Sezione II

Fasce e settori del lessico

ISABELLA CHIARI* – ALESSANDRO OLTRAMARI° – GUIDO VETERE[2]
(* Sapienza Università di Roma, °CNR – ISTC, [2]IBM Centro di Studi Avanzati di Roma)

Di cosa parliamo quando parliamo fondamentale? Lessemi, accezioni, sensi e ontologie*

1. INTRODUZIONE

Nelle storia delle scienze linguistiche, la semantica, intesa come indagine del rapporto tra le unità del lessico e la realtà esterna al linguaggio, ha sempre rappresentato una delle maggiori difficoltà, ed è stata oggetto di atteggiamenti anche decisamente scettici (De Mauro, 1965). La discussione sulla teoria del significato, e in particolare il suo sviluppo nella filosofia del linguaggio novecentesca, ha prodotto varie ipotesi, ma non risultati largamente accettati (Picardi, 1999). Nondimeno, lo sviluppo delle moderne tecnologie dell'informazione, e in particolare del web, richiedono oggi, con grande urgenza, che la semantica delle lingue naturali sia in qualche modo esplicitata. Tale urgenza deriva dalla necessità di abilitare il dialogo tra i sistemi automatici che, connettendosi attraverso le reti telematiche, gestiscono molti aspetti della nostra vita quotidiana, come ad esempio le operazioni bancarie, il commercio, i procedimenti amministrativi, il turismo, le relazioni sociali, l'intrattenimento. La visione di una rete globale di automi in grado di intendersi sui concetti sottostanti a tutte queste attività è detta appunto Semantic Web (Berners-Lee, *et al.,* 2001). Questa visione, che si è sviluppata nelle comunità di ricerca informatica a partire dall'ultimo decennio del secolo scorso, conferisce un ruolo fondamentale all'ontologia. Per ontologia, nel contesto del Semantic Web, si intende in genere un tipo di costruzione logico-descrittiva intesa a modellare e rappresentare le entità e i processi del mondo in cui vivono le applicazioni informatiche. Le ontologie in uso nel Semantic Web, tuttavia, si configurano per lo più come terminologie di natura linguistica, in cui le accezioni sono trattate alla stregua di

* Il progetto della sperimentazione è stato ideato, discusso e revisionato congiuntamente dai tre autori. Della presente pubblicazione il paragrafo 1 è stato elaborato da Guido Vetere; il paragrafo 2.1 congiuntamente elaborato da Isabella Chiari e Guido Vetere; i paragrafi 1.2; 2.3; 3.1-3.3 e 4 sono stati elaborati da Isabella Chiari; il paragrafo 2.3 è stato elaborato da Alessandro Oltramari. La gestione del database è stata supervisionata da Guido Vetere, la creazione e ideazione di TMEO è di Alessandro Oltramari; i tre autori sono inoltre responsabili della revisione delle classificazioni. Gli autori ringraziano in particolare per la sperimentazione: N. Amabile, V. Arena, V. Cristini, M. D'Auria, F. De Giusti, V. Di Marco, L. Mascara, A. Napoleone, F. Riccardi, T. Taboga, E. Vanni. Si ringraziano inoltre per lo sviluppo del database e interfaccia P. Cangialosi e A. Mencancini.

predicati logici e le relazioni lessicali (ad esempio l'iponimia) sono tradotte in assiomi (ad esempio l'inclusione). Tale "promozione" del senso linguistico allo statuto di elemento logico-formale, anche quando non sia del tutto ingenua, appare non priva di criticità. Vi sono tuttavia anche proposte di ontologie cosiddette "fondazionali" per le quali, piuttosto che rivolgersi al linguaggio, si adottano prospettive metafisiche e/o cognitive. L'obiettivo di queste ontologie è quello di vincolare il significato inteso dei predicati che occorrono nei modelli informatici, siano essi più o meno derivati da fatti linguistici, rispetto ad alcune categorie, che sono mutuamente vincolate da relazioni (ad esempio dipendenza, parte), a cui sono annessi assiomi che derivano per lo più dalla tradizione ontologico-formale del Novecento.

2. Semantica e ontologia: Il progetto Senso Comune

Questo contributo intende discutere da un punto di vista teorico e applicativo il problema della integrazione tra risorse lessicali, soprattutto di tipo lessicografico, e rappresentazioni della realtà e delle conoscenze mediante modelli formali. Uno dei principali obiettivi auspicati per la realizzazione del cosiddetto Semantic Web, e più in generale per dar conto in maniera più accurata delle conoscenze linguistiche e delle conoscenze *tout court* in qualche modo espresse e verbalizzate testualmente, è l'integrazione tra diversi tipi di informazione rilevante dal punto di vista linguistico, tenendo conto da una parte della polisemia lessicale e testuale e dall'altra dell'utilità di associare alle rappresentazioni più specificatamente linguistiche rappresentazioni di altri tipi di conoscenze (De Mauro, 2009). In questa direzione si sono mossi tra i primi i creatori di WordNet (Fellbaum, 1998), intuendo tra l'altro la necessità che tali risorse fossero rese liberamente disponibili in formato sorgente. Per la lingua italiana Senso Comune (SC, www.sensocomune.it) sviluppa un progetto che mira alla costruzione di una base di conoscenza linguistica dell'italiano con la cooperazione di utenti/parlanti attraverso il web. Si tratta di un primo passo, anche se non l'unico, nella direzione di integrare e restringere il divario, puramente artificiale, tra dimensione semantica, pragmatica e informazione di sfondo enciclopedico e contestuale. Delle diverse possibilità di far interagire ontologia, semantica e risorse lessicali (Prévot, *et al.,* 2010), SC ha scelto, nella sperimentazione qui presentata, la marcatura dell'informazione linguistica in una risorsa lessicale esistente con una mappa ontologica molto generale[1].

