Restructuring a Taxonomy of Literary Themes and Motifs for More Efficient Querying

FAHAD KHAN

Istituto di Linguistica Computazionale "A.Zampolli" – ILC
Consiglio Nazionale delle Ricerche – CNR

SILVIA ARRIGONI

Università Ca' Foscari Venezia, Dipartimento di Studi Umanistici

FEDERICO BOSCHETTI

Istituto di Linguistica Computazionale "A.Zampolli" – ILC
Consiglio Nazionale delle Ricerche – CNR

FRANCESCA FRONTINI

Istituto di Linguistica Computazionale "A.Zampolli" – ILC
Consiglio Nazionale delle Ricerche – CNR

Abstract

In this paper we describe ongoing work in the restructuring of a tagset originally organised as a taxonomy and used to annotate literary themes and motifs in a corpus of classical works of poetry from a number of different traditions. We show how such a tagset can be rendered more efficient and useful through the appropriation of ideas and techniques from lexical semantics and ontology design. The newly redesigned tagset is described with examples showing how the new design is much more expressive than the old taxonomy; furthermore, an example query is described in order to demonstrate how more refined semantic searches can be carried using the new version of the taxonomy. The final result is, we hope, a resource that will be useful not only for the specific project for which it was developed but one that is well-designed and well-documented enough to be of use for other similar semantic annotation tasks.
Keywords: themes and motifs; taxonomy; hyponymy; lexical semantics; ontology.

Resumo

Neste artigo descrevemos o trabalho em curso de reestruturação de um conjunto de etiquetas originalmente organizado sob a forma de taxonomia e usado para anotar temas e motivos literários num *corpus* de obras clássicas de poesia de diferentes tradições. Mostramos como este conjunto de etiquetas se pode tornar mais útil e eficiente através da apropriação de ideias e técnicas da semântica lexical e da criação de ontologias. O conjunto reconstruído de etiquetas é descrito através de exemplos que mostram como a nova estrutura é muito mais expressiva do que a antiga taxonomia; além disso, descreve-se um exemplo de *query* para demonstrar como se podem realizar pesquisas semânticas mais finas usando a nova versão da taxonomia. O resultado final é um recurso útil não apenas para o projeto específico para o qual foi desenvolvido, mas cremos que está suficientemente bem desenhado e documentado para ser útil para tarefas de anotação semântica semelhantes. **Palavras-chave:** temas e motivos; taxonomia; hiponímia; semântica lexical; ontologia.

MATLIT 4.2 (2016): 11-27. ISSN 2182-8830
http://dx.doi.org/10.14195/2182-8830_4-2_1

1. Introduction

The semantic tagging of texts is often performed at the level of individual words or lexemes and is frequently used to assist in the process of word sense disambiguation by specifying the particular semantic field(s) that each token in a text belongs to. But semantic tagging can also be carried out at other levels of textual organisation. In this article we will be concerned with the semantic annotation of a corpus of poems in which the textual unit to be annotated is the poetic line, or even in certain cases (given the brevity of the poems in the corpus) the entire poem itself. The corpus in question consists of poems taken from Classical Greek, Latin, Italian and Arabic anthologies and the process of annotation is part of an ongoing Italian national project *Memorata Poetis*, which we will discuss in greater detail in the next section. Our focus in this article will be on the tagset used in the annotation—a tagset which was originally organised as a taxonomy—and on the process of restructuring which we felt was necessary to make the tagset more amenable to automatic processing and querying. The idea, in brief, was to redesign the taxonomy to permit more efficient access to data about the interpretations of texts in the corpus, in order to ulimately facilitate more involved kinds of semantic analyses of the texts themselves. The work that we describe in this article is the initial stage in a wide ranging case study into the application of ideas and practices from lexical semantics and ontology modelling, as well as the computational resources produced by these fields, to the study and analysis of literary texts. We believe that this places our work firmly within the tradition of the digital humanities, understood as a 'transdiscipline' that mobilises 'the tools and unique perspectives enabled by digital technology' (Dacos 2011).

The structure of the article is as follows. In the next section, Section 2, we will give a description of the tagset as well as a brief overview of its use in the *Memorata Poetis* project. In the following section, Section 3, we describe the process of restructuring the tagset, with particular emphasis on the tagset's main structuring relation, hyponymy. In Section 4 we describe the new 'ontologically' structured version of the tagset and outline the kinds of queries that could be made using it. In the final section we discuss the current status of our work and future plans.

2. The Taxonomy of Themes and Motifs

The work that we detail in this paper takes place within the ambit of *Memorata Poetis*[1], a project whose overall goal is to create a multilingual semantic search engine for the comparative analysis of literary themes, motifs and

[1] http://www.memoratapoetis.it/.

figures in poetic works ranging across a number of different literary traditions. The project had a strong initial focus on poetic inscriptions (*Carmina Epigraphica*) in Greek, Latin, Italian and Arabic and the idea was to tag a corpus of epigraphic works that varied not only according to language and to cultural provenance but also according to age, genre and type, and with respect to their relationships to 'high' literary genres such as the lyric, elegiac etc. Other languages such as English and Old Lithuanian were added to the corpus of texts afterwards.

