INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

UMI

A Bell & Howell Information Company
300 North Zeeb Road, Ann Arbor MI 48106-1346 USA
313/761-4700 800/521-0600

Université d'Ottawa · University of Ottawa

KNOWLEDGE EXTRACTION TECHNOLOGY FOR TERMINOLOGY

by

Laura Davidson

School of Translation and Interpretation
University of Ottawa

under the supervision of

Ingrid Meyer, Ph.D
School of Translation and Interpretation
and
Douglas Skuce, Ph.D.
School of Information Technology and Engineering

Thesis submitted to
the School of Graduate Studies and Research
of the University of Ottawa
in partial fulfillment of the requirements
for the degree of M.A. (Translation)

Acknowledgements

I would first like to express my thanks to my thesis director, Ingrid Meyer, for her constructive guidance throughout the entire thesis process. I especially appreciated the timeliness with which she provided her feedback.

Many thanks also to my thesis co-director, Doug Skuce, for his valuable input from a computer science perspective and for providing me with access to the Artificial Intelligence lab at the School of Information Technology and Engineering, where I performed most of my research.

I also wish to thank Judy Kavanagh, the programmer of IKARUS and the Text Analyzer, for her patient and expert help during my research. Of all the people involved with my thesis, Judy had to spend the most time with me—a commendable feat in itself! :-) And thanks for all the cat stories too, Judy!

I wish to thank David Miller for providing me with the solution to a particular computer problem and, in effect, saving my thesis. The solution was elegantly simple, but the problem was major in that it brought half my thesis research to a screeching halt.

It is also a pleasure to thank Professor Sylvie Lambert for revising the French translation of my abstract and for her promptness in doing so.

And to my Mother, Maureen, goes a most sincere thanks for all her moral support, encouragement, and infinite supply of *warm fuzzies* throughout my entire university adventure, including my undergrad years. Had I thrown in the towel at any point during that time, I would never have had the wonderful opportunity to show her this MA thesis and say "Look what I made at school today, Mom!"

Abstract

Terminologists scan large amounts of specialized texts to discover the terms for the concepts in a given subject field and to extract knowledge-rich contexts. These contexts make explicit, by means of linguistic structures, the semantic relations that exist between the concepts. Developing the subject field's conceptual network is called *concept analysis*. To carry out concept analysis, many terminologists are still using paper-based corpora. Yet this is time-consuming and error-prone.

This thesis explores a semi-automatic approach to concept analysis that involves electronic corpora and knowledge extraction technology. My research focussed on a program called the Text Analyzer (TA), which I tested for its effectiveness in retrieving knowledge-rich contexts from French and English electronic corpora in the subject field of composting.

I first discovered the linguistic patterns that French and English use to express three semantic relations. The TA was then programmed with these patterns to be able to extract knowledge-rich contexts from the corpora. I then tested the TA's extraction capabilities and prepared statistics showing its effectiveness. Analysing the test results revealed ways to enhance the TA's performance. As a small follow-up experiment, I added more patterns to the TA and again tested its extraction effectiveness, which was improved. Up to that point, my focus was on *lexical* patterns. As part of the follow-up experiment, I also performed an exploratory test of the potential of *grammatical* patterns for knowledge extraction.

This research revealed that much work is still needed to produce highly effective knowledge extraction programs. Even so, the statistics were encouraging and showed this technology's potential for dramatically reducing the time terminologists spend scanning corpora.

Résumé

Le terminologue effectue le dépouillement d'une quantité énorme de documentation spécialisée pour repérer des termes désignant les concepts et pour acquérir des contextes riches en connaissances. Ces contextes montrent, par le biais des structures linguistiques, les relations sémantiques (c'est-à-dire notionnelles) qui relient les notions. Le terme *analyse conceptuelle* désigne l'activité d'établir le réseau notionnel du domaine en question. Pour effectuer cette analyse, bon nombre de terminologues se servent toujours de corpus imprimés. Cette méthode, pourtant, prend du temps et est susceptible d'erreurs.

La présente thèse explore l'utilité de la technologie qui extrait des connaissances des corpus électroniques pour l'automatisation partielle de l'analyse conceptuelle. J'ai concentré mes recherches sur un logiciel connu sous le nom de *Text Analyzer* (Analyseur de textes). J'ai soumis cet outil à des essais destinés à vérifier son efficacité pour l'extraction de contextes riches en connaissances, présents dans un corpus anglais et dans un corpus français, les deux traitant du domaine du compostage.

J'ai commencé par découvrir l'éventail de structures linguistiques employées par l'anglais et le français pour exprimer trois relations notionnelles. J'ai ensuite ajouté ces structures au code de l'Analyseur de textes pour extraire des corpus des contextes riches en connaissances. Après avoir fait tourner l'Analyseur de textes, j'ai dressé des statistiques pour montrer son efficacité. Une analyse des résultats des essais a laissé voir plusieurs façons d'améliorer l'efficacité de l'outil. Dans le cadre d'une étude complémentaire, j'ai ajouté à l'Analyseur de textes des structures linguistiques supplémentaires pour ensuite vérifier à nouveau son efficacité d'extraction, cette

dernière se voyant bel et bien améliorée. Jusqu'ici, j'ai concentré mes efforts sur les structures *lexicales*. Dans le cadre de l'étude complémentaire, j'ai aussi testé quelques structures *grammaticales* pour leur efficacité d'extraction des contextes riches en connaissances.

Ces recherches indiquent qu'il reste beaucoup à faire pour pouvoir créer des logiciels capables d'extraire des connaissances à un très haut degré. Les résultats de mes essais sont tout de même prometteurs et valident ainsi l'utilité de cette technologie, l'avantage principal étant sa capacité de réduire le temps que les terminologues consacrent à dépouiller des corpus.

Table of Contents

List of Figures

Introduction

> In earlier times [i.e. before the printing press], to possess an idea or a fact meant keeping it a secret, having the power to prevent others from knowing it....the first postal services were designed for the security of the state. Physicians and lawyers locked their knowledge in a learned language. The government helped craft guilds exclude trespassers from their secrets. But the printing press made it harder than ever to keep a secret. (Boorstin: 1983, 409)

The existence of the Internet is evidence that humans have made an about-face since the times before the printing press: rather than trying to "hoard" our knowledge, we are striving to keep up with the increasing demand for knowledge and cannot seem to pump it out fast enough. Moreover, most civilizations, rather than being largely illiterate and uninformed, are now suffering instead from information overload.

Terminology is a field that endeavours to collect, process, and disseminate in an organized, intelligible fashion the lexical items of specialized fields of knowledge. Language specialists such as translators clamour for bilingual terminologies covering the domains they work in. Computers and telecommunications have increased the speed at which translators (among other professionals) are expected to produce. Terminology must keep up with this growing demand brought on by computers and by the phenomenon known as globalization, creating in turn what Marshall McLuhan back in the 1960's called "the global village."

For the field of terminology to compete in this modern "sweatshop" of knowledge production, supply and demand, the traditional paper-based methods used in terminology work must be abandoned in favour of time-saving computerized methods. This thesis explores the potential of knowledge extraction technology for terminology.

One of the main things that terminologists look for in a corpus (i.e. a body of text) are *knowledge-rich contexts*. These contexts have two functions. First, they help terminologists understand the concept designated by a term and the ways in which the concepts in that field relate to one another. Second, they serve as textual support on term records. Knowledge extraction programs could semi-automate this time-consuming task. In other words, a computer program, rather than terminologists themselves, could be set to analyse millions of words and retrieve the knowledge-rich contexts.

The essence of these contexts are the *semantic relations* which hold between the concepts. Each semantic relation can be expressed by a variety of linguistic structures consisting of one or more words that indicate to readers what relation is being dealt with at that moment. For instance, in the previous sentence, I used *consisting of* to introduce the *parts* of a linguistic structure, namely "one or more words". The structure *consisting of* and its inflectional variations (such as *consists of*, *consist of*, etc.) stand between two slots, X and Y, which in turn can be filled with any grammatically and logically suitable words that denote a whole and its parts.

This "part of" relation is also called *meronymy*. The two other relations that this thesis deals with are *hyponymy* (the generic-specific relation) and *functionality*. My research involved first *discovering* various linguistic structures that express these relations, and then *testing* their usefulness in a computer program for extracting knowledge-rich contexts automatically.

OBJECTIVES

The purpose of this thesis was to explore the potential that knowledge extraction technology has for terminology work. Terminologists are concerned with concept analysis (i.e. discovering the network of relations between the concepts in subject fields). Concept analysis enables terminologists to understand the subject field and therefore be able to prepare adequate term records, that important vehicle of knowledge transmission. Being able to "delegate" to a computer the search and retrieval of necessary semantic (i.e. conceptual) information from a multi-million word corpus would greatly facilitate their work.

To carry out this exploration, I used (and contributed to the development of) the Text Analyzer (TA), one component of an experimental computer program called IKARUS, which stands for Intelligent Knowledge Acquisition and Retrieval Universal System (Skuce: 1996b). This program is under development in the Artificial Intelligence lab[1] at the University of Ottawa's School of Information Technology and Engineering.

The "Conceptual Operations" function of the TA allows users to extract (from an electronic corpus treating a given subject field) sentences that express the semantic relations that hold between the concepts in that field. The TA searches in the corpus for sentences that contain the search word (a term entered by the user) and any one of a list of previously programmed linguistic structures/patterns that express semantic relations. Before I began the work for this thesis, the TA was equipped with only five English patterns for each of the following relations:

[1] The Artificial Intelligence lab has recently been renamed LAKE (Language Analysis for Knowledge Engineering.

hyponymy, meronymy, functionality, and synonymy. For the first three relations, my goal was to augment the number of English patterns and provide French ones, thereby *enhancing* the TA's knowledge extraction capabilities for English and *enabling* it to do the same for French.

METHODOLOGY

Before starting the practical work for this thesis, I first had to do some learning. My four initial learning tasks were as follows. It was necessary for me to:

- gain a deeper understanding of terminology work
- gain a deeper understanding of semantic relations
- learn how to use all functions of the TA
- learn a satisfactory amount of the UNIX operating system, since the TA is UNIX-based

The practical work required me to:

- choose a subject field and build French and English electronic corpora for that field
- discover the terms in that subject field
- discover linguistic patterns in French and English that express hyponymy, meronymy, and functionality
- work with the programmer to equip the TA's Conceptual Operations function with the patterns and then do a pre-test to discover and solve any "bugs"
- perform the final testing and prepare effectiveness statistics
- add to the TA further patterns discovered during analysis of the final test results
- perform a "follow-up" test of the TA after enhancements (i.e. further pattern additions), prepare new effectiveness statistics and compare them to the previous ones to determine if enhancements brought about improvements.
- perform a brief exploration into knowledge extraction using *grammatical* patterns

ORGANIZATION OF THE THESIS

Chapter 1 discusses the field of terminology. It begins with a look at traditional terminology work and explains why there is a need for change. The chapter then investigates the benefits of using computers in terminology work and the concept of taking a "knowledge engineering" approach to terminology. This chapter draws on the work of researchers in terminology, linguistics, and computer science.

Chapter 2 explores three semantic relations in depth (hyponymy, meronymy, and functionality) and makes clear their importance for terminology. The chapter draws on studies by semanticists, linguists, and computer scientists of the linguistic patterns that express semantic relations. Following that is a discussion of how these patterns can be used for computerized knowledge extraction and why this technology is relevant to terminology. Included in this chapter is a section explaining how my study differs from previous research.

Chapter 3 describes in detail the practical work of choosing a field, building the corpora, determining the terms, discovering linguistic patterns that express semantic relations, programming the TA, testing the knowledge extraction tools, preparing effectiveness statistics, and suggesting ways to enhance the TA even further.

Chapter 4 is a "follow-up" to Chapter 3. The TA was enhanced with more patterns, discovered during the previous testing. A modest test was performed to determine if the enhancements brought about improvements in the TA's extraction effectiveness. Finally, a brief exploratory experiment was performed using *grammatical*, as opposed to *lexical*, patterns.

Chapter 5 provides a conclusion and avenues for future research.

Chapter 1 NEW STRATEGIES FOR TERMINOLOGY

1.0 INTRODUCTION

We are witnessing a number of knowledge explosions: increased knowledge in established fields and the birth of entirely new fields. Terminologists must keep on top of the linguistic developments resulting from knowledge explosions, but traditional terminology methods are often too slow for the speed at which terminologists are now expected to produce. This chapter looks at terminology work and explains why knowledge extraction technology has great potential for this field.

1.1 NEW STRATEGIES FOR TERMINOLOGY

Many terminologists are still using paper-based documentation in their work, whether for background reading or for scanning and concept analysis. In fact, "it is still common to find new, high-quality terminological dictionaries that are compiled largely from a paper-based corpus" (Meyer and Mackintosh: 1996, 2). This documentation takes such forms as encyclopedias, textbooks, specialized dictionaries, standards, monographs and journals. The use of paper documentation is slow work. Since terminologists have begun the migration to electronic documents (corpora), corpus analysis tools that are designed specifically for terminology work are needed.

Although terminologists are using computers for knowledge representation (i.e. term banks), they rarely take full advantage of the potential of the computer as a tool to help them in the process of terminology. In addition, while they prepare "concept trees" for their own use, they

normally do not provide the concept trees to users. As a result, users of traditional dictionaries or term banks cannot effectively gain as broad an understanding of the field as the terminologists do because the former cannot see at a glance where a concept fits into the system of concepts. Definitions "ne donnent qu'un aperçu fragmentaire du domaine" (Meyer and McHaffie: 1994, 431). This fragmentation should be avoided because terms do not exist in isolation. As Lyons (1968, 443) says, "the sense of a lexical item may be defined to be, not only dependent on, but identical with, the set of relations which hold between the item in question and other items in the same lexical system." In addition, tight time constraints can limit terminologists' understanding of the subject field as they are working because they may not be able to read through all the paper documentation in the time allotted. In short, traditional methods paired with tight deadlines may lead both terminologists and users of terminological works to deal with concepts in isolation, resulting in a limited understanding of the subject field.

Fortunately, with computer technology improving by leaps and bounds, and with the advent of the World Wide Web (WWW), terminologists can now build electronic corpora more easily than in the past. Electronic texts can be quickly downloaded from the WWW. As well, computer systems being developed in the field of knowledge engineering, a sub-field of Artificial Intelligence, can be useful to terminologists: corpus analysis tools (such as the TA) can be tailored to terminology work. Knowledge bases are another means of providing a more structured and logical way of storing and using terminological information, and knowledge management systems allow for effective ways of handling large amounts of information. This thesis, however, will look only at knowledge extraction technology for corpus analysis.

It should be noted here that the possibility of incorporating computerized methods into terminology work is not simply a novelty. The current reality of downsizing, rationalization and funding cutbacks places "on the chopping block" many activities in businesses and institutions that cannot be financially justified. Otman (1989, 66) warns that "certaines tâches du terminographe peuvent—et doivent---être...automatisées, si l'on veut rationaliser et rentabiliser l'activité terminologique." Therefore, terminology departments will have to start integrating computers as much as possible into *each* stage of the terminology process in order to increase productivity (without sacrificing quality) in order to reduce their draw on company/departmental funds. If not, terminology units risk being severely handicapped or cut out entirely. It can be inferred from this that terminology work is seen by many as a luxury, not as a necessity. The opposite is true, however, because "internationalization has...become a fact of life" (Picht and Draskau:1985, 24), bringing with it the crucial need for terminological standardization. Terms, as Sager (1994a, 7) says, are the means of knowledge transfer. According to Sager (1990, 2), terminology "is vital to the functioning of all sciences."

Therefore, "if progress [is] not to be neutralized by stagnation" (Picht and Draskau: 1985, 24), indeed if communication is to be successful among countries, cultures, even between specialists in the same country and subject field, then the importance of terminology must be recognized. The onus is nevertheless on terminologists to take advantage of the available ideas and technology to make their work more effective and efficient.

1.2 MODERN TOOLS AND METHODS FOR TERMINOLOGY

1.2.1 Computers and terminology

Relatively recently, the potential for computers as a useful tool for terminology work has been explored. "L'outil informatique est devenu un levier indispensable de l'activité terminologique" (Otman: 1989, 63). Otman explains that "Le terme *terminotique* est maintenant communément accepté pour dénommer ce mariage entre la 'carpe informatique' et le 'lapin terminologique'" (1989, 64). This is part of a relatively recent phenomenon that Sager (1994b, xvii) describes as follows:

> Early approaches to language were descriptive....the natural science approach led to the fragmentation of linguistics into its many applied branches of sociolinguistics, ethnolinguistics, psycholinguistics, lexicography, terminology, language acquisition etc, which have laid the foundations for an engineering view of language.

Because of this, we now see papers such as "Applying Knowledge-Engineering Technology to Terminology" (Meyer 1991). Meyer compares the tasks of the terminologist and those of the knowledge engineer and explains that, because of the similarity, terminologists can benefit from technological developments in knowledge engineering.

Terminologists must discover the concepts in a subject field and the semantic relations that hold between them. Computer tools that aid in this process can only benefit terminologists by semi-automating this task. Hence the value of research into the linguistic structures that express semantic relations: a computerized corpus analysis tool can be programmed with these structures, which then act as probes, retrieving the knowledge-rich contexts where the structures occur in a

given corpus. Terminologists can then build up the conceptual network using the extracted contexts, showing how the concepts in the field are interrelated. There are potential problems with using these linguistic structures for knowledge extraction, however, and will be discussed in Chapter 3.

1.2.2 Corpora and corpus-analysis tools

1.2.2.1 What is a corpus?

A corpus, to use Atkins *et al*'s (1992, 1) modified definition appearing in Meyer and Mackintosh (1996, 4), "is a collection of electronic texts...built according to explicit design criteria for a specific purpose." To this, I also add the following from Sinclair's (1991, 171) definition: "a collection of naturally-occurring language text, chosen to characterize a state or variety of a language."

A corpus will be built based on the specific needs of the area of inquiry that will use it. However, several things can said about corpora in general. According to McEnery and Wilson (1996, 22), "In building a corpus of a language variety, we are interested in a sample which is maximally representative of the variety under examination, that is, which provides us with as accurate a picture as possible of the tendencies of that variety, including their proportions." It is important that documents be obtained from a wide range of sources in order to attenuate the potential of skewing, i.e. to minimize the influence of a single style of writing or mode of expression. For a similar reason, "...a corpus needs to contain many millions of words" (Sinclair: 1991, 19). A corpus can be either fixed in length or constantly added to. The latter is known as a

monitor corpus and is "...primarily of importance in lexicographic work..." (McEnery and Wilson: 1996, 22). As well, a corpus can be composed of either spoken or written text or a combination of both. Regarding period, Sinclair (1991, 18) states that "Most corpora attempt to cover a particular period of time..." Another consideration is whether the corpus will hold samples of documents or whole documents. However, with computer storage capability increasing as it is, building a corpus with whole documents is not the problem it once was.

Terminological corpora will naturally differ from lexicographic ones because the users in both cases have different needs. (To prepare the information in the present paragraph, I have drawn on Meyer and Mackintosh 1996, pp 9 to 14). Since terminologists need to "acquire both linguistic and conceptual information, the corpus needs to be as linguistically and conceptually *rich* as possible...." Because terminologists deal with specialized fields, "domain experts should play a role" in building the corpus; lexicographers, however, generally do not require outside help. A terminological corpus can be smaller than a lexicographic one, primarily because there is not the time in a terminology project to build one of lexicographic size. "The *individual texts* in the [terminological] corpus must be complete, which is not always the practice in lexicography." Terminologists require whole texts because of their need to gain subject-field knowledge. Unlike lexicographers, terminologists must "delimit the domain that the corpus is to represent." This is difficult because a) boundaries between fields are not clear-cut, and b) a decision must first be made on how many general concepts to include in the finished terminological work. A terminological corpus should be composed of a variety of genres and authors from instructional, advanced and popularized texts on the given subject field; in this way, "differing degrees of

technicality" will be accounted for in the corpus. Regarding the age of texts, "terminographers are primarily interested in very current texts" because of their focus on new terms and concepts. Texts to be included should generally be original, not translations, and the majority should be written by native speakers of the language(s) under study.

1.2.2.2 Early attitudes towards corpora

The corpus approach to language study dates back to the 1800's. Before Chomsky, corpora had been used for studying language acquisition and spelling conventions, for language pedagogy, comparative linguistics, and syntax and semantics (McEnery and Wilson 1996, 2-3). However, "Chomsky...changed the direction of linguistics away from empiricism and towards rationalism in a remarkably short period of time" (McEnery and Wilson 1996, 4). The difference between rationalism and empiricism is the following: rationalism is "based on artificial behavioural data, and conscious introspective judgements" whereas empiricism is based on "the observation of naturally occurring data" (McEnery and Wilson 1996, 4). "Chomsky suggested that the corpus could never be a useful tool for the linguist, as the linguist must seek to model language competence rather than performance" (McEnery and Wilson 1996, 5).

The interesting thing to realize is that, even if Chomsky's early view of corpus-based research still held sway today, it would nevertheless be invalid for the kind of research done in this thesis. Terminological research must be performed on naturally occurring data (i.e. real-life corpora comprised of text books, articles, etc.) because the terminologist has to discover the current state of the knowledge in a given subject field as expressed in writing by subject field

experts. Introspective judgments, therefore, have little place in terminology. Furthermore, for a computer program to extract knowledge from such a corpus, it must be equipped with linguistic patterns (that express the relations between concepts in the field) that experts do indeed use, as opposed to cognitively plausible patterns dreamed up by a linguist. The latter patterns would be useless for extracting knowledge from a real-language corpus if they are never used by the experts who wrote the texts making up the terminological corpus.

1.2.2.3 Corpus use

According to Sinclair (1991, 14), "More and more people in every branch of information science are coming to realize that a corpus as the sample of the living language, accessed by sophisticated computers, opens new horizons." In the positive words of McEnery and Wilson (1996, 2), "A corpus-based approach can be taken to many aspects of linguistic inquiry."

In spite of the benefits of an electronic corpus-based approach, the field of terminology, as an area of linguistic inquiry, has been slow to follow the example set by lexicography. As Rundell and Stock wrote (1992c, 51), "...the sheer wealth of data to which lexicographers now have access cannot fail to revolutionize the dictionary-making process." The same thing can be said for the introduction of the corpus-based approach into the field of terminology.

The beauty of the corpus-based approach is its verifiability. On this point, McEnery and Wilson (1996, 13) state the following: "Leech argues that the corpus is a more powerful methodology [than introspection] from the point of view of the scientific method, as it is open to objective verification of results." At the end of that sentence, the authors added a footnote, which

I would also like to provide here: "One is irresistibly drawn to remembering Galton's[2] words, 'until the phenomena of any branch of knowledge have been submitted to measurement and number, it cannot assume the dignity of a science." As Sagan (1996, 25) states, "Science is more than a body of knowledge; it is a way of thinking." When experiments can be replicated and results verified, one's endeavours are a great deal more "dignified" and credible.

1.2.2.4 Corpus analysis tools

"Can you imagine searching through an 11-million-word corpus...with nothing more than your eyes? The whole undertaking becomes prohibitively time-consuming. It also becomes very expensive and error prone" (McEnery and Wilson: 1996, 10). A concern in early corpus linguistics was how to process the data. Before computers, researchers naturally *had* to rely on humans, with all the attendant problems. Modern computers and corpus analysis tools, however, make the concern about handling large corpora a non-issue. In fact, the advent of the computer has made "the term *corpus*...now almost synonymous with the term **machine-readable** corpus" (McEnery and Wilson: 1996, 14). Sinclair (1991, 27), as well, recognizes the value of computers in the processing of natural language texts: "...the quality of linguistic evidence is going to be improved out of all recognition, because of the power of the computer in data management".

How a given field of inquiry processes the corpus will naturally be dictated by the needs of that field. In turn, the corpus analysis tools to aid in the processing will be designed to accomplish specific tasks related to those needs. In general, there are two approaches to corpus analysis:

---

[2] Sir Francis Galton (1822-1911) was an English scientist and a cousin of Charles Darwin.

quantitative and qualitative. McEnery and Wilson (1996, 62) explain the difference:

> Whereas in quantitative research we classify features, count them, and even construct more complex statistical models in an attempt to explain what is observed, in qualitative research the data are used only as a basis for identifying and describing aspects of usage in the language and to provide 'real-life' examples of particular phenomena.

In short, quantitative analysis is statistical, and qualitative analysis is identification and description.

