Senso Comune
a Community Knowledge Base for the Italian Language
Introduction

- Senso Comune is an ongoing project to build open knowledge base of the Italian language
- Collaborative research initiative freely supported by a multi-disciplinary community
  - Univ. of Rome ‘Sapienza’ and ‘Tor Vergata’, Bologna, Bolzano, Pavia, Trento, et al.
  - Italian National Research Council (CNR) ISTC, ILC
  - Fondazione Bruno Kessler, Trento
  - IBM Center for Advanced Studies Italy
- Non-profit organization led by Prof. Tullio De Mauro [www.sensocomune.org](http://www.sensocomune.org)
Objectives

- Build an open, collaborative knowledge base of the Italian language (i.e. a possibly incomplete database with a schema specification that allows automated reasoning)
- Collect information from both dictionary sources and skilled people (scholars, researchers, practitioners, etc)
- Formalize linguistic knowledge
  - Morphological and lexical information
  - Semantic specifications through ontologies
  - Thematic roles and frames
- Develop a specific platform for linguistic knowledge acquisition
- Distribute open, standardized linguistic data
Approach

- Start from the core lexicon of De Mauro’s dictionary
  - ~2,000 most common Italian lemmas (90% coverage)
  - ~13,000 senses
- Allow (qualified) users to enrich / modify the content
  - New lemmas, senses, usage instances, lexical relations, etc
  - Ontological classification of each sense
- Complete the coverage
  - ~100,000 lemmas
  - ~200,000 senses, including technical ones
Lexicon and Ontology

- Specific focus on the lexicon – ontology interplay [Oltramari and Vetere, 2008]
  - To what extent linguistic senses bear ontological commitments?
- Our position:
  - Linguistic constructs ≠ truth-valued logic constructs
  - Linguistic senses and ontology concepts are (in principle) disjoint
  - A (partial) mapping function leads from senses to concepts
  - Lexical relationships (e.g. synonymy, hyponymy) are not immediately (nor necessarily) reflected in ontological axioms (e.g. equivalence, inclusion)
- Main differences w.r.t. WordNet
  - A-priori ontological backbone
  - Clear distinction between senses and concepts
  - Formal and focused account of the conceptual level
  - Elements of frame semantics
The Ontology Behind Senso Comune

- Inherits from DOLCE (CNR ISTC)
  - "Nominalistic" subset (no universals)
  - Reified classes and relationships (characterizations)
- Modules
  - Foundational ontology
  - Morpho–syntactic structure
  - Semasiological structure
  - Semantic Relations and frames
Semasiological Model

Abstract Entity

Information Object

Lemma

Meaning Record

Water

Expression

“A liquid etc.”

Water-1

{ Noun }

Linguistic Property

definition

part

lexical relation

gram.spec.

Characterization

Substance

Concrete Entity

Meaning

Action

Drink-

Water-

Water-#

“Drink- is Meaning and characterizes only Action”

generated from categorized meaning records

mapping punning

usage instance

Part

annotations

“The boy [drinks] a glass of water”

Drink-#

Water-1 ⊆ Meaning and characterizes only Substance

generated from categorized meaning records

categorization

categorization

{ Noun }

Characterization

gram.spec.

Part

definition

usage instance

“A liquid etc.”

Water

expression

Part

Lexical relation

gram.spec.

Part

Definition

Expression

“A liquid etc.”

Water
Semasiological Model

- Separating linguistic senses and relationships (e.g. synonymy, hyponymy, and antinomy) from their ontological counterparts (e.g. inclusion, disjointness) is at the basis of our model. This separation prevents linguistic facts to be directly mapped to logic propositions, thus relieves linguistic meanings the burden of embodying ontological commitments [Vetere and Oltramari, 2008].

- We distinguish between meanings as registered in dictionaries from the concepts they refer to (if any). The former are instances of the class MeaningRecord (InformationObject), while the latter are subclasses of Meaning (Characterization). Basically, MeaningRecords are instantiated in Dictionaries, while Meaning classes are instantiated in linguistic acts/texts.

- Annotating a MeaningRecord instance with an Ontology class amounts at introducing a Meaning subclass which is restricted to that class.

- Mapping between MeaningRecords (instances) and Meanings (classes) can be done by annotations, punning, etc. In any case, formal semantics of mappings can be specified in different ways.

