

AN OPEN KNOWLEDGE BASE FOR ITALIAN LANGUAGE IN A COLLABORATIVE PERSPECTIVE

Chiari I, A. Gangemi, E. Jezek, A. Oltramari, G. Vetere, L. Vieu



<http://www.sensocomune.it/>

- Sapienza Università di Roma
- Université Paris 13 Sorbonne Paris
- Università di Pavia
- Carnegie Mellon University
- IBM Center for Advanced Studies
- IRT-CNRS, Université Toulouse III

INTRODUCTION

- The notion of document has undergone significant changes in the last ten years going toward an integrated approach that includes all kinds of data (text and multimedia), and giving strong relevance to the way textual and non textual data are organized, annotated, interlinked and managed in open and closed environments. The area of computational lexicography and that of computer-aided traditional lexicography has been deeply affected by recent changes in IT and by the possibilities of integrating different representations of linguistic data in a single environment giving the user (human and machine) different interpretative perspectives over the same data sources.
 - **WordNet**
 - **FrameNet**
 - **VerbNet**
 - **RDF- and OWL-based implementations**
 - **MultiWordNet**
 - **ItalWordNet**
 - link between semantic entities from lexica, like senses, synsets, frames, roles, etc., and semantic entities from ontologies.

OUTLINE

- General introduction to the Senso Comune project, its purpose, features and current state of development and
 - how the different layers are integrated in the general architecture of SC
 - how the architecture is represented in the platform and in the downloadable resource
- Exemplify and discuss how collaborative annotation is currently used for tagging different kind of data and what are the tools developed in order to support annotation

SENSO COMUNE: THE PROJECT

- started in **2006** as a collective initiative promoted by a group of researchers, lead by linguist Tullio De Mauro, interested in the development of an open lexical resource for the Italian language.
- Among those who contributed to the design and development of Senso Comune : Tullio De Mauro (President), Guido Vetere (vice-President), Diego Calvanese, Isabella Chiari, Aldo Gangemi, Nicola Guarino, Elisabetta Jezek, Maurizio Lenzerini, Malvina Nissim, Alessandro Oltramari, Laure Vieu, Fabio Massimo Zanzotto.
- **Senso Comune** (literally “common sense”, but more specifically intended as “common semantic knowledge”) is not conceived merely as an electronic dictionary but as a knowledge base where different kinds of data are integrated, connected and annotated
- Offer formalized representation of linguistic knowledge such as lexical and morphological information, semantic specifications through ontologies, and thematic roles and frames. Senso Comune is devoted to the distribution of linguistic data in an open and standardized form.
- The resource is available for download in a specific XML format under a Creative Commons Attribution-Non Commercial-Share Alike 3.0 License.