There are different ways of preparing a corpus for use with corpus analysis tools. A corpus can be either unannotated (in its raw, "naturally occurring" state) or annotated (having extra information added to it by researchers). Both forms of the corpus are used differently. On an unannotated corpus, frequency analyses and concordances can be performed. In frequency analysis, any or all word-forms can be counted and then the output sorted alphabetically, by frequency, or by first occurrence. A concordance program can retrieve all instances of a given search word (neatly aligned in a column) along with a certain amount of context for each instance. Many allow a KWIC (key word in context) concordance of several words on either side of a search word; others offer, as well, the option of longer contexts. Different concordancers allow for various kinds of output sorting. In general, the function of a concordancer is to allow researchers to detect patterns of language use, which would be virtually impossible to do if they were to simply read the documents contained in the corpus.

There are a number of ways to annotate a corpus, such as part-of-speech tagging, parsing, and semantic annotation. McEnery and Wilson (1996, 24) claim that "...the utility of the corpus is considerably increased by the provision of annotation." However, I posit that it is the needs of the

particular researchers that will determine whether to annotate the corpus, and if annotated, which types of annotation will be more valuable.

The ultimate purpose of electronic corpora and the tools necessary to analyse them is to semi-automate (ideally, to automate completely) the work in a given area of linguistic inquiry. And in our increasingly fast-paced world, time is of the essence. To conclude this section, I cite Rundell and Stock (1992a, 14):

> As computers perform more and more of the routine work *and* make possible increasingly fine-grained analyses of the language, lexicographers [and of course terminologists] will simultaneously be liberated from drudgery and empowered to focus their creative energies on doing what machines cannot do....And the proper function of language experts...is to *interpret* that evidence [from corpora], to select and synthesize what is significant and appropriate, and so to mediate between the corpus and the end-user of the materials they produce."

1.2.3 The World Wide Web as a source of electronic text

Terminologists need documentation for two reasons: to find terms and to find contexts that indicate what the terms mean and how they behave (e.g. collocations, morphology.) These tasks in the context of traditional paper-based terminology work are very time-consuming. Terminologists must physically find and retrieve the sources they need, either by walking or driving to where the documentation is located. Having access to the World Wide Web (WWW), however, allows terminologists to sit in front of a computer screen and find documentation in electronic form, obviating the need to leave the office. The more time that passes, the greater the range and quantity of information that will become available on the WWW.

"With a magnitude measured in millions of pages, the World Wide Web is not an easy medium to master....Sifting through this much material...clearly calls for automated search tools of some sort" (Venditto: 1996, 79). The tools used to find the desired Web information are called search engines. A wide range of these tools are currently available to researchers and allow various types of searches on any electronic text. Once researchers learn how to use these tools effectively, the research process can be greatly facilitated, in comparison to traditional methods.

1.2.3.1 Quality of information on the WWW

No discussion of a documentation resource would be complete without a look at the quality of what researchers can find on the Web. This is especially important for terminology work because the documentation selection stage, as Cole (1987, 79) says, is a crucial step in the terminological research project: the choice of documentation directly determines the research results (and ultimately the quality of the works produced).

At present, the Internet is an "open system"; input can come from anywhere. And so it is with the WWW, being a part of the Internet. Anyone, anywhere can publish anything on the Web. A search for information on a given topic will yield a collection of documents of varying accuracy and validity. As Smith (1996) points out, "the Net contains a great deal of information, but much less knowledge; a great deal of noise, but little signal." As a result, Smith says that researchers "waste a great deal of time sifting the good from the bad."

This waste of time is not the only disadvantage, though. Smith talks about the "principle of least effort" as put forth by Thomas Mann, General Reference Librarian at the US Library of

Congress. According to this principle, most researchers tend to be satisfied with the most easily available sources, regardless of quality. This is a hazard, then, for Web research, owing to the quantity of information of dubious quality and the temptation of selecting the first pages presented in a Web search. As a result of all this, Smith says that "the near instant accessibility of on-line material---ostensibly one of the Internet's greatest strengths---suddenly becomes a weakness."

However, this is only the current state of the Internet, and changes occur rapidly in this new field. Unfortunately, though, most prognostications do not discuss the future *quality* of information; discussion is limited to the fact that more and more information is being made available, in more new and dazzling ways, including enhanced multimedia. Obviously, increased quantity and multimedia are a boon to researchers, but articles dealing with these phenomena do not address the *quality* issue. One potential solution to the problem of accessing large quantities of Web junk is electronic, or virtual, libraries. Librarians and researchers can work together to select high-quality sources and make them available all in one place on the Web.

1.2.4 Knowledge engineering

1.2.4.1 What is knowledge engineering?

Knowledge engineering involves acquiring, formalizing and refining knowledge so that it can be used by machines or people. Knowledge engineers are concerned with building knowledge bases (specialized, easily navigable storehouses of information) on specialized topics. Their goal is to organize and present knowledge in such a way that the knowledge system responds to the needs of the user (whether human or machine) and allows a "connected" understanding of the

topic rather than a fragmented understanding of pieces of knowledge in isolation. This also implies logical, intelligent navigation through information.

World knowledge is increasing at an astronomical rate. As Picht and Draskau (1985, 24) say,

> The 19th century was remarkable for the giant strides with which scientific progress advanced and found practical applications. This situation led to a vast need for terminology, and it soon came to be realized that these explosive developments likewise called for the organization of knowledge.

This is still the case today. Not only is knowledge itself continually increasing, but also the amount of knowledge being made *available* is increasing, and the WWW is adding to that phenomenon. Skuce (1996c) says,

> today's systems for storing and sharing knowledge on a large scale are undergoing rapid change, driven mainly by the success of the World Wide Web and its search engines. Soon a vast array of material will become available....We seek better ways of dealing with this impending revolution.

Information is virtually useless if it cannot be a) accessed and b) accessed efficiently. Skuce (1996b) says that "most on-line knowledge is currently kept in files that are hard to structure, classify, browse and find....We need systems to improve management of this kind of knowledge."

The School of Information Technology and Engineering at the University of Ottawa has for several years been developing and refining a knowledge management system (kms) in preparation for the above-mentioned information revolution. This kms was originally born with the name CODE (Conceptually Oriented Design Environment) and has been called a "knowledge processor". This system was tested (for its potential as a tool for terminology) in a pilot project at the Terminology and Linguistic Services Directorate at the Department of the Secretary of State of Canada.

The program was refined, simplified and "rewritten" so that it runs in Netscape on the WWW. It has also been renamed IKARUS (Intelligent Knowledge Acquisition and Retrieval Universal System). One component of IKARUS is the Text Analyzer (TA), a corpus analysis tool (Kavanagh, 1995). The TA was used in much of the research for this thesis, and a description of this tool is provided in Section 1.2.5.1.

1.2.4.2 Linking knowledge engineering with terminology

As Meyer and Paradis (1991, 3) say, "knowledge engineering implies a focus on specialized concepts." The three main activities are the acquiring, formalizing and refining of knowledge. Knowledge engineers glean information from texts and experts, turn that information into a useable format for its intended user, and finally, continually refine (correct) and update the formalized information when subject field knowledge changes or deepens.

The main tasks of a terminologist mesh almost perfectly with those of a knowledge engineer. Cole's paper (1987) outlines the tasks of the terminologist: background reading, documentation selection, scanning, collation of data and finally the adding, updating or removing of term records. The first three correspond to knowledge acquisition, since this is how the terminologist learns about the subject field, its system of concepts and extracts the terms and contexts for the next stage of a project. The collation of data and production of term records correspond to knowledge formalization: the terminologist must organize the information in a way that will be useful for the end user. Finally, the ongoing modification of terminology works (eg. dictionaries and term banks) corresponds to knowledge refinement, which can be called quality control; information in a knowledge base, as in terminology works, must keep up with advances in human knowledge if they are to effectively serve the purpose for which they were created.

Terminologists, then, can be seen as a kind of knowledge engineer, and it obviously follows from this that they can only benefit from tools being developed in a knowledge engineering context. The practice of borrowing from other fields is not new to terminology.

1.2.5 Knowledge extraction technology

Before knowledge can be acquired, formalized and refined, it naturally has to be extracted, or retrieved, from the body of text in which it is embedded. When the body of text in question is electronic, then corpus analysis tools come into play to help researchers derive meaningful data from their corpora. Terminologists have the specific need of discovering and representing the concepts in a field and the conceptual network underlying the texts written in a given subject field. The Text Analyzer (TA) component of the IKARUS program has the ability to extract sentences showing the semantic relations that hold between concepts, thereby having the potential to help semi-automate this kind of knowledge extraction necessary in terminology. Following is a description of the TA and its functionalities.

1.2.5.1 The Text Analyzer

1.2.5.1.1 What is the Text Analyzer?

The Text Analyzer (TA) is a type of corpus-analysis tool that enables users to extract and analyse certain kinds of information contained in electronic documents (Kavanagh, 1995). Currently, the TA operates on the WWW, meaning that any researcher with access to standard browsers (for example, Netscape 3.0 or higher) can use the program.

1.2.5.1.2 Potential users of the TA

The TA is proposed as a tool for any person or group of people "whose job requires them to search for knowledge in documents" (Kavanagh: 1995, 2). The program's developers name

specifically, among others, terminologists as a group of people who would benefit from this technology. Traditionally, terminologists examine vast amounts of text (a process called scanning), looking for terms and discovering the conceptual network of a given subject field. Their job would be greatly facilitated in terms of decreased time and increased productivity if the scanning could be at least semi-automated.

1.2.5.1.3 Functions of the TA

The TA has a number of operations, not all of which came into play during the present study. The main operations are the following:

Preprocessing

a) sentence delimiting
b) part of speech tagging
c) finding and grouping compound nouns

Main Processing

d) frequency operations
e) concordance
f) collocations
g) conceptual operations

Figure 1 below shows the main menu of the TA with the Operations menu expanded.

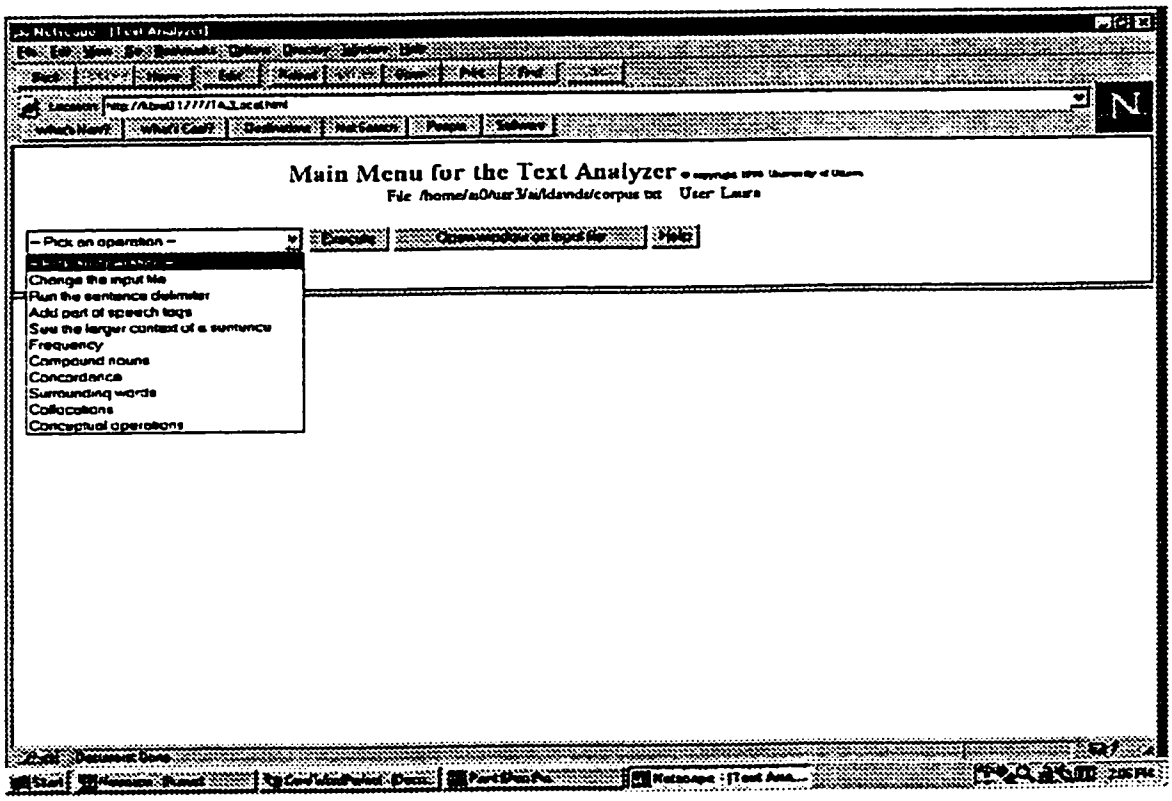


Figure 1—The Text Analyzer (main menu)

Sentence delimiting is a preprocessing operation that must be performed on a text before it can be used in the TA. This operation divides the text into individual sentences and numbers them in the order they appear in the text.

Part-of-speech tagging is another preprocessing operation. It is required only for some functions (such as finding and grouping compound nouns).

Finding and grouping compound nouns, the last preprocessing operation, is completely optional. The TA looks for all sequences of consecutive words (or nouns if text is tagged) that are at least two words long. The grouping part of the function "links" the words in the potential compounds (by replacing the spaces with the "=" sign) so that the compounds can be used for

other functions, such as concordancing (because the concordancer recognises only single words or linked compounds).

The frequency operations analyse the text for frequencies of either a specific word, all single words, word pairs or word triples. The user can specify a minimum frequency, if desired. As well, the user can have the program ignore common and relatively unimportant words (such as *the*, *January*, and *is*) that the programmer has previously entered into the program. Frequency analysis is helpful for determining potential terms, the assumption being that a text treating a given subject field will use the terms from that field relatively frequently.

The concordancer searches for all instances in the text of a search word. The user can specify the output to be a KWIC concordance (key word in context) or a full sentence concordance.

The collocation operation looks for other words in the vicinity of a search word, within a range of five words on either side. The output is a table indicating the relative position of neighbouring words and their frequencies. The assumption is that words appearing consistently in the same position relative to the search word are collocations.

Finally, the conceptual operations were the most important component of the TA for the present thesis. The TA attempts to find expressions of the following four semantic relations involving a search word: hyponymy, meronymy, synonymy and functionality. As mentioned in the Introduction, one of the goals of this thesis was to *enhance* the functionality of these operations for English text and *enable* the TA to perform the same operations on French text. To perform these functions, the TA searches for instances of the search word in association with linguistic patterns that express the four relations. For example, when users perform a meronymic search for a given term, the TA looks for all sentences that contain the term AND any of a set of linguistic patterns (such as *part of*) that express meronymy. These operations are only as good as the

quality and number of patterns with which the program is equipped; a human still has to analyse the output to weed out the "noise" from the useful sentences.

The next chapter describes the three semantic relations studied in this thesis: hyponymy, meronymy and functionality. Since semantic relations are at the core of knowledge-rich contexts, it is important to understand these relations and the structures (words or word combinations) that express them.

Chapter 2 EXPLORING SEMANTIC RELATIONS

2.0 INTRODUCTION

The concept of semantic relations is not new. Lyons (1968, 443) discussed the "priority of sense-relations". Lyons (1977) covers several of these in detail. Evens *et al* (1980) dealt with various relations and their use for classification in different fields. Palmer (1981) discussed a number of relations. Felber (1984) talked about logical and ontological relationships and synonymy. Cruse (1986) devoted entire chapters to individual relations, such as meronymy and taxonomy. Winston *et al* (1987) elaborated a "taxonomy of part-whole relations". Iris *et al* (1988) dealt with the problems of the part-whole relation. Miller (1990) discussed relations and their importance to the WordNet project.

The idea that there exist recurring linguistic structures to express semantic relations is not new either. These structures have been designated by different terms, depending on the researcher. Lyons uses *formulae*; Cruse referred to *diagnostic frames* or *test frames*; Winston *et al* simply use *frames*; Flowerdew (1992) talks about *linguistic structures* that make certain definitional information salient (and break these down into *boosters* and *downgraders*). More recently, Pearson (1996) refers to *hinges*. Ahmad and Fulford (1992) call them *knowledge probes* and used them for computerized knowledge extraction.

What *is* relatively new is the concept of using these linguistic structures in computerized text analysis. The Ahmad and Fulford study examined structures expressing synonymy, hyponymy, meronymy (they call it the partitive relation), as well as causal and material relations with a view to determining their efficacy in extracting knowledge from electronic corpora. The

Pearson study focussed on superordinacy/hyponymy as it is linguistically expressed in formal definitions in order to, in turn, extract formal definitions from corpora.

This chapter explores semantic relations, the linguistic structures by which they can be expressed, and the ways in which these structures can be used for computerized knowledge extraction from electronic corpora.

2.1 SEMANTIC (CONCEPTUAL) RELATIONS

2.1.1 Semantic relations and the field of terminology

According to Robison (1970, 273), "Human language can be viewed as a vehicle for describing relationships". In the field of terminology specifically, determining concepts, conceptual relations, and conceptual networks is vital. In the words of Felber, "any terminology work should be based on concepts and not on terms" (1984, 116). He defines *concepts* as "the mental representations of individual objects....not only of beings or things...but...also of qualities...actions...and even of locations, situations or relations." (1984, 115). Lyons' definition is the following: "By concept is to be understood an idea, thought or mental construct by means of which the mind apprehends or comes to know things....concepts mediate between words and objects" (1977, 110).

Concepts do not exist in a vacuum, though; they have meaning only by being set in relation to other concepts. In fact, "the sense of a lexical item may be defined to be, not only dependent upon, but identical with, the set of relations which hold between the item in question and other items in the same lexical system" (Lyons: 1968, 443). Ahmad and Fulford reflect this as

well: "The terms of a domain do not exist in isolation, they are related to one another in a variety of ways" (1992, 1). In any subject field, then, there are semantic links among the concepts. The web of such links constitutes what is known as the *conceptual network* of the subject field. In French, this network is called the *réseau notionnel*. In the words of Rondeau (1981, 85), "une notion se délimite par le biais de ses rapports avec d'autres notions et par la place qu'elle occupe dans un réseau notionnel."

Terminologists engage in concept analysis to determine the relations between the concepts that the terms designate, for the purpose of understanding the subject field and adequately representing that knowledge on term records. According to Meyer (1994, 7):

> High-quality concept analysis is a *sine qua non* for high-quality terminology work: without some understanding of the conceptual structures underlying the domain, the terminologist cannot properly carry out many of the practical tasks related to the production of a vocabulary.

During a terminology project, terminologists aim for as full an understanding of the domain as possible. Sager (1994b, 43) explains the difference between passive and full understanding: "We understand 'passively', when we have only a vague idea of the place of a concept in the knowledge space. We understand 'fully' when we know the precise place of a concept in relation to others". We can consider a given subject field as a special knowledge space, and "pour un domaine donné, chaque notion occupe une place définie à l'intérieur d'un système organisé de relations" (Larivière: 1996, 410).

Moreover, conceptual analysis of the subject field is not an isolated activity in a terminology project; rather, it permeates all aspects of the project. In the words of Meyer (1993, 140), "Concept management is crucial to all stages of terminology work as it is practised today

and will become even more so for the next generation of term banks." Being able to manage concepts adequately and establish conceptual networks presupposes a thorough understanding, on the part of the terminologist, of the various semantic relations and the ways in which relations between concepts are expressed in real language.

2.1.2 Linguistic patterns that express semantic relations

In the documentation of any subject field, definitions and explanatory material[3] are used to make explicit the semantic relations that exist between concepts in that subject field. Each semantic relation can be expressed by a variety of linguistic structures consisting of one or more words that indicate to readers what relation is being dealt with at that moment. The pattern *consisting of*, used in the previous sentence, expresses meronymy, i.e. the "parts of" relation. This structural pattern is, therefore, a powerful semantic device that makes it clear in the reader's mind that the relation being expressed between concept X and concept(s) Y is meronymy.

2.1.3 Definitions as expressions of semantic relations

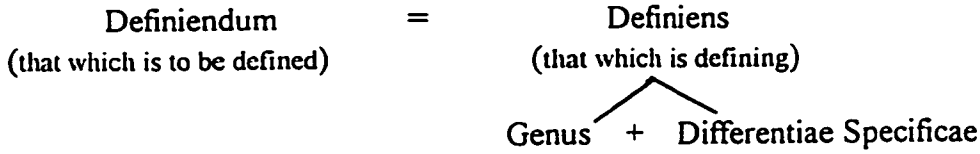
Austin considers definitions to be a type of performative utterance (1962, 65). He explains that the term *performative utterance* "indicates that the issuing of the utterance is the performing of an action—it is not normally thought of as just saying something" (1962, 6-7). For example, a person uttering the sentence "I promise to call you tomorrow" is making a promise by saying so.

[3] By *explanatory material*, I mean any definition-like information that cannot be classified as a formal definition.

Said another way, by uttering the above sentence, a person is doing two things: saying something *and* making a promise. The word *promise* in this sentence is called a *performative verb*.

Pearson (1996, 817-818) explains that performative utterances are not always prefaced by the performative verbs that act as markers, such as "I promise..." or "I swear..." Indeed, definitions generally do not appear with the explicitly stated performative verb phrase "I hereby define..." signalling that a definition or some explanatory material is about to follow. Readers are nevertheless able to recognize a definition when they are reading a text. How? Austin explains that illocutionary acts are conventional acts, i.e. "an act done as conforming to convention" (1962, 105). In the case of definitions, those conventions take the form of a) an equational structure and/or b) formulaic linguistic patterns or structures that mark the utterance as a definition.

As Picht and Draskau (1985, 51) explain, "the definition has a considerable similarity with the mathematical equation." Indeed, following Aristotle, the structure of a formal definition can be written as a simple equation:



In natural language, the "=" sign is usually expressed by a verb such as *is* or *consists of*. For example: ***a car is a means of transportation that has four wheels and travels on roads***, which distinguishes *car* from other means of transportation such as *motorcycle* or *boat*. The diagram above shows the classical definition structure (called an *intensional definition*, in the terminology

literature). However, as Picht and Draskau (1985, 51) point out, "with certain modifications to the right hand side of the equation, this structure may also be applied to other types of definitions." One type of definition described by Picht and Draskau (1985, 52) that has relevance here is the *extensional definition*.

The extensional definition involves naming the units that make up a whole or naming the list of specific instances of a generic concept. Consider the following examples:

a) North America is a continent made up of three countries: Canada, the United States, and Mexico.

b) Types of operating systems include DOS, Unix, and Windows.

Regarding the order of elements in a definition, Pearson (1996, 818-819) points out (with reference to the classical definition structure) that "contrary to what one might normally expect...the term which is being defined can appear before or after the defining statement", i.e. on either side of the definitional equation. For example, "The black substance made from the remains is called humus" (Pearson: 1996, 823).

At this point, it should be pointed out that the *definiens* side of a definitional equation can contain actual terms from the same subject field in which the *definiendum* occurs. Ahmad and Fulford (1992, 18) note that "definitions largely comprise descriptions of semantic relations holding between terms [concepts]." For example, consider the following sentence which defines the concept *compost process*, from the field of composting: "The compost process is a partial breakdown of organics by microorganisms such as bacteria and fungi." The sentence fragment following the *is a* structure contains the words *organics*, *microorganisms*, *bacteria*, and *fungi*,

which are, themselves, terms from the field of composting. This is useful for terminologists during the stage of term identification and the fleshing out of their concept tree of the domain.

Loffler-Laurian (1990, 14-18) deals with definitions as expressed in scientific discourse. She describes five *catégories définitoires* corresponding to five ways of answering the question "Qu'est-ce que c'est?". This question can be answered by a) naming the item (Denomination); b) providing a lexical equivalent having a larger extension (Equivalence); c) indicating the characteristics of the item, usually through adjectivisation (Characterisation); d) describing the parts/components of the item (Analysis); and e) revealing the item's function (Function). Loffler-Laurian (1990, 19-20) also points out which types of scientific discourse the various types of definitions can be found in. She explains that, in highly specialised scientific discourse, definitions are very rare, but when provided, they tend to be from all definitional categories *except* Function. In semi-popularised discourse, Analysis and Function definitions are most common. Popularised discourse (the goal of which is to inform the reader, yet lending the appearance of *scientificité*) seems to provide all types of definitions except Characterisation. In pedagogical discourse, the reader can expect to find Denomination, Characterisation, and Analysis, with Equivalence and Function being uncommon.

The information outlined in the preceding paragraph is evidence of the importance, in terminology, of including in the corpora texts having different levels of technicality. Terminologists, seeking as much information about the given subject field, will need access to *all* types of definitions, and not just to those provided in, say, text written by and for specialists.

