La costruzione di una base di conoscenza lessicale per la lingua latina: LatinWordnet

Il progetto di costruzione di una base di conoscenza lessicale per la lingua latina di cui qui si tratta nacque, alla fine del 2004, dalla constatazione che non erano ancora disponibili rappresentazioni elettroniche della conoscenza semantica di quella lingua: la creazione di una simile base di conoscenza apre oggi alla possibilità di impiegare tecniche di analisi proprie del *Natural Language Processing* sui testi latini.
La necessità di fornire al calcolatore dati in grado di rendere elaborabile il testo elettronico nella sua dimensione semantica discende dai principi di funzionamento stessi della macchina e dalle peculiarità dell'oggetto testo.
Il testo elettronico, infatti, presenta una sua materialità nella codifica in impulsi che permette di renderlo memorizzabile nell'elaboratore e, allo stesso tempo, può essere rappresentato graficamente sullo schermo del computer perché continui ad essere leggibile al lettore umano, al quale la codifica binaria rimane nascosta. Tuttavia, la memorizzazione di un testo nel computer non è altro che un processo di transcodifica che sostituisce i significanti con altri convenzionalmente equivalenti per una necessità causata da un cambiamento del mezzo. L'aver memorizzato un testo in un elaboratore elettronico non fa dell'elaboratore il destinatario del messaggio, né più né meno di quanto lo sia il foglio su cui scriviamo con penna ed inchiostro. Il testo è quindi immagazzinato, ma non elaborabile.
La scrittura è un atto grafico, l'atto finale di un processo che, a partire dal senso, codifica quest'ultimo attraverso la sintassi e il lessico di un linguaggio naturale e ne dà una convenzionale rappresentazione nell'ordinamento lineare di una sequenza di *glifi*. L'ente unificatore che permette di stabilire l'equivalenza semantica di ciascun codice è il soggetto-lettore, interprete in grado di superare attraverso la bontà delle sue inferenze l'ambiguità delle codifiche linguistiche e grafiche.
Inferenza e significazione sono processi umani che nella *semiotica cognitiva* di Peirce, come essa viene definita da Massimo Bonfantini, vengono accomunati:

> Un segno, o *representamen*, è qualcosa che sta a qualcuno per qualcosa sotto qualche rispetto o capacità […] Definisco un *Segno* come qualcosa che da un lato è determinato da un *oggetto* e dall'altro determina un'idea nella mente di una persona, in modo tale che quest'ultima determinazione, che io chiamo *interpretante* del segno, è con ciò stesso mediatamente determinata da quell'*oggetto*[1]

La funzione *interpretante* è un fatto puramente umano ed è bene per questo distinguere tra interpretazione di un testo e processi di elaborazione elettronica di un testo: il testo rappresentato a video, infatti, continua ad essere oggetto di inferenza e di interpretazione per il lettore umano, né più né meno del testo cartaceo, ma il testo come insieme di dati è un oggetto elaborabile da parte del *computer* soltanto nella dimensione del significante.
In altre parole il formato dati *testo* non fornisce una rappresentazione del messaggio adeguata all'elaborazione automatica del contenuto, né il *computer* ha possibilità di realizzare una semiotica intesa con Hjelmslev come individuazione dei rapporti che intercorrono tra espressione e contenuto.
Dunque per poter rendere elaborabili alcune delle caratteristiche del testo che, pur essendo evidenti al lettore umano, non possono essere immediatamente oggetto di elaborazione elettronica è necessario operare una ulteriore transcodifica che attraverso l'aggiunta di segni possa consentire una diversa rappresentazione del contenuto la cui espressione sia sottoponibile al processo di elaborazione, per mezzo di formalismi logici e matematici che modellizzino una teoria del significato. Si tratta in altre parole di realizzare, all'interno dello spazio testuale, una sorta di segnaletica che permetta la creazione di una mappa per l'elaborazione di elementi e di dimensioni

