From Text to Ontology: Extraction and Representation of Conceptual Information

Antonio Moreno & Chantal Pérez

Departamento de Filolgía Inglesa, Francesa y Alemana
Universidad de Málaga, Spain
[amo; mph]@uma.es

Abstract

In this paper we report work carried out within the OncoTerm[1] project, whose final aim is the development of a terminological database on the subject domain of cancer, as well as the publishing of the contents of this database on a website, which is to offer other related services. In this project we pay special attention to the extraction and formal representation of the conceptual structures underlying the domain of oncology. This approach to terminological work has determined to a large extent the tools and methodologies that we have deployed to attain our goals. In this paper we focus on some of the techniques used for the extraction of conceptual information from our special purpose corpora and on the representation and formalization of such information in the knowledge-based terminology management system designed and developed by one of the authors.

1. Introduction

In the OncoTerm project we have sought to manipulate both domain knowledge and the terminology associated with this domain; for this reason, commercial, off-the-shelf Terminology Management Systems (TMS) would not meet our requirements in terms of Knowledge Representation (KR), and therefore we set out to develop our own TMS. A first approach to this was the development of a database schema under a commercial relational database management system, the shortcomings of which soon became apparent. The system we have developed, OntoTerm™, is a full-fledged TMS that integrates KR techniques, allowing the creation and management of a hierarchically organized body of knowledge, capable of expressing user-defined relations among concepts with inheritance of properties, a visual representation of the hierarchy, and a number of other features.

Another important methodological assumption underlying our project is that domain-specific knowledge is not isolated from general world knowledge. Putting this into practice, however, implies the creation of large-scale language-independent resources, i.e. *ontologies*, that would fall outside the scope of a terminology project of the kind we have undertaken. Our approach has been to re-use an already existing such resource, the ontology created at the Computing Research Laboratory in New Mexico State University (USA) for the Knowledge-Based Machine Translation project known as Mikrokosmos. It is not the aim of the present

[1] This research has been partly carried out within the framework of the project *OncoTerm: System of oncological information and resources*, funded by the Spanish Ministry of Education (DGICYT) under code number PB98-1342.

paper to describe the adaptation of this resource to OntoTerm; this, and the integration process of domain-specific and general knowledge from the Mikrokosmos ontology can be found in Moreno Ortiz & Perez Hernandez (2000). The description of the Mikrokosmos ontology itself, can be found, for example, in Mahesh & Nirenburg (1995) and Mahesh (1996).

Another consequence of our knowledge-driven approach to terminology is that we should make sure that the knowledge that we plan to represent is as accurate and descriptive as possible. Extraction of conceptual information has been repeatedly quoted in the literature as a bottleneck in the compilation of terminologies (Ahmad & Fulford 1992; Meyer & Mackintosh 1996). Even though term extraction from text corpora has concentrated a lot of research effort in the recent years, with a few notable exceptions[2], not so much has been done for knowledge extraction, perhaps because the tools available for the representation of this sort of information were not widely available. In our project we have attempted to apply a number of techniques from corpus linguistics to knowledge extraction from domain-specific corpora. In what follows, we focus on the description of these techniques, but we also offer a brief description of the formalization and representation of the knowledge thus acquired and its integration in the Mikrokosmos ontology in OntoTerm.

2. Corpus-based extraction of conceptual information

The first step of our investigation was the compilation of the *special purpose corpus*[3] to be used to obtain domain knowledge. We gathered from various paper and electronic sources a series of oncology-related texts and used both internal and external criteria to select the texts included in the corpus. In the selection process, we took into consideration several criteria proposed by other scholars (Sinclair 1992:5; Bowker 1996:39; Meyer & Mackintosh 1996: 267; Pearson 1998: 52) and combined them with others specific to our research project. Those criteria refer to the quality, quantity, simplicity and documentation of the texts included[4]. Other important criteria relate to the linguistic status of the texts, their relevance to the subject domain, their age and the communicative setting in which the texts were produced and consumed.

