Verso un lessico computazionale aperto per la lingua italiana

Il progetto Senso Comune

Guido Vetere
Associazione Senso Comune
Via Nomentana 56
Roma, Italia
guido.vetere@sensocomune.it

ABSTRACT

La mancanza di un lessico computazionale aperto per la lingua italiana limita l'accesso alle risorse informative in rete nel nostro Paese. Questo riguarda anche il patrimonio informativo della Pubblica Amministrazione, la cui accessibilità è da molti anni indicata come una delle priorità per migliorare il rapporto tra cittadini e amministrazioni.
Senso Comune è un progetto per la costituzione di un lessico computazionale aperto della lingua italiana. Il progetto è sviluppato da un'associazione senza fini di lucro. Il lessico sarà acquisito attraverso una piattaforma cooperativa e distribuito in formato aperto.

1. INTRODUZIONE

Le parole della lingua naturale sono la principale chiave di accesso alle risorse informative digitalizzate. Parole contenute nei testi, usate per descrivere dati e fissare riferimenti, parole che significano in genere cose ben precise, sia pure in una grande varietà di contesti linguistici e culturali. Tuttavia ancora oggi, per la maggior parte dei sistemi informativi, le parole sono semplici sequenze di caratteri alfabetici, e non veicoli di significato. E' noto che questo limita notevolmente la reperibilità delle risorse disponibili in rete, e questo vale anche, naturalmente, per l'accesso del cittadino al patrimonio informativo della Pubblica Amministrazione, per gli scambi tra Amministrazioni, o per la gestione dei documenti nei singoli Enti.
Le parole incontrano i loro significati per lo più in modo obliquo e tangenziale: usiamo costruzioni diverse per riferirci alla stessa cosa, e, per converso, ciascuna parola può significare cose assai differenti a seconda del contesto in cui viene scambiata. I linguaggi specialistici, come quelli della Pubblica Amministrazione, pur cercando di limitare e disciplinare tale complessità, non possono tuttavia eliminarla, e dunque finiscono per moltiplicare i "giochi linguistici" in cui le parole sono coinvolte. Nasce dunque l'esigenza, per chi sviluppa sistemi informativi, di catturare in qualche modo l'incontro tra l'espressione linguistica e il suo significato, ciò a cui ci si riferisce in genere con termine *semantica*. Da questa esigenza si è andata sviluppando la visione del cosiddetto *Semantic Web*[1], insieme di ricerche, standard e tecnologie in grado oggi di indicare soluzioni certamente parziali (e talvolta ingenue), ma già utili per il conseguimento di progressi nel trattamento concreto delle risorse informative. Sistemi complessi e di grande rilevanza sociale come quelli della Pubblica Amministrazione potrebbero dunque consentire ai loro utenti un accesso più fluido, trasparente e intelligente alle risorse informative se queste fossero corredate da annotazioni formali riguardanti il loro significato.
Modellare la semantica di una lingua, cioè il rapporto tra le espressioni del linguaggio (*significanti*) ed un'appropriata rappresentazione del loro contenuto concettuale (*significati*) è notoriamente complesso e controverso [5]. Muovendo dal versante dell'espressione, il dizionario tradizionale cerca solitamente di essere esaustivo per quanto concerne sfumature di senso ed esempi d'uso, con ricco corredo di informazione lessicale, grammaticale ed etimologica. Tuttavia, esso affida la specificazione del significato a descrizioni linguistiche informali, fatte di parole, cioè dello stesso tipo di oggetti che si intende definire. Questo li rende pressoché inutili come supporto nel trattamento della lingua da parte di sistemi automatici [4]. D'altro canto, sul versante dei modelli concettuali formali che si propongono come teorie logiche del contenuto, detti correntemente *ontologie*, oggi diffusi anche attraverso linguaggi logico-descrittivi standardizzati (ad es. OWL [2]), si può osservare come tali risorse siano solitamente molto carenti sul piano della lessicalizzazione.
Un approccio più specifico alla modellazione della semantica del linguaggio naturale è contenuto in lessici computazionali come WordNet [3] [8] per l'inglese ed il suo corrispettivo multilingue EuroWordNet [13]. In queste risorse, i sensi linguistici sono chiaramente enucleati e messi reciprocamente in rapporto mediante relazioni semantiche (es. sinonimia, antonimia, iponimia, meronimia). In genere, si riconosce una corrispondenza tra gruppi di sinonimi (detti *synset*) e concetti (cosiddetti) ontologici, cioè proprietà degli enti nel dominio di riferimento, e corrispettivamente si tende ad assimilare le relazioni semantiche a relazioni logiche (inclusione, disgiunzione) ovvero ontologico-formali (parte). In questo modo,