[1] C. S. PEIRCE, *Semiotica*, Torino, Einaudi, 1980, p. 132 e p.194

del testo che vanno oltre la reticente sequenza simbolica e di individuare algoritmi per il trattamento di questi segni secondo un ben preciso modello semiotico.
Il testo elettronico, quindi, per poter essere in qualche modo elaborato dal calcolatore nella dimensione del significato deve essere "mappato" o come nel caso del *markup* di tipo *strongly embedded* deve contenere segni che conducano il tracciamento di una mappa: la logica e la semantica modale[2] elaborano una teoria dei *mondi possibili* che mostra la facoltà del testo di poter essere rappresentato in un altro testo, sulla base di atteggiamenti proposizionali di un terzo interpretante.
Le operazioni di marcatura semantica che rendono possibile la simulazione della competenza interpretativo-referenziale nel calcolatore sono possibili solo a partire dalla presenza di strumenti di gestione e catalogazione delle informazioni lessicali che siano in grado di rappresentare la versione elettronica della memoria semantica umana.
La ricerca atta a costruire basi di conoscenza lessicale in grado di fornire una riproduzione della complessità delle relazioni tra le componenti del linguaggio umano ha portato alla creazione di strutture di dati complesse denominate "reti semantiche".
Le reti semantiche sono basate sull'idea che la conoscenza possa essere rappresentata attraverso concetti correlati per mezzo di varie relazioni. Un *network* semantico, dunque, pone l'accento sulla creazione di una struttura indipendente dall'espressione e che possa modellizzare i rapporti a livello di contenuto: una rete semantica è quindi composta da un insieme di nodi e archi. Gli archi sono etichettati in base al tipo di relazioni che rappresentano; i dati di fatto relativi ad un determinato nodo, come le sue caratteristiche (colore, dimensione ecc.), sono spesso inseriti in una struttura di dati chiamata cornice (ingl. *frame*). Ciascuna voce di un frame è chiamata zoccolo[3] (ingl. *slot*). Il *frame* di una rosa può essere, per esempio, così schematizzato:

(**rosa**
(*ha-colore* rosso)
(*altezza* 60 cm)
(*è-un* fiore)
)

In questo caso il *frame rosa* è un singolo nodo di una rete semantica che mostra una relazione IS-A (è-un) con il nodo *fiore*. Gli *slot ha-colore* e *altezza* contengono proprietà individuali della rosa[4].
Fino ad ora sono stati sviluppati numerosi sistemi per la comprensione del linguaggio naturale e per la costruzione automatica di network semantici per la rappresentazione della conoscenza presente in un testo[5]. I problemi più frequenti sono legati alla rappresentazione degli argomenti riguardanti lo spazio o il tempo: per esempio risulta difficile immagazzinare le informazioni presenti in una frase come "lunedì scorso la rosa è cresciuta di trenta centimetri ed è diventata più alta di tutto nel

[2] A tale proposito: S. KRIPKE, *Naming and Necessity*, Oxford, Basil Blackwell, 1980.
[3] Cfr. M. MINSKY, "A Framework for Representing Knowledge", In *The Psychology of Computer Vision*, cur. Winston, 211--277, McGraw Hill, New York, 1975.
[4] In questo contesto è d'obbligo il richiamo alla teoria dei *database* relazionali che utilizzano una modellizzazione molto simile per descrivere i dati trattati. Le reti semantiche specializzano il modello relazionale orientandolo alla rappresentazione dei sistemi linguistici.
[5] Per una panoramica: R. C. SCHANK, *Conceptual Information Processing*, New York, Elsevier Science Inc., 1975; R. C. SCHANK, and W. LEHNERT, "Computer Understanding of Stories", In *Human and Artificial Intelligence*, cur. Klix, 135-139, North-Holland, Amsterdam, 1979; F. GOMEZ, and C. SEGAMI, *The Recognition and Classification of Concepts in Understanding Scientific Texts*, «Journal of Experimental Theoretical Artificial Intelligence», 1989, 1, 1, pp 51--77; F. GOMEZ, and C. SEGAMI, *Classification-Based Reasoning*, «IEEE Transactions on Systems, Man and Cybernetics», 1991, 21, 3, pp 644--659.