Section 2.2 below describes the relations of hyponymy, meronymy and functionality and includes a discussion on previous research related to linguistic structures, appearing within definitions and explanatory material, that express these relations.

2.2 THE THREE SEMANTIC RELATIONS STUDIED IN THIS THESIS

2.2.1 Choosing which relations to study

There are many interesting semantic relations that provide avenues for study: hyponymy, meronymy, functionality, synonymy, antonymy, causality, etc. However, the present thesis is limited to the first three.

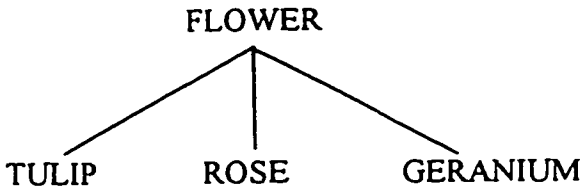
Given that the secondary purpose of this thesis is to contribute to the development of the Text Analyzer (TA), I limited myself to working with, at most, the four relations that the TA can already handle: hyponymy, meronymy, functionality, and synonymy. I later decided not to study synonymy after all because the source (TERMIUM) I used to discover the linguistic patterns for the chosen relations turned out to be poor in explicitly stated synonymous relationships.

While it may at first seem that studying only three relations is not very extensive, at second glance, we realize that each relation is complex, as described in the following sections.

2.2.2 Hyponymy

Hyponymy is a relation of inclusion. In the words of Lyons (1968, 453), "...the meaning of *tulip* [or *rose* or *geranium*] is said to be 'included' in the meaning of *flower*." To take another example from Lyons (1968, 453), "the 'meaning' of *scarlet* is said to be 'included' in

the 'meaning' of *red*." When represented hierarchically, "The 'upper' term is the *superordinate* and the 'lower' term the *hyponym*" (Palmer: 1981, 85).



Therefore, the term *tulip* is a hyponym with respect to *flower*, while *tulip*, *rose* and *geranium* are co-hyponyms, i.e. hyponyms of the same superordinate term. Looking at this from the opposite angle, *flower* is the superordinate with respect to *tulip*, *rose* and *geranium*. (It should be noted that the designation of a term as a superordinate or hyponym is relative, not absolute. For example, *tulip* is hyponym of *flower*, but in turn would be a superordinate of *parrot tulip*).

This relation is also called *generic-specific* or *taxonomy*[4]. Lyons (1977, 291) defines hyponymy as "the relation which holds between a more specific, or subordinate, lexeme and a more general, or superordinate, lexeme." Lyons (1968, 453) explains that "This relationship...has been formalized by certain semanticists in terms of the logic of classes". Indeed, "From the time of Aristotle this relation [taxonomy] has been central to the process of definition. The classical form...starts with a *genus* which is...superordinate to the term to be defined. To this are added the

[4] Cruse views taxonomy slightly differently. He says it "may be regarded as a sub-species of hyponymy: the taxonyms of a lexical item are a sub-set of its hyponym" (1986, 137). While the present thesis recognizes Cruse's distinction, it will not be elaborated on here.

differentiae which distinguish this term from related terms" (Evens *et al*: 1980, 119). The "related terms" just mentioned are co-hyponyms.

Regarding the specific linguistic structures that express hyponymy, semanticists and other scholars have indicated some as part of their discussion on semantic relations. Cruse claims that "A useful diagnostic frame for taxonomy is: An X is a kind/type of Y" (1986, 137). The X and Y are "slots" that represent any logically and grammatically correct terms such as "A spaniel is a kind of dog". Cruse raises an issue that is relevant to the problem of *noise* (i.e. sentences that contain this structure but do not express the relation in question). He states that "the expression *kind of* is not univocal, and it is necessary to be able to recognize those senses which are irrelevant for the diagnosis of taxonomy" (1986, 137-8). For example, "he was wearing a kind of flattened, three-sided turban—I don't know exactly what it was" (1986, 138). This statement is simply an approximate description of an unfamiliar item and should not be mistaken as hyponymy. Cruse also proposes diagnostic frames for co-taxonymy (co-hyponymy): "An X is a kind of Y, and a Z is another kind of Y" and "...and so is a Z". The expressions *another kind of* and *so is* both imply that X is not the only member of the class Y, that Z, too, belongs in this class. Cruse also suggests a frame for verbs, which, as he says, "seem to show hierarchical structuring to a more limited extent than nouns" (1986, 139). For example, he mentions "X-ing is a way of Y-ing", as in "walking is a kind of moving".

Lyons, too, has determined some of the linguistic structures indicating hyponymy. "When...hyponymy holds between nouns, it is possible to insert syntactically appropriate expressions containing them in place of X and Y in the following formula 'X is a kind of Y'...

'sort' and 'type' may be substituted for 'kind' in colloquial English" (1977, 292). He also suggests that "There are many other more specific lexemes...which may be employed...eg. 'shade' in 'Crimson is a shade of red" (1977, 292-3). Indeed, this is a hint that, within subject fields, there may be structures that are particular to that field, as discussed later in this thesis. Regarding co-hyponymy, Lyons says that "When a noun X is superordinate to more than one hyponym...such expressions as the following will be accepted as meaningful: 'cows and other (kinds of) animals'" (1977, 293). This implies that cows are not the only member in the class of animals, which is obvious to most people, but for terminologists working in an unfamiliar field, this type of implied knowledge is very informative by indicating that there are further hyponyms to watch for.

Ahmad and Fulford (1992, 13) prepared a list of such potential structures for hyponymy (and four other relations) and tested some of them for effectiveness on a corpus. The list includes *form%*, *type of*, *is a*, and *group** [5]. Pearson, who dealt with hyponymy only, as expressed in formal definitions, discovered different "hinges" (1996, 820) depending on whether X comes before the defining statement or after. Some examples of hinges are: *X is/are*, *X consist(s) of*, *X is/are defined as*, and *is called a X*. Pearson (1996, 821) expresses concern about focussing adverbs such as *generally* and *usually* because they "prevent statements from having generic reference....When a term is described as *usually* having a particular characteristic, it is not possible to conclude that it *always* has this characteristic". I believe that excluding sentences containing these adverbs is fine under ideal conditions. However, the fact that "X is *usually* blue" still

[5] Ahmad and Fulford use the % and * symbols as wild cards (in a computer program) to allow for the different inflectional variations of the probes. The % symbol stands for a single character; the * symbol stands for any number of characters.

constitutes knowledge, and as such is valuable information for terminologists, who are interested in the actual state of knowledge as it exists in a subject field, and not in ideal expressions of relations.

Borillo (1996, 113) writes that "En linguistique, on s'intéresse en général à la relation d'hyperonymie dans le cadre de la construction de dictionnaires, pour l'élaboration de la définition des mots". She conducted a study, the objective of which was to "voir comment la relation d'hyperonymie se manifeste linguistiquement dans les textes, i.e. sous quelles formes lexico-syntaxiques elle trouve à s'exprimer" (Borillo: 1996, 114) and later "repérer dans les textes des structures très locales susceptibles d'être interprétées comme des mises en relation de type hyperonymique" (Borillo: 1996, 121). The study revealed a wide variety of French structures, including (but not limited to) the following:

-Na est un NX [where Na is the hyponym and NX is the hypernym]
-Na est un sorte/espèce/type/variété de NX
-Na est un NX très ADJ
-Na et autres NX, Na et tout autre NX
-NX et plus particulièrement Na
-Na, le NX le plus Adj
-NX comme (par exemple) Na

Hearst (1992) conducted a similar study for English and started with the following list of patterns: *such...as*, *or other*, *and other*, *including*, and *especially*. The Hearst study was also concerned with describing a method for automatic discovery of new patterns.

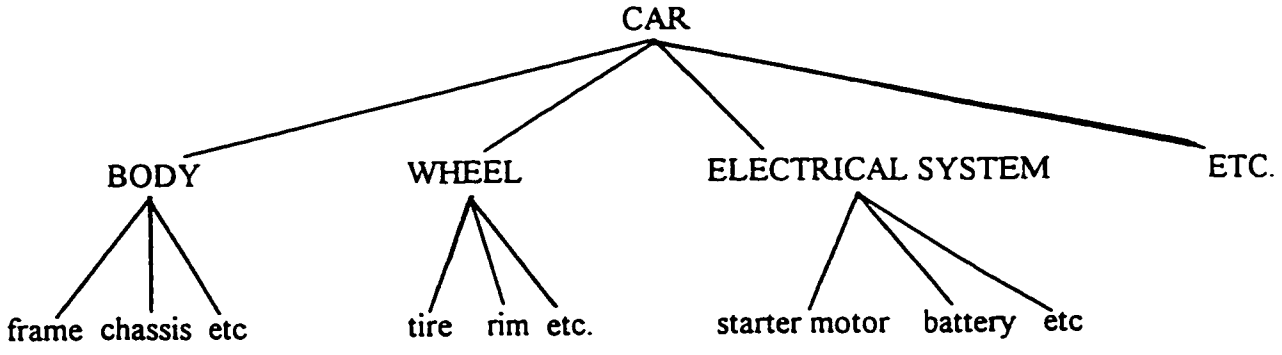
2.2.3 Meronymy

In the words of Cruse (1986, 157), "there is no doubt about the central importance of fully integrated and cohesive physical objects, with well-differentiated parts, in the concepts of 'part'

and 'whole'". Van Campenhoudt (1996, 60) describes the nature of meronymy as follows: "Le principal élément relationnel est la *connexion* du tout et de ses parties". He also explains the importance of meronymy for knowledge-based terminology: "Une base de connaissances terminologiques devrait notamment être à même de gérer les relations méronymiques à travers différentes tâches" (1996, 79).

Cruse (1986, 158) provides an amusing description of the difference between "piece" and "part": "The contrast between parts and pieces is potentially operative even with highly integrated wholes such as animal bodies: there is a clear difference between such a body hacked to pieces, and one carefully dissected into its parts". It is important to note at this point that meronymy is not limited to physical objects. In the words of Cruse, "one may also speak of parts of non-concrete entities such as events, actions, processes, states, and abstract nominal notions" (1986, 172). An example of this is the following: "The steps in WordPerfect 7.0 for italicizing text are a) highlighting the desired text, b) clicking the italics button on the toolbar, and c) clearing the highlighting."

The following diagram shows how part-whole relationships can be represented hierarchically:



There are different ways in which something can be a part of something else, and this has led scholars to claim that meronymy is a complex relation. Iris *et al* say, "the part-whole relation should be treated as a collection of relations, not as a single relation" (1988, 262). On this point, Miller asserts that "Such observations raise questions about how many different 'part of' relations there are" (1990, 256). Iris *et al* (1988) describe "four models of the part-whole relation": the functional component (the most common); the segmented whole; collections and members; and sets and subsets. Winston *et al* (1987) elaborated a "taxonomy of part-whole relations" that involves six types of meronymy:

a) component / integral object e.g.: handle/cup
b) member / collection e.g.: tree/forest
c) portion / mass e.g.: grain/salt
d) stuff / object e.g.: steel/bike
e) feature / activity e.g.: dating/adolescence
f) place / area e.g.: oasis/desert

They also propose three "relation elements": functional, homeomerous, and separable, for which each meronymy type is either yes (+) or no (-). The relational element "functional" indicates that the part has a function with respect to its whole; "homeomerous" means that the part is identical to the other parts making up the whole; and "separable" means that the part can be separated from the whole. For example, the relation between handle/cup is "component/integral object" and is functional (+), homeomerous (-), and separable (+). This means that the handle has a specific function with respect to the whole, does not resemble the other parts of the cup, and can be separated from the whole. In a later study (Chaffin *et al*, 1988), four more relational elements are added.

Given that there is more than one way X can be part of Y, it stands to reason that there exists a variety of linguistic structures that can express meronymy. Winston *et al* focus only on *part of* and its relatives: The X is part of the Y; X is partly Y; X's are part of Y's; X is part of Y; the parts of a Y include Xs, etc, but they grant that "knowledge of parts and wholes can be expressed in many specialized ways" (1987, 417-418). They do propose such words as *component*, *member*, *portion*, and *feature*, all which can substitute for *part* in specific instances (1987, 430).

Cruse discusses test-frames for meronymy, such as "A Y has Xs/ an X" (1986, 160) and "X is a part of a Y" (1986, 161), but says either one on its own is not a reliable indicator of meronymy; both can be used to express non-meronymic relations, such as "changing diapers is part of being a mother" (although I would argue that the large task of *motherhood* could perhaps be broken down into smaller tasks (parts) such as *changing diapers*). He states that "The parts of Y include the X/Xs" is a "test-frame that does not leak" (1986, 161); however, this is an ideal structure that will not always occur in actual language in use.

In the French language, Larivière points out *comprend* and *se compose de* as structures that come under the category of *composition* (1996, 412).

Miller explains that three types of meronymy are coded into WordNet: W_m is a component part of W_h; W_m is a member of W_h; and W_m is the stuff that W_h is made from, and that the *is a component of* relation is the most frequent (1990, 256).

Evens *et al* have found that "In English the part-whole relation seems to be expressed most often with *have*, *of*, or the possessive" (1980, 188-9). Lyons (1977, 312) points out that

earlier works have claimed the same thing: "part-whole relations between lexemes are bound up with a particular sub-class of possessive constructions, exemplified by such semantically...related phrases and sentences as 'John's right arm' and 'John has a right arm'".

The Ahmad and Fulford study (1992) tested the efficiency of 17 linguistic structures expressing what they refer to as the partitive relation. Their list includes *consist* of*, *constituent%*, *factor%*, *zone%*, *contain**, and *component%* (1992, 15). Interesting to note is that Ahmad and Fulford see *material* as a relation separate from meronymy (1992, 13), whereas other scholars consider it to be a type of meronymy.

Jackiewicz (1996) conducted a study in French whose purpose was to "montrer par quels moyens lexicaux la langue individualise les parties d'un tout" (51). This study focussed on verbs and nouns and classified the findings into three main categories: words that express 1) *action de composition d'objet*, 2) *action de décomposition d'objet*, and 3) *types des parties*. The first category includes such words as *réunir*, *joindre*, *ensemble*, *liaison*, *construction* and *agrégat*. A sample sentence is "Le granite résulte de la **fusion** de roches détritiques" (Jackiewicz: 1996, 58). The second category includes *désassembler*, *diviser*, and *fragmentation*. An example given is "L'eau **s'analyse en** oxygène et hydrogène" (60). The third category includes *composant*, *élément*, and *stade*, and a sample sentence is "L'opium figure parmi les **ingrédients** essentiels de la Lamaline" (61).

The ultimate conclusion to be drawn from this discussion is that meronymy is a complex relation. In addition, there appears to be no consensus on either the types of meronymy or on the

linguistic structures that express it. For the purposes of this thesis, I adopt the view of meronymy that Winston *et al* take.

2.2.4 Functionality

Unlike hyponymy and meronymy, the relation of function has not been studied in great depth. Miller sees function as a characteristic of a concept and describes it as follows: "A functional feature of a nominal concept is intended to be a description of something that instances of the concept normally do, or that is normally done with or to them" (1990, 257). Not all concepts, though, necessarily have a function or purpose. Miller asks us to consider how unnatural it is to say "the function of a canary is to fly" (1990, 257). Although flying is a normal activity for birds, it can hardly be considered its function. On the other hand, "Particularly among the human artifacts there are things that have been created for a purpose...[and can be] defined both by structure and use" (Miller: 1990, 259). This has great relevance for terminology, which deals with various fields of human endeavour which naturally involve human artifacts.

There are two "ways" that a concept can function. As mentioned above, an object can have a specific function with respect to the larger object it belongs to (such as wheels, which allow a car to move along a road). However, other objects can have a function outside of a part-whole relationship with respect to an "independent" object, but still within the same subject field. For example, an axe is used for cutting and splitting logs. Obviously the axe is not part of the log it is cutting, but the concepts *axe* and *log* are related, nonetheless, through the relation of the functionality of the axe.

Quite surprisingly, the Ahmad and Fulford (1992) study did not treat function at all. Larivière suggests that the constructions *vise à*, *utilisé pour*, *sert de*, and *sert à* express functionality in the French language (1996, 412).

I conclude that, even if *function* has not been greatly studied in lexical semantics, it is nonetheless important in terminology for helping understand the meaning of many concepts, particularly artifacts.

2.3 USING LINGUISTIC PATTERNS FOR KNOWLEDGE EXTRACTION

If we consider the linguistic structures expressing semantic relations to be devices, we realize that they can be very useful tools for knowledge extraction from texts. A terminologist analysing a text and seeking to understand the domain in terms of what the objects described in the text are a part of, or what its own parts are, can scan the text for occurrences of *consists of* used in combination with the terms that designate those objects. If that terminologist had a computer program equipped with this linguistic device to perform this scanning automatically, the task could be carried out extremely quickly on huge quantities of electronic text. Supply that computer program with a selection of these linguistic devices, and a program of this sort has potential for all terminologists, who spend a large percentage of the time in a terminology project trying to understand the subject field, i.e. determining the relations between concepts.

Definitions and explanatory material in the documentation of a subject field bear the responsibility of making explicit the conceptual relations that hold between the concepts in that field. Definitions appear in dictionaries, but can also be nested in a textbook, article, etc. Pearson

(1996, 817) says that "certain specialised text types contain a combination of...metalanguage statements and that many...are in fact complete or partial definitions of terms used in texts." As part of a study on semantic relations, Ahmad and Fulford (1992) analysed the "linguistic means used by authors to encode semantic relations in a text." They discovered that each of the relations of hyponymy, meronymy and synonymy (among others) can be expressed in a variety of ways. For example, two patterns for expressing hyponymy are

a) X is a type of Y
b) X is a species of Y

where X is a hyponym of Y, and Y is therefore the superordinate. The term *tulip* can fit into the X-slot; *flower* can fit into the Y-slot. As a result of their exploration of language in use, Ahmad and Fulford compiled a "lexical archive" (1992, 7) of these linguistic patterns found for each of the relations they dealt with. They refer to these patterns as *knowledge probes*, which is an apt designation because when these patterns are used in a suitable corpus analysis tool, they can identify/extract terms semi-automatically, along with definitional information about those terms, i.e. knowledge-rich contexts.

2.3.1 Relevance for terminology

For terminologists, concept analysis is the activity of determining the relations between concepts in a subject field, thereby establishing a conceptual network for that subject field.

As mentioned in Chapter 1, many terminologists are still using paper-based documentation to carry out their work, but this is a *very* slow process. It also tends to cause terminologists to deal with concepts in isolation when they should be trying to see the "big picture". Obviously,

using corpus analysis tools for extracting information from electronic texts is a speedy way of carrying out concept analysis, especially if the tools are equipped with the linguistic structures that express semantic relations.

2.4 HOW THE PRESENT STUDY DIFFERS FROM PREVIOUS RESEARCH

In the above sections, I outlined studies similar to the one I carried out (described in the following chapter). My study differed from these in various ways, however, as described below.

The information presented by Cruse (1986) and Lyons (1968, 1977) was theoretical, and focussed on cognitively plausible means of expressing semantic relations. My own study was empirical in that, to discover how French and English express relations, I searched real text for linguistic patterns actually used to express relations. Furthermore, Cruse and Lyons did not intend to use these patterns for automatic knowledge extraction, whereas that was indeed *my* ultimate goal.

The Pearson (1996) study looked at hyponymy only, specifically the expression of formal definitions. I chose to research the relations of hyponymy, meronymy and functionality. The Pearson paper was "part of a broader investigation into the possibility of producing specifications for the (semi-) automatic identification and retrieval of terminological definitions from corpora" (Pearson: 1996, 817). Some testing was done, but the results were only briefly presented in the conclusion. By comparison, my study was relatively extensive, and I present the supporting statistics.

The Jackiewicz (1996), Borillo (1996), and Hearst (1992) studies were similar in that they all discovered linguistic structures with a view to using them for semi-automatic knowledge extraction from corpora. Each was concerned with only a single relation while I investigated three. None of these studies carried out statistical analysis of effectiveness, whereas my does.

The Ahmad and Fulford study (1992) most closely resembled mine. They researched five relations and the linguistic structures expressing them. They proceeded to test the effectiveness of individual patterns as *probes* for knowledge extraction. I modelled part of my testing after theirs. Where my study differs is in its testing of the entire collection of patterns for a given relation. In other words, I studied how effective each group of patterns as a whole was for extracting knowledge.

Two areas where my study differs from all the above are 1) I researched both French and English whereas the other studies were concerned with one or the other language, and 2) I went beyond *lexical* structures to explore the usefulness of *grammatical* patterns as well.

2.5 CONCLUSION

We humans have a natural and excellent *conceptual* understanding of the semantic relations between concepts. If we did not, we would not be able to communicate. However, there is still the need for research into *how* those relations are specifically expressed in real language, if we are going to put this kind of information to work extracting knowledge from electronic corpora. Not only do we need to know all the specific structures expressing each semantic relation; we also need to discover what else those same structures express in order to develop

"filters" to prevent useless information from being extracted. All this implies a detailed analysis of individual linguistic patterns for a given language. To add to this work, we can then proceed to the same study for other languages, if we want to broaden the capabilities of knowledge extraction technology.

The next chapter presents my own such study of French and English.

Chapter 3 DISCOVERING AND TESTING LINGUISTIC PATTERNS

3.0 INTRODUCTION

This chapter explains how I discovered and tested French and English linguistic patterns that typically express the semantic relations of hyponymy, meronymy, and functionality. The source for *discovering* the patterns was the1996 CD-ROM version of TERMIUM. The corpora used for *testing* them for their efficacy in the semi-automatic extraction of knowledge-rich contexts were a French one and an English one in the subject field of Composting. I used the World Wide Web (WWW) as the sole source for the texts in my corpora with the intention of showing the potential of the WWW as a "virtual corpus" for building real terminology corpora.

The research for this chapter was carried out in four stages: choosing a subject field and building the corpora; discovering linguistic patterns for three semantic relations; preliminary testing; and final testing. These four stages are described in Sections 3.1 to 3.4 respectively. Section 3.5 presents the results and analysis of the final testing, and Section 3.6 provides some considerations for future research.

3.1 CHOOSING A SUBJECT FIELD AND BUILDING THE CORPORA

3.1.1 Choosing a subject field

The task described in this chapter involves building and using two corpora. In *lexicography*, corpora are composed of a wide range of *general language* texts. However, since this thesis deals with the usefulness of knowledge extraction for *terminology* work, the texts used

to build the corpora had to contain *specialized* information on a single subject field (in order to parallel the kind of corpus that would be used by real terminologists).

3.1.1.1 Criteria affecting the choice of subject field

The choice of subject field rested on three criteria. First, it had to be a subject that was reasonably well represented on the WWW. By this, I mean that there were a variety of documents on the subject (in both French and English), originating from more than one source. Second, the documents available had to be of an instructional nature, thereby increasing the likelihood that they would contain definitional and explanatory information. Instructional texts generally explain the concepts in the subject field and their relation to other relevant concepts. Third, the field chosen had to be relatively new and of immediate importance to people today, again to parallel the normal situation in terminology. However, the field could not be so new that most of its terms would not appear in the 1996 version of TERMIUM. (The reason for this will become clear in Section 3.2.1.) Finally, choosing this subject field fits in with my philosophy of maintaining a broad perspective towards knowledge, rather than a narrow focus: I wanted a field *outside* the area of computing, since most of my previous domain expertise was in the latter.

3.1.1.2 Suitability of "Composting" as the subject field

With these criteria in mind, I decided on the subject field of "Composting". The WWW provided a range of French and English documents on this topic, all explaining the science and the "how to" of composting. The subject is not entirely new; what *is* recent, though, is people's

awareness of the detrimental impact that humans can have on the environment. Because of the importance of this issue, provincial governments, among other organizations, are starting to make available on the WWW such information as I found, detailing the process of and reason for composting, as well as how to do it.

3.1.1.3 A few remarks about "Composting"

Some persons unfamiliar with composting may have an overly simplistic view of it, imagining that it is as easy as throwing one's banana skins in the backyard and watching them rot. In reality, composting draws on the knowledge and terminology of waste management, physics, soil sciences, biology, and chemistry. In addition, composting is undertaken not simply on a small scale by environmentally conscious individuals, but also on a large scale by municipal and commercial composting operations.