[1] http://www.w3.org/2001/sw/
[2] http://www.w3.org/TR/owl-features/
[3] http://wordnet.princeton.edu/

si giunge a modelli integrati ontologico-linguistici come ad esempio OntoWordNet [11]. Questo tipo di risorse sono in grado di offrire supporto ad applicazioni di *information retrieval* semantici, alimentati anche attraverso sistemi di estrazione di informazione basati su analisi molto sofisticate del linguaggio naturale. Tuttavia, per la loro complessità, le risorse ontologico-linguistiche sono scarse e generalmente carenti. Inoltre, mentre il mondo anglofono può giovarsi di WordNet come una risorsa aperta e gratuita[4], i lessici computazionali per l'italiano sono proprietari e non disponibili per la collettività.
Da queste considerazioni nasce il progetto Senso Comune [5]. L'iniziativa, portata avanti da un'associazione senza fini di lucro fondata da un gruppo studiosi di logica, ontologia e linguistica, partecipata da una larga comunità scientifica, presieduta dal Prof. Tullio De Mauro e sostenuta dalla Fondazione IBM Italia, si propone di costruire una base di conoscenza linguistica della lingua italiana da rendere disponibile come in formato aperto [6]. Essa consisterà in un lessico computazionale con adeguate strutture informative per rappresentare in modo formale, e dunque comprensibile alla macchina, le complesse conoscenze sottostanti alla lingua. Raccogliere e organizzare tale conoscenza, e sviluppare metodi di ragionamento su di essa, può essere un fattore determinante per l'evoluzione del Web italiano verso un sistema più ricco, funzionale ed accessibile. La Pubblica Amministrazione, in particolare, potrà giovarsi di questa risorsa per facilitare l'accesso alla propria informazione e migliorare il proprio grado di integrazione.
Come procedere all'acquisizione e alla sistemazione di una conoscenza così vasta e complessa come quella della lingua? In sintesi, l'idea, che illustreremo nel presente articolo, è quella di partire dal lessico di base dell' italiano, portarlo ad un adeguato livello di formalizzazione, costruire la sua controparte semantica sulla base di un'ontologia fondamentale, ed infine aprire la risorsa al contributo dell'intera comunità dei parlanti.

2. IL MODELLO DI SENSO COMUNE

Senso Comune si basa su un modello per la rappr4sentazione integrata di informazione lessicale (forme linguistiche, accezioni, relazioni lessicali) ed informazione ontologica (classi, attributi, relazioni logiche). Il modello è diviso in elementi (vedi Fig. 1) tra loro correlati, alla base dei quali si pone un *metamodello* che definisce il linguaggio logico-descrittivo con il quale i modelli sono redatti. Nelle sezioni che seguono illustreremo alcune tra le caratteristiche di ciascun elemento.

2.1 Metamodello

Il *metamodello* di Senso Comune è la logica descrittiva DL-Lite [2]. Rispetto agli scopi applicativi si Senso Comune, è stato analizzato che tale logica offre un rapporto ottimale tra espressività e computabilità. Il linguaggio di modellazione UML [7] (in particolare, Class Diagram) è stato adottato come sintassi concreta diagrammatica per lo sviluppo del modello, sulla base di una relazione formale nota tra questo e DL-Lite

[4]il permesso di usare, copiare, modificare, e distribuire sia il software sia il database è garantito a tutti, a titolo gratuito, per qualsiasi scopo
[5]www.sensocomune.it
[6]è allo studio una forma di licenza Creative Commons
[7]http://www.uml.org/

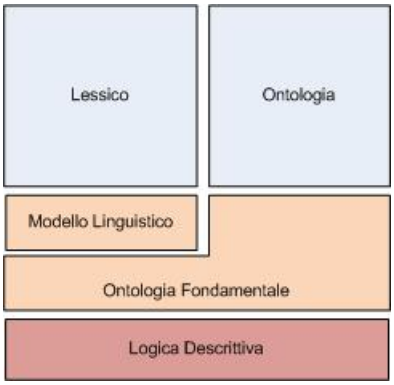

Figure 1: Il modello di Senso Comune

(Tabella 1).
Essenzialmente, DL-Lite è una logica descrittiva trattabile per scrivere ontologie e interrogare basi di conoscenza con efficienza paragonabile e quella del DBMS relazionali. Per ottenere tale efficienza, rispetto ad altri linguaggi di specificazione di ontologie, DL-Lite limita l'uso di costrutti come ad esempio la quantificazione universale, la disgiunzione e l'enumerazione. D'altro canto, è dimostrato [3] che tali costrutti non sono in pratica utilizzabili per accedere con precisione ai volumi di dati caratteristici delle basi di dati documentali, che rappresenta lo scenario d'uso più tipico per una risorsa come Senso Comune.
Come per ogni logica descrittiva, alla base di DL-Lite vi sono costrutti per rappresentare *Concetti* (classi) e *Ruoli* (relazioni binarie), l'inclusione, la quantificazione esistenziale sui ruoli e la negazione. Inoltre, si introduce una struttura sintattica per restringere le capacità espressive del linguaggio. Questa si basa sulla distinzione tra:

AtomicConcept : concetti atomici (A)

BasicConcept : concetti di base (B)

GeneralConcept : concetti generali (C)

AtomicRole : ruoli atomici (P)

BasicRole : ruoli di base (Q)

GeneralRole : ruoli generali (R)

ValueDomain : dominio di attributo (D)

Tali elementi sono legati dalle seguenti corrispondenze:

- Concetti:

$$B \leftarrow A \mid \exists R$$
$$C \leftarrow B \mid \neg B$$

- Ruoli:

$$Q \leftarrow P \mid P^-$$
$$R \leftarrow R \mid \neg R$$

dove il costrutto P^- è utilizzato per rappresentare *l'inverso* di un ruolo atomico. Dei ruoli, inoltre, si può indicare la

funzionalità (cardinalità = 1).

DL-Lite permette di definire *assiomi di inclusione* della forma:

$$B \sqsubseteq C \quad Q \sqsubseteq R$$

ovvero è possibile specificare relazioni di inclusione che coinvolgono concetti (ruoli) base a sinistra e concetti (ruoli) generali a destra. Tale limitazione è cruciale per assicurare la trattabilità delle basi di conoscenza DL-Lite.

Gli *assiomi di appartenenza* a concetti e ruoli degli oggetti quantificati si esprimono nella forma usuale:

$$A(a) \quad D(a) \quad P(a,b)$$

La semantica formale di DL-Lite, infine, è definita con una struttura del primo ordine in modo del tutto analogo alle logiche descrittive standard [1].

2.2 Ontologia

Le risorse ontologico-linguistiche di origine lessicografica come WordNet sono in genere costruite a partire dall'analisi del linguaggio. La struttura tassonomica fondamentale di queste risorse consiste in una gerarchia di iponimia ottenuta mediante l'indagine del lessico. Questo in genere fa sì che ai concetti (*synset*) di massima generalità si giunga senza tener conto di alcuna distinzione categoriale di natura ontologica. La sistemazione ontologica di tali risorse viene eventualmente effettuata a posteriori, come nel caso di OntoWordNet [11].
Senso Comune parte da una diversa impostazione. Un piccolo numero di concetti è assunto *a priori* come struttura ontologica fondamentale rispetto alla quale tutte le nozioni semantiche della risorsa vengono a definirsi. Questa ontologia di base deriva da DOLCE [10] e comprende attualmente circa trenta concetti e venti relazioni binarie. Di seguito le principali categorie.

Entity ($\in$ **AtomicConcept**): concetto di massima generalità.

Concrete ($\sqsubseteq$ **Entity**) : entità qualificate nello spazio-tempo (es. oggetti, eventi).

Abstract ($\sqsubseteq$ **Entity**) : entità non qualificate nello spazio-tempo (es. proposizioni).

Object ($\sqsubseteq$ **Concrete**) : entità concrete che presentano una unità essenziale ed un'esistenza autonoma; non hanno parti temporali anche se le loro proprietà possono cambiare nel tempo (es. una nave).

Event ($\sqsubseteq$ **Concrete**) : entità concrete estese nel tempo, che possono avere parti temporali (es. una corsa).

Quality ($\sqsubseteq$ **Entity**) : aspetti percepibili delle entità concrete, non facenti parte delle stesse ma da esse esistenzialmente dipendenti (es. un colore).

Tra queste categorie ontologiche e le *parti del discorso* dell' italiano (sostantivo, verbo, aggettivo, etc) vi è una relazione complessa. Benché non si possa stabilire tra queste una sicura corrispondenza, tuttavia la relazione tra parti del discorso e categorie ontologiche sarà presa in considerazione come una delle euristiche per la collocazione dei sensi linguistici rispetto all'ontologia fondamentale.

