Structuring a Domain Vocabulary in a General Knowledge Environment

Nilda Ruimy
Istituto di Linguistica Computazionale
Consiglio Nazionale delle Ricerche
Via G. Moruzzi 1- 56124
Pisa, Italy
nilda.ruimy@ilc.cnr.it

Abstract

The study which is reported here aims at investigating the extent to which the conceptual and representational tools provided by a lexical model designed for the semantic representation of general language may suit the requirements of knowledge modelling in a domain-specific perspective. A general linguistic ontology and a set of semantic links, which allow classifying, describing and interconnecting word senses, play a central role in structuring and representing such knowledge. The health and medicine vocabulary has been taken as a case study for this investigation.

1. Introduction

The study of biomedical language has raised an increasing interest during the last few years and has led, through the creation of biomedical thesauri, database, ontologies and semantic networks, to the semantic definition and categorization of the specific concepts of this domain, let us mention, for example, the Gene Ontology, the MeSH thesaurus, the UMLS Metathesaurus, Semantic Network and SPECIALIST lexicon etc.. The study which is reported here has obviously no ambition to compete with such valuable initiatives. It merely aims at investigating the extent to which the conceptual and representational tools provided by a lexical model tailored to the semantic representation of general linguistic knowledge, namely the SIMPLE model, may suit the requirements of domain knowledge modelling. A general linguistic ontology and a set of semantic relations, which enable classifying, describing and interconnecting concept instances, play a central role in structuring and representing such knowledge.

The health-related vocabulary has been taken as a case study for this investigation. In this paper, structuring is investigated of the health-related vocabulary encoded in PAROLE-SIMPLE-CLIPS – the largest, generic, electronic lexicon of Italian that includes core subsets of vocabulary from various subject domains. An outline of the semantic characterization of entities and events that make up this domain vocabulary is provided. The organization and interrelation of entity-denoting words, and their connection to events is illustrated. A fragment of semantic network[1] (Appendix 1) built according to the most relevant paradigmatic and syntagmatic relations is visually presented, whereby nodes are concept instances (word senses) and edges represent labelled semantic relations connecting word sense pairs. A glance at EuroWordNet (EWN) semantic links then shows that capturing additional relevant information will be possible when the Italian lexicons built according to EWN and SIMPLE lexical models, viz. ItalWordNet and PAROLE-SIMPLE-CLIPS, are linked (Ruimy & Roventini, 2006).

2. Lexical Knowledge Structuring Devices

In accordance with the SIMPLE model, the semantic layer of the PAROLE-SIMPLE-CLIPS is structured in terms of an ontology and lexical entities are characterized and interconnected by means of a rich set of semantic features and relations.

The SIMPLE ontology is a multidimensional type system which has been designed, by combining top-down and bottom-up approaches, for the multilingual lexical encoding of concrete and abstract entities, properties and events. It consists of 157 language- and domain-independent semantic types and is based on both hierarchical and non-hierarchical conceptual relations. The ontology encompasses two different kinds of semantic types: the *simple* and the *unified* types. While *simple* types (i.e. one-dimensional) can be fully characterized in terms of a taxonomic relation to a parent type, *unified* types (i.e. multi-dimensional), besides the subsumption relation, also incorporate orthogonal meaning dimensions. Such organization along multiple dimensions of meaning contributes to avoid an overloading of hyperonymic relations, which constitutes one of the main drawbacks of traditional type systems (Guarino, 1998).

Multidimensionality is expressed in the SIMPLE ontology by means of the *Extended Qualia Structure*, a revisited version of the Generative Lexicon representational tool which played a crucial role in defining the distinctive properties and differentiating the degree of complexity of SIMPLE semantic types.

