Terminographic definition and concept representation[1]

Pamela Faber
University of Granada

One of the objectives of the research project OncoTerm (PB98-1342) is to facilitate the translation of medical texts within the domain of Oncology by elaborating a bilingual terminological database, based on the information extracted from specialized texts as well as medical dictionaries. Medical concepts are organized in categories represented by templates, which are systematically applied to all category members. The definitional information within each term entry is thus totally coherent with the information regarding other terms within the same conceptual category. This is conducive towards the specification of a language of terminographic definition, which is concise, consistent and applicable not only to the domain of oncology, but also extensive to other medical domains and other languages.

1. Introduction

One of the principal tasks of terminology management is the representation of the conceptual structure of domains of specialized knowledge. In the research project ONCOTERM (PB98-1342) we are in the process of elaborating a bilingual terminological database of the specialized domain of Oncology, based on the information extracted from a corpus of scientific texts as well as the entries in medical dictionaries. In our database, the terminographic definitions play an important role, since they are conceived as the natural language translation of the conceptual structure of the domain. However, in order for this to be so, they must be coherent in regards to both their micro- and macrostructure. Such coherence can only be the result of a previous analysis of textual data.

One source of information that we have used to specify interconnections between specialized medical concepts is that encoded in dictionaries. In our opinion, the information in dictionaries constitutes a lexical-conceptual network that is in direct relation with the knowledge expressed. The analysis of dictionary entries shows how terms are structured in chains such as the following:

(1) {particle beam radiation therapy}@→{external radiation therapy}@ →{radiation therapy}@ →{treatment}@ →{event}@→{ALL}

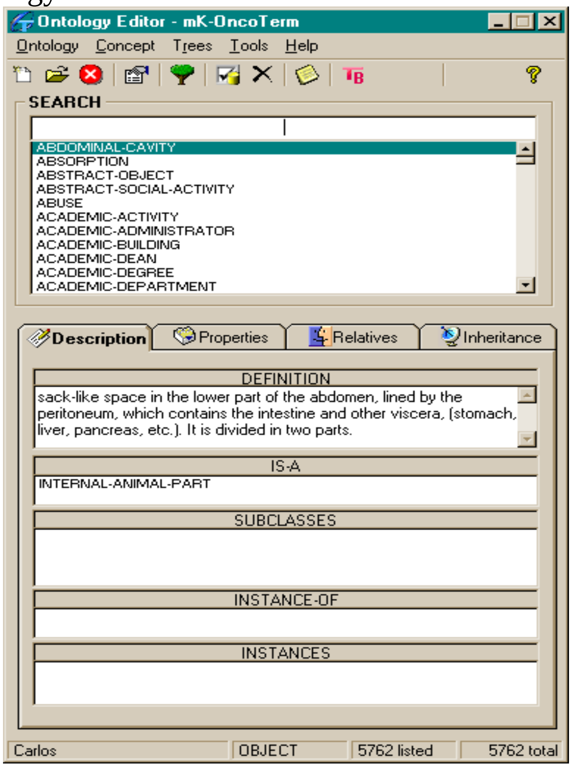
Such lexical chains reveal knowledge parameters, which are specified in the differentiating features of interrelated terms (Meijs and Vossen 1992: 144-145).

The conceptual organization of our termbase is derived from such information, which we are using to extend the Mikrokosmos Ontology, an already existing knowledge resource, developed by the Computing Research Laboratory of the University of New Mexico. We use the conceptual information in this ontology, which is considered non-language-specific, as the means to link terms in different languages. In the Mikrokosmos approach, an ontology is conceived as a language-neutral body of knowledge about the world. It constitutes a repository of primitive symbols used in meaning representation, which are interconnected by means of a rich system of semantic and discourse-pragmatic relations defined among the concepts (Mahesh and Nirenburg 1995: 1).

The data of our project is represented and stored in a terminology management system called Ontoterm (www.ontoterm.com), computer application that has been developed and implemented by one of the members of our group, Prof. Moreno Ortiz of the University of Malaga (Moreno Ortiz 2000ab; Moreno Ortiz y Pérez Hernández 2000). The necessity of elaborating a new resource for the project was the result of deficiencies perceived in existing systems, such as Multiterm. Ontoterm allows the reusability of the Mikrokosmos Ontology, which we are in the process of extending. At the last count, our ontology has 5833 concepts, of which 397 are relations.

