Sweetening Ontologies cont’d:
Aligning Bottom-up with Top-down Ontologies

Elisabetta Jezek
Università di Pavia

CREOL-JOWO
Medical University of Graz
Sept. 25, 2019
Content and goal of the talk

- We investigate an issue at the interface between language and ontology.
- We run an experiment in which we align the corpus-based (bottom-up) system of semantic types developed in the T-PAS resource with the upper-level foundational (top-down) ontology DOLCE.
- We limit the experiment to the Endurant domain.
- The goals is to highlight the distinctions and similarities between the two systems from a cognitive and application-based perspective.
- What we learned and future work.
By applying the methodology of Corpus Pattern Analysis (Hanks 2013) to the analysis of corpus evidence for about 1600 average polysemic Italian verbs, with the goal of acquiring their recurrent semantic structures (e.g. Human partecipa in Activity), we have compiled a list of 180 semantic types to characterize the semantic preferences of verbs for each argument position in each verb sense.
These semantic types (Event, Location, Food, Vehicle, etc.) are obtained from manual clustering of lexical items found in the argument positions of verbal structures in the corpus: they can thus be seen as human judgments about the selectional preference of verbs.
partecipare

[Human | Human Group] partecipare a [Activity]

[Human] | [Human Group] prende parte, contribuisce o semplicemente è presente a [Activity]
<table>
<thead>
<tr>
<th>Details</th>
<th>Left context</th>
<th>KWIC</th>
<th>Right context</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>#2695929 « Desidero esserci - ha aggiunto il premier - . &lt;/s&gt; &lt;/s&gt;</td>
<td>Parteciperò alla campagna elettorale al fianco del nostro candidato Rutelli e</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>#7242056 ogni suo errore , specialmente considerando il fatto che tu non</td>
<td>partecipare alla competizione e nessuno garantisce che saresti stato in gra</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>#13239303 038 Località : Lo spazio che c’è in te Cara Vale ... &lt;/s&gt; &lt;/s&gt; per</td>
<td>partecipare vedi le istruzioni della pagina precedente (è spiegato in manier</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>#16609005 LEA A tutti gli abitanti e &amp; # 047 ; o proprietari di via Garibaldi 2</td>
<td>Partecipate tutti all’assemblea condominiale che si terrà domenica 22.09.9</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>#17328061 ( tel. 0337/912703 ) Due sono le trasmissioni con giochi a cui si</td>
<td>partecipare inviando una cartolina postale con la risposta della settimana :</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>#24248633 di sono interessati studenti e professori che magari pensano di</td>
<td>partecipare al concorso negli anni successivi . &lt;/s&gt; &lt;/s&gt; Ovviamente essere</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>#25238821 « a Colignola e un’altra a Verona , in Via Stella , e alle ricerche</td>
<td>parteciparono militari americani di Verona e Vicenza con auto civili e furono m</td>
<td></td>
</tr>
</tbody>
</table>
Ontological categories vs. Semantic types

- These types look very much like ontological categories; however, instead of being stipulated, they are induced by the analysis of selectional properties of verbs.

- Despite the obvious correlations, the methodology underlying the identification of semantic types in T-PAS differs from the way categories are defined in resources such as the DOLCE ontology.

- While “aiming at capturing the ontological categories underlying natural language and human common sense” (cf. Masolo, Borgo, Gangemi, Guarino, Oltramari 2003) DOLCE does not derive the categories from systematic observation and clustering of linguistic data.
Research questions

- Are semantic types obtained through corpus analysis of selectional preferences of verbs similar to speculative categories defined primarily on the basis of axiomatization?
- If not, how do they differ from a cognitive and application-based perspective?
- Aligning the semantic type inventory of T-PAS to the categories of DOLCE.
DOLCE does not commit to a strictly referentialist metaphysics and aims at capturing the ontological categories underlying natural language and human commonsense (Gangemi et al. 2002).