According to the SIMPLE model, the semantic content of a word sense is therefore expressed through its membership in an ontological type; semantic similarity between word senses thus implies their sharing a semantic type. The membership in a SIMPLE semantic type inherently triggers the instantiation of a rich bundle of semantic features and

[1] http://www.ilc.cnr.it/clips/events_entities_medicine_domain.pps

relations[2]. Among the semantic features, let us only mention here the 'domain' information, which enables to relate a semantic unit to the area of knowledge it is used in. Among the semantic relations, on the other hand, are the sixty ones that build up the *Extended Qualia Structure* (Appendix 2). The *Extended Qualia* relations allow to express fine-grained distinctions for describing the componential aspect of a word's meaning along different points of view – its general characterization, composition, origin and function – and for capturing the nature of its relationships to other word senses. Qualia relations link either intracategorial or cross-categorial semantic units: they enable to organize and interrelate entities through taxonomic and partonomic semantic links and to connect them to events strictly related to their meaning by means of non-taxonomic links that supply contextual /collocational information[3]. The isa relation, on the other hand, enables to generalize over properties shared by lexical entries and to subdivide concept-denoting lexical units which share a semantic type, thus creating virtual subtypes[4].
In the following, particular focus is placed on the types of lexical relationships expressible by the SIMPLE *Extended Qualia Structure*.

3. Medicine Domain in the Lexicon

In the PAROLE-SIMPLE-CLIPS lexicon, the ontological classification of lexical units belonging to the sphere of medicine[5] – which are all retrievable through their 'domain' information label – ranges over the main type hierarchies of the SIMPLE ontology, i.e. Concrete entity, Abstract entity and Event.

3.1. Medical Specialties

Medical specialties, i.e. discipline denoting entities, which are clustered under the DOMAIN semantic type, are characterized, where relevant, by conceptual part-whole correlations through meronymic and holonymic constitutive relations and connected, through an associative link, to conceptually related words: [*cardiochirurgia* isa *disciplina*] [–[6] is_a_part_of *chirurgia*], [– concerns *cuore*], [– concerns *intervento*] (cardiac surgery, discipline, surgery, heart, operation); [*oncologia* concerns *tumore*] (oncology, cancer).

3.2. Healthcare Operators and Consumers

Humans related to the health domain are distinguished into 'healthcare operators' and 'medical patients' in terms of a different ontological classification.

3.2.1. Healthcare Operators

Healthcare operators/providers are ontologically classified under the semantic type PROFESSION and linked through a telic relation to their typical occupational activity: [*chirurgo* isa *medico*] [– is_the_activity_of *operare*] (surgeon, physician, operate). On the other hand the connection to their domain of activity: *urologo, urologia* (urologist, urology) is provided by the 'domain' information feature.

3.2.2. Medical Patients

Medical Patients, which are instances of the semantic type PATIENT_OF_EVENT, are linked by an agentive relation to the typical event they underwent (or are undergoing) and from which the human-denoting word is, most of the time, morphologically derived: [*ammalato* isa *persona*] [– agentive_prog[7] *ammalarsi*] (sick person, person, fall ill), [*ustionato* agentive *ustionare*] (burnt person, burn). Related to patients are the events they are affected by, i.e. the diseases or disorders: [*lebbroso* isa *malato*] [– affected_by *lebbra*] (leper, leprosy); [*diabetico* affected_by *diabete*] (diabetic, diabetes); [*cardiopatico* affected_by *cardiopatia*] (cardiopathic, cardiopathy). Diseases and disorders, in turn, are connected to their symptoms.

3.3. Health Conditions

Entities that can be construed as symptoms span over different sub-hierarchies of events. They are felt either as phenomena, perceptions or non relational acts. PHENOMENON-typed symptoms are related, where possible, to their effect and to the affected entity, through constitutive relations: [*prurito* causes *arrossamento / irritazione*] (itch, reddening / irritation), [– affects *pelle / mucosa*] (skin / mucosa). PERCEPTION-typed symptoms are related to the 'instrument' of perception: [*dolore* instrument *senso*] (pain, sense). Those classified as NON_RELATIONAL_ACT are linked in the constitutive role to the affected physiological function and to the 'instrument' body part: [*tosse* affects *respirazione*], [– instrument *gola*] (coughing, respiration, throat).
Diseases and disorders, which are subsumed by the dedicated semantic type DISEASE, are linked through constitutive relations to the affected body part, to the illness effect and, where relevant, to the typically affected subject. Wherever possible, diseases are moreover related by means of an agentive relation to their causal agent, relation that enables to create somehow a taxonomy of diseases: [*parotite* isa *malattia*] [– typical_of *bambino*], [– affects *ghiandola*], [– causes *gonfiore*], [– caused_by