2.1 Description of the Tagset

Due to the fact that the *Memorata Poetis* project set out to study the treatment of poetic themes and motifs across diverse traditions, and also in view of the epigrammatic nature of the source texts themselves, the decision was made to carry out the tagging at a more abstract level of textual organisation than at the word level, that is, at the level of the poetic line and/or the epigrammatic poem in its entirety. The tagset used in *Memorata Poetis*, known as the Taxonomy of Themes and Motifs [TTM], was initially developed by philologists and specialists in the literary analysis of Ancient Greek and Latin texts. Its design was directly influenced by the indices used in traditional anthologies of classical poetry and was therefore founded upon long established, pre-digital, practices in the semantic annotation and categorisation of literary texts.

There are approximately 1,250 tags in the *Memorata Poetis* tagset all of which are labelled in Latin, which serves as an interlingua for the project. These tags are arranged in a taxonomic structure in which more specific concepts at a greater depth in the taxonomy are associated with, and thus subsumed by, more general concepts, using a generic notion of relevance to link concepts together. This arrangement is very similar to the organisation of semantic fields, in which certain 'representative' concepts are used as topic names and serve to cluster together other concepts which are usually more specific. This kind of taxonomic organisation has, in the past, been referred to as a *semantic field taxonomy* as, for example, in work on the large scale semantic lexicon UCREL (Archer et. al. 2004). It is also closely related to the classification of lexical items in a thesaurus. As an example the following tags are associated with the 'semantic field' of *Amor* [Love] in the TTM and are subsumed under it: *Voluptas* [Pleasure], *Mors voluntaria* [Suicide], *Crimen* [Guilt], and *Puer* [Boy].

The TTM is divided into six different thematic areas: *Animalia* [Animals], *Arbores et virentia* [Trees and Plants], *Homines* [Men], *Dei et heroes* [Gods and Heroes], *Loca* [Places], and finally *Res* [Things]. Each of these topics represents a top node in the taxonomy. As there is no single top node, and there are no 'transversal' relations between the nodes in the six different subtaxonomies dominated by these six top nodes either, the TTM is, in graph theoretical terms, a forest of labelled trees. This lack of interconnectedness raises a number of problems for the usability of the taxonomy as we shall discuss below. Each of the TTM's six separate sub-taxonomies is further arranged into three layers, in decreasing order of generality. So that, for example, one of the paths through the taxonomy can be described in terms of the labels of the nodes visited as follows: (*Homines, Laudatio, Amicitiae*). Figure 1 below represents a schematic of the TTM.

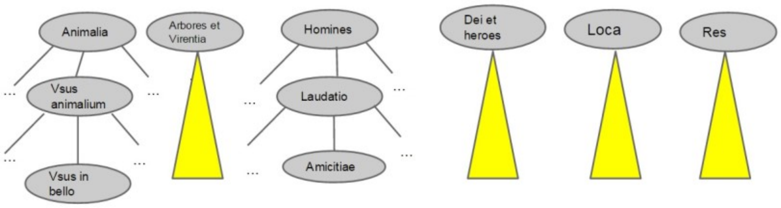

Figure 1. The Taxonomy of Themes and Motifs.

There are a number of clear benefits to having a hierarchical structure, and even one as simple as the TTM's, in a large tagset over having no structure at all. For a start it makes the members of the tagset much more accessible to those carrying out the actual work of annotation. Indeed the TTM tags are arranged as a taxonomy in the annotation interface itself which makes them much easier to locate visually. In addition, by imposing a hierarchical structure on the tagset, here based on a general notion of relevance, we enable certain, limited, kinds of inference to be made on the texts annotated with that tagset. For example, if a line, *l1*, in a poem is tagged with the name of a particular kind of flower, say *Rosa* [Rose] and another line, *l2*, in another poem or perhaps the same poem, has been tagged with the name of another kind of flower, say *Lilium* [Lily], then clearly there is an important semantic/thematic link between the two lines: they're both talking about flowers. This is represented in the TTM by the fact that both of these tags are subsumed by the more general tag *Flores* [Flowers]. We can therefore capture the common floral subject matter of both lines without having to tag either line separately with the tag *Flores*. Unfortunately in the present case one has to be careful when making such kinds of inference since the TTM 'relevance' relation merges a number of different semantic or ontological relations such as hyponymy and meronymy into one single relation. And to return to the TTM example given above, it may very well be the case that boys and guilt are generally associated with love, but that doesn't necessarily mean that every poem or poetic line tagged with either *Puer* or *Crimen* will inevitably be about love.

