Sweetening Ontologies Cont'd: Aligning bottom-up with top-down ontologies

Elisabetta JEZEK [a]
[a] *University of Pavia, Department of Humanities*
Section of Theoretical and Applied Linguistics
jezek@unipv.it

Abstract. This paper addresses an issue at the interface between language and ontology. Specifically, we report the results of the alignment we performed between the T-PAS ([1]) and DOLCE categories ([2]), and discuss the distinctions and similarities we observed from a cognitive and application-based perspective. The motivation for our work lies in the different nature of the two resources; while T-PAS is a bottom-up system, in which semantic types are identified by manual clustering the fillers of argument positions of verbs gathered from large corpora, DOLCE is top-down ontology, in which categories are not based on extensive linguistic evidence and are stipulated on formal grounds. The preliminary results of the alignment task reveal that the most general types in T-PAS can be mapped fairly well into DOLCE's upper level. Two substantial issues remain open, the mapping of the Abstract category and the treatment of systematic polysemy. The experiment also assesses the anthropic character of the bottom-up system compared to the top-down system, and the fine-grained granularity of the first compared to the second. On the other hand, the taxonomy of DOLCE is ontologically more solid than the T-PAS hierarchy. The resulting alignment benefits both sides.

Keywords. semantic types, ontological category, endurant, corpus analysis, anthropic concepts

1. Introduction

By applying the methodology of Corpus Pattern Analysis ([3]) to the analysis of corpus evidence for about 1200 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.[1] These semantic types (EVENT, LOCATION, FOOD, VEHICLE, PART OF BODY, etc.) are obtained from manual clustering of lexical items found in the argument positions of verbal structures in the corpus:[2] they can thus be seen as human judgments abut the selectional preference of verbs as shown in the corpus. These type look very

[1] The analysis is carried out within the T-PAS (Typed Predicate-Argument Structures) project, which aims at building an inventory of semantic structures for Italian verbs for linguistic analysis and text processing ([1]).

[2] For example, for ACTIVITY in the structure above we have: *gara, riunione, selezione, manifestazione, seduta, cerimonia, conferenza, votazione, elezione, celebrazione, esequia, competizione, maratona, discussione, messa, festa*, and so forth.

much like ontological categories; however, instead of being stipulated, they are induced by the analysis of selectional properties of verbs, and reflect the way in which humans talk about events and states of entities through language. Despite the obvious correlations, the methodology underlying the identification of types in T-PAS differs from the way categories are defined in resources such as the DOLCE ontology (Descriptive Ontology for Linguistic and Cognitive Engineering), which, despite "aiming at capturing the ontological categories underlying natural language and human common sense' ([4]) does not derive the categories from systematic observation and clustering of linguistic data.

In this paper, we report the results of an experiment of alignment of the semantic type system developed in T-PAS to the DOLCE category system. The goals of this exercise are to verify the possibility of the alignment, and to identify the similarities and differences between a corpus-based and a speculative system of categories, highlighting the distinctions and similarities between the two from a cognitive and application-based perspective. It is expected that both resources can benefit from this analysis, and consequently be improved. We regards such issues of significant importance for applied ontology.

The structure of the contribution is as follows. Section 2 provides a description of the T-PAS semantic type system focusing on the aspects that are relevant for the present discussion; section 3 presents an overview of the content and structure of DOLCE; section 4 reports the results of the alignment between the two resources and the analysis of the results. Finally, section 5 reports the concluding observations together with directions for further research.

2. System of Semantic Types in T-PAS

The system of semantic types in T-PAS currently contains 180 semantic types (as of August 2019). The list is organized in a hierarchy in order to identify the appropriate level of specificity of the selectional properties of individual verbs. The taxonomic structure is based on human judgments, and on the manual comparison of the members of the set of lexical items that instantiate each type in the corpus. The only relation represented in the system is subsumption. The part_of relation is currently not represented. This has consequence as to how types such as PART OF BODY are currently positioned in the structure. Fig. 1 reports the top level of the system.

