet, invece, esso risulta eccessivamente improntato alla struttura di WordNet (e non potrebbe essere altrimenti), per poter rappresentare propriamente un'ontologia. Inoltre l'ipotesi su cui si basa il modello è che tra gli stessi concetti, in lingue diverse, intercorrono le stesse relazioni: per quanto riguarda le lingue italiano e inglese, per le quali la sovrapposizione è sul 99% dei synset, e, in generale, per le lingue occidentali, questo può anche essere accettabile, ma risulta difficile pensare che questo modello sia estendibile anche a lingue di origine diversa.

D'altra parte, la struttura più semplice di MultiWordNet, e la minore complessità del modello implementato, si riflette in una maggiore compatibilità nella rappresentazione delle diverse lingue e facilità di ricerca e navigazione delle relazioni. Questo, insieme con la possibilità di ampliare facilmente l'ontologia con nuovi concetti, fa di MultiWordNet uno strumento lo stesso valido in quelle applicazioni che richiedano di poter effettuare ricerche interlinguistiche, senza la necessità di un'eccessiva precisione linguistica. Tra l'altro la struttura di MultiWordNet, proprio per questo suo carattere di estensione e non di integrazione multilingua, mostra una maggiore grado compatibilità con quella WordNet.

Per quanto riguarda EuroWordNet, infine, esso è un progetto molto più ambizioso, che mira a creare un'ontologia completa delle lingue europee. La sua struttura mostra presupposti migliori per l'integrazione di nuove ontologie, anche perché è stata progettata ad hoc per questo scopo. Nell'architettura di EuroWordNet sembrano essere più bilanciate le parti cosiddette language dipendent e language indipendent, consentendo perciò un discreto grado di indipendenza per le strutture della lingue rappresentate, aspetto, questo, che in MultiWordNet è poco sviluppato.

5.2 Conclusioni

Nell'ambito del presente elaborato sono state studiate da vicino le architetture dei due database lessicali MultiWordNet ed EuroWordNet. Essi rappresentano i due approcci di ricerca più importanti nell'obiettivo di estendere le ontologie linguistiche con il concetto del multilinguismo.

L'analisi è stata condotta separatamente sui due database. Si è cercato poi di ridurli ad una forma comune: allo scopo, è stato utilizzato il modello E/R che, per semplicità di notazione e potenzialità espressiva, si prestava bene al nostro scopo. I diagrammi E/R sono stati poi, per completezza, tradotti in schemi relazionali, secondo le regole di traduzione standard.

In un secondo momento, avendo a disposizione gli schemi E/R, è stato possibile mettere a confronto le due ontologie multilingua. Dal confronto è emerso come queste presentino un alto grado di compatibilità a livello di schema concettuale, in particolare si è visto come sia possibile, negli schemi E/R, ritrovare le strutture di MultiWordNet all'interno di EuroWordNet. Perciò nel passaggio successivo, quello di delineare una schema comune atto a contenere entrambi i database, si è scelto di utilizzare proprio EuroWordNet.

Ovviamente un progetto completo dovrebbe tener conto anche di problematiche quali l'eliminazione della ridondanza, ossia dei concetti presenti in entrambe le ontologie; si dovrebbe pensare anche a come risolvere potenziali contraddizioni nelle informazioni presenti nei due database. Si dovrebbero infine fare tutte le scelte relative al progetto e all'implementazione fisica del database. Questo esula dall'ambito di questa tesi, il cui obiettivo era quello di valutare la possibilità dell'integrazione e a tracciarne le linee guida essenziali.

5.4 Sviluppi futuri

Per quanto riguarda gli sviluppi futuri, il primo è già evidente nella conclusione: si tratta cioè di progettare e implementare il processo di integrazione delle ontologie lessicali di MultiWordNet ed EuroWordNet.

Sarebbe inoltre interessante valutare meglio la compatibilità tra le relazioni presenti in MultiWordNet ed EuroWordNet, in particolare per vedere se ci sono contraddizioni; si potrebbe estendere il confronto anche a WordNet. Quest'ultimo punto è particolarmente importante in previsione della futura integrazione di EuroWordNet nel sistema SEWASIE: se i risultati delle analisi confermassero le conclusioni raggiunte in questo elaborato si potrebbe anche pensare di utilizzare MultiWordNet come passaggio temporaneo tra WordNet ed EuroWordNet, trasferendo solo in un secondo momento le relazioni definite.

Bibliografia

[1] Giovanni Malvezzi. Estrazione di relazioni lessicali con WordNet nel sistema MOMIS. Tesi di laurea. Università di Modena e Reggio Emilia, Facoltà di Ingegneria, corso di laurea in Ingegneria Informatica, 1999-2000. http://sparc20.dsi.unimo.it/tesi/index.html

[2] Veronica Guidetti. Intelligent Information Integration systems: extending lexicon ontology. Tesi di laurea. Università di Modena e Reggio Emilia, Facoltà di Ingegneria, corso di laurea in Ingegneria Informatica, 2001-2002. http://sparc20.dsi.unimo.it/tesi/index.html

[3] A. G. Miller. Wordnet: A lexical database for english. Communications of the ACM, 38(11):39-41, 1995. http://www.cogsci.princeton.edu/~wn/

[4] D. Beneventano, S. Bergamaschi, S. Castano, A. Corni, R. Guidetti, G. Malvezzi, M. Melchiori and M. Vincini, Information Integration: the MOMIS Project Demonstration, Proc. Int. Conf. on Very Large Data Bases VLDB-2000 (Cairo, Egypt, 2000). http://www.dbgroup.unimo.it/pubs.html

[5] S. Bergamaschi, D. Beneventano, F. Guerra, A. Fergnani, F. Mandreoli, R. Martoglia, S. Bernardini. Specification of the general framework for the multilingual semantic enrichment processes and of the semantically enriched data stores. http://www.dbgroup.unimo.it/pubs.html

[6] Piek Vossen. (1996) Right or wrong: combining lexical resources in the EuroWordNet project. Proceedings of Euralex-96 International Congress. E' disponibile all'indirizzo http://www.illc.uva.nl/EuroWordNet/docs.html (Sezione "Papers related to EuroWordNet", P004).

[7] E. Pianta, L. Bentivogli, C. Girardi. MultiWordNet, developing an aligned multilingual database. Proceedings of the First International Conference on Global WordNet, Mysore, India, January 21-25, 2002. http://multiwordnet.itc.it/english/publications.html

[8] Vossen P. (ed.). EuroWordNet, General Document. E' disponibile all'indirizzo http://www.illc.uva.nl/EuroWordNet/docs.html (Deliverable D032D033/2D014).

[9] P. Diez-Orzas, P. Forest, M. Louw. High-level structure of the EuroWordNet database. A Novell ConceptNet-based semantic network. (1996). E' disponibile all'indirizzo http://www.illc.uva.nl/EuroWordNet/docs.html (Deliverable D007).

[10] Nicola Guarino. Formal ontologies and information systems. In Proceedings of the International Conference on Formal Ontology in Information Systems (FOIS'98), Trento, Italy, june 1998.

[11] Alain Fergnani. Ontology dinamics for semantic web: the MOMIS approach. Tesi di laurea. Università di Modena e Reggio Emilia, Facoltà di Ingegneria, corso di laurea in Ingegneria Informatica, 2001-2002. http://sparc20.dsi.unimo.it/tesi/index.html

[12] T. R. Gruber. A translation approach to portable ontology specifications. In Knowledge Acquisition, volume 5, pages 199-220, 1993.