The TTM does not include any instances of multiple inheritance[2] and so, in cases where more than one topic is relevant for a particular tag, the tag in question ends up being subsumed under only one of these. This means for instance that there is no relation between the concepts *Mors* [Death] and *Mors animalium* [Death of Animals], the former of which is subsumed under *Animalia*. So if we were to search a text tagged using the TTM for mentions of 'death' without wanting to distinguish between the death of humans and the death of animals then we would have to use a query that included both tags. Another TTM design decision that was problematic from our point of view was the fact that, e.g., *Laudes Animalium* [Praises of Animals] and *Laus Florum* [Praise of Flowers], as well as the several other types of praise listed in the TTM, were not related together since they fell under different sub-taxonomies, and nor was there any overarching node capturing the general concept of praise in the TTM. So that while it is true that the arrangement of the TTM is well motivated from a philological point of view and that it enables the organisation of the tagset under salient topic headings, there are several aspects of this arrangement that render the TTM less *usable* as a digital resource than it might otherwise be. We discuss this in more detail next.

2.2 Viewing the TTM as a Digital Resource

The development of a domain ontology, if it is to be done properly, can be a time consuming and labour intensive process and so unless there are some clear advantages to doing so, that is unless it genuinely serves to assist (in this case) philologists and literary researchers in carrying out their work, then there's little point in going to all the extra effort. One of the things which we hope to show as an outcome of this work, and which we will attempt to argue for in the paper, is that there are indeed some important benefits to putting in this kind of work. But before we proceed any further it's important to clarify what is it that we intend when we talk about developing an ontology "properly", that is, what it means for an ontology to be well constructed: for instance why shouldn't the 'relevance' based semantic field taxonomy structure discussed above suffice? In order to answer this it is necessary to realise that there are two, related, issues at stake here. The first is that if a taxonomic resource like the TTM is to be useful in a wider context than just a single project and, in particular, if it is to be used to tag texts belonging to a different set of languages or cultural traditions—and this for example was the case with *Memorata Poetis* since the intention was always to expand the use of the tagset to texts in other languages—then it should be easily reusable by and accessible to a wider community of users. One of the easiest ways of ensuring this is by developing resources with reference to, if not a pre-

[2] By multiple inheritance here we refer to cases in which a child node has two or more parent nodes; in other words, a situation in which a node belongs to two categories at the same time.

existing standard (whether that's *de facto* or official), then at least to widely adopted or widely recognised ideas and practices pertaining to resources of the same type or to those that are similar in salient respects. In the present case we felt that it was important to refer back to recent work in the fields of lexical semantics and ontology engineering.

The second issue is one that was alluded to above and concerns the fact that once we start viewing a taxonomic resource like the TTM as a computational artifact—that is rather than as something that just happens to be stored as a digital file but which otherwise is essentially no different from a traditional thesaurus or taxonomy—then this inevitably leads to considerations of efficiency and ease of access, e.g., how easy is it to write code to allow the automatic extraction of information from the dataset and how quickly will this code run? How can I best link together the different parts of my datasets to other external datasets to allow me to search them together? That is, we are forced to look at it as a digital resource to be processed automatically and with varying degrees of human intervention. In many cases it will transpire that although the organisation of a taxonomy or an ontology is regarded as informative and intuitive by its human users, in reality the information contained within it is difficult to extract and requires some sort of reasonably involved preprocessing.

So that, based on these and a number of other factors, we decided to carry out the restructuring of the TTM by explicitly taking into consideration its status as a computational resource and against the background of previous work on what we felt were related resources, such as wordnets and domain ontologies. But before we go on to describe this step it is important to emphasise that as the work outlined in this paper is tentative and is only in its initial stages—and in view of the fact that it wasn't always clear that techniques that are used primarily to design or evaluate wordnets or, for example, biomedical ontologies would apply to literary studies—we felt that it was crucial that the decisions made throughout the design process were guided by the needs and requirements of the end users, in this case the philologists and literary experts who, we expect (and hope), will use the tagged texts and the tagset as a resource in their work. This means that we have worked in close collaboration with the potential end users of the tagset throughout and that we made the eliciting and specification of requirements and the formulation of use cases a priority in our work.

3. Restructuring the TTM

One of the first issues that we wanted to fully clarify after deciding to proceed with the restructuring of the TTM was the question of the status of the TTM from a formal point of view, and related to this was the question of what formal status the new restructured version of the TTM, which we will refer to as the TTM2 in what follows, should have.

First, though, we will give some definitions. The word *taxonomy* derives from the Ancient Greek τάξις (taxis) meaning ‘order’ or ‘arrangement’ and is generally used to describe the classification of a domain usually (but not always) in terms of a hierarchy with a single subsumption relation. The word ontology[3] on the other hand is usually defined in the information sciences as a shared conceptualisation of a domain or domains. Constructing an ontology in this sense entails formulating the description of a domain, often in some kind of formal language, making sure to include those entities and relations which are most relevant to the information needs of potential users. However, definitions of terms like taxonomy and ontology are general enough that they are helpful only up to a certain point and in many contexts they are used interchangeably. We will take the common noun ‘taxonomy’ to refer to a hierarchical arrangement of a set of terms using a single subsumption relation; and ‘ontology’ to mean a more general formal description of a domain that may encompass a number of different relations. The TTM then can be viewed as a taxonomy that classifies a reasonably large collection of Latin words and multiword expressions from a series of different domains (these are, to recapitulate, [Animals], [Trees and Plants], [Men], [Gods and Heroes], [Places], [Things]). The criterion according to which terms were included in the tagset was based on the fact that they served to describe the content or meaning of poems or of specific poetic lines in a wide-ranging corpus of classical poetry.