It is interesting to note, as well, that some terms used in composting are actually common general language words (like *grass* or *weeds*). While these terms themselves do not change their basic meaning when used in a composting context, the underlying concepts do take on a greater importance within the field and are viewed from a different angle. Grass, for example, is an excellent source of nitrogen (but a reader would not find this as part of a definition of *grass* in a general unilingual dictionary). This is important knowledge, though, because a compost pile must have the proper carbon-to-nitrogen (C/N) ratio, which presupposes that the builder of the pile knows which raw materials contain what elements and in what quantities. We must, therefore, not underestimate the value of everyday terms when they are used in a specialized subject field.

3.1.2 Building the corpora used for term identification and testing of linguistic patterns

3.1.2.1 Source of electronic text

Corpora are bodies of text that **must** be in electronic (machine-readable) form in order to be processed by a computer. What Sinclair (1991, 14) stated about getting text into a computer is still the case today: "There are three normal methods of text input...: a) adaptation of material already in electronic form; b) conversion by optical scanning (machine reading); c) conversion by keyboarding." The latter is, by far, the least desirable because of the time it would take to produce a corpus of reasonable size. Option b) is much more time-efficient; in addition, the cost of scanners is decreasing rapidly. However, building a corpus with text already in machine-readable form is the ideal method.

The advent of the Internet, specifically the WWW, has meant the ready availability of information on a wide variety of subjects—all in machine-readable form. While a certain amount of information on the WWW is of dubious quality, researchers can find documents from reputable sources as well. To build my corpora, I culled my texts from the WWW because of the convenience of using information already in electronic form, and to demonstrate that terminologists too would find this technology a valuable tool for their work.

3.1.2.2 Internet tools used to build the corpora

For finding texts on the WWW, I used the Alta Vista search engine. This tool has the reputation of being a very productive engine (eg. Venditto, 1996), retrieving vast numbers of WWW addresses for websites that it considers relevant to the user's search word or phrase. It is

true that Alta Vista can retrieve from its database what may at first appear to be an overwhelming number of documents. (Users of this search engine are familiar with being informed that, for example, 2,000 documents were found for their query!)

Two considerations, however, can override this information overload "anxiety". First, Alta Vista presents the user with a ranked, not random, list of document addresses. They are listed in descending order of relevance (based on four ranking criteria), with the most relevant at the top. Although Alta Vista's relevancy ranking is not considered to be the most effective (Venditto: 1996, 81), in my own experience with this tool, I have found that the first 40 to 50 addresses are the most relevant. Any appearing further down on the list can usually be ignored.

The second point to consider about a search engine that retrieves *more* than a researcher needs is that it is infinitely easier to ignore an irrelevant document than it is to find relevant ones that have been missed or by-passed. Engines such as Yahoo or Magellan provide a single document or a short list of documents under very precise topic headings. It is the engine's staff who determine what topic headings to file documents under. This prior categorization takes the control out of the users' hands: users cannot always know how and where the staff have categorized its web sites. Alta Vista, on the other hand, retrieves all documents containing the users' search word or phrase, which is why the list of document addresses is so long for most queries. As a result, users can decide how useful or not a given document will be[6].

[6] This is not to say that Alta Vista is "better than" Magellan or any other search engine. Each has its merits and search philosophy. Researchers themselves must decide what type of search they want to perform during a given project—which obviously implies that researchers become familiar with a variety of these important tools (Venditto: 1996, 81).

3.1.2.3 Nature of the texts and size of the corpora

From all the potential documents that Alta Vista provided me, I chose only those documents[7] dealing with composting as the main topic. Fortunately, all of these were instructional, which meant that the semantic relations between concepts were made explicit. Any texts in which composting was mentioned only briefly, I rejected.

The reader may not be able to tell from looking at the Webliography that some of the documents in one language were translations of the other. The Government of New Brunswick, for example, posted the same composting information in both French and English. Other documents, though, were available in either English or French only.

As to size, the English corpus was 27,065 words, or 86 pages of text. The French one was 21,741 words, or 74 pages. The French corpus is smaller because, at present, information on the WWW is still predominantly in English. In light of that fact, it was encouraging to have been able to build a French corpus that was at least close in size to the English.

3.1.2.4 Problem: electronic French text in various computer operating systems

At this point, I encountered and solved a small technical problem that I feel should be outlined to aid future researchers who may wish to perform a similar study using the same combination of computer programs as I did. The problem was small in that its solution was very

[7] A "Webliography" of the WWW document addresses (along with the organization providing the documents) is provided at the end of the thesis.

simple, once I actually discovered it[8]. The problem was a major one, however, in that, without the solution, I would not have been able to complete much of the work necessary for this thesis.

When users need text only from a WWW document in Netscape, they simply save the document with a *.txt* extension. Users may not realize that the text is saved in ANSI (Windows) format, not in ASCII (DOS) format. When the user opens the file in WordPerfect for Windows, a dialogue box confirms the conversion from ASCII format. Users who think the file *is* in ASCII format will select "yes", and the file will open. If the text is written in French, all accented characters will be represented with an incorrect character or symbol. Writing a macro to convert all "bizarre" characters into the correct ones solves the immediate problem in Windows, but is completely unnecessary. It also complicates the issue if the user needs to work with the French file in UNIX because the file is in ASCII format; French text *must* be in ANSI format to preserve the accented characters in UNIX.

The solution comes in knowing that text files saved from Netscape are already in ANSI format, and thus are "UNIX ready". (Any French text files not in ANSI format would have to be resaved as such to be used in UNIX.) To open the file in Windows so that it displays accented characters correctly, the user must, when presented with the dialogue box confirming conversion from ASCII, open the drop-down menu of file formats and select ANSI (Windows) text. This allows the user to work with French documents in Windows while leaving them in UNIX-friendly format.

---

[8] It was David Miller, a Master's graduate from the School of Translation, who provided me with the main solution.

3.2 DISCOVERING THE LINGUISTIC PATTERNS THAT EXPRESS HYPONYMY, MERONYMY AND FUNCTIONALITY

3.2.1 TERMIUM as the source of linguistic patterns

The source for discovering linguistic patterns was the 1996 version of TERMIUM. I rejected the idea of analysing specific text types where definitional information is common (such as text books, scholarly articles, and magazines) in favour of TERMIUM for the very reason that this term bank is itself a distillation of such information culled from the above-mentioned types of sources. TERMIUM employs terminologists, who are specially trained to do this work. Therefore, TERMIUM is a repository of high-quality real-life knowledge-rich material. I analysed the following textual support sections on the French and English sides of 76 term records in the subject field of composting: *definition*, *context*, *example*, and *observation*. In Section 3.2.3, I explain how I decided which terms to use.

3.2.2 Nature of the textual support fields on TERMIUM records

This section is based on information from the 1984 in-house Terminologist's Handbook used at TERMIUM (Handbook: 1984).

The Handbook defines *textual support* as "any material in the form of running text entered on a record to give information on a term or concept." TERMIUM provides four types of textual support: a definition, a context, a usage sample (also called *example*), and an observation. Individual records, however, may or may not have the full complement of fields.

The definition field contains a formal definition, i.e. the class that the concept belongs to, along with its distinguishing characteristics. Definitions can be a) handcrafted by terminologists

based on their research, b) composed of chunks of information taken from various sources, or c) quoted verbatim from sources, where the given definition is sufficiently complete.

The context field provides the user with the term in either a defining, explanatory or associative context. Defining contexts are not quite formal definitions, but do provide enough information so the user can identify the concept of the term. Explanatory contexts illustrate the place or use within the subject field of the concept in question. Finally, in associative contexts, the term occurs in association with other terms or concepts in the given subject field. In this way, the user can determine the field and possibly the concept's place in the field.

Usage samples usually give little or no information as to the meaning of a term. They attempt to demonstrate that a term exists or to show grammatical behaviour. Even though usage samples may not provide meaning information about the term, they may still provide linguistic patterns. For example, the meronymic pattern *la teneur* was found in the following usage sample on the term record for *potassium*: "La teneur du potassium varie localement dans l'eau de mer." This sentence says very little about potassium itself, but nonetheless yielded a linguistic pattern. And discovering linguistic patterns was the purpose of this stage of the research.

Finally, the observation field provides explanations or supplementary information on a term. The information for this field can be either taken from a source or provided by the terminologists from what they have discovered about the term.

3.2.3 Determining which terms in TERMIUM to analyse

At present, TERMIUM does not have the feature whereby users can search for terms by subject field; users are limited to searching for individual terms or words and then seeing how many and in which domains those terms belong to. Therefore, in order to determine my nomenclature, I proceeded with the following steps.

First, I ran my previously prepared French and English composting corpora through the TA's "word frequencies" operation. I prepared a printout of the frequencies of single words, word pairs, and word triples. With this information in hand, I chose the most frequent words or word combinations that fulfilled each of the following criteria:

a) were meaningful terms in the field of composting;

b) appeared as headwords in TERMIUM; and

c) their TERMIUM records contained some textual support (i.e. something written in at least one of the definition, context, example or observation fields.

To these three, I added a fourth criterion in the case of terms chosen from the French frequency analysis. A given TERMIUM record contains a French and English term for the concept in question. I performed the frequency analysis on the English corpus first. Since some of the documents in the French corpus were translations of the English (or vice versa), there is obviously a "duplication" in the terms used in the French corpus. Therefore, the fourth criterion is d) French records chosen must not already have been selected during my investigation of the English frequency analysis. For example, the French term *compostage* does not appear in the list on the previous page because it shares a TERMIUM record with *composting*, and this record, therefore,

had already been analysed. The entire selection procedure provided me with 76 records for 68 composting terms. Figure 2 below is a list of the term records used and the frequencies of terms in the Composting corpora.

Figure 2—Composting Terms in TERMIUM used for Discovering Linguistic Patterns

English Terms from Composting Corpus / TERMIUM French Equivalents	Frequency in English Composting Corpus
anaerobic condition(s) / condition anaérobie	4
aeration / aération	13
bacteria / bactéries	39
[bacterium / bactérie]	
calcium carbonate / carbonate de calcium	2
carbon / carbone	36
carbon/nitrogen ratio / rapport carbone/azote	25
clay soil(s) / sol argileux	3
compost / criblé de décharge	433
[compost / compost]	
composting / compostage	209
composting system / bac à compostage	4
decomposer(s) / décomposeur, détritiphage	11
[decomposer(s) / décomposeur]	
decomposition / décomposition	27
diffusion / diffusion	13
diffusion coefficient / coefficient de diffusion	4
distilled water / eau distillé	3
earth / terre	13
fertilizer / engrais	11
finished compost / compost fini	9
fungi / eumycètes	20
garbage / déchets de cuisine	16
[municipal waste / déchets urbains]	
[household garbage / déchets ménagers]	
heat cycle / cycle thermodynamique	2
humus / humus	10
insect(s) / insecte	14
kitchen waste(s) / résidus domestiques	4
litter / litière	30
manure / fumier	47
microbial activity / activité microbienne	6
micro-organism(s) / microorganisme	21
moisture / humidité	85
moisture content / état hygrométrique	20
[moisture content / teneur en eau]	
mortality / mortalité	43
nitrogen / azote	68
nutrient(s) / substance nutritif	35
organic acid(s) / acide organique	3
organic matter / matière organique	24
organic material / matière organique	13
organic waste(s) / débris organiques	6
oxygen / oxygène	39
particle size / grosseur de grain	6
peat moss / mousse de tourbe	7
pH / pH	33
plant pathogen(s) / agent pathogénique des plantes	2

potting mixes (mixture) / terreau 3
sandy soil / sol léger 2
sewage sludge / boues d'épuration 2
soil / sol 118
[soil / sol]
solid waste(s) /déchet solid 6
soil amendment / amendement des sols 3
soil conditioner / amendement synthétique 3
weed(s) / plante nuisible 20
yard waste(s) / résidus de jardin 5

French Terms from Composting Corpus / TERMIUM English Equivalents ..... Frequency in French Composting Corpus

andainage / windrowing 3
coliforme(s) / coliform 2
digesteur(s) / digester 3
élément(s) fertilisant(s) / plant nutrient 7
[élément(s) fertilisant(s) / nutrient element]
gazon / lawn 19
géotextile(s) / geotextile 1
herbes / herbs 15
méthane / methane 2
papier journal / newsprint 3
phosphate naturel / rock phosphate 4
phosphore / phosphorus 21
porosité effective / effective porosity 2
potassium / potassium 14
réacteur / reactor 24
terre végétale / topsoil 2
tondeuse / lawnmower 3
vermicompostage / vermicomposting 8

The reader will note that the number of records consulted (76) exceeds the number of terms selected (68). This is due to the occasional existence, in TERMIUM, of more than one record for a given term, and where the equivalent in the other language may or may not be different on each record. For example, the term *compost* has two records because the concept is slightly different in each case; the French equivalent on one is *compost*, whereas on the other it is *criblé de décharge*.

Finally, I analysed the above-mentioned fields on all 76 term records. I searched for linguistic patterns expressing the semantic relations of hyponymy, meronymy and functionality. Figure 3 below is a list of the patterns found, along with a sample sentence extracted from

TERMIUM for each pattern. The patterns in the list have been "fine tuned". A detailed discussion of the fine-tuning process and the use of the asterisk "wild card" is found in Section 3.3.3.1.3.

3.2.4 Size of the TERMIUM record set

When it was time to determine which and how many term records to analyse for discovering the patterns that express the relations I chose, an important question arose: How many records are enough? First, the reason why quantity is a consideration at all has to do with exhaustiveness. For me to be able to state with any degree of confidence that I had discovered the majority of patterns that appear in TERMIUM, I would have had to analyse the majority of the records in the term bank. TERMIUM contains approximately 1,045,500 term records. The impracticality of analysing even 60% of these can be seen immediately.

The reader will recall that I analysed only 76 records, which by comparison seems like a paltry amount. However, since I am attempting to "parallel" a terminology project, I first had to limit myself to the terms of a single subject field. Then I had to work only with the composting terms that appeared in TERMIUM and that fulfilled the criteria outlined in Section 3.2.3. This narrowing down ultimately left me with 76 records. Just for comparison purposes, though, even if I had arrived at 10,000 records to analyse, this figure is still only 9/10 of 1 percent of all the records in TERMIUM—hardly a representative sample, indeed! For the purposes of this thesis, the analysis of 76 records provided me with a significant number of patterns to produce useful test results. However, it is clear that, in a large study, far more should be analysed.

Figure 3—Linguistic Patterns found in TERMIUM [9]

Hyponymy—English

Pattern	Sample Sentence from Termium
such as	compost: Organic residues, or their mixture, **such as** peat, manure, or discarded plant material and soil, placed in a pit, moistened, and allowed to become decomposed....
refer*...[20 characters].....to	Coliform: **referring** to aerobic and facultative anaerobic, gram negative, non-spore-forming bacilli in the colon....
following groups...[100 characters]...:	The **following groups** of elements are considered essential: (1) nitrogen, phosphorus, and potassium, usually called primary nutrients ...; (2) calcium, magnesium, and sulfur, called secondary nutrients... .[the colon is important in pattern]
is defined as / are defined as	Geotextiles **are defined as** permeable textiles used in conjunction with soils or rocks.
includ*	Herbs **include** and occasionally may designate any of the following: aromatic, culinary, medicinal and fine herbs.
is classified as / are classified as	Reactors **are classified as** homogeneous or heterogeneous.
types	A number of proprietary **types** [of geotextiles] are available...including Mirafi, Typor....
descriptive of	Anaerobic condition: [is] **descriptive of** a condition in which dissolved oxygen, nitrate and nitrite are absent.
is a	Carbon **[is] a** nonmetallic, chiefly tetravalent chemical element
is the / are the	carbon/nitrogen ratio (C/N ration): **[is]** the ratio of the weight of organic carbon to the weight of total nitrogen in a soil or in an organic material.
is any / are any	micro-organism: **[is]** any plant or animal of microscopical size.
is called / are called	iron, copper, manganese, boron, zinc, chlorine, ad molybdenum, which **are called** minor or micronutrients
and other / and any other (co-hyponymy)	organic material: compounds composed of carbon, hydrogen, **and other** elements with chain or ring structures.
occur*...[20 characters]....as	It [calcium carbonate] **occurs** in nature **as** aragonite, calcite, chalk, limestone, lithographic stone, marble, marl and pure [...] travertine.
term used	Earth is a **term used**...for soil that can be used for cultivation.
constitute*	Fertilizers **constitute** a special group of soil amendments.
a general term for	Litter [is] **a general term for** the layer of loose organic debris...that accumulates in wooded areas.
among	**Among** the many important organic acids are acetic, CH_3COOH, and oxalic, $H_2C_2O_4$, acids, and phenol, C_6H_5OH...

[9] Readers will note that, in some of the sample sentences from TERMIUM, one word of a multi-word pattern appears in square brackets. The brackets and the word inside are my addition to account for the elliptical format of some of Termium's definitional information. To conserve space, Termium definitions leave implicit some obvious information. For example, in the definition "micro-organism: any plant or animal of microscopic size", the verb "is", which would appear in front of "any" is implicit.

especially	
(....[100 characters]....)	[Borrowed from French hyponymy patterns, because parentheses are typographic, not linguistic (strictly speaking); therefore not language specific..]

Meronymy—English

contain / containing / contains / contained	Fertilizers are materials that **contain** one or more nutrient elements essential for plant growth.
made from	Geotextiles are woven, non-woven or knitted fabrics...**made from** synthetic fibres of polyamide (nylon)...
component of	Methane is a chief **component of** natural gas.
combin*	Newsprint is a grade of paper, **combining** high percentages of ground wood pulp.
occur*...[100 characters]...in	Phosporus **occurs in** minerals...and in all living matter.
consist*	compost: a mixture that **consists** largely of decayed organic waste, resulting from the degradation of waste.
mak*...[30 characters]...up	oxygen: A colorless, odorless gas at NTP, which **makes up** 20.99% by volume of the air in the atmosphere.
consituent* of	In the combined form the element [nitrogen] is a **constituent of** all proteins.
composed of	Organic Materials: Compounds **composed of** carbon, hydrogen, and other elements with chain or ring structures.
possess	Enzymes **possess** many ionizable groups so that pH changes may alter the conformation of the enzyme, the binding of the substrate, and the catalytic activity of the groups in the active site of the enzyme.
part of	Yard rubbish [is] essentially a **part of** combustible rubbish.
divided into	The earth's surface...**is divided into** soil and nonsoil.
with	Organic acid: a chemical compound **with** one or more carboxyl radicals (COOH) in its structure.
in	The moisture **in** cheese influences its rate of ripening, its pH, its flavor, and its nutritive value.

Functionality—English

designed for	A digester is a tank **designed for** anaerobic fermentation of biomass.
essential to	Plant nutrients...**are essential to** growth...
used	a) Plant nutrients...**are used** by the plant in the elaboration of its food and tissue. b) Geotextiles...**are used** in civil engineering works as drainage blankets... c) Phosporus...**is used in** the manufacture of matches and incendiary bombs. d) Topsoil...**is used to** [used...to] topdress roadbanks, gardens, and lawns.
functions of	Other **functions of** geotextiles include erosion control, filtration....
made...[30 characters]... for	Newsprint is **made** especially **for** use in the printing of newspapers.
for use in / for using in	Newsprint is made especially **for use in** the printing of newspapers
in order to	composting bin: A plastic container into which people recuperate vegetable waste, **in order to** recycle it and get natural fertilizers or soil conditioners.

uses to which	The variety of technological **uses to which** microorganisms are put is enormous. In brewing and baking, the alcohol and carbon dioxide produced in fermentation by yeasts are utilized.
essential for	Oxygen is **essential for** the life of most organisms on earth.
needed for	Oxygen.... is **needed for** decomposition of refuse.
utilize	Large, well-constructed and well-maintained generators **utilize** the biogas to provide power for the plant.

Hyponymy—French

constitu*	Les bactéries coliformes **constituent** le groupe des bactéries du côlon.
il s'agit d*	Coliformes. **Il s'agit de** microorganismes producteur d'acide et de gaz...
(...[100 characters]...) Parentheses function like "comme" or "tel que"	...en fournissant des éléments fertilisants majeurs (azote, phosphore, potassium).
type*	Il existe deux **types** principaux de geotextiles, perméables et imperméables.
comprennent	Les bonnes herbes **comprennent** les plantes suivantes:....
forme*	Le phospore existe sous différentes **formes** dans l'organisme vivant.
est un*	bactérie: **[est un]** microorganisme unicellulaire, microscopique, présentant un cytoplasme sans mitochondries et contenant de l'acide diaminopimélique.
est l*	activité microbienne: **[est l']** ensemble des changements biochimiques résultant du métabolisme des organismes vivants.
descriptif d*	condition anaérobie: **[est] descriptif d'**une condition dans laquelle l'oxygène dissous, les nitrates et les nitrites sont absents.
désigné...[30 characters]...sous le nom	le compostage où les ordures sont abandonnées a elles-mêmes n'est guère complet avant deux ou trois ans, et avant de l'utiliser, il faut souvent passer au tamis le produit final **désigné** généralement **sous le nom** de "criblé de décharge".
on désigne...[30 characters]... sous le nom	déchets urbains: **on désigne sous le nom** de "déchets urbains" les ordures déposées dans tous récipients [...] les résidus provenant du nettoiement de la cité et les objets abandonnés sur ses voies.
sont les / sont des	déchets ménagères: **[sont les]** déchets solides issus de la vie domestique.
sont aussi des	microorganisme: **ils sont aussi des** agents actifs dans les fermentations et participent à la décomposition des matières organiques.
constitue* ce que l'on appelle	les résidus ménagers **constituent ce que l'on appelle** les gadoues vertes [ou] ordures ménagères.
est tout*	microorganisme: **[est] tout** organisme vivant visible seulement au microscope.
s'appelle	La courbe de la tension (pression) de l'eau du sol en fonction de la teneur en eau **s'appelle** la courbe de rétention d'eau, parfois la courbe caractéristique de l'eau du sol.
variétés	Carbon...présente une grande diversité de **variétés** allotropiques, notamment : formes cristallines comme le diamant et le graphite; formes amorphes comme les carbones.
comme	...formes cristallines **comme** le diamant et le graphite...
tel que / telle que / tels que / telles que	...les noirs de carbones **tels que** les noirs de fumées et les noirs thermiques...
terme...[20 characters]...désigne	terre: **terme** qui **désigne** aussi bien le sol de culture que notre planète.
parmi	**Parmi** les acides organiques, citons l'acide acétique, CH_3COOH, et l'acide oxalique, $H_2C_2O_4$...

sortes...[30 characters]...:	On distingue deux **sortes** de materiaux à la surface de la terre...: le sol et le non-sol.
on qualifie de	**On qualifie de** sol le matériau naturel, non consolidé, minéral ou organique, qui peut maintenir la croissance des plantes.
catégorie...[30 characters]...d*	Les engrais forment une **catégorie** particulière d'amendements

Meronymy—French

constitu*	a) composteur domestique rotatif, l'«Enviro-Cycle» [...], **constitué** d'un tonneau de plastique pourvu d'une ouverture et de grilles de ventilation. b) oxygen: corps gazeux diatomique (O_2), **constituant** le cinquième de l'atmosphère terrestre.
contenir	Certains déchets biologiques risques par ailleurs de **contenir** des germes pathogènes.
contenant	bactérie: Microorganisme unicellulaire, microscopique, présentant un cytoplasme sans mitochondries et **contenant** de l'acide diaminopimélique.
contien*	Chaque cellule [d'un microorganisme] **contient** plusieurs milliers d'enzymes, chacune étant capable de catalyser un type de réactions chimiques.
contenu*	Quantité d'eau gravitaire...**contenue** dans un terrain aquifère.
combiné*...[20 characters]...avec	Il s'agit de grandes cuves cylindriques **combinées** souvent **avec** une partie inférieure conique....
compose / composent / composant	...les corps et éléments chimiques qui les **composent**
composé*...[30 characters]...d*	Gazon: végétation dense, **composée d'** herbes courtes et fines.
renferm*	Les engrais simples, **renfermant** un seul element fertlisant majeur
trouv*...[30 characters]...dans	a) On le **trouve** essentiellement sous forme de chlorure **dans** les évaporites avec.... b) Méthane **se trouve dans** le gaz de charbon et dans le gaz naturel.
à base d*	Papier journal: papier apprêté, **à base de** pâte au bisulfite et de pâte mécanique.
teneur d*	**La teneur du** potassium varie localement dans l'eau de mer.
teneur* en	Sol dont la **teneur en** argile est relativement élévée.
muni...[30 characters]...d*	La tondeuse à gazon est **munie de** pièces travaillantes rotatives montées sur....
comprend	Elle [la tondeuse] **comprend** souvent une trémie de collecte de l'herbe coupée.
comprennent	Eaux ménagères...**comprennent** les eaux de cuisine, de lessive, de toilette, etc.
chargé*...[30 characters]...d*	L'eau est...véhicule qui n'apparaît pas à l'état pur, mais **chargé de** substances minérales et organiques.
présentant	bactérie: DEF*Microorganisme unicellulaire, microscopique, **présentant** un cytoplasme sans mitochondries et contenant de l'acide diaminopimélique.
pourvu* d*	composteur domestique rotatif, l'«Enviro-Cycle» [...], constitué d'un tonneau de plastique **pourvu** d'une ouverture et de grilles de ventilation.
incluant	déchets urbains: déchets solides issus de la vie domestique, **incluant** [...] les déchets [...] des établissements industriels et commerciaux [...] les résidus [...] des écoles, casernes, hôpitaux, prisons.
entre dans la constitution d*	Son rôle biologique [celui de l'azote] est important, car il **entre dans la constitution des** amines, des acides aminés, des protéines, des amides, des alcaloïdes, des dérivés du pyrrole comme la chlorophylle et l'hémoglobine.

font partie d* ????	les virus **font partie des** microorganismes.
ayant	Bactéries: Groupe important d'organismes monocellulaires microscopiques, actifs sur le plan métabolique, **ayant** un noyau diffus (non séparé) généralement autonomes et se multipliant habituellement par fission binaire.
riche*...[30 characters]...en	Phosphate naturel: roche suffisament pure et **riche en** phosphates calciques.
trace*...[30 characters]...d*	Eau distillée: eau privée des gaz dissous, de ses impuretés minérales et organiques, (mais non des **traces de** silice et de plomb)
fraction...[30]... d*	Humus: **fraction des** matières organiques qui reste dans le sol après décomposition de la plus grande partie des débris végétaux et animaux incorporés dans le sol.
comport*	Legalement, ils peuvent **comporter** jusqu'à 70 pour 100 d'humidité.
avec	Mélange de terre ou de matières inertes **avec** des matières organiques fermentées ou susceptibles de fermenter.