- Lexical relations are predicated on MeaningRecords; hence they do not have a direct ontological import. Any correspondence (e.g. hyponymy > inclusion) should be introduced based on suitable heuristics. Also, attributes of MeaningRecords instances (e.g. glosses, grammatic features, usage marks, rhetoric marks, etymology, etc) do not affect the mapped Meaning class (if any).

- Different MeaningRecords instances (e.g. from different dictionaries) may be mapped to the same Meaning class. This way, the model may accommodate meaning records coming from different sources, that might use different sets of attributes (e.g. different usage marks).
Supporting Senses Classification with TMEO

[Oltramari, 2010]
Technicalities

- Description Logic underlying all modules
  - Formal semantics and decidability
  - Well understood computability / expressiveness
  - Compliance with standards of knowledge representation and automated reasoning

- Native OWL 2 specification
  - http://www.sensocomune.org/ontologies/SensoComune.owl
  - http://www.sensocomune.org/ontologies/SensoComuneLexicon.owl
  - http://www.sensocomune.org/ontologies/SensoComuneSemantics.owl

- UML derivative models to map with Java and Relational DBMS
Architecture
Screenshots
An Experiment

- About 4,500 core substantival senses were classified by students and supervisors
- Identifying the ontological commitment in linguistic senses turned out to be hard in many cases
- Confidence of classifications was rated
  - 59% accepted
  - 33% controversial
  - 8% rejected
- Most controversial concept: SOCIAL_OBJECT

[Chiari, Oltramari and Vetere, 2010]

<table>
<thead>
<tr>
<th>CLASSE</th>
<th>accettate</th>
<th>%</th>
<th>controversy</th>
<th>%</th>
<th>non accettate</th>
<th>%</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>PERSON</td>
<td>443</td>
<td>89.0%</td>
<td>52</td>
<td>10.4%</td>
<td>3</td>
<td>0.6%</td>
<td>498</td>
</tr>
<tr>
<td>ANIMAL</td>
<td>18</td>
<td>85.7%</td>
<td>3</td>
<td>14.3%</td>
<td>0</td>
<td>0.0%</td>
<td>21</td>
</tr>
<tr>
<td>NATURAL_OBJECT</td>
<td>177</td>
<td>83.9%</td>
<td>22</td>
<td>10.4%</td>
<td>12</td>
<td>5.7%</td>
<td>211</td>
</tr>
<tr>
<td>ARTIFACT</td>
<td>478</td>
<td>81.7%</td>
<td>86</td>
<td>14.7%</td>
<td>21</td>
<td>3.6%</td>
<td>585</td>
</tr>
<tr>
<td>PHYSIC_STATE</td>
<td>35</td>
<td>81.4%</td>
<td>6</td>
<td>14.0%</td>
<td>2</td>
<td>4.7%</td>
<td>43</td>
</tr>
<tr>
<td>ACTION</td>
<td>350</td>
<td>76.6%</td>
<td>98</td>
<td>21.4%</td>
<td>9</td>
<td>2.0%</td>
<td>457</td>
</tr>
<tr>
<td>TEMPORAL_QUALITY</td>
<td>130</td>
<td>76.5%</td>
<td>33</td>
<td>19.4%</td>
<td>7</td>
<td>4.1%</td>
<td>170</td>
</tr>
<tr>
<td>AGENT</td>
<td>18</td>
<td>75.0%</td>
<td>6</td>
<td>25.0%</td>
<td>0</td>
<td>0.0%</td>
<td>24</td>
</tr>
<tr>
<td>SUBSTANCE</td>
<td>83</td>
<td>74.1%</td>
<td>25</td>
<td>22.3%</td>
<td>4</td>
<td>3.6%</td>
<td>112</td>
</tr>
<tr>
<td>NATURAL_PROCESS</td>
<td>33</td>
<td>73.3%</td>
<td>12</td>
<td>26.7%</td>
<td>0</td>
<td>0.0%</td>
<td>45</td>
</tr>
<tr>
<td>CORPORAL_STATE</td>
<td>23</td>
<td>71.9%</td>
<td>8</td>
<td>25.0%</td>
<td>1</td>
<td>3.1%</td>
<td>32</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CLASSE</th>
<th>accettate</th>
<th>%</th>
<th>controversy</th>
<th>%</th>
<th>non accettate</th>
<th>%</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>SOCIAL_OBJECT</td>
<td>61</td>
<td>35.7%</td>
<td>96</td>
<td>56.1%</td>
<td>14</td>
<td>8.2%</td>
<td>171</td>
</tr>
<tr>
<td>ENTITY</td>
<td>16</td>
<td>51.6%</td>
<td>15</td>
<td>48.4%</td>
<td>0</td>
<td>0.0%</td>
<td>31</td>
</tr>
<tr>
<td>PLACE</td>
<td>94</td>
<td>53.1%</td>
<td>79</td>
<td>44.6%</td>
<td>4</td>
<td>2.3%</td>
<td>177</td>
</tr>
<tr>
<td>PROCESS</td>
<td>37</td>
<td>50.0%</td>
<td>31</td>
<td>41.9%</td>
<td>6</td>
<td>8.1%</td>
<td>74</td>
</tr>
<tr>
<td>FUNCTION</td>
<td>26</td>
<td>37.7%</td>
<td>28</td>
<td>40.6%</td>
<td>15</td>
<td>21.7%</td>
<td>69</td>
</tr>
<tr>
<td>EVENT</td>
<td>90</td>
<td>55.6%</td>
<td>65</td>
<td>40.1%</td>
<td>7</td>
<td>4.3%</td>
<td>162</td>
</tr>
<tr>
<td>STATE</td>
<td>68</td>
<td>45.3%</td>
<td>59</td>
<td>39.3%</td>
<td>23</td>
<td>15.3%</td>
<td>150</td>
</tr>
<tr>
<td>PSYCHIC_STATE</td>
<td>113</td>
<td>61.4%</td>
<td>64</td>
<td>34.8%</td>
<td>7</td>
<td>3.8%</td>
<td>184</td>
</tr>
<tr>
<td>MENTAL_PROCESS</td>
<td>49</td>
<td>62.8%</td>
<td>27</td>
<td>34.6%</td>
<td>2</td>
<td>2.6%</td>
<td>78</td>
</tr>
<tr>
<td>QUALITY</td>
<td>241</td>
<td>57.2%</td>
<td>140</td>
<td>33.3%</td>
<td>40</td>
<td>9.5%</td>
<td>421</td>
</tr>
<tr>
<td>SOCIAL_GROUP</td>
<td>53</td>
<td>64.6%</td>
<td>26</td>
<td>31.7%</td>
<td>3</td>
<td>3.7%</td>
<td>82</td>
</tr>
</tbody>
</table>
Work in Progress