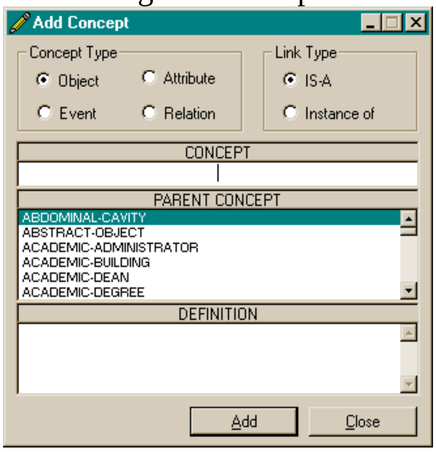
The architecture of Ontoterm has two main modules: an ontology editor and a termbase editor. In the ontology editor is where the modelling of conceptual structure takes place:

() Ontology editor



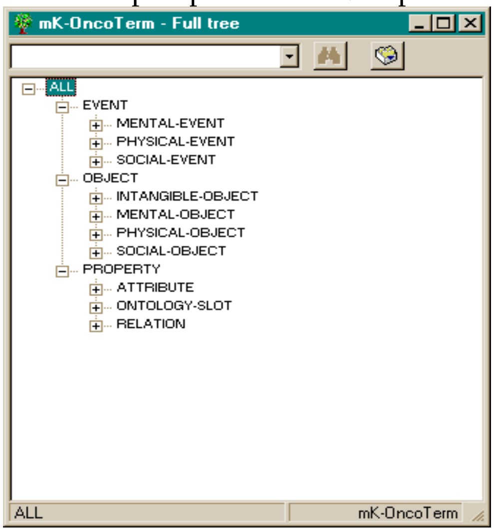
The ontology editor shows the list of concepts in alphabetical order. Each concept appears with its description, properties, relations with other concepts, and inheritance. This module has two submodules. The first submodule () is where new concepts are added to the ontology. When a new concept is added, it is necessary to indicate its parent concept so that its location in the ontology is properly specified.

(5) Submodule for adding new concepts



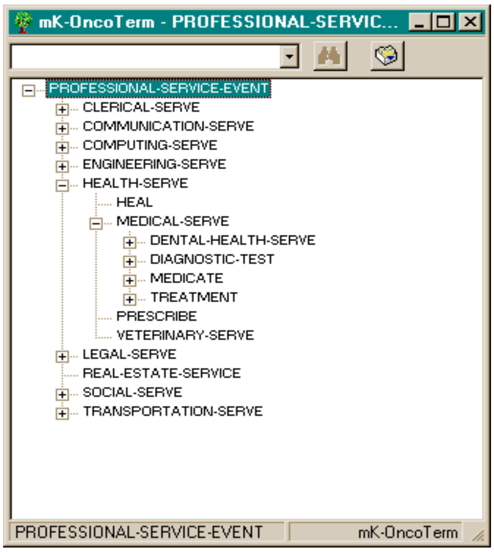
The conceptual hierarchy elaborated in Ontoterm can be visualised in the second submodule in which all of the concepts appear in a tree structure. The top concepts of the Mikrokosmos ontology are shown in (6):