The corpora that resulted from our selection process were two:

- the *cancer corpus*: 30 million words of cancer-related texts, with different levels of technicality and addressed to different audiences and
- the *leukemia subcorpus*: about half a million words of texts dealing specifically with leukemia and its treatment, directed to both specialists and patients and their families.

In addition to these two corpora, we have used a general language corpus commercially available, the *British National Corpus* (BNC), a 100 million word collection of samples of written and spoken language from a wide range of sources, designed to represent a cross-section of current English language. This corpus has been used as a reference corpus, with a

[2] See, for instance, the work carried out by the TIA working group and the proceedings of the EKAW'2000 Workshop on Ontologies and Texts available in http://www.irit.fr/wsontologies2000.

[3] Following Pearson (1998: 48), we use the term *special purpose corpus* to refer to "a corpus whose composition is determined by the precise purpose for which it is to be used", so it does not necessarily have all the properties of a *reference corpus* (for example, *representativeness*), although the language used in a *special purpose corpus* may not either deviate from the linguistic norm, as it would be the case in a *special corpus*.

[4] For a complete discussion on corpus selection criteria and a full list of the texts included in our corpora see Pérez Hernández (2000).

view to establish comparisons between the language used in general and specialised texts.

To extract conceptual information from our corpora we started by applying analysis techniques that have proved useful in corpus linguistics and corpus lexicography. We tool used to process the corpus is *WordSmith Tools*, a very powerful and versatile piece of software designed by Michael Scott and distributed by Oxford University Press.

We consider the corpus as a tool assisting the terminographer in the delimitation and organization of the subject domain (i.e. oncology). Thus, our first goal was to process the corpus to obtain not only term-candidates, but to identify other lexical items representing concepts relevant to the subject domain, as well as their distinguishing attributes and the relations that hold between concepts.

2.1. Extracting relevant concepts: key words and links between key words

In our approach, the first step to take in the search for conceptual information is the identification of *key words*, that is, words whose frequency is unusually high in comparison with some norm. To do so, we compared the word list in each special purpose corpora with the one in the BNC, our reference corpus. Key words are calculated by comparing the frequency of each word in the smaller of the two word lists (the one in the oncology corpus) with the frequency of the same word in the reference wordlist[5]. Key words are not necessarily the most frequent words in the texts, but those whose frequency is higher than expected in comparison with general language texts. In this way, key words provided a useful way to characterise the subject domain of our texts.

The study of the key words in the two special purpose corpora allowed us to identify a number of important conceptual areas relevant for the representation and organization of the knowledge pertaining to the domain of oncology, together with the key concepts integrated in each conceptual area. Table 1 shows the first 100 key words obtained by comparing the *leukemia corpus* with the BNC:

N	WORD	FRQ	LEUK LST %	KEY NESS
1	CELLS	4.415	0,86	30.763
2	LEUKEMIA	2.512	0,49	26.158
3	CELL	2.974	0,58	20.184
4	MARROW	1.560	0,30	14.661
5	LYMPHOMA	1.239	0,24	11.667
6	CHEMOTHERAPY	1.190	0,23	11.112
7	BLOOD	1.854	0,36	9.336
8	DISEASE	1.762	0,34	9.271
9	HODGKIN'S	846	0,16	8.882
10	THERAPY	1.193	0,23	8.483
11	MYELOMA	833	0,16	8.465
12	ACUTE	1.146	0,22	7.741
13	TREATMENT	1.687	0,33	7.692
14	BONE	1.083	0,21	6.999
15	TRANSPLANTATION	635	0,12	6.666
16	CHRONIC	950	0,18	6.579
17	CLL	601	0,12	6.176
18	REMISSION	702	0,14	6.080
19	AML	584	0,11	6.021
20	T	1.504	0,29	5.630
21	DOSE	794	0,15	5.297
22	LYMPHOCYTIC	508	0,10	5.068
23	CLINICAL	808	0,16	4.602
24	MYELOID	448	0,09	4.571