giardino". Le informazioni che riguardano una caratteristica presente in tempi diversi o la posizione relativa di un oggetto sono difficilmente registrabili in una rete semantica[6].
Uno degli esempi più completi di rete semantica è costituito da WordNet[7], un sistema disponibile pubblicamente, che contiene *frame* specificamente orientati alla rappresentazione delle parole: a partire dal riconoscimento della natura del tutto accidentale dell'ordinamento dei dizionari attraverso *spelling*, nel modello di WordNet le parole sono organizzate per blocchi di significato, denominati *synset*, che raccolgono tutti i lemmi che lessicalizzano lo stesso concetto; i *synset* sono collegati tra loro per mezzo di relazioni che includono, assieme alla sinonimia, anche l'iponimia, la meronimia e l'antinomia. L 'ipo-iperonimia mette in relazione significati subordinati e superordinati fornendo così una struttura gerarchica di concetti. La relazione meronimica induce una gerarchia delle parti sull'insieme dei significati. In questo modo il livello lessicale è chiaramente separato da quello concettuale e questa distinzione è rappresentata dal *medium* semantico-concettuale e dalla relazione semantica che uniscono rispettivamente *synset* e parole. Le relazioni presenti tra i verbi permettono di mettere in luce relazioni di *implicazione* (ingl. *entailment*) e di troponimia. Due verbi sono correlati dall'*implicazione* nel momento in cui il primo verbo implichi il secondo: per esempio la coppia comprare-pagare. La troponimia è la relazione presente nel momento in cui due attività collegate da implicazione avvengono allo stesso tempo: un esempio è la coppia zoppicare-camminare.

Il progetto di costruzione di una rete semantica per il latino ha avuto come basi di partenza due modelli: WordNet e MultiWordNet[8], progetto sviluppato dall'Istituto Trentino di Cultura[9] inteso a realizzare una rete semantica multilingue.

Il modello adottato dal progetto MultiWordNet (MWN) consiste nel costruire le reti semantiche specifiche per un linguaggio mantenendo il più possibile le relazioni semantiche disponibili nella WordNet di Princeton (PWN). Ciò è ottenibile costruendo i nuovi *synset* in corrispondenza dei *sysnset* della PWN, ogni volta che ciò sia fattibile, e importando le relazioni semantiche dai corrispondenti *synset* inglesi; in questo modo si ipotizza che se esistono due *synset* nella PWN e una relazione che li collega, la stessa relazione leghi i corrispondenti *synset* in una lingua diversa. Secondo Vossen[10], il modello di MWN (o "modello a espansione", *expand model*) garantisce un elevato grado di compatibilità tra differenti *wordnet*. Per constatare questo fatto basta considerare che la costruzione di qualsiasi rete semantica necessariamente implica un gran numero di decisioni soggettive (e discutibili). Così se due reti semantiche sono costruite indipendentemente per due diverse lingue, mostreranno differenze che dipendono solo parzialmente dalle differenze tra le due lingue: alcune non banali discrepanze strutturali dipenderanno, infatti, da scelte soggettive o da criteri di costruzione differenti. Il modello di MWN minimizza queste differenze aderendo strettamente ai modelli di costruzione di PWN.

Il modello MWN presenta anche degli inconvenienti potenziali: il rischio più serio è quello di forzare una eccessiva dipendenza sulla struttura lessicale e concettuale di uno dei linguaggi coinvolti[11]. Questo rischio può essere scongiurato permettendo alla nuova rete semantica di divergere, quando necessario, dalla struttura di PWN.

---

[6] Su tale problema è centrata un interessante disamina intitolata "Representational Thorns" in D. B. LENAT, and R. V. GUHA, *Building Large Knowledge-Based Systems; Representation and Inference in the Cyc Project*, Boston, Addison-Wesley, 1989.

[7] Cfr. G. A. MILLER, R. BECKWITH, C. FELLBAUM, D. GROSS, K. J. MILLER, *Introduction to Wordnet: An on-Line Lexical Database*, «International Journal of Lexicography», 1990, 3, 4, pp 235--244; C. FELLBAUM (cur.), *Wordnet: An Electronic Lexical Database (Language, Speech, and Communication)*, Cambridge, MA, USA, MIT Press, 1998.

[8] Cfr. EMANUELE PIANTA, LUISA BENTIVOGLI, CHRISTIAN GIRARDI, "MultiWordNet: Developing and Aligned Multilingual Database",In *Proceedings of the First International Conference on Global WordNet*, Mysore, India, January 21-25, 2002, pp. 293-302.

[9] Oggi "Fondazione Bruno Kessler".

[10] Cfr. P. VOSSEN, "Right or Wrong: Combining Lexical Resources in the Eurowordnet Project", In *Proceedings of Euralex-96*, ed. Gellerstam, Jarborg, Malmgren, Noren, Rogstrom and Papmehl, 715--728, Goetheborg, 1996.

[11] Cfr P. VOSSEN, cit., p. 718.