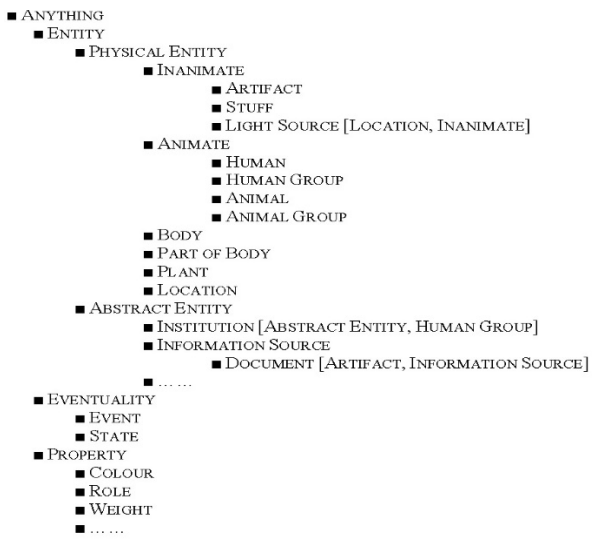

Figure 1. *Top-level* of the T-PAS system with a selection of *leaf* types

As shown in Fig. 1, the starting point of the hierarchy is the type ANYTHING. The top level has ENTITY, EVENTUALITY (in Emmond Bach's [5] 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, not visible in the Figure), 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 braches, whereas PROPERTY has, *inter alia*, COLOR, ROLE, and WEIGHT as subtypes. The system includes multiple inheritance, as with DOCUMENT, which is typed as an ARTIFACT and as an INFORMATIONAL SOURCE. For our current purposes, we do not discuss the domains of EVENTUALITY and PROPERTY, and focus our attention on PHYSICAL ENTITY.

3. DOLCE

Fig. 2 below depicts the structure of DOLCE, with focus on the *Endurant* branch.

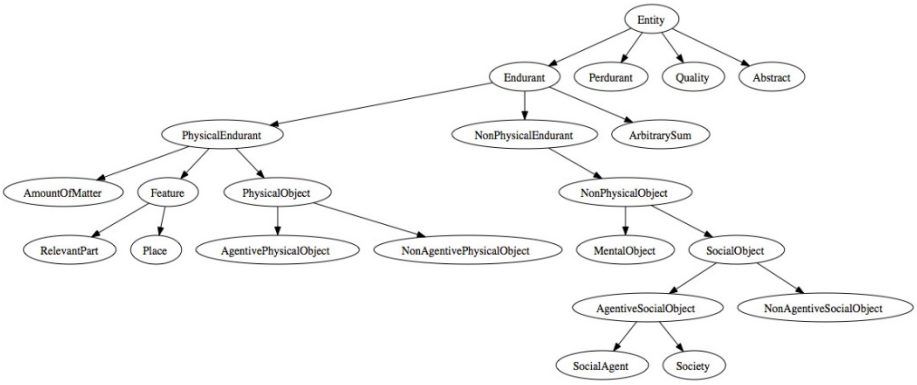

Figure 2. Taxonomy of DOLCE's categories (fragment)

The upper level of DOLCE distinguishes between *Endurant*, *Perdurant*, *Quality* and *Abstract*. While *Endurants* are entities that exist over time (objects), *Perdurants* are entities that happen over time (events and states). The main relation between the two is that of "PARTICIPATION": an *Endurant* exists at a given time by participating in a *Perdurant*. *Qualities* are inherent in the entities we can perceive or measure: shapes, colors, sizes, sounds, smells, and so on. They manifest themselves with entities and exist as long as the entity exists. *Abstracts* are entities without spatial or temporal qualities. Within *Endurants*, DOLCE distinguishes between *Physical* and *NonPhysical*, depending on whether or not they have direct spatial qualities. Child nodes of *Physical* are *AmountOfMatter*, *PhysicalObject* and *Feature*, distinguished on the basis of Unity and the relation of dependency (cf. [6, 7]). In particular, *PhysicalObjects* are *Endurants* with Unity, whereas *AmountsOfMatters* are *Endurants* without Unity. *PhysicalObject* and *AmountsOfMatter* do not depend on other objects, while *Feature* depends on another object, its "host". Examples of *Features* are *RelevantPart*, such as a bump in an object, and *Place*, such as a hole in a piece of cheese. *PhysicalObjects* are analyzed in *Agentive* and *NonAgentive* depending on whether or not they may have intentions. In DOLCE the *AgentivePhysical* are made of *Non-agentivePhysical*: for example, a person is made up of an organism ("CONSTITUTION" relation). *NonPhysicalObjects* ("abstract" in common parlance) are distinguished in *MentalObjects* and *SocialObjects* depending on whether or not they are generically dependent on a community of agents. *SocialObjects* are further divided into *Agentive* and *NonAgentive*. *AgentiveSocialObjects* are for example companies, like Sony. *NonAgentiveSocialObjects* are laws, norms, peace treaties, etc.; they are generically dependent on *Societies*.

4. Mapping T-PAS into DOLCE

After presenting the content and structure of T-PAS and DOLCE, in this section we report the results of the tentative alignment of the system of semantic types in T-PAS (in red in Fig. 3) to the categories and structure of DOLCE (in blue, accompanied by the definitions). As previously clarified, the attention is focused on the *Endurant* branch, in particular *PhysicalEndurant*. The alignment is depicted in Fig. 3, which is followed by comments on the individual operations.