N	WORD	FRQ	LEUK LST %	KEY NESS
51	CYTOGENETIC	243	0,05	2.371,1
52	ALPHA	394	0,08	2.338,8
53	PERIPHERAL	362	0,07	2.317,4
54	MEDICINE	469	0,09	2.308,5
55	LEUKAEMIA	322	0,06	2.286,9
56	REGIMEN	264	0,05	2.282,0
57	HEMATOLOGY	214	0,04	2.246,5
58	PLATELETS	272	0,05	2.240,5
59	INDUCTION	319	0,06	2.136,5
60	PLASMA	344	0,07	2.111,8
61	ALLOGENEIC	203	0,04	2.108,6
62	LYMPHOID	230	0,04	2.058,6
63	MALIGNANCIES	213	0,04	2.046,5
64	LEUKEMIAS	191	0,04	2.005,1
65	DONOR	299	0,06	1.994,5
66	REFRACTORY	216	0,04	1.958,4
67	ABNORMALITIES	269	0,05	1.891,1
68	NORMAL	654	0,13	1.888,9
69	ASSOCIATED	587	0,11	1.883,7
70	INTERFERON	237	0,05	1.861,8
71	CUTANEOUS	184	0,04	1.810,4
72	TRANSPLANT	263	0,05	1.793,4
73	CLONAL	184	0,04	1.783,1
74	TOXICITY	217	0,04	1.744,2

[5] WordSmith tools uses the Log Likekihood statistical test of significance designed by Ted Dunning.

25	SURVIVAL	789	0,15	4.448
26	MEDIAN	576	0,11	4.073
27	LYMPHOMAS	418	0,08	4.011
28	CANCER	784	0,15	3.952
29	CML	378	0,07	3.781
30	MYELOGENOUS	350	0,07	3.674,3
31	LEUKEMIC	351	0,07	3.671,1
32	STEM	562	0,11	3.669,2
33	TUMOR	365	0,07	3.663,1
34	GENE	638	0,12	3.662,9
35	RELAPSE	424	0,08	3.599,2
36	LYMPHOCYTES	402	0,08	3.303,2
37	AUTOLOGOUS	319	0,06	3.252,5
38	APOPTOSIS	345	0,07	3.249,5
39	LYMPHOBLASTIC	315	0,06	3.167,0
40	DIAGNOSIS	519	0,10	3.069,4
41	ONCOLOGY	311	0,06	3.012,0
42	MULTIPLE	557	0,11	2.990,5
43	MALIGNANT	383	0,07	2.964,3
44	CANCERLIT	282	0,05	2.960,4
45	CHROMOSOME	386	0,08	2.904,2
46	TREATED	693	0,13	2.740,2
47	PROGNOSTIC	293	0,06	2.670,9
48	RESULTS	859	0,17	2.552,2
49	RELAPSED	260	0,05	2.456,0
50	LYMPH	288	0,06	2.394,1

75	GRADE	369	0,07	1.703,5
76	ETOPOSIDE	165	0,03	1.667,4
77	RESPONSE	597	0,12	1.663,3
78	PROGNOSIS	219	0,04	1.656,0
79	RADIOTHERAPY	213	0,04	1.654,6
80	INDUCED	315	0,06	1.622,1
81	HEMATOPOIETIC	147	0,03	1.543,2
82	ANEMIA	142	0,03	1.490,7
83	REGIMENS	164	0,03	1.483,7
84	RECEPTOR	231	0,04	1.464,0
85	CYCLOPHOSPHAMIDE	150	0,03	1.460,4
86	ANTIBODIES	243	0,05	1.442,3
87	HAEMATOL	137	0,03	1.438,2
88	PROTEIN	347	0,07	1.435,0
89	CHROMOSOMES	209	0,04	1.432,7
90	RADIATION	286	0,06	1.418,5
91	TUMORS	137	0,03	1.401,8
92	SERUM	269	0,05	1.386,2
93	ANALYSIS	551	0,11	1.369,7
94	ABSTRACTS	174	0,03	1.367,3
95	RANDOMIZED	148	0,03	1.330,4
96	PHASE	373	0,07	1.308,7
97	IRRADIATION	166	0,03	1.291,1
98	SPLEEN	163	0,03	1.283,7
99	NON	664	0,13	1.277,8
100	VITRO	205	0,04	1.268,6