It is not based on empirical evidence, but it has a formal structure defined on ontological principles and axioms that the T-PAS system of semantic types does not possess.
Top level of the T-PAS System of Semantic Types
with a selection of leaf types

- **Anything**
  - **Entity**
    - **Physical Entity**
      - **Inanimate**
        - **Artifact**
        - **Stuff**
        - **Light Source** [Location, Inanimate]
      - **Animate**
        - **Human**
        - **Human Group**
        - **Animal**
        - **Animal Group**
    - **Body**
    - **Part of Body**
    - **Plant**
    - **Location**
  - **Abstract Entity**
    - **Institution** [Abstract Entity, Human Group]
    - **Information Source**
      - **Document** [Artifact, Information Source]
- **Eventuality**
  - **Event**
  - **State**

**Property**
- **Colour**
- **Role**
- **Weight**
- **...**

Elisabetta Jezek Università di Pavia
Aligning Bottom-up with Top-down Ontologies
The starting point of the T-PAS taxonomy is the type **Anything**.

The top level has **Entity**, **Eventuality** (in Emmond Bach’s terminology) and **Property** as branches.

The main distinction in the domain of the **Entity** is between **Physical** and **Abstract Entity**.
Physical Entity is further distinguished in Inanimate, Animate, Body, Part of Body, Plant and Location.

Body, Part of Body and Plant are considered ambiguous with respect to animacy, and therefore classified as subtypes of Physical Entity.

Artifact forms a large and articulated branch of Inanimate (34 nodes in total), together with the sister note Stuff (17 nodes).
The system contains no type for **Natural Kind** (as opposed to **Artifact**) nor a type for **Individuated Entity** (as opposed to **Stuff**).

The prevailing distinction in the domain of **Physical Entity** is between **Animate** and **Inanimate**.

This finds motivation in the role that this distinction plays in language, in particular in defining the semantic preferences that verbs impose on their arguments.
The domain of EVENTUALITY has EVENT and STATE as main branches, whereas PROPERTY has, inter alia, COLOR, ROLE, and WEIGHT as subtypes.

The system includes multiple inheritance.

For our current purposes, we do not discuss the domains of EVENTUALITY and PROPERTY, and focus our attention on Physical entity.
DOLCE top level distinguishes between Endurant, Perdurant, Quality and Abstract.

An Endurant participates in a Perdurant: for example a person (Endurant) may participate in a discussion (Perdurant).

Qualities inhere to entities; every entity comes with certain qualities (color, smell, size, weight etc.), which exist as long as the entity exist.

Abstracts are entities with no spatial nor temporal qualities.
Within *Endurant*, DOLCE distinguishes between *Physical* and *Non-physical* (according to whether they have direct spatial qualities).

Within *Physical*, a distinction is drawn between *Amount of Matter*, *Object*, and *Feature*, based on the notion of Unity and the relation of Dependence.

*Object* are *Endurants* with Unity, *Amounts of Matter* are *Endurants* with no Unity (none of them is an essential whole).
Objects and Amounts of Matter are not dependent on other objects, while Features are dependent on another object, their host.

Examples of Features are Relevant Parts such as a bump, and Places such as a hole in a piece of cheese, the underneath of a table etc.

Physical Objects are divided into Agentive and Non-agentive according to whether or not they have intentions.

Agentive Objects are constituted by Non-agentive Objects: for example, a person is constituted by an organism.
Further Distinctions in DOLCE 3/3

