Generazione di una rete semantica da un dizionario. Criteri e procedure

Daniele Christen

ABSTRACT

Questo articolo descrive dei criteri e delle procedure volte ad estrarre in maniera automatica delle informazioni semantiche da un dizionario monolingue in formato elettronico. Questa operazione è parte di un progetto il cui obiettivo è la creazione di una base di dati lessicale strutturata su più livelli, morfologico, sintattico e semantico.
Il trattamento informatico del dizionario sorgente ha inoltre prodotto delle informazioni sull'organizzazione del dizionario stesso, sulle caratteristiche stilistiche dei compilatori e sulle occorrenze delle parole usate nelle definizioni. Si tratta di dati che, anche in termini statistici, potrebbero essere oggetto di un'analisi a parte, di interesse lessicografico, in quanto consentono di rilevare quali termini sono maggiormente usati nelle definizioni, venendo a rivestire un ruolo vicino a delle entità semantiche *primitive*. Inoltre, i dati inerenti alle definizioni circolari e ai termini di una definizione dei quali viene a mancare la rispettiva definizione nello stesso dizionario potrebbero suggerire l'adozione di nuovi criteri di razionalità lessicografica.
Dopo discusso brevemente il rapporto fra dizionario e rete semantica, l'articolo mostra i criteri e le procedure adottate per la traduzione in una rappresentazione formale del contenuto semantico delle definizioni del dizionario. Sono esaminate sia la nozione di "primitivo semantico" e alcune delle più diffuse tipologie di definizioni che occorrono nel dizionario-fonte. Si discute inoltre il carattere problematico di alcuni tipi di definizione. Nell'ultima parte è affrontata la pertinentizzazione del significato delle parole che occorrono nelle parafrasi (un tipico problema di *Word Meaning Disambiguation*) e sono mostrate alcune delle regole impiegate per applicare una disambiguazione controllata.

Generazione di una rete semantica da un dizionario. Criteri e procedure

Daniele Christen

1. Introduzione

Questo articolo descrive dei criteri e delle procedure volte ad estrarre in maniera automatica delle informazioni semantiche da un dizionario monolingue in formato elettronico. Questa operazione è parte di un progetto il cui obiettivo è la creazione di una base di dati lessicale strutturata su più livelli, morfologico, sintattico e semantico (ivi comprese le informazioni relative al dominio o ambito d'uso, al registro linguistico ecc.) coerentemente ordinata in base ai significati dei lemmi. La coerenza interna del database lessicale è data dal fatto che tutti i dati sono stati estratti da un medesimo dizionario-fonte[1].

Una risorsa che integri i diversi livelli della desrizione del lessico è fondamentale per un analizzatore sintattico[2], perché i significati delle parole sono associati ai dati morfologici necessari al POS-tagging e a informazioni quali la struttura argomentale (valenza), le restrizioni di sottocategorizzazione, ecc. indispensabili per un'analisi sintattica competa e corretta. La rete semantica costituisce inoltre una risorsa per altre applicazioni: è usata per esempio dai sistemi di ricerca e di classificazione delle informazione contenute in documenti non strutturati (*Information Retrieval*).

Il trattamento informatico del dizionario sorgente ha inoltre prodotto delle informazioni sull'organizzazione del dizionario stesso, sulle caratteristiche stilistiche dei compilatori e sulle occorrenze delle parole usate nelle definizioni. Si tratta di dati che, anche in termini statistici, potrebbero essere oggetto di un'analisi a parte, di interesse lessicografico, in quanto consentono di rilevare quali termini sono maggiormente usati nelle definizioni, venendo a rivestire un ruolo vicino a delle entità semantiche *primitive*. Inoltre, i dati inerenti alle definizioni circolari e ai termini di una definizione dei quali viene a mancare la rispettiva definizione nello stesso dizionario potrebbero suggerire l'adozione di nuovi criteri di razionalità lessicografica.

L'estrazione delle informazioni semantiche è completamente automatica ed è applicata alle definizioni del dizionario preliminarmente sottoposte a un'analisi sintattica integrale[3]. Inoltre procede in base a regole prevalentemente deterministiche, laddove la tendenza attualmente dominante nell'ambito del *semantic tagging* tende a fondarsi su

---

1 Le reti più note reti semantiche disponibili (per l'italiano *ItalWordNet* e *MultiWordNet*, allineate con la *Princeton Word Net*) non sono corredate di informazioni sintattiche. Il progetto CLISP, avviato nel 2000, costituisce invece una risorsa che integra i diversi livelli del lessico (http://www.ilc.cnr.it/clips/). La base lessicale prodotta nell'ambito del presente progetto unifica i dati morfologici, sintattici e semantici estratti da un unico dizionario-fonte, indicizzandoli in base al significato.

2 Vedi: D.Christen, *Syntagma. A linguistic Approach to Parsing*.

3 L'idea di creare una base lessicale in funzione di applicazioni di linguistica computazionale è stata perseguita seguendo gli approcci più diversi almeno da trent'anni a questa parte: vedi per es. A. Zampolli e A. Cappelli 1983; Lesk 1986; R.J. Byrd, N. Calzolari 1987. Per una rassegna (anche bibliografica) più recente rinvio a: Vossen 1992; James Pustejovsky ,Peter Anick ,Sabine Bergler 1993; R. Bindi., N. Calzolari, M. Monachini; V. Pirrelli e A. Zampolli 1994; J. Véronis e N. Ide 1995; S. Granger e M. Paquot 2012.

criteri prevalentemente statistici[4].
L'etichettatura semantica (*semantic tagging*) ha per oggetto sia i termini che occorrono in una data definizione sia le relazioni tra i termini e risulta da un'interpretazione delle categorie morfo-sintattiche e delle relazioni sintattiche contenute nella definizione di un lemma. Nell'ambito di questa operazione sono stati necessariamente affrontati i classici problemi legati alla disambiguazione del senso delle parole (*Word Meaning Disambiguation*), di cui questo articolo propone qualche strategia risolutiva, ristretta all'ambito delle parafrasi lessicografiche.
L'elaborazione informatica dei dati derivati dal dizionario ha inoltre portato all'estrazione automatica di tassonomie e di rappresentazioni degli oggetti che tendono a configurarsi come ontologiche. Da questo punto di vista, il progetto ha dei punti di contatto con i progetti volti alla generazione automatica di tassonomie e di ontologie (*Taxonomy Induction* e *Ontology Learning*)[5].

Per quanto riguarda il problema delicato della riduzione delle relazioni semantiche espresse dal dizionario a una rappresentazione formale di predicati primitivi, si può osservare che, da Aristotele alle tassonomie medievali fino a Leibnitz, da Greimas alle semantiche cognitive e generative, fino a Mel'cuk, Jackendoff, Pustejovsky; da *WordNet* a *dbpedia* e agli standard delle ontologie formali, a dispetto dei diversi paradigmi filosofici e della varietà delle teorie, il catalogo di etichette usate per categorizzare entità lessicali e/o concettuali e le relazioni semantiche rimane sostanzialmente invariato. E resta comunque problematico e irriducibile il rapporto tra i "nomi e le cose"[6], dal momento che “la langue n'est pas une nomenclature”, come insegna Saussure, e le espressioni linguistiche possono ritagliare sezioni sempre diverse del *continuum* (Hjelmslev) dell'esperienza; possono essere oggetto della ben nota “creatività linguistica”, in virtù della loro “onniformatività” (De Mauro), essere riutilizzate in co-testi (Petöfi) e contesti nuovi per significare cose diverse (Pustejovsky), o per dire cose diverse da quelle che sembrano dire (Austin); e la loro formalizzazione risulta talora impossibile (Grice).
Se, come nel caso di questo progetto, ci si fonda sul principio che descrivere il senso di una parola è farne una parafrasi , cioè tradurla in altre parole (Greimas), la spiegazione rimane necessariamente prigioniera del codice stesso che ci si prova a spiegare. Il rischio ben concreto è quello di perdersi in una “semiosi infinita” (Pierce e Eco): un universo di illusoria comprensione del linguaggio, dove i segni rinviano ad altri segni, ma da cui rimangono escluse le componenti fondamentali del senso e dell'intelligenza: l'esperienza del mondo e l'intenzionalità (Searle). Se “il senso di una parola è il suo uso nel linguaggio (Wittgenstein), questo “uso” deve implicare l'attrito con il mondo extralinguistico.
Questa consapevolezza teorica, a mio avviso, non delegittima però una ricerca empirica

---

4 Con qualche eccezione: per es. Padó e Lapata 2007 dottano un approccio ibrido, che comprende anche l'uso delle informazioni sintattiche dei testi analizzati; il loro articolo riporta anche un'ampia rassegna, a cui rimando, della bibliografia relativa al modello vettoriale di prossimità semantica. Questo metodo è fondato sull'*ipotesi distribuzionale* del lessico, che postula che alla similarità distribuzionale delle parole corrisponda una loro prossimità o similarità semantica: vedi un'applicazione del modello per es. nelle ricerche di Baroni e Zamparelli 2010 e Baroni e Lenci 2010. Un esame critico si trova in Sahlgren 2008. Peraltro la recente integrazione di criteri composizionali nel modello distribuzionale (N. Pham, R. Bernardi, Y.-Z. Zhang and M. Baroni 2013) mostra la precarietà di un approccio puramente statistico alla semantica. Sulla necessità di utilizzare delle informazioni sintattiche per le operazioni di semantic tagging si è inoltre già espressa la curatrice di VerbNet (Palmer....)

5 Vedi per es. Velardi, Faralli e Navigli 2012 e la relativa bibliografia a cui rimando.

6 Il riferimento è ovviamente a M. Foucault, *Les Mots et les Choses. Une archéologie des sciences humaines*, 1966, che esamina il carattere paradigmatico e ideologico dei sistemi ontologici.

tesa a esplorare fino ai suoi confini la possibilità di dare a una macchina la possibilità di tradurre in un qualche (arbitrario) formalismo le espressioni linguistiche, per poterlo utilizzare per confrontare tra loro espressioni linguistiche diverse e individuarvi eventuali corrispondenze, in modo da assolvere dei compiti che assomigliano a ciò che fa un essere umano quando riconosce che un testo dice che Tizio compie una certa azione, che Caio possiede determinate qualità o che due testi trattano il medesimo argomento.

Dopo discusso brevemente il rapporto fra dizionario e rete semantica (parte 2) mostrerò i risultati della procedura di estrazione delle informazioni semantiche dal dizionario-fonte e il trattamento preliminare a cui sono state sottoposte. Nelle parti successive spiegherò i criteri e le procedure adottate per la traduzione in una rappresentazione formale del contenuto semantico delle definizioni del dizionario. Discuterò in particolare la nozione di "primitivo semantico" (parte 4) e descriverò alcune delle più diffuse tipologie di definizioni che occorrono nel dizionario-fonte. Mostrerò i meccanismi usati per tradurre le relazioni sintattiche delle parafrasi in relazioni semantiche (parte 5) e presenterò la rappresentazione formale adottata: il *Semantic Frame* (parte 6). Si offrirà anche l'occasione di esemplificare il carattere problematico di alcuni tipi di definizione. In seguito affronterò la pertinentizzazione del significato delle parole che occorrono nelle parafrasi: un tipico problema di *Word Meaning Disambiguation* (parte 7). Illustrerò quindi alcune delle regole impiegate per applicare una disambiguazione controllata.

2. Reti semantiche e dizionari

Il ruolo delle reti semantiche nell'ambito delle tecnologie dell'informazione e della conoscenza è universalmente riconosciuta. Le reti semantiche trovano un impiego nella linguistica computazionale (NLP), come supporto ai processi di parsing (analisi grammaticale/sintattica), nei sistemi di IR, di tagging automatico di documenti e così via.
Una rete semantica consente per esempio di inferire dalla conoscenza che la classe di oggetti chiamati "babbuino" appartiene alla classe (*type)* "scimmia, la verità dei predicati mammifero("babbuino"), animale("babbuino"), e quindi di far ereditare a tutti gli individui della classe "babbuino" le proprietà caratteristiche che pertengono alle classi sovraordinate. Consente inoltre di collegare il nome "arresto" all'azione di "arrestare", che comporta due attori, cioè ruoli-tematici (agente/causa e paziente), chiarendo così il valore semantico di "rapinatore" nell'espressione l'arresto del rapinatore" o di "guasto" nella frase "il guasto ha causato l'arresto del motore". Una rete semantica permette di collegare tra loro parole anche etimologicamente lontane tra loro, come "vendere", "comprare", "acquisto", "acquirente", "merce", "prezzo", assegnando a ognuno di questi termini la sua funzione in uno *script* (Schank e Abelson 1977) che descriva una transazione economica. Ci dice, per esempio, che "martello" è uno strumento e che è associato soprattutto a verbi come "battere", "picchiare", "inchiodare"; che il "falegname" è un artigiano e che l'oggetto principale della sua attività è il "legno". Da questo punto di vista, i confini fra una rete semantica ricca e articolata, in cui confluiscano anche i nomi propri (*Named Entities*), e una enciclopedia diventano estremamente labili[7].
Le reti semantiche più note attualmente sono state realizzate perlopiù manualmente o annotando dati preliminarmente trattati mediante dispositivi informatici, spesso

[7] Sui confini labili tra *dizionario* e *enciclopedia*, dal punto di vista della filosofia del linguaggio, vedi U. Eco 1987.

impegnando decine di persone lungo l'arco di molti anni. I redattori di reti semantiche si trovano spesso a dover risolvere gli stessi problemi che sono familiari ai lessicografi, in particolare con le classi "vaghe", e cioè quando/se determinate proprietà pertengono ad alcuni gruppi di individui o alcuni individui della classe, ma non necessariamete a tutti, e dove quindi occorrono delle definizioni vaghe, sfumate, *fuzzy* (mediante espressioni come "solitamente", "spesso", "normalmente", "talora"), conformi a una definizione prototipicale degli oggetti, con tutte le complicazioni di calcolo del valore di verità di un enunciato che ciò comporta[8].

Un altro aspetto problematico che occorre affrontare nella creazione di una rete semantica è la polisemia: una stessa parola può avere significati anche profondamente diversi a seconda del contesto. Perciò la rete deve collegare fra loro non i lemmi, ma i loro significati, solitamente ben distinti nei dizionari e a cui si associano spesso strutture argomentali (ovvero sintattiche) diverse[9]. Un verbo come "leggere" cambia valenza a seconda del significato: "Paolo sa leggere" (monovalente) vs "Paolo legge un libro (bivalente); nel primo caso l'espressione denota una capacità, nel secondo un'attività, e quindi un oggetto come "giornale" è correlabile i maniera pertinente solo alla seconda accezione (vedi anche l'*Annesso* 3 in appendice a questo articolo, che mostra la ramificazione dei significati dei temini che definiscono un lemma).

L'idea di derivare una rete semantica direttamente dalle definizioni di un dizionario appare quindi come una soluzione a prima vista abbastanza immediata. Non sono infatti mancati i tentativi in questo senso, usando anche degli strumenti computazionali[10]. Ciò che a prima vista incoraggia l'impresa è il fatto che è possibile fare una descrizione tipologica (da un punto di vista strutturale e lessicale) e quindi un inventario tendenzialmente chiuso dei tipi di definizione date nei dizionari.

La loro tipologia è inoltre quasi sempre legate alla categoria grammaticale del lemma che definiscono. Le definizioni di verbi tendono ad iniziare con un verbo all'infinito; quelle di un nome o di un aggettivo iniziano solitamente con un sintagma nominale, la cui testa è un sinonimo o un iperonimo del definendo, completato da aggettivi, da sintagmi preposizionali o da frasi relative che ne esprimono le caratteristiche specifiche.

Gli ostacoli che incontra chi volesse estrarre delle informazioni semantiche direttamente dalle parafrasi definitorie di un dizionario sono però molteplici. Nonostante le analogie che consentono di riconoscere un insieme abbastanza ben delimitato di strutture linguistiche impiegate per la definizione dei lemmi, sussiste comunque un grande numero di varianti sia a livello distribuzionale sia a livello lessicale (vedi sezione di questo articolo). Per es. il nome, sinonimo o iperonimo del definendo, con cui inizia una parafrasi, può essere preceduto da un determinante o da un aggettivo ("tigre: grande felino ecc.") e le caratteristiche possono essere espresse da un semplice aggettivo, da participio ma anche da una frase relativa complessa che richiede un'analisi specifica. Nel caso dei verbi, dove la definizione ricorre solitamente a un verbo iperonimo a cui sono aggiunte delle espressioni avverbiali di tipo modale che ne circoscrivono il senso, la formulazione di queste caratteristiche conosce un'ampia gamma di variazioni: avverbi, gruppi preposizionali, frasi subordinate di vario tipo; il loro significato può inoltre essere prettamente modale ("correre" è definito come un "procedere" "velocemente" o "di fretta"), ma anche causale-efficiente o causale-finale ("spremere" "schiacciare" "per

[8] Per un esame complessivo del problema della categorizzazione degli oggetti nella psicologia cognitiva e di riflesso nella linguistica e nella semiotica, rimando a Violi 1997. Sul fronte linguistico il problema è sollevato per la prima volta da Labov 1977.