[1] L'uso di ontologie ha la doppia funzione di fornire una rappresentazione di una concettualizzazione di un dato dominio e contribuire alla costruzione di applicazioni che automatizzino processi di ragionamento sulla struttura semantica di testi nel dominio dato. Esiste inoltre una funzione di adeguatezza computazionale che riguarda l'efficacia ed applicabilità dell'ontologia nella

2.1. *Il progetto Senso Comune*

Senso Comune (Oltramari e Vetere, 2008, Oltramari, *et al.,* 2010) è una base di conoscenza della lingua italiana aperta al contributo dei parlanti attraverso il web, disponibile in formato sorgente. La risorsa lessicale è strutturata in modo da consentire il suo impiego, oltre che per la consultazione, anche in sistemi di trattamento automatico della lingua, caratteristica che la distingue dai dizionari tradizionali. Il contenuto di SC è disposto in una base di dati improntata a un modello lessicale. Il modello, nella sua generalità, si ispira ad alcune recenti proposte per le risorse linguistiche computazionali, come quella del Lexical Markup Framework [LMF], esso tuttavia è stato disegnato per supportare più specificatamente i processi di acquisizione e di interrogazione. Per quanto riguarda l'acquisizione, la base di conoscenza di SC è stata inizialmente popolata con le accezioni con marca d'uso *fondamentale* del *Grande Dizionario Italiano dell'Uso* (De Mauro, 1999). In fasi successive, gruppi di studenti del corso di Linguistica computazionale della Facoltà di Scienze Umanistiche dell'Università Sapienza hanno provveduto a revisionare e arricchire il contenuto[2].

costruzione di software (Evermann e Fang, 2010: 392). L'uso di ontologie per diverse applicazioni che vanno dal Semantic Web, commercio elettronico, *knowledge management*, ecc. ha polarizzato negli ultimi anni l'interesse anche sul tema della valutazione, comparazione e test delle ontologie in relazione alle concettualizzazioni che gli utenti fanno del dominio dell'ontologia (Brewster e O'Hara, 2007, Evermann e Fang, 2010). Le ontologie infatti devono poter essere discusse dal punto di vista qualitativo sulla base anche della condivisibilità effettiva delle concettualizzazioni su cui si basano da parte da utenti profani. Le ontologie dovrebbero dunque avere una sorta di requisito di plausibilità d'uso, ossia la *cognitive ontology quality*: "ontology, as a formal description of a domain, must conform to the way in which the domain is perceived and understood by a human observer" (Evermann e Fang, 2010: 392). Questa caratteristica è dunque legata alla qualità semantica percepita da chi usa l'ontologia. La centralità della dimensione semantica e di quanto venga effettivamente catturato di essa da una ontologia è diventato uno dei temi più delicati, molto oltre le questioni di metalinguaggio, formalizzazione e interoperabilità (Berners-Lee, *et al.*, 2006, Hu, *et al.,* 2007)

[2] La politica di apertura nei confronti dei contributi degli utenti è, specialmente nella fase attuale, orientata a garantire la massima qualità e controllabilità. Pertanto, a differenza da quanto avviene per risorse enciclopediche come la Wikipedia, l'autorizzazione a modificare o ampliare la base di conoscenza è data attualmente a specifici gruppi di lavoro o a seguito di una verifica delle credenziali dei potenziali contributori. In ogni caso, qualsiasi utente della rete può commentare i contenuti presenti.

Figura 1. *Schermata di accesso a SC*

Senso Comune attualmente si compone di un nucleo costituito dalla fascia del vocabolario fondamentale di 2.071 lemmi che coprono più del 90% delle occorrenze di testi parlati e scritti (De Mauro, 1980). Il progetto, sotto la supervisione di Tullio De Mauro, intende integrare la risorsa lessicografica con altre risorse linguistiche in un progetto collaborativo che si svolge attraverso una piattaforma accessibile al sito web dell'Associazione (www.sensocomune.it). Uno degli obiettivi di SC è quello di mettere in relazione le informazioni linguistiche di natura lessicale e semantica del dizionario con una rappresentazione delle conoscenze, sotto forma di una ontologia formale. Tale relazione è spesso complessa da mappare per specifiche differenze teoriche che soggiacciono alle modalità di rappresentazione delle conoscenze linguistiche e delle ontologie. In particolare è evidente che non si tratta di modelli sovrapponibili, soprattutto per le caratteristiche tipiche degli elementi lessicali quali sinonimia e soprattutto polisemia.

In SC, ciascuna accezione (ma, attualmente, solo quelle appartenenti alla categoria dei sostantivi) può essere associata a un concetto definito in un'ontologia. L'ontologia di SC è basata su una versione semplificata dell'ontologia DOLCE (Masolo, *et al.,* 2002). Il significato di questa classificazione è un aspetto cruciale dell'intero progetto. L'intuizione che si intende cogliere è che l'uso regolare di una parola in determinati contesti, usualmente esprimibile in una definizione lessicografica ed esemplificabile mediante frasi tipiche, implica una qualche sorta di "impegno" verso un certo tipo di entità. La natura di questo impegno non rientra negli scopi del progetto, dunque la categorizzazione non comporta né un atteggiamento realista, per cui gli enti, a priori, informerebbero il linguaggio, né un atteggiamento costruttivista, per cui le categorie ontologiche sarebbero prodotto del linguaggio stesso. In SC, l'ontologia si pone piuttosto come costrutto ipotetico che resta in attesa di riscontri empirici, lasciando da parte il problema teorico e filosofico di cosa eventualmente giustifichi e fondi la relazione tra il linguaggio e una certa teoria della realtà.