(6) Submodule for concept representation: Top concepts



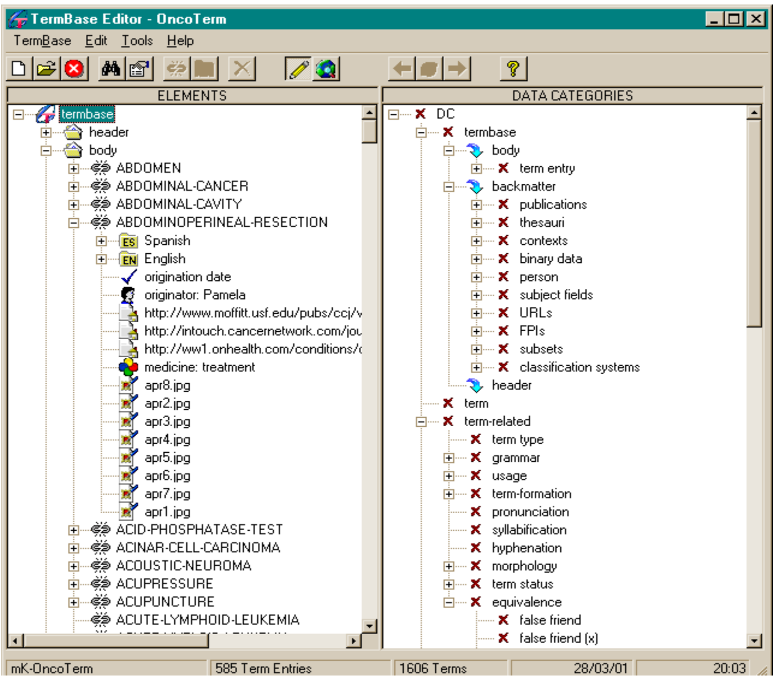
Evidently, specialised concepts are integrated at more specific levels of the ontology. For example, concepts such as TREATMENT or DIAGNOSTIC-TEST, are children of MEDICAL-SERVE.

(7) Submodule for the visualisation of concepts: HEALTH-SERVE



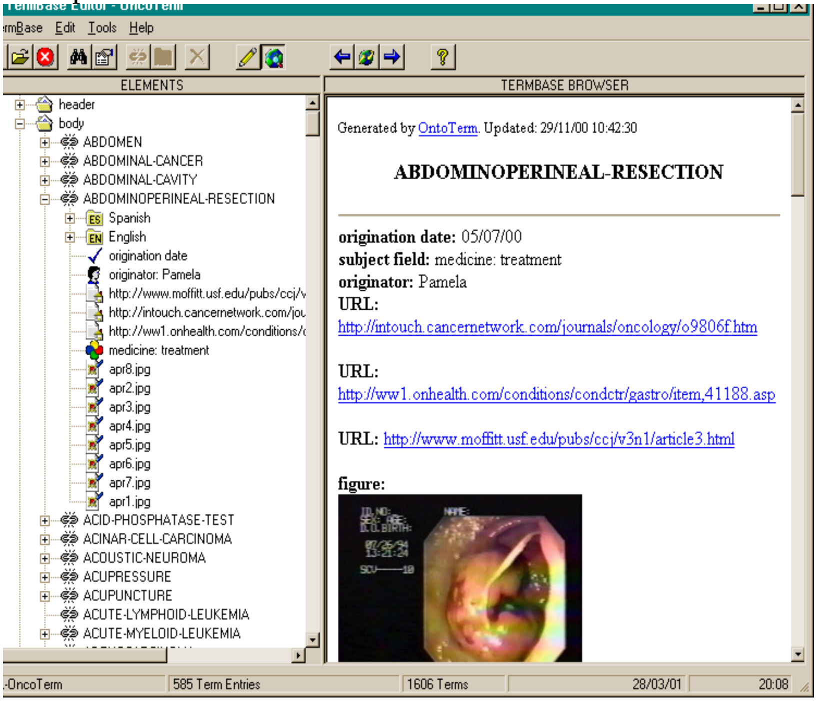
The second module of Ontoterm is the Termbase Editor, which stores the terminological information for the concepts in the ontology. This module complements the ontology editor and describeses the linguistic representation/s of each concept with the data categories specified in ISO 12620.

() Termbase editor



The information contained in the termbase can also be represented grafically in the form of HTML reports.

() HTML Report



2. Conceptual Organization

The organization of concepts is an activity that should be carried out systematically, and have a theoretical base, something that has not always been the case in Terminology. One way of structuring conceptual information is evidently to formulate an *ad hoc* list of conceptual domains and subdomains, and then to assign terminological units to the categories considered most suitable. However, this type of *top-down* processing, which often has been used in the elaboration of thesauri and databases, is entirely dependent on the intuition of a particular group of lexicographers/terminologists. Alternatively, one can use an exclusively bottom-up approach and work upwards to arrive at a list of conceptual categories by starting from available data in texts and dictionary entries,. Our approach entails a combination of the two methods. We have used as the starting point for our bilingual database an unstructured list of 2,500 cancer-related terms in English and Spanish. The top-concepts were arrived at through the analysis of definitions as well as through calculations of frequency in the corpus of specialized texts that we have elaborated. The texts belonging to our corpus encode the following communicative contexts:

(2) Communicative contexts represented in the OncoTerm corpus

SENDER	⇨	RECEIVERS
SPECIALIST	⇨	**Medical Students**
	⇨	**Health Professionals**
	⇨	**Specialists**
	⇨	**Oncologists**

In this way, we have managed to identify eight conceptual categories, the majority of which are applicable to other medical domains (Faber 1999: 99; Faber y Mairal 1999):

(3) Conceptual Categories

- ❑ DIAGNOSTIC_PROCEDURE ❑ TUMOR ❑ INSTRUMENT ❑ DRUG
- ❑ BODY_PART ❑ TREATMENT ❑ SPECIALIST ❑ MEDICAL_INSTITUTION

We have integrated these categories in the Mikrokosmos Ontology at the concept node MEDICAL-SERVE. Initially, we thought that *symptom*, *risk factor* and *side effect* could also be separate categories in the ontology. However, on observing that ASBESTOS, SMOKING and SUNLIGHT (*risk factors*) or COUGH and FATIGUE (*symptoms*) were already included in the ontology at more primary levels, we decided that this information would be better represented in the form of relations. For example, SYMPTOM-OF and its inverse, HAS-SYMPTOM:

(4) Conceptual relations

	CONCEPTUAL RELATION ⇓	
COUGH	**SYMPTOM-OF**	LUNG_CANCER
LUNG_CANCER	**HAS-SYMPTOM**	COUGH
SMOKING	**RISK-FACTOR-OF**	LUNG_CANCER
LUNG_CANCER	**HAS-RISK-FACTOR**	SMOKING

Our assertion is that the internal structure of each category and the knowledge parameters upon which such structure is based should be explicit in the descriptions of its member concepts. As shall be seen, the nuclear part of the definition of each concept is indicative of the category to which it belongs, as well as its level of specificity.

2.1. Terminographic definition and conceptual structure: TREATMENT

According to Bejoint (1997: 19-20) definitions have never been given due importance in Terminology. In many termbases, definitions are simply inserted in a cut-and-paste fashion from other dictionaries, termbases or knowledge resources, without taking into consideration both their internal as well as external coherence. However, definitions are mini-knowledge representations, and accordingly, the organization of information encoded in definitions should be structured both in regards to its perceptual salience well as in regards to the definitions of other related concepts within the same category.

For example, within the domain of ONCOLOGY, TREATMENT, is one of the top concepts. Its immediate subordinate concepts are RADIATION THERAPY, CHEMOTHERAPY, SURGICAL PROCEDURE, and BIOLOGICAL THERAPY. The position of RADIATION THERAPY in the domain is firstly evident in the fact that the nucleus of its definition in different dictionaries is *treatment*:

(5) Radiation therapy definitions

RADIOTHERAPY / RADIATION THERAPY	
HarperCollins Medical Dictionary	the **treatment** of disease by any radioactive substance or radiant energy.
The Cancer Dictionary	the use of high-energy penetrating rays or subatomic particles to **treat** or control disease.
Stedman's Concise Medical Dictionary	**treatment** with x-rays or radionuclides.
On-line Medical Dictionary (www.graylab.ac.uk/omd)	**treatment** with high energy radiation from X-rays or other sources of radiation

Although these dictionaries differ as to the type and quantity of information in their definitions of the same concept, they all coincide in designating *radiation therapy* as a kind of TREATMENT. If we wish to elaborate an adequate definition, it is not viable to merely copy one of these definitions in our termbase. Rather it is necessary to consider the knowledge parameters that define the conceptual category, and use this information as a blueprint for all of the definitions of its member concepts.