[9] Vedi per es. Leks 1986; J. Véronis e N. Ide 1995.

[10] Vedi nota 3, *supra.* p.1 e nota 4. p.2.

estrarne il succo"), come pure temporale ("mietere" è "tagliare" il grano "quando è maturo").
Anche sul piano lessicale le definizioni non sono sempre improntate a una coerenza intratestuale (non solo a causa dei diversi stili redazionali dei compilatori, ma perché è indirizzato a un lettore umano, non certo a una macchina). Per esempio nella scelta del termine iperonimo si trovano sensibili variazioni: un oggetto può essere definito uno "strumento", un altro un "arnese", ma questi due termini appaiono come sinonimi dal momento che la definizione di "arnese" inizia con "strumento" e viceversa. Certe specie di animali sono definiti ricorrendo a una nomenclatura esatta in termini zoologici, ma con un diverso riferimento per quanto riguarda la classificazion (specie, classe, famiglia, ordine,...); e anche in questo caso mediante strutture sintattiche molto eterogenee: "mammifero acquatico ecc." o "animale domestico della specie dei felini...".
È inoltre evidente che i termini utilizzati nella definizione, con quel particolare senso contestuale, devono essere correlati a loro volta con altri nodi della rete, cioè con altri significati, il che può risolversi nelle definizioni circolari a cui nessun dizionario da noi esaminato sfugge.
Per risolvere questi problemi ho adottato la strategia di sottoporre le definizioni del dizionario a un'analisi grammaticale (*POS-tagging)* e sintattica integrale (mediante un *parser*), e di applicare le procedure per l'estrazione dei dati semantici alle parafrasi definitorie così analizzate.

3. Dal dizionario al *database* lessicale

La scelta del dizionario-fonte ha considerato in particolar modo la presenza di informazioni preziose per il *parser,* in particolare la valenza, che nel nostro caso si differenzia a seconda dei significati. Il dizionario in questione limita la menzione della valenza ai soli verbi (i curatori assicurano una futura estensione di questo dato ad altre categorie).
A questo proposito posso segnalare che tra i prodotti secondari scaturiti dal presente progetto c'è un applicativo che può essere impiegato per estrarre da vasti corpora di testi delle informazioni sul comportamento sintattico delle parole, compresi i dati come la valenza, la reggenza, la sottocategorizzazione anche di nomi, aggettivi.
La versione elettronica del dizionario-fonte è stata dapprima oggetto di una scansione, da parte di un modulo informatico costruito *ad hoc*, che ha portato all'estrazione e alla classificazione di tutte le informazioni contenute nelle voci: dalla sillabazione, alla categoria grammaticale, dalle caratteristiche morfologiche (comprese forme particolari del plurale o del femminile, spesso ristrette a particolari significati della parola) a quelle sintattiche (valenza, restrizioni di sottocategorizzazione, possibilità che certi argomenti siano inespressi, ecc.). Inoltre sono state estratte e registrate separatamente: le informazioni relative a un uso settoriale (come "matematica", "diritto", "sport"), al registro linguistico e alla funzione retorica ("famigliare", "ironico", "figurato"); i sinonimi e i contrari; le locuzioni idiomatiche in cui appare la parola data; gli esempi e le *parafrasi definitorie* del lemma, cioè quelle che costituiscono la propriamente la definizione del significato data dal dizionario (che d'ora in poi chiamerò semplicemente *parafrasi*)[11].

[11] La centralità della nozioni di parafrasi nello studio della semantica (e della linguistica in generale) è affermata e argomentata con determinazione da Mel'čuk 2012, I 2: 45-78. Come si vedrà, il modello qui presentato, che risulta da una ricerca autonoma iniziaza nel 1989, presenta molti punti di contatto con il modello Meaning-Text di Mel'čuk.

Nel database sono state inoltre registrate delle informazioni supplementari, come il *referente tipico*, che è solitamente introdotto con una formula come "detto di...." ("abbaiare: detto di cane...."), e la reggenza (la preposizione e la categoria subordinata) di molti lemmi, ricavabile dagli esempi. Qui di seguito è riportato l'elenco delle informazioni estratte:

- varianti del lemma ("televisione", "tivù", "tv"...)
- sillabazione
- categorie grammaticali assumibili
- paradigmi della flessione, compreso il diverso valore semantico di certe forme (*i bracci vs le braccia*)
- l'eventuale posizione del lemma in parole polirematiche e locuzioni idiomatiche
- caratteristiche e restrizioni sintattiche (valenza, reggenza, sottocategoriz zazione, ristrutturazione, ecc.)
- la parafrasi delle varie accezioni del lemma
- informazioni relative all'uso settoriale, al registro, ai valori connotativi
- informazioni relative alle parole che coprono tipicicamente certi ruoli tematici ("abbaiare": *detto di "cane"* indica l'agente tipico del senso letterale di questo verbo)
- esempi
- i termini correlati ("acqua" > "idrico"), quando menzionati

Tav. 1

Le parole ricevono un indice progressivo (ID) e le diverse accezioni di una medesima parola sono contrassegnate da un indice (MNG). I dati sono distinti e indicizzati in base al significato, il che significa che le caratteristiche morfologiche, sintattiche e semantiche sono correlate a una specifica accezione di un lemma dato (e non genericamente al lemma). Questa caratteristica si rivela cruciale per esempio durante le procedure di analisi sintattica (*parsing*), in quanto le strutture sintattiche concorrono a selezionare il senso delle parole in una data frase e, reciprocamente, il significato delle parole opera selettivamente sulle strutture e interpretazioni sintattiche applicabili alla frase.
Dai circa 53'000 lemmi della lingua italiana contenuti nel dizionario-fonte sono state estratti poco più di 105'000 significati, ognuno corredato dei dati indicati sopra.

Classi chiuse:	1'134
Nomi, Aggettivi e Avverbi (lemmi):	53'000
Significati individuati e registrati	105'000

Tav. 2

Oggetto delle operazioni successive sono state le *parafrasi*, isolate dal contesto della voce del lemma. Esse sono state dapprima sottoposte a un'analisi sintattica mediante un *dependency parser* che peraltro impiega le informazioni morfologiche (POS-tagging) e sintattiche risultanti dall'estrazione descritta qui sopra[12].
Il *parser* è stato opportunamente adattato alle particolari strutture linguistiche usate in un dizionario (frasi con verbo reggente all'infinito, sintagmi nominali, preposizionali, aggettivali autonomi, cioè sganciati dalle strutture sintattiche nelle quali usualmente

[12] Il constraint based dependency parser che utilizza questi dati, svilupparo dai miei colleghi della Parsit (www.parsit.it) a Torino (Italy), ha ottenuto il miglior punteggio al concorso Evalita 2011. Il fatto di poter disporre di infomazioni di tipo sintattico ha dato al nostro parser un indubbio vantaggio rispetto ai sistemi che procedono inmaniera puramente statistica (data-driven).

compaiono come complementi e modificatori). Riporto qui sotto un esempio di definzione, la sua riduzione a records di dati e infine analisi sintattica di una delle sue parafrasi prodotta dal *parser* (in formato pseudo-CoNLL):

amadriade [a-ma-drì-a-de] s.f. 1 Ninfa dei boschi che, secondo la mitologia greca, nasceva e moriva con l'albero che le era sacro 2 zool. Grossa scimmia africana con folta criniera, muso canino, coda a ciuffo # sec. XV

:ID 1462 :MNG 3641 :LEX "amadriade" :PARFR "Ninfa dei boschi che, secondo la mitologia greca, nasceva e moriva con l'albero che le era sacro" :CAT 200 :AUX :RFL 0 :TRN 0 :VAL "" :GEN 2 :NUM 1 :FRMPL ""

:ID 1462 :MNG 3642 :LEX "amadriade" :PARFR "Grossa scimmia africana con folta criniera, muso canino, coda a ciuffo" :CAT 200 :AUX :RFL 0 :TRN 0 :VAL "" :GEN 2 :NUM 1 :FRMPL ""

0	amadriade		amadriade	NOUN		3642	MEANING
1	Grossa	grosso	ADJ	2	RMOD		
2	scimmia	scimmia	NOUN	0	TOP		
3	africana	africano	ADJ	2	RMOD		
4	con	con	PREP	6	CONN		
5	folta	folto	ADJ	6	RMOD		
6	criniera	criniera	NOUN	2	RMOD		
7	,	,	PUNCT	8	COORD-BASE		
8	muso	muso	NOUN	6	COORD2ND-BASE		
9	canino	canino	ADJ	8	RMOD		
10	,	,	PUNCT	11	COORD-BASE		
11	coda	coda	NOUN	8	COORD2ND-BASE		
12	a	a	PREP	13	CONN		
13	ciuffo	ciuffo	NOUN	11	RMOD		

Riporto come secondo esempio l'analisi di un verbo ("iscrivere"):

iscrivere [i-scrì-ve-re] v. (irr.: coniug. come scrivere) # v.tr. [sogg-v-arg-prep.arg] Registrare, inserire qlcu. o qlco. in un registro, in una lista: iscrivi il mio nome nell'elenco dei partecipanti; i. la spesa sul registro dei conti; estens. associare qlcu. a una scuola, una società e sim.: i. il figlio alla prima classe del liceo # iscriversi # v.rifl. [sogg-v-prep.arg] Compiere le formalità necessarie per essere ammesso a un'organizzazione o a un'attività come membro, socio, partecipante: i. a un partito # sec. XIV

I dati estratti:

(:ID 21003 :MNG **50283** :LEX "iscrivere" :PARFR "Registrare, inserire qlcu. o qlco. in un registro, in una lista" :TD "" :CAT 100 :AUX "avere" :RFL 0 :TRN 2 :VAL "[sogg-v-arg-prep.arg]" :GEN 0 :NUM 0 :FRMPL "")

(:ID 21003 :MNG **50284** :LEX "iscrivere" :PARFR "associare qlcu. a una scuola, una società" :TD "" :CAT 100 :AUX "avere" :RFL 0 :TRN 2 :VAL "[sogg-v-arg-prep.arg]" :GEN 0 :NUM 0 :FRMPL "")

(:ID 21003 :MNG **50285** :LEX "iscrivere" :PARFR "Compiere le formalità necessarie per essere

ammesso a un'organizzazione o a un'attività come membro, socio, partecipante" :TD "" :CAT 100 :AUX "essere" :RFL 1 :TRN 3 :VAL "[sogg-v-prep.arg]" :GEN 0 :NUM 0 :FRMPL "")

La formalizzazione delle informazioni grammaticali e sintattiche ricavate dalla definizione:

ID 21003 :LEX "iscrivere" :MNG **50283** :CAT 100 :AUX avere :TRN 2 :RFL 0 :FRMPL :TD :SEM1 :CTRL F :RAIS F :PSTV F :PREP (***in,su***) :TIPlex (***elenco,registro***) :TIPid (13031,31555)
:VAL [(:FNCT **sogg** :PSET (:PRP (:CAT NP :RGG () :OPT F :MDV () :SEM ()))
(:FNCT **v** :PSET (:PRP (:CAT VP :RGG () :OPT F :MDV () :SEM ()))
(:FNCT **arg** :PSET (:PRP (:CAT NP :RGG () :OPT F :MDV () :SEM ()))
(:FNCT **prep.arg** :PSET (:PRP (:CAT PP :RGG () :OPT F :MDV () :SEM ()))]

:ID 21003 :LEX "iscrivere" :MNG **50284** :CAT 100 :AUX avere :TRN 2 :RFL 0 :FRMPL :TD :SEM1 :CTRL F :RAIS F :PSTV F :PREP (**a**) :TIPlex () :TIPid (2)
:VAL [(:FNCT **sogg** :PSET (:PRP (:CAT NP :RGG () :OPT F :MDV () :SEM ()))
(:FNCT **v** :PSET (:PRP (:CAT VP :RGG () :OPT F :MDV () :SEM ()))
(:FNCT **arg** :PSET (:PRP (:CAT NP :RGG () :OPT F :MDV () :SEM ()))
(:FNCT **prep.arg** :PSET (:PRP (:CAT PP :RGG () :OPT F :MDV () :SEM ()))]

:ID 21003 :LEX "iscrivere" :MNG **50285** :CAT 100 :AUX essere :TRN 3 **:RFL 1** :FRMPL :TD :SEM1 :CTRL F :RAIS F :PSTV F :PREP () :TIPlex () :TIPid ()
:VAL [(:FNCT **sogg** :PSET (:PRP (:CAT NP :RGG () :OPT F :MDV () :SEM ()))
(:FNCT **v** :PSET (:PRP (:CAT VP :RGG () :OPT F :MDV () :SEM ()))
(:FNCT **prep.arg** :PSET (:PRP (:CAT PP :RGG (***a***) :OPT F :MDV () :SEM ()))]

Si può osservare come ognuna delle tre accezioni è corredata di un ricco sistema di dati grammaticali (pensati in particolar modo per essere impiegati da un parser di alto livello, *rule-based*), come l'ausiliare, la riflessività (:RFL), la valenza con i sue argomenti (diversi a seconda dell'accezione), *controllo, sollevamento*.
Ogni argomento della *valenza* è inoltre provvisto di una serie di restrizioni, in particolare: *reggenza* (:RGG), *sottocategorizzazione* (:CAT), l'indicazione se il costituente è obbligatorio o facoltativo (:OPT). Possono essere inoltre istanziate delle restrizioni per quanto riguarda il modo verbale (:MDV), nonché *restrizioni semantiche* (per es. animato o inanimato) per ognuno degli argomenti (:SEM).
Nel caso di MNG 50283 sono anche registrate alcune parole (ricavate dagli esempi della definizione) che occupano tipicamente la posizione di prep.arg. per quella particolare accezione del verbo ("elenco","registro").

Le definizioni analizzate dal parser (riporto l'analisi di solo due delle tre parafrasi):

0	iscrivere	iscrivere	VERB	50283	MEANING
1			NOUN	2	SUBJ
2	Registrare	registrare	VERB	0	TOP INFINITE
3	,	,	PUNCT	5	COORD-BASE
4			NOUN	5	SUBJ 0.1
5	inserire	inserire	VERB	3	COORD2ND-BASE INFINITE
6	qualcuno	qualcuno	PRON	5	OBJ
7	o	o	CONJ	7	COORD-BASE
8	qualcosa	qualcosa	PRON	7	OBJ
9	in	in	PREP	11	CONN

10	un	un	ART		11	INDEF
11	registro	registro	NOUN	5	RMOD	
12	,	,	PUNCT	15	COORD-BASE	
13	in	in	PREP	15	CONN	
14	una	una	ART		15	INDEF
15	lista	lista	NOUN	5	RMOD	
0	iscrivere	iscrivere	VERB	50284	MEANING	
1			NOUN	2	SUBJ	
2	associare	associare	VERB	0	TOP	INFINITE
3	qualcuno	qualcuno	PRON	2	OBJ	
4	a	a	PREP	6	CONN	
5	una	una	ART		6	INDEF
6	scuola	scuola	NOUN	2	RMOD	
7	,	,	PUNCT	9	COORD-BASE	
8	una	una	ART		9	INDEF
9	società	società3	NOUN	2	RMOD	

Nella fase successiva le definizioni sintatticamente analizzate sono tradottenella r*ete semantica di base*. Le sezioni che seguono illustrano alcune caratteristiche delle definizioni del dizionario che possono essere sfruttate per definirne ed etichettarne semanticamente il contenuto.