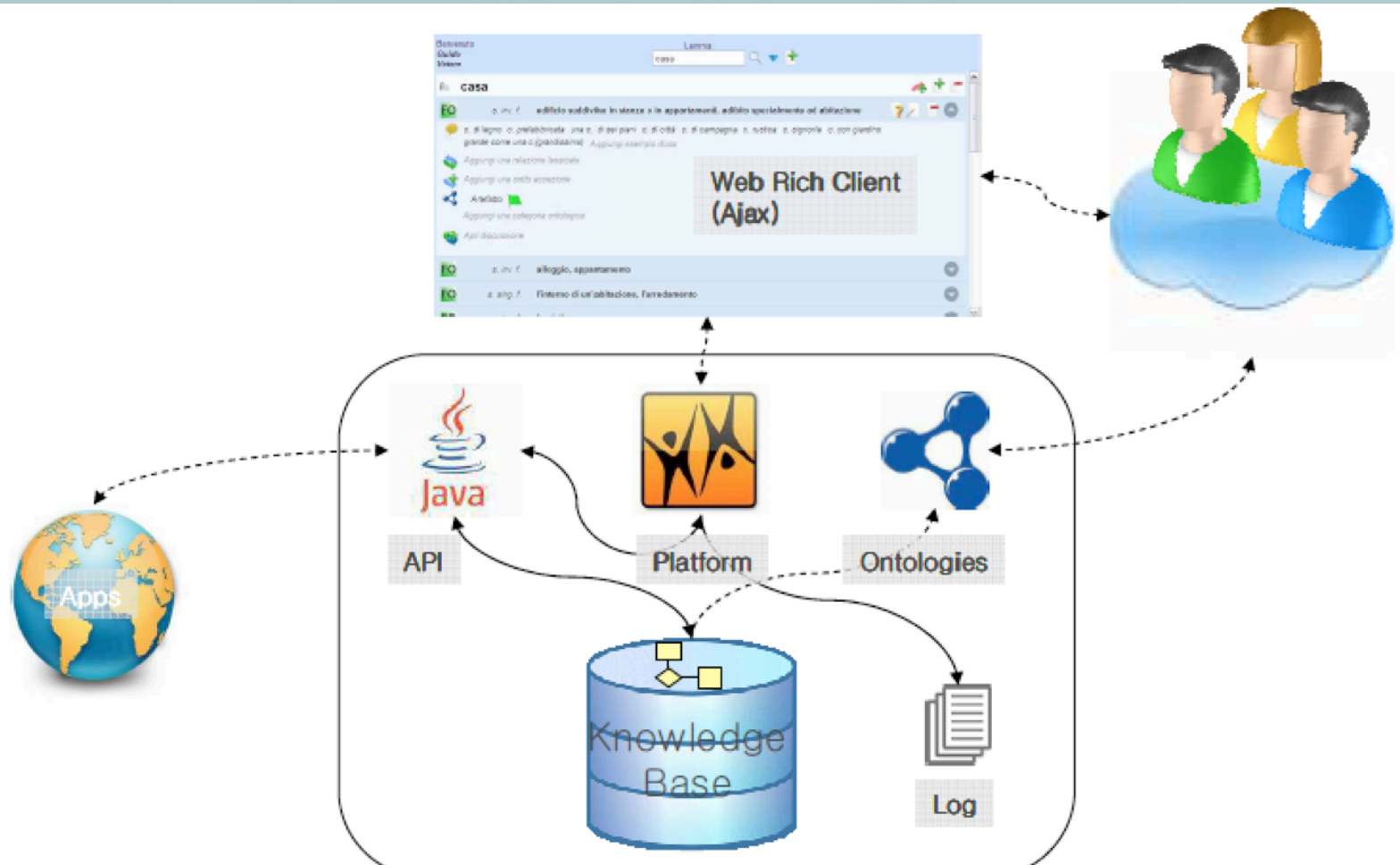
BUILDING THE RESOURCE

- The starting point of the resource was the so-called fundamental vocabulary of Italian language, containing the top-ranked **2,000 lemmas** (and about **13,000 senses**) extracted from frequency lists of written and spoken Italian and covering about **90%** of the occurrences of any written or spoken text.
- remodelled in order to fit into a formal representation (meta-model) specifically designed for integrating lexical and ontological information
- shared online in 2009, hosted by the Center for Advanced Studies of IBM Italia
- the resource was later enriched with the classification of **4,586 senses of basic nouns** (1,111 nouns) by means of a small set of predefined ontological categories.
- The resource has been recently integrated with the Italian version of **MultiWordNet**, whose content has been arranged to fit in the Senso Comune meta-model.

THE WEB-BASED PLATFORM

- the platform manages an information system structured around a set of Java modules, which link the one another by way of programming interfaces (API).
- Application modules are arranged in two architectural layers: the front-end (which handles users' interactions) and the back-end, which manages data accesses and transactions.
- The database where the resource is stored has a schema which efficiently supports complex queries, such as, for instance, looking for all the lexical entries whose meanings are classified with a given ontology concept. Also, integrity of data records is preserved, based on RDBMS standard functionalities.
- An Open Source object-relational mapping utility (Hibernate) ensures a smooth alignment of data records with programming structures.