Un altro importante vantaggio del modello MWN è che possono essere utilizzate procedure automatiche per velocizzare la costruzione dei *synset* corrispondenti e per l'individuazione delle divergenze tra PWN e la rete semantica che si sta costruendo. In tutte queste procedure la stessa PWN può essere usata utilmente come risorsa.

La costruzione di LWN (LatinWordNet) si è basata principalmente su una procedura automatica per l'assegnazione.

Seguendo il modello MWN, il nostro obiettivo è quello di costruire, ogniqualvolta sia possibile, un *synset* latino che sia sinonimo (semanticamente corrispondente) con i *synset* di PWN. Se ciò non è possibile, si è individuata una idiosincrasia Inglese-a-Latino o Latino-a-Inglese[12].

I *synset* sinonimi latini possono essere costruiti seguendo tre differenti strategie:

- La prima strategia è basata sui traducenti dall'inglese al latino. Per ciascun *synset* di PWN *S*, cerchiamo un gruppo di traducenti che siano i sinonimi delle parole inglesi di *S*. Se non è possibile costruire nessun *synset* sinonimo latino di *S* si è trovata una idiosincrasia lessicale inglese-a-latino.
- La seconda strategia è basata sui gruppi di traducenti Latino-a-Inglese. Per ciascun senso σ di una parola latina *L*, si cerca un *synset* di PWN che includa almeno un traducente inglese di *L* e si costituisce un legame tra *L* e *S*. Quando la procedura è stata applicata a tutti i significati della parola latina, possiamo costruire la classe di equivalenza di tutti i gruppi di parole latine che sono state collegate con lo stesso *synset* di PWN. Ciascun gruppo nella classe di equivalenza è il *synset* latino sinonimo con alcuni *synset* di PWN. Se per un gruppo di sinonimi latini non c'è alcun *synset* sinonimo in PWN, si è trovata una idiosincrasia lessicale Latino-a-Inglese.
- La terza strategia sfrutta la natura multilingue di MWN e i due dizionari di macchina latino-inglese e latino-italiano: attraverso di essa le parole latine che risultano avere come traduzione parole inglesi e parole italiane che sono contrassegnate dallo stesso identificativo di *synset* vengono attribuite al medesimo *synset* della rete LatinWordnet con lo stesso identificativo. Esse infatti presentano con certezza la lessicalizzazione latina del concetto espresso dai traducenti nelle due lingue moderne[13].

Il miglior allineamento tra la WordNet di Princeton e quella latina può essere ottenuto utilizzando entrambe le strategie per cercare di validare i risultati incrociandoli.

Trovare collegamenti tra i significati delle parole latine e i *synset* di PWN è un processo complesso e lungo, anche se è sempre molto più rapido rispetto alla costruzione da zero dei *synset* latini, della loro organizzazione in una rete semantica e del metterli in corrispondenza con i *synset* di PWN. Per ciascun significato latino, il lessicografo dovrebbe cercare i gruppi di traducenti equivalenti in un dizionario bilingue, trovare tutti i *synset* che contengono questi traducenti equivalenti, valutare con attenzione il significato di questi *synset* (sinonimi, glosse, relazioni semantiche) e, infine, decidere quale *synset* di PWN, se esiste, è sinonimo del significato latino della parola. Per alcuni significati di parola il lessicografo potrebbe dover valutare decine di *synset* di PWN.

Per aiutare il lessicografo nel suo lavoro è stata realizzata una procedura che sceglie, per ciascun significato di una parola latina, i *synset* di PWN che più probabilmente hanno un significato

---

[12] Per una trattazione estesa relativa al problema dei *lexical gap* si vedano: L. BENTIVOGLI, and E. PIANTA, "Looking for Lexical Gaps", In *Proceedings of the ninth EURALEX International Congress*, ed. Heid, Evert, Lehmann and Rohrer, 663--669, Stuttgart, 2000; L. BENTIVOGLI, E. PIANTA, and F. PIANESI, "Coping with Lexical Gaps When Building Aligned Multilingual Wordnets", In *Proceedings of LREC-2000, 2nd International Conference on Language Resources and Evaluation*, ed. Gavrilidou, Crayannis, Markantonatu, Piperidis and Stainhaouer, 993--997, Atene, 2000.