Un aspetto interessante consiste inoltre nell'osservare l'effetto dell'ontologia sull'organizzazione della risorsa lessicale. È importante notare che in SC le categorie ontologiche sono esplicitamente assegnate a ciascuna accezione. Questo rappresenta un'importante differenza rispetto a ciò che caratterizza altre importanti risorse linguistiche, come ad esempio WordNet (Fellbaum, 1998), in cui l'organizzazione concettuale emerge dalle relazioni lessicali (principalmente quelle di sinonimia e iponimia), e gli impegni ontologici sono eventualmente indagati a posteriori (Gangemi, *et al.,* 2008). La concettualizzazione a priori delle accezioni, secondo un'opportuna categorizzazione, offre agli utenti di SC la possibilità di valutare l'efficacia e la specificità delle singole accezioni rispetto alla mappa dell'ontologia di SC.

2.2. *Sperimentazione sui sostantivi del vocabolario fondamentale*

L'obiettivo della sperimentazione è osservare la possibilità di associare a ciascuna accezione dei lemmi con marca fondamentale (FO) una categoria ontologica e verificare quali sono i problemi concreti che si incontrano in tale operazione. In particolare ci siamo concentrati su 1.111 sostantivi appartenenti al vocabolario fondamentale (circa il 53,6% dei 2.071 lemmi con etichetta FO del De Mauro, 2000). Il compito consisteva nell'associare a ciascuna delle 4.586 accezioni con marca fondamentale dei sostantivi FO una sua classe ontologica[3], tra le macroclassi previste per i sostantivi.

La sperimentazione ha previsto tre fasi: I FASE: Classificazione primaria, in assenza di istruzioni per preservare una categorizzazione di senso comune, da parte di 12 soggetti; Discussione e rimodulazione dell'ontologia. II FASE: Revisione della classificazione, da parte dei tre autori e di 4 sperimentatori del primo gruppo; Attribuzione delle classi di confidenza (*accettato, controverso, non accettato*); Discussione. III FASE: Ultima revisione della coerenza generale, a campione.

Uno dei problemi teorici più interessanti è indagare in che senso la categorizzazione proposta possa dirsi di senso comune. Come abbiamo detto, l'ontologia di SC si collega debolmente al database lessicale e l'impegno epistemologico relativo al tale connessione e alla struttura stessa delle categorie e relazioni dell'ontologia rimane volutamente opaca per l'utente, che è chiamato, sulla base del suo senso comune ad associare elementi del database lessicale con elementi dell'ontologia. L'ontologia proposta è stata inoltre costruita anche utilizzando il

[3] La sperimentazione non ha dunque analizzato tutte le accezioni dei lemmi FO, ma solamente quelle anch'esse marcate FO, escludendo dunque un gran numero di accezioni appartenenti alle altre fasce del vocabolario, come alto uso, alta disponibilità, comune, tecnico-specialistico.

feedback degli sperimentatori che per primi hanno usato l'ontologia nella sua prima versione semplificata.

Per fare un esempio, il lemma sostantivale *peso* possiede diverse accezioni ciascuna delle quali rimanda a diverse classi ontologiche: l'accezione come "corpo soggetto alla forza di gravità" (*caricare, portare, sollevare pesi*) è classificata come OGGETTO, "oggetto, specialmente metallico, opportunamente graduato che serve nelle operazioni di pesatura" come ARTEFATTO, mentre "senso di pesantezza dovuto a cattiva digestione" come STATO CORPOREO, "condizione, situazione che reca disagio, fastidio, sofferenza fisica o morale" come STATO PSICHICO, "autorità, prestigio, influenza" (*il peso di una posizione sociale, il peso del casato*) come IDEA, ecc.

Il sistema di tutoraggio (vedi 2.3) può essere attivato a scelta dell'utente per guidarlo alla classificazione attraverso una serie di domande (*Puoi toccare, vedere, annusare, gustare, o percepire un "peso"? Puoi contare diversi "pesi"? un "peso" è costruito da un agente?*, ecc.), sensibili allo storico delle risposte, fino alla classificazione soddisfacente per l'utente.

2.3. *Il sistema di tutoraggio TMEO*

La realizzazione di ontologie richiede normalmente una fase di rappresentazione formale ad alto livello, in cui vengono definite le distinzioni ontologiche in un ambito semantico più o meno generale[4], e una fase di popolamento a basso livello, in cui strati di conoscenza sempre più specifici e legati ad uno o più domini "ereditano" la struttura ontologica di riferimento[5]. Se questo processo di progettazione e sviluppo permette di ottenere modelli concettuali cognitivamente più adeguati e formalmente più rigorosi rispetto alle tecniche automatiche (altresì dette di "ontology learning"), la creazione manuale di ontologie rimane più dispendiosa in termini di costi umani relativi ai tempi di realizzazione: in particolare, il popolamento dei livelli più specialistici di un'ontologia richiede un'interazione costante tra ontologo ed esperto di dominio, nel corso della quale il primo assiste il secondo nell'attività di elicitazione della conoscenza secondo adeguati vincoli concettuali e formali.

In questo contesto, la metodologia TMEO (Tutoring Methodology for the Enrichment of Ontologies) è stata sviluppata per mantenere il livello qualitativo della creazione manuale di ontologie limitandone i tempi di realizzazione.

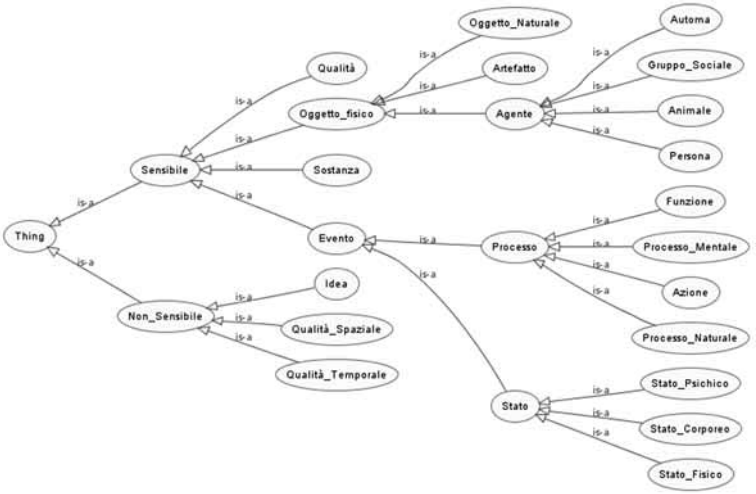
---

[4] Dal punto di vista intensionale, tale struttura si declina in predicati unari (ad esempio, "(essere) casa", "(essere) automobile", "(essere) albero") e predicati n-ari (ad esempio "essere-parte-di", "partecipare-a", etc.