Figure 3. Mapping T-PAS into DOLCE. DOLCE categories are marked in black and italic, together with their definitions, whereas TPAS semantic types are in grey and small caps.

Endurant vs. Entity

The first high-level observation is that DOLCE's *Endurant* category is a node that aligns very well with the T-PAS organization. DOLCE *Endurant* corresponds to ENTITY in T-PAS. 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 of the system, as the word ENTITY is often used in the linguistic literature in a way which excludes events. Note that ANYTHING is T-PAS has a specific interpretation as it stands for all semantic types that play the role of participant in the event described by the verb selecting them ("PARTICIPATION" relation between endurants and perdurants in DOLCE).

Endurants and the *Object* / *Stuff* distinction

DOLCE *PhysicalEndurant* corresponds to PHYSICAL ENTITY in T-PAS; the internal organization of the two nodes, however, differs. *Amount of Matter* is a sister node of *PhysicalObject* in DOLCE (together with FEATURE), while in T-PAS, its closest equivalent (STUFF) is a type of PHYSICAL ENTITY (INANIMATE PHYSICAL ENTITY, see Fig. 3). The solution in DOLCE appears more adequate, as the animate/inanimate distinction apparently applies only to objects with Unity.

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) (see [8]), and the "PARTHOOD" relation, which defines the relation between a body and its parts, are not represented in T-PAS, in which 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 *NonPhysicalEndurant* (entity with temporal properties such as *Mental* and *Social Object*, see Fig. 2); these two categories appear in different nesting levels. There is no possible one-to-one alignment in this case. For application purposes, the two DOLCE's category can be conflated into T-PAS ABSTRACT ENTITY, inasmuch 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.

Agents and the animate / inanimate distinction

The *Agent* label is used in DOLCE to express a potential agent, that is, a living being endowed with intentions, beliefs, and desires. *Physical objects* that have intentionality (i.e. the ability to direct / handle objects or states of the world) are called *Agentive*, those that have no intentionality are called *NonAgentive*.

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 [9], 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; it does not include PLANT but it includes the taxonomy of the animal kingdom (ANIMAL and ANIMAL GROUP). The animal kindgdom 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 (see Fig. 1); typically these are verbs of sound emission such as *barking* (DOG), or verbs of motion such as *galloping* (HORSE). Linnaeus categories such as MAMMAL are not present, as no verb has been identified yet that selects for it.

Feature and parasitic entities

DOLCE assumes the category *Feature* for "parasitic entities" that are constantly dependend 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.

Locations

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*). 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.

Natural kinds / Artifactual Types distinction

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. Note that from a linguistic perspective, the distinction between individuated and mass appears to be the most represented formally in the world's languages, that is, the grammatical behavior of nouns appears to be primarily determined by their encoding individuals or masses (cf. [10, 135]).

Types vs. roles

T-PAS has FOOD and BEVERAGE as types of ARTIFACT. In [7, 218] it is observed that "nothing is necessarily food, and just about anything is possibly food". In other words, *food* is considered a role that an entity can play in a food event, not an ontological

category (roles being anti-rigid properties that characterize the way something participates to a contingent event). While sharing the same theoretical position, 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, also in the light of the presence of artificial food, that is, man-made entities to be consumed as food.[3]

Systematic polysemy

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. [10] for an overview), for example "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"). These cases 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 (instanciated 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 currently not represented in DOLCE: see, however, the proposal in [11] to formally represent *complex categories* that 'overlap' with disjoint domains of entities using the ontological relationship of "CONTITUTION" and the notion of coincidence.[4]

Although it is used in other lexical ontologies such as WordNet, multiple inheritance is normally avoided in formal ontologies because it introduces incoherence and inconsistency from an ontological perspective and it creates problems for calculating inferences. It makes the hierarchy more like a graph than a tree and the inference is more complex. 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.

Leaf Categories

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 requested by verb selectional behaviour: the set of categories is not finite and may increase as long as 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 types. Child categories in T-PAS disclose the anthropocentric character of the type

[3] Note that T-PAS also contains a semantic type called ROLE in the PROPERTY branch, used for words and expressions that denote functions, for example Prime Minister in the expression: "Mattarella has appointed Giovanni Conte Prime Minister" (see Fig. 1).

[4] Specifically, [11] proposes to characterize such cases as ontological constructs based on coincidence relations between entities of disjoint categories.

system, that does not reflect the state of the art in scientific knowledge but how everyday speakers talk about the world: for example, the Artifact node includes WEAPON (and its subtypes FIREARM and BOMB), three subtypes of VEHICLES (ROAD, FLYING and WATER), FOOD, BEVERAGE, BUILDING, GARMENT, FURNITURE, and peculiar types such as FLAG and STRING.