SC: THE PLATFORM



SC: THE MODEL

- Senso Comune's model is specified in a set of “networked” ontologies comprising
 - **a top level module**, which contains basic concepts and relations
 - a small set of ontological categories relevant for lexical semantics are drawn. In particular, we developed a simplified OWL-lite version of **DOLCE** (Descriptive Ontology for Linguistic and Cognitive Engineering) (Gangemi et al., 2002)
 - **a lexical module**, which models general linguistic and lexicographic structures
 - **a frame module** providing concepts and axioms for modelling the predicative structure of verbs and nouns.
- These ontologies have been given an OWL specification, thus leveraging the expressive power of Description Logic (DL)

ANNOTATION AND EVALUATION

- The first enrichment procedure performed on the lexical resource was the association of each of 4,586 word senses
- The experimentation was conducted by a group of graduate students of computational linguistics who worked separately on different sets of word senses and further discussed classification problems.
- To enrich the knowledge base, though, language **users have been given access to the lexical level only.**
- The experimentation conducted using the Senso Comune platform was carried out in three phases:
 - (I) **Unsupervised common sense classification;**
 - (II) **Revision** of the classification (lead by Chiari, Vetere and Oltramari and four students) with the additional task of giving a confidence evaluation to the classification using three tags (accepted, controversial, not accepted) and discussion;
 - (III) **Final revision of consistency** in classification actions

EXAMPLE TASK

- PESO (noun)
 - “corpo soggetto alla forza di gravità” (*caricare, portare, sollevare pesi*) **OBJECT**
 - “oggetto, specialmente metallico, opportunamente graduato che serve nelle operazioni di pesatura” **ARTIFACT**
 - “senso di pesantezza dovuto a cattiva digestione” **PHYSICAL STATE**
 - “condizione, situazione che reca disagio, fastidio, sofferenza fisica o morale” **MENTAL STATE**
 - “autorità, prestigio, influenza” (*il peso di una posizione sociale, il peso del casato*) **IDEA**

CATEGORIZATION AND META-COGNITIVE AWARENESS

- Since ontological categorization is not a simple task and involves complex metalinguistic and cognitive operations significant control check strategies were introduced by giving experimenters the possibility of associating a **confidence labels** to their choices asserting whether their classification was perceived as fully confident or problematic or ultimately very uncertain.
- We further checked **inter-annotator agreement**, and observed what categories and association tasks were mostly accepted as common by different annotators, what produced more disagreement, and what were perceived as hazardous.
- **OBJECTIVE**
 - Obtain word sense annotation
 - Critical cases and scheme discussion
 - Methodological and theoretical insight
 - Multicategorization
 - Ontological categories and their abstractness

Lemma: terra

aneta su cui si svolge la vita dell'uomo, terzo in ordine di distanza dal Sole attorno al quale percorre la sua orbita di rivoluzione

gli abitanti della T. [Aggiungi esempio d'uso](#)