[5] L'ereditarietà si traduce estensionalmente in inclusione tra classi: ad es. *tigre* è sottoclasse di *animale* ed eredita quindi tutte le sue proprietà definitorie, per cui si dice appunto che *tigre* è un animale con certe caratteristiche (felino, carnivoro, etc.).

Ispirandosi sia nel nome che nella pratica ai Dialoghi di Platone, in cui il Socrate guida i suoi discepoli nel percorso che conduce alla Verità, l'obiettivo ben più circoscritto di TMEO è quello di sfruttare le proprietà rilevanti dello "strato alto" di un'ontologia per generare domande da porre agli esperti di dominio (eventualmente riuniti in comunità Wiki, come nel caso di SC). Nello specifico, il compito di TMEO all'interno del progetto è di guidare il parlante alla selezione della categoria ontologica più adeguata per classificare i lemmi fondamentali della lingua italiana: sequenze di risposte a domande calibrate su categorie e relazioni ontologiche di alto livello inducono in modo semi-automatico il cosiddetto "popolamento" dell'ontologia. Come si può notare in Figura 2, nell'attuale ontologia di SC, tutte le distinzioni fondamentali possono essere ricondotte alla macro-suddivisione tra la classe "sensibile", che designa cioè le entità percepibili tramite i cinque sensi (edifici, canzoni, partite di calcio, etc.) e la classe "non-sensibile", riferita cioè alle entità che diremmo "astratte" come le idee, i numeri, i contenuti dei nostri enunciati, etc.

Figura 2. *Struttura gerarchica della classificazione*



Si consideri il seguente esempio, facilmente riproducibile consultando la piattaforma online di SC. Per classificare il lemma *scarpa* nell'accezione di "parte dell'abbigliamento, di cuoio o di altro materiale, che copre e protegge il piede gener. fino alla caviglia o poco sopra", il sistema pone le seguenti domande, sensibili alla risposta precedente:

Figura 3. *Esempio di tutoraggio con TMEO*

TMEO: Puoi percepire *scarpa* con almeno uno dei cinque sensi?
Utente: SI
TMEO: Puoi contare diversi (esempi di) *scarpa* ?
Utente: SI
TMEO: Puoi assistere, partecipare, subire o essere il soggetto di *scarpa*?
Utente: NO
TMEO: *scarpa* può decidere autonomamente per raggiungere un obiettivo, interagendo eventualmente con altri agenti (non necessariamente dello stesso tipo)?
Utente: NO
TMEO: *scarpa* è costruito/definito/inventato da un agente?
Utente: SI
TMEO: *scarpa* ha un'esplicita funzione sociale (es. una legge)?
Utente: NO

A questo punto, dopo aver elaborato la sequenza di risposte date dall'utente, il sistema classifica l'accezione del lemma *scarpa* secondo la categoria ontologica ARTEFATTO. La classificazione avviene dunque senza l'interazione diretta tra ontologo ed esperto di dominio, semplificando e incapsulando le tecniche di modellazione concettuale note al primo in una sequenza di vincoli concatenati sotto-forma di domande in linguaggio naturale da porre al secondo.

3. RISULTATI

I risultati della sperimentazione sulle 4.586 accezioni dei 1.111 lemmi sostantivali del vocabolario fondamentale si possono raggruppare in tre principali categorie di analisi di cui di seguito presenteremo alcuni dati: dati sulla distribuzione delle accezioni nelle categorie ontologiche; dati sulla problematicità della categorizzazione; dati sulla varietà di classi ontologiche dei lemmi presi in esame.

3.1. *Le classi ontologiche e la loro distribuzione*

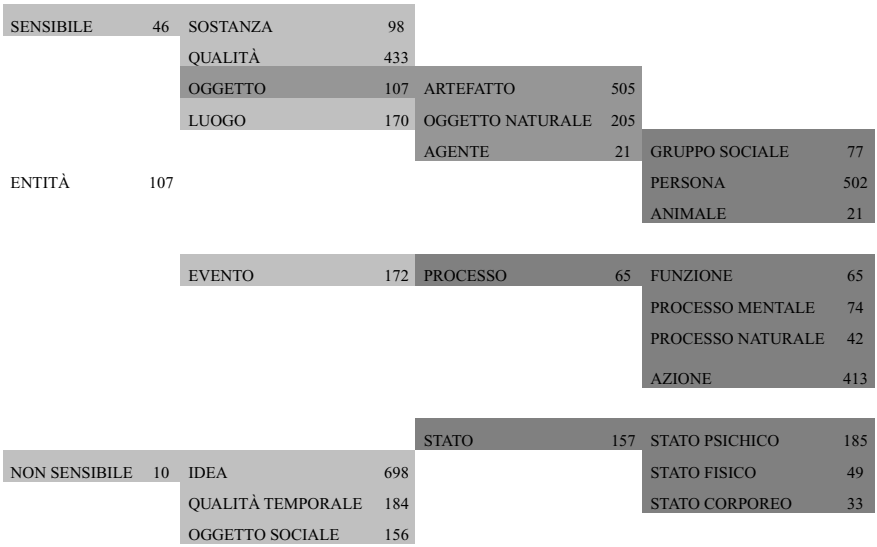
Per rispondere alla domanda nel titolo di questo contributo si può iniziare osservando i dati relativi alla popolosità delle diverse classi ontologiche rappresentate nel vocabolario fondamentale, come si vede in Tabella 1. Osservando i dati puramente quantitativi appaiono come classi molto ben rappresentate soprattutto IDEA, ARTEFATTO, PERSONA, QUALITÀ, AZIONE.