4. Primitivi semantici nel contesto di questo modello

L'etichettatura semantica del contenuto delle definizioni dei lemmi (e quindi del lemma stesso) consiste in due operazioni chiaramente distinte. La prima riguarda l'etichettatura delle parole. La seconda riguarda l'etichettatura delle relazioni fra le parole all'interno di una data parafrasi. Una nozione che entra necessariamente in gioco nella scomposizione semantica delle parafrasi lessicografiche, è la nozione di "termine "primitivo", cioè il limite inferiore di scomposizione raggiungibile.
Nella prossima sezione discuterò il senso che la nozione di "primitivo" viene ad assumere in questo modello. Nelle sezioni successive esaminerò, nell'ordine, dapprima l'etichettatura delle "parole" e poi quella delle relazioni semantiche tra le "parole".
Ciò che interessa qui è la nozione di "primitivo" semantico in termini puramente linguistici, anzi, lessicografici, e non in termini filosofici o di psicologia cognitiva, anche se queste dimensioni sono ovviamente tra loro collegate. In pratica, come postulato anche nel modello *Meaning-Text* di Mel'čuk, con il quale il presente modello ha molte analogie, i termini di una parafrasi possono a loro volta essere scomposti nei termini che compongono la loro parafrasi; e questa scomposizione può essere ripetuta ciclicamente finché con si raggiungono i termini "primitivi" contingenti, cioè quelli arbitrariamente stabiliti dall'economia stessa del dizionario[13].
La prospettiva da cui intendo affrontare la *vexata quaestio* dei primitivi non è quella deduttiva, oggetto delle piu diverse teorie filosofiche e linguistiche, ma quella induttiva ed empirica, fondata sul sistema lessicografico del dizionario-fonte che è alla base di questo progetto[14].

[13] Vedi: Mel'čuk 2012, II, 4, in particolare 184-188.

[14] Miller e Johnson-Laird 1976 hanno proposto un inventario di primitivi semantici e di relativi campi semantici a partire da un'analisi lessicologica; vedi anche Johnson-Laird 1983, cap. 15. Altre teorie tendono

Se si esamina il materiale lessicale impiegato dal dizionario si possono stabilire almeno due criteri complementari per individuare i termini che hanno i valori semantici più ampi o più generici e che quindi in qualche maniera assumono il ruolo di entità primitive. Ambedue i criteri sono di natura prettamente statistica, ma forniscono comunque delle informazioni già di per sé significative.
Il primo criterio è il numero di significati attribuiti dal dizionario a un certo lemma. Si può facilmente immaginare che, quanto più ampio è il ventaglio di accezioni e di usi di una parola, tanto più generico è il suo senso.
Il secondo criterio è il numero di occorrenze di una parola nelle definizioni delle altre parole del dizionario. Anche in questo caso, è lecito supporre che il carattere in qualche modo "primitivo" del termine è proporzionale alla frequenza con cui occorre nelle parafrasi definitorie.
Il secondo criterio può essere ulteriormente affinato, correlando il dato puramente quantitativo della frequenza con quello qualitativo del ruolo semantico svolto nella definizione. Riporto parte della tabella con i lemmi che presentano il numero più alto di occorrenze, ed mi limito qui, per motivi di spazio, a qualche osservazione inerente ai dati più significativi. Le parole che occorrono in assouto più frequentemente nelle definizioni del dizionario sono le seguenti (mi limito i lemmi con più di 15'000 occorrenze):

PAROLA	RELAZIONE SEMATICA	FREQUENZA
essere	TYPE_OF	168637
fare	TYPE_OF	127672
avere	TYPE_OF	95839
parte	HAS_OBJ	84376
dare	TYPE_OF	41'791
mettere	HAS_QUALITY	35468
luogo	HAS_MATTER,HAS_SPACE,HAS_DIMENSION	35314
cosa	HAS_OBJ	34577
potere (verbo e nome)	TYPE_OF	33202
prendere	TYPE_OF	30822
persona	HAS_DEST,HAS_CAUSE	30784
acqua	HAS_AGNT	29299
tempo	HAS_SPEC	28383
stato	HAS_MATTER,HAS_SPACE,HAS_DIMENSION	26206
corpo	TYPE_OF	24459
punto	TYPE_OF	23264
forma	HAS_MATTER,HAS_SPACE,HAS_DIMENSION	21258
andare	HAS_QUALITY	21084
parte	HAS_OBJ	18840

invece a postulare i primitivi semantici, sostenendone addirittura il carattere innato (Fodor 1980), e proponendone degli inventari di numero inevitabilmente variabile e dal carattere inesorabilmente provvisorio. Ho fatto una rassegna delle teorie sulla nozione di primitivo semantico in un'altra sede (D. Christen 1999) e quindi non la discuterò ulteriormente qui.

vita	HAS_DEST,HAS_MANNER	17817
usare	TYPE_OF	17407
tenere	HAS_QUALITY	17149
atto	TYPE_OF	16613

È necessaria l'avvertenza che la relazione semantica che appare accanto a un termine in questa tabella, non è correlata esattamente con la frequenza. Un esame più rigoroso, che ci riserviamo di fare in altra sede, distingue le occorrenze secondo la relazione semantica che intrattengono con le altre parole nella parafrasi definitoria. Inoltre questa tabella non distingue la frequenza per categorie. I dati completi sono naturalmente disponibili per ulteriori analisi. L'esame di questi grezzi dati statistici può comunque già essere oggetto di qualche considerazione.
Tolti i verbi "essere" e "avere", di cui in questa tabella non sono stati distinte le occorrenze come ausiliare (che non sono significative al nostro proposito) e quelle come verbo (significative), notiamo che le parole con la frequenza più alta nelle definizioni è data da quelle che effettivamente possono aspirare al ruolo di *primitivi semantici* anche solo da un punto di vista intuitivo. Verbi come "fare" (.... occorrenze), "dare" (..), "potere" (....), "prendere" (...) e nomi come "cosa" (.... occorenze), "luogo" (....), "persona" (....), possono quindi essere assunti empiricamente come primitivi semantici, che costituiscono dunque il limite della scomposizione semantica delle definizioni.
Queste entità primitive potranno essere in seguito oggetto di una formalizzazione ulteriore, che li riduca alla loro *forma logica*, ricorrendo ai termini della logica modale che consenta di rendere conto delle loro implicazioni temporali (per esempio di verbi primitivi come "diventare", che implicano stati qualititivi diversi di un medesimo oggetto lungo l'asse del tempo rappresentabili mediante dei predicati modali) e le loro presuppsizioni logiche[15].
Occorre anche osservare che, se ci si propone una rigorosa definizione dei termini primitivi (nell'ambito del dizionario-fonte) i termini dovrebbero essere sottoposti a un esame più attento, per individuare quelli che nelle parafrasi sono sinonimi: come per esempio "aumentare", "crescere". ma anche "aumento" e "crescita", la cui disseminazione è il mero prodotto della *variatio* stilistica dei redattori e i cui indici di frequenza andrebbero quindi sommati per valutarne la consistenza, in termini quantitativi, di primitivi semantici.
Mi propongo di tornare su questo argomento in maniera più completa e rigorosa in un'altra occasione. Intanto, i dati ricavati da questo esame hanno portato alla definizione di un primo e provvisorio, in questa fase sperimentale del progetto, insieme di etichette semantiche di base (o *primitive*) che ho impiegato in via del tutto empirica nelle fasi successive del progetto, che illustrerò nelle prossime sezioni.

5. Etichetattatura semantica delle unità lessicali e delle loro relazioni

Nelle prossime sezioni discuterò l'assegnazione di etichette semantiche alle entità (nel senso più ampio del termine) che hanno una realizzazione lessicale nel dizionario-fonte (sezione 5) e alle relazioni tra queste entità (sezione 6).
Il formalismo adottato e le relative etichette sono prettamente empiriche ed arbitrarie e

[15] Jackendoff 1983 e 1990; Chierchia e McConnel-Ginet 1993; Moss e Tiede 1997.

possono essere adeguate, in un momento successivo, agli *standards* proposti nell'ambito della descrizione semantica del lessico in linguistica computazionale[16].

5.1 Parole riferite a oggetti, eventi, qualità e azioni

La prima operazione consiste nell'assegnazione di alcune categorie di base a delle parole che occorrono frequentemente nelle definizioni e che costituiscono delle entità in qualche modo "primitive", a partire dalle quali vengono definite le altre. Nella sezione precedente abbiamo menzionato alcuni criteri che hanno permesso di postulare alcune entità e relazioni primitive, all'interno dell'universo semantico di *questo* dizionario, in base alla frequenza delle parole. Le categorie di base applicate qui coincidono perlatro spesso, perlomeno a livello sostanziale, con quelle più diffuse nelle tassonomie e ontologie correnti. Inoltre, come si è detto, esse costituiscono solo un punto di partenza provvisorio, visto che la struttura ontologica del "mondo" rappresentato nel dizionario si genererà automaticamente applicando le procedure di analisi semantica.
Alcune delle categorie di base (*type_of*) qui adottate sono "cosa" (*thing*); "persona" (*person*), "azione" (*action*), "qualità" (*quality,* estesa anche a modalità di eventi e azioni), "atto di parola" (*speech-act*), percezione (*perception),* "cambiamento" (*change*) con le sue sottospecificazioni "spazio", "qualità". Alcune di queste categorie sono attribuite immediatamente in base alla parola che funge da testa della definizione: "persona", "luogo" sono *tags primitivi* automaticamnete ereditati dai significati di cui aprono la definizione; un verbo di movimento o un nome derivato da un verbo di movimento è etichettato come *type_of(change,place)*; un verbo come "diventare" equivale a *(change, quality).* La definizione di un nome introdotto dalla formula "atto di" seguita da un verbo, consente di etichettare il nome come indicante un'azione: *type_of(action).* Ma le etichette di base possono essere attribuite anche secondo le caratteristiche grammaticali e morfologiche del lemma base: un verbo, se non classificato diversamente in base a un'identità parola-etichetta, corrisponde solitamente a un'azione; un aggettivo a una qualità, un nome a una cosa generica (*thing*). I nomi e gli aggettivi definiti da una frase che inizia con il pronome "chi" mostrano chiaramente di riferirsi a una persona. Ma non c'è una corrispondenza sistematica fra categoria grammaticale e concetto: un nome deverbale, come visto, può indicare un'azione; un aggettivo può indicare una qualità ma anche l'agente di un verbo ("fuggitivo") o il suo paziente (nel caso di aggettivi corrispondenti al participio passato di verbo transitivo); un verbo può indicare un cambiamento di stato o di forma o di dimensione, che non è propriamente un'azione.
Le etichette assegnate provvisoriamente durante un primo ciclo di esplorazione dei significati del dizionario, sono poi sostituite o arricchite con delle etichette più precise e specifiche durante i cicli successivi.

5.2 La relazione di iper/iponimia: la relazione *type_of*

Individuare le parole che hanno un grado di genericità più alto di altre è relativamente facile secondo il nostro approccio: come si è visto, basta classificare secondo l'indice di frequenza le parole che formano la testa dei sintagmi che reggono sintatticamente le

[16] Vedi per es. le raccomandazioni del gruppo di lavoro EAGLES (http://www.ilc.pi.cnr.it/ EAGLES96/ EAGLESLE.PDF); di EuroWordNet (Vossen et al. The eurowordnet base concepts and top ontology, 1998) e ACQUILEX (Technical Annex).

definizioni.
La relazione di base che quindi si stabilisce tra il lemma-base e il primo termine che funge da testa della parafrasi (o i termini coordinati) è quindi genericamente la relazione *type-of* (equivalente al *is_a* di WordNet), che corrisponde il più delle volte con una relazione di iperonimia. Si tratta di una relazione binaria: *type_of*(a,b) dove il primo termine è l'iponimo e il secondo l'iperonimo:

type_of ("fabbro","artigiano")
type_of ("artigiano","person")
type_of ("tigre", "felino")
type_of ("correre","andare").

Come si può subito notare, alcuni termini diversi (per es. "luogo" e "spazio") hanno una frequenza ambedue sufficientemente alta da poter essere considerati iperonimi di altre parole. Associando ad ambedue l'etichetta di "luogo" (*place*), questa caratteristica semantica si propaga a tutti i lemmi la cui definizione abbia come testa uno di questi due termini. Da questo punto di vista, la rete semantica risulta più economica del dizionario-fonte.
Una seconda procedura fa la scansione, lemma per lemma, dei valori *y* associati al secondo argomento della relazione *type_of*(x,*y*), cercandone le occorrenze come primo argomento della medesima relazione negli altri lemmi type_of(*y*,z) e inferendone per transitività l'attribuzione della classe di genericità più alto al lemma di partenza:

$$\text{type\_of}(x,y) \wedge \text{type\_of}(y,z) \rightarrow \text{type\_of}(x,z)$$

Questa procedura, ripetuta ciclicamente finché non si producono nuove relazioni transitive, genera delle catene di termini iperonimi, posti gerarchicamente secondo un grado sempre maggiore di genericità, che vengono a configurarsi come una classificazione ontologica (come un albero di Porfirio) dei significati dei termini associati alle parole.
In base al loro contenuto referenziale si può dunque procedere a una prima classificazione delle parole e della forma delle loro definizioni.
Come si evince da questa sommaria rassegna, la relazione di iperonimia svolge un ruolo fondamentale nelle definizioni del dizionario. Questa caratteristica è usata dal programma che genera la rete semantica per creare automaticamente delle catene iperonimiche (una serie gerarchica di relazioni *type-of*), ossia delle tassonomie e degli "alberi di Porfirio" degli oggetti (per esempio: *babbuino* > *scimmia* > *mammifero* > *vertebrato* > *animale* > *organismo*)[17]. Dato che di ogni oggetto, sui diversi piani di generalità, sono espresse le principali caratteristiche specifiche, è possibile derivare dalle definizioni del dizionario un'ontologia.

5.3 Nomi

Come già detto, non esiste (e non può esistere) nel dizionario una uniformità tra la categoria grammaticale del lemma-base e quella del definendo: i nomi sono esemplari da questo punto di vista, in quanto possono designare referenti sostanziali, e quindi essere

[17] Vedi, in appendice a questo articolo, gli schemi semantici di "scimmia", "mammifero" "vertebrato" (Annesso 1).

parafrasati mediante un sintagma nominale; ma possono designare azioni o eventi e quindi essere parafrasati mediante verbo (solitamente introdotto da un *predicato di attribuzione* come "atto di"), oppure qualità di oggetti, ed essere parafrasati mediante un aggettivo. Da un sommario esame risultano le sguenti tipologie:

a) Nomi che si riferiscono a esseri umani sono generalmente introdotti dall'iperonimo (primitivo) “persona” o dal pronome personale “chi”, seguiti da modificatori (aggettivi o frase relativa) che ne danno le caratteristiche specifiche, relative al sesso, all'età, allo stato sociale, alla funzione pubblica, all'attività esercitata.
b) Nomi (e anche aggettivi) riferiti a comportamenti di persone o a stati in cui una persona figura nel ruolo di paziente sono solitamente introdotti da un sinonimo; o anche dal pronome “chi” seguito da una frase relativa che specifica il comportamento o la stato caratteristico; oppure direttamente da un participio passato (con diatesi passiva) del verbo che indica l'azione di cui il referente è paziente.
c) Nomi riferiti a animali e vegetali o a cose che sono oggetto di una classificazione scientifica, tecnica o istituzionale, sono definiti perlopiù da un iperonimo (spesso specifico dell'ambito o dominio d'uso) e da modificatori (aggettivi, sintagmi preposizionali, frasi relative). Nomi con referenti più generici sono descritti mediante la medesima struttura sintattica e introdotti solitamente da sinonimi o iperonimi. Lo stesso vale per termini astratti, riferiti a idee, valori, concetti, movimenti culturali, ma anche attività).
d) Nomi che esprimono un'azione (in generale nomi deverbali) sono introdotti dalla formula “atto/azione di” seguita dal relativo verbo; oppure da una frase infinitiva che descrive l'azione.
e) Nomi che indicano proprietà caratteristiche, qualità e modalità, e che spesso hanno un corrispettivo aggettivo, sono di solito definiti direttamente da una frase relativa (“ciò che....”, o semplicemente “che....”).
f) Sono inoltre frequenti le parafrasi in cui si mostra un esplicito atto di denominazione: la parola (solitamente un nome) è presentata come “denominazione/nome/titolo di....” seguita dall'entità così denominata. In alternativa viene usata la formula “detto di....”, che spesso è ridotta alla semplice preposizione “di” in apertura di definizione (“di stabilimento balneare”).