5. Concluding Observations

We illustrated the main features of the system of semantic types developed in T-PAS starting from the analysis of linguistic evidence from corpora, and we compared it with DOLCE foundational ontology, whose categories are not the result of a data-driven approach.

In general, the study 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. Two substantial issue emerge: first, 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; secondly, the phenomenon of systematic polysemy is currently not represented in DOLCE, while it is systematically encoded as multiple inheritance in T-PAS.

Furthermore, the study shows that 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.

In the alignment described in the previous section, DOLCE, whose structure satisfies the OntoClean methodology, provides several hints for improvement of the general architecture of the type system in T-PAS, where the only relation is subsumption: for example, the relation of "PARTHOOD" and the relation of "CONSTITUTION" will be helpful to improve the current treatment of BODY, PART OF BODY, HUMAN GROUP, and ANIMAL GROUP. Conversely, from an applied perspective, T-PAS provides the "leaf categories" that DOLCE does not possess - being DOLCE an upper level ontology. The benefit of the experiment can be confirmed as reciprocal.

As regards future work, a first step concerns the completion of the alignment with DOLCE's *NonPhysicalEndurant*, *Perdurant* and *Quality* branches, not included here. A second step concerns systematic polysemy, and the implementation of a distinct treatment of complex categories [11] and metonymic shifts [12]. Finally, a third step regards the evaluation of corpus-derived semantic types agains automatically obtained clusters of argument fillers, in line with works such as [13] and more recently - in a distributional semantic framework - [14].

Acknowlegments

We would like to thank three reviewers for their helpful comments. All remaining errors and infelicities are our own.

References

[1] Jezek, Elisabetta & Magnini, Bernardo & Feltracco, Anna and Bianchini, Alessia & Popescu, Octavian. T-PAS: A resource of Typed Predicate-Argument Structures for Linguistic Analysis and Semantic Processing. In Calzolari, Nicoletta et al. (ed) *Proceedings of the Ninth International Conference on Language Resources and Evaluation* (LREC'14), 890-895. Paris, ELRA, 2014.

[2] Masolo, Claudio & Borgo, Stefano & Gangemi, Aldo & Guarino, Nicola & Oltramari, Alessandro. WonderWeb Deliverable D18: Ontology Library (final). *IST Project 2001-33052 WonderWeb: Ontology Infrastructure for the Semantic Web*, 2003.

[3] Hanks, Patrick. *Lexical Analysis: Norms and Exploitations*. Cambridge MA: The MIT Press, 2013.

[4] Gangemi, Aldo & Guarino, Nicola & Masolo, Claudio & Oltramari, Alessandro & Schneider, Luc. Sweetening Ontologies with DOLCE. In *Proceedings of the International Conference on Knowledge Engineering and Knowledge Management*: 166-181. Berlin – Heidelberg: Springer, 2002.

[5] Bach, Emmon. The Algebra of Events. *Linguistics and Philosophy*, 9 (1986), 5-16.

[6] Guarino, Nicola & Welty, Christopher A. Evaluating Ontological Decisions with OntoClean. *Communications of the ACM* **45**(2) (2002), 61-65, New York, ACM Press.

[7] Guarino, Nicola & Welty, Christopher A. An overview of OntoClean In *Handbook on Ontologies*: 151-171. Berlin – Heidelberg, Springer, 2009.

[8] Gaio, Silvia & Borgo Stefano & Masolo Claudio & Oltramari Alessandro & Guarino Nicola. Un'introduzione all'ontologia DOLCE. AIDA informazioni **28**(1-2) (2010), 107-125.

[9] Guarino Nicola. BFO and DOLCE: So Far, So Close. Cosmos + Taxis **4** (4) (2017), 10-18.

[10] Jezek, Elisabetta. *The Lexicon: An Introduction*. Oxford, Oxford University Press, 2016.

[11] Arapinis, Alexandra & Vieu, Laure. A plea for complex categories in ontologies. *Applied Ontology* **10**(3-4) (2015), 285-296.

[12] Jezek, Elisabetta & Vieu, Laure. 2014. Distributional analysis of copredication: Towards distinguishing systematic polysemy from coercion. In Basili Roberto et al (eds.) *First Italian Conference on Computational Linguistics* CLiC-it 2014, 219-223. Torino: Accademia University Press, 2014.

[13] Snow, Rion & Jurafsky, Daniel & Ng, Andrew Y. Semantic Taxonomy Induction from Heterogenous Evidence. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, 801–808. Sydney, Australia: Association for Computational Linguistics, 2006.

[14] Jezek, Elisabetta & Ponti, Edoardo Maria & Magnini Bernardo. Evaluating distributional representations of verb semantic selection. In Harry Bunt (ed) *Proceedings of the 15th Joint ACL - ISO Workshop on Interoperable Semantic Annotation* (ISA-15), IWCS 2019, Gothenburg University, 2019.