lessicale

zione

ontologica

insieme delle

traferma, in c

sta estensione di terreno, regione, territorio

zione, paese, patria

Valuta la classificazione per l'accezione

☐  Classificazione verificata

☒  Classificazione in fase di discussione

☐  Classificazione problematica

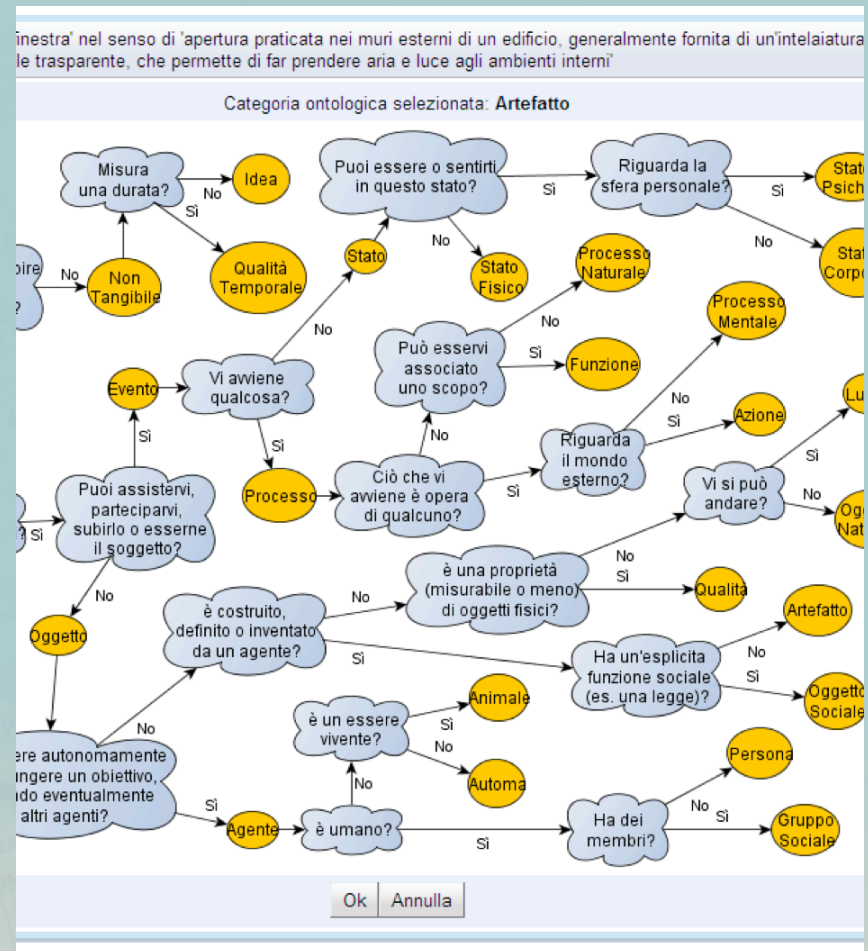
 

COLLABORATIVE ANNOTATION

- Collaborative annotation was thus achieved by the inclusion of discussion strategies that aimed at increasing cooperativeness and awareness:
- (a) collective discussion,
- (b) explicit annotation of controversial cases and
- (c) confidence evaluation
- (d) man-machine tutoring tools

TMEO (TUTORING METHODOLOGY FOR THE ENRICHMENT OF ONTOLOGIES)

- TMEO is a classification system based on broad foundational distinctions from DOLCE-Spray, a simplified version of DOLCE, which can be implemented with a synoptic map or with a sequential question answering procedure: it helps the user/editor to select the most adequate category of the reference ontology as the super-class of the given lexicalised concept: different answer paths lead to different mappings between the lexicon and the (hidden) ontological layer



TMEO: QUESTION ANSWERING PROCEDURE



Classificazione di 'cane' nel senso di 'animale domestico molto comune, diffuso in tutto il mondo, usato per la caccia, la difesa, nella pastorizia, come animale da compagnia o per altre attività'

Entità


Puoi percepire cane con almeno uno dei cinque sensi?





☐ Y

☐ N

Non classificato  


Aggiungi una categoria ontologica

 *Apri discussione*

FO s. m. sing. in espressioni negative, nessuno    

2011 Jan 29 16:32:27 Visualizzati 1 risultati su 1 totali v. 1.9.0

Senso Comune contiene attualmente:



- 31945 lemmi; di cui 2059 con almeno un'accezione
- 14046 accezioni; di cui 13161 fondamentali 7 di alto uso 73 di alta disponibilità 188 comuni 412 tecnico specialistiche
- 586 relazioni lessicali