[13] In altre parole una parola data una parola latina sarà attribuita a quei *synset* che costituiscono l'intersezione dei gruppi di *synset* individuati dai rispettivi gruppi di traducenti: vale a dire quei *synset* ai quali rimandano sia i traducenti italiani, sia i traducenti latini.

compatibile. Nel miglior caso la procedura sceglie solamente il candidato corretto, e il lessicografo deve solamente confermare la selezione. Nel peggiore dei casi la procedura trova solo candidati erronei o non può trovare alcun candidato e il lessicografo deve operare manualmente. Nella maggior parte dei casi la procedura trova una rosa ristretta di candidati che includono quello corretto, e il lessicografo deve confermare la scelta corretta e rifiutare quelle errate. In altre parole l'algoritmo aiuta il lessicografo a focalizzare quale sia il *synset* di PWN più adatto.

La procedura-assegnazione prende come *input* uno dei sensi della sezione Latino-a-Inglese del dizionario di macchina e fornisce in *output* un gruppo di candidati, ciascuno dei quali è descritto da una coppia del tipo , dove *punteggio di certezza* (PC) misura il grado di certezza nel legame tra il significato della parola latina e il *synset* di PWN. Solo i candidati con un PC più alto di una certa soglia vengono proposti al lessicografo. Scegliere il livello di soglia è una questione di bilanciare precisione e richiamo (vedi capitolo relativo all'information retrieval in generale). Maggiore è la soglia, minore è la probabilità che candidati erronei siano proposti (alta precisione), ma è anche maggiore la possibilità che la scelta più idonea non sia inclusa nel gruppo dei candidati (basso richiamo).

Per un determinato significato di parola listato nel dizionario latino-inglese, la procedura-assegnazione considera il gruppo di parole inglesi che vengono proposte come traducenti equivalenti per quel significato e trova tutti i *synset* contenenti almeno un traducente equivalente. Questi *synset* costituiscono il gruppo di candidati (GCand) che deve essere collegato con il significato di parola latina dell'*input*. Possiamo riassumere il primo passo dell'algoritmo dicendo che esso calcola i GCand del significato di una determinata parola latina. Il resto dell'algoritmo consiste nell'ordinare i GCand calcolando il PC di ciascuno dei suoi *synset*.

L'ordinamento dei GCand è basato su una serie di regole per stabilire i legami: ogni regola, se applicata con successo a un candidato, alza il suo PC. Si deve notare che il *PC parziale*, contribuito da ciascuna regola, varia a seconda di fattori specifici alla regola. Accanto al dizionario di macchina, vengono utilizzate dalle regole anche altre risorse, come la sezione italiana di MultiWordNet e un dizionario italiano-latino, un dizionario dei sinonimi latini e la stessa PWN.

Le strategie individuate per la costruzione dei legami sono quattro:

- probabilità generica: la regola di probabilità generica si basa sulla supposizione che solo un elemento nel GCand è il corretto candidato per legare il senso di una parola latina. Di conseguenza si può supporre che maggiore è la cardinalità del GCand, minore è la probabilità che ciascun candidato sia quello esatto. La cardinalità del GCand dipende dal grado di ambiguità delle parole che sono proposte come traducenti equivalenti del significato della parola di *input*. Se c'è un solo *synset* nel GCand, ciò significa che tutti i traducenti equivalenti della parola di input sono monosemici: è quindi altamente probabile che l'unico *synset* nel GCand sia sinonimo del significato della parola di *input*[14].
- traduzione incrociata: questa regola si basa sulla supposizione che se colleghiamo un significato di parola al corretto *synset* attraverso un traducente equivalente, è probabile che almeno alcuni dei sinonimi del traducente, presenti PWN, abbiano la parola di input come traducente equivalente inglese-latino. Si prenda ad esempio il latino *punctum, i*: quando riferito a insetti, si traduce come "sting". "Sting", però, appartiene a 4 *synset* di PWN: *sting*, *stinging*; *pang*, *sting*; *sting*, *bite*, *insect bite*; *bunco*, *bunco game*, *sting*. Solo il terzo *synset* è sinonimo della parola latina. Se guardiamo ai sinonimi di *sting* nel terzo *synset* possiamo trovare che la sezione inglese-latino dà *punctum* come traduzione di *bite*. Riassumendo, la regola della

[14] Cfr. il criterio monosemico usato in J. ATSERIAS, S. CLIMENT, X. FARRERES, G. RIGAU, and H. RODRÍGUEZ, "Combining Multiple Methods for the Automatic Construction of Multilingual Wordnets", In *Recent Advances in Natural Language Processing Ii. Selected Papers from Ranlp '97*, cur. Nicolov and Mitkov, 327--340, John Benjamins, Amsterdam & Philadelphia, 1997.