Functionality—French

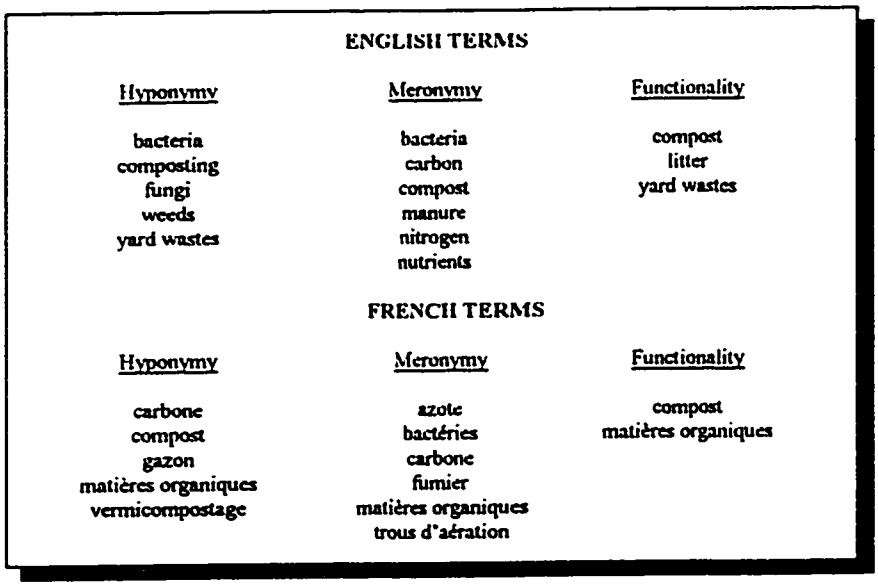
utilis*	Ne peuvent y subsister que des organismes hétérotrophes **utilisant** les débris organiques. .
serv*...[20 characters]... à serv*...[20 characters]... au	Une tondeuse est un appareil **servant à** la coupe mécanique du gazon.
sert*...[20 characters]...à sert*...[20 characters]...au sert*...[20 characters]...pour	mousse de tourbe: espèce de tourbe dont on se sert pour alléger le sol
permet*	Un digesteur est un appareil **permettant** la fermentation anaérobie de la biomasse.
joue*...[20 characters]...un rôle	Le géotextile peut **jouer un rôle** hydraulique...
fonction	Le géotextile peut jouer un rôle hydraulique (**fonction** drainante, fonction filtre)...
employé*...[20 characters]...comme	Phosphate naturel: roche suffisamment pure et riche en phosphates calciques pour être **employée** directement **comme** engrais phosphate.
utilisé*...[20 characters]...comme	compost: mélange de résidus divers d'origine végétale ou animale, mis en fermentation lente afin d'assurer la décomposition des matières organiques, et **utilisé comme** engrais et comme amendement.
utilisé*...[20 characters]... pour	Ce terreau [i.e. compost] est ensuite **utilisé pour** fertiliser le sol et le jardin.
utilisable*...[20 characters]... comme	Le compostage utilise la fermentation aérobie des ordures ménagères en vue de la préparation d'un compost **utilisable comme** amendement en agriculture.
nécessaire à nécessaire au	oxygène: Corps gazeux diatomique (O_2), constituant le cinquième de l'atmosphère terrestre et **nécessaire à** la respiration.
a pour objet	La granulométrie **a pour objet** la mesure de la taille des particules élémentaires qui constituent les ensembles de grains [...] et la définition des fréquences statistiques des différentes tailles de grains dans l'ensemble étudié.
affecte*...[20 characters]...à affecte*...[20 characters]...au	On l'**affecte** généralement à des reboisements ou à des pâturages permanents si la teneur en argile le rend trop lourd.
essentiel*...[20 characters]...à essentiel*...[20 characters]...au	...elements **essentiels à** la croissance des plantes. (In other words, some elements have the function of being indispensible building blocks for plant growth.)

3.3 PRELIMINARY TESTING

3.3.1 Choosing which terms to use for pre-testing the patterns

Even though 76 term records (for the 68 different terms) seems like a very small sample with respect to TERMIUM, analysing all occurrences of all the terms, for all the relations, in my corpora would have been too time-consuming. For example, the French term *compost* alone occurs 259 times. Therefore, for the purposes of this thesis, I settled on approximately five terms for each relation in each of French and English. This would give me a potential of 30 tests to perform. (Note, however, that some terms were tested for more than one relation.)

Determining which particular terms to use for each relation and language involved several steps. First, I used a concordance program to search for all sentences in my corpora that contained the 68 terms. Out of these sentences, I wrote down those which involved the terms in a hyponymic, meronymic or functionality relation. In this way, I had a record of all relevant sentences in the corpora. I then chose those terms that yielded the greatest number of sentences for a given relation type. For example, the term *manure* occurred 14 times for the relation of meronymy. Figure 4 is a list of all the terms chosen for the testing:



ENGLISH TERMS

Hyponymy	Meronymy	Functionality
bacteria	bacteria	compost
composting	carbon	litter
fungi	compost	yard wastes
weeds	manure	
yard wastes	nitrogen	
	nutrients	

FRENCH TERMS

Hyponymy	Meronymy	Functionality
carbone	azote	compost
compost	bactéries	matières organiques
gazon	carbone	
matières organiques	fumier	
vermicompostage	matières organiques	
	trous d'aération	

Figure 4—Terms used in the testing of patterns

3.3.2 Pre-testing the linguistic patterns

Since the TA is still in the experimental stage, it was necessary to pre-test the Conceptual Operations function once the programmer had added to this function the linguistic patterns I had discovered in TERMIUM. I worked closely with Judy Kavanagh, the TA's programmer, during this stage of the research. We had to ensure that the TA would extract from my composting corpora those sentences that were there for any given relation. Because I already had a record of all relevant sentences in the corpora, I was able to determine which ones the TA was not extracting. To do so, I simply compared the list of sentences that *were* extracted by the TA to my complete record of relevant sentences. We then proceeded to fix the bugs that caused the TA to ignore certain relevant sentences.

Pre-testing the patterns also enabled me to determine which patterns needed to be "fine tuned". For example, inflectional variations had to be accounted for. A pattern such as "teneur en" found in TERMIUM and used as is in the testing of patterns to extract meronymy would exclude the inflectional variation "teneurs en" that appears in the French corpus. Therefore, if I had not previously noted down all relevant sentences, I would not have known that sentences containing "teneurs en" were being excluded during the final testing stage.

3.3.3 Results and analysis of pre-testing

3.3.3.1 Problems encountered and solutions

In the pre-testing, some relevant sentences were not extracted. Analysis of this problem revealed five reasons why the TA missed these sentences: case sensitivity, imperfect sentence delimiting, inflectional variations, patterns involving intervening text, and size of the search window. I explain all five below.

3.3.3.1.1 Case sensitivity

The TA's conceptual operations function is case sensitive. This meant that if the search term was *nitrogen*, sentences starting with *Nitrogen* were ignored. Hence, I performed two searches per term, one capitalized, one not[10]. In addition, we determined that the linguistic patterns themselves, when entered into the TA's programming, are case sensitive as well. As a result, any sentences in the corpus containing the search term, but starting with a capitalized

[10] The *obvious* solution is actually to reprogram the TA to be not case sensitive; however, given that the TA's developers were under time constraints in other areas of their work, reprogramming the TA for this minor issue was not immediately possible. Therefore, the most effective solution here, for my own research, was simply to perform a search for the capitalized and non-capitalized forms of my terms.

linguistic pattern, were not retrieved. The programmer solved this problem by entering into the TA both the capitalized and non-capitalized version of those patterns that could conceivably begin a sentence.

3.3.3.1.2 Sentence delimitation

Another problem stemmed from the fact that the sentence delimiting function is not 100 % accurate. Recall that this function divides the text up into sentences and numbers them. To perform this, the TA does not simply look for text that lies between two periods. That would lead to much sentence fragmentation because periods can often be found in abbreviations that appear mid-sentence, to name one possibility. Therefore, "the sentence delimiter program uses various heuristics to try and overcome these problems" (Kavanagh: 1995, 26). Nevertheless, the output is still less than perfect and fragmentation does occur at times. This inaccuracy affected my own study when a few relevant sentences (which I knew to be in the corpora) were not retrieved; they had been fragmented in such a way that the search term was in the first "sentence" and the pattern was in the second "sentence". To overcome this problem, we had to attack the symptom rather than the cause. In other words, I "ignored" the fragmentation and manually extracted the complete sentence. In an ideal situation, the programmer would have the time to improve the sentence delimiting function. However, the developers had other higher priority issues to deal with.

3.3.3.1.3 Inflectional variations

Earlier in the thesis I mentioned the need to "fine tune" the patterns to account for such things as inflectional variations. The programmer referred to this process as *tweaking*. The

example I gave earlier involved the French pattern *teneur en* and *teneurs en*. The actual pattern that I happened to find in TERMIUM was the singular. However, when we innocently programmed just the singular into the TA, the conceptual operation for meronymy missed a sentence that contained *les teneurs en*. To tell the TA that we wanted to retrieve both forms of this pattern, the programmer, when entering it into the programming, used the asterisk (*) as a wild card: *teneur* en*. The asterisk represents zero or more characters immediately following a string of characters; in this case, the string is made up of the letters t-e-n-e-u-r.

Many verbs, too, had to be tweaked. For example, the pattern *refers to* had to be entered as *refer* to* in order for the TA to extract relevant sentences that may contain the verb as *refer*, *refers*, *referring* or *referred*.

One interesting problem arose, though, when the programmer entered the meronymic verb *contain** into the TA. Obviously, we were interested in obtaining instances of *contain*, *contains*, *containing*, and *contained*. As well as these, however, we ended up with every sentence containing the search term AND the word *container*, i.e. an overwhelming number of irrelevant sentences! To eliminate this problem, we decided that, rather than indicating *contain** in the TA's programming, it was better to enter each of four forms of *contain*. A similar instance occurred for the French meronymic verb *composer*. We started by using *compos** to allow for the various conjugations...until we ALSO retrieved every sentence containing *compost* and *compostage*, giving us literally hundreds of sentences! We ended up tweaking this pattern by eliminating *compos** and entering in each of the verb conjugations we were interested in.

3.3.3.1.4 Patterns involving intervening text

Another part of the patterns that had to be tweaked was the distance between the words of a multi-word pattern such as *désigné sous le nom*. There are sentences where this pattern occurs as is. On other occasions, however, the author of the text may have written something like "...désigné *généralement* sous le nom de..." We had to allow for such intervening text in those patterns where it is possible to occur. In the list of patterns (see Figure 3), I have indicated, using the following format, the maximum amount of space that has been allotted in the TA: *désigné*...[30 characters]...sous le nom*. (Note that the TA counts spaces as characters as well.)

The difficult question, though, is how much space should actually be allotted? Is 30 characters enough for this particular pattern? Isn't it possible that a long clause could appear there? For instance, it is not inconceivable to see a sentence such as "Le concept X est désigné, selon les experts dans les domaines de la biologie et de l'anthropologie, sous le nom de Y." In this case, the 30-character maximum is exceeded, and this sentence would be ignored. Determining the answers to these questions for each pattern would require a lengthy, detailed study using huge corpora (for French and English) and a concordancing program.

3.3.3.1.5 Size of the search window

A similar consideration was necessary in determining the context horizon or "search window" for the search term itself. In other words, what is the ideal maximum distance on either side of a search term that the TA should have to look in search of a linguistic pattern? Originally, the programmer assigned the arbitrary search window of 25 characters. The character count starts at the beginning of the search word and ends at the beginning of the pattern (or vice versa,

depending on which comes first, the pattern or the word). This distance of 25 characters can be either on the left side of the search word or the right. Important to note is that, during the search, if the end (or beginning) of the sentence comes before the maximum distance has been reached, the search stops; the TA does not look beyond the boundaries of the sentence.

In the pre-testing, I started out using this arbitrary maximum of 25 characters. However, I noted that many sentences were not retrieved simply because the search word and pattern appeared farther apart than 25 characters. For example, in a sentence like the following, the term *gazon* and the meronymic pattern *riche en*, occur 59 characters apart. "Le **gazon** coupé : il a tendance à se compacter et il est très **riche en** azote...". To determine the optimal maximum search window for the present study, I looked through all the relevant sentences in my corpora and found the sentence whose composting term and semantic pattern appeared the farthest distance apart, i.e. 83 characters. I then asked the programmer to use the round number of 85 characters as the new search window size.

3.3.4 The effect of changes in search window size on retrieval effectiveness

For virtually any search for sentences expressing the given semantic relations, a combination of relevant and irrelevant sentences (i.e. hits and noise, respectively) will be retrieved. When the specified maximum distance between search word and pattern is increased, obviously more relevant sentences will be extracted. However, along with the increased number of hits also comes an increased amount of noise. Since the calculations of retrieval effectiveness (discussed in Section 3.5.1) are based on the amount of hits, noise, and misses (defined in Section

3.4.1), the ultimate figures for retrieval effectiveness may change each time there is a change in the size of the search window.

For the present study, I was able to determine the optimal search window size of 85 characters because a) my corpora are relatively small, and b) I dealt with only a small number of terms. It is likely, though, that for *each pattern* there is an optimal search window size—beyond which only noise would be retrieved. However, discovering these optimums for use in a "real" terminology project would take considerable research. It was beyond the scope and purpose of this thesis to delve any further into this issue.

3.4 FINAL TESTING

After solving the problems identified in the pre-testing, I proceeded to the final testing. It involved the same set of French and English composting terms as the pre-testing (see Figure 4).

3.4.1 Hits, misses, and noise

In order to later calculate the retrieval effectiveness of each search I performed, I needed certain data: the number of hits, noise, and misses. For each of the terms in both languages that I chose for the three relations, a selection of sentences was extracted. I analysed these sentences, determining which were *hits* (sentences containing the search term and a linguistic pattern that *did* express the relation in question) and *noise* (sentences with the search term and a linguistic pattern expressing something *other than* the relation in question).

Further to that, I had to find the *misses* (sentences containing the search word used in a given relation , but where the relation is *not* expressed with one of the linguistic patterns in the TA). Obviously, misses are not extracted because the TA would have no way of knowing they were in the corpora, since the relation was expressed by linguistic means other than what the TA was programmed with. For example, the following missed sentence uses the term *gazon* in a hyponymic relation manifested by the linguistic pattern *ou d'autre* (which is not already in the TA):

> Ou vous pouvez tout simplement reconstituer la pile de compost à l'aide d'autres herbes de tonte de **gazon ou d'autres** matières vertes.

In order to find the misses for a given term, I performed a concordance to extract all instances of that term. I then proceeded to read each sentence, looking for those where the given relation was expressed *without* one of the linguistic patterns in the TA. The misses themselves can prove useful because they may provide more patterns that could be included in a program such as the TA (such as the pattern *ou d'autre** from the sentence above).

I then compiled a list of the hits, noise, and misses for each term tested. By way of an example, Figure 5 below is the compilation of hits/noise/misses for the term *yard wastes*

HITS

1. To help meet that requirement , North Carolina passed l law that prohibits depositing organic **yard wastes such as** leaves , grass clippings , or tree trimmings in the state 's landfills .
2. Some **yard wastes , such as** wood chips , are very difficult to compost fully and are therefore not suitable for incorporation into the soil .
3. Typical **yard wastes , such as** leaves or tree bark , may contain less then 1 percent nitrogen and phosphorus (dry weight) , whereas animal wastes may contain nearly 2 percent nitrogen and even higher percentages of phosphorus and potassium

MISSES

1. Woody or "brown" yard wastes, like tree trimmings and autumn leaves, can be shredded and used as mulch around plants and on plants.
2. Food wastes, as well as green yard wastes like vegetable tops and grass clippings, can be dug into the ground.

NOISE

1. What **is the** " laziest " way to compost **yard wastes** ?

Figure 5—Hits/noise/misses from a hyponymic search using *yard wastes*

3.4.2 Coincidences

During this testing stage, I made an interesting discovery. Among the sentences that were extracted for three of my terms (*bacteria*, *carbon*, and *trous d'aération*), not only were there hits and noise, but also something that I have labelled "coincidences". By my definition, a coincidence is a sentence that contains the search word and one of the TA's linguistic patterns where the semantic relation in question is expressed not by the pattern, but some other linguistic means. Consider the following sentence:

> Nitrogen is a crucial **component of** proteins , and **bacteria** , whose biomass is over 50% protein, need plenty of nitrogen for rapid growth.

This sentence was extracted during a search for expressions of the meronymic relation involving the search word *bacteria*. It was extracted because it contains *component of*, one of the patterns

programmed into the TA. However, the meronymy that actually involves the term *bacteria* is "...bacteria, whose biomass is over 50% protein..." (meaning that bacteria are made up of 50% protein). It was simply a happy coincidence that the TA retrieved this sentence because it happened to contain *component of* within 85 characters of the search term. Although this sentence cannot truly be considered a hit for statistical purposes, in a terminology context it most assuredly is a hit; a valuable piece of meronymic information about bacteria is conveyed here. And that is the whole purpose of using a knowledge extraction tool in the first place. Nevertheless, when calculating retrieval effectiveness, I did not include the coincidences in with the hits. What is needed in the future is a way of dealing with coincidences statistically.

3.5 RESULTS AND ANALYSIS OF THE FINAL TESTING

3.5.1 Retrieval effectiveness

3.5.1.1 Recall and Precision

After all the testing was completed, it was necessary to derive some meaning from the data collected for hits, misses, and noise. In the field of information retrieval, the most common measures of retrieval effectiveness are *recall* and *precision*. According to Frakes (10), "Recall is the ratio of relevant documents retrieved for a given query over the number of relevant documents for that query in the database....Precision is the ratio of the number of relevant documents retrieved over the total number of documents retrieved." I will explain this in a way more meaningful to this thesis.

First, where the above definition says *documents* and *database*, the word *sentences* and *corpus* respectively can be substituted. Second, when we ask "What is the *recall* for a particular search?", we are asking, "Out of all the relevant sentences in the corpus, how many did our search actually retrieve?". Third, by asking, "What is the *precision*?", we are asking "Out of all the sentences that were extracted during a search, how many were hits and how many were just noise."

To determine the values for both precision and recall for each search, the following formulae are used:

$$recall = \frac{hits}{hits + misses} \qquad precision = \frac{hits}{hits + noise}$$

In the field of information retrieval, "Both recall and precision take on values between 0 and 1." (Frakes, 10). However, to make the resulting values more meaningful to the readers of this thesis, I have decided to express my own recall and precision values as percentages. For example, a precision value of .86 for a given search would be expressed in this thesis as 86%— meaning that, out of all the sentences extracted for that search, 86% of them were hits. A recall value of .55 would indicate that out of all relevant sentences in the corpus, 55% of them were actually retrieved during the search. Figure 6 below presents the recall and precision figures for each search performed.

Figure 6—Recall and Precision Values

English	French
Hyponymy	Hyponymy
bacteria recall: 4/21 = 19% precision: 4/ 7 = 57%	carbone recall: 2/2 = 100% precision: 2/4 = 50%
composting recall: 4/17 = 24% precision: 4/22 = 18%	compost: recall: 2/3 = 67% precision: 2/16 = 13%
fungi recall: 5/5 = 100% precision: 5/7 = 71%	gazon: recall: 3/5 = 60% precision: 3/4 = 75%
weeds recall: 4/6 = 67% precision: 4/6 = 67%	matières organiques: recall: 3/5 = 60% precision 3/7 = 43%
yard wastes recall: 3/5 = 60% precision 3/4 = 75%	vermicompostage recall: 3/3 = 100% precision: 3/3 = 100%
AVERAGES: recall = 54% precision = 58%	**AVERAGES: recall = 77% precision = 56%**
Meronymy	Meronymy
bacteria recall: 6/7 = 86% precision: 6/13 = 46%	azote recall: 10/11 = 91% precision: 10/10 = 100%
carbon recall: 5/15 = 33% precision: 5/11 = 45%	bactéries recall: 3/3 = 100% precision: 3/4 = 75%
compost recall: 34/50 = 68% precision: 34/128 = 27%	carbone recall: 6/6 = 100% precision: 6/6 = 100%
manure recall: 14/20 = 70% precision: 14/19 = 74%	fumier recall: 6/8 = 75% precision: 6/15 = 40%
nitrogen recall: 18/35 = 51% precision: 18/32 = 56%	matières organiques recall: 4/5 = 80% precision: 4/5 = 80%
nutrients recall: 6/9 = 67% precision: 6/12 = 50%	trous d'aération recall: 3/3 = 100% precision: 3/3 = 100%
AVERAGES: recall = 63% precision = 50%	**AVERAGES: recall = 91% precision = 83%**
Functionality	Functionality
compost recall: 4/19 = 21% precision 4/10 = 40%	compost recall: 3/13 = 23% precision: 3/ 19 = 16%
litter recall: 1/3 = 33% precision: ½ = 50%	matières organiques recall: 2/5 = 40% precision: 2/3 = 67%
yard wastes recall: 1/4 = 25% precision: 1/1 = 100%	
AVERAGES: recall = 26% precision = 63%	**AVERAGES: recall = 32% precision = 42%**

3.5.1.2 Improving recall and precision

Terminologists using a knowledge extraction program would ideally want 100% recall and 100% precision for every search. In other words, they would want to be able to retrieve all the relevant sentences in the corpus, without extracting any "junk", or noise. This is the ideal situation for *any* information retrieval task. Unfortunately, however, "recall and precision are inversely related. That is, when precision goes up, recall typically goes down and vice-versa." (Frakes, 10-11). Said another way, one can be improved, but usually at the expense of the other.

When applied to a study such as the present one, we can raise the recall value by increasing the search window size or programming the TA with more patterns for the relation in question. That would indeed bring in more relevant sentences. Along with them, however, would come more instances of noise, with a resulting decrease in the precision value. Any work done to improve precision would jeopardize recall to some degree.

3.5.1.3 Importance of recall for terminology work

This may at first seem like a losing battle. However, for terminology work, I believe there is a way to deal with this conundrum. I would propose that the focus be on finding ways to improve recall rather than precision. High recall values mean that most of the relevant sentences are being extracted from the corpus for a given search. There would be a lot of noise too, but it is *infinitely* easier to ignore a sentence that proves useless than it is to read through the entire corpus looking for relevant sentences that had been missed. Having to do the latter would defeat the purpose of using a knowledge extraction program in the first place.

3.5.1.4 Nature of statistics

As Mark Twain wrote in his autobiography, "There are three kinds of lies: lies, damned lies, and statistics." While I myself do not see statistics as lies, I do believe that they can be misleading. They can be manipulated to exaggerate or downplay the importance of what they are measuring. Therefore, in the interest of honesty, I believe it is important to include here a brief discussion about the precision and recall statistics from my own study.