In this respect, it is important to emphasize that definitions are not just given information, but are constructions in which the knowledge parameters specified confer different types of focus on an entire domain, as is evident in the following representation of *radiation therapy*:

(6) Radiation therapy: Definitional hierarchy

TREATMENT	
radiation therapy treatment involving the use of high-energy rays to damage cancer cells and stop them from growing and dividing.	
[INSIDE] ⇓	**Internal radiation therapy** radiation therapy in which radioactive material is placed in or near tumor.
[BODY_ PART] ⇓	**intracavitary radiation therapy** internal radiation therapy in which radioactive material is placed in a body cavity (e.g. uterus, chest, vagina, etc.)
[PELVIS/ ABDOMEN]	**intraperitoneal radiation therapy** intracavitary radiation therapy in which radioactive material is placed directly in the pelvic and abdominal cavities. This is a treatment for ovarian cancer. [...]
[OUTSIDE] ⇓	**External radiation therapy** radiation therapy in which the high-energy rays used to treat the cancer patient come from a machine, located at some distance from the body.
[BEAM_ TRAJECTORY]	**stereotactic radiation therapy** (external) radiation therapy in which a number of precisely aimed beams of ionizing radiation coming from different directions meet at a specific point, delivering the radiation treatment to that spot. [...]
[BEAM_TYPE]	**particle beam therapy** (external) radiation therapy in which high energy radiation is delivered to tumor cells through the use of fast-moving subatomic particles (protons and neutrons). [....]

In (6) we can see a partial representation of the subdomain of RADIATION THERAPY. The outside column shows the type of conceptual information represented in the differentiating information of the concepts. Rather than copying definitions from other sources, we have elaborated the definitions ourselves and extracted the differentiating information by means of corpus analysis, in which concordances are grouped to show different conceptual distinctions.

(7) Radiation therapy concordances

RADIATION_THERAPY	
RADIATION_SOURCE_LOCATION	
1	516-520, 1981. Bagshaw MA: External **radiation therapy** of carcinoma of prosta
2	r near the tumor. Also called internal **radiation therapy** or implant radiation.
↳ **INTERNAL_RADIATION_THERAPY**	
3	] 4. Laser therapy or interstitial **radiation therapy** for endobronchial
4	ntimeters thick: 1. Intracavitary **radiation therapy**. In most instances, 6,
↳ **EXTERNAL_RADIATION_THERAPY**	
↳ **BEAM_TRAJECTORY**	
5	c radiosurgery and stereotactic **radiation therapy**. stereotaxis (
↳ **BEAM_TYPE**	
6	il and intraoperative electron beam **radiation therapy** on the outcome of pati
7	radiosensitizers, or particle-beam **radiation therapy**. [14-17] 4. Isotre
8	to chemotherapy.[7-9] Fast neutron beam **radiation therapy** or accelerated hyper

9	ltileaf collimator.[25] Proton-beam **radiation therapy** is also under investig
10	al with mixed-beam (neutron/photon) **radiation therapy**, compared to standard
	RADIATION_BEAM_TARGET
	↳ **BODY_PART**
11	m the NWTS-3 demonstrate that abdominal **radiation therapy** does not provide sig
12	metastases are identified, whole brain **radiation therapy** (30 gray in 2 gray f
13	s incorporating chemotherapy plus chest **radiation therapy** are listed below. Th
14	pts to lower the dose of craniospinal **radiation therapy** to 2,340 cGy have resu
15	tive chest wall and regional lymph node **radiation therapy** are undergoing reass
16	Jett JR, McGinnis WL, et al.: Thoracic **radiation therapy** alone compared wit
	↳ **FIELD_COVERAGE**
17	of whole pelvis versus small-field **radiation therapy** for carcinoma of prost
18	us 2 months of ABVD plus extended-field **radiation therapy** is being conducted b
19	arbell NJ, Silver B, et al.: Wide-field **radiation therapy** with or without ch
	RADIATION_ADMINISTRATION
	↳ **DOSAGE**
20	ged remissions. The need for low-dose **radiation therapy** is under study.[1] The
21	pse after initial wide-field, high-dose **radiation therapy** have a good prognosi
22	using accelerated superfractionated **radiation therapy** for advanced squamous
23	e.[6,11-18] 3. Novel fractionation **radiation therapy** clinical trials are un
24	l evaluation: 1. Hyperfractionated **radiation therapy** to improve tumor contr
	↳ **INTENSITY**
25	ned 5-fluorouracil and supervoltage **radiation therapy** of locally unresectabl
26	years' experience with megavoltage **radiation therapy**. Cancer 37(6): 2605-26
27	ectron-beam irradiation or orthovoltage **radiation therapy** may be used to pal