Table 1: Key words in the Leukemia corpus (compared with the BNC corpus)

All these key words refer to concepts essential to the subject domain, and the conceptual areas in which they can be grouped give the terminographer a good estimate of the central knowledge areas that should be covered in the formal representation of the domain. It is also very interesting to note that many of the key words are adjectives, verbs or deverbal nouns. This is important because they represent properties of other concepts or relations denoting the way they relate to each other. Among the conceptual areas obtained with the study of key words we highlight the following:

- **Body parts and organic substances:** *cells, cell, marrow, blood, bone, stem, gene, lymphocytes, chromosome, lymph, platelets, plasma, proteins, antibodies, chromosomes, serum, spleen.* In this group we can also include two specialised adjectives, or related to a condition in the skin, *cutaneous,* and the other to the process of blood formation: *hematopoietic*.
- **Diseases and pathological processes:** *leukemia, disease, myeloma, lymphomas, cancer, apoptosis, leukaemia, malignancies, leukemias, abnormalities, tumor* and *anemia*. To these, we should add those that denote *types of* leukemia: *acute, chronic, lymphocytic, myeloid, myelogenous, lymphoblastic, multiple* and *lymphoid*.
- **Treatments**: *chemotherapy, therapy, treatment, transplantation, treated, dose, regimen, donor, receptor, transplant, radiotherapy, phase, radiation*, *irradiation* and *induction,* and two specialised adjectives: *autologous* and *allogeneic* that appear in combination with (*bone marrow) transplantation/transplant* and *(blood*) *transfusions* to refer to different *types of* transplantations.
- **Drugs used in leukemia treatment**: *interferon, etoposide, cyclophosphamide*.
- **Diagnosis and prognosis**: *remission, survival, relapse, diagnosis, prognostic, results, response, prognosis, analysis* and the adjectives *relapsed* and *refractory*.
- **Properties**: *leukemic, malignant, clonal, cytogenetic, toxicity, peripheral, clinical, normal*

and *median.*

By computing other key words (comparing the leukemia corpus with the oncology corpus and, in turn, the latter with the reference corpus, the BNC), the list of concepts integrated in each conceptual area was further improved, and a number of other minor subject areas were identified.

A second technique to extract conceptual information from our corpora concerns the study of *links* between key words. Links are the co-occurrences of key-words within a collocational span. This turned out to be a very productive way to identify different types of a number of conceptual relations holding between concepts. Table 2 shows a sample of the first 60 key words linked to the key word *leukemia*:

N	WORD	LINKS	= %
1	ACUTE	1.014	40,37
2	CHRONIC	703	27,99
3	LYMPHOCYTIC	477	18,99
4	MYELOGENOUS	346	13,77
5	CELL	343	13,65
6	MYELOID	285	11,35
7	CELLS	282	11,23
8	LYMPHOBLASTIC	264	10,51
9	LYMPHOMA	196	7,80
10	AML	150	5,97
11	TREATMENT	120	4,78
12	T	120	4,78
13	CLL	97	3,86
14	HAIRY	92	3,66
15	CML	91	3,62
16	THERAPY	91	3,62
17	MARROW	83	3,30
18	PROMYELOCYTIC	78	3,11
19	CHILDHOOD	74	2,95
20	MYELOMA	60	2,39
21	TREATED	59	2,35
22	DISEASE	58	2,31
23	REMISSION	53	2,11
24	BLOOD	49	1,95
25	REFRACTORY	47	1,87
26	CANCER	45	1,79
27	CHEMOTHERAPY	45	1,79
28	RELAPSED	45	1,79
29	BONE	40	1,59
30	MYELOMONOCYTIC	40	1,59