.

traduzione incrociata considera i sinonimi presenti in PWN di un traducente che crea il collegamento e calcola un PC parziale che è proporzionale al numero di sinonimi che hanno la parola italiana come traducente dall'inglese al latino.

- corrispondenza della glossa: un gruppo di regole di collegamento sfrutta le informazioni contenute nella glossa inglese che introduce la maggior parte del dizionario di macchina inglese-latino. La glossa può contenere un campo semantico specifico, un sinonimo, un iperonimo, o una specificazione di contesto d'uso. Queste informazioni possono essere utilizzate in vario modo.

L'informazione relativa al campo semantico è sfruttata grazie ad una risorsa sviluppata parallelamente a MWN, cioè la marcatura di tutti i *synset* di PWN con una etichetta relativa al campo semantico[15]. La glossa del dizionario contiene una etichetta relativa al campo semantico e se questa etichetta corrisponde a un *synset* individuato come candidato, allora il candidato ottiene un maggiore PC. Le varianti nelle etichette dei campi semantici sono gestite attraverso un tabella di corrispondenze.

Quando le glosse contengono parole o frasi, si cerca un corrispondente tra di esse e le parole contenute nelle glosse di PWN. Per fare ciò, si estraggono i lemmi delle parole inglesi delle glosse, e si controlla la loro presenza nelle glosse del traducente equivalente in PWN. La forza della corrispondenza dipende dal grado di ambiguità del traducente. Maggiore è la polisemia, minore è il peso attribuito alla corrispondenza.

Il meccanismo ha due estensioni basate sul fatto che le glosse spesso specificano il genere della parola che stanno definendo al posto di un sinonimo. La prima estensione cerca una corrispondenza tra una parola latina e un iperonimo del suo traducente equivalente. Il secondo meccanismo cerca una corrispondenza tra una parola latina e una parola inglese contenuta nella glossa di un iperonimo del *synset* candidato. Se la corrispondenza tra la parola latina e la parola inglese viene ottenuta attraverso uno dei meccanismi indiretti il PC parziale sarà più basso rispetto all'individuazione diretta.

- Intersezione di *synset:* questa regola sfrutta il fatto che i gruppi di traduzione possono includere più traducenti equivalenti, che sono ovviamente sinonimi. Se uno dei traducenti equivalenti è ambiguo, possiamo usare gli altri traducenti equivalenti per disambiguare. In pratica, la regola prende i differenti gruppi di candidati che sono accessibili attraverso diversi traducenti equivalenti e li interseca. I *synset* che sono nell'intersezione ottengono un PC. Per esempio la parola latina *pila* è tradotta nel suo senso metaforico come "pillar, mainstay". La parola pillar appartiene a 5 *synset* di PWN, mentre *mainstay* appartiene a tre *synset*. C'è però un solo *synset* che li contiene entrambi.

Una volta terminata l'assegnazione si ottiene una struttura reticolare che fornisce una modelizzazione dei rapporti semantici tra le parole: il modello di stoccaggio dei dati in MultiWordNet riflette i principali elementi teorici della rete semantica multilingue. Il *database* è costruito sull'idea che esista un gruppo di dati comuni a tutte le lingue e altri specifici di ciascuna lingua. Nell'implementazione le relazioni *semantiche* di PWN sono contenute in un modulo chiamato COMMON-DB, mentre le relazioni *lessicali* per il latino e per l'inglese sono conservate in altri due moduli LATIN-DB e ENGLISH-DB. In altre parole l'informazione relativa a quali lemmi appartengano ai *synset* si trova nei *database* delle lingue, mentre l'informazione relativa alle relazioni tra i synset, che rimangono costanti tra le lingue, è immagazzinata nel COMMON-DB. La corrispondenza tra i *synset* realizzati nelle diverse lingue si ottiene utilizzando sempre lo stesso codice identificatore: i synset di lingue diverse che hanno lo stesso codice di identificazione appartengono al medesimo *multisynset*. Il COMMON-DB descrive le relazioni tra i multisynset di

[15] Cfr. B. MAGNINI e G. CAVAGLIÀ, "Integrating Subject Field Codes into Wordnet", In *Proceedings of LREC-2000, 2nd International Conference on Language Resources and Evaluation*, cur. Gavrilidou, Crayannis, Markantonatu, Piperidis e Stainhaouer, Atene, 2000, pp. 1413--1418.