The reader will note, upon perusing the statistics, that some searches resulted in promising numbers. For example, the search for meronymy using the French term *bactéries* shows a recall value of 100% and a precision value of 75%. This is very impressive (and misleading) on its own. However, a look at the actual number of sentences involved gives a different picture. First, there were only three relevant sentences in the entire corpus with the meronymic relation involving *bactéries*. The TA did retrieve them all, but that is not a guarantee that it would have done the same if the corpus had contained 325 such relevant sentences. Therefore, the 100% recall value would be much more significant if the actual numbers had been 325 out of 325, rather than only 3 out of 3.

Second, the precision value for the same search was 75%, indicating that this search was reasonably accurate (not much noise). The statistics work to the program's *dis*advantage in this case. Note that three out of the four sentences retrieved were relevant; therefore, only one sentence was noise. But since only four sentences were involved, a single sentence is "worth" 25% of the total—a huge portion. If the total sentences extracted had been 125, for example, and again only one had been noise, the precision value would have been 99%!

The conclusion to be drawn from these observations is that a study dedicated solely to statistical analysis of a knowledge extraction program would have to use large corpora where it is more likely that a greater number of sentences would be extracted during each search.

Regarding the analysis of small amounts of data, Carl Sagan lists "statistics of small numbers" (1996, 214) as one of "the most common and perilous fallacies of logic and rhetoric" (1996, 212). Drawing hard and fast conclusions from limited information is not possible, and arguments that attempt to do so are fallacious. How, then, does the present thesis (which is based on a small-scale analysis) reconcile itself with this fact? The answer is clear when readers recall the purpose of the thesis: I am not claiming to prove here and now that, without a doubt, knowledge extraction technology in its present state is perfect for terminology work. Rather, I am attempting to show the *potential* of this technology for terminology work and provide a springboard for future research in this area.

3.5.2 Effectiveness of the patterns individually

The previous section on retrieval effectiveness dealt with the patterns as a group for each relation. For example, consider the 91% recall figure for the patterns during a meronymic search for the French term *azote*. The patterns as a group retrieved 91% of all relevant sentences in the corpus. However, this figure reveals nothing about *which* of the 28 patterns in that group came into play nor about the relative efficacy of each pattern that *did* come into play. I define *efficacy*, here, as the following: the number of *hits* retrieved by a given pattern over the total number of sentences retrieved by that pattern, expressed as a percentage.

The following sections are an analysis of the output (i.e. sentences extracted) from the final testing. The purpose is to determine which of the patterns came into play and, of these, which retrieved hits and which were *too* productive by harvesting a large amount of noise. (Readers may find it helpful to refer to Figure 3 presented earlier in the thesis while they read through the analysis.) The calculation method used to determine the efficacy of individual linguistic patterns is similar to that used in the Ahmad and Fulford study (1992).

3.5.2.1 Hyponymic patterns

3.5.2.1.1 English hyponymic patterns

There were 21 English patterns[11] for hyponymy discovered and entered into the TA. However, the hyponymic searches revealed that only eight of these patterns appeared in the English corpus expressing hyponymic relations in association with the five composting terms used for the searches. Figure 7 below shows the patterns that did come into play, along with the number of hits and noise each retrieved (i.e. their productivity) and their efficacy.

[11] Inflectional variations are not counted in this figure. For example, *is the* and *are the* I have considered one pattern. The asterisk wild card could not be used in this instance and therefore the two conjugations had to be programmed individually.

Hyponymic Pattern	Hits	Noise	Total (hits + noise)	Efficacy (%) (hits / total)
and other	5	6	11	45%
especially	1	2	3	33%
includ*	1	1	2	50%
is a	1	4	5	20%
is the / are the	2	5	7	29%
is / are classified as	1	0	1	100%
such as	7	4	11	64%
(parentheses)	2	4	6	33%

Figure 7—Productivity and efficacy of English hyponymic patterns

The two most efficient patterns is this case were *is/are classified as* and *such as*, meaning that they harvested proportionately more hits than noise. However in terms of "workload", the two patterns that brought in the highest quantity of useful information are *and other* and *such as*. The pattern *is a* produced the poorest results. This is to be expected, though, because this pattern is very versatile in real language. For example, it can act as an "=" sign in subjective completion: "A computer is a girl's best friend." The *is a* pattern is also used in many sentences expressing non-useful hyponymic relations. Consider the following sentence: "Windows 95 is a good Christmas present." This sentence would be useful only to someone preparing a hierarchy of gifts[12]. Researchers looking to find information on Windows 95 as a piece of technology would consider this sentence as noise.

---

[12] And, as such, raises the issue of multi-dimensionality, which is beyond the scope of this thesis. Those interested in learning more about multi-dimensionality will find Lynne Bowker's M.A. thesis an excellent springboard into the subject: "Guidelines for Handling Multidimensionality in a Terminological Knowledge Base." School of Translation and Interpretation, University of Ottawa, 1992.

Interesting to note is the pattern *and other*. Sentences extracted by this pattern have to be read with care so as not to be misunderstood. This will become clear upon consideration of the following sentence, which was retrieved in the search for English hyponymy using *bacteria* as the search word:

> Some species scavenge on decaying vegetation, some feed on **bacteria**, fungi, protozoa **and other** nematodes, and some suck the juices of plant roots, especially root vegetables.

This sentence is noise, not a hit. It appears to *imply* that bacteria, fungi and protozoa are types of nematodes, but in *reality* this is not the case. The importance of context is demonstrated here, because the *some species* at the beginning of the sentence refers to *some species of nematodes*. The terminologist would be aware of this only by reading the preceding sentence in the corpus. The piece of information relayed in the sample sentence is that some species of nematodes eat bacteria, fungi and protozoa, and are sometimes "cannibalistic", eating others of their own kind, i.e. other nematodes.

Another issue revealed by this is the problem of dealing with a subject field that is unfamiliar to the researcher. A sentence such as "...grass, leaves, branches and other animals..." would pose no problem. Being familiar with these concepts prevents us from assuming that grass, leaves and branches are kinds of animals. But how is a terminologist unfamiliar with biology to know that bacteria, fungi and protozoa are NOT kinds of nematodes, even though the sentence seems to say that? The solution is for terminologists to continually rely on the surrounding text in the corpus and not just on the words immediately before and after the semantic pattern to avoid misreading this type of sentence.

Regarding the patterns that did not happen to come into play, it would be wrong to conclude that they are useless. They simply did not happen to appear in association with the given search words in this particular corpus. Therefore, conclusions can only be drawn from data on those that did appear.

3.5.2.1.2 French hyponymic patterns

For French hyponymic patterns, 24 were found during the analysis of the TERMIUM records and programmed into the TA. Of these, only seven came into play during a search for hyponymy. Figure 8 below shows the statistics for these seven patterns.

Hyponymic Pattern	**Hits**	**Noise**	**Total** (hits + noise)	**Efficacy (%)** (hits / total)
comme	3	4	7	43%
est un*	5	5	10	50%
est l*	0	5	5	0%
forme*	1	1	2	50%
il s'agit d*	1	0	1	100%
sont les / sont des	1	2	3	33%
(parentheses)	2	3	5	40%

Figure 8—Productivity and efficacy of French hyponymic patterns

These statistics show that *il s'agit d** had the best precision, retrieving no noise in this study. However, the patterns that brought in the most information were *comme* and *est un**. These two also harvested a proportionately high quantity of noise. Recall, though, that noise can be ignored; what terminologists need is as much relevant information as possible. Interesting to note is that

the pattern *est l** in this study retrieved nothing *but* noise. Yet I hesitate to conclude that this patterns is useless; it is simply another instance revealing that a study on a much larger corpus would produce more telling data than in the present study.

3.5.2.2. Meronymic patterns

3.5.2.2.1 English meronymic patterns

Of the 14 English meronymic patterns programmed into the TA, nine brought in results. Figure 9 presents the data for these patterns.

Meronymic Pattern	Hits	Noise	Total (hits + noise)	Efficacy (%) (hits / total)
combin*	0	1	1	0%
component of	1	0	1	100%
composed of	0	1	1	0%
consist*	0	1	1	0%
constituent* of	1	0	1	100%
contain/contains/containing/contained	25	2	27	93%
in	49	93	142	35%
mak*...up	1	0	1	100%
with	8	35	43	17%

Figure 9—Productivity and efficacy of English meronymic patterns

From these statistics, we see that three of these patterns have a "perfect score" of 100%: *component of*, *constituent* of* and *mak*...up*. In each case, only one sentence was retrieved and was a hit. These three patterns have the potential to be useful.

Interesting to note, however, are the three patterns that earned a 0% score: *combin**, *composed of*, and *consist**. Our language instincts tell us that these patterns are "very meronymic", but the statistics above seem to imply that there are other relations they can express. However, when we look at the actual sentences that these patterns appeared in, we see that they do indeed express meronymy, but that the meronymic relation does not involve the search word. And that is why the sentence was noise and not a hit. For example, the following sentence was retrieved in a meronymic search using *compost* as the search word:

> Garden suppliers sell **compost** starters or "activators", often **composed of** high-nitrogen fertilizers.

Strictly speaking, this sentence is noise because I wanted part-whole relations involving *compost*. The sentence was harvested because the search word and the pattern *composed of* occurred within 85 characters of each other. However, the actual meronymic relation involves the term *compost starter* (or *compost activator*), which apparently contains high-nitrogen fertilizers. In a terminological context, this sentence could be classified as a hit; *compost starter* is a bona fide term in the field of composting, and as such, terminologists would need information on this concept, including what it is made of.

The statistics above show that an extremely effective pattern was *contain* and all its conjugations: out of 27 sentences retrieved, 25 were hits and only 2 were noise. Two conclusions can be drawn from this. First, *contain* is used very frequently in real language to express meronymy. Second, it seems to be specific to meronymy, i.e. it is not used to express other relations. In fact, the two "noise" sentences that this pattern retrieved did express a part-whole

relation, but just not involving the search word. For example, the following sentence was extracted in the meronymic search for *compost*:

> Raw materials suspected of **containing** significant pesticide residuals should be withheld from **compost** for horticultural uses.

The word *containing* is indeed expressing meronymy, but the term *compost* is not part of the meronymic relation.

The last of the English patterns to be discussed are the two tiny, yet ubiquitous pronouns *in* and *with*. These words were highly productive in that they brought in many sentences. However, they are extremely versatile. English uses them in many different ways other than to express meronymy; hence their great potential for extracting noise.

The pronoun *in* yielded an efficacy rate of only 35%. What I discovered, during the analysis of the sentences harvested by this word, is that it frequently occurs as part of a longer unit. For example, in all eight hits extracted during a search using *nitrogen* as the search word, the word *in* was part of the following multi-word units: *rich in*, *high in*, and *low in*. A sample sentence is as follows: "Young, green plants are very high in nitrogen." The same phenomenon is true for the three hits involving *in* and the search word *carbon*. And looking at the hits involving *in* and the search word *bacteria* reveals two more potential multi-word patterns: *present in* and *found in*. Other similar patterns discovered are *appear in*, *stored in*, and *live in*. It may be tempting to conclude that *in* should not stand on its own as a pattern in a knowledge retrieval program. However, many hits did contain *in* by itself. Consider, for example, the following sentence extracted with *compost* as the search word:

> Information on the amounts of metals such as zinc, iron, lead, nickel, and cadmium **in** the **compost** should also be studied.

Granted, the pattern *in* alone generates much noise. One interesting instance of noise was this sentence: "Is there a compost pile in your past?" This is cute, but constitutes noise regardless of how one looks at it. As my own statistics show: only 35% of the sentences extracted were hits, meaning that 65% were noise. A more in-depth study is necessary to determine if multi-word patterns with *in* could bring in enough information that *in* by itself could be dropped.

The other productive pronoun is *with*, which scored even lower than *in* in terms of efficacy: only 17% of the sentences retrieved were hits. My analysis of the output sentences showed that *with* is even more versatile than *in*. The word *with* can mean many things, including, but not limited to, the following: *accompaniment* ('Mix the compost with the potting soil'), *means* ('hit the nail with a hammer'), *manner* ('with care'), *has as parts* ('a roof with guttering is recommended'), and *has as a characteristic or property* (materials with high carbon-to-nitrogen ratios). Obviously, for a meronymic search, researchers are interested in *with* only when it means *has as parts*, and possibly when it means *accompaniment*. To understand the why the latter may be acceptable, consider the following sentence extracted from the English corpus: "Another thing that often comes with manure is a supply of weed seeds." By virtue of having been eaten by the animal in question, weed seeds may be part of the manure.

I tend to remain pessimistic about the usefulness of *with* as a pattern for knowledge extraction. Given the extremely low amount of information it brought in (17%; hence 83% noise) and the fact that the hits themselves did not yield much information, I would propose dropping *with* entirely. An example of a hit containing *with* (where *nitrogen* was the search word) is the

following: "...materials with a high nitrogen content, such as fresh grass clippings..." While this does tell the researcher that grass clippings contain nitrogen, it is actually the word *content* that has the greater meronymic power, and as such, this sentence would have been extracted if *content* had been one of the patterns[13]. In short, eliminating *with* as a pattern would sacrifice only a minuscule amount of knowledge, while vastly reducing the amount of noise harvested during a meronymic search.

3.5.2.2.2 French meronymic patterns

Of the 28 French meronymic patterns discovered and programmed into the TA, only 10 came into play. Figure 10 below indicates the productivity and efficacy of those 10.

Meronymic Pattern	**Hits**	**Noise**	**Total** (hits + noise)	**Efficacy (%)** (hits / total)
avec	0	6	6	0%
ayant	1	0	1	100%
comprend	1	0	1	100%
constitu*	2	2	4	50%
contenir	1	0	1	100%
contien*	4	0	4	100%
fraction...d*	1	0	1	100%
muni*...d*	3	1	4	75%
riche*...en	5	0	5	100%
teneur* en	13	2	15	87%

Figure 10—Productivity and efficacy of French meronymic patterns

[13] Section 3.5.4 covers the subject of further potential patterns to be considered for addition to the TA—and the pattern *content* is one of these potential additions.

The efficacy percentages for these patterns look promising. In general, what they are saying is that these patterns bring in much useful information without dragging in a mountain of noise. The pattern *teneur*...en* in particular was very productive, with 13 hits and only two instances of noise. The proportionately low amount of noise seems to indicate that *teneur*...en* is specific to meronymy. In fact, the two instances of noise were indeed expressions of meronymy, but the meronymic relation simply did not involve the search word.

The next productive patterns, in descending order, were *riche*...en*, *contien** and *muni*...d**, none of which brought in any noise. It may be assumed from this that these patterns, like *teneur*...en,* are specific to meronymy and do not have a number of meanings outside the meronymic one.

The pattern *constitu**, however, proves to be less reliable. This was owing to the fact that one possible form of this pattern is *constitue* when it means *représente* (or simply *est*). As such, hyponymy is being expressed. Indeed, the two instances of noise in my study exhibited this very phenomenon:

> 1. Le fumier composté **constitue** un produit stable et hygénique, plus facile et sécuritaire à manipuler...que le fumier non composté."
>
> 2. L'entreposage du fumier pendant l'hiver...**constitue** donc une étape qui ne peut généralement pas être éliminée."

There were four patterns that each brought in one hit and no noise , and thus had an efficacy rate of 100%. From my own intuition of the French language, I would claim that three of them (*fraction...d**, *contenir*, and *comprend*) have the potential to be useful. However, I am less confident about the pattern *ayant*, which, like the English *with* is a very versatile word. In fact, it

may have been stretching the limits of meronymy to include as a hit the one sentence that was extracted by *ayant*. It is as follows:

> Le compostage du fumier nécessite d'abord une bonne régie à l'étable afin d'obtenir un fumier ayant de bonnes propriétés tout en optimisant la conservation des éléments fertilisants.

Under a stricter definition of meronymy, this sentence could safely be counted as noise, thereby increasing the total noise for this pattern to two instances and reducing the efficacy rate to 0%. If, in a subsequent study, this pattern is found to bring in a proportionately high amount of noise, it may be advisable to eliminate this pattern.

Turning to the pattern *avec*, we see that the efficacy rate was 0%—only noise was retrieved in my study. An analysis of the instances of noise revealed that the relations being expressed with *avec* were *accompaniment* and *means*, and not *has as parts*. A sample sentence is as follows: "Recouvrez les matières organiques **avec** le compost existant ou **avec** de la terre." Based, then, on the previous analysis of English *with* and the assumption that *with* and *avec* are used in a more or less comparable manner in their respective languages, it may be safe to assume that *avec*, too, can be considered for elimination as a pattern for computerized knowledge extraction.

3.5.2.3 Functionality patterns

3.5.2.3.1 English functionality patterns

On an individual basis, the English functionality patterns turned out to be somewhat disappointing. First, only four out of the eleven patterns in the TA came into play. Second, three

of them proved to be useless in my study. Figure 11 below shows the productivity and efficacy for these four patterns.

Functionality Pattern	Hits	Noise	Total (hits + noise)	Efficacy (%) (hits / total)
essential to	0	1	1	0%
functions of	0	1	1	0%
in order to	0	1	1	0%
used	6	6	12	50%

Figure 11—Productivity and efficacy of English functionality patterns

The sentence retrieved by the pattern *essential to* was noise in the present study. However, the semantic relation expressed in the sentence was indeed functionality; it was simply the case that the relation did not involve the search word *compost*:

> "...oxygen diffusion through the smaller pores and into the aqueous film surrounding **compost** particles is **essential** to maintaining aerobic conditions for the active microorganisms."

It is the functionality of *oxygen diffusion*, or its role in the composting process, that is explained in this sentence. It is clear that the pattern *essential to* does indeed have potential for extracting knowledge for the semantic relation for functionality. Therefore, it should not be dropped, in spite of its statistically poor results in the present study.

A similar phenomenon involves the pattern *functions of*; the sentence it extracted expresses functionality, but just not with the search word *compost*. The sentence is as follows:

> "One of the principal **functions of** mixing and turning the compost is to redistribute moisture to minimize this preferential airflow and the nonuniform decomposition that results."

Therefore, the pattern *functions of* is a potentially useful pattern.

Regarding the pattern *in order to*, its ability to express functionality is not immediately obvious. That, combined with its apparent poor performance in the present study, may lead to a hasty dismissal of this pattern. However, a closer look at a) the original sentence in TERMIUM in which I found the pattern and b) the sentence it extracted in the testing will reveal this pattern's potential for expressing functionality. The following is the TERMIUM sentence defining the term *composting bin*:

> "A plastic container into which people recuperate vegetable waste, in order to recycle it and get natural fertilizers or soil conditioners."

The pattern *in order to* can be replaced by the words *for the purpose of*, thereby making it obvious linguistically that it is the purpose, or function, of a composting bin that is described. Indeed, the sentence can even be re-worded to make this more evident still, without changing the meaning: "The function of a composting bin is to recycle recuperated vegetable waste into natural fertilizers or soil conditioners."

The next sentence is the one extracted during the testing part of the present study:

> "You will still need to dry some of the compost **in order to** measure the moisture content..."

This sentence answers the question "What is the purpose (or function) of drying some compost", the function being that drying allows you take a moisture content reading. Even though *drying some of the compost* is not actually a term from this subject field, the sentence nevertheless shows that the pattern *in order to* can express functionality and therefore has potential as a pattern for knowledge extraction.

The final pattern that came into play and was most productive in the present study was *used*. The fact that its efficacy rate was only 50% shows that this pattern has the potential to harvest much noise along with the useful sentences. An analysis of the extracted sentences reveals that perhaps the word *used* by itself should not be the pattern, but rather it may be better to include an accompanying particle. Out of the six hits, three of those sentences have *used* occurring with *as*, as in the following sentence: "Mortality compost can be **used as** a nutrient source for crops...." The patterns *used for*, *used in* and *used on* occurred in the rest of the hits. The instances of noise contain *used* in the multi-word units *used by*, *used up*, *used in*, *used to*, as well as *used* on its own.

A more exhaustive study would be better able to determine the usefulness of these variations on the theme of *used*, but I believe it is safe to conclude that *used* should be part of multi-word patterns with specific particles in order to reduce the amount of noise.

3.5.2.3.2 French functionality patterns

Of the 14 French functionality patterns programmed into the TA, four came into play. The table displaying their productivity and efficacy values appears below:

Functionality Pattern	**Hits**	**Noise**	**Total** (hits + noise)	**Efficacy (%)** (hits / total)
fonction	1	1	2	50%
permet*	0	2	2	0%
serv*...à / serv*...au	2	1	3	67%
utilis*	2	13	15	13%

Figure 12—Productivity and efficacy of French functionality patterns

The two "top performers" were *fonction* and *serv*...à / serv*...au*. The hit retrieved by *fonction* is expressing the functionality of (or at least *one* function of) *compost* through analogy:

> "Une quantité abondante de compost ajoutée au sol fera **fonction** d'éponge, absorbant l'eau de la pluie et la relâchant lors des secheresses."

The instance of noise (a title) retrieved clearly shows another use of the word *fonction* that does not express functionality: "Évolution du volume des piles de **compost** en **fonction** du temps." This is a mathematical use of *fonction*; English *function* is used in a similar manner. Nevertheless, *fonction* has the potential as a knowledge extraction pattern.

The pattern *serv*...à / serv*...au* appeared to be particularly effective, without harvesting too much noise. The instance of noise in this case, though, does express function, but simply not involving the search word *compost*: "Les sac de feuilles **servent** souvent de pare-vents pour les bacs à compost." Therefore, this pattern, along with its related variation *sert*...à / sert*...au / sert*...pour*, will prove useful for extracting functionality knowledge.

Concerning the pattern *permet**, it yielded poor results, as seen in Figure 12 above. Nevertheless, I believe that this pattern does have potential, in light of the clear instance of functionality evidenced in the sentence extracted from TERMIUM: "Un digesteur est un appareil **permettant** la fermentation anaérobie de la biomasse." The word *permettant* can be adequately replaced here by the words *utilisé pour* without any loss of meaning.

The final French functionality pattern to be discussed is *utilis**, whose results are much poorer than one may have expected. As with the corresponding English pattern *used*, the French pattern would probably bring in much less noise if accompanied by a particle. Indeed, analysing

the 13 instances of noise involving *utilis** revealed that each instance contained this pattern on its own, appearing as either *utiliser*, *utilisez* or *utilisé*. One noise sample is as follows:

> "Lorsque le compost est prêt à être **utilisé**, la température de la pile n'augmentera pas à plus de 43°C, quelle que soit la fréquence du retournage."

Three of the patterns programmed in the TA do have an accompanying particle: *utilisé*...comme*, *utilisé*...pour*, and *utilisable*...comme*. They did not happen to come into play in the present study, though, because they are too limiting; the first word in each of these multi-word patterns should be shortened to *utilis**. Indeed, one of the two hits obtained involved the words *utilisez pour*. (The second hit would not have been harvested without *utilis** on its own, but it is debatable whether it should have even been classified as a hit anyway, so little information does it bring in: "La terre peut **utiliser** toutes les matières organiques que vous pouvez lui fournir...")

3.5.3 Quality of the hits

At this point, it is necessary to dispel the illusion that all hits are created equal. Up to now, I have spoken of hits as absolutes, as if all were good. This is not the case. Not every hit obtained will necessarily be a perfect (i.e. a complete) expression of the relation in question for a given search term. The nature of real language in use is such that an author will provide as much information in a single explanatory sentence as required at that moment. Later in the same paragraph or page, the author may add another explanatory sentence, providing the reader with additional information as it becomes pertinent to do so.

The implication of this for terminology is that terminologists, using a knowledge extraction program, may have to build a composite profile for a term by taking chunks of

semantic information from several sentences. As such, terminologists will not discard an extracted sentence that has potential as a hit simply because it is not "perfect".

To illustrate the phenomenon of knowledge spread out over several sentences and the varying quality of hits, I will use the three hits harvested during a hyponymic search for the French term *vermicompostage*:

> 1. Le vermicompostage **est une** façon viable de composter les résidus de la cuisine à l'intérieur des bâtiments.
>
> 2. Le vermicompostage **est une** activité possible toute l'année.
>
> 3. Le vermicompostage **est une** méthode de compostage faisant appel à l'activité des vers rouges pour décomposer la matière organique.

It is evident that neither sentence 1 or 2 can stand on its own as a definition. Sentence 3 can, but that does not mean that first two should be ignored. Sentence 1 and 2 are hits because they contribute additional information about vermicomposting that expands the reader's understanding of this concept. Sentence 2 is a conclusion drawn from sentence 1: vermicomposting can be done indoors; *therefore* it can be done all year long. This information further differentiates vermicomposting from other types of composting; hence the reason why sentence 1 and 2 are hits even though, individually, neither is an adequate definition or complete expression of hyponymy.