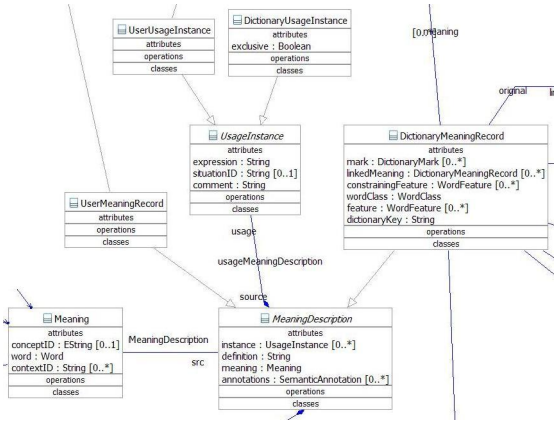

Figure 2: Modello linguistico: significati

2.3 Lessico

Il modello per la rappresentazione di informazione lessicale di Senso Comune è una estensione dell'ontologia di base consistente in un insieme di concetti astratti che rappresentano nozioni lessicografiche. Nel corso dell'analisi, è emersa la necessità di rappresentare sia la struttura lessicografica tipica del dizionario tradizionale, sia strutture specifiche per la raccolta di informazione linguistica da utenti non specialisti, e di conciliare le due prospettive. Questo ha portato ad un modello di rappresentazione più esteso e per certi versi più complesso di quelli conosciuti allo stato dell'arte, come il Lexical Markup Framework [9] attualmente in corso di standardizzazione presso ISO[8]. In ogni caso, le parti essenziali del modello lessicale di Senso Comune potranno facilmente essere messe in corrispondenza con LMF.
Oltre a fornire l'apparato di rappresentazione per le strutture morfologiche, il modello del lessico rappresenta *significati* e *relazioni tra significati*.

2.3.1 Significati

Il diagramma delle classi mostrato in Fig.2 mostra il modo in cui sono modellati i significati delle parole.

Meaning ($\sqsubseteq$ **Abstract**) : è la forma reificata della fondamentale relazione di significazione, indipendente da qualsiasi descrizione (vedi **MeaningDescription**). La relazione di significazione associa una parola (o espressione poliremática) al concetto in una ontologia e (possibilmente) ai contesti entro i quali la significazione si realizza.

MeaningDescription ($\sqsubseteq$ **Abstract**) : è la descrizione di un significato. Contiene una frase di definizione, un insieme di esempi d'uso, un insieme di annotazioni semantiche.

[8]International Organization for Standardization, http://www.iso.org

Table 1: UML e DL-Lite

UML	DL-Lite
Class	A
Association, Attribute ($\neq PrimitiveType$)	P,P^-
Attribute (PrimitiveType)	D
InstanceSpecification	$A(a)$
LiteralString	$D(d)$
Slot (definingFeature.type $\neq PrimitiveType$)	$P(a,b)$
Slot (definingFeature.type $= PrimitiveType$)	$D(a,b)$
Generalization	$B \sqsubseteq C$
cardinality = 1	$funtc(P)$

UserMeaningRecord ($\sqsubseteq$ MeaningDescription) : rappresenta le descrizioni di significati nella forma specificata dagli utenti.

DictionaryMeaningRecord ($\sqsubseteq$ MeaningDescription) : rappresenta le descrizioni dei significati nella forma lessicografica caratteristica del dizionario.

UsageInstance ($\sqsubseteq$ Abstract) : rappresenta gli esempi d'uso di un determinato significato legati ad una particolare descrizione.

UserUsageInstance ($\sqsubseteq$ UsageInstance) : rappresenta gli esempi d'uso specificati dagli utenti.

DictionaryUsageInstance ($\sqsubseteq$ UsageInstance) : rappresenta gli esempi d'uso nella forma caratteristica del dizionario.

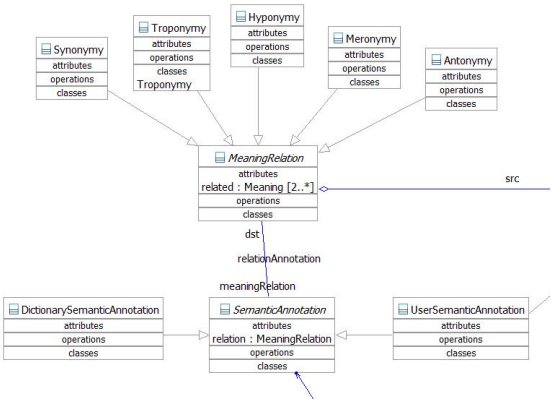

Figure 3: Modello linguistico: relazioni

Si noti come la classe **Meaning** rappresenti l'accezione linguistica come associazione tra espressione e contenuto concettuale. Quest'ultimo è rappresentato dall'identificativo di uno (ed un solo) concetto ontologico, così da consentire la definizione di una funzione:

$$\sigma : Meaning \rightarrow Concept$$

In particolare, σ non è iniettiva (diversi **Meaning** possono riferirsi allo stesso **Concept**), né suriettiva (non tutti i concetti devono avere controparte lessicale), né comunque totale (può non essere definita per alcuni valori del dominio). Tutto ciò che si richiede è che, se definito, il contenuto concettuale di un'accezione sia univoco.

2.3.2 *Relazioni lessicali*

Il diagramma delle classi mostrato in 3 rappresenta le relazioni binarie tra significati. In particolare le relazioni prese in considerazione sono : sinonimia, troponimia, iponimia, antonimia e meronimia. Le classi componenti sono:

MeaningRelation ($\sqsubseteq$ Abstract) : è la reificazione della relazione tra significati. Contiene un riferimento a ciascuno dei significati coinvolti nella relazione.