5.4 Aggettivi

Le definizioni di aggettivi (Adj) sono il più spesso espresse mediante un aggettivo iperonimo modificato da un avverbio che ne restringe/delimita il valore semantico (esempio) oppure da frasi relative (RelClause) con diatesi attiva o passiva. In questi ultimi casi, che esamineremo più da vicino qui, la testa dellla definizione è data da un verbo (V):

a1) adesivo: "Che *aderisce*, che *si attacca*"
a2) solvibile: "Che *può* pagare"
a3 indeciso: "che *è* aperto a varie possibilità"
a4) iodico: "Che *è* a base di iodio"
a5) menagramo: "Chi *è ritenuto* capace di portare sfortuna"
a6) grande: "che *ha* dimensioni maggiori rispetto a un'altra dello stesso tipo"
a7) grande: "Che *supera* la misura ritenuta normale in volume, altezza, quantità,

ampiezza, capienza, forza, intensità, durata"

Riconosciamo, nella costruzione sintattica di queste definizioni, alcune tipologie di diversa complessità:

a1) è dato da due verbi coordinati con diatesi attiva, che riconducono l'aggettivo a delle azioni. Ciò significa che il referente del nome, a cui in una qualunque frase fosse associato questo aggettivo ("adesivo"), ha la proprietà di compiere le azioni di "aderire" e di "attaccarsi", cioè ne rivestirà il ruolo tematico di agente. Il lemma, nella sua accezione di aggettivo, qui identificata dal campo *meaning (*MNG), sarà registrato nella base di dati semantica come:

a1) :LEX "adesivo" :MNG 2300
:CAT adj :TYPE_OF "quality" :AGENT_OF ("aderire","attaccarsi")

a2) ha come testa un V(1) modale (come "dovere", "volere", "sapere" "essere capace di" e relativi sinonimi) che ha come argomento il verboV(2) "pagare". La presenza del modale impone l'attribuzione di un predicato formale predefinito :MODAL che assume come valore il contenuto lessicale del verbo modale (in questo caso "potere"), dal quale dipenderà logicamente il predicato AGENT_OF:

a2) :LEX "solvibile" :MNG 36000
:CAT adj :TYPE_OF "quality" :MODAL "potere"
:AGENT_OF, POTENTIAL "pagare"

Se in un'espressione linguistica a un N sarà associato l'aggettivo "solvibile", se ne potrà dedurre che ha la proprietà di essere potenzialmente capace di compiere l'azione di "pagare".

a3) è dato dal copulativo "essere" che ha come argomento il sintagma aggettivale complesso (AdjP) la cui testa è "aperto". Quando la definizione di un aggettivo rinvia ad un altro aggettivo (con o senza la mediazione sintattica della copula "essere"), la procedura impone la relazione HAS_QUALITY che assumerà come valore il contenuto lessicale dell'aggettivo, il cui modificatore, dato dal sintagma preposizionale "a varie possibilità", sarà registrato nella sua forma sintattica, senza ulteriori elaborazioni durante il primo ciclo di estrazione:

a3) :LEX "indeciso" :MNG 15000
:CAT adj :TYPE_OF "quality" :HAS_QUALITY (N,"aperto"
(:UA "a varie possibilità" [si omette qui l'analisi]))

a4) è dato dal V -copula "essere" e da un sintagma preposizionale che, con le sue varianti riscontrate nel nostro dizionario di riferimento ("composto di", "costituito da" ecc.), è associabile, in una prima fase, al predicato HAS_PART che indica una relazione di meronimia:

a4) :LEX "iodico" :MNG 18000
:CAT adj :TYPE_OF "quality" :HAS_PART (N,"iodio")

In una successiva fase di elaborazione della base di dati semantica, in base all'associazione:

"iodio" → TYPE_OF "matter"

predicato HAS_PART potrà essere sostituito o completato dalla relazione MATTER (N,"iodio").

a5) ha come testa un verbo (con diatesi passiva) il cui argomento è un complemento predicativo formato dall'aggettivo con valore modale "capace"+"di"+Vinf. Il predicato principale "ritenere" appartiene alla famiglia di sinonimi con la medesima struttura valenziale [sogg-v-compl.pred] che comprende anche "considerare", "prendere+per", "reputare".
A questi verbi viene assegnato il predicato a tre argomenti ATTRIBUTION(AGENT, PATIENT, ATTRIBUTE). Dato che la diatesi è passiva, il referente del lemma-base occuperà la posizione del secondo argomento (ruolo tematico di PATIENT), il terzo argomento è saturato dal complemento predicativo, mentre il primo è saturato da quello che la grammatica tradizionale chiama complemento d'agente ("da"+SN), qualora fosse espresso.
Nel nostro caso la definizione è resa più complessa dal fatto che il complemento predicativo è un'espressione modale ("essere capace di", sinonimo di "potere") a cui è applicata la formalizzazione già descritta più sopra:

a5) :LEX "menagramo" :CAT adj :TYPE_OF "quality"
:ATTRIBUTION(nil, N, (MODAL "potere" AGENT_OF, POTENTIAL "portare" (HAS_OBJ "sfortuna")))

a6) ha come testa il verbo "avere", con diatesi attiva (per cui il pronome relativo ha la funzione di soggetto), che ha come argomento un SN complesso un costrutto al quale associamo il predicato HAS_QUALITY.
Il valore della relazione è saturato dal SN, la cui testa è il N "dimensioni" modificato dall'attributo "maggiori" e da un altro SP complesso.
Nella prima fase del processo è fomalizzata solo la testa del SN con il suo modificatore, e la parte non ulteriormente analizzata è registrata in un campo apposito:

a6) :LEX "grande" :CAT adj :TYPE_OF "quality"
:HAS_QUALITY "dimensione" (:MOD "maggiore"
:UA ("rispetto a un'altra dello stesso tipo"))

Trascurando per il momento l'analisi della specificazione "rispetto a un'altra dello stesso tipo", è interessante rilevare già qui che la definizione di "maggiore" (Adj) dice:

a8) *maggiore* : Più grande, superiore

che porta, per quanto riguarda la prima parte della parafrasi ("più grande") a una relazione quasi circolare (se si ignorano i modificatori), mentre la seconda conduce alle definizioni dell'aggettivo:

a9) *superiore*: Che sta più in alto, che sta sopra
a9') *superiore*: Maggiore per altezza, grandezza, numero, intensità

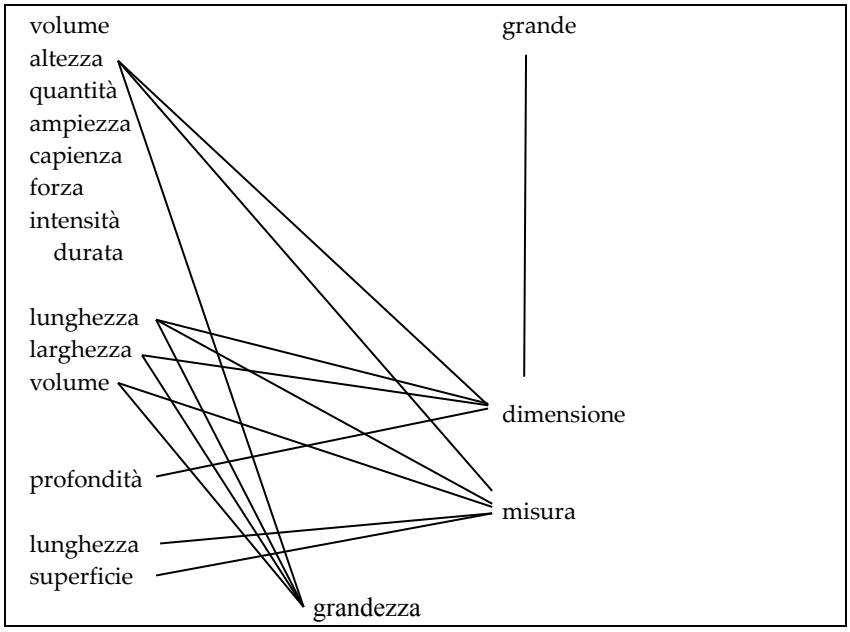
di cui la seconda acquista maggiore pertinenza atraverso la mediazione dei nomi:

a10) *grandezza*: Dimensioni, mole di un oggetto considerato in lunghezza,
larghezza, altezza, volume
a11) *dimensione*: Ciascuna delle misure che determinano l'estensione di un
corpo nel piano o nello spazio (lunghezza, larghezza, altezza o profondità)

in cui co-occorrono le parole "dimensione" (che legano questi termini all'aggettivo "grande"), "altezza", "grandezza", a cui si legano altri termini più specifici ("lunghezza", "altezza", "larghezza" ecc.). Si nota inoltre la sinonimia tra il secondo significato di "maggiore" (a8) e quello di "superiore" (a9').
La seconda definizione di "grande" (a7) è anch'essa espressa mediante una F relativa, la cui testa è data dal V "superare", un predicato bivalente che qui ha come primo argomento (soggetto) la variabile N saturabile dal nome di cui l'aggettivo è attributo; e come secondo argomento un SN complesso (contenente una F relativa participiale associabile al predicato ATTRIBUTION) la cui testa è la parola "misura":

a7) :LEX "grande" :CAT adj :TYPE_OF "quality"
:AGENT_OF "superare"(N, "misura"
(:ATTRIBUTION(nil,"misura","normale")),
:UA ("in volume, altezza, quantità, ampiezza, capienza, forza, intensità, durata"))

dove il terzo argomento della relazione ATTRIBUTION porrà, a livello di interpretazione, i tipici problemi della *fuzzy logic* (come determinare la misura "normale" per un certo tipo di oggetti? Da che "dimensione" o "misura" in poi un certo oggetto può essere considerato "grande"?).



Tav. 3

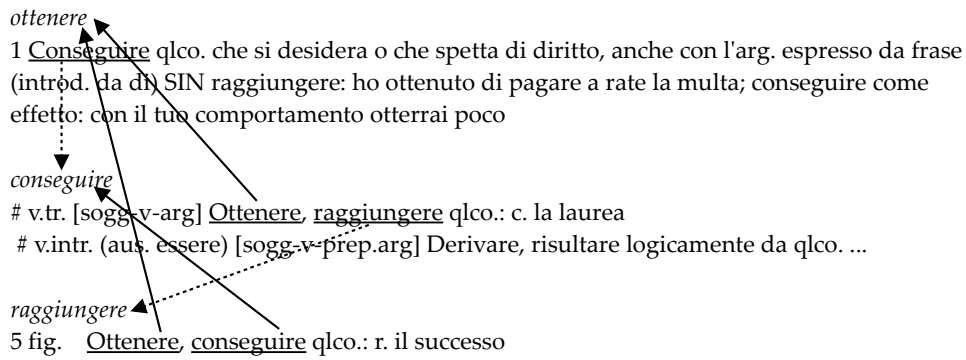
La parte non analizzata (:UA) in questa fase dell'estrazione, potrà essere formalizzata in una seconda fase di estrazione, quando sarà disponibile anche l'analisi semantica di una delle accezioni del nome:

a12) *misura*: Insieme delle dimensioni (lunghezza, altezza, superficie,
volume, ecc.) di un oggetto o del corpo umano

che si riallaccia a sua volta ai nomi e agli aggettivi visti sopra, che vengono quindi a costituire un vero e proprio campo semantico, ovvero una sotto-rete semantica (Tav. 3).

5.5 Verbi. Sinonimia e pseudo-sinonimia

I verbi sono solitamente spiegati mediante una frase infinitiva, il cui verbo è un iperonimo o un sinonimo del lemma-base. Soprattutto per i verbi sono frequenti le
Fig.1
definizioni circolari, per cui *a* è definito con *b* e la definizione di *b* rinvia ad *a*.
Un esempio è dato dai verbi "conseguire, "ottenere", "raggiungere":



La circolarità è evidente:

ottenere > *conseguire* > *ottenere*
ottenere > *conseguire* > *raggiungere* > *conseguire*
ottenere > *conseguire* > *raggiungere* > *ottenere*

Ciò può avvenire anche in casi in cui i due termini non sono indicati esplicitamente come sinonimi: "raggiungere" è marcato come sinonimo di "ottenere", ma la relazione tra "conseguire" e "raggiungere" e quella tra "ottenere" e "conseguire" e non sono rilevate, nonostante la seconda sia mediata da "raggiungere". Chiameremo queste relazioni delle relazioni di *pseudo-sinonimia*.
La circolarità funziona da *stopper* nel processo automatico di decomposizione semantica delle definizioni.

5.6 Definizioni problematiche

In qualche caso la circolarità diventa insidiosa, in particolare quando l'occorrenza del

primo termine nella definizione del secondo assume un significato che non è riconducibile in maniera plausibile ai significati attribuitigli nella sua definizione. Un esempio è dato dal valore del termine "prestazione", in particolare nel linguaggio settoriale del diritto:

prestazione 1 Messa a disposizione della propria opera, della propria competenza, delle proprie capacità: p. professionale || p. d'opera, attività svolta in un rapporto di lavoro dipendente 2 Risultato ottenuto, rendimento fornito in un'attività SIN performance: la squadra ha fornito una buona p. 3 (spec. pl.) Rendimento di una macchina, di un dispositivo, ecc.: p. di un'automobile 4 dir. Oggetto e contenuto di un'obbligazione

obbligazione (...) 3 dir. Rapporto giuridico che costringe un soggetto (debitore) a effettuare una determinata prestazione a favore di un altro soggetto (creditore): o. pecuniaria

Nell'ambito del diritto, una "prestazione" è definita come "contenuto" (genericamente iponimo) di una "obbligazione". La definizione di "obbligazione" vede un *soggetto1* che si trova a dover "effettuare" una "prestazione" (si immagina di tipo "patrimoniale") a favore di un *soggetto2*. La definizione circolare risulta qui aberrante, perché sarebbe come dire che una prestazione patrimoniale è "oggetto e contenuto" di se stessa.

Altrettanto problematiche sono le definizioni in cui un termine appare come contenuto di una disciplina omonima, e dove quindi il lemma che viene definito appare, in un'accezione diversa (ma ciò è sottinteso) nella definizione stessa:

diritto 1 Insieme di principi codificati allo scopo di fornire ai membri di una comunità regole oggettive di comportamento su cui fondare una ordinata convivenza...
2 Disciplina che ha per oggetto di studio i principi, le norme costitutive del diritto SIN giurisprudenza

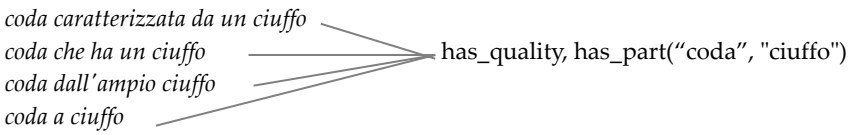
In questi casi il sistema informatico deve applicare delle regole che tengono conto delle caratteristiche testuali del documento che sta analizzando, in particolare il principio (che si auspica sia applicato con coerenza in tutto il dizionario) per cui la definizione di una disciplina che ha per contenuto un oggetto omonimo segue, nell'ordinamento delle definizioni, la definizione del suo oggetto. In questo caso l'oggetto di dititto(1) è il diritto(1).
L'assegnazione di indici (almeno nella versione elettronica del dizionario) che rimandino alle accezioni pertinenti risolverebbe anche le ambiguità correlate, che si presentano per esempio in una definizione come la seguente, in cui occorre il termine "mitologia", ma non è evidente determinare quale delle due accezioni di "mitologia" presentate dal dizionario sia quella pertinente:

ninfa 1 Nella mitologia classica, creatura femminile divinizzata, abitatrice di boschi e acque

mitologia 1 Il complesso dei miti di una certa cultura o epoca: m. greca
2 Scienza dei miti; studio dei miti

5.7 Relazioni semantiche

Da un esame delle relazioni fra le parole all'interno delle parafrasi del dizionario emergono immediatamente alcune caratteristiche importanti.
Nonostante la grande varietà di espressioni usate dai redattori del dizionario, le relazioni possono essere solitamente ricondotte a un insieme piuttosto ristretto, a cui è possibile assegnare una stessa etichetta. Occorre perciò fare un inventario delle molteplici varianti che esprimono una medesima relazione. Queste vengono quindi a costituire delle vere e proprie *famiglie* di espressioni:



Ciò non esclude naturalmente, accanto a queste relazioni frequentissime, la presenza di relazioni molto particolari, per le quali non conviene concepire una corrispondente relazione formale (un *tag*), per non incorrere in una proliferazione incontrollabile delle relazioni primitive.
L'esempio precedente mostra inoltre che, in una prima fase dell'interpretazione semantica delle parafrasi, certe relazioni risultano necessariamente ambigue (come in questo caso tra meronimia e attribuzione di una qualità), almeno finché non si è giunti a definire il valore semantico dei termini della relazione[18].
Una parte importante del materiale verbale usato nelle definizioni non ha una "consistenza ontologica" o referenziale e può essere direttamente formalizzato come una relazione di *attribuzione*. L'oggetto dell'attribuizione è una delle relazioni di base. Per esempio le formule "formato /composto /costituito da..." possono essere direttamente ridotte alla relazione di meronimia (*has_part*); la formula "in maniera/in modo..." seguita da un aggettivo è un'attribuzone di qualità che lega un verbo, rispettivamente un nome, a una sua maniera o qualità (*has_quality*). C'è poi una famiglia di espressioni ("usato /impiegato per....", "adibito a...", "volto a....", "che serve a ...") usate nelle definizioni di nomi che riconducono tutti all'attribuzione di una funzione o di uno scopo (*has function, has goal*), espressi dal verbo o dal nome che segue. Lo stesso vale per i corrispondenti nomi deverbali ("per uso/ impiego domestico"). All'attribuzione di una relazione di causa (*has_cause*) è dedicata una famiglia specifica di espressioni che comprendono "dovuto a...", "causato da...", "conseguente a..." , che possono essere facilmente inventariate. Espressioni come "in grado di...", "capace di..." indicano una potenzialità: il soggetto è potenzialmente agente dell'azione espressa dal verbo che segue, come per es. nella definizione di "animale", descritto come "organismo sensibile in grado di muoversi spontaneamente", che è formalizzata come segue:

```
2.4  1    ATTRIBUTION  "in grado"        ADJ
          [AGNT_OF, POTENTIAL(2.5,ACTION,CHANGE,PLACE,"muovere")]
2.5  4    AGNT_OF, POTENTIAL  "muovere"        VERB
          [TYPE_OF(ACTION, CHANGE,PLACE,"muovere"),
          HAS_QUALITY(2.6, QUALITY,"spontaneamente")]
```

[18] Questa operazione di disambiguazione, che passa necessariamente per la selezione del significato congruente che i termini della relazione ("coda" e "ciuffo", nell'esempio) hanno nel contesto di una data definizione, è argomento delle sezioni.