From the start we decided to try and limit the number of new items we added to the tagset and to instead focus our attention on the semantic relations between the tags since we felt that these were the source of most of the issues that we found during our preliminary analysis of the TTM. One of the first and most important design decisions that we had to make therefore was the choice of the core relation between the tags in the taxonomy: this would serve to provide a backbone for the rest of the taxonomy. We will discuss this further in the next section. Another important decision was which format to use for storing and working with this new version of the taxonomy. In the end we decided on the Web Ontology Language (OWL 2). It was an easy choice to make for a number of reasons, not least of which was the availability of numerous off-the-shelf tools for working with OWL 2 ontologies including reasoning engines such as Pellet. The OWL 2 language is based on the RDF data model and so gives us the further possibility of adding TTM2 to the Linked Open Data cloud[4] in the future, and thereby enriching the tagset with other resources from the cloud, as well as—and this was especial-

[3] The word is usually written with a small ‘o’ to distinguish it from the word ‘Ontology’ as a technical philosophical term.
[4] http://lod-cloud.net/.

ly important for us—permitting us the use of the powerful SPARQL query language[5].

3.1 Hyponymy and the Relation of Taxonomy

As we stated above, the 'relevance' relation that is used to structure the TTM collapses together various different semantic relations in one. Probably the best known of these individual semantic relations is that of *hyponymy*. We take the following, fairly standard, definition of hyponymy from Cruse's *Lexical Semantics* (Cruse 1986):

> A lexical item l is a hyponym of another lexical item m if the sentence "A is f(l)" entails (but is not in its own turn entailed by) "A is f(m)".

The formula *f(x)* is here understood as an indefinite expression representing the minimal syntactic elaboration necessary for the lexical item *x* to function as the complement of 'to be', e.g., for the noun 'dog' this indefinite expression would be 'a dog' as in 'Toby is a dog', for the adjective 'crimson' on the other hand a suitable indefinite expression might be 'a crimson ball'. In the former case then we can use the fact that 'Toby is a dog' entails 'Toby is a mammal' to posit that 'dog' is a hyponym of 'mammal'; similarly on the basis of the fact that 'This is a crimson ball' entails 'This is a red ball' and other similar sentences we can derive the fact that 'crimson' is a hyponym of 'red'. Hyponymy then has the big advantage that it gives a simple criterion, in terms of sentence frames, for deciding whether two lexical items are semantically related to each other; and this would seem to make it a good candidate for structuring taxonomies. In Figure 2 we give an instance of a basic taxonomy where the nodes are linked together using the relation of hyponymy.

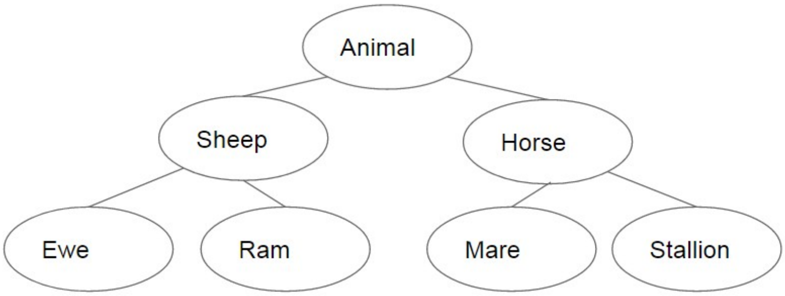

Figure 2. Example Taxonomy.

Hyponymy is often identified with the ontological ISA relation (the ISA or IS_A relation is a common name for the subsumption relation used as the

[5] http://www.w3.org/TR/sparql11-query/.

main structuring relation in ontologies); it is also used as the main 'backbone' relation in the well-known lexical resource Wordnet (Miller, 1995). There are, however, several problems with its use as the structuring relation in a "well-formed taxonomy". The most serious of these concerns the fact that, like the TTM relevance relation, hyponymy is too *gross* a relation and gathers together too many different relations in one. For instance, Cruse (1986) observes that taxonomies like that in Fig. 2 actually manifest two *different* kinds of classification: the first kind marks the division of the class *animal* into *sheep* and *horse*, the other the division of *sheep* into *ewe* and *ram*, and *horse* into *mare* and *stallion.* Cruse's claim is that this contradicts the *principle of differentiation*, which is, as he argues, an important property of taxonomies. The principle of differentiation allows us, in principle at least, to extend each path indefinitely, at least until we reach classes with one member or none, so that 'Arabian Horse' is subsumed by 'Horse' and 'Seglawi' is subsumed by 'Arabian Horse', and so on. This isn't possible with the separation of *Horse* into *Mare* and *Stallion.*

Instead Cruse goes on to suggest a different principal semantic relation for taxonomies, which he calls the *taxonomy* relation[6] and which we have adopted for the work outlined in this paper. In order to motivate the definition of this particular semantic relation for taxonomies—and in so doing to motivate the choice of taxonomy as our principal semantic relation—we will briefly look at an ontology design methodology called OntoClean (Guarino and Welty, 2004) which also recommends against the kinds of conceptual over-generalisations which are manifested in the definition of hyponymy given above and which are frequent in a resource such as WordNet.