[2] For an exhaustive description of the information content of a semantic unit, see Ruimy et al. (2002).
[3] Such links, which are intimately related to the word's predicative structure, are most useful for sense disambiguation.
[4] It is worth noting that, besides the relationships expressed by qualia relations, synonymic, polysemic and derivational links are also encoded in the semantic representation of an entity.
[5] We do not only intend terms denoting anatomical parts, pathologies or medical procedures, but also actors (agents and patients), instruments, locations, etc.
[6] '–' stands for the first member of the previous relation.
[7] The qualia relation 'agentive_prog' links an individual (Semantic Unit 1) to the ongoing action/event (Semantic Unit 2) he is performing or undergoing.

virus] (parotitis, disease, child, gland, swelling, virus) while [*malaria* caused_by *parassita*] (malaria, parasite).

3.4. Anatomical Parts

Clustered under the semantic type BODY_PART are the (affected) anatomical parts which all bear the semantic label 'Anatomy' and are characterized by meronymic and holonymic links: [*mano* is_a_part_of *braccio*], [– has_as_part *dito / palma / dorso*] (hand, arm, finger, palm, back).

3.5. Medical Procedures

Medical procedures, i.e. acts typically performed by healthcare operators, are typed as PURPOSE_ACT and sub-classified according to their specific nature by means of the hypernymic relation: [*amniocentesi* isa *esame*] (amniocentesis, test), [*isterectomia* isa *intervento*] (hysterectomy, operation) or [*chemioterapia* isa *terapia*] (chemotherapy, therapy). Their goal is expressed in the telic role by means of the 'purpose' relation: [*vaccinare* purpose *prevenire*] (vaccinate, prevent), [*biopsia* purpose *diagnosticare*] (biopsy, diagnose). The medical instrument used for such procedures is specified, where relevant: [*ecografia* instrument *ecografo*] (ultrasound, echograph).

3.6. Medical Instruments

Specific instruments are characterized as to their artifactual nature, partonomic properties and function, according to the definition of the type INSTRUMENT, which they belong to. Instruments are also related to their typical user, e.g.: [*bisturi* isa *strumento*], [– created_by *fabbricare*], [– has_as_part *lama*], [– used_for *incidere*], [– used_by *medico*] (lancet, instrument, make, blade, cut, doctor). In this domain, instruments used by healthcare operators are obviously more numerous than those concerning medical patients, as e.g. [*termometro* isa *strumento*], [– created_by *fabbricare*], [– used_for *misurare*], [– measures *temperatura*] (thermometer, measure, temperature).

3.7. Drugs and Medications

As to drugs and medications, besides being sub-classified through the hypernymic relation, e.g.: *antibiotico / antinfiammatorio / analgesico / antipiretico*, etc., (antibiotic / anti-inflammatory / analgesic / fever-reducer) they are related to the event they are administered for – from the generic *curare* (heal) to more precise events, e.g. *anestetizzare* (anaesthetize) – and, where appropriate, to the specific affection to be cured: e.g.: [*antistaminico* used_against *allergia*] (antihistamine, allergy). Their characterization is therefore rather granular, e.g. [*antibiotico* isa *farmaco*], [– produced_by *microrganismo*], [– used_against *infezione*], [– used_for *prevenire*], [– used_for *curare*] (antibiotic, medicine, microrganism, infection, prevent, cure); [*morfina* isa *sostanza*], [– derived_from *oppio*], [– used_as *analgesico*] (morphine, substance, opium, analgesic).

3.8. Locations

Entities denoting places where healthcare is provided are ontologically classified as BUILDING and defined as to their mode and purpose of creation by means of agentive and telic relations. They are semantically related, where appropriate, to their meronyms or holonyms through specific constitutive relations: [*ospedale* isa *edificio*] [– created_by *costruire*], [– used_for *ricoverare / curare*], [– concerns *medico / infermiere / malato*], [– has_as_part *reparto*] (hospital, building, build, hospitalize, heal, doctor, nurse, patient, division) while [*lebbrosario* isa *ospedale*], …, [– concerns *lebbroso*]. Such conceptual entities are also metonymically related to their corresponding polysemic senses denoting healthcare institutions and body of medical workers, classified under the INSTITUTION and HUMAN_GROUP types respectively.