Dove il fatto che il predicato *agent_of* è accompagnato dal predicato *potential* gioca un ruolo importante nel calcolo del valore di verità della definizione in termini di *fuzzy logic*, per cui è anche un gatto paralitico e un mulo costretto a camminare a forza di bastonate restano pur sempre degli animali che soddisfano la definizione.
Analoghe a queste, sono le espressioni che attribuiscono valori quantitavi agli oggetti (dimensioni, velocità ecc.), la cui interpretazione può essere suffragata dalla co-occorrenza di termini che denotano una misura ("litro", "metro", "grado", ecc.).
Certe definizioni sono introdotte dalla formula "tipo/specie di...", che esplicita la relazione tra il lemma e il termine che figura come specificatore della testa della definizione ("tipo"):

> 1 zool. (al pl., iniziale maiusc.) Tipo di animali, al quale appartiene anche l'uomo, caratterizzati dalla presenza di uno scheletro interno...

la cui traduzione immediata in termini semantici contiene i predicati di attribuzione "tipo di", "caratterizzata da" (*attribution, quality, has_part*) e "presenza" che è un'attribuzione di esistenza (*has_existence*). L'espressione "di cui fa parte anche l'uomo" risulta già tradotta in una relazione ambigua tra meronimia e iponimia (*has_part, has_token*). Come detto, i predicati di attribuzione possono essere cancellati dallo schema concettuale. L'attribuzione della relazione type_of("tipo") risulterebbe peraltro ridondante. Perciò un meccanismo di "sollevamento" sposta il contenuto del predicato di attribuzione al posto del predicato di attribuzione stesso. L'entità lessicale che esprime l'attribuzione viene cancellata dalla rappresentazione formale del concetto, per lasciar posto unicamente alla relazione semantica di base:

```
60.0    "vertebrato"    NOUN  103952
60.1    [TYPE_OF(60.2,THING,ANIMAL,"animale"),
        HAS_TOKEN(60.17,THING,PERSON,"uomo")]
60.2    [TYPE_OF(THING,ANIMAL,"animale"),
        HAS_PART(60.5,THING,"scheletro")]
60.5    [TYPE_OF(THING,"scheletro"),
        HAS_QUALITY(60.6,QUALITY,"interno")
```

Nell'operazione di etichettatura semantica delle relazioni che occorrono fra i termini, la struttura sintattica della stesse ha un ruolo fondamentale. Molte delle relazioni di base sono ricavabili immediatamente dalle relazioni sintattiche che intercorrono tra i costituenti della parafrasi. Così, la relazione tra un nome e un aggettivo o una frase relativa è immediatamente interpretabile come l'attribuzione di una qualità (*has-quality*), che, nel caso della frase relativa, può consistere in un'azione o in uno stato espresso dal verbo, di cui il referente è agente oppure paziente (nel caso di relative con diatesi passiva).
Le funzioni sintattiche di soggetto e di oggetto di verbi agentivi attivi possono essere tradotte perlopiù nei ruoli di agente e di oggetto o paziente (ulteriori distinzioni possono intervenire a un livello successivo dell'analisi). Il dizionario sorgente inoltre indica spesso il valore semantico degli argomenti (soprattutto nei verbi trivalenti, dove il terzo argomento può ricoprire vari ruoli attanziali) e dei costituenti avverbiali facoltativi (usando delle espressioni come "spesso con specificazione della provenienza", per es. per i verbi di movimento).

Più problematica è l'attribuzione di un'etichetta ai costituenti introdotti da preposizioni, dato che queste non hanno un significato univoco. In questi casi vengono assegnate tutte le relazioni semantiche solitamente introdotte dalla preposizione, rinviando alla fase successiva la selzione della relazione pertinente. Avverbi e frasi gerundive indicano una maniera (ovvero una qualità) di un'azione o di un evento. Il contenuto della qualità è dato dal valore semantico del modificatore stesso, che può indicare delle circostanze di luogo o di tempo, come pure degli attributi di maniera.

5.8 La struttura sintattica delle parafrasi

Un esame della struttura sintattica delle parafrasi porta inoltre alle seguenti conclusioni:

i) è impiegato un numero relativamente ridotto di costrutti sintattici;
ii) tutte le costruzioni sintattiche hanno la particolarità di essere dei costituenti che solitamente non appaiono indipendentemente nel discorso formale, dato che sono tipicamente dei costituenti dipendenti, cioè SN, PP, SAdj, frasi infinitive e frasi relative.
iii) la coordinazione è usata in maniera massiccia, sia sul piano principale della definizione dove occorrono spesso più iperonimi coordinati tra loro, sia sul piano delle caratteristiche specifiche dell'oggetto, che solitamente sono presentate in forma di enumerazione.

Se ne conclude che l'analisi sintattica automatica delle parafrasi è possibile, ma il *parser* deve essere adattato a questi particolari costrutti: deve accettare come espressioni grammaticali autonome dei costituenti che solitamente non appaiono indipendentemente; può inoltre escludere frasi "normali" che hanno come testa un verbo al modo finito. Particolare precisione è poi richiesta nella gestione delle strutture coordinate, un settore delicato della sintassi, dove le ambiguità di aggancio sono piuttosto la norma che l'eccezione.

6. La rappresentazione formale delle definizioni del dizionario

Una definizione viene tradotta in quello che definisco uno *schema concettuale* (*Semantic Frame*), a sua colta costituito da *unità concettuali* (*Semantic Units*) legate tra loro da relazioni semantiche di vario tipo. La relazione tra il lemma-base e il termine che costituisce la testa sintattica della definizione è solitamente di tipo iperonimico, quindi una relazione *type-of*. A titolo esemplificativo riprendiamo le due parole ("amadriade" e "iscrivere") citate all'inizio di questo articolo. La loro rappresentazione formale, a questo livello della traduzione semantica della struttura sintattica mostrata nella sezione 2, assume il seguente aspetto:

```
3203.0  "amadriade"   NOUN 3642
                      [TYPE_OF(THING,"amadriade" ),
                      TYPE_OF(3203.1,THING,"scimmia" ),
                      HAS_QUALITY(3203.2,QUALITY,"grosso" ),
                      HAS_QUALITY(3203.3,QUALITY,"africano" ),
                      HAS_PART(3203.4,THING,"criniera" ),
```

```
                        HAS_PART(3203.6,THING,"muso" ),
                        HAS_PART(3203.8,THING,"coda" )]
3203.1  "scimmia"       [TYPE_OF(THING,"scimmia" )]
3203.2  "grosso"        [TYPE_OF(QUALITY,"grosso" )]
3203.3  "africano"      [TYPE_OF(QUALITY,"africano" )]
3203.4  "criniera"      [TYPE_OF(THING,"criniera" ),
                        HAS_QUALITY(3203.5,QUALITY,"folto" )]
3203.5  "folto"         [TYPE_OF(QUALITY,"folto" )]     ]
3203.6  "muso"          [TYPE_OF(THING,"muso" ),
                        HAS_QUALITY(3203.7,QUALITY,"canino" )]
3203.7  "canino"        [TYPE_OF(QUALITY,"canino" )]
3203.8  "coda"          [TYPE_OF(THING,"coda" ),
                        HAS_QUALITY,HAS_PART(3203.9,THING,"ciuffo" )]
3203.9"ciuffo"          [TYPE_OF(THING,"ciuffo" )]
```

```
41.0  "iscrivere"   VERB      50283
                    [TYPE_OF(ACTION,"iscrivere" ),
                    TYPE_OF(41.1,ACTION,"registrare" ),
                    HAS_OBJ/PATENT(41.2,PERSON,"qualcuno" )
                    HAS_OBJ/PATENT(41.3,THING,"qualcosa" )
                    HAS_SPACE,HAS_TARGET(41.4,THING,"registro"
                    HAS_SPACE,HAS_TARGET(41.5,THING,"lista" ]

42.0  "iscrivere"   VERB      50284
                    [TYPE_OF(ACTION,"iscrivere" ),
                    TYPE_OF(42.1,ACTION,"associare" ),
                    HAS_OBJ/PATIENT(42.2,THING,PERSON,"qualcuno" ),
                    HAS_TARGET(42.3,THING,ORG,PERSON,COLLECT,"scuola" ),
                    HAS_AGENT(42.4,THING,"tR$" )]
```

In questa prima fase della generazione, le relazioni puntano a delle semplici parole (delle stringhe prive di una determinazione). La maggior parte di esse è dotata di più significati, in virtù della natura polisemica dei segni linguistici, i cui valori semantici sono in questa fase ancora indeterminati. L'*annesso* 3 (in appendice) esemplifica il problema mostrando la diramazione dei significati dei termini che definiscono il lemma "abaco".

Quindi, perché gli schemi semantici (Semantic *Frames*) acquistino consistenza, occorre identificare e selezionare quale delle accezioni della parola a cui punta una data relazione semantica sia congruente[19].

Il tentativo di selezionare i significati pertinenti all'interno di una data definizione basandosi su criteri meramente statistici si è rivelato fallimentare. È stata quindi necessaria l'elaborazione di un complesso algoritmo che tiene conto di tutte le informazioni rese disponibili dal dizionario, a partire da quelle esterne alla parafrasi (come il dominio d'uso, i referenti tipici ecc.), fino alle relazioni semantiche definite durante la prima fase qui descritta. Spesso è necessario recuperare anche delle informazioni relative alla struttura sintattica, e la scelta di operare sulle parafrasi preventivamente analizzate dal parser si è mostrata decisiva. Alla descrizione di questa operazione, complessa e delicata la parte seguente di questo articolo.

[19] Sulle discordanze manifestate da persone, nello svolgimento di questo compito di annotazione semantica, vedi Fellbaum C., Delfs L., Wolff S., Palmer M., 2004.

7. Pertinetizzazione del contenuto delle parafrasi

7.1 Un problema di *Word Meaning Disambiguation*

Nelle sezioni precedenti ho mostrato come è possibile interpretare e tradurre in relazioni semantiche le categorie e i legami sintattici che strutturano la parafrasi. Dopo la prima fase di elaborazione, lo schema concettuale (*Semantic Frame* = SF), in cui si traduce la parafrasi di un particolare significato di un certo lemma, si presenta come un sistema di concetti (*Semantic Units* = SU) provvisti di un'etichetta semantica, le cui relazioni con le altre unità presenti nella parafrasi sono a loro volta state interpretate e provviste di un *tag* semantico.
Ma l'entità lessicale cui fa capo ogni concetto non è ancora definita, e punta a tutti i significati che potenzialmente può assumere. Il database che contiene le parafrasi dei lemmi è strutturato e indicizzato in base ai significati dei lemmi (identificati dall'indice MNG). Il compito che occorre a questo punto affrontare è quello di definire il valore semantico dei termini che compongono la parafrasi: il che equivale a selezionare quei significati (individuati dall'indice MNG) che hanno la maggiore congruenza nel co-testo della definizione. Ho scelto di non eliminare dal CF i significati delle CU che non soddifsano i criteri, ma piuttosto di assegnare loro un punteggio, che può essere anche negativo. Tutti i significati potenzialmente assumibili dai temini (CU) restano così disponibili per altre operazioni sulla rete semantica, mentre solo quelli con un punteggio positivo saranno presi in considerazione dagli applicativi che usano la rete semantica (per esempio un *parser* sintattico nei compiti di disambiguazione, un sistema di *Information Retrieval* o un *Question Answring System*).
Il punteggio è assegnato in base a dei criteri eterogenei che combinano delle strategie di semantica distribuzionale con dei dati che provengono dalla definizione del lemma e che sono estranei alla parafrasi (per es. l'ambito d'uso settoriale di una parola o il suo registro linguistico), e delle informazioni che provengono dalle parafrasi stesse, a diversi livelli e che operano selettivamente: dalle categorie grammaticali, che possono essere ammesse o non ammesse in un determinato contesto semantico, alle caratteristiche morfologiche (certe accezioni di un lemma sono per esempio ristrette all'uso del plurale), fino ai legami sintattici che risultano dal parsing preliminare e a quelli semantici definiti nella prima fase della generazione del database semantico.
Nella prossima sezione presenterò le nozioni di base del formalismo che adotto in questa descrizione. Nelle successive illustrerò più da vicino i dati che vengono utilizzati e alcune delle regole che sono applicate[20].

7.2 Notazione: termini e relazioni assiomatiche

/a/	: una parola (anche polirematica)
a	: una particolare accezione della parola /a/. Equivale a: /a/(*a*).
A	: l'insieme dei significati assumibili dalla parola /a/: A = {*a1, a2, a3*...}
D^a	: l'insieme delle parole contenute nella definizione di un'accezione del lemma /a/. D(a) = {/b/(*b1,b2*), /c/(*c1,c2,c3*)...}
S^a	: insieme dei significati pertinenti delle parole contenute nella definizione del lemma /a/. S^a = S(*a*) = {*b2,c1,*...}

[20] Un articolo che presenta queste regole e le relative procedure in maniera completa è in preparazione: *Rule-based Semantic Tagging. An Example.* (forthcoming).

$a = s(S^a)$: il significato *a* è funzione (s) delle relazioni semantiche tra i significati pertinenti S^a della definizione D^a

head(/b/,D^a) : la parola /b/ è la testa (sintattica) della definizione del lemma /a/. La testa della definizione può anche essere data da termini coordinati: head((b,c), a).

hyper(/b/,*a*) : la parola /b/ è un iperonimo o pseudo-sinonimo del significato "a" della parola /a/. Condizioni:
se head(/b/, D^a) → hyper(/b/,a), e, per la transitività dell'iperonomia:
se head(/c/, D^b) ∧ head(/b/, D^a) → hyper(/c/,a)

syn(*a*,*b*) : due significati sono sinonimi o pseudo-sinonimi se la definizione dell'uno rinvia a quella dell'altro: se head(/a/, D^b) ∧ head(/b/, D^a) → syn(c,a)

rel^a(/b/,/c/) : una qualunque relazione semantica tra due parole /b/ e /c/ all'interno di una data definizione D^a del lemma /a/.
Esempio: type_of(/b/,/c/), has_quality(/b/,/c/), has_part(/b/,/c/).
Uno dei termini della relazione può essere lo stesso lemma di base /a/, per esempio quando la definizione è introdotta da un pronome: coreferente con il lemma stesso, come in "pelle": "quella che ricopre...".

rel^a(*b*,*c*) Forma assunta della relazione, quando è individuata l'accezione pertinente dei termini /b/ e /c/ all'interno della definizione /a/.
Vale allora l'equivalenza: $rel^a(b,c) = s(S^a) = a = a(/a/)$

inc_score(*a*) : incrementa il punteggio di attendibilità dell'accezione "a" nell'ambito di una data definizione.

dec_score(*a*) : decrementa il punteggio di attendibilità dell'accezione "a" nell'ambito di una data definizione.