At times, a hit may be too general to be of any real use in preparing a record for a term. For example, consider the following sentence extracted during a meronymic search using the term *carbon*:

> Understanding which materials are high **in** carbon, and which are high in nitrogen will help you build a pile with a good balance of ingredients for decomposition.

The strict meronymic information to be obtained from this sentence is "there are materials that

contain much carbon." On its own, this is too general. However, the sentence has tremendous value for expanding terminologists' understanding of the subject field; it alerts them to the fact that carbon and nitrogen are important to the composting process, and it is therefore necessary to discover what materials in particular contain these elements. Logically, the terminologist can assume that, if carbon and nitrogen are that important to composting, somewhere in the corpus will be found statements mentioning the terms for specific materials that contain carbon and nitrogen. A search for these specific terms, will lead the terminologist to two more terms: the collective term for materials high in carbon (*brown materials*) and those high in nitrogen (*green materials*).

In conclusion, sentences retrieved from a corpus by a knowledge extraction program must be carefully assessed for their potential as hits in terminology work. Care must be taken not to delete potential hits simply because they do not fit someone's definition of an ideal expression of a given semantic relation.

3.5.4 Analysis of the misses

This section will analyse a range of sentences in the corpora that were NOT retrieved by the TA, but which do express the semantic relations under study involving the search terms used. The purpose of this analysis is to determine why the TA bypassed these sentences and how the TA can be improved such that it *would* pick up these (and similar) sentences in subsequent testing, thereby improving recall. (Readers are reminded that recall is the number of relevant sentences retrieved in relation to the total number of relevant sentences in the corpus.)

In general, there were three reasons why a relevant sentence was missed. The first reason involves the search word and linguistic pattern appearing too far apart in the sentence. In other words, the maximum search window of 85 characters was exceeded. Second, the semantic relation in question was expressed by linguistic patterns *other than* those programmed into the TA. The TA was programmed with those patterns I discovered in 76 TERMIUM records; it is highly likely that analysing a greater number of records would have provided a larger list of patterns to include in the TA and reduce the number of misses. The third reason involves the semantic relations being expressed by linguistic patterns that are grammatical rather than lexical.

3.5.4.1 Distance between search word and pattern

The issue of the 85-character search window being exceeded is fairly straightforward. The following sentence was missed during a meronymic search using *carbon* as the search term:

> **Carbon** is both an energy source (note the root in our word for high energy food: carbohydrate), and the basic building block **making up** about 50 percent of the mass of microbial cells.

The term *carbon* and the pattern *mak*...up* occur 126 characters apart. The first solution that may spring to mind is to increase the search window size. It is important to remember, though, that any change in search window size will affect precision and recall. If increasing the size to a very large number results in one extra hit but 100 more instances of noise, it is hardly worth increasing the size. A more exhaustive study is needed to determine an optimum size.

3.5.4.2 Exhaustivity of linguistic patterns

There are many lexical means for expressing semantic relations. Admittedly, my limited analysis of TERMIUM uncovered only a small list of possibilities. Therefore, there was bound to exist in the corpora sentences expressing hyponymy, meronymy and functionality by means of patterns different than those I discovered. The following sentence, which the TA missed during a hyponymic search for the French word *gazon*, is an example:

> Ou vous pouvez tout simplement reconstituer la pile de compost à l'aide d'autres herbes de tonte de gazon **ou d'autres** matières vertes.

Because of the pattern *ou d'autres*, the terminologist learns that *(tonte de) gazon* is a type of *matière verte*, an obvious expression of hyponymy. Therefore, *ou d'autres* (and, I hypothesize, the pattern *et d'autres*) could be added to the TA in order to retrieve all sentences expressing hyponymy in this way. Figure 13 below is a list of potential new linguistic patterns, discovered by analysing the sentences missed by the TA.

Hyponymy	Meronymy	Functionality
another method method* of method*...for methods...: (colon is important) and various other like (= such as)	content* source ingredients built out of proportion* of high- -rich its (possessive) reserve of	act* like utilized as serve* as
****** c'est de / c'est un* ou d'autres / et d'autres par exemple	****** source*...de	****** l'utilisation

Figure 13—Possible new French and English patterns, as found in the misses

3.5.4.3 Grammatical vs lexical patterns

The issue of lexical versus grammatical patterns did not arise earlier in this thesis because only an analysis of the misses brings it to light. Lexical patterns are words or word combinations whose meaning expresses a given semantic relation, as seen in all the examples given so far. However, there are instances where the underlying *part of speech*, rather than the actual *lexical item*, plays a large role in the expression of that relation. Consider the following sentence from the English corpus:

> A vile smell around the compost tells you that anaerobic bacteria are moving in, and the pile may simply need to breathe.

We can ask the hyponymic question "What *kind* of bacteria?" The answer is "*anaerobic* bacteria". Obviously the word *anaerobic* indicates the kind of bacteria in question. However, it would be virtually useless to enter this word as a pattern for future knowledge extraction because how many other hyponyms in English are classified by their "anaerobicness"? The "secret formula" for extracting this kind of knowledge involves using the underlying part of speech, in this case the adjective. What is needed is a computer program that can find occurrences of the search word accompanied by a preceding adjective. Hyponymy lends itself well to this particular grammatical pattern. Naturally, this type of retrieving would result in some noise as well, such as sentences discussing, say, "interesting" bacteria or "strange" bacteria.

Functionality is frequently expressed by means of a term+verb construction. The following sample sentence is taken from the French corpus:

> Le **compost améliorera** la texture des sols argileux et sablonneux et leur redonnera les éléments nutritifs essentiels.

This sentence clearly describes one of the functions of compost, but without an explicit marker such as "est utilisé pour". The way to retrieve this type of sentence is to enable a computer program to find sentences where the search word is the subject of a following verb. The program will also have to allow for words intervening between the search word and its verb. The following functionality sentence shows why:

> Le **compost** de couleur foncée **attire** la chaleur du soleil qui rechauffe le sol du jardin, et qui prolonge de quelques jours notre courte saison de croissance.

It is conceivable that an entire clause, or even several, can be appear between the subject and its verb. Research is needed to determine how far apart they can occur.

In conclusion, the sentences missed by the TA in the present study have revealed a number of ways to enhance the retrieving capabilities of the TA, and thereby improve future recall values.

3.5.5 Walkthrough with one term

This section provides a practical illustration of the potential usefulness of knowledge extraction technology for terminology work.

First, let us examine what terminologists would have to go through *without* the aid of a knowledge extraction program. The paper-based method of scanning text in terminology work is an exceedingly long way of finding all sentences containing a given term. In addition, it is error-prone because human concentration is not perfect; there is always the possibility that the terminologist will (among other errors) overlook perfectly valid sentences. It is even likely that terminologists may have to go through their documentation more than once. Using a concordance program on electronic text instead greatly reduces the time spent on scanning and the chance of

overlooking occurrences of a given term; all instances of a term and the sentences in which they appear can be retrieved from a corpus and displayed for the terminologist in a matter of seconds.

However, a concordance program does not discriminate. It obviously retrieves useful and useless sentences alike. Faced with a lengthy list of retrieved sentences for a particular term, terminologists still must read every single one to determine which ones they will use in their work. To reduce even further the time that terminologists spend in their work, a knowledge extraction tool can perform the task that concordancers cannot: it can perform a sort of triage on the corpus sentences containing a specified term and offer up the knowledge-rich ones while passing over the useless ones.

For demonstration purposes, I will use the English term *compost* from my own research in a brief comparative analysis of scanning using a concordancer versus the TA, a knowledge extraction program.

First of all, a concordance of the search term *compost* provided 470 sentences---and this was from a corpus of 27,065 words, or 86 pages of text. Assuming it takes a terminologist 15 seconds to read each sentence and determine whether to accept or reject it, he/she would require about two hours to get through 470 sentences. Note that this time figure is ideal in that it does not account for *any* pauses whatsoever on the part of the terminologist. To allow for the "humanness" of the terminologist, it is not unreasonable to add another 30 or 40 minutes to the original two hours. And this is for a single term only.

Using the TA, on the other hand, resulted in a manageable list of just 128 potential knowledge-rich sentences for the term *compost*. The terminologist would still have to read

through each of them, but the time needed to do so is now reduced to only 30 minutes (an 80% time saving)---a significant improvement indeed! This 80% saving becomes more significant when larger corpora are involved. Work that would take five days for terminologist to get through using a concordancer would be reduced to only a single day using a knowledge extraction tool. In addition, recall that the TA is in the experimental stage; system enhancements would reduce the amount of noise and increase the number of knowledge-rich sentences extracted.

Thus, terminologists, rather than "wasting" their time on simply separating the linguistic wheat from the chaff (a necessary process in paper-based or concordance-based work) could devote this huge time-saving to improving the quality of their term records and increasing production.

3.6 CONCLUSIONS

3.6.1 Suggestions for future research

3.6.1.1 The problem with ambiguity

The preceding study may give the impression that semantic relations are clear-cut categories or that a linguistic pattern used to express a particular relation is used to express *only* that relation. However, neither of the above is true. Semantic relations and linguistic patterns are prone to ambiguity.

First, not all researchers agree on what constitutes meronymy, for example. Cruse (1986, 175) states that "Entities such as groups, classes and collections stand in relations which resemble meronymy with their constituent elements." Winston *et al* (1987, 423), however, have this to say:

"Collections must be distinguished from classes. The class-member relation is not a meronymic relation because it is not expressed by "part" but by "is"..."

An example from my own research will help illustrate the confusion that sometimes arises when trying to determine what relation is being expressed in a given sentence. The following two sentences appear on a single TERMIUM record and are describing the same phenomenon. However, one seems to be expressing hyponymy, while the other, meronymy:

> a) The earth's surface (the material that is to be classified) is divided into soil and nonsoil.
>
> b) On distingue deux sortes de matériaux à la surface de la terre (objet de la classification): le sol et le non-sol.

The first seems meronymic: both soil and non-soil are *parts of* the earth's surface. The second is hyponymic: soil and non-soil are *kinds of* surface material.

Does this mean, then, that a relation that appears to hold between two concepts is not inherent and absolute, but rather, is dependent on point of view or the way the sentence is worded? To answer this question, much detailed research is needed to work out the subtleties of language (and maybe even those of human perception).

Then comes the problem of categorizing linguistic patterns, which can also be ambiguous. If we do decide that sentence a) and b) above are both expressing hyponymy, then the pattern *divided into* (and its inflectional variants *divide into*, *division into*, etc.) is a hyponymic pattern. What happens if we then encounter a sentence such as the following one from TERMIUM:

> [Insects are] small arthropod animals characterized, in the adult state, by division of the body into head, thorax, and abdomen, three pairs of legs on the thorax, and, usually, two pairs of membranous wings.

This is very obviously expressing meronymy, i.e. the parts of an insect. Conclusion: the pattern *divid* into* is used in real language to express both hyponymy and meronymy. A knowledge extraction program would have to include the pattern in both places. As a result, a search for hyponymy will necessarily retrieve, as well, all the meronymic expressions using *divid* into*, which would constitute noise. This is an unavoidable situation until researchers are able (if possible) to tease out the subtle differences between meronymic and hyponymic sentences using this pattern. The same can be said for other areas of relation overlap. I believe that this would imply a knowledge extraction program capable of performing some amount of semantic analysis—a step far beyond simple character-string recognition.

Related to the above problem involves assuming that a given well-known pattern is used to express only those relations we normally associate it with. For example, what relation does the pattern *define* express? The automatic answer is probably *hyponymy*, as in "Geotextiles are defined as permeable textiles used in conjunction with soils or rocks" (*genus* plus *differentiae specificae*).

This answer is not wrong. But neither is it totally right, because the word *define* is not as "clean" as we would like; it is potentially ambiguous. For example, in the context of Java programming language, "to define" means "to create", "to call into existence". Classes and interfaces are *defined* within packages. This meaning of *define* is similar to that seen in, say, WordPerfect. When users wants to create columns in a document, the columns must first be "defined", i.e. created by the user specifying how many, their width, type, etc. The relation

expressed in this case is that of "creation" or perhaps "generation": two concepts are related by the fact that one creates the other.

The conclusion is that, across subject fields, some patterns can "switch" categories and end up expressing a different semantic relation. Again, much research is needed on large amounts of text in various subject fields to determine which patterns are relatively fixed and which are ambiguous.

The above discussion about relation overlap tends to paint a negative picture of the fuzziness among relations and patterns expressing them. This fuzziness is not really a major concern in a terminology context, however, because regardless of *how* sentences expressing semantic relations are extracted, they will be useful anyway. For example, a terminologist performs a meronymic search for *bacteria*. If the pattern *part of* is programmed into the knowledge extraction program, a sentence such as "Bacteria are part of the animal kingdom" may be retrieved. This sentence is expressing class membership; hence *hyponymy*. This is valuable information for the terminologist, who is not about to discard it simply because it was harvested during a search for *meronymy*.

3.6.1.2 Lexical vs Grammatical patterns

Up to this point, the work for this thesis concentrated on linguistic patterns that are *lexical*, i.e. words (or character strings) that express particular relations. Another area for further study is *grammatical* patterns—those parts of speech in certain positions in a sentence that can express those relations. Consider the following sentence:

> The oil filter removes abrasive particles that enter the lubrication system before they can cause excessive wear.

It is clear that the oil filter's function is being explained. However, there are no lexical patterns in this sentence that should be entered into a knowledge extraction program; it is the underlying part of speech that performs the role of expressing function: term + action verb.

Chapter 4, Section 4.2 of this thesis presents a brief exploration into using grammatical patterns for knowledge extraction, but is by no means exhaustive.

3.6.2 General conclusions

The first conclusion to be drawn from my research, involving an experimental program called the Text Analyzer, is that knowledge extraction technology can play an important role in the field of terminology. I have been able to demonstrate, by partially simulating a terminology project, that at least some of the conceptual analysis can be semi-automated. The focus for future research on this technology and other systems should be on obtaining optimal recall values by discovering as many lexical and grammatical patterns as possible for each relation and determining the best search window size for each pattern.

The second conclusion is the following: Before humans can program a computer with the "knowledge" required for effective, reliable text analysis, we will need to do a great deal of research into the intricacies and subtleties of language so we ourselves can fully understand semantic relations in general and how they are expressed in real language in particular.

Chapter 4 ENHANCING THE TEXT ANALYZER

4.0 INTRODUCTION

The preceding chapter focussed on discovering and testing a selection of lexical patterns for knowledge extraction using the Text Analyzer. By way of a brief follow-up, the present chapter describes subsequent enhancements made to the TA's English lexical patterns and explores knowledge extraction using grammatical patterns.

Regarding lexical patterns, the TA's "Conceptual Operations" function was enhanced with patterns discovered during an analysis of the misses from the previous testing. After the programmer added these new patterns to the TA, a new set of knowledge extraction tests was performed using a subset of the original English terms and the same corpus from the previous testing. New precision and recall statistics were prepared and compared with the first statistics to determine if the enhancements improved the effectiveness of the English "Conceptual Operations" function.

Regarding grammatical patterns, the TA has the ability to extract sentences containing adjectives or verbs associated with a given search word. These operations do not require specific patterns to be entered into the TA's programming code. To prepare the corpus for grammatical searches, the user simply runs it through the TA's part-of-speech tagger. Two tests were performed on the English corpus: a hyponymic search using the "search for adjectives in the same sentence" operation and a functionality search using the "verbs following the search word" operation. Statistics were prepared for these two tests to determine the effectiveness of these two grammatical methods of knowledge extraction.

4.1 ENHANCING THE TA'S LEXICAL PATTERNS

The purpose of further enhancing and testing the TA's "Conceptual Operations" function was to "push" the TA as far as possible, i.e. to attempt to improve its knowledge extraction effectiveness with lexical patterns within the context of this study. The implication is that, if improvements in effectiveness can be obtained with minor enhancements, future comprehensive, in-depth work in this area will lead to major improvements. This in turn would bring knowledge extraction technology closer to being implemented in terminology work.

4.1.1 Delimiting the follow-up testing

Given that this enhancement work was intended to be simply a brief follow-up to the previous work, I focussed only on the English corpus and terms. In addition, I dealt with only a subset of the 13 English terms used in the original testing.

The lexical patterns used in the previous testing were obtained from 68 TERMIUM records. The 15 *new* patterns used to further enhance the TA were discovered from analysing the misses from the previous testing[14]. I worked with eight out of the 13 original English terms since it was the misses from these eight terms that provided the 15 new lexical patterns. Therefore, subsequent testing with these terms could potentially show improvement in effectiveness. Figure 14 below shows the terms used in the new testing, and Figure 15 lists the 15 new patterns.

[14] Recall that many of the missed sentences were missed simply because they contained the search word and expression of the particular semantic relation, but this relation was expressed by lexical means *other than* those already programmed into the TA.

Hyponymy	Meronymy	Functionality
composting weeds yard wastes	carbon compost nitrogen nutrients	compost

Figure 14—Subset of English terms used for testing the TA enhancements

Hyponymy	Meronymy	Functionality
method* and various other like (= such as)	content* source* ingredient* built out of proportion* of high- -rich its (possessive) reserve of	act* like utilized as serve* as

Figure 15—New patterns discovered in the misses

4.1.2 Testing and statistics

Analysing the misses for potential new patterns was a simple process and the 15 new patterns were then added into the TA's programming code. As in the work from the previous section, the programmer and I performed a "pre-test" once the new patterns were in the TA. The purpose of this was to ensure that the misses in question were indeed being extracted by the TA and to fix any "glitches" that prevented their extraction. (Section 4.1.3 below outlines problems encountered.)

Once the problems were resolved, I performed a search for each of the eight terms. In advance, I could safely predict that the *recall* values would be improved because sentences that had been missed before would now be retrieved. In other words, out of all the relevant sentences in the corpus for each particular search, the TA would retrieve more this time than it did previously. Indeed, in all eight cases, the recall values increased. What I was *not* able to predict with certainty ahead of time was the effect that the enhancements would have on the *precision* values. I was rather pessimistic, given Frakes' claim (1992, 10-11) that "when precision goes up, recall typically goes down and vice versa". Therefore, I assumed that the new patterns would retrieve so much extra noise as to negate the benefits of the increased number of hits they harvested. I was pleasantly surprised: in seven out of the eight tests, the precision values *increased*. For the eighth test, precision at least remained the same, rather than decreasing. Interesting to note is the new statistics on the functionality relation. Although they improved, they remained low compared to the statistics for the other relations.

See Figure 16 below for a table comparing previous statistics with those obtained from the enhancement testing.

Figure 16—Statistics compared (lexical patterns)

Precision/recall BEFORE enhancements		Precision/recall AFTER enhancements	
Hyponymy		Hyponymy	
composting	recall: 4/17 = 24% precision: 4/22 = 18%	composting	recall: 10/17 = 59% precision: 10/20 = 50%
weeds	recall: 4/6 = 67% precision: 4/6 = 67%	weeds	recall: 5/6 = 83% precision: 5/7 = 71%
yard wastes	recall: 3/5 = 60% precision: 3/4 = 75%	yard wastes	recall: 5/5 = 100% precision: 5/6 = 83%
Meronymy		Meronymy	
carbon	recall: 5/15 = 33% precision: 5/11 = 45%	carbon	recall: 16/18 = 89% precision: 16/21 = 76%
compost	recall: 34/50 = 68% precision: 34/128 = 27%	compost	recall: 52/56 = 93% precision: 52/156 = 33%
nitrogen	recall: 18/35 = 51% precision: 18/32 = 56%	nitrogen	recall: 34/36 = 94% precision: 34/50 = 68%
nutrients	recall: 6/9 = 67% precision: 6/12 = 50%	nutrients	recall: 9/9 = 100% precision: 9/15 = 60%
Functionality		Functionality	
compost	recall: 4/19 = 21% precision: 4/10 = 40%	compost	recall: 6/13 = 46% precision: 6/15 = 40%

4.1.3 Problems encountered

As mentioned in Section 4.1.2 above, several problems were revealed during the "pre-test" with the new patterns. Each "surface" problem resulted from a single fundamental problem of conflicting patterns in a given sentence. I will present here only one of these surface problems.

The following miss that should have been retrieved in the new meronymic search using *compost* as the search term was not being retrieved:

> Use materials readily available; special **ingredients**, such as manure, are not necessary to make good **compost**.

This sentences explains that manure is a possible (but not vital) ingredient of compost. The sentence was not retrieved even though it contains the new pattern *ingredient** and the search term *compost*. The programmer determined that there is a conflict in the programming. The word combination *mak*...up* is a pattern already in the TA. The TA detected the word *make* in the sentence above, but because it is not followed by *up*, the TA rejected the sentence without checking other possibilities.

This problem could only have been revealed through increasing the number of patterns in the TA; eventually there will be sentences in a corpus that contain more than one pattern and can potentially cause problems. The programmer explained that the fundamental problem is caused by faulty logic in the coding, i.e. by the way in which the program code asks the TA to perform its searches—it does not allow the TA to analyse the sentence beyond the pattern it chooses to look at first. Resolving the problem for the present study would have been too time-consuming. However, future programmers can benefit from the discovery of this problem by knowing that their code must allow the program to be less restrictive in its analysis of a sentence.

4.2 EXPLORING KNOWLEDGE EXTRACTION USING GRAMMATICAL PATTERNS

Analysing the misses from the previous testing revealed not only potential new lexical patterns but also sentences that expressed semantic relations by *grammatical* rather than *lexical* means. Being able to extract this type of sentence would further enhance the usefulness of the TA.

Consider the relation of functionality. The following sentence explains the function of an oil filter for an engine:

> The oil filter removes abrasive particles that enter the lubrication system before they can cause excessive wear.

It is clear that the function is being explained, but what part of this sentence could be entered into a knowledge extraction program for future use? The word *removes* appears to be a meaningful sentence element, but it would be unwise to include the pattern *remov** in a program such as the TA because it is much too versatile. Furthermore, many concepts do not have as a function removing other things. Using this lexical pattern would result in very low recall AND precision values.

For knowledge extraction purposes, then, it is the underlying part of speech that performs the role of expressing function. If a concept has a function, it will *perform* that function. In language, the performance can be expressed by a verb describing an action and following the term in question relatively closely.

Another semantic relation that can be expressed grammatically is hyponymy. For example, hyponyms of the term *filter* may be expressed by an adjective[15] preceding the term: *oil filter*, *air*

[15] It may be tempting to conclude, from the examples given further in the sentence, that only nouns used adjectively (i.e. nominal adjectives) can express hyponymy (e.g. *oil* filter). However, attributive adjectives work as

filter, *gas filter*, which are types of filters. The TA is capable of performing this type of hyponymic search and the functionality search outlined above, provided that the corpus in question has been run through the part-of-speech tagger. This operation is described in the following section.

4.2.1 Preparing the corpus for grammatical searches

Before the corpus can be searched grammatically for instances of hyponymy and functionality (by looking for adjectives or verbs), it must be processed by the part-of-speech tagger. This operation analyses the corpus and labels each word with its part of speech. According to Kavanagh (1995, 29), the particular part-of-speech tagger incorporated into the TA, namely the Brill tagger, "can be used on unfamiliar text with an accuracy of about 96%."

Once the corpus has been processed in this way, it is ready to be searched on the basis of verbs following a given search word and adjectives occurring in the same sentence as the search word.

4.2.2 The TA's grammatical search capabilities

One operation of the TA allows users to extract sentences that contain adjectives (both attributive and nominal) appearing in the same sentence as a specified search word. This means that a search performed for, say, the term *filter* will retrieve all sentences containing *oil filter*, *air filter*, and any other kind of filter discussed in that corpus.

---

well (e.g. *thermophilic* bacteria are a kind of bacteria).

Naturally, there is the potential for noise extraction too. A sentence describing an *amazing filter* would not be an instance of hyponymy. In addition, since the TA will extract sentences with an adjective occurring anywhere in each sentence, it is possible that *filter* may stand unmodified, while another word in the sentence carries the adjective. These, too, would be noise. It would have been more useful for the present study if the TA could restrict the search to only those sentences where the adjective immediately precedes the search word[16].

Another TA operation lets users search for sentences where a verb follows the search word. This is useful for searching grammatically for the function of a concept because not all expressions of functionality contain the explicit markers such as *is used for* or *serves as*; some sentences directly state (by means of a following verb) what a given concept does. The TA is already programmed to allow for up to two words intervening between the search word and its verb (and therefore preventing the TA from rejecting such useful sentences as "Compost *always* provides nutrients..."). In addition, this particular operation allows users to tell the TA to ignore common verbs, if desired. Selecting this option can reduce the amount of noise the TA would retrieve.