4. Enhancing the Network

The fragment of semantic network presented below[8] consists of 112 representative lexical units belonging to 24 SIMPLE ontological types and connected by means of 112 hypernymic and 105 non-hierarchical semantic links, expressed by 28 different types of qualia relations[9], besides synonymic and derivational links.

Further additional information might be captured by borrowing some EuroWordNet semantic relations encoded in the ItalWordNet lexical database. Of particular interest in the EWN model are the relations that allow to capture quite straightforwardly the connection existing between an event and its typical participants[10], its typical location and used instrument, e.g. [*operare* involved_agent *chirurgo*], [– involved_patient *paziente*], [– involved_location *ospedale*], [– involved_instrument *bisturi*] (operate, surgeon, patient, hospital, lancet) or conversely, through the reverse 'role' relations that link concrete or abstract entities to an event. On the other hand, the 'near_synonym' relation, used in IWN to characterize two synsets linked by a close relation (but yet not close enough as to collapse their respective variants in a unique synset) would enable retrieving data by allowing to link strictly related concepts such as *ospedale* and *clinica* (clinic) that are hardly definable as real synonyms, but whose relationship should desirably be expressed.

5. Final Remarks

PAROLE-SIMPLE-CLIPS being a general language lexicon, the number of lexical units that are related to the health domain is obviously restricted (about 4,000 lexical items

[8] Due to a manual layout, the graphical representation provided here is rather entangled because of the many edge crossings. For a better comprehension of the relationships holding among concept instances, a visualization tool should be used.

[9] Clearly, the whole set of health-related words encoded in the PAROLE-SIMPLE-CLIPS lexicon is interconnected by a higher number of semantic relations.

[10] Such a link allows to infer the relationship holding between those participants.

out of which 1,116 are fully encoded[11]); on the other hand the absence, to date, in the lexicon, of multi-word units which, we are fully aware of, are central to the medical domain is a severe limitation. Besides, the semantic characterization of medical terms was performed by lexicographers that are not field experts.

Beyond these limitations, this paper aimed at highlighting the fact that the SIMPLE lexical model, which efficiently meets the complex structuring and representational needs of general knowledge, is also able to adequately define the terms used to describe and represent an area of knowledge. In particular, the intention was to show that the SIMPLE general ontology, as it is, already allows to structure reasonably enough a domain knowledge but also to emphasize its versatility, flexibility and cross-domain portability.

In designing SIMPLE ontology, provision was in fact made to allow for the creation of new (language / domain-specific, or even more granular) types, without altering the overall ontological architecture. As to semantic relations, which constituted the semantic vocabulary for describing the distinctive properties of the types, their set was extendible as well. Actually, in order to face new representational needs emerging from the extension of lexical coverage, a few new types and relations were unproblematically added during the development of the large resource that implements this model.

Built for general knowledge structuring, SIMPLE ontology is likely to be too granular for domain knowledge requirements. The type system can however be easily simplified by discarding the most specific classes of concepts. It can also be customized, by identifying knowledge areas of primary importance to a domain and whereby fine-grained types and relations are crucially needed as well as less crucial zones where a top level classification is deemed sufficient.

Consensual, since they were designed to cope efficiently with multilingual knowledge representation and well-defined[12] in virtue of their multidimensional nature – expressed in terms of semantic relations, SIMPLE semantic types may, in our view, provide a classification and formal description of the basic concepts relevant to the medicine domain. They are able to account for their properties and restrictions and for their relationships, thanks in particular to the expressiveness of qualia relations for knowledge representation. A rather straightforward customization process, whereby hierarchies of types and more specific semantic relations are created, could render the SIMPLE ontology able to fully capture the conceptual organization of the medical domain. Such a structuration of a domain lexical knowledge might provide, on our opinion, a relevant contribution to the automatic extraction of relational data in an information retrieval application, while contributing to promote the reusability of general ontologies.

[11] Entries non-fully encoded are ontologically classified and bear relevant semantic features but are not, for the time being, linked to other word senses through semantic relations.

[12] Each conceptual type is associated to a *template*, i.e. a schematic structure that provides an explicit characterization of the type by means of a structured cluster of defining properties thus stating the constraints on type assignment.