7.3 Iperonimia

Come ho mostrato spiegando la prima fase della generazione della rete semantica, nelle definizioni ha un ruolo rilevante, anche in termini quantitativi, il legame di iperonimia o di pseudo-iperonimia. Questa proprietà delle definizioni è sfruttata dall'algoritmo che pertinentizza il valore delle parole all'interno di una parafrasi, in quanto le proprietà del referente di un iperonimo sono spesso esplicitate anche a livello dei sui iponimi.
Ciò consente di confrontare il contenuto di iponimi diversi tassonomicamente dipendenti da un medesimo iperonimo. Inoltre permette, come già rilevato, di generare delle catene iperonimiche. La prima regola R0 dell'algoritmo genera le coppie di parole legate da iperonimia, che può può rivelarsi una pseudo-sinonimia (vedi sezione 5.5).

R0) hyper(a,b)
gorilla [go-rìl-la] s.m. inv.
1 Grande scimmia africana, con pelle nera ricoperta da pelo grigio scuro....
2 fig. Uomo dal fisico possente e dai modi grossolani
3 fig. Guardia del corpo: essere scortato dai g. # a. 1875; a. 1970 (3)

scimmia [scìm-mia] s.f.
1 Denominazione generica di *mammiferi* con corpo coperto di peli....
2 gerg. Sbornia: prendersi delle s. terribili; nel gergo della droga, crisi di astinenza

mammifero [mam-mì-fe-ro] agg., s.
s.m. Denominazione generica di animale che ha le ghiandole mammarie # a. 1855

Nell'esempio, le parole “gorilla”>”scimmia” > “mammifero” > “animale” formano una catena iperonimica, cioè un ramo di un sistema tassonomico:
La parola "vertebrato", che non è entrata nella tassoomia durante il primo ciclo di estrazione, è recuperata durante il secondo ciclo, che integra questi dati con quelli forniti dal frame seguente:

```
101692.0  0        "vertebrato"    NOUN103953
          [HAS_TOKEN(101692.2,THING,"mammifero"),
          HAS_TOKEN(101692.3,THING,"rettile"),
          HAS_TOKEN(101692.4,THING,"uccelli"),
          HAS_TOKEN(101692.5,THING,"anfibi"),
          HAS_TOKEN(101692.6,THING,"pesce")]
```

che risulta dalla traduzione (e successiva eleminazione: vedi sezione ....) dell'espressione usata nel dizionario "...ne fanno parte mammiferi, rettili, uccelli, anfibi, pesci").
Il risultato è una lista che correla il lemma "amadriade" coni significati che potenzialmente ne sono degli iperonimi, a livelli diversi della tassonomia:

```
"amadriade";3642;NOUN;HYPER;"scimmia";85532;NOUN
"amadriade";3642;NOUN;HYPER;"scimmia";85533;NOUN
"amadriade";3642;NOUN;HYPER;"scimmia";85534;NOUN
"amadriade";3642;NOUN;HYPER;"mammifero";55168;NOUN
"amadriade";3642;NOUN;HYPER;"vertebrato";103953;NOUN
"amadriade";3642;NOUN;HYPER;"animale";4770;NOUN
"amadriade";3642;NOUN;HYPER;"animale";4771;NOUN
"amadriade";3642;NOUN;HYPER;"animale";4772;NOUN
"amadriade";3642;NOUN;HYPER;"animale";4773;NOUN
"amadriade";3642;NOUN;HYPER;"animale";4774;NOUN
"amadriade";3642;NOUN;HYPER;"organismo";64296;NOUN
"amadriade";3642;NOUN;HYPER;"organismo";64297;NOUN
"amadriade";3642;NOUN;HYPER;"organismo";64298;NOUN
```

In una fase successiva sarà selezionato ulteriormente il significato che costituisce l'iperonimo più appropriato (vedi sezione 6).

7.4 Regole fondate su dati esterni alla parafrasi

Una prima serie di regole utilizza i dati della definizione esterni alla parafrasi vera e propria. Si tratta di informazioni relative al dominio (o settore) d'uso di un dato significato di una parola, alla determinazione del registro linguistico, ecc. (vedi sezione 3 di questo articolo).
Come già rilevato da altri[21], il dominio (e il relativo sottocodice settoriale) a cui appartiene un determinato uso di una parola può giocare un ruolo determinante nella disambiguazione e naturalmente anche nella pertinentizzazione di un termine all'interno di una parafrasi. Possono presentarsi quattro situazioni diverse (E1-E4) e occorre terner conto del fatto che un termine /b/ della definizione D[a] di un significato "a" ha normalmente un grado di generalità più alto di "a", soprattutto se si tratta della testa della

[21] Magnini, Strapparava, Pezzulo e Ghiozzo 2003; Bentivogli, Girardi e Pianta 2003.

definizione: per cui i significati di /b/ saranno iperonimi o sinonimi del significato "a". Se dunque il significato "a" è ristretto a un certo dominio, un certo significato "b" di /b/ può appartenere o non appartenere a quel dominio, e il suo punteggio sarà incrementato. Nel caso opposto, se "a" non ha restrizioni di dominio, e il significato "b" appartiene a un dato dominio, (cioè è più ristretto), non può essere pertinente nel contesto dato; e quindi il suo punteggio sarà decrementato. Se ambedue i significati sono legati a un dominio, e questi non fossero identici o non fossero compatibili (come per esempio "economia" e "diritto"), il punteggio di "b" diminuisce.

E1) se $(/b/ \in D^a) \wedge domain(a) = domain(b) \rightarrow inc\_score(b)$

E2) se $(/b/ \in D^a) \wedge domain(a) = x \wedge domain(b) = \varnothing \rightarrow inc\_score(b)$

E3) se $(/b/ \in D^a) \wedge domain(a) = \varnothing \wedge domain(b) = x \rightarrow dec(score)$ oppure score=score

E4) se $(/b/ \in D^a) \wedge domain(a) <> domain(b) \rightarrow dec\_score(b)$

Se due termini sono in relazione sinonimica esplicita (dichiarata dalla definizione), allora il punteggio viene incrementato (E5, E6); lo stesso vale quando coincide la determinazione del registro linguistico (E7) oppure quanto i due significati hanno il medesimo referente (E8, E9); in caso contrario il punteggio rimane invariato:

E5) se $(/b/ \in D^a) \wedge synonym(a) = synonym(b) \rightarrow inc\_score(b)$

E6) se $(/b/ \in D^a) \wedge (synonym(a) = /b/ \vee synonym(b) = /a/) \rightarrow inc\_score(b)$

E7) se $(/b/ \in D^a) \wedge register(a) = register(b) \rightarrow inc\_score(b)$

E8) se $(/b/ \in D^a) \wedge (refers\_to(a) = refers\_to(b)) \rightarrow inc\_score(b)$

E9) se $(/b/ \in D^a) \wedge (refers\_to(a) = /b/ \vee refers\_to(b) = /a/) \rightarrow inc\_score(b)$

La verifica sul referente è estesa anche agli iperonimi dei due termini:

E10) se $(/b/ \in D^a) \wedge hyper(c,a) \wedge (refers\_to(a) = /b/ \vee refers\_to(b) = /a/) \rightarrow inc\_score(b)$
Se il significato "a" del lemma-base /a/ è parte di una locauzione idiomatica e il significato "b" di un suo termine appartiene anch'esso a una locuzione, ma diversa, allora "b" non è considerato un significato pertinente nella definizione di "a". Dato che una locuzione idiomatica o una parola polirematico sono solitamente parafrasate mediante dei termini più generali, che non fanno parte di una locuzione, il punteggio rimane invariato. Se invece il termine "b" appartiene a una locuzione e il definendo no, allora il punteggio è negativo, in quando non ho riscontrato casi in cui una parola sia parafrasata mediante una locuzione:

E11) se $(/b/ \in D^a) \wedge idiom(a) <> idiom(b) \rightarrow dec\_score(b)$

E12) se $(/b/ \in D^a) \wedge idiom(a) = \varnothing \wedge idiom(b) = x \rightarrow dec(score)$ o

Si tratta di dati certamente significativi in sé, e di cui occorre tener conto quando ci sono, ma che da un punto di vista quantitativo sono piuttosto limitati. Sui ca. 106'000 significati individuati nel dizionario sorgente, quelli a cui è attribuito un dominio sono poco più di 16'000. Le determinazioni di registro sono ca. 6'000; le locuzioni idiomatiche 15'400; le attribuzioni di un referente tipico 5'400. Da un punto di vista quantitativo, questi dati non bastano per operare una pertinentizzazione estesa su un corpo così vasto di lemmi e di significati.

7.5 Regole fondate su dati interni alla parafrasi

Le regole seguenti hanno per oggetto il contenuto della parafrasi, cioè i termini e le loro definizioni. Queste regole sono quelle risultate più produttive sul piano della pertinentizzazione del valore delle parole all'interno delle parafrasi. Le regole, di cui mostreremo le più importanti, operano dei confronti incrociati a diversi livelli:

a) tra il lemma base /a/ e la definizioni D^b di un termine /b/ della definizione "a"
b) tra le definizioni D^b , D^c del valore dei termini /b/, /c/ contenuti nella definizione di "a"
c) tra le definizioni degli iperonimi di "a" e la definizione D^b di un valore del termine /b/
d) tra il lemma /a/ e la definizione degli iperonimi di "b"
e) tra gli iperonimi dei significati dei termini /b/ e /c/ contenuti nella definizione "a".

Come si vedrà, queste regole incrementano notevolmente le possibilità di attribuire un grado di pertinenza al valore dei termini della parafrasi, soprattutto perché l'esplorazione è estesa anche agli iperonimi delle parole (dei significati) che entrano in gioco. Dato che la regola R0 (sezione 9) ha generato delle catene iperonimiche, queste possono, in teoria, essere esplorate fino al loro estremo.

Il metodo ha una compenente statistica, in quanto la pertinenza di un certo significato rispetto a un altro è misurata sulla co-occorrenza dei termini nelle rispettive definizioni (o in quelle dei loro iperonimi). Ma se si limitasse a questo rilievo, produrrebbe dei risultati anche incongruenti, per esempio considerando compatibili la nozione di "società di credito" le definizioni di diverse locuzioni("denaro sporco", "denaro fresco") citate fra le accezioni di "denaro", a causa della mera co-occorrenza della parola "attività"[22]:

(:ID_disc 10189 :ID-MNG 24487 :LEX "credito" :PARFR "delle società che esercitano tale attività" :TD "
(:ID_disc 11084 :ID-MNG 26537 :LEX "denaro" :PARFR "quello ottenuto attraverso attività illecite" :TD "denaro sporco"
(:ID_disc 11084 :ID-MNG 26539 :LEX "denaro" :PARFR "apporto di capitale a un'attività" :TD "denaro fresco"

Per aumentare le garanzie di correttezza delle co-occorrenze rilevate e per selezionare quelle pertinenti sono introdotte delle verifiche sul contesto sintattico in cui le parole sono inserite. Il punteggio viene incrementato unicamente se il termine in esame condivide, nei due diversi contesti, anche delle caratteristiche strutturali e formali: per esempio: se è implicato nella stessa relazione semantica con l'elemento da cui dipende sintatticamente (il che funziona anche con diatesi diversa, dato che l'interpretazione semantica ha già "normalizzato" le strutture passivizzate) (R...); se il rispettivo governor, nei due diversi contesti, è semanticamente identico o prossimo (R...); se sono rispettate

[22] Per una discussione di questo problema vedi Véronis e Ide 1995.

eventuali restrizioni morfologiche che accompagnano un dato significato della parola in esame (R...).

R1) se $(/b/ \in D^a) \wedge (/c/ \in D^a) \wedge (/c/ \in D^b) \rightarrow$ inc_score(b)
Se nella definizione "b" di un termine /b/ della parafrasi del definendo "a" occorre anche un altro termine /c/ del definendo, allora aumenta il grado di pertinenza del significato "b" nella definizione di "a".

Esempio 1:
gorilla s.m. inv. 1 Grande scimmia africana, con pelle nera ricoperta da pelo grigio scuro....

pelo
1 Appendice epidermica flessibile e filiforme, costituita da sostanza cornea e tessuto connettivale, che si sviluppa sulla pelle dell'uomo e di molti mammiferi

Il significato "Appendice epidermica..." di /pelo/ è confortato dalla presenza di /pelle/ nella sua definizione.

Esempio 2: *accreditare*
b) # [sogg-v-arg-prep.arg] comm. Segnare a credito di qlcu. una somma di denaro

credito
c2) estens. somma cui si ha diritto

somma
c) 2 estens. Quantità complessiva, soprattutto di denaro.

Nella definizione di una delle parole che compongono la definizione del lemma-base appare un altro termine contenuto nella definizione del lemma-base. In questo caso il fenomeno si verifica indipendentemente con due parole diverse: nella definizione di "credito"(c2) appare la parola /somma/ contenuta nella definizione di base; nella definizione di "somma"(c) compare il termine /denaro/ che c'è anche nella definizione di base. "Credito"(2) e "somma"(c) incrementano ambedue il loro punteggio.

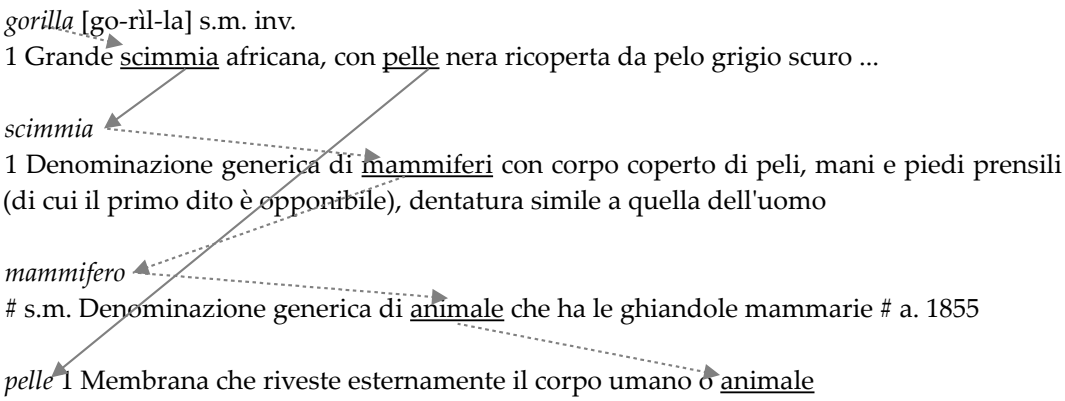
Esempio 3: *pelle:* 1 Membrana che riveste esternamente il corpo umano o animale SIN cute, epidermide

corpo
3 Organismo umano o animale SIN fisico, corporatura

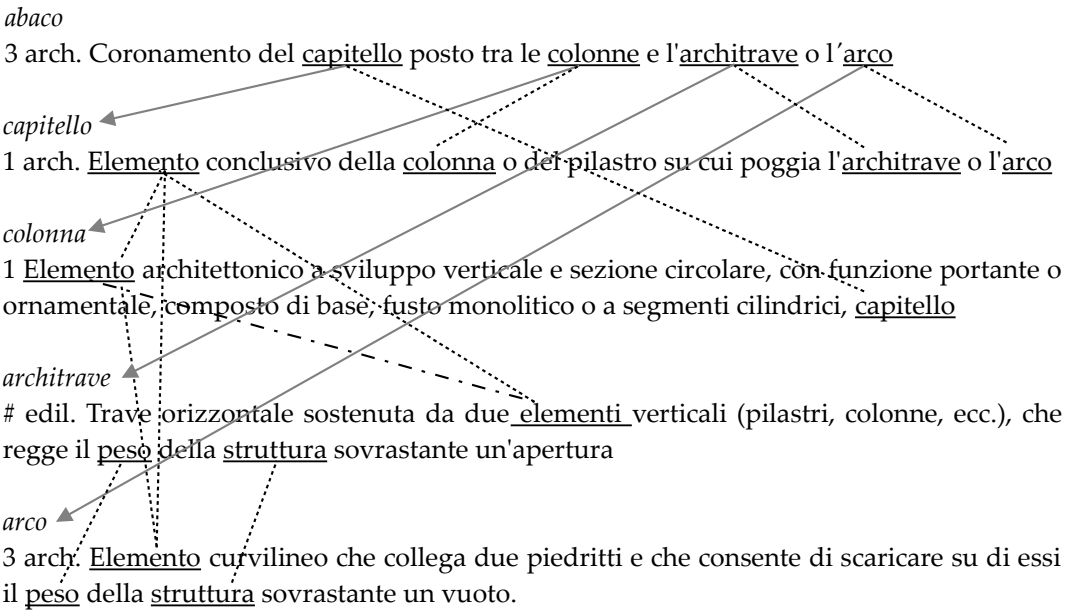
Nell'esempio 3 sono addirittura due (/umano/ e /animale/) le parole della definizione del lemma base ("pelle") che compaiono nella definizione di un terzo termine (/corpo/). L'accezione "corpo"(3) acquista quindi due punti nell'ambito di questa definizione di "pelle".