What are the possible motivations for instituting this kind of rigour when it comes to designing ontologies? Well, nowadays ontologies tend to be encoded in formal languages like OWL 2 or in F-logic, for which there exist a number of specialised inference engines. These inference engines make it possible to automatically check ontologies for consistency and to render explicit information that is otherwise implicit in the resource. If however we are to use formal ontologies to automatically derive implicit knowledge then it is essential to pay attention to the kinds of, often subtle, distinctions between different conceptual categories and relations that can have important logical consequences, but which usually go unnoticed or unremarked in day to day situations. Take the well known 'Student/Person' example for instance. 'Student' is obviously a hyponym of 'Person', and so it would seem reasonable to subsume 'Student' under 'Person' in an ontology containing both concepts. Similarly a 'Social Entity' is a hyponym of 'Group (of People)' and 'Physical Object' is a hyponym of 'Amount of Matter'. But in each of these cases we are arguably sense pairing together two different kinds of entities. For instance, if I qualify as a student then clearly I fulfill all the con-

[6] Note that the relation of taxonomy should be distinguished from the use of the noun taxonomy that we gave above to describe a hierarchical classification of words.

ditions to be a student, but it is entirely possible that given a slightly different set of circumstances I might not have been a student—and therefore I will not qualify as a student in all possible worlds: on the other hand it is not the case that a person in this actual world might be a non-person in some possible world and still qualify as the same kind of entity; and so on this account 'Student' and 'Person' seem to be two different kinds of properties. Again, it is clear that any amount of matter that loses or exchanges some part of itself thereby ceases to be the very same amount of matter, on the other hand, an artifact like a broom or a washing machine can be respectively the same broom or washing machine even if parts are lost or exchanged; a very similar argument can be made about 'Social Entity' and 'Group of People'. Guarino and Welty argue that it is necessary to take these kinds of difference into consideration in the design of a formal ontology, since otherwise we will inevitably end up deriving mistaken knowledge from poorly designed ontological resources; for this reason they developed the OntoClean ontology design methodology (Guarino and Welty 2004). OntoClean has so far proven to be an extremely influential set of normative guidelines within the ontology engineering community.

It is important to bear in mind here that although we probably won't want to subsume 'Physical Object' under 'Amount of Matter' there is still clearly an important relationship between them: one that we might want to describe by saying that instances of the one class are composed of by members the other. Similarly: although according to OntoClean we should avoid subsuming 'Student' under 'Person', we will likely still want to store the information that a person can, under specific circumstances, have the role of a student with a new relation that can be specified using axioms describing the domain and range of the relation, e.g., whether it's functional, whether it is the sub-relation of another relation, etc. OWL 2 enables us to do this by using so called Object Properties, that is, relations between OWL 2 individuals. With these formal tools we can in effect tease out the different strands of meaning that are clustered together in the hyponymy relation, and as a consequence make the semantic structure of the taxonomy much more transparent.

Overall then we can think of a well-designed ontology as having one core, backbone relation that supplies it with a basic skeletal structure and which we can subsequently adorn with other salient kinds of semantic relation, thereby specifying, for example, information about different varieties of meronymy or describing other facets of commonsense or world knowledge in a more compartmentalized fashion. And this is the tack that we decided to take with regards to the design of TTM2.

We decided to adopt Cruse's relation of taxonomy which is defined similarly to the definition of hyponymy given above, with the difference that we use the following diagnostic frame: given two lexical items X, Y, it is the case that X is a *taxonym* of Y if the following sentence frame is considered as semantically unproblematic by native speakers of English:

> 'An X is a kind/type of Y'.

Of course we were also able to exploit other kinds of linguistic and/or ontological evidence based on, for example, the OntoClean guidelines in deriving this backbone taxonomy. We found that Cruse's criterion of taxonomy was a good enough approximation to the OntoClean guidelines for our purposes, although, OntoClean for example, recommends against including the class of properties which they call *attributions*, such as the class of red things (Guarino and Welty, 2004), whereas in the case of poetic themes and motifs and especially in the TTM, such properties are quite commonly encountered, e.g., *Lumina Amantis* [Lover's Eyes] and *Arbores deis deabusque consecratae* [Trees consecrated to the gods and goddesses].

Once more we should make a proviso here. We were careful throughout the process of redesigning the tagset to ensure that the meaning and literary context of the tags was fully taken into account: and so we tried to ensure that we received feedback on each successive version of our redesign; this feedback has been so far very positive and has led to a fruitful dialogue between researchers from different backgrounds and with different skillsets and expectations about what such work might ultimately achieve.