Tabella 1. *Popolosità delle classi ontologiche dei sostantivi FO*

Classe ontologica	Accez. FO	%	Classe ontologica	Accez. FO	%
IDEA	689	15,02%	ENTITÀ	107	2,33%
ARTEFATTO	505	11,01%	SOSTANZA	98	2,14%
PERSONA	502	10,95%	GRUPPO SOCIALE	77	1,68%
QUALITÀ	433	9,44%	PROCESSO MENTALE	74	1,61%
AZIONE	413	9,01%	FUNZIONE	65	1,42%
OGGETTO NATURALE	205	4,47%	PROCESSO	65	1,42%
STATO PSICHICO	185	4,03%	STATO FISICO	49	1,07%
QUALITÀ TEMPORALE	184	4,01%	SENSIBILE	46	1,00%
EVENTO	172	3,75%	PROCESSO NATURALE	42	0,92%
LUOGO	170	3,71%	STATO CORPOREO	33	0,72%
STATO	157	3,42%	ANIMALE	21	0,46%
OGGETTO SOCIALE	156	3,40%	AGENTE	21	0,46%
OGGETTO	107	2,33%	NON SENSIBILE	10	0,22%

Tuttavia il dato grezzo così presentato non dà conto della complessità soggiacente sia alla gerarchia delle classi del modello ontologico, sia della complessità dei processi di categorizzazione che sono governati da 'preferenze' per certi livelli di astrazione piuttosto che altri. Come si è visto nello schema in Figura 2 l'ontologia usata nella sperimentazione presenta una struttura ad albero per cui le classi sono contenute l'una nell'altra e vanno lette appunto in tale relazione (vedi Figura 4). Il fatto che alcune classi, come OGGETTO, siano meno popolose delle classi meno astratte, come ARTEFATTO o OGGETTO NATURALE non significa che gli 'oggetti' sono meno tipici del vocabolario fondamentale, poiché la classe OGGETTO include le altre e per essere valutata bisogna tener conto della relazione tra classi nonché della preferenza generale per categorie di livello basico, quali appunto ARTEFATTO, PERSONA, ecc.

Figura 4. *Popolosità delle classi nella gerarchia*



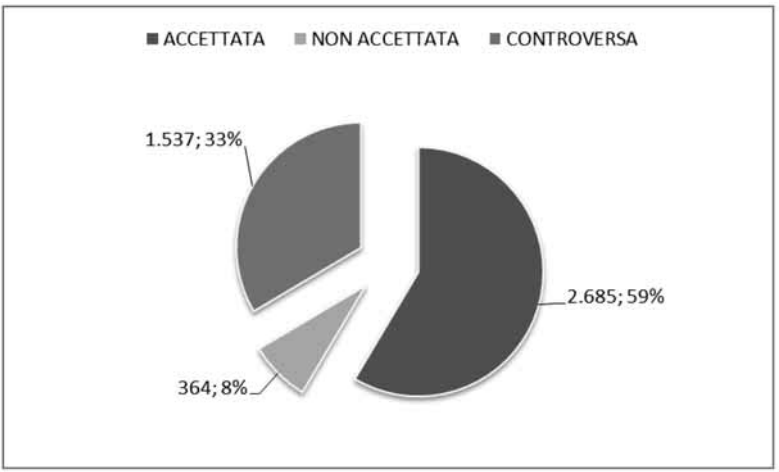
Vediamo ad esempio nella relazione quantitativa tra categorie superiori e inferiori della gerarchia (ad esempio AGENTE e GRUPPO SOCIALE, PERSONA, ANIMALE; oppure PROCESSO e le sue sotto tipologie FUNZIONE, PROCESSO MENTALE, PROCESSO NATURALE, AZIONE; così come STATO e le rispettive STATO FISICO, STATO PSICHICO, STATO CORPOREO) si preferiscono in genere gli elementi che si collocano sul polo più determinato. Nella maggioranza dei casi le classificazioni preferite sono quelle sull'estremo più concreto. Il caso della preferenza per una macro-classe si verifica in genere laddove la glossa dell'accezione, così come si trova nella risorsa lessicografica usata, comprenda più sottocategorie, come nel caso del lemma *animale* nell'accezione "ogni organismo vivente dotato di sensi e movimento" classificato AGENTE, poiché nella classificazione si applica sia a PERSONA che ad ANIMALE; simile il caso del lemma *passaggio* nell'accezione "il mutare stato, condizione" classificato PROCESSO, poiché può essere MENTALE, NATURALE, AZIONE, e dunque una diversa classificazione avrebbe escluso una parte delle classi che effettivamente nell'uso corrispondono a quella accezione. Le caratteristiche di numerosità vanno dunque interpretate sulla scorta di almeno due parametri: le caratteristiche della glossa e la tendenza a preferire categorie di livello base (o prototipi) ed evitare categorie sopra-ordinate.

I dati di popolosità tuttavia non sono sufficienti da soli a dare conto della problematicità della classificazione, anche perché le attribuzioni a una classe possono essere più o meno sicure e più o meno condivise.