Although the radiation therapy concordances in (7) are not exhaustive, they offer a template on the basis of which the rest of the subdomain can be modeled. Among other things, they inform us that *external radiation therapy*, as the default value for the domain, has a greater variety of conceptual distinctions than internal radiation therapy. The types of internal radiation therapy that appear in the corpus are *intracavitary radiation therapy* and *interstitial radiation therapy*. In contrast, *external radiation therapy* shows conceptual distinctions such as the kind of radiation used, the trajectory of the beam, the part of the body targeted, the extension of the targeted area, as well as the dose and intensity of the radiation. As can be seen, the corpus data is in consonance with the definitional hierarchy for radiation therapy proposed in (6).

The corpus data also shows the existence of other secondary conceptual distinctions, which are related to the status of radiation as an EVENT. These distinctions are related to contextual factors such as when the radiation is delivered in combination with other treatments (*preoperative radiation therapy, postoperative radiation therapy*), or its function according to the phase of the illness (*curative radiation therapy, palliative radiation therapy*).

2.2. Category templates

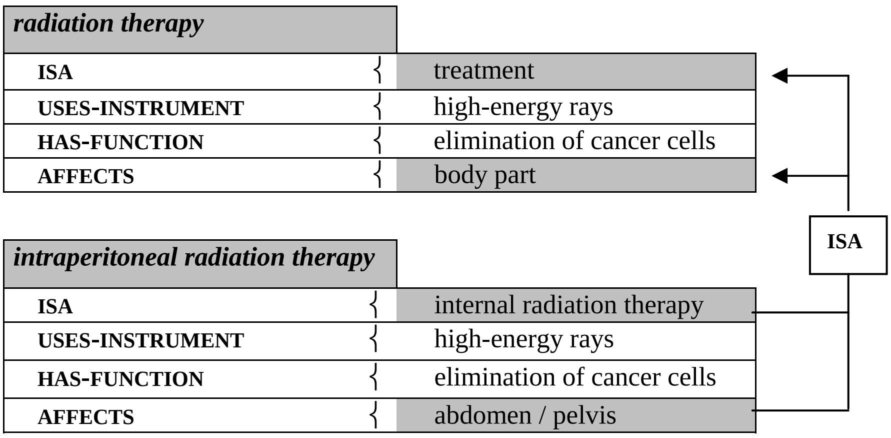
The internal structure of each conceptual category is represented by a set of conceptual relations, which acts as a model of the category in question. The template for TREATMENT would be the following:

(8) Category template: Treatment

CONCEPTUAL CATEGORY	**CONCEPTUAL RELATION**
TREATMENT	**IS-A**
	USES-INSTRUMENT
	HAS-FUNCTION
	HAS-LOCATION

If this basic template is projected onto *radiation therapy*, the values generated for the relations are in direct correspondence to the information in the definitions of the terms. In a similar way, the template is also valid for subordinate terms such as *intraperitoneal radiation therapy*. When the template is applied, more specific values are generated. Inheritance is thus evident at various levels.

(9) Category template: intraperitoneal radiation therapy



Each conceptual category thus has a prototypical template of conceptual relations, which can be used to format the definition of terms. The definitions are the natural language translation of the conceptual relations.

3. Conceptual categories and the verbal lexicon

Terminological studies normally focus on concepts, which in most cases are linguistically represented by nominal forms. However, both in the comprehension and structure of specialized discourse, verbs play an important role. This is due to the fact that a considerable part of our knowledge is composed of EVENTS and STATES, many of which are linguistically represented by verbs[2]. For a true understanding of specialized texts, this type of processual knowledge cannot be ignored.

Both specialized and general communication use the same inventory of verbs, though in specialized communication the arguments are generally terms. Another important difference is that in general language discourse these verbs can be highly polysemic, while in specialized communication their meaning is restricted. This restriction of meaning is also extensive to the semantic characteristics of their arguments

For example, *respond* is a verb that frequently appears in our corpus in reference to the category of TREATMENT . It can have the following meanings in general language discourse:

(10) *Respond:* general language meanings

RESPOND	
Longman Dictionary of English Language and Culture	**1** to say or write (something) in reply
	2 to do something in answer.
	3 to get better as a result of a treatment.