N	WORD	LINKS	= %
31	ASSOCIATED	37	1,47
32	LINEAGE	36	1,43
33	TYPE	35	1,39
34	MYELOBLASTIC	35	1,39
35	LYMPHOID	34	1,35
36	MULTIPLE	32	1,27
37	RELAPSE	31	1,23
38	TYPES	30	1,19
39	RESULTS	30	1,19
40	HODGKIN'S	30	1,19
41	DIAGNOSED	30	1,19
42	APL	29	1,15
43	ABSTRACT	29	1,15
44	CHROMOSOME	28	1,11
45	RELATED	27	1,07
46	DIAGNOSIS	25	1,00
47	MYELODYSPLASTIC	25	1,00
48	PHASE	25	1,00
49	LYMPHOCYTES	25	1,00
50	CLINICAL	24	0,96
51	GENE	24	0,96
52	NON	23	0,92
53	MALIGNANT	23	0,92
54	POSITIVE	23	0,92
55	SYNDROME	23	0,92
56	CHROMOSOME	22	0,88
57	RISK	22	0,88
58	SURVIVAL	22	0,88
59	COMPLETE	21	0,84
60	RESIDUAL	21	0,84

Table 2: Key words linked to the Key word ***leukemia***

These links pointed out different types of relevant information to characterize the concept of *leukemia*:

- **Types of leukemia**: according to the way the disease develops: *acute* and *chronic*; according to the age of the patient suffering the disease: *childhood*; according to the type of cell involved in the disease: *lymphocytic, myelogenous, myeloid, lymphoblastic, hairy, promyelocytic, myelomonocytic, myeloblastic* and *lymphoid.*
- **Response to treatment**: *relapsed, residual, refractory* and the deverbal noun *remission* (which, in turn, frequently co-occurs with key word number 59 in the nominal group *complete remission*).
- **Body parts**: *cell, cells, T, marrow, blood, bone, lineage, chromosome, lymphocytes, gene, chromosomes.*

- **Related diseases**: *lymphoma, (multiple) myeloma, disease, cancer, Hodgkin's, myelodysplastic, syndrome.*
- **Treatment and diagnosis**: *treatment, therapy, treated, chemotherapy, diagnosed, diagnosis, risk, survival.*

With the study of key words, links among key words and *associates* (key words being linked by appearing frequently, not just in a collocational span, but in the same text) we were able to identify the areas of knowledge central to our subject domain and the concepts conforming each area. However, the organization of the domain and precise characterisation of each individual concept required a more detailed study of the nature of the relations and properties that characterise each concept. In the following section we show briefly how we searched for that kind of detailed information.

2.2. Knowledge probes: semantic relations and linguistic structures

We searched the corpus to find linguistic structures pointing at *knowledge rich* contexts (in the sense of Meyer and Mackintosh 1994), that is, those parts of the texts where the authors include key information about the concepts in the domain and their relations. Some other scholars have previously investigated the usefulness of linguistic structures to locate semantic relations (see, for instance Ahmad & Fulford 1992; Kavanagh 1995; Davidson 1998). For instance, the linguistic structure *X is a Y* is usually denotes a generic/specific relation. However, in the literature revised, no indication was made of how a terminographer could access those linguistic structures to further investigate the corpus. In our approach, we decided to use the key words and links between them as *knowledge probes*, searching in the corpus for all the co-occurrences of a particular key word with a linked key word.

With this technique we looked up, for instance, all the occurrences in which the key word *leukemia* is linked to the key word *diagnose.* We widened our search with wildcards to the string **diagnos*** to extract both nominal and verbal uses. The study of the concordance lines retrieved gave us a wealth of information about the diagnostic aspects of *leukemia*; we reproduce here some of them:

The diagnosis of acute *leukemia* requires the demonstration of leukemic cells in the bone marrow, peripheral blood, or extramedullary tissues.

Smoldering AML refers to a syndrome in which the diagnostic features of acute leukemia are present, but the disease follows an indolent or subacute course.

Special histochemical stains should be performed on bone marrow specimens of all children with acute *leukemia* to confirm their diagnosis. The stains most commonly used include myeloperoxidase, PAS, Sudan Black B, and esterase.

Hairy cell *leukemia* may be difficult to diagnose early because it symptoms are vague and resemble those of other illnesses.

Bone marrow aspiration evaluates the stem cells that mature into normal blood cells. The procedure is used to diagnose *leukemia* and to check the response to treatment.

Untreated adult acute myeloid *leukemia* (AML) is defined as newly diagnosed leukemia with no prior treatment.