MWN. Quindi, tutte le informazioni semantiche che sono indipendenti dalla lingua possono essere aggiunte al COMMON-DB[16].

Sì è mostrato come il modello di dati di MWN rappresenti le costanti concettuali presenti in lingue differenti. Tale modello di dati, inoltre, evidenzia anche le divergenze semantiche tra le lingue[17]. Inoltre, anche se si mantengono le relazioni semantiche evidenziate da PWN come base del COMMON-DB, è possibile aggiungere nuove relazioni o modificare quelle esistenti. La possibilità di modificare le relazioni semantiche di PWN e di rappresentare le idiosincrasie concettuali nei linguaggi specifici è stata implementata attraverso dei moduli aggiuntivi che sovrascrivono, senza modificarli fisicamente, i dati originali di PWN. Il COMMON-DB infatti contiene tutte le relazioni semantiche originali di PWN e una risorsa chiamata COMMON-ADD-ON che ne riscrive un parte. Ciascuna lingua contiene un language-ADD-ON che specifica le relazioni semantiche che sono proprie di quella lingua.

Le peculiarità lessicali sono codificate all'interno delle aggiunte specifiche di ciascuna lingua. Se c'è prova che la lessicalizzazione di un determinato concetto manchi in una lingua, nella sezione lessicale del *database* di quella lingua viene inserita una etichetta vuota per quel nodo[18]. Per la rappresentazione delle differenze denotative e dei gap lessicali, vengono seguite due diverse strategie: se il nodo vuoto corrisponde a una differenza denotativa, una o più relazioni vicine vengono usate per collegare il nodo ad un *synset* più generico o a molti synset più specifici. Se il nodo vuoto corrisponde a un gap lessicale, viene riportata nella glossa del nodo vuoto una parafrasi di traduzione appropriata, preceduta dalla parola chiave *TE* (*Translating Equivalent*). Le relazioni più vicine vengono inserite nella risorsa linguistica aggiuntiva specifica della lingua in questione.

Ciascun *database* linguistico contiene anche un modulo con informazioni lessicografiche relative ai collegamenti tra i sensi delle parole e i *synset*.

Per quel che riguarda le relazioni, tutte quelle semantiche sono state importate da PWN e sono disponibili assieme alle relazioni più vicine, cioè le nuove relazioni specifiche di ciascuna lingua che sono state aggiunte nella MWN per rappresentare le differenze denotative.

L'attuale implementazione della parte latina di MWN si basa sull'aggiunta di un modulo[19] in grado di rendere indipendente il livello grafico/ortografico dall'individuazione dei lemmi. In pratica per ciascun lemma di dizionario è stata introdotta una grafia normalizzata, associata ad un numero espandibile di grafie alternative. Nei *synset* della parte latina non vengono registrati direttamente i lemmi, ma dei codici identificativi: in questo modo possono essere utilizzate diverse grafie per la rappresentazione dello stesso lemma, e sono inoltre collegate all'interno del *synset* anche tutte le realizzazioni morfologiche della flessione dei lemmi. L'implementazione della base dati è stata effettuata attraverso un *database* relazionale, in modo da permettere l'interfacciamento con il sistema di IR in maniera versatile, sfruttando le possibilità di consultazione anche attraverso un ambiente distribuito.

La consistenza della base dati è di 9378 lemmi collocati in 8973 *synset* con 143701 archi di relazione: la copertura lessicale e i risultati dell'assegnazione automatica sono in fase di valutazione e di controllo. Lo strumento è consultabile attraverso il sito della Fondazione Bruno Kessler[20] che ha sviluppato e messo a disposizione l'interfaccia per effettuare il *browsing* della rete semantica latina contestualmente a quelle realizzate per altre lingue.

Stefano Minozzi

---

[16] In particolare le relazioni relative ai campi semantici.
[17] Nella fattispecie i *gap* lessicali.
[18] Il termine "nodo" è usato in quanto MWN si compone di una struttura reticolare, dove i lemmi sono inseriti come nodi e le relazioni semantiche costituiscono i collegamenti tra i nodi.
[19] Non disponibile online.
[20] http://multiwordnet.itc.it/online/multiwordnet.php