Synonymy ($\sqsubseteq$ MeaningRelation) : rappresenta la sinonimia tra significati. Due significati si definiscono sinonimi se sono percepiti con la stessa valenza semantica in tutti i contesti d'uso. Il giudizio di sinonimia può essere problematico. Ad esempio, *bambino* nel senso fondamentale di *essere umano tra la nascita e lŠinizio della fanciullezza* è sinonimo di *bimbo*: tuttavia quest'ultima espressione è usata in un contesto più familiare ed esprime un maggior coinvolgimento emotivo.

Troponymy ($\sqsubseteq$ MeaningRelation) : rappresenta la troponimia tra significati. Un troponimo descrive un particolare modo di fare qualcosa (riferito solo ai verbi). Ad esempio: camminare è un troponimo di andare, poiché camminare è un particolare modo di andare.

Hyponymy ($\sqsubseteq$ MeaningRelation) : rappresenta la relazione di generalità/specificità: si definisce *iponimo* un'accezione che esprime un significato più specifico. Ad esempio: *cane* nel senso di *animale domestico* è un iponimo di *mammifero*, poiché un cane è conosciuto come un determinato tipo di mammifero. Anche nel caso dell'iponimia, tuttavia, il giudizio lessicale dell'utente può rivelarsi problematico.

Antonymy ($\sqsubseteq$ MeaningRelation) : rappresenta la relazione di contrarietà, tipicamente tra gli aggettivi, come ad esempio *bello* e *brutto*. In alcuni casi questa può corrispondere alla *disgiuntezza logica*.

Meronymy ($\sqsubseteq$ MeaningRelation) : rappresenta il rapporto parte/tutto, dove *meronimo* è la parola che descrive la parte. Ad esempio, sono legati da questa relazione *braccio* (nel senso di braccio umano) e *corpo umano*.

Si osservi che, in Senso Comune, relazioni lessicali come la sinonimia, iperonimia/iponimia, antonimia, generalmente

percepite come manifestazione linguistica di relazioni logiche di identità, implicazione e mutua esclusione, non sono necessariamente tradotte in relazioni ontologiche. Questa corrispondenza viene istituita a seguito di un processo di attenta valutazione (analisi ontologica) delle indicazioni lessicali fornite dagli utenti. Meronimia e troponimia, ad esempio, richiedono un attento confronto con le nozioni ontologiche di parte, in accordo con proprietà formali (es. transitività) che difficilmente sono percepite dall'utente (e dal lessicografo).

3. SVILUPPO DELLA RISORSA

La base di conoscenza Senso Comune sarà inizialmente popolata con circa 10.000 accezioni dei 2075 lemmi del vocabolario di base di De Mauro [6], alle quali sarà fatta corrispondere una controparte concettuale basata sulle categorie ontologiche descritte in 2.2.Risorse ontologico-linguistiche dell' italiano esistenti, come ad esempio EuroWornet [13] o i glossari tecnico-specialistici della Pubblica Amministrazione, potranno essere integrate mediante un'adeguata conversione e rese disponibili in formato aperto.
Una volta costituito il nucleo di accezioni fondamentali, la base di conoscenza di Senso Comune verrà sviluppata per mezzo di una piattaforma cooperativa aperta al contributo (controllato) del parlante.

3.1 Acquisizione della terminologia di base

L'acquisizione della terminologia di base è in corso di completamento. A partire dai lemmi estratti in forma testuale dal dizionario, l'obiettivo è quello di costruire corrispettive istanze della classe **LexicalEntry**. La strategia di conversione consiste nel passare per un formato intermedio: viene creato un file `XML` in cui i contenuti del dizionario sono annotati in maniera tale da poter identificare il tipo di contenuto. Successivamente, attraverso un processo di compilazione, si ottiene la popolazione della base di dati.
L'idea di procedere all'annotazione manuale del formato testuale del dizionario è stata scartata a seguito di un analisi di fattibilità: infatti i tempi di strutturazione dei soli lemmi fondamentali (2075 lemmi) sono stati valutati in circa tre anni in lavoro. Studiando il dizionario a livello di struttura, inoltre, ci si è accorti che neanche un approccio completamente automatico sarebbe stato fattibile. Infatti, rispetto al modello di Senso Comune, il formato testuale del dizionario si è rivelato non sufficientemente regolare per consentire la progettazione di un parser per la sua conversione automatica nel formato desiderato. Per questi motivi è stato scelto un approccio semi-automatico: un parser genera la migliore approssimazione del formato XML desiderato, mentre una figura di linguista si occupa di correggere eventuali errori e arbitrare sulle migliori soluzioni in casi di incertezza (Fig.4. In particolare, la distinzione tra casi d'uso e sfumature di senso non si può evincere con regolarità dalla strutture sintattica del formato testuale dizionariale.