4. *A Description of TTM2*

We made the decision to use DOLCE-lite,[7] a lightweight version of the popular DOLCE upper level ontology (Gangemi et. al., 2004), to provide TTM2 with its highest and most abstract level of concepts, since the original tagset lacked a lot of these more general concepts.[8] This means that the backbone taxonomy of TTM2 (that is the ontology with only the taxonomy relations between nodes) is a single tree with a unique beginner node instead of a forest as was the case before. Using an already existing upper level ontology not only saved us the time and effort of creating a new set of upper level concepts from scratch, but it also serves to facilitate the future integration of TTM2 with other similar semantic resources that also use DOLCE as an upper level. DOLCE-Lite was chosen both because of its popularity as an upper level ontology and the fact that it has a very strong theoretical basis while at the same time being simpler and easier to navigate than DOLCE itself.

[7] www.loa.istc.cnr.it/ontologies/DOLCE-Lite.owl.

[8] We didn't use the class *physical objects* from DOLCE but instead use a related class *physical and supernatural objects*.

As mentioned previously, we used Cruse's relation of taxonomy to structure the backbone of the taxonomy and then added other relations (OWL 2 object properties) encoding various different kinds of conceptual relevance relation between the tags. Several of the most important design decisions were guided directly by the necessities of poetic analysis. For instance, we used ideas from speech act theory to help classify various aspects of poetic rhetoric thus providing us with a useful means of grouping together instances of praise, blame, etc. in the TTM2. In addition we kept tags such as *Fama et gloria* [Fame and Glory] and *Maritus et Uxor* [Husband and Wife] that represented conjunctions in the TTM2 for pragmatic reasons, since such conjunctions often have a significance over and above that of their individual conjuncts taken together. In cases where the separate conjuncts did not have their own tags, we augmented TTM2 with a tag for each separate conjunct.

We tried to avoid the addition of new tags to the ontology as far as possible, but in a number of instances, such as in the example of the conjuncts, the case for augmenting the tagset with new tags was in fact strong enough: this was particularly true in those cases where the creation of a more general conceptual node was desirable to faciliate certain very useful types of corpus query. To return to the example of the various forms of praise listed under unrelated categories in the TTM, we created a new node *Laus* subsuming each individual kind of praise so that *Laus animalium*, *Laus poetarum*, *Laus amicitiae*, *Laus artium* etc all now fall under *Laus*.

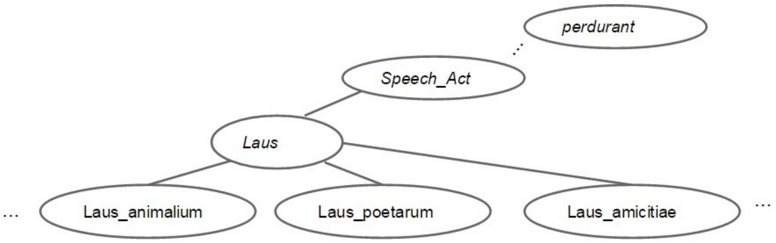

Figure 3. Laus in the TTM (Nodes not in the original have labels in italics).

Then, when relevant, we related each kind of Laus to its object using a new OWL 2 object property, *hasPraiseObject,* which we introduced especially for this purpose and whose domain we fixed as *Laus* using the following OWL 2 axiom:[9]

ObjectPropertyDomain(:hasPraiseObject :Laus)

We represented the restriction that instances of the class *Laus_animalium* can only praise instances from the class of animals, and that instances of the

[9] The axiom has been serialized using the OWL Functional Syntax.

class *Laus_poetarum* can only praise instances of the class of poets respectively using the axioms below:

> EquivalentClasses(:Laus_AnimaliaObjectAllValuesFrom(:hasPraiseObject :Animalia))
> EquivalentClasses(:Laus_poetarumObjectAllValuesFrom(:hasPraiseObject :Poetae))

There were a number of other similar cases to Laus in the TTM2 such as for example, *Adlocutiones* [Addresses], *Dirae* [Damnations] which we were able to treat in like manner. So that for instance we placed all the different types of address such as *Adlocutio ad lectorem* [Address to the Reader], *Adlocutio ad patriam* [Address to the homeland] under *Adolocutiones*, regardless of where they were placed previously, and defined an object property *hasAddressee.*

One of the most important upper level aspects of the classification is the distinction between perdurants (entities that are in a technical sense time bound) and endurants (entities whose essential properties are stable over time); these categories were inherited from the DOLCE upper ontology.[10] Endurants are further classified into *physical_and_supernatural_objects* and *non-physical_object*. This latter category includes *social_object*, under which we place the several different arts and disciplines and uses of things that are found in TTM, e.g., *Vsus animalium in rebus domesticis* [Uses of animals for domestic purposes], *Vsus arborum etc in magia* [Uses of trees, etc, in magic]. Many of the concepts that had previously been included under *Homines* are now to be found under *nonphysical_object* and in a lot of cases are under *social object*, such concepts include *Mores* [Customs], *Artes et disciplinae* [Arts and disciplines], *Amicitia* [Friendship], *Leges* [Laws] and other social institutions.