How is *leukemia* diagnosed? To find the cause of a person's symptoms, the doctor asks about the patient's medical history and does a physical exam.

The differential diagnosis of LGL *leukemia* should be considered in two different contexts: diseases associated with CD56 expression and those associated with reactive LGL proliferation.

Diagnosis of *leukemia* is supported by findings of the medical history and examination, and examining blood under a microscope. Leukemia cells can also be detected and further classified with a bone marrow aspiration and/or biopsy.

Correct diagnosis of acute promyelocytic *leukemia* (APL) requires proof of the translocation (15;17)(q24;q11), which appears to be absolutely specific for this particular type of myeloid disorder.

With this type of combined searches we were able allocate our concepts in their position in the knowledge structure of our subject domain, and we could complete their description by

means of properties assigned to each concept and the relations established between concepts. We needed, however, an appropriate tool that allowed us to represent and formalize in a coherent and consistent way all this information.

3. Ontology-based representation of conceptual information

OntoTerm is the knowledge-based multilingual terminology management system that we have developed to represent and manage the conceptual and terminological information of our subject domain. We have used this application to represent the knowledge of our subject domain, integrated in the Mikrokosmos ontology mentioned in the introduction. The overall architecture of the application may be said to stand on two primary modules: the Ontology Editor, where ontology construction is carried out, and the Termbase Editor, where lexical mappings and term description take place. A number of other tools facilitate browsing, navigating, and reporting the ontology. A fundamental design principle has been to isolate the user from the application's internal operations. Thus, users can concentrate on the construction of the conceptual structures relevant for their domain, rather than on how this knowledge is encoded.

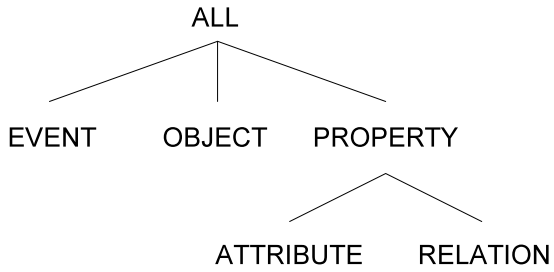
OntoTerm runs under MS Windows 9x/NT/2000, requiring the usual hardware for this platform's applications. An evaluation version is now available for download from the OntoTerm web site (http://www.ontoterm.com/).

The Ontology Editor is where conceptual modelling takes place and it can be said to be the main component of the application, as it is not possible to enter a term in the Termbase Editor without assigning it to an already existing concept in the ontology. This does not mean that an ontology must be fully developed in order to start terminological work, as the user may choose to perform both tasks concurrently, mapping terms in the Termbase Editor to concepts defined in the Ontology Editor as they are entered. This is possible because, although termbase files exist independently of ontology files, the former have information about the location of the latter and, conversely, an ontology has information about all termbases that are using it. Whenever a concept is added, deleted or modified in the Ontology Editor, all termbases assigned to this ontology are updated accordingly.

Because OntoTerm considers an ontology as a reusable resource, i.e. a repository of structured knowledge that is able to support a number of lexical resources, concepts can be marked to belong to a specific project or subset. The user can then choose to list only those concepts belonging to a specific subset.

OntoTerm has been designed to support the kind of ontology envisaged in the Mikrokosmos project. Therefore, even though users can create new ontologies, these must follow a number ontological assumptions:

- The ontology must be self-defined, i.e., each and every one of the entities making up the ontology must, in turn, be defined in it. This, in fact, is a meta-ontology which includes actions and assumptions about the ontology that determine ontology content.
- A given minimal upper model is imposed, i.e., each and every concept must fall into one of the following categories:



- Events and objects are defined by their position in the hierarchy and by their properties.
- Properties are inherited down the hierarchy unless otherwise specified (inheritance blocking or overriding).
- Relations and properties must be defined in terms of their *domain* (the set of concepts they can be used to define) and their *range* (the values they are liable to take).