However, in specialized medical discourse, three meanings are reduced to one:

(11) *Respond*: restrictions

RESPOND (meaning in specialized discourse)	to get better as a result of a treatment.

In both kinds of texts, the verb imposes structure on the discourse since it determines the number of arguments in each proposition, as well as their semantic characteristics and function. However, what is invariably overlooked is the fact that predicates in specialized texts are also instrumental in the analysis of conceptual structure. Verbs can be related to the conceptual categories that characterize the arguments they normally appear with in medical texts. In such cases, we have even found that verbs can trigger entire hierarchies of concepts and provide useful information in the structure of conceptual categories.

The following selection of concordances show that that *respond* as a predicate is monosemic in medical texts and is indicative of TREATMENT:

(12) *Respond*: Concordances

RESPOND			
1	r estrogen and progesterone receptors	**respond**	best to progestin therapy. Among
2	aid that stage I and stage II cancers	**respond**	equally well to radiation or sur
3	onal cell cancer of the urethra may	**respond**	favorably to the same chemothera
7	of evidence: 3iiiDi] For patients who	**respond**	to neoadjuvant chemotherapy, loc

8	w transplantation, some patients will	**respond**	to interferon alfa.6 Infusions
9	onding patients who relapse usually	**respond**	to retreatment with interferon a
10	after relapse. The primary group may	**respond**	to high-dose chemotherapy and au
11	atonin may affect the way tumor cells	**respond**	to chemotherapy and radiation th
13	nd secondary refractory patients who do	**respond**	to induction chemotherapy, but
14	esponse rate of 20% in those who do not	**respond**	to standard progesterone therapy
15	topenia or hemolytic anemia who fail to	**respond**	to alkylating agents and pre
16	resected or metastatic tumors failed to	**respond**	to chemotherapeutic agents f
17	r other mechanical problems expected to	**respond**	to antineoplastic therapy
18	Tamoxifen: Some patients (18%) will	**respond**	to tamoxifen (20 milligrams
19	bHCG and AFP. Certain of these tumors	**respond**	to platinum-based combination ch
21	" tumor resection. Low-grade tumors may	**respond**	to various chemotherapeutic re
22	ow and serious infections that do not	**respond**	to antibiotics.20 Prophylactic o
23	on: Patients whose disease does not	**respond**	to combined radiation therapy an
24	effective.1,3,4 Patients who do not	**respond**	to a cisplatin-based combination
25	nodal relapse. Patients who do not	**respond**	to induction chemotherapy (about
27	indolent lymphoma: half of patients	**respond**	to a four-dose treatment program
28	TIC – to treat painful conditions that	**respond**	to nerve blocks (e.g., celia
30	emotherapy at the time of CNS relapse	**respond**	to second-line chemotherapy.13
31	the patients treated. Patients who	**respond**	usually demonstrate improvement
32	with metastatic disease at diagnosis	**respond**	well to the therapy given to pat
33	is continued; if the tumor did not	**respond**	, an alternative regimen is used.
34	th newly diagnosed medulloblastoma will	**respond**	, at least partially, to chemot
36	dermal tumors and pineoblastomas will	**respond**	, at least partially, to chemothe
37	and vincristine.1 If the tumor fails to	**respond**	, it may be a benign lesion.
38	and reticulocyte counts usually do not	**respond**	. The effect of GM-CSF treatment

The concordances show that in the specialized texts of our corpus, *respond* has two arguments, each one representative of a specific conceptual category:

(13) Basic proposition

First Argument	Predicate	Second Argument
(condition)$_{\text{Affected}}$	RESPOND	(treatment)$_{\text{Effector}}$

Furthermore, if we analyze the concordances for *respond* in (12), the terms that fill the argument slots can be organized in hierarchies, which activate different areas of the ontology. In this sense, the information extracted from our corpus is thus one of the sources for the conceptual structure of the domain. In (14) concordance data has been arranged to make explicit the conceptual category of TREATMENT:

(14) Treatment hierarchies based on corpus data

First Argument	Predicate	Second Argument
DISEASE ∑ painful condition	**RESPOND**	**TREATMENT**
CANCER ∑ small cell lung cancer [IS-A] ∑metastatic breast cancer [IS-A]		∑ retreatment ∑ second treatment ∑ four-dose treatment
TUMOR ∑ tumor cells [PART_OF] ∑ metastatic tumor [IS-A] ∑ unresectable tumor [IS-A] ∑ small cell carcinoma [IS-A] ∑ myeloma [IS-A] ∑ medulloblastoma [IS-A] ∑ epenymoma [IS-A] ∑ pinelblastoma [IS-A] ∑ lymphoma [IS-A]		**THERAPY** ∑ GM-CSF treatment standard therapy ∑ initial therapy ∑ ***chemotherapy*** ∑ neoadjuvant chemotherapy ∑ high-dose chemotherapy ∑ second-line chemotherapy ∑ induction chemotherapy ∑ anticancer drug ∑ chemotherapeutic agent ∑ alkylating agent ∑ platinum-based combination
PATIENT ∑ cancer patient [IS-A]		∑ cisplatin-based combination ∑ Cyclophosfamide ***radiation therapy*** ∑ combined radiation therapy ***Biological therapy*** ∑ GM-CSF treatment ∑ Interferon ∑ Interferon alfa 6

4. Conclusions

To structure the domain of ONCOLOGY, we have used both dictionaries and specialized texts to arrive at an inventory of conceptual categories as well as templates that characterize the definitions of the concepts within the category. Although these templates are flexible enough to permit the manipulation of ideas, they also have constraints. This necessarily means a well-defined system of attributes, entities and relations. In our opinion, terminographic definitions are the vehicle for specifying such a system of constraints, based on a system of templates. Each template is anchored to a well-designed conceptual ontology by means of which terms are related to each other on the basis of an underlying model of the world. As Nirenburg and Raskin (in press) so accurately point out, only then is it possible to justify the postulation of a certain number of theoretical concepts, a set of roles and features, and a prescribed range of values.

[1] This research was carried out within the framework of ONCOTERM (PB98-1342), funded by the Spanish Ministry of Education.

[2] As part of another research project, we designed a semantic network, which reflected the underlying conceptual meaning of the verbal lexicon. Our objective was to study the potential that this type of lexical structure has for knowledge representation. Accordingly, we analyzed and classified approximately 8,000 verb in English and Spanish, structuring them onomasiologically in hierarchical lexical domains according to the premises of the Functional-Lexematic Model (Faber and Mairal 1998, 1999). The structure of these lexical domains is implicit in the definitions of their members, a definitional stem based on the *Decomposition Principle* of Mel'cuk (1988).

5. References

Bejoint, Henri. (1997) “Regards sur al définition en terminologie” in *Cahiers de Lexicologie* 70, 1997-1, 19-26.

Faber, P. (1999) “Conceptual analysis and knowledge acquisition in scientific translation”. *Terminologie et Traduction 2 (1999)*, 97-123.

Faber , P. and R. Mairal (1998) Dominios y esquemas de predicado: hacia una productividad léxica. In Gerd Wotjak (ed.) *Teoría del Campo y Semántica Léxica*. Frankfurt: Peter Lang.

Faber, Pamela and Ricardo Mairal Usón (1999) *Constructing a lexicon of English verbs*. Berlin: Mouton de Gruyter.

Mahesh, K. and S. Nirenburg (1995) “A situated ontology for practical NLP”. In *Proceedings on basic ontological issues in knowledge sharing*. International Joint Conference on Artificial Intelligence (UCAI-1995), August 1995, Montreal, Canada.

Meijs, W. and P. Vossen. (1992) “In so many words: Knowledge as a lexical phenomenon”. In Pustejovsky, J. and S. Bergler (eds). *Lexical semantics and knowledge representation*. Berlin: Springer.

Mel'cuk, I. (1988) “Semantic description or lexical units in an explanatory combinatorial dictionary: Basic principles and heuristic criteria”. *International Journal of Lexicography* 1 (3), 165–188.

Nirenburg, S. and V. Raskin (in press) “Ten choices for lexical semantics”. To appear in *Computational Linguistics*.