3.2. *Il grado di certezza della classificazione*

Per tener conto di questi aspetti abbiamo aggiunto una segnalazione soggettiva che combinasse i due fattori della sicurezza e condivisione nella categorizzazione. Dalla fase di revisione accanto a ogni accezione si è apposta una marca che segnala la concordanza senza dubbi degli sperimentatori su una data classe (accettata), la presenza di una o più ipotesi definite, ma non concordate tra gli sperimentatori (controversa) e la assenza di una ipotesi qualsivoglia che soddisfi la complessità dell'accezione (non accettata). Un esempio di categorizzazione accettata da tutti i revisori è abito come "vestito, capo di abbigliamento" classificato come ARTEFATTO; affare come "caso giudiziario o politico di grande rilievo pubblico" è classificato OGGETTO SOCIALE, ma controverso (per una possibile sovrapposizione con EVENTO) per cui i revisori l'hanno segnalata come problematica per qualche aspetto; buongiorno segnala una mancanza di categoria ontologica appropriata (si può pensare eventualmente ad AZIONE, ma non si tratterebbe propriamente di una categorizzazione di senso comune) ed è classificato come non accettato; stesso dicasi per capitale come "città in cui si trovano gli organismi centrali di uno stato", classificato come OGGETTO SOCIALE, ma non accettato perché sovrapposto significativamente con LUOGO.

Figura 5. *Grado di certezza delle classi*



Come si vede in Figura 5 il 59% delle accezioni è assegnata con sicurezza e condivisione a una determinata classe ontologica, mentre il 33% delle accezioni

risultano controverse, e del tutto non accettabili l'8%. I dati più specifici si trovano in Tabella 2.

Osservando le classi che raccolgono un grado di certezza più ampio (che va da un 81% fino a 68%) troviamo in ordine ANIMALE, PERSONA, OGGETTO NATURALE, ARTEFATTO, SOSTANZA, AZIONE, QUALITÀ TEMPORALE. Con una alta percentuale di non accettazione troviamo invece ENTITÀ, SENSIBILE, NON SENSIBILE, FUNZIONE, OGGETTO, STATO, IDEA.

Tabella 2. *Classi e grado di certezza di attribuzione*

Classe	**Accettate**	**%**	**Controverse**	**%**	**Non accettate**	**%**	**Totali**
AGENTE	14	66,67%	7	33,33%	0	0,00%	21
ANIMALE	17	80,95%	4	19,05%	0	0,00%	21
ARTEFATTO	377	74,65%	106	20,99%	22	4,36%	505
AZIONE	291	70,46%	113	27,36%	9	2,18%	413
ENTITÀ	14	13,08%	28	26,17%	65	60,75%	107
EVENTO	88	51,16%	77	44,77%	7	4,07%	172
FUNZIONE	19	29,23%	32	49,23%	14	21,54%	65
GRUPPO SOCIALE	49	63,64%	25	32,47%	3	3,90%	77
IDEA	301	43,69%	301	43,69%	87	12,63%	689
LUOGO	90	52,94%	76	44,71%	4	2,35%	170
NON SENSIBILE	0	0,00%	8	80,00%	2	20,00%	10
OGGETTO	45	42,06%	39	36,45%	23	21,50%	107
OGGETTO NATURALE	157	76,59%	36	17,56%	12	5,85%	205
OGGETTO SOCIALE	53	33,97%	92	58,97%	11	7,05%	156
PERSONA	405	80,68%	94	18,73%	3	0,60%	502
PROCESSO	30	46,15%	30	46,15%	5	7,69%	65
PROCESSO MENTALE	41	55,41%	31	41,89%	2	2,70%	74
PROCESSO NATURALE	28	66,67%	14	33,33%	0	0,00%	42
QUALITÀ	222	51,27%	171	39,49%	40	9,24%	433
QUALITÀ TEMPORALE	125	67,93%	52	28,26%	7	3,80%	184
SENSIBILE	21	45,65%	13	28,26%	12	26,09%	46
SOSTANZA	71	72,45%	23	23,47%	4	4,08%	98
STATO	66	42,04%	69	43,95%	22	14,01%	157
STATO CORPOREO	20	60,61%	12	36,36%	1	3,03%	33
STATO FISICO	32	65,31%	15	30,61%	2	4,08%	49
STATO PSICHICO	109	58,92%	69	37,30%	7	3,78%	185
Totale	*2.685*	*58,55%*	*1.537*	*33,52%*	*364*	*7,94%*	*4.586*

Cumulando i dati con marca 'controversa' e 'non accettata', come indicatore di forte problematicità della classe o dell'attribuzione troviamo che, con più del 50% di incertezza, si collocano le classi presenti nella Tabella 3, troviamo inoltre che la media di problematicità complessiva delle classi si colloca al 47,70%.

Tabella 3. *Classi fortemente problematiche (>50%)*

Classe	**Controverse + Non accettate**	**% Contr. + Non accett.**
NON SENSIBILE	10	100,00%
ENTITÀ[6]	93	86,92%
FUNZIONE	46	70,77%
OGGETTO SOCIALE	103	66,02%
STATO	91	57,96%
OGGETTO	62	57,95%
IDEA	388	56,32%
SENSIBILE	25	54,35%
PROCESSO	35	53,84%

Come si può notare le classi con maggiore incertezza complessiva (a volte molto popolose) si trovano in rami alti della gerarchia ontologica, ossia sono in genere astratti. Classi come STATO, OGGETTO, SENSIBILE, PROCESSO e, ancora più, SENSIBILE, NON SENSIBILE non solo risultano piuttosto povere, ma sono accompagnate nella gerarchia da sottocategorie che invece risultano particolarmente popolose (si pensi per OGGETTO alle sottocategorie forti PERSONA, OGGETTO NATURALE, ARTEFATTO, ANIMALE). Quando esiste la possibilità di attribuire una etichetta più determinata infatti questa viene preferita e il grado di sicurezza con la quale si gestisce la categoria astratta diminuisce progressivamente.