This tool has proved to be flexible enough to represent the kind of knowledge that we have been able to extract using the techniques shown above. Unlike other systems, it does not rely on a limited set of primitive properties on which to build complex ones. Instead, users can choose to define their own properties that best define the specialized domain they are trying to represent. Thus, representing a given domain-specific attribute or relation of the kind discussed above, consists of finding the appropriate position in the hierarchy, i.e., its parent(s) concept(s), and delimiting its domain and range. It can then be used to describe any object or event concept in its domain.

Table 3 below summarizes the information contained in the Ontology Editor for one of the concepts of our subject domain, LEUKEMIA-OF-UNSPECIFIED-CELL-TYPE. In our conceptual hierarchy, this concept is a daughter of MALIGNANT-NEOPLASM-OF-LYMPHATIC-AND-HEMATOPOIETIC-TISSUE, which is in turn a daughter of HAVE-NEOPLASM, a type of DISEASE-EVENT. In the table below we reproduce the subclasses of this concept and the relations and attributes that connect this concept with others in the ontology. As it can be seen in this example, some of the relations and attributes are assigned locally, that is, directly to the concept, whereas others are inherited from the information already assign to other concept higher in the hierarchy. It is also important to highlight that both relations and attributes are concepts themselves, which contain specific information about their *domain,* that is, the set of concepts they can define and their *range*, their possible values.

The kind of ontology that results from such modelling work may become fairly complex. In order to facilitate revision tasks, OntoTerm provides a number of tools to navigate, edit, and report the information contained in its ontologies, including an on-line HTML report generator that can be used to publish the whole ontology.

ISA	MALIGNANT-NEOPLASM-OF-LYMPHATIC-AND-HEMATOPOIETIC-TISSUE
SUBCLASSES	ACUTE-LEUKEMIA-NOS CHRONIC-LEUKEMIA-NOS SUBACUTE-LEUKEMIA-NOS
RELATIONS	**AFFECTS-PHYSIOLOGICAL-SYSTEM**: LYMPHATIC-SYSTEM BONE-MARROW **DIAGNOSED-WITH** : BLOOD-TEST; BONE-MARROW-BIOPSY BONE-MARROW-ASPIRATION **HAS-RISK-FACTOR** : NUCLEAR-ENERGY ANTICANCER-DRUG BENZENE **HAS-SYMPTOM** : SWEAT INFECTION BECOME-TIRED FLU
INHERITED RELATIONS	**AFFECTS-BODY-PART** : PHYSIOLOGICAL-SYSTEM (from DISEASE-EVENT) TISSUE (from DISEASE-EVENT) **AFFECTS-TISSUE** : TISSUE (from HAVE-NEOPLASM) **CAUSED-BY :** CARCINOGEN (from HAVE-NEOPLASM) **DIAGNOSED-WITH-THE-USE-OF** : DIAGNOSTIC-DEVICE (from HAVE-NEOPLASM) **EXPERIENCER** : PATIENT (from DISEASE-EVENT) **LOCATION** : TISSUE (from HAVE-NEOPLASM) PHYSIOLOGICAL-SYSTEM (from HAVE-NEOPLASM) **SIDE-EFFECT-OF** : DRUG (from DISEASE-EVENT) **SYMPTOM-OF** : HAVE-NEOPLASM (from DISEASE-EVENT) **THEME** : ANIMAL (from DISEASE-EVENT) **TREATED-WITH** : MEDICAL-MATERIAL (from HAVE-NEOPLASM) MEDICAL-ARTIFACT (from HAVE-NEOPLASM) **TREATED-WITH-THE-USE-OF** : DRUG (by default) (from DISEASE-EVENT)
INHERITED ATTRIBUTES	**DIRECTION-OF-CHANGE :** NEGATIVE (from CHANGE-STATE) POSITIVE (from CHANGE-STATE) **TYPE-OF-CHANGE :** PHYSICAL (from DISEASE-EVENT) MENTAL (from DISEASE-EVENT)

Table 3: The concept LEUKEMIA-OF- UNSPECIFIED-CELL-TYPE *in the Ontology Editor.*

4. Conclusion

In this paper we have detailed some of the techniques used in our research to extract conceptual information from special-purpose corpora, as well as the tools we employ to represent it. In the extraction process, we have adopted a semi-automatic approach that uses a very complete and powerful software tool used in corpus linguistics and corpus lexicography. Although our current efforts aim at fully automating the knowledge extraction process, the use of such a tool has the advantage of being commercially available and extremely user-friendly. With regards to OntoTerm, the knowledge-based terminology management system we have used to represent conceptual information, it seems clear that its design and facility of use make it possible the integration of knowledge representation techniques in the terminology management process, with a minimum training on the part of the terminographer.