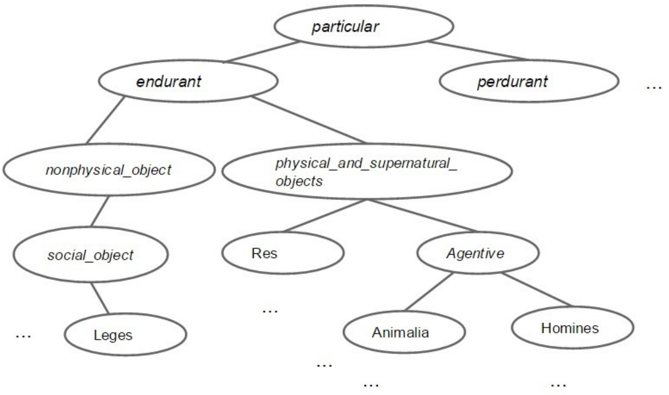

Figure 4. Some of the concepts under the DOLCE category endurant.

[10] Note: the entries that were added to the ontology have an English label in what follows, the original entries were in Latin.

The concept *physical_and_supernatural_objects* which is partitioned into *Agentive* and *Res* (which is usually translated as 'things', but really we mean physical things in this context) includes most of the entries that were previously to be found under the categories *Animalia* [Animals] and *Arbores_et_virentia* [Trees and Plants] such as e.g., *Amphibia*, *Pisces*, *Lilium* and *Papaver*. *Animalia* itself is now under *Agentive*, and *Arbores_et_virentia* under *Res*. *Res* also includes *Pontes* [Bridges] and categories like *Lapides_et_Metalla* [Stones and Metals], *Corpus* [the Body], and *Vtensilia* [Utensils]. The *Agentive* category is used to classify entities that demonstrate agency, e.g., *Homines* [Men/People], *Dei_et_Heroes* [Gods and Heroes].

We categorise the following classes as perdurants: *Convivium*[Banquet], *Servitus* [Slavery], *Annus* [Year]. In order to further categorise these classes we used Zero Vendler's event classification of events as states, achievements, accomplishments and activities (Vendler 1957). So we classify the following concepts as accomplishments: *Bellum* [War], *Iura_et_negotia* [Tribunals and trials], and *Vita* [Life]. The category of *speech act* also comes under *accomplishment*, and we use this classify such concepts as *Fascinum* [Spell], *Laus* [Praise], *Querella* [Complaint], and *Adlocutiones* [Addresses]; these categories are particularly useful when it comes to annotating epigrams. Under *achievement* we have *Partus* [Birth], *Mors* [Death], and *Victoria* [Victory]. *activity* contains *Adulterium* [Adultery], *Masturbatio* [Masturbation], and *Saltationes* [Dances]. Finally *state* includes *Virginitas* [Virginity], *Infantia* [Childhood], and *Libertas* [Freedom]. The other two top level nodes (under the DOLCE category *particular*) besides *endurant* and *perdurant* are *quality* (which subsumes concepts such as *human_qualities* and *Deformitas_et_foeditas* [Deformity and ugliness]) and *space-region* (which subsumes *Loca* [Location]).

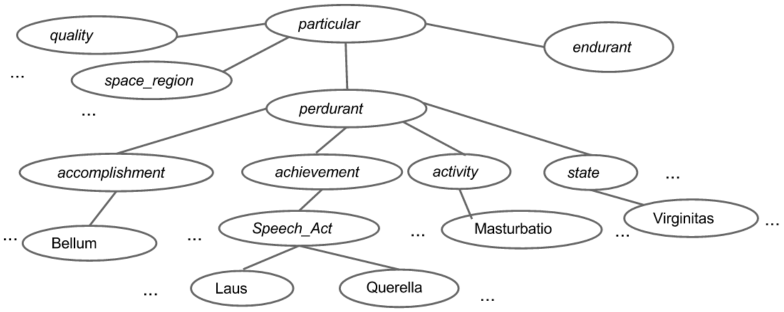

Figure 5. Some of the concepts under the DOLCE category perdurant.

4.1 Querying

The redesign of the TTM was carried out largely with the intention of making querying the corpus much more efficient and straightforward, and having decided on OWL 2 for the format of the TTM2 we were able to make use of the SPARQL semantic web query language. SPARQL allows a great deal of flexibility in writing queries based on the graph structure of RDF datasets. This is useful especially in the case of speech acts such as *Laus* and *Adlocutio*, where it permits the specification of the kind of thing being praised or addressed as part of the query; so that when combined with information on the authors or dates this constitutes a powerful tool for searching a corpus. For instance, assuming that each TTM2 tag is explicitly associated with the ID's of all the lines which have been tagged with it (at the time of writing this has not yet been implemented) and that this information is stored in our dataset as RDF triples, then we can easily write a SPARQL query to find all of the poems which praise inanimate classes of physical objects, e.g., flowers and buildings, as opposed to animate entities such as persons. This would have been difficult with the previous taxonomy even if it had been directly converted into OWL without changing the original hierarchy. In the TTM2 each act of praise is classified under *Laus* and linked to its object via the *hasPraiseObject* relation when relevant. We can then write a SPARQL query of the following form in order to capture our information request:

```
SELECT ?line ?praisetype
WHERE {
?line hasTag ?praisetype
?praisetype owl:equivalentClass _:R .
_:R a owl:Restriction .
_:R  owl:onProperty   :hasPraiseObject.
_:R  owl:allValuesFrom  ?q  .
?q rdfs:subClassOf :res .
}
```