4.2.3 Testing and statistics

The purpose of this testing was simply to provide a exploratory glimpse at searching grammatically for expressions of semantic relations. For testing the two operations described

[16] Interestingly enough, the TA *was* originally programmed to search for adjectives preceding the search word, but its functionality has since been changed to look for adjectives *anywhere* in sentences containing the search word.

above, I chose a single term for each. For a grammatical search for hyponymy, I used *bacteria*; for functionality, I used *compost*. The testing was straightforward and problem-free. Figure 17 below shows the hits/noise statistics resulting from the two searches.

Hyponymy (based on adjectives)	**Functionality** (based on following verb)
bacteria hits: 7 noise: 24 (8 adj. preceding) (16 adj. anywhere in sentence)	compost hits: 8 noise: 39

Figure 17—Hits and noise extracted with grammatical patterns

Note that, under hyponymy, I have further classified the instances of noise depending on where in the sentence the adjective occurred. The purpose of this is to show that, had the TA been able to restrict the search to adjective *immediately preceding* the search word, there would have been only eight instances of noise rather than 24. In either case, there are still seven new hits to add to the original four retrieved lexically (as seen in Figure 6).

Regarding functionality, we see that searching grammatically brings in a proportionately large amount of noise. Recall, however, that terminologists can easily ignore noise. The perspective to take, here, is that eight hits were retrieved during the *grammatical* search for the functionality of *compost*, which can be added to the six hits retrieved during the *lexical* search.

Therefore, being able to search a corpus for expressions of semantic relations using grammatical patterns further enhances the TA's value for semi-automating terminology work.

Chapter 5 CONCLUDING REMARKS

5.0 DIRECTIONS FOR FUTURE RESEARCH

In spite of the amount of work accomplished for this thesis, I feel I have only touched on the research necessary for significantly enhancing knowledge extraction technology. The following is a list of potential avenues for future research.

1. **Improving sentence delimiting for greater accuracy in text processing.** Errors in sentence delimiting lead to inaccurate knowledge extraction. For example, a hit would not be extracted if the sentence delimiter broke it into two sentences, with one part containing the search word and the other containing the lexical pattern.

2. **Discovering even more lexical patterns for the three relations covered in this thesis.** The patterns used in this thesis were only a beginning and were discovered in a relatively small body of text. Future research would undoubtedly reveal even greater lists of patterns for these three relations.

3. **Studying other semantic relations and discovering lexical patterns that express them.** Hyponymy, meronymy, and functionality are not the only relations that are relevant to terminology work. Future studies can look also at synonymy and causality, for example.

4. **Determining which patterns are "universal" and which are domain specific.** Practical constraints prevented me from researching this area for my thesis. Some patterns like *is a species of* may be a hyponymic pattern limited to biological sciences and related fields. Only extensive research will lead to firm conclusions on the applicability range of individual patterns.

5. **Delving further into research on *grammatical* patterns expressing semantic relations.** This thesis studied only two possibilities: adjectives for hyponymy and verbs for functionality. In addition, future studies may also look at ways to filter out noise, since my study showed that the potential for noise extraction using grammatical patterns is high.

6. **Researching all of the above for languages other than French and English.** For knowledge extraction technology to have global applicability (which it must, if it is to be used for terminology work), it obviously must be able to process text written in the world's major languages.

5.1 CONCLUSIONS

The main goal of this thesis was to show how knowledge extraction technology can be useful for terminology work. In doing so, my work contributed to the development of the Text Analyzer, an experimental knowledge extraction program. It should be noted that the work for this thesis involved enhancing a single program (the TA), but is not in fact limited to the TA; it can be applied as well to other corpus analysis tools and, as such, my research has universal validity.

Kavanagh (1995, 105-106) indicated that "we would like to see the following enhancements made to the Text Analyzer", one of which was "Discovering more *knowledge probes* for conceptual operations." The work accomplished for the present thesis has done just that. In addition to significantly building up the TA's archives of *English* lexical patterns for three out of four of its conceptual relations (hyponymy, meronymy and functionality), I created archives of *French* lexical patterns for the same relations, thereby providing the TA with second-language capabilities. The obvious implication of this is that the TA and similar programs can be further developed to perform conceptual operations on electronic texts in many other languages as well.

It was proposed (Kavanagh 1995, 2) that the TA has potential value for terminologists (among other language specialists), who "must examine large quantities of text, often in the millions of words..." By partly simulating a terminology project, I have shown in this thesis how the terminological tasks known as scanning and conceptual analysis can eventually be semi-automated by means of a knowledge extraction program. The statistics derived from testing the effectiveness of lexical patterns for knowledge extraction, while admittedly being based on small

corpora, are encouragingly high, and knowledge extraction using grammatical means only increases the value of knowledge extraction programs. Therefore, my thesis demonstrates that further development of this technology with a view to incorporating it into terminology work is a realistic pursuit.

Not only is it realistic, it is also much needed, given the pressure currently being put on terminologists from two sides. On the one hand, corporate and government cutbacks are forcing terminology as a field to reduce production time. On the other hand, we find an increasing demand for terminologies and lexicons from language specialists, *themselves* attempting to meet ever-tighter production deadlines.

But, just as pressure creates a diamond out of carbon, so the stresses placed on terminology will transform it into a more solid and valuable field of endeavour.

Bibliography

AHMAD, K. and S. COLLINGHAM. (1996). "Renewable Terminology." *Euralex '96 Proceedings*, Part II. Eds. M. Gellerstam *et al.* Göteborg: Göteborg University, Department of Swedish. 759-769.

AHMAD, K. and H. FULFORD. (1992). *Knowledge Processing: 4. Semantic Relations and their Use in Elaborating Terminology.* (Computing Sciences Report). Surrey: University of Surrey.

ATKINS, B.T.S., J. CLEAR, and N. OSTLER. (1992). "Corpus Design Criteria". Journal of Literary and Linguistic Computing 7(1): 1-16.

AUSTIN, J. L. (1962). *How to Do Things with Words.* Cambridge (Mass.): Harvard University Press.

BORILLO, A. (1996). "Exploration automatisée de textes de spécialité: repérage et identification de la relation lexicale d'hyperonymie." *Linx*, no. 34/35. Nanterre: Université de Paris X Nanterre. 113-124.

BOORSTIN, D.J. (1983). *The Discoverers.* New York: Random House.

CHAFFIN, R., *et al.* (1988). "An Empirical Taxonomy of Part-Whole Relations: Effects of Part-Whole Relation Type on Relation Identification." *Language and Cognitive Processes*, Vol. 3(1). 17-48.

COLE, W. (1987). "Terminology: Principles and Methods." *Computers and Translation*, vol 2. Ed. W.P. Lehmann. Sarasota: Paragidgm Press, Inc. 77-87.

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

DUBUC, Robert. 1992. *Manuel pratique de terminologie*, 3rd edition. Montreal: Linguatech.

EVENS, M., *et al.* (1980). *Lexical-Semantic Relations: A Comparative Survey.* Edmonton: Linguistic Research, Inc.

FELBER, H. (1984). *Terminology Manual.* Paris: International Information Centre for Terminology (Infoterm).

FLOWERDEW, J. (1992). "Salience in the Performance of One Speech Act: The Case of Definitions." *Discourse Processes* 15. 165-181.

FRAKES, W.B. (1992). "Introduction to Information Storage and Retrieval Systems." *Information Retrieval Data Structures and Algorithms*. Eds. W.B. Frakes and R. Baeza-Yates. New Jersey: Prentice Hall, Inc.

[HANDBOOK] "The Terminologist's Handbook." (1984). Unpublished, in-house manual used by TERMIUM staff.

HEARST, M. A. (1992). "Automatic Acquisition of Hyponyms from Large Text Corpora." *Proceedings of the Fourteenth International Conference on Computational Linguistics*. Nantes: Coling-92.

IRIS, M. A., *et al.* (1988). "Problems of the part-whole relation." *Relational models of the lexicon: Representing knowledge in semantic networks*. Ed. Martha W. Evens. Cambridge: Cambridge University Press.

JACKIEWICZ, A. (1996). "L'expression lexicale de la relation d'ingrédience (partie-tout)." *Faits de langues: revue de linguistique*, no. 7. Paris: Ophrys. 53-62.

KAVANAGH, J. (1995). "The Text Analyzer: A Tool for Knowledge Acquisition from Texts." Master's Thesis. Dept. of Computer Science, University of Ottawa. 1995.

LARIVIERE, L. (1996). "Comment formuler une définition terminologique." *Meta*, Vol. 41(3). Montréal: L'Université de Montréal. 405-418.

LAURIAN, A. (1983). "Typologie des discours scientifiques: deux approches." *Études de linguistique appliquée*, no. 51. Paris: Didier. 8-20.

LYONS, J. (1968). *Introduction to Theoretical Linguistics*. Cambridge: Cambridge University Press.

LYONS, J. (1977). *Semantics: Volume 1*. Cambridge: Cambridge University Press.

McENERY, T. and A. WILSON. (1996). *Corpus Linguistics*. Edinburgh: Edinburgh University Press.

MEYER, I. (1993). "Concept Management for Terminology: A Knowledge Engineering Approach." *Standardizing Terminology for Better Communication: Practice, Applied Theory, and Results*. Eds. Richard A. Strehlow and Sue Ellen Wright. Philadelphia: American Society for Testing and Materials. 140-151.

MEYER, I. (1994). "Linguistic Strategies and Computer Aids for Knowledge Engineering in Terminology." *L'actualité terminologique/Terminology Update*, Vol. 27(4). Ed. M. Valiqette. Ottawa: Public Works and Government Services Canada. 6-10.

MEYER, Ingrid, and Kristen MACKINTOSH (1995). In press 1996. "The Corpus from a Terminographer's Viewpoint." *International Journal of Corpus Linguistics*, Vol 1(2). Amsterdam/Philadelphia: John Benjamins Publishing Company.

MEYER, Ingrid, and Bruce McHAFFIE. 1994. "De la focalisation à l'amplification: nouvelles perspectives de représentation des données terminologiques." *TA-TAO: Recherches de pointe et applications immédiates*. Montreal: AUPELF-UREF.

MEYER, I. and L. PARADIS. (1991). "Applying Knowledge-Engineering Technology to Terminology: A Pilot Project." *L'actualité terminologique/Terminology Update*, Vol. 24(2). Ed. M. Valiqette. Ottawa: Department of the Secretary of State of Canada. 3-8.

MEYER, I., *et al.* (1997). "Systematic Concept Analysis within a Knowledge-Based Approach to Terminology." *Handbook of Terminology Management*, Vol. 1. Eds. Sue Ellen Wright and Gerhard Budin. Amsterdam/Philadelphia: John Benjamins Publishing Company. 98-118

MILLER, G. A. (1990) "Nouns in WordNet: A Lexical Inheritance System." *International Journal of Lexicography*, Vol. 3(4). Oxford: Oxford University Press.

OTMAN, G. (1989). "Terminologie et intelligence artificielle." *Banque des mots*, Special edition (Dec. 1989). 63-95.

PALMER, F. R. (1981). *Semantics*. (2nd ed.) Cambridge: Cambridge University Press.

PEARSON, J. (1996). "The Expression of Definitions in Specialised Texts: A Corpus-based Analysis." *Euralex '96 Proceedings*, Part II. Eds. M. Gellerstam *et al.* Göteborg: Göteborg University, Department of Swedish. 759-769.

PICHT, Heribert, and Jennifer DRASKAU. 1985. *Terminology: An Introduction*. Guilford: University of Surrey.

ROBISON, H. R. (1970). "Computer-Detectable Semantic Structures." *Information Storage and Retrieval*, Vol. 6. Oxford: Pergamon Press. 273-288.

RONDEAU, G. (1981). *Introduction à la terminologie*. Montréal: Centre Educatif et Culturel inc.

RUNDELL, M. and P. Stock. (1992a) "The Corpus Revolution." *English Today*, Vol 8 (2). Cambridge: Cambridge University Press. 9-14.

RUNDELL, M. and P. Stock. (1992b) "The Corpus Revolution." *English Today*, Vol 8 (3). Cambridge: Cambridge University Press. 21-32.

RUNDELL, M. and P. Stock. (1992c) "The Corpus Revolution." *English Today*, Vol 8 (4). Cambridge: Cambridge University Press. 45-51.

SAGAN, C. (1997). *The Demon-Haunted World: Science as a Candle in the Dark.* New York: Ballantine Books.

SAGER, Juan C. (1990). *A Practical Course in Terminology Processing.* Amsterdam/ Philadelphia: John Benjamins Publishing Company.

SAGER, Juan C. (1994a). "Terminology: Custodian of Knowledge and Means of Knowledge Transfer." *Terminology*, Vol 1, 1. Eds. Helmi B. Sonneveld and Kurt L. Loening. Amsterdam/Philadelphia: John Benjamins Publishing Company.

SAGER, J. (1994b). *Language Engineering and Translation: Consequences of Automation.* Amsterdam/Philadelphia: John Benjamins Publishing Company.

SINCLAIR, J. (1991). *Corpus, Concordance, Collocation.* Oxford: Oxford University Press.

SKUCE, D. (1996a). *IKARUS Details.* http://www.csi.uottawa.ca:80/~kavanagh/Ikarus/ IkarusDetails.html Oct. 1996 (accessed Oct. 15, 1996).

SKUCE, D. (1996b). *IKARUS: Intelligent Knowledge Acquisition and Retrieval Universal System.* http://www.csi.uottawa.ca:80/~kavanagh/Ikarus/Ikarus4.html Oct. 1996 (accessed Oct. 15, 1996).

SKUCE, D. (1996c). "Integrating Web-based Documents, Shared Knowledge Bases, and Local Document Searching for User Help."

SMITH, Jim. (1996). "Sceptics Column." *Ariadne: The Web Version*, Issue 1. http://ukoln.bath.ac.uk/ariadne/issue1/jim/intro.html Jan 17, 1996 (accessed Sept. 18, 1996).

VAN CAMPENHOUDT, M. (1996). "Recherche d'équivalents et structuration des réseaux notionnels: Le cas des relations méronymiques." *Terminology*, Vol 3(1). Amsterdam/Philadelphia: John Benjamins Publishing Company. 53-83.

VENDITTO, G. (1996). "Search Engine Showdown: IW Labs Test Seven Internet Search Tools." *Internet World* (May 1996), 79-86.

WINSTON, M., *et al.* (1987). "A Taxonomy of Part-Whole Relations." *Cognitive Science* 11(4). 417-444.

Webliography

URLs OF DOCUMENTS USED IN THE CORPORA

Note: All documents accessed April 1997

English

Government of New Brunswick

http://www.gov.nb.ca/environm/comucate/compost/magic.htm
http://www.gov.nb.ca/environm/comucate/compost/back.htm
http://www.gov.nb.ca/environm/comucate/compost/comworks.htm
http://www.gov.nb.ca/environm/comucate/compost/usecomp.htm
http://www.gov.nb.ca/environm/comucate/compost/nurcp.htm
http://www.gov.nb.ca/environm/comucate/compost/build.htm
http://www.gov.nb.ca/environm/comucate/compost/combin.htm
http://www.gov.nb.ca/environm/comucate/compost/otways.htm
http://www.gov.nb.ca/environm/comucate/compost/qacom.htm
http://www.gov.nb.ca/environm/comucate/compost/indoors.htm
http://www.gov.nb.ca/environm/comucate/compost/wormcom.htm
http://www.gov.nb.ca/environm/comucate/compost/explor.htm
http://www.gov.nb.ca/environm/comucate/compost/shortcom.htm

---

http://www.shepherdseeds.com/sgsweb/library/grow_guides/ggs1/compostingpart1/compostingpart1.html
http://www.shepherdseeds.com/sgsweb/library/grow_guides/ggs1/compostingpart1/compostingpart12.html
http://www.shepherdseeds.com/sgsweb/library/grow_guides/ggs1/compostingpart1/compostingpart13.html
http://www.shepherdseeds.com/sgsweb/library/grow_guides/ggs1/compostingpart1/compostingpart14.html
http://www.shepherdseeds.com/sgsweb/library/grow_guides/ggs1/compostingpart1/compostingpart15.html
http://www.shepherdseeds.com/sgsweb/library/grow_guides/ggs1/compostingpart1/compostingpart16.html
http://www.shepherdseeds.com/sgsweb/library/grow_guides/ggs1/compostingpart1/compostingpart17.html
http://www.shepherdseeds.com/sgsweb/library/grow_guides/ggs1/compostingpart1/compostingpart18.html
http://www.shepherdseeds.com/sgsweb/library/grow_guides/ggs1/compostingpart1/compostingpart19.html
http://www.shepherdseeds.com/sgsweb/library/grow_guides/ggs1/compostingpart1/compostingpart110.html
http://www.shepherdseeds.com/sgsweb/library/grow_guides/ggs1/compostingpart2/compostingpart2.html
http://www.shepherdseeds.com/sgsweb/library/grow_guides/ggs1/compostingpart2/compostingpart22.html
http://www.shepherdseeds.com/sgsweb/library/grow_guides/ggs1/compostingpart2/compostingpart23.html
http://www.shepherdseeds.com/sgsweb/library/grow_guides/ggs1/compostingpart2/compostingpart24.html

---

http://www.bae.ncsu.edu/bae/programs/extension/publicat/wqwm/psfact11.html

http://www.bae.ncsu.edu/bae/programs/extension/publicat/wqwm/ebae171_93.html

http://www.bae.ncsu.edu/bae/programs/extension/publicat/wqwm/ag473_14.html

http://www.bae.ncsu.edu/bae/programs/extension/publicat/wqwm/ebae202_94.html

---

Cornell Composting

http://www.cfe.cornell.edu/compost/Note.html
http://www.cfe.cornell.edu/compost/invertebrates.html
http://www.cfe.cornell.edu/compost/microorg.html
http://www.cfe.cornell.edu/compost/monitor/monitortemp.html
http://www.cfe.cornell.edu/compost/monitor/monitorph.html
http://www.cfe.cornell.edu/compost/monitor/monitormoisture.html
http://www.cfe.cornell.edu/compost/calc/rightmix.html
http://www.cfe.cornell.edu/compost/odors/factors.html
http://www.cfe.cornell.edu/compost/oxygen/oxygen.transport.html
http://www.cfe.cornell.edu/compost/calc/cn_ratio.html

French

Government of New Brunswick

http://www.gov.nb.ca/environm/comucate/compost/magic.htm
http://www.gov.nb.ca/environm/comucate/compost/guide.htm
http://www.gov.nb.ca/environm/comucate/compost/pile.htm
http://www.gov.nb.ca/environm/comucate/compost/utile.htm
http://www.gov.nb.ca/environm/comucate/compost/recette.htm
http://www.gov.nb.ca/environm/comucate/compost/nupile.htm
http://www.gov.nb.ca/environm/comucate/compost/conten.htm
http://www.gov.nb.ca/environm/comucate/compost/autres.htm
http://www.gov.nb.ca/environm/comucate/compost/questo.htm
http://www.gov.nb.ca/environm/comucate/compost/inter.htm
http://www.gov.nb.ca/environm/comucate/compost/vermi.htm
http://www.gov.nb.ca/environm/comucate/compost/exper.htm
http://www.gov.nb.ca/environm/comucate/compost/cours.htm

---

http://www.geagri.qc.ca/parde/compadap.html

---

http://www.geagri.qc.ca/parde/fumferme.html

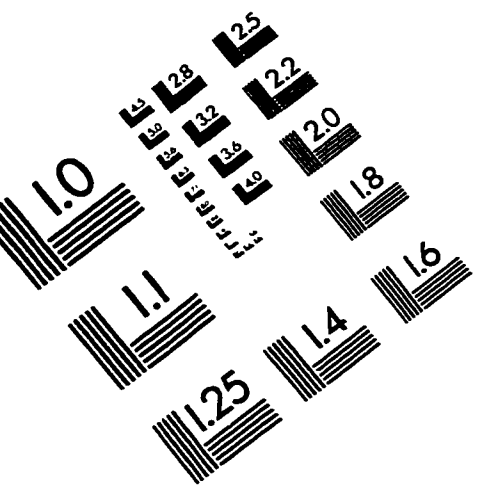
---

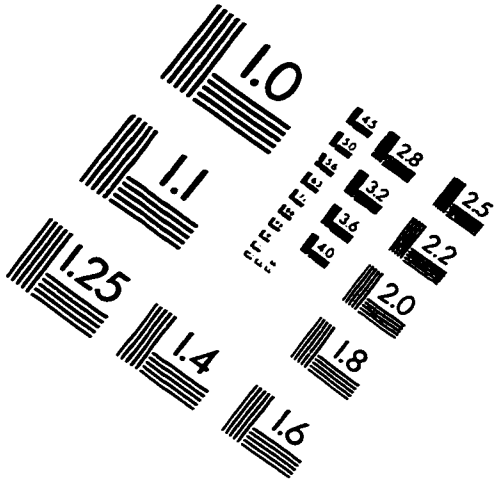
http://www.wul.qc.doe.ca/biospher/truc/truc_00000_f.html#comp

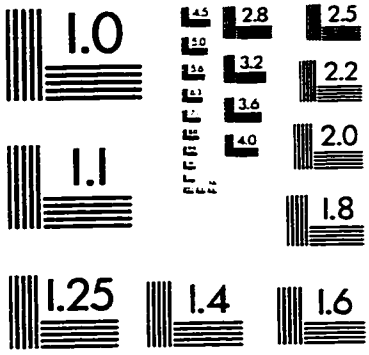
---

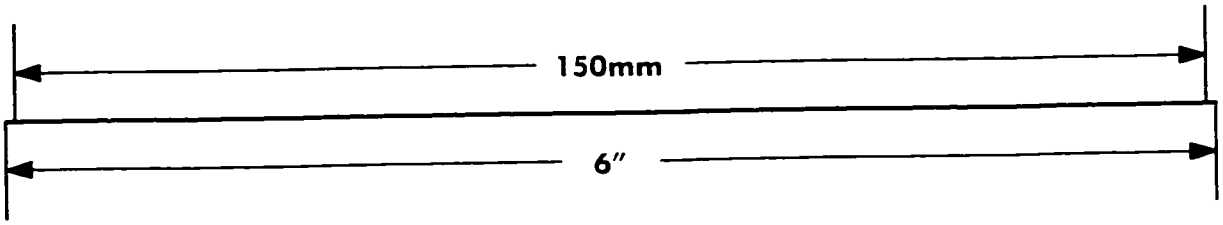
http://helios.emse.fr/~brodhag/TRAITEME/fich17_3.htm

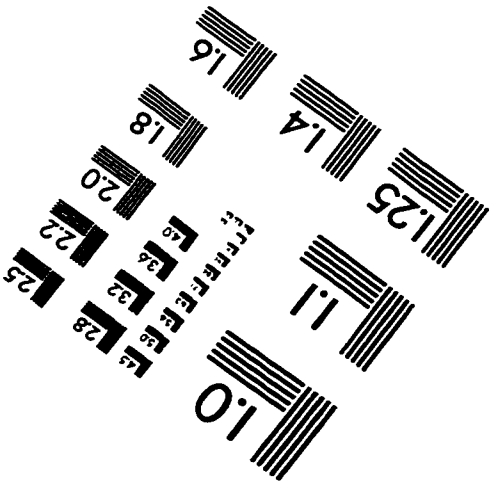
IMAGE EVALUATION
TEST TARGET (QA-3)

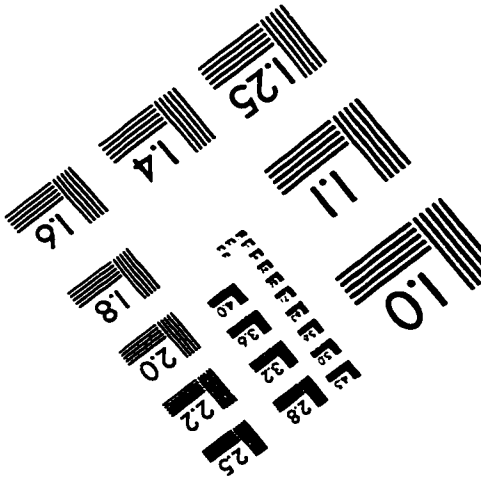












APPLIED IMAGE, Inc
1653 East Main Street
Rochester, NY 14609 USA
Phone: 716/482-0300
Fax: 716/288-5989