References

AHMAD, K. & H. FULFORD (1992). *Knowledge Processing 4. Semantic Relations and their use in the elaboration of Terminology*. Computing Sciences Report. University of Surrey.

AHMAD, K., A. DAVIES, H. FULFORD & M. ROGERS (1994). "The Elaboration of Special Language Terms; the Role of Contextual Examples, Representative Samples and Normative Requirements", in *Euralex'92 Proceedings I.* Tampere: Studia Translatologica: 139-150.

BOWKER, L. (1996). "Towards a corpus-based approach to terminography". *Terminology*. Vol. 3(1): 27-52. Amsterdam: John Benjamins Publishing Company.

CABRÉ, M. T. (1999b). *La terminología: representación y comunicación: elementos para una teoría de base comunicativa y otros artículos.* Universitat Pompeu Fabra. Institut Universitari de Lingüística Aplicada.

DAVIDSON, L. (1998). *Knowledge Extraction Technology for Terminology*. MA Thesis, Translation Department. University of Ottawa.

GUARINO, N. (1998a). "Formal Ontologies and Information Systems", in N. Guarino (ed.) *Formal Ontologies and Information Systems*. Amsterdam: IOS Press, pp: 3-15.

MAHESH, K. (1996). *Ontology Development for Machine Translation: Ideology and Methodology.* NMSU. Computing Research Laboratory. Technical Report MCCS-96-292. New Mexico.

MAHESH, K. & S. NIRENBURG (1995). A Situated Ontology for Practical NLP. In Proceedings of the Workshop on Basic Ontological Issues in Knowledge Sharing. International Joint Conference on Artificial Intelligence (IJCAI-95), August 1995. Montreal, Canada.

MEYER, I. & K. MACKINTOSH (1996a). "The Corpus from a Terminographer's Viewpoint". *International Journal of Corpus Linguistics*, Vol. 1 (2), pp: 257-285.

MEYER, I., K. ECK & D. SKUCE (1997). "Systematic Concept Analysis within a Knowledge-Based Approach to Terminology", in S. E. Wright & G. Budin (eds.) *Handbook of Terminology Management*. Vol. 1 Amsterdam/Philadelphia: John Benjamins, pp: 98-118.

MORENO ORTIZ, A. (2000a) "Managing Conceptual and Terminological Information in a User-friendly Environment". Proceedings of OntoLex 2000. Workshop on Ontologies and Lexical Knowledge Bases.

MORENO ORTIZ, A. (2000b) "OntoTerm: un sistema abierto de representación conceptual". Proceedings of the XVI SEPLN Conference (Spanish Society for NLP).

MORENO ORTIZ & C. PÉREZ HERNÁNDEZ (2000). "Reusing the Mikrokosmos Ontology for Concept-Based Multilingual Terminology Databases". Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC 2000 Athens, Greece. 31 May -2 June), pp: 1061-1067.

PEARSON, J. (1998). *Terms in Context.* Studies in Corpus Linguistics Vol. 1. Amsterdam/Philadelphia: John Benjamins.

PÉREZ HERNÁNDEZ, C. (2000). *Explotación de los Corpora Textuales Informatizados para la Creación de Bases de Datos Terminológicas.* PhD Thesis. University of Málaga.

SINCLAIR, J. M. (1992). "The Automatic Analysis of Corpora", en J. Svartvik (ed.) *Directions in Corpus Linguistics.* Proceedings of Nobel Symposium 82, Stockholm, 4-8 August 1991. Berlin/New York: Mouton de Gruyter, pp: 379-398.