R2) se $(/b/ \in D^a) \wedge (\text{hyper}(/x/,/a/)) \wedge (/x/ \in D^b) \rightarrow$ inc_score(b)
Se un iperonimo del definendo "a" appare nella definizione di un suo termine /b/, allora aumenta il grado di pertinenza di quella particolare accezione "b" del termine /b/.

Nell'esempio seguente, il significato “pelle”(1) acquista un punto in quanto la sua definizione contiene un iperonimo (di 3° grado) di "gorilla”(1).



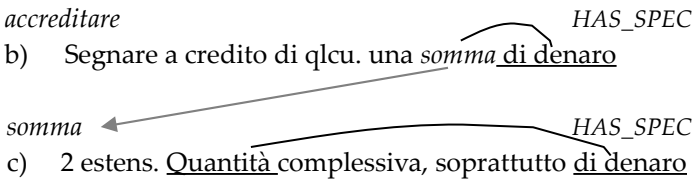
R3) se $(/b/ \in D^a) \wedge (/c/ \in D^a) \wedge (/d/ \in D^b) \wedge (/d/ \in D^c) \rightarrow$ inc_score(b,c)
Confronto fra le definizioni di due termini della definizione-base (select02). Se le rispettive definizioni di due termini del definendo presentano un elemento lessicale comune, il valore di queste due accezioni incrementa il suo grado di pertinenza.



Quattro termini della parafrasi "abaco", /capitello/, /colonna/, /architrave/ e /arco/, contengono, in una delle loro molteplici definizioni, un termine comune (/elemento/): ciò porta a incrementare il punteggio di queste accezioni; in questo caso, ognuna delle accezioni interessate riceve addirittura tre punti, in quanto il termine comune è ribadito da altri tre elementi.
La co-occorrenza del termine /struttura/ (e del modificatore /soprastante/) e quella del termine /peso/ nelle accezioni “architrave” e “arco”(3) porta a un ulteriore incremento del loro punteggio.
Il significato “colonna”(1) riceve inoltre un punto supplementare per la regola R1: il termine /capitello/, della definizione-base, occorre infatti anche dentro la definizione di

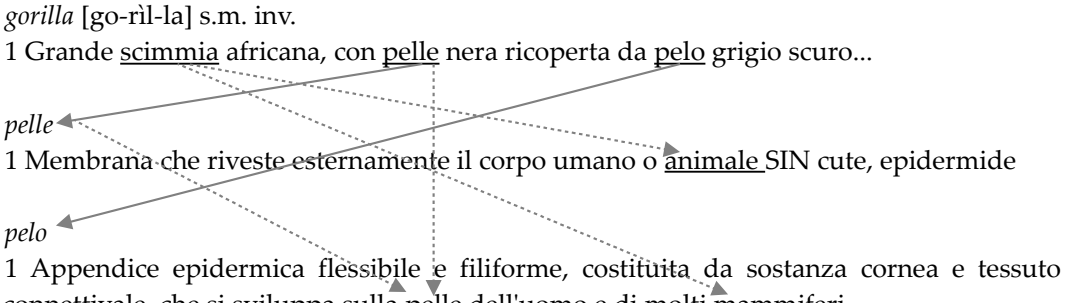
“colonna”(1).
Per la stessa regola R1 il significato “capitello”(1) aumenta di tre punti, dato che nella sua definizione appaiono i termini "colonna", “architrave” e “arco” della definizione-base.
Inoltre le parole che hanno in comune il dominio “architettura” (arch.) incrementano ulteriormente il loro punteggio.

R4) se $(/b/ \in D^a) \wedge (/c/ \in D^a) \wedge (\text{head}(/b'/,D^b)) \wedge (\text{rel}^a(/b/,/c/)) = \text{rel}^b(/b'/,/c/)))$
$\rightarrow$ inc_score(b,c)



Se nella parafrasi del definendo "a" tra due termini /b/ e /c/ intercorre la medesima relazione semantica che si ha nella parafrasi di uno di questi termini /b/, fra la testa /b'/ della parafrasi (che corrisponde a un iperonimo del termine /b/) e il termine /c/, allora l'accezione "b" acquista pertinenza nel contesto del significato "a".

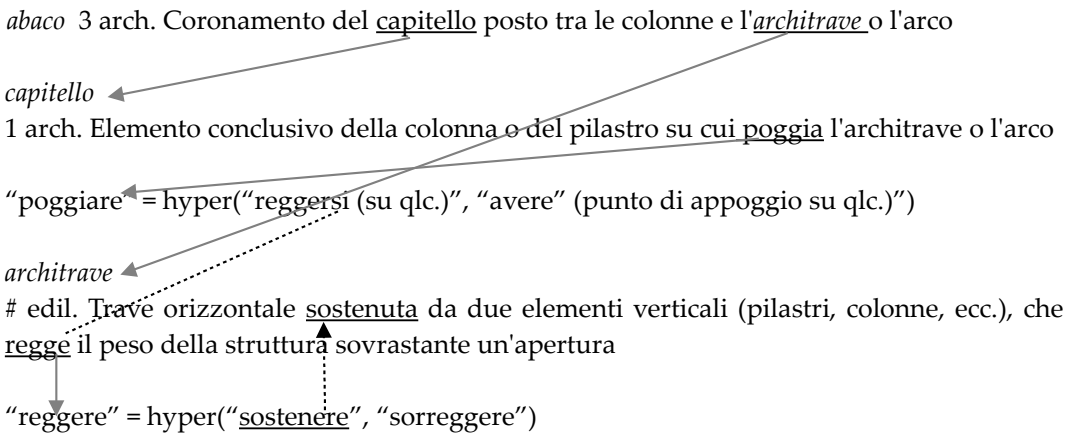
R5) se $(/b/ \in D^a) \wedge (/c/ \in D^a) \wedge (\text{hyper}(/e/,/d/)) \wedge (/e/ \in D^b) \wedge (/e/ \in D^c) \rightarrow$ inc_score(b,c)
Se le definizioni di due termini /b/ e /c/ della parafrasi di "a" hanno in comune l'iperonimo o l'iponimo /e/ di un terzo termine /d/ (virtuale), allora ambedue i termini /b/ e /c/ aumentano il loro grado di pertinenza:



Nell'esempio precedente, due significati, “pelle” (="membrana") e “pelo” (="appendice"), di termini contenuti nella parafrasi di "scimmia", hanno in comune, nelle reciproche definizioni, le parole /animale/ e /mammifero/ legate tra loro da un rapporto di ipo/iperonimia. Ambedue i significati ("pelle" e "pelo") incrementano quindi il loro punteggio nell'ambito della definizione "a". Il fatto, poi, che l'iponimo in comune (/animale/ e /mammifero hanno come iponimo "scimmia") è casualmente il definendo, consente l'applicazione della regola R2 sia all'uno sia all'altro dei significati dati di "pelle" e "pelo", e porta a un ulteriore incremento del loro grado di pertinenza. In questa definizione si applica anche le regola R1 (/pelo/ in questa definizione di /pelle), per "pelo" nel senso di "appendice epidermica" incrementa ulteriormente il suo grado di attendibilità.

R6) se $(/b/ \in D^a) \wedge (/c/ \in D^b) \wedge (\text{hyper}(/d/,/b/)) \wedge (/d/ \in D^c) \rightarrow$ inc_score(b)

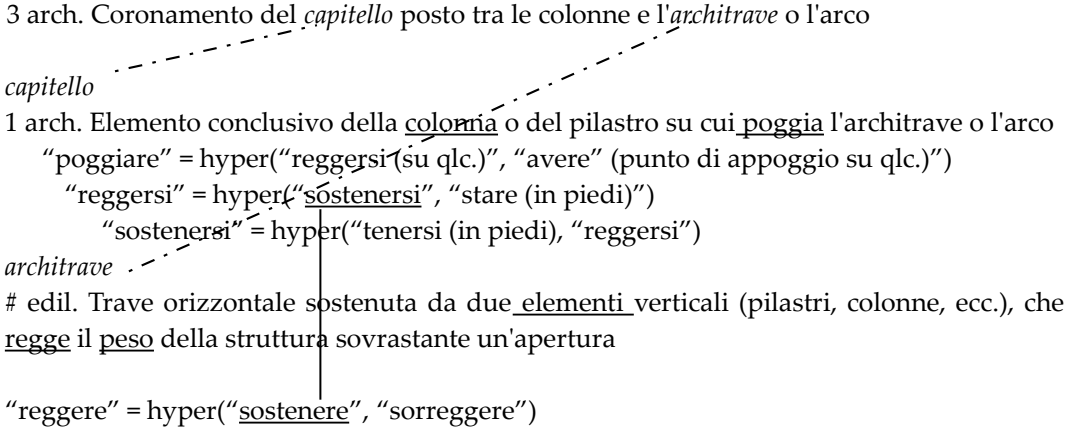
Confronto fra gli iperonimi delle definizioni di due termini della definizione-base:



L'iperonimo ("reggere"), di una particolare accezione del verbo /poggiare/ presente nella definizione di "capitello"(1), occorre nella definizione di "architrave". Questa regola contempla la mera co-occorrenza dei termini, senza badare alla struttura sintattico-semantica in cui sono inseriti, e che in questo caso chiederebbe che si stabilisca l'equivalenza tra il pattern "reggersi"(x,"su"(y)) = reggere(y,x).

R7) se $(/b/ \in D^a) \wedge (/c/ \in D^a) \wedge (/d/ \in D^b) \wedge (/e/ \in D^c)$
$\wedge$ (hyper(/x/,/d/)) $\wedge$ (hyper(/x/,/e/)) $\rightarrow$ inc_score(b,c)
Confronto fra le definizioni di due termini della definizione-base (select02).

Esempio 1
abaco
3 arch. Coronamento del *capitello* posto tra le colonne e l'*architrave* o l'arco



Le parole "capitello" e "architrave" hanno in comune il verbo "sostenere" che è un iperonimo comune ai verbi "poggiare" e "reggere" che compaiono in una delle loro definizioni. Di conseguenza queste due accezioni incrementano il loro punteggio all'interno della definizione di "abaco"(3). Questo punto si aggiunge a quello conseguito in base alla regola R6.

Questa sezione, come ho anticipato (p.) non poteva ambire alla completezza per motivi di spazio. Tornerò su questo argomento in un'altra sede, illustrando anche le regole che

comportano l'applicazione di trasformazioni sintattiche, che comportano un salto di categoria, a partire da correlazioni fra lemmi derivati.

8. Conclusioni

In questo articolo è stato presentato un progetto volto all'estrazione automatica di una rete semantica da un dizionario. L'operazione è motivata dal fatto che la rete semantica va a integrare un database lessicale, estratto dal medesimo dizionario-fonte, che contiene i dati morfologici e sintattici dei medesimi lemmi, indicizzati in base al loro significato. Il database lessicale è quindi una risorsa (contenente circa 105'000 significati per 53'000 lemmi della lingua italiana) internamente coerente, utilizzabile da un parser sintattico e da altri applicazioni di linguistica computazionale (NLP), per es. dei sistemi di *Information Retrieval*.
I dati ricavati dall'analisi computazionale del dizionario hanno anche un interesse lessicografico: consentono di individuare le espressioni e i termini impiegati con maggiore frequenza nelle definizioni e di rivacare dei criteri che ottimizzino la coerenza interna dei dizionari.
L'etichettatura semantica del contenuto delle definizioni date dal dizionario è governata da regole, e si fonda su una preliminare analisi sintattica delle definizioni. Le relazioni sintattiche tra i termini di una data definizione sono interpretate dal sistema e tradotte in relazioni semantiche, dando origine a degli schemi concettuali i cui nodi sono le unità concettuali, cioè le parole che costituisono la definizione. La grande varietà di formulazioni usate nel dizionario è spesso riconducibile alla medesime relazioni semantiche di base. Ciò consente di ridurre l'informazione a dei legami che in si possono considerare in qualche maniera dei lefgami di base o "primitivi". Le definizioni via iperonimia consentono la generazione di catene di significati via via più generali, cioè di tassonomie. Inoltre, dato che gli oggetti sono definiti in base alle loro caratteristiche specifiche, è spesso possibile giungere a uno schema che coincide con una descrizione ontologica.
Gli schemi di significato (*Semantic Frames*) così ricavati dalle definizioni contengono però in un primo tempo solo delle relazioni tra significanti, ognuno dei quali può rinviare a più di un significato. Occorre quindi pertinentizzare i termini nel co-testo di una data definizione. Questa seconda fase del progetto, descritta in maniera solo sommaria in questo articolo, ha quindi l'obiettivo di selezionare i significati congruenti dei termini della definizione. Anche questa seconda fase è in larga parte fondata su regole, dal momento che un metodo meramente statistico si è dimostrato insufficiente. Un prossimo articolo ne fornirà una descrizione più completa.
Chi scrive confida di poter applicare i medesimi criteri e analoghe procedure per collegare tra loro dizionari di lingue diverse, rendendo così la rete semantica una risorsa multilingue.

9. Bibliografia

9.1 Teoria linguistica e semantica (bibliografia essenziale)

CHIERCHIA, G. e McCONNELL-GINET, S., 1993. *Significato e grammatica. Semantica del linguaggio naturale,* Padova, Franco Muzzo Editore. (2000. *Meaning and Grammar.* MIT Press , Cambridge,

Mass.)

CHIERCHIA, G., 1997. *Semantica,* Bologna , Il Mulino

DE MAURO, T., 1989 (1965). *Introduzione alla semantica,* Roma-Bari, Laterza

DE MAURO, T., 1968. *Per una teoria formalizzata del noema lessicale* in DE MAURO 1971: 115-160; (anche in DE MAURO 1989: 235-282), ,

DE MAURO, T., 1971. *Senso e significato. Studi di semantica teorica e storica,* Bari, Adriatica Editrice

ECO, U., 1984. *Semiotica e filosofia del linguaggio,* Torino, Einaudi

FILLMORE, C, 1968, The case f or case, in *Bach, E. - Harms, R.T.* (*eds*.). *Universals in linguistic theory*, Holt, Rinehart & Winston, pp. 1-88.

GREIMAS, A.J., 1996. *Del senso* (orig. *Du sens,* Paris, 1970), Milano, Bompiani

HORSTKOTTE, G., 1982. *Sprachliches Wissen: Lexikon oder Enzyklopädie?,* Bern-Stuttgart-Wien, Verlag Hans Huber

JACKENDOFF, R., 1989. *Semantica e cognizione,* Bologna , Il Mulino (*Semantics and Cognition.* MIT Press, Cambridge, Mass.. 1986).

JACKENDOFF R., 1999. *Semantic Structures*. MIT Press, Cambridge.

KEMPSON, R.M., 1981. *Semantica* , Bologna, Il Mulino.

LABOV, W., 1977. *Il continuo e il discreto nel linguaggio,* Bologna, Il Mulino.

MANZOTTI, E., 1973. *Un nuovo modello di semantica generativa,* "Studi italiani di linguistica teorica e applicata", 1973, n.3, 451-472.

McCAWLEY, J.D., 1968. *The role of semantics in a grammar*. In *Universals in Linguistic Theory,* E. Bach and R. Harms (eds.), 124-169. New York: Holt, Rinehart.

MEL'CUK I.A., 2012. *Semantics. From Meaning to Text*.

PRANDI M., 2004 *The Building Blocks of Meaning*. *Ideas for a Philosophical Grammar.* John Benjamins ed. Amsterdam/ Philadelphia.

PRIETO, L.J., 1967. *Principi di noologia. Fondamenti della teoria funzionale del significato,* Roma, Ubaldini

PRIETO, L., 1976. *Pertinenza e pratica. Saggio di semiotica,* Milano, Feltrinelli

PUSTEJOVKSY. J., 1995. *The Generative Lexicon.* MIT Press, Cambridge

SEARLE, J.R., 1980. *Minds, Brains and Programs. The Behavioral and the Brain Sceinces,* in TONFONI, G. (a cura di), 1984.

SEARLE, J.R., 1994. *La riscoperta della mente,* Torino, Bollati Boringhieri.

TAYLOR, J.T., 1996. *L'incomprensione linguistica,* Roma-Bari, Laterza.

VIOLI, P., 1997. *Significato ed esperienza,* Milano, Bompiani.

WITTGENSTEIN, L., 1974. *Ricerche filosofiche,* Torino, Einaudi.

WITTGENSTEIN, L., 1990. *Grammatica filosofica,* Firenze, La Nuova Italia.