- *Non-physical Objects* ("abstracts" in common parlance) are divided into *Social Objects* and *Mental Objects* according to whether or not they are generically dependent a community of agents.
- *Social Objects* are further divided into *Agentive* and *Non-agentive*.
- *Agentive Social Objects* are for example *Societies* such as *Sony*.
- *Non-agentive Social Objects* are *laws, norms, peace treaties* ecc., which are generically dependent on *Societies*. 
- **Endurant** live in time (and can therefore exhibit changes) by participating in a Perdurant -> ENTITY
- **Physical Endurant** have direct spatial qualities
  - Amount of Matter: Endurants with no unity, none of them is an essential whole, change identity when they change parts (merelyologically invariant) -> STUFF
    - SOLID
      - MATERIAL
        - CLOTH
        - THREAD
        - METAL
        - DUST
        - SOIL
    - FLUID
      - LIQUID
        - BEVERAGE [ARTIFACT, LIQUID]
        - ALCOHOLIC DRINK
        - WINE
        - WATER [BEVERAGE, LIQUID]
      - WATER
      - VAPOUR
        - GAS
        - AIR
        - SMELL
- **Physical Object** Endurants with unity, merelyologically variant, non dependent on other objects
- **Agentive** Endurants with intentions, constituted by non-Agentive Physical Objects (spatially co-localized with them) -> Animate
  - Human
  - Human Group
    - Institution [Human Group, Abstract Entity]
    - Business Enterprise
  - Animal
    - Cat
    - Cow
    - Horse
    - Dog
    - Sheep
    - Goat
    - Snake
    - Spider
    - Bird
    - Insect
    - Fish
  - Animal Group
- **Non-Agentive** Endurants without intentions
  - Inanimate
    - Artifact
    - Weapon
    - Bomb
    - Firearm
    - Beverage [Artifact, Liquid]
      - Alcoholic Drink
      - Wine
      - Water [Beverage, Liquid]
    - Food
    - Building [Artifact, Location]
    - Garment
    - Artwork
    - Movie [Artwork, Performance]
      - includes video
    - Musical Composition [Concept, Artwork]
      - Picture
    - Document [Artifact, Information source]
      - Agreement [Document, Speech Act]
    - Machine
      - Vehicle
        - Road Vehicle
        - Water Vehicle
        - Flying Vehicle
      - Computer
    - Device
      - Software
      - Container
      - Engine
      - Flag
      - Furniture
      - Image
      - Medium [Artifact, Abstract], e.g. radio,
          TV, the Press
      - Sound Maker e.g. alarm clock, bell
      - Musical Instrument
      - String
      - Ball
      - Drug
  - Body
  - Parts of the Body
  - Plant
  - Location
    - Natural Landscape Feature
      - Waterway includes lakes, the sea,
        rivers and streams
        - Waterway [Watercourse, Route] e.g
        canals, navigable rivers
        - Hill
        - Route includes roads, railways
          - Waterway [Route, Watercourse] e.g
          canals, navigable rivers
        - Area includes geographical area, e.g. states
        - Building [Location, Artifact]
        - Light Source [Location, Inanimate]
    - Feature parasitic entities constantly dependent on physical objects - their hosts (not spatially co-localized with them)
      - Relevant Part e.g. bump, damage
      - Place e.g. crack, hole, opening, window,
        doorway
      - Aperture

---

Mapping T-PAS onto DOLCE (excerpt)
DOLCE *Endurant* category is a node that aligns very well with the T-PAS organization.

DOLCE *Endurant* corresponds to *Entity* in CPA.

On the other hand, *Entity* is the label used in DOLCE for the top node, which corresponds to *Anything* in T-PAS. We regard Anything as a better term for the top node as Entity is often used in linguistics in a way which excludes Events.

*Anything* is T-PAS stands for all semantic types that play the role of participant in the event described by the verb selecting them (PARTICIPATION relation).
DOLCE *Physical Endurant* corresponds to *Physical Entity* in T-PAS; the internal organization of the two nodes, however, differs.

*Amount of Matter* is a sister node of *Physical Endurant* in DOLCE, while in T-PAS its closest equivalent *stuff* is a type of *physical entity (inanimate physical entity).*

It seems reasonable to move *stuff* (and its subtypes) higher in the T-PAS taxonomy.

The solution in DOLCE appears more adequate, as the animate/inanimate distinction apparently applies only to objects with Unity.
Endurants and the Object/Stuff distinction

- In T-PAS, **Body** and **Part of body** are child nodes of **Physical Entity**, and sister nodes of **Animate** and **Inanimate**.
- The **CONSTITUTION** relation, used in DOLCE for co-located entities, as in the case of a person (agent) and its organism (not agent), and the **PARTHOOD** relation, which defines the relation between a body and its parts, are not represented in T-PAS.
- The only relation between the semantic types is the **IS_A** relation.
- In the future it would be convenient to expand the relations in T-PAS to include **CONSTITUTION** and **PARTHOOD**.
Abstracts and the tangible/intangible distinction

- **Abstract entity** in T-PAS defines all intangible entities.
- DOLCE distinguishes among *Abstract* (entity without temporal qualities, such as mathematical objects) and *Non Physical Endurant* (entity with temporal properties such as *Mental* and *Social Object*);
- These two categories appear in different nesting levels.
There is no possible one-to-one alignment in this case.