CURRENT ANNOTATION ENRICHMENT

- Current work focuses of enriching the resource with data-induced verbal frames. The target corpus consists of about **8,000 usage examples** associated with the fundamental senses of the **verb lemmas** in the resource.
- The annotation task involves tagging the usage instances with **syntactic** and **semantic** information about the participants in the frame realized by the instances.
- Users are given:
 - a hierarchical taxonomy for semantic roles
 - Definitions and examples
 - Decision trees
 - TMEO tutoring

ANNOTATION OF THE SEMANTIC ROLE

**FO**

v.tr. seguito da aggettivo o da complemento di argomento, venire a conoscenza tramite la lettura

**FO**

v.tr. amare la lettura, dedicarsi a essa



nel tempo libero #0 leggo gli italiani leggono poco è uno che legge moltissimo [Aggiungi esempio d'uso](#)



[Aggiungi una relazione lessicale](#)



[Aggiungi una sotto accezione](#)



[Apri discussione](#)

**FO**

v.tr. pronunciare a

**FO**

v.tr. intendere, int

**FO**

v.tr. di un manosc

**FO**

v.tr. analizzare, in

**FO**

v.tr. decifrare, interpretare segni o scritture non

**FO**

v.tr. figurato intuire, indovinare pensieri, intenzior

**FO**

v.tr. figurato indovinare i pensieri più intimi e ripos

**FO**

v.tr. FO - fondamentali prevedere il futuro o presun

Esempio d'uso:

gli italiani leggono poco

Commento:

Conferma

gli italiani

☐ Predicato

Tipo di sintagma:

Nominale

Dipendenza sintattica:

Soggetto

Ruoli tematici:

Agente



[Aggiungi una ruolo tematico](#)

Categoria ontologica:

Persona

gli

italiani

italiano : nativo o abitante dell'Italia

COLLABORATION IN SENSO COMUNE

- To build a semantic resource through a cooperative process, Senso Comune follows two main paths:
 - a) axiomatized top-level ontological categories and relations are introduced and maintained by ontologists in order to constrain the formal interpretation of lexicalised concepts (top-down direction);
 - b) users are asked to enrich the semantic resource with linguistic information through a collaborative approach (bottom-up direction).
 - A special tutoring system supports users by interacting with users on the basis of question-answering mechanisms (i.e. TMEQ)
- Web-based cooperative platform. The platform shares a number of key features with wikis:
 - Editing through browser
 - Rollback mechanism (versioning of saved changes is available, so that an incremental history of the same resource is maintained)
 - Controlled access and different role read/write privileges
 - Collaborative editing and discussion forum support
 - Emphasis on linking
 - Search functions

SC AND WIKIS

- a rich interactive and WYSIWYG Web interface that is tailored to linguistic content;
- annotations are encoded directly into the Senso Comune meta-model;
- support of multiple annotation and conflict on annotation agreement signalling;
- support of annotation processes with the aid of the tutoring system supports on the basis of question-answering mechanisms – i.e. TMEO (A Tutoring Methodology for the Enrichment of Ontologies).
- Senso Comune presently allows a **number of annotation actions and user generated content**: adding new lemmas, adding new word senses, associating word senses to multiple ontological categories, adding usage examples, annotating usage examples with syntactic and semantic information about the participants in the frame realized by the instances, including argument/adjunct distinction, adding and modifying lexical relations (synonymy, hyponymy, hyperonymy, meronymy, holonymy) to word senses.

THE ROLE OF COLLABORATIVE ANNOTATION IN THE PROJECT

- Collaboration is also enhanced by the annotation of **self-evaluative notes** regarding the confidence annotators pose in the single tagging action performed, stimulating self-reflection and metalinguistic and metacognitive awareness and also discussion among different annotators.
- Collaborative annotation is conceived within Senso Comune as a productive **restructuring method for the overall architecture of the web-based environment**, but also for the progressive refinement and reorganisation of the **theoretical** issues posed by the **integration of the three main representation modules**. Being annotation an interpretative task, it is a primary source for observing phenomena and re-planning links and relationships among layers.

THANK YOU!

<http://www.sensocomune.it/>