9.2 Semantica computazionale

ALLEN, S. e PETÖFI, J.S. (eds.), 1979. *Aspects of automatized text processing,* Papiere zur Textlinguistik, Hamburg, Buske

AGIRRE E., RIGAU G., 1995. A Proposal for Word Sense Disambiguation using Conceptual Distance. *Proceedings of the International Conference on Recent Advances in Natural Processing. Velingrad*

AGIRRE E., RIGAU G., 1996. A Proposal for Word Sense Disambiguation using Conceptual Density. *Proceedings of the International Conference on Recent Advances in Natural Processing. Velingrad*

AGIRRE E. and EDMONDS Ph. G. (eds), 2007, Word sense disambiguation: algorithms and applications, Springer.

ARTALE A. , MAGINI B., STRAPARAVA C., 1997. *Lexical Discrimination with the Italian Version of Wordnet.*

BAKER C.F., FILLMORE CH.J, CRONIN B., 2003. *The Structure of the Framenet Database*, "International Journal Lexicography" 16 (3): 281-296.

BARNBROOK G., DANIELSSON P., MAHLBERG,M. (eds.), 2004) *Meaningful Texts: The Extraction of Semantic Information from Monolingual and Multilingual Corpora* , London - New York: Continuum International Publishing Group Ltd.
BARONI M. and LENCI A., 2010. *Distributional Memory: A general framework for corpus-based semantics.* Computational Linguistics 36 (4): 673-721.

BARONI M. and ZAMPARELLI R, 2010. *Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space*. *Proceedings of the Conference on Empirical Methods in Natural Language Processing,* 2010, East Stroudsburg PA: ACL, 1183-1193

BENTIVOGLI L., GIRARDI C., PIANTA E., 2003., *The MEANING Italian Corpus.* Corpus Linguistics 2003 Conference, 2003, pp. 103-112. (Corpus Linguistics 2003 Conference, Lancaster, United Kingdom)

BINDI R., CALZOLARI N., MONACHINI N., PIRELLI V., ZAMPOLLI A., 1994. *Corpora and computational lexica. Integration of different methodologies of lexical knowledge acquisiton*. "Literary and Linguistic Computing", 9, 1 (1994), 29-46.

BYRD R.J., CALZOLARI N., CHODROW S., KLAVANS J., NEFF M.S. and RIZK O.A., *Tools And Methods For Computational Lexicology.* "Computational Linguistics", Volume 13, Numbers 3-4, July-December 1987

FELLBAUM C., DELFS L., WOLFF S., PALMER M., 2004. *Word meaning in dictionaries, corpora and the speaker's mind*. In Barnbrook, Danielsson, Mahlberg (eds.), *Meaningful Texts: The Extraction of Semantic Information from Monolingual and Multilingual Corpora*, London - New York: Continuum

International Publishing Group Ltd.: 31-38

GRANGER S., PAQUOT Magali., 2012. *Electronic Lexicography.* Oxford University Press.

LESK, M., 1986. Automated Sense Disambiguation Using Machine-readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone. Proceedings of the 1986 SIGDOC Conference.

MAGNINI B., STRAPPARAVA C., (1994), *Costruzione di una base di conoscenza lessicale per l'italiano basata su ItalWordNet*, Atti del XXVIII Congresso della Società di Linguistica Italiana, Palermo, 415-418

MAGNINI B., STRAPPARAVA C., PEZZULO C., GHIOZZO A., 2003. *The Role of Domain Information in Word Sense Disambiguation*. Journal of Natural Language Engineering (on Sensval-2), 9.

MARKOWITZ, J., AHLSWEDE, T., EVENS, M., 1986. *Semantically significant patterns in dictionary definitions*. ACL Conference Proceedings, 112-119.

MCCARTHY D. AND CARROLL J., 2003. Disambiguating nouns, verbs, and adjectives using automatically
acquired selectional preferences. *Computational Linguistics*, 29(4):639–654, December.

MILLER G.A. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

PHAM N., BERNARDI R., ZHANG Y.-Z. Zhang and BARONI,M.. 2013. S*entence paraphrase detection: When determiners and word order make the difference. Proceedings of the Towards a Formal Distributional Semantics Workshop at IWCS 2013*, East Stroudsburg PA: ACL: 21-29

PUSTEJOVSKY J. (1991). *The Generative Lexicon*, "Computational Linguistics", 17: 71-77

PUSTEJOVSKY J., BERGLER S., 1992. *Lexical semantics and knowledge representation.* First SIGLEX Workshop, Berkeley, CA, USA, June 1991. Springer.

PUSTEJOVSKY J., (ed.), 1993. *Semantics and the Lexicon*. MIT Press, Cambridge, MA.

PUSTEJOVSKY J., ANICK P., BERGLER S., *Lexical semantic techniques for corpus analysis.* Computational Linguistics - Special issue on using large corpora: II archive Volume 19 Issue 2, June 1993, pages 331-358
MIT Press Cambridge, MA, USA

PUSTEJOVKSY. J., 1995. *The Generative Lexicon.* MIT Press, Cambridge

RESNIK, P. (1997) Selectional Preference and Sense Disambiguation, *In Proceedings of ACL Siglex Workshop on Tagging Text with Lexical Semantics, Why, What and How?*, Washington, 1997.

SAHLGREN M., 2008. *The Distributional Hypothesis. From context to meaning: Distributional models of the lexicon in linguistics and cognitive science*. Rivista di Linguistica, 20,1: 33-53 .

SINCLAIR J., BALL J. (1996). *EAGLES Preliminary Recommendations onText Typology.* (url).

VÉRONIS J., NANCY IDE. N., 1995. *Large Neural Networks for the Resolution of Lexical Ambiguity*. In *Computational Lexical Semantics* (P. Saint-Dizier ed.) .Cambridge University Press, 1995.

VOSSEN P., 1992. *The automatic construction of a knowledge base from dictionaries: a combination of techniques*, in: H. Tommola, K. Tarantola, T. Salmin Tolonen, J. Schopp (eds) *Proceedings of the 5th*

Euralex International Congress on Lexicography, Tampere, Finland, 1992.

VOSSEN P., et. al. 1998, *The EuroWordNet base concepts and top ontology.* Final Version (http://www.researchgate.net/publication/228594694_The_eurowordnet_base_concepts_and_top_ontology/file/79e4150604a1a45d3d.pdf&sa=X&scisig=AAGBfm1N879XyUTOujqoP5Wd8EcIs2JcWg&oi=scholarr&ei=Mu1vUcqNC8rb7AbLtoDwDw&ved=0CC0QgAMoADAA)

WILKS, Y., CHARNIAK, E. (eds)., 1976 *Computational Semantics. An Introduction to Artificial Intelligence and Natural Language Understanding*. Amsterdam: North-Holland

WILKS, Y., D. FASS, C. GUO, J. MACDONALD, T. PLATE, B. SLATOR, 1990. *Providing Machine Tractable Dictionary Tools*. In J. PUSTEOVSKY (ed.), *Semantics and the Lexicon*. MIT Press, Cambridge, MA.

WILKS, Y., SLATOR, B., GUTHRIE, L. (1996) *Electric Words: dictionaries, computers and meanings*. Cambridge, MA: MIT Press.

WILKS, Y., BREWSTER, C., 2009. *Natural Language Processing as a Foundation of the Semantic Web*. Now Press: London

ZAMPOLLI A., CAPPELLI A., 1981. (eds.), *The possibilities and limits of the computer in producing and publishing dictionaries*. Proceedings of the European Science Foundation Workshop, Pisa, 1981, Istituto di Linguistica Computazionale - Giardini, Roma - Pisa, 1984, pp. 77-82. (Linguistica Computazionale, 3, 1983).

Annesso 1: ***Semantic Frames***

Schemi semantici delle definizioni di "scimmia", "mammifero", "vertebrato"

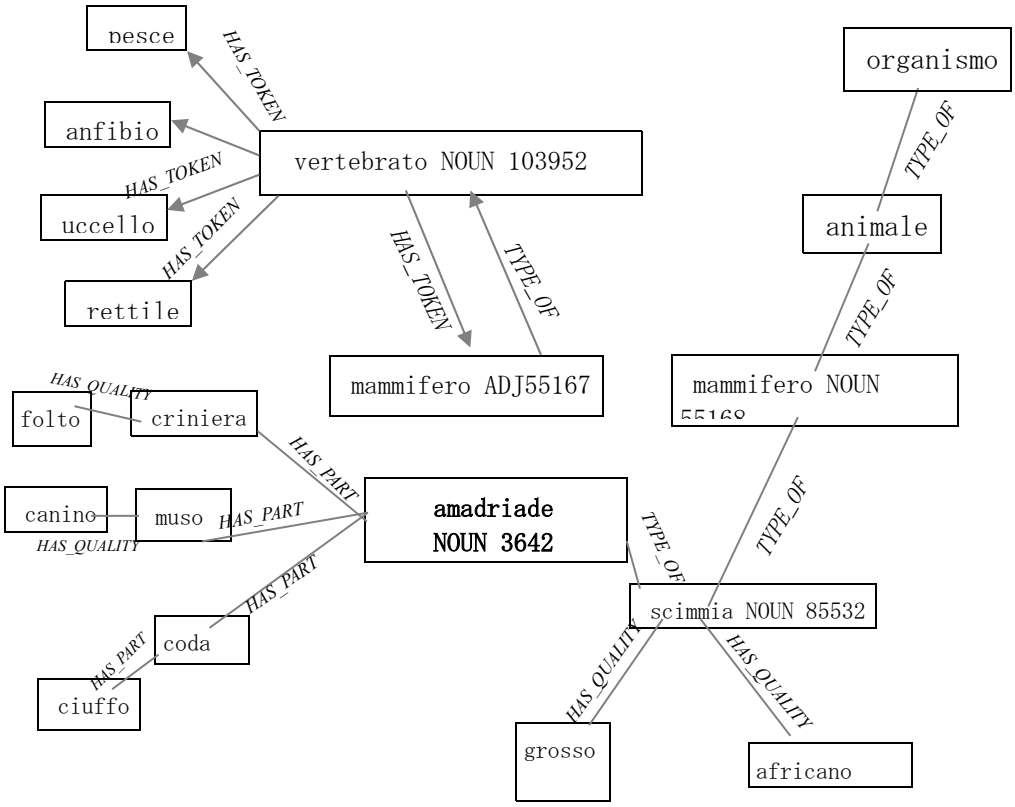
42.0	"scimmia"	NOUN85532 [TYPE_OF(THING,"scimmia"), TYPE_OF(42.1,NAME_OF,"denominazione"), TYPE_OF(42.3,THING,"mammifero"), HAS_QUALITY(42.2,QUALITY,"generico"), HAS_PART(42.4,THING,"corpo")]
42.1	"denominazione"	[TYPE_OF(NAME_OF,"denominazione"), HAS_QUALITY(42.2,QUALITY,"generico"), HAS_PART(42.4,THING,"corpo")] [HAS_PART(42.8,THING,"piede")] HAS_PART(42.10,THING,"dentatura")]
42.2	"generico"	[TYPE_OF(QUALITY,"generico")]
42.3	"mammifero"	[TYPE_OF(THING,"mammifero")]
42.4	"corpo"	[TYPE_OF(THING,"corpo"), HAS_QUALITY(42.5,QUALITY,"coperto"),
42.5	"coperto"	[TYPE_OF(QUALITY,"coperto")]]
42.6	"pelo"	[TYPE_OF(THING,"pelo")]
42.7	"mani"	[TYPE_OF(THING,"mani")]
42.8	"piede"	[TYPE_OF(THING,"piede"), HAS_QUALITY(42.9,QUALITY,"prensile")]
42.9	"prensile"	[TYPE_OF(QUALITY,"prensile")]]
42.10	"dentatura"	[TYPE_OF(THING,"dentatura"), HAS_QUALITY(42.11,QUALITY,SIMILARiTY,"simile")]
42.11	"simile"	[TYPE_OF(QUALITY,SIMILARiTY,"simile"), HAS_SPEC(42.12,THING,"quella")]]
42.12	"quella"	[TYPE_OF(THING,"quella"), HAS_SPEC(42.13,THING,PERSON,"uomo")]
42.13	"uomo"	[TYPE_OF(THING,PERSON,"uomo")]

27.0	"mammifero"	NOUN55168 [TYPE_OF(THING,"mammifero"), TYPE_OF(27.1,NAME_OF,"denominazione"), TYPE_OF(27.3,THING,ANIMAL,"animale"),] TYPE_OF(THING,ANIMAL,"animale" 4770 0), TYPE_OF(2.1,THING,"organismo" 4770 0), TYPE_OF(3.1,THING,"bestia" 4771 [TYPE_OF(THING,ANIMAL,"animale"), HAS_QUALITY(27.4,ACTION,HAS_PART,"avere")]
27.4	"avere"	[TYPE_OF(ACTION,HAS_PART,"avere"), HAS_OBJ/PATIENT(27.5,THING,"ghiandola"), PART_OF(27.7,THING,"ghiandola"), HAS_PART,HAS_TOKEN(27.9,THING,"che")]
27.5	"ghiandola"	[TYPE_OF(THING,"ghiandola"), HAS_QUALITY(27.6,QUALITY,"mammario")]
27.6	"mammario"	[TYPE_OF(QUALITY,"mammario")]

```
57.0   "vertebrato"      NOUN 103952
                         [TYPE_OF(THING,"vertebrato" ),
                         TYPE_OF(57.2,THING,ANIMAL,"animale" )]
                         [DOMAIN("zoologia","","","scienze" )]
57.1   "Tipo"            [TYPE_OF(THING,ATTRIBUTION,TYPE,"tipo" ),
                         TYPE_OF(57.2,THING,ANIMAL,"animale" ),
                         HAS_TOKEN(57.16,THING,PERSON,"uomo" ),
57.2   "animale"         [TYPE_OF(THING,ANIMAL,"animale" ),
                         HAS_PART(57.5,THING,"scheletro" )]
                         AS_PART(57.5,THING,"scheletro" )]
57.5   "scheletro"       [TYPE_OF(THING,"scheletro" ),
                         HAS_QUALITY(57.6,QUALITY,"interno" ),
                         HAS_QUALITY(57.7,ATTRIBUTION,SPACE,"coincidere" ),
                         SPACE_OF(57.8,THING,"asse" )]
57.6   "interno"         [QUALITY(QUALITY,"interno" )]
57.7   "coincidere"     [TYPE_OF(ATTRIBUTION,SPACE,"coincidere" ),
                         SPACE_OF(57.8,THING,"asse" ),
                         HAS_SPACE(57.9,THING,"colonna" ),
                         HAS_SPACE(57.13,THING,"spina" )]
57.8   "asse"           [TYPE_OF(THING,"asse" )]
57.9   "colonna"        [TYPE_OF(THING,"colonna" ),
                         HAS_QUALITY(57.10,QUALITY,"vertebrale" ),
                         HAS_QUALITY(57.11,QUALITY,"spina" )]
57.10  "vertebrale"     [QUALITY(QUALITY,"vertebrale" )]
57.11"spina"             [TYPE_OF(THING,"spina" ),
                         HAS_QUALITY(57.12,QUALITY,"dorsale" )]
57.12  "dorsale"        [QUALITY(QUALITY,"dorsale" )]
57.15  "appartenere"    [TYPE_OF(ACTION,ATTRIBUTION,TOKEN,"appartenere"          ),
                         HAS_TOKEN(57.16,THING,PERSON,"uomo" )]
57.16  "uomo"           [TYPE_OF(THING,PERSON,"uomo" )]
```

```
58.0   "vertebrato"      NOUN103953
                         [TYPE_OF(THING,"vertebrato" ),
                         HAS_PART,TOKEN_OF(58.2,THING,"mammifero" ),
                         HAS_PART,TOKEN_OF(58.3,THING,"rettile" ),
                         HAS_PART,TOKEN_OF(58.4,THING,"uccelli" ),
                         HAS_PART,TOKEN_OF(58.5,THING,"anfibi" ),
                         HAS_PART,TOKEN_OF(58.6,THING,"pesce" )]]
```

Annesso 2: tassonomie generate



Annesso 3.

Diramazione dei significati dei termini che definiscono il lemma "abaco"

abaco [à-ba-co] s.m. (pl. -chi)
1 *Tavoletta per calcoli in uso nell'antichità e nel Medioevo;*
- *estens. aritmetica, calcolo*

2 *mat. Metodo grafico per rappresentare i valori di una funzione di più variabili: a. cartesiano*
3 *arch. Coronamento del capitello posto tra le colonne e l'architrave o l'arco # sec. XIII*