From an applied perspective, the two DOLCE’s category can be conflated into T-PAS **Abstract entity** as the latter does not draw a distinction between intangible entities with or without temporal qualities.

Such a modeling decision, however, is far from being without consequences.
The *Agent* label is used in DOLCE for a potential agent, that is, a living being endowed with intentions, beliefs, and desires.

In T-PAS, agent is not present, as it is considered a role assumed by a human in an eventuality rather than a type - a thematic role in linguistic terms, which, according to Guarino 2017, corresponds to the processual role theorized by Loebe.

Therefore, the DOLCE *Agentive / NonAgentive PhysicalObject* distinction does not have a direct equivalent in T-PAS.
The closest type to which DOLCE’s AgentivePhysicalObject can be associated in T-PAS is ANIMATE.

In T-PAS animate subsumes, among others, HUMAN and HUMAN GROUP (squadra); it does not include PLANT but it includes the taxonomy of the animal kingdom (ANIMAL and ANIMAL GROUP).
- The animal kingdom differs from the scientific taxonomy of Linnaeus.

- T-PAS includes semantic types for animals for whom there exists a verb that selects the class or species as argument.

- Typically these are verbs of sound emission such as *to bark* (**dog**), or verbs of motion such as *to gallop* (**horse**).

- Linnaeus categories such as **mammal** are not present, as no verb has been identified yet that selects for it.
DOLCE assumes the category Feature for parasitic entities that are constantly dependent on physical objects (their so-called Hosts).

Feature subsumes Place (holes in a cheese) and Relevant part (bumps or edges).

T-PAS does not have a type that matches Relevant Part but has Aperture as a type of Location, which can be aligned to DOLCE’s Place category.
In T-PAS we find the semantic type Location, which is used for both natural places and artifactual ones (an island, a parking lot).

DOLCE has the category Place, which, however, does not correspond to T-PAS Location.

In DOLCE, the spatial dimension is considered a Quality of an entity (specifically Spatial Location > Spatial Region).
Locations

- There is therefore no direct mapping between the two systems as regards the type **Location**.
- From a linguistic point of view, the solution in T-PAS appears more apt to account for the geographical entities denoted by words that qualify as independent entities: *mountains, lakes, islands*, and so forth.
Neither DOLCE nor T-PAS draw a distinction between manufactured objects and natural, mind-independent entities.

T-PAS has Artifact as a type of Inanimate but does not have its counterpart natural kind.

DOLCE has neither one nor the other.

In the Entity branch of T-PAS and the Endurant branch of DOLCE the prevailing distinction is that between concrete and abstract, and between individuated (i.e. with Unity) and mass (without Unity).
The distinction between natural kind and artifactual type is orthogonal to the other categories: for example, **stuff** in T-PAS subsumes both natural entities (metal) and artifacts (cloth), **location** subsumes both natural entities (hill) and artifactual ones (route), and so forth.

From a linguistic perspective, the distinction between individuated and mass appears to be the most represented formally in the world’s languages.

The grammatical behavior of nouns appears to be primarily determined by their encoding as individuals or masses (Jezek 2016, 135).
T-PAS has **Food** and **Beverage** as types of artifact.

In (Guarino and Welty 2009, 218) it is observed that "nothing is necessarily food, and just about anything is possibly food".

Food is considered a role that an entity can play in a food event (roles being anti-rigid properties that characterize the way something participates to a contingent event).
While sharing this theoretical stance, in T-PAS it is believed that there being a large numbers of verbs selecting for the two types (currently 78 for FOOD and 11 for BEVERAGE) it is pragmatically useful to keep the two labels in the repertoire of semantic types.

This is also motivated also by the presence of artifactual food, that is, man-made entities which purpose is to be consumed as food.
Systematic polysemy is the phenomenon whereby a word or expression exhibits an alternation of meanings that is also exhibited by other words in the lexicon, so that this alternation can be considered regular (in Apresjan’s terms: cf. Jezek 2016 for an overview),

Examples are content/container in the case of *glass*, *dish*, *bottle* ('break a glass’ vs. ’drink a glass’), and physical object/information in the case of *book*, *letter*, *novel* (’The book is heavy to carry around’ vs.’ The book examines the life of Dante’).