3.2 La piattaforma cooperativa

Una volta costituito nella sua forma di base, il lessico computazionale verrà esteso attraverso una piattaforma cooperativa che condivide molte delle caratteristiche di quelle conosciute oggi come "wiki". Un wiki è un software basato sul Web che permette a tutti i visitatori di definire il contenuto del sito. Questo fa del wiki una piattaforma semplice

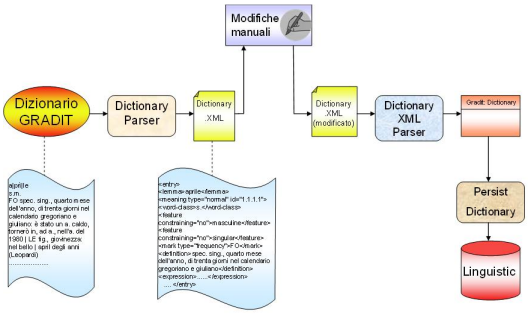

Figure 4: Il processo di acquisizione dei lemmi di base

e facile da usare per lavori cooperativi su testi e ipertesti[7]. Attualmente, sono disponibili in rete un gran numero di sistemi wiki con diversi scopi e tipi di utenza, i quali però condividono i seguenti aspetti caratterizzanti:

- Editing attraverso il browser: i contenuti sono, normalmente, redatti attraverso il browser e dunque senza alcun software aggiuntivo con una (relativamente) semplice sintassi.
- Meccanismo di rollback: i cambiamenti dei contenuti di un wiki sono versionati ogni volta che vengono salvati, quindi, le versioni precedenti delle pagine sono mantenute.
- Accesso non restrittivo: nella maggior parte dei sistemi wiki, l'accesso è completamente libero: chiunque accede ad una pagina del sistema può correggere, modificare, completare o cancellare qualsiasi contenuto.
- Editing collaborativo: molti sistemi wiki forniscono supporto per l'editing collaborativo attraverso forum di discussione, indici dei cambiamenti e altro.
- Enfasi sul *linking*: le pagine in un wiki sono, solitamente, fortemente connesse ad altri ipertesti.
- Funzioni di ricerca: praticamente tutti i sistemi wiki permettono una ricerca sul contenuto di tutte le pagine.
- Caricamento di altri contenuti: oltre la creazione di contenuti testuali diversi wiki permettono di caricare contenuti multimediali come immagini, codice di programmi, etc.

Gli aspetti critici dei sistemi wiki sono principalmente tre:

1. Difficoltà a mantenere la neutralità delle informazioni. Nonostante l'impegno a rappresentare il punto di vista neutrale, è difficile ottenere un accordo completo e reale su tutti i temi trattati. I responsabili del progetto invitano i lettori a segnalare i contenuti che per la loro percezione e la loro sensibilità non appaiono neutrali, per poter essere in grado di intervenire.
2. Qualità dei contenuti. Questo secondo aspetto riguarda i casi in cui gli argomenti non sono trattati nella maniera

più completa o più aggiornata o maggiormente comprensibile. Anche in questo caso, come nel precedente, si fa affidamento sulle segnalazioni da parte dei lettori o sugli interventi diretti dei responsabili.

3. Vulnerabilità ad attacchi malevoli. Gli attacchi malevoli sono azioni tese volontariamente a danneggiare i contenuti o ad arrecare disagio nei lettori con l'introduzione di termini o concetti offensivi o semplicemente fuori luogo.

Wiktionary[9] è un progetto nato con lo scopo di produrre un dizionario libero e multilingue, contenente significati, etimologie e pronunce, e dunque rappresenta il più vicino termine di raffronto tra Senso Comune e la galassia dei wiki. Iniziato il 3 maggio 2004, ha già più di 70.000 lemmi. In generale, una voce del Wikizionario contiene:

- un template iniziale, che specifica la lingua a cui appartiene il lemma ed il tipo di parola, (ad esempio se è un sostantivo italiano, oppure un verbo in inglese, e così via)
- il lemma seguito da informazione morfologica.
- un elenco numerato che esprime i diversi significati del lemma. In alcuni casi il significato può essere preceduto dall'ambito di applicazione, ad es.: (informatica) per lemmi di argomento inerente l'informatica.