Diverso il caso di IDEA, classe molto popolosa con 388 ipotesi di attribuzione, e di OGGETTO SOCIALE (156) collocate entrambe sul versante relativamente più specifico della gerarchia. La difficoltà di categorizzazione di questa classe non è dunque il frutto della struttura della gerarchia ontologica, ma di un problema soggiacente che coinvolge il prototipo stesso del concetto di IDEA e la sua 'non sensibilità'. Tale problematicità è probabilmente connessa anche con la difficile orga-

[6] La popolosità di ENTITÀ, che nella gerarchia è la classe più astratta in assoluto, in realtà è un indicatore di difficoltà di classificazione, poiché laddove non fosse stato possibile individuare una classe appropriata si è ricorso all'elemento più indietro nella tassonomia.

nizzazione interna di ciò che risulta 'interiore'[7], classica della inafferrabilità della organizzazione semantica.

3.3. *La varietà di classe ontologica dei lemmi*

Un ultimo dato interessante riguarda la varietà di classi ontologiche attribuite a ciascun lemma. La varietà ci permette infatti di valutare quanto le accezioni FO dei lemmi presi in esame siano distribuite su classi ontologiche diverse. Escludendo dal computo i 219 lemmi che hanno una sola accezione, si è proceduto a osservare per ciascun lemma la sua capacità di essere associato nelle sue accezioni a diverse classi ontologiche.

Tabella 4. *Varietà delle classi ontologiche per lemma*

Classi ontologiche	**Lemmi**	**Media nr. accezioni**	**% sul totale**
1	182	2,91	20,38%
2	313	3,33	35,05%
3	188	5,00	21,05%
4	108	7,36	12,09%
5	51	8,49	5,71%
6	26	10,62	2,91%
7	18	12,28	2,02%
8	5	17,40	0,56%
10	1	21,00	0,11%
12	1	26,00	0,11%
Totale	**893**		**100,00%**

Complessivamente, come si vede in Tabella 4, vi è una relazione proporzionale tra il numero delle accezioni e la varietà di classi ontologiche. La popolosità delle classi invece non è correlata alla varietà.

I lemmi che sono monocategoriali sono solamente 182 (ossia poco più del 20%): *balcone, calza, coltello, ingegnere, ministro, nipote, principessa, rabbia*. E

[7] Infatti tra i non sensibili, l'unica categoria che funziona senza grandi problematicità è QUALITÀ TEMPORALE, che raccoglie astratti fortemente convenzionalizzati socialmente come varie accezioni prototipiche di: *anno, aprile, attimo, avvenire, domani, domenica, epoca, momento, secolo, settembre, settimana, tempo, termine*.

per quanto riguarda le classi ontologiche più popolate da lemmi mono-categoria sono, in ordine decrescente: PERSONA (52), ARTEFATTO (27), IDEA (18), AZIONE (14). La maggioranza dei lemmi si attestano su uno, due o tre classi ontologiche, con agli estremi i casi di *punto* (8 classi distribuite su 22 accezioni), *storia* (10 classi per 18 accezioni), *vita* (12 classi per 26 accezioni) che, oltre ad avere un numero molto alto di accezioni, presentano una altissima varietà di classi ontologiche di appartenenza.

4. DISCUSSIONE

Uno dei problemi di fondo che è necessario affrontare nel momento in cui si vogliano collegare risorse lessicali e ontologie è la forte assiomatizzazione di queste ultime, in confronto al carattere sfumato delle prime. L'operazione proposta in questo lavoro è una sorta di ribaltamento problematico dell'applicazione di una ontologia, la quale impone solitamente impegni di esplicitazione e applicazione che qui volutamente sono stati messi da parte. Si è cercato di vedere, attraverso l'applicazione 'ingenua', di senso comune, delle classi dell'ontologia, quali siano i problemi soggiacenti a tale applicazione e quali i luoghi in cui i meccanismi di categorizzazione svolgono un ruolo di filtro centrale.

Un problema che emerge fortemente dalla sperimentazione è quello della relazione complessa, non omologica, che si istituisce tra l'ontologia, il lessico e i processi cognitivi di categorizzazione che servono a mettere in relazione i due piani. La centralità della dimensione cognitiva (Evermann, 2005) si rivela anche e soprattutto nel momento in cui sperimentatori umani si ingaggiano in una operazione esplicita di categorizzazione. Emergono da questo punto di vista alcune interessanti regolarità che possono e interpretate nello sfondo della linguistica cognitiva (Berlin e Kay, 1969, Markman e Wisniewski, 1997, Rosch e Lloyd, 1978, Rosch, *et al.,* 1976, Ungerer e Schmid, 1996). In particolar modo affiora, anche nel caso di categorizzazioni molto generali, come quelle qui applicate, una preferenza per un livello basico di categorizzazione, che coinvolge categorie culturalmente salienti, che raggruppano caratteristiche di idealizzazione più prototipiche e largamente riconosciute. Le categorie basiche mostrano infatti una maggiore numerosità rispetto a quelle sovraordinate.

Sono emersi inoltre specifici problemi di classificazione, laddove il modello ontologico difettava di categorie soddisfacenti: esempi di questo tipo sono i prodotti scientifici, le unità di misura, gli oggetti semiotici (accezioni di lemmi come *messaggio, discorso, scritto, biglietto*); gli usi pragmatico-discorsivi (*guarda caso! Dio mio!*). Questioni complesse vengono anche aperte con l'associabilità di diverse classi alla stessa accezione in situazioni piuttosto regolari. Casi di questo tipo sono i gruppi luogo/artefatto/funzione (per accezioni di *bagno, albergo*, ecc.), ma anche luogo/oggetto naturale (*bosco* come "terreno", ecc.), processo/azione/risultato, sostanza/artefatto.