In Jezek and Vieu 2014 we identified corpus evidence supporting the view that the second example is an instantiation of a particular kind of of systematic polysemy called inherent polysemy.
Systematic polysemies are currently treated in T-PAS through multiple inheritance, that is, a semantic type inherits from more than one type, and each subsumption relation implicitly represents one of the types that are conflated in the ambiguous class.

For example, the DOCUMENT type (instantiated by nouns such as libro 'book' and lettera 'letter') inherits from the artifact type and the information source type.

This is a case of cross domain multiple inheritance, as the two types are situated in different branches of the type system (PHYSICAL ENTITY and ABSTRACT ENTITY respectively).
Systematic polysemy is to my knowledge currently not represented in DOLCE.

Aparinis and Vieu 2015 propose to formally represent complex categories that overlap with disjoint domains of entities using the ontological relationship of CONSTITUTION and the notion of coincidence.
Although it is used in other lexical ontologies such as WordNet, multiple inheritance is normally avoided in formal ontologies.

It introduces incoherence and inconsistency from an ontological perspective and it creates problems for calculating inferences.

However, for the purposes of natural language processing task such as sense disambiguation we believe that multiple inheritance constitutes a valid ad interim solution to the problem of systematic polysemy, until a formal characterization is standardized.
- DOLCE is an upper level ontology, and for this reason it does not comprise fine-grained child categories.
- T-PAS comprises just as many child categories as they are required by verb selectional behaviour: the set of categories is not finite and may increase as long as new the analysis of new verbs requires new semantic types.
At present the deepest type in the system is the type wine (stuff > fluid > liquid > beverage > alcoholic drink > wine), motivated by a specific sense of the verbs *invecchiare* 'to grow old' and *maturare* 'to ripen'.

The largest node is Artifact, with 35 subtypes.
Child categories in T-PAS disclose the anthropocentric character of the type system, that does not reflect the state of the art in scientific knowledge but rather how everyday speakers communicate with each other and talk about the world (what they use when they do it).
As an example, the Artifact node includes: WEAPON (puntare 'point at’) and its subtype BOMB (denonare 'detonate’); three subtypes of VEHICLES (ROAD, FLYING and WATER), FOOD, BEVERAGE, BUILDING, GARMENT (indossare 'wear’), FURNITURE (arredare 'furnish’), and peculiar types such as FLAG (sventolare) and STRING (slegare 'untie’, stringere 'tighten’).
### System of Semantic Types

<table>
<thead>
<tr>
<th>Query</th>
<th>Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>detonare</td>
<td>2</td>
</tr>
<tr>
<td>esplodere</td>
<td>1</td>
</tr>
<tr>
<td>scoppiare</td>
<td>1</td>
</tr>
</tbody>
</table>
The exercise shows that the analysis based on linguistic evidence induces semantic types that can be linked to the upper level of a top-down ontology like DOLCE quite successfully, at least as far as the Endurant domain is concerned.
Concluding Observations

- One substantial issue emerge:
- The category abstract in T-PAS maps to two disjoint classes in DOLCE (Abstract and NonPhysicalEndurant) and there is no straightforward one-to-one alignment in this case;
Insights on the language/cognition interface.

The data-driven inventory of types in T-PAS is populated by semantic types that point to cognitive categories that are relevant to human communication, which do not necessarily match scientific classifications: hence the anthropic character of the T-PAS system.
Future work

- Completion of the alignment with DOLCE.
- Implementation of a distinct treatment for systematic polysemy and inherent polysemy?
Ongoing work

- Validation of corpus-derived semantic types against automatically obtained clusters of argument fillers in a distributional semantic framework.
Open question: are selectional preferences as identified through manual clustering of corpus evidence truthmakers of ontological categories?
Thank you for listening!
The Lexicon
An Introduction
Elisabetta Ježek
9780199601547
Paperback
28 January 2016
Oxford Textbooks in Linguistics