If we also assume that our dataset contains information about the author of the work, its date, language, and genre in the form of RDF triples (and since it's RDF it is simple to link to other relevant datasets that give us additional background information), then it becomes clear that the TTM2 provides the basis for an extremely useful means of accessing the information in a corpus via semantic queries.

5. Current Status and Further Work

The first release of the TTM2 is currently being finalised and an initial classification for all of the original nodes in the tagset has been made (although a small number of others have been added to an updated version of the original *Memorata Poetis* tagset since we started working on it, for example those relating to the Islamic religion). For each tag in the TTM2 we plan to add information about the tag that it was subsumed under in the TTM (where appropriate) as a kind of "topic" feature; this means that we can also leverage the semantic field information contained in the TTM in the TTM2.

The TTM2 hasn't yet been used as part of the actual *Memorata Poetis* annotation process; but there are currently plans to integrate it as part of the annotation process. In addition we intend to carry out at least one further process of feedback elicitation from the future users of the tagset and to conduct an evaluation of the TTM2 as a means of querying texts tagged with the *Memorata Poetis* tagset. We want to study how well such a taxonomical organisation can assist in the formulating of queries which users will want to make in the course of their research.

Further, we are planning on mapping TTM2 to the Princeton Wordnet and to the Ancient Greek and Latin wordnets (Bizzoni et al 2014), (Minozzi 2009), (McGillivray 2010); in the case of the Latin Wordnet this will allow us to enrich the TTM2 with information about Latin antonyms and synonyms. We also hope to enrich the TTM2 with links to other linked open datasets (such as the geo-historical gazetteer Pleiades for toponyms, and DBpedia for other entities). This is preparation for the eventual release of TTM2 as Linked Open Data. Throughout the design process for TTM2 we have striven to create a resource that will be useful not only for *Memorata Poetis* but one that is well-designed and well-documented enough to be of use for other similar semantic annotation tasks and make a valuable contribution to the Linked Open Data cloud.

References

ARCHER, Dawn, Paul Rayson, Scott Piao, and Tony McEnery (2004). "Comparing the UCREL Semantic Annotation Scheme with Lexicographical Taxonomies." *Proceedings of the 11th EURALEX (European Association for Lexicography) International Congress (Euralex 2004).* Ed. Geoffrey Williams and Sandra Vessier. Volume III. Lorient, France: Université de Bretagne Sud : 817–827.

BIZZONI, Yuri, Federico Boschetti, Riccardo Del Gratta, Harry Diakoff, Monica Monachini, and Gregory Crane (2014). "The Making of Ancient Greek WordNet." *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14).* Ed. Nicoletta Calzolari et al.: ELRA,. Reykjavik, Iceland : 1140–1147.

CRUSE, D. Alan (1986). *Lexical Semantics*. Cambridge: Cambridge University Press.

DACOS, Marin (2011). "Manifeste Des Digital Humanities." *Billet*. 13 Mar. 2016. http://tcp.hypotheses.org/318.

GANGEMI, Aldo, Nicola Guarino, Claudio Masolo, Alessandro Oltramari, and Luc Schneider (2002). "Sweetening Ontologies with DOLCE." *Knowledge Engineering and Knowledge Management: Ontologies and the Semantic Web*. Lecture Notes in Computer Science 2473. Ed. Asunción Gómez-Pérez and V. Richard Benjamins. Springer Berlin Heidelberg. 166–181. 13 Mar. 2016.

GUARINO, Nicola, and Christopher A. Welty (2009). "An Overview of OntoClean." *Handbook on Ontologies*. International Handbooks on Information Systems. Ed. Steffen Staab and Rudi Studer. Springer Berlin Heidelberg. 201–220. 13 Mar. 2016.

MCGILLIVRAY, Barbara (2010). "Automatic Selectional Preference Acquisition for Latin Verbs." *Proceedings of the ACL 2010 Student Research Workshop*. *ACM Digital Library*. Stroudsburg, PA, USA: Association for Computational Linguistics. 73–78.

MILLER, George A (1995). "WordNet: A Lexical Database for English." *Commun. ACM Digital Library. ACM* 38.11: 39–41.

MINOZZI, Stefano (2009). "The Latin WordNet Project." *Latin Linguistics Today. Akten Des 15. Internationalem Kolloquiums Zur Lateinischen Linguistik*. Ed. Peter Anreiter and Manfred Kienpointner. Vol. 137. Innsbruck, Austria: 707–716.

VENDLER, Zeno (1957). "Verbs and Times." *Philosophical Review* 66.2: 143–160.