Si aprono dunque alcune strade da percorrere in parallelo. Dal punto di vista teorico generale: osservare le condizioni di possibilità della mappatura tra usi lessicali e ontologie; conciliare gli scopi lessicografici tradizionali con l'integrazione in applicazioni computazionali, tenendo conto della complessità e indeterminatezza dei confini tra i sensi. E inoltre interessante sarebbe proiettare la classificazione in maniera controllata sui testi, passando dal livello sistemico a quello dell'uso effettivo. Per quanto riguarda l'ontologia, gli obiettivi saranno: affinare l'ontologia in modo da permettere classificazioni agevoli delle classi più problematiche, le astratte specialmente non sensibili; si sta valutando inoltre se permettere la marcatura con più di una classe ontologica, laddove necessario, e se indirizzarsi verso una categorizzazione sovraordinata. Per quanto riguarda la risorsa lessicale si dovrà riflettere sull'opportunità di dividere alcune accezioni in sottoaccezioni che tengano conto delle diverse classi ontologiche; confrontare l'applicabilità dell'ontologia formale in senso stretto con le etichette applicate mediante classificazione di senso comune e separare i criteri di 'confidenza' e 'concordanza' dell'annotazione delle classi. Dal punto di vista tecnico e di usabilità sarà necessario riflettere sulle condizioni di miglioramento degli strumenti tecnici e degli strumenti di supporto e/o eventuale addestramento degli utenti nei compiti di classificazione.

Riferimenti Bibliografici

Berlin Brent, Kay Paul, *Basic color terms: their universality and evolution*, Berkeley, Oxford, University of California Press, 1969 (1991).

Berners-Lee Tim, Hall Wendy, Hendler James A., O'Hara Kieron, Shadbolt Nigel, Weitzner Daniel J., *A Framework for Web Science.* «Foundations and Trends in Web Science», a. 1, n. 1, 2006, pp. 1-130.

Berners-Lee Tim, Hendler James, Lassila Ora, *Semantic Web*, «Scientific American Magazine», May, 2001, pp. 29-37.

Brewster Christopher, O'Hara Kieron, *Knowledge representation with ontologies: Present challenges-Future possibilities*, «International Journal of Human-Computer Studies», a. 7, n. 65, 2007, pp. 563-568.

De Mauro Tullio, *Introduzione alla semantica*, Bari, Laterza, 1965.

De Mauro Tullio, *Guida all'uso delle parole*, Roma, Editori Riuniti, 1980.

De Mauro Tullio, *Grande Dizionario Italiano dell'uso*, Torino, UTET, 1999.

De Mauro Tullio, *Dizionario della lingua italiana*, Milano, Paravia, 2000.

De Mauro Tullio, *Basi di conoscenze e banche dati lessicali*, Istituto della Enciclopedia Italiana. XXI secolo. Comunicare e rappresentare, Istituto della Enciclopedia Italiana, Roma, 2009, pp. 253-308.

Evermann Joerg, *Towards a cognitive foundation for knowledge representation*, «Information Systems Journal», 15, 2005, pp. 147-178.

Evermann Joerg, Fang Jennifer, *Evaluating ontologies: Towards a cognitive measure of quality*, «Information Systems», a. 4, n. 35, 2010, pp. 391-403.

Fellbaum Christiane, *WordNet: an electronic lexical database*. Cambridge, Mass, MIT Press, 1998.

Gangemi Aldo, Navigli Roberto, Velardi Paola, *The OntoWordNet Project: Extension and Axiomatization of Conceptual Relations in WordNet*, «Proceedings of On the Move to Meaningful Internet Systems OTM2003», Catania, Springer Verlag, 2008, pp. 820- 838.

Hu Bo, Dasmahapatra Srinandan, Lewis Paul, Shadbolt Nigel, *On capturing seman- tics in ontology mapping*, «Proceedings of the 22nd national conference on Artificial intelligence», Vancouver, British Columbia, Canada, 2007, pp. 311-316.

Markman Arthur B., Wisniewski Edward J., *Similar and Different: The Differentiation of Basic-Level Categories*, «Journal of Experimental Psychology: Learning, Memory, and Cognition», a. 1, n. 23, 1997, pp. 54-70.

Masolo Claudio, Gangemi Aldo, Guarino Nicola, Oltramari Alessandro, Schneider Luc, *Sweetening Ontologies with DOLCE*, «Knowledge Engineering and Knowledge Management. Ontologies and the Semantic Web, 13th International Conference, EKAW 2002», Siguenza, Spagna, 2002, pp. 166-181.

Prévot Laurent, Oltramari Alessandro, Borgo Stefano, *Interfacing ontologies and lexi- cal resources,* in Huang Churen, Calzolari Nicoletta, Gangemi Aldo, Lenci Alessandro, Oltramari Alessandro, Prévot Laurent (a cura di), *Ontology and the Lexicon*, Cambridge, Cambridge University Press*,* 2010, pp. 229-245.

Oltramari Alessandro, Vetere Guido, *Lexicon and Ontology Interplay in Senso Comune,* «Proceedings of OntoLex 2008» (Workshop on Ontologies and Lexical Resources), 6th International Conference on Language Resources and Evaluation, Marrakech, Marocco, 2008.

Oltramari Alessandro, Vetere Guido, Lenzerini Maurizio, Gangemi Aldo, Guarino Nicola, *Senso commune*, «Proceedings of LREC 2010 7th International Conference on Language Resources and Evaluation», May 17-23, Valletta, Malta, 2010.

Picardi Eva, *Le teorie del significato*, Roma, Laterza, 1999.

Rosch Eleanor, Lloyd Barbara B., *Cognition and categorization*, Hillsdale, Erlbaum, New York, London, Wiley, 1978.

Rosch Eleanor, Mervis Carolyn B., Gray Wayne D., Johnson David M., Boyes-Braem Penny, *Basic objects in natural categories*. «Cognitive Psychology», 3, 8, 1976, pp. 382-439.

Ungerer Friedrich, Schmid Hans-Jo, *An introduction to cognitive linguistics*, Harlow, Longman, 1996.