In generale, dato il carattere multilinguistico di Wiktionary, è necessario specificare la lingua a cui tale termine appartiene. Vi è inoltre un uso intensivo dei *template* che rende l'editing del lemma alquanto complesso per l'utente non esperto. Ad esempio, è necessario specificare correttamente il *template* che indica il tipo di parola (o parte del discorso) in oggetto. Le voci comprendono in genere anche le seguenti sezioni:

- la sillabazione della parola in oggetto;
- la pronuncia del termine in oggetto, definita secondo gli standard IPA;
- l'etimologia della parola in oggetto;
- i sinonimi della parola in oggetto;
- i contrari della parola in oggetto;
- eventuali proverbi e modi di dire correlati alla parola in oggetto;
- per un significato particolare può essere aggiunta una o più frasi a carattere esplicativo;
- la traduzione nelle diverse lingue della parola in oggetto;
- la coniugazione o la declinazione completa;
- eventuale collegamento al termine Wikipedia corrispondente.

[9]http://it.wiktionary.org/

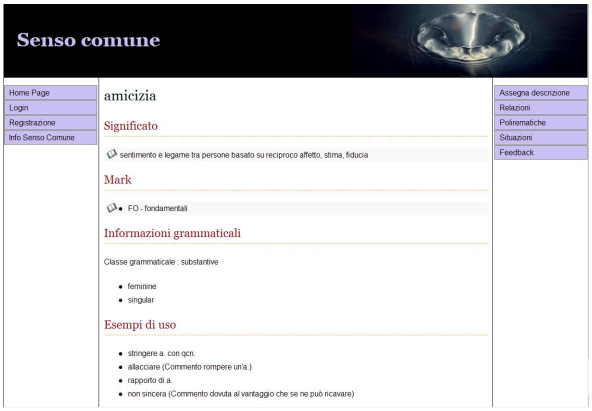

Figure 5: Prototipo del sistema Senso Comune

La principale limitazione del Wiktionary, attualmente, consiste nella struttura dizionariale, che è stata modellata secondo la corrispondenza lemma-pagina, essendo le singole accezioni ridotte a sezioni testuali e non dotate di specifico identificatore. Questo fa sì, ad esempio, che le relazioni di sinonimia o di alterazione (così come qualsiasi relazione lessicale) si possano istituire solo al livello del lessema, mentre è noto che esse riguardano l'accezione.

A causa di tali limitazioni, si è stabilito di sviluppare Senso Comune con una piattaforma applicativa wiki specifica e originale (Figura 3.2).

Il sistema, attualmente nella fase di prototipazione, si fonda su una base di dati relazionale ottenuta dal modello linguistico (2.3) integrato con un ragionatore DL-Lite appositamente studiato per operare su ontologie di grande volume. La possibilità lasciate sono differenti a seconda si tratti di operare su una descrizione di significato derivata da un dizionario oppure inserita da altri utenti.

Dopo aver visualizzato tutte le informazioni legate alla descrizione del significato ricercato, l'utente potrà decidere di inserire un nuovo lemma, una nuova accezione, una nuova relazione lessicale, o lasciare un feedback (ad es. per indicare la propria familiarità con le accezioni o le relazioni lessicali presenti). Al contrario, l'utente non sarà chiamato direttamente a definire la parte concettuale del lessico (ontologia), avendo dunque come accesso alla semantica quello della definizione di relazioni lessicali. All'inserimento di una nuova accezione, il sistema creerà automaticamente uno specifico concetto corrispondente, cercando di posizionarlo nel modo più appropriato rispetto all'ontologia fondamentale, anche ed eventualmente mediante l'interazione con l'utente. Successivamente, le relazioni lessicali espresse dagli utenti saranno utilizzate per inferire corrispondenze (anche semi-automaticamente) sul versante ontologico, con eventuali fusioni (o scissioni) tra concetti.

4. SENSO COMUNE PER LA PUBBLICA AMMINISTRAZIONE

Con la legge 241 del 1990, la trasparenza, la comunicazione, l'accesso alla documentazione amministrativa e l'informazione

sui procedimenti sono stati riconosciuti come una delle priorità della Pubblica Amministrazione. Le informazioni contenute nei documenti delle amministrazioni pubbliche sono state considerate nella loro rilevanza per i cittadini e le imprese, ed il loro accesso è stato dichiarato come diritto per i cittadini, le imprese, la società civile in genere. Negli anni '90 è iniziato un lavoro sul linguaggio dei testi normativi e amministrativi, confluito poi nella direttiva Dipartimento della Funzione Pubblica del 2001 sulla semplificazione linguistica. Un lavoro esteso, approfondito e tecnicamente accurato sul lessico amministrativo ed il suo uso effettivo per l'accesso, anche e soprattutto telematico, al patrimonio informativo della P.A., tuttavia, non è ancora stato intrapreso. L'iniziativa denominata "Chiaro" ha prodotto tra le altre cose un glossario online [10] che tuttavia non si configura come un lessico computazionale. Glossari e tesauri di termini tecnico-specialistici dell'amministrazione sono stati sviluppati in diversi progetti, mancando tuttavia un'iniziativa organica di integrazione, diffusione ed utilizzo della conoscenza terminologica in forma utile per l'accesso al patrimonio informativo della P.A.