- Ontology of Semantic Relations and Frames
  - Goals
    - Provide a formal (DL) characterization
    - Represent users’ linguistic knowledge
    - Support NLP tasks with efficient reasoning
  - Issues
    - Cope with higher-order features
    - Syntactic–Semantic binding requires co-reference?
References

- G. Vetere, 2008: Verso un lessico computazionale aperto per la lingua italiana, PAAL2008
Credits

- **Project coordination**: Guido Vetere (IBM)
- **Linguistics**: Isabella Chiari (Uni Roma Sapienza, coord.), Elisabetta Jezek (Uni Pavia), Fabio M. Zanzotto (Uni Roma Tor Vergata)
- **Logic and Ontology**: Diego Calvanese (Uni Boziano), Nicola Guarino, Aldo Gangemi (CNR ISTC), Maurizio Lenzerini (Uni Roma Sapienza), Alessandro Oltramari (Carnegie Mellon University), Guido Vetere (IBM)
- **Design and Development**: Alessandro Faraotti (coord.), Daniele Chermaz, Ilaria Gorgia, Michele Minno, Fabrizio Smith, Giuliano Iacobelli, Piero Cangialosi, Andrea Mencancini, Carlo Ferrari (IBM)

- **Founders**
  - Tullio De Mauro
  - Aldo Gangemi
  - Nicola Guarino
  - Maurizio Lenzerini
  - Malvina Nissim
  - Guido Vetere

- **Board**
  - Tullio De Mauro (President)
  - Diego Calvanese
  - Isabella Chiari
  - Aldo Gangemi
  - Nicola Guarino
  - Alessandro Oltramari (Secretary)
  - Guido Vetere (Vice President)