References

Alonge, A., Calzolari, N., Vossen, P., Bloksma, L., Castellon, I., Marti, T., Peters, W. (1998). The Linguistic Design of the EuroWordNet Database, Special Issue on EuroWordNet. In N. Ide, D. Greenstein, P. Vossen (eds.), 'Computers and the Humanities', XXXII, 2-3, 91--115.

Guarino, N. (1998) Some Ontological Principles for Designing Upper Level Lexical Resources, in Proceedings of the First International Conference on Language resources and Evaluation, 527--534, Granada.

Lenci, A. et al. (2000). SIMPLE Linguistic Specifications, Deliverable D2.1, ILC-CNR, Pisa.

Lenci, A. (2000). Building an Ontology for the Lexicon: Semantic Types and Word Meaning, Workshop on Ontology-Based Interpretation of Noun Phrases, Kolding-Denmark.

Lenci A., Calzolari N., Zampolli A. (2003). SIMPLE: plurilingual semantic lexicons for natural language processing. In A. Zampolli, N. Calzolari, L. Cignoni, (eds.), Computational Linguistics in Pisa, Special Issue, XVIII-XIX, (2003). Pisa-Roma, IEPI. Tomo I, 323--352.

Miller, G., Beckwith, R., Fellbaum C., Gross D., Miller K.J. (1990). Introduction to WordNet: An On-line Lexical Database, International Journal of Lexicography, III, 4, 235--244.

Pustejovsky J. (1995). The Generative Lexicon, The MIT Press, Cambridge, MA.

Pustejovsky J. (1998). Specification of a Top Concept Lattice, ms., Brandeis University.

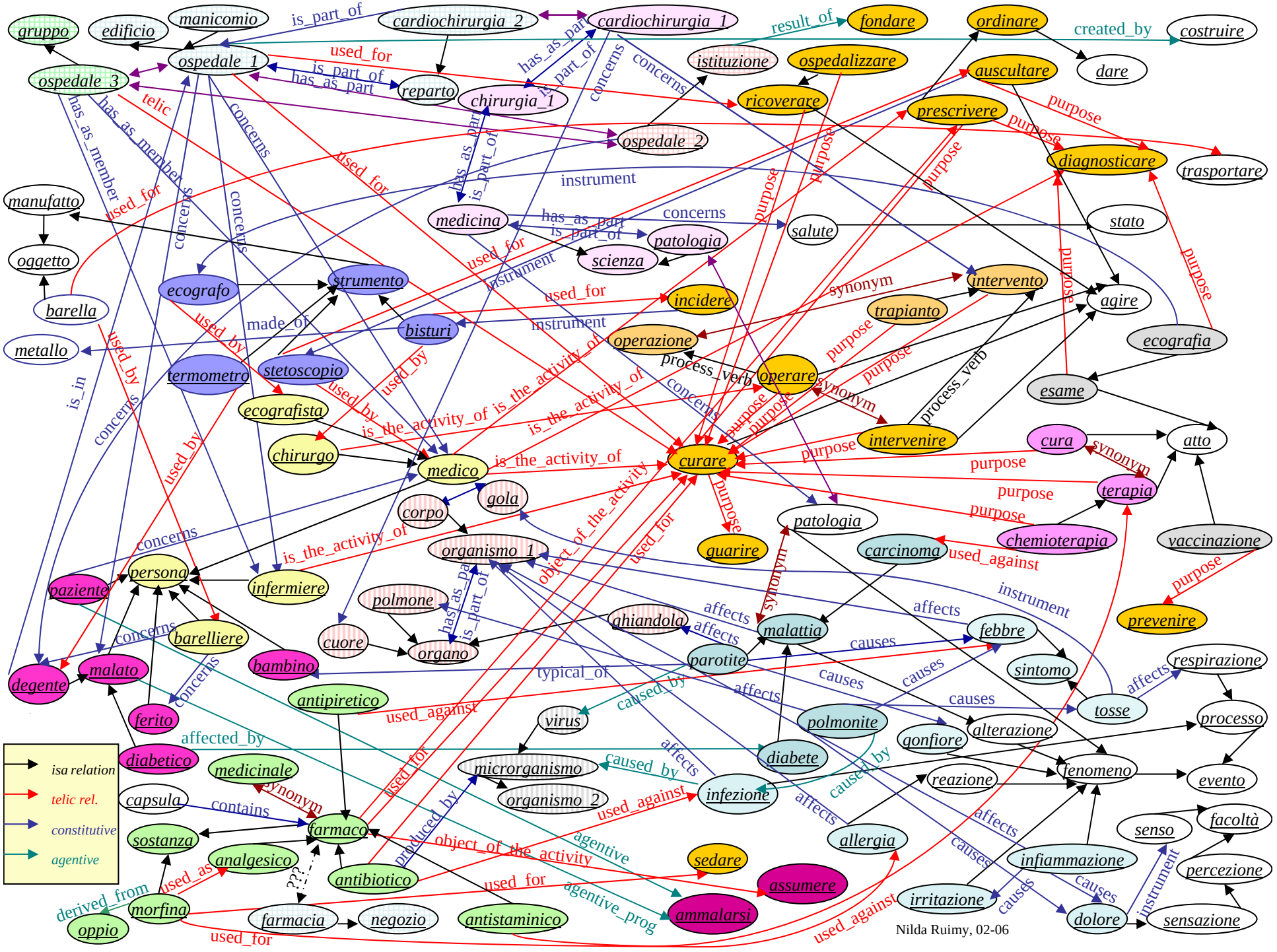
Roventini, A., Alonge, A., Bertagna, F., Calzolari, N., Cancila, J., Girardi, C., Magnini, B., Marinelli, R., Speranza, M., Zampolli, A. (2003). ItalWordNet: Building a Large Semantic Database for the Automatic Treatment of Italian. Computational Linguistics in Pisa, Special Issue, XVIII-XIX, Pisa-Roma, IEPI. Tomo II, 745--791.