Progetti più specifici per migliorare l'accesso alle risorse informative della Pubblica Amministrazione sono stati di recente condotti utilizzando l'approccio cosiddetto "a faccette" (*faced based*) [12]. Questo consiste nell'identificare un certo numero di attributi rilevanti per classi definite di documenti e nel definire i valori ammessi per ciascuno di essi. In seguito, ogni documento è definito da un vettore di valori (eventualmente estratti con procedure automatiche) ed è quindi possibile compiere operazioni di classificazione e reperimento. Le *facet* possono eventualmente essere in corrispondenza con ontologie ed attraverso di esse supportare procedure di *reasoning* e funzioni avanzate di ricerca.

Il contributo di Senso Comune ad una Pubblica Amministrazione "aperta e libera" è duplice. Da una parte, la piattaforma di acquisizione della risorsa, attualmente in prototipazione, può essere utilizzata per lo sviluppo di una base di conoscenza di nozioni comuni alle funzioni pubbliche e alle loro utenze. Glossari più ampi e strutturati di quelli attualmente esistenti potranno essere sviluppati in modo cooperativo ma controllato, per divenire risorse linguistico-concettuali condivise tra le Pubbliche Amministrazioni, e tra queste ed i cittadini. D'altra parte, lessico computazionale aperto contenente terminologia tecnico-specialistica della Pubblica Amministrazione parte di Senso Comune, per la sua specifica forma, potrà supportare lo sviluppo di applicazioni (ad es. knowledge management systems) dotate di sofisticate procedure di ragionamento automatico, in grado di operare con efficienza su ampie basi di dati. Infine, grazie alla sua gratuita disponibilità, Senso Comune potrà favorire lo sviluppo industriale nel Settore Pubblico garantendo condizioni di accesso anche ai soggetti industriali che operano prevalentemente su piattaforme Open Source.

5. RINGRAZIAMENTI

Al Presidente Tullio De Mauro e al Direttivo dell'Associazione Angelo Failla, Aldo Gangemi, Nicola Guarino, Maurizio Lenzerini e Malvina Nissim per la loro revisione. Ad Ilaria Gorga (IBM Italia), Alessandro Oltramari (CNR), Rita Plantera (Università di Roma I), Fabrizio Smith (Università di Roma III) per il materiale messo a disposizione.

[10] http://www.funzionepubblica.it/chiaro/glossario.htm

6. REFERENCES

[1] F. Baader, D. Calvanese, D. L. Mcguinness, D. Nardi, and P. F. Patel-Schneider, editors. *The Description Logic Handbook : Theory, Implementation and Applications*. Cambridge University Press, January 2003.

[2] D. Calvanese, G. De Giacomo, M. Lenzerini, R. Rosati, and G. Vetere. Dl-lite: Practical reasoning for rich dls. In *Proc. of the 2004 Description Logic Workshop (DL 2004)*, volume 104 of *CEUR Electronic Workshop Proceedings, http://ceur-ws.org/Vol-104/*, 2004.

[3] D. Calvanese, G. D. Giacomo, D. Lembo, M. Lenzerini, and R. Rosati. Tractable reasoning and efficient query answering in description logics: The dl-lite family. *J. Autom. Reasoning*, 39(3):385–429, 2007.

[4] I. Chiari. *Introduzione alla linguistica computazionale*. Laterza, Bari, 2007.

[5] T. De Mauro. *Introduzione alla semantica*. Laterza, Bari, 1965.

[6] T. De Mauro, editor. *Grande Dizionario dell'Italiano dell'Uso*. UTET, Torino, 1999.

[7] A. Ebersbach, M. Glaser, and R. Heigl. *Wiki : Web Collaboration*. Springer, November 2005.

[8] C. Fellbaum, editor. *WordNet, An Electronic Lexical Database*. MIT Press, Boston, 1998.

[9] G. Francopoulo, N. Bel, M. George, N. Calzolari, M. Monachini, M. Pet, and C. Soria. Lexical markup framework (lmf) for nlp multilingual resources. In *Proceedings of the COLING-ACL Workshop on Multilingual Lexical Resources and Interoperability*, pages 1–8, 2006.

[10] A. Gangemi, N. Guarino, C. Masolo, A. Oltramari, and L. Schneider. Sweetening ontologies with dolce. In *Proceedings of EKAW*, pages 21–29, 2002.

[11] A. Gangemi, R. Navigli, and P. Velardi. The ontowordnet project: extension and axiomatization of conceptual relations in wordnet, 2003.

[12] L. Rosati, M. E. Lai, and C. Gnoli. Faceted classification for public administration. In *Workshop Semantic Web Applications and Perspectives (Ancona, 11 dicembre 2004)*, 2004.

[13] P. Vossen, editor. *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Kluwer Academic Publishers, Dordrecht, 1998.