Ruimy N., Monachini M., Distante R., Guazzini E., Molino S., Ulivieri M., Calzolari N., Zampolli A. (2002). CLIPS, a Multi-level Italian Computational Lexicon, LREC 2002, in Proceedings of the Third International Conference on Language Resources and Evaluation, Vol. III, 792--799, Las Palmas de Gran Canaria.

Ruimy N., Monachini M., Gola E., Calzolari N., Del Fiorentino M.C., Ulivieri M., Rossi S. (2003). A computational semantic lexicon of Italian: SIMPLE . In A. Zampolli, N. Calzolari, L. Cignoni, (eds.), Computational Linguistics in Pisa, Special Issue, XVIII-XIX, (2003). Pisa-Roma, IEPI. Tomo II, 821--864.

Ruimy N., (2006). Merging two Ontology-based Lexical Resources, LREC2006, in Proceedings of the Fifth International Conference on Language Resources and Evaluation, Genoa.

Appendix 1: Fragment of semantic network of health-related vocabulary



Nilda Ruimy, 02-06

Appendix 2: The 60 Extended Qualia relations

The 28 relations which provide the structure of the fragment of network (Appendix 1) by linking 112 health-related word senses are marked in grey.

Extended Qualia Relations			
Formal role	Constitutive role	Agentive role	Telic role
isa	**has_as_property**	**derived_from**	**used_for**
antonym_comp	**related_to**	**resulting_from**	**purpose**
mult_opposition	**constitutive**	**agentive_prog**	**object_of_the_activity**
antonym_grad	**typical_of**	**affected_by**	**used_as**
antonym	**quantifies**	**agentive_experience**	**indirect_telic**
	is_in	**result_of**	**is_the_activity_of**
	measures	**source**	**used_against**
	concerns	**created_by**	**is_the_ability_of**
	property_of	**agentive**	**used_by**
	uses	**caused_by**	**telic**
	resulting_state		**is_the_habit_of**
	has_as_effect		
	typical_location		
	affects		
	feeling		
	precedes		
	measured_by		
	kinship		
	is_a_part_of		
	instrument		
	has_as_part		
	successor_of		
	produces		
	contains		
	has_as_colour		
	is_a_follower_of		
	made_of		
	causes		
	is_a_member_of		
	lives_in		
	has_as_member		
	produced_by		
	constitutive_activity		
	relates		