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.

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.

ProQuest Information and Learning
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
800-521-0600

Université d'Ottawa · University of Ottawa

A COMPARATIVE STUDY OF HYPERNYMIC PATTERNS FOR KNOWLEDGE EXTRACTION

by

Tricia Morgan

School of Translation and Interpretation
University of Ottawa

under the supervision of

Ingrid Meyer, Ph.D.
School of Translation and Interpretation

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)

Emerging semi-automatic knowledge extraction tools make it possible for terminologists to efficiently locate contexts in text that illustrate the meaning and usage of terms. Semi-automatic knowledge extraction makes use of a pre-programmed set of linguistic structures (patterns) which frequently show relationships between concepts.

This thesis contributes to the linguistic research on patterns that is needed for tool development. More specifically, this thesis focuses on the patterns for one specific relation: hyperonymy. The research involved two comparative studies of hypernymic patterns, one *intra*-linguistic (French) and the other *inter*-linguistic (French-English). The first study involved discovering French patterns in two separate domains, and comparing their nature and frequency between domains.

The second study involved comparing the stylistic differences between the French patterns discovered and a pre-existing set of English patterns.

This research has revealed that comparative linguistic research on hypernymic patterns can provide useful information for developing terminological knowledge extraction tools.

ACKNOWLEDGEMENTS

There are so many people who have helped me in so many ways with this thesis. First of all, I would like to thank my advisor, Dr. Ingrid Meyer, for all of her help, suggestions, and encouragement.

I would also like to thank Dr. Bossé-Andrieu for the suggested "reading material" in comparative linguistics, and for answering all my questions.

I am also greatly thankful to Teresa Reguly, Ph.D. candidate in molecular genetics at Mount Sinai Hospital Research Institute, for her expert and prompt help with all of my questions on genetics, no matter how ridiculous they sounded to her.

Judy Kavanaugh, programmer of the Text Analyzer, is also to be thanked for her demonstrations of the software and help with individual questions.

Many thanks also to Sandrine Colle, for correcting the French translation of my abstract and for answering the few questions I had on French usage.

Also, thanks to Kristen Mackintosh for the use of her Chicago Manual of Style and for convincing me that I could indeed finish my thesis some day.

Perhaps the person to whom I am most indebted is my roommate, Liz, who allowed me to take over the dining room table for my thesis, and who also constantly encouraged me and put a positive spin on any situation.

ABSTRACT

Terminologists are required to scan a wide variety of texts to search for contexts that illustrate the meaning and usage of terms. Traditionally, this entailed the time-consuming process of manually reading large amounts of paper-based documentation to extract the most interesting contexts. However, emerging semi-automatic knowledge extraction tools make it possible to locate contexts much more efficiently. Semi-automatic knowledge extraction makes use of a pre-programmed set of linguistic structures (patterns) which frequently show relationships between concepts.

This thesis contributes to the linguistic research on patterns that is needed for tool development. More specifically, this thesis focuses on the patterns for one specific relation: hyperonymy. The research involved two comparative studies of hypernymic patterns, one *intra*-linguistic (French) and the other *inter*-linguistic (French-English). The first study involved discovering French patterns in two separate domains, and comparing their nature and frequency between domains.

The second study involved comparing the stylistic differences between the French patterns discovered and a pre-existing set of English patterns.

This research has revealed that both intra-linguistic and inter-linguistic comparative linguistic research on hypernymic patterns can provide useful information for the development of terminological knowledge extraction tools.

RÉSUMÉ

Les terminologues sont appelés à dépouiller un large éventail de textes pour repérer les contextes qui démontrent le sens et l'usage d'un terme. Traditionnellement, ceci impliquait que le terminologue examine de près un grand nombre de documents sur papier pour en retirer les contextes les plus intéressants du point de vue terminologique. Cependant, des logiciels informatiques en développement pour l'extraction semi-automatique permettent de retrouver des contextes d'une façon plus efficace. Les logiciels se servent d'une liste de structures linguistiques (patrons) qui montrent fréquemment les relations entre concepts.

Cette thèse contribue à la recherche linguistique sur les patrons nécessaire au développement des logiciels. En particulier, cette thèse se concentre sur ces patrons dans le cadre d'une relation : l'hypéronomie. La recherche vise à effectuer deux études comparatives de ces patrons, dont une *intra*linguistique (français), et l'autre *inter*linguistique (français-anglais). La première étude impliqua la découverte de patrons français dans deux domaines séparés et la comparaison des patrons et leur fréquence dans les deux domaines.

La deuxième impliqua la comparaison des différences stylistiques entre les patrons français et une liste existante de patrons anglais.

Cette recherche montre que les études de linguistique comparée (intralinguistique et interlinguistique) sur les patrons peuvent mettre à jour des informations utiles pour le développement de ces logiciels dans le cadre de la recherche terminologique.

TABLE OF CONTENTS

INTRODUCTION

Loci communes (literally "common places") refer to very general types of rhetorical arguments that can be applied in any domain. Aristotle classifies all *loci* that can serve as premises for dialectical or rhetorical syllogisms as *loci* relating to accident, species, property, sameness, and definition (Perelman 83-84). According to Selinker *et al*, "One of the most important and frequently employed rhetorical functions is that of 'definition'; this function is basic to the scientific thinking and reporting process" (1976, 39). This last *locus*, definition, plays an extremely important role in terminology, and will be the focus of this thesis.

Terminologists have always preferred finding definitions for terms in actual sources, as opposed to constructing definitions themselves. Finding definitions has become increasingly easy with the advent of the information age and the development of new computer technology for knowledge extraction. Whereas in the past, terminologists manually scanned texts for useful contexts, knowledge extraction tools allow terminologists access to the speed and capabilities of a computer to sift through electronic sources. This thesis attempts to make a modest contribution to the development of knowledge extraction tools for terminology. Its contribution is primarily *linguistic* (as opposed to *computational*) in nature.

Though definitions can come in many forms, traditionally, they are often expressed according to the format set down by Aristotle, that is *genus* + *differentia* (Eck, 1993). A *genus* is often referred to as an *hypernym*, and definitions constructed according to this formula can be called *hypernymic contexts*[1]. Aristotle's formula is still

[1] Hypernymic contexts will be discussed in more detail in Chapter 1.

applied today by lexicographers, terminologists, scientific and technical writers, and anyone else who needs to define terms. One of the aims of knowledge extraction tools is to locate these hypernymic contexts in large bodies of electronic texts by using recurrent words, groups of words, or paralinguistic features of the text (paralinguistic patterns). For instance, consider the following sentence:

> Le lecteur de disquettes est un appareil d'enregistrement (...) sur un disque amovible, la disquette.

Here, the pattern *est un* signals that *lecteur de disquettes* is an hyponym of *appareil d'enregistrement* (the hypernym). My research involved discovering these patterns in French, viewing how frequently they occur, identifying issues and problems for tool development, and comparing these French patterns to an established set of English patterns.

OBJECTIVES

General Objective

The general goal of this thesis is to add to the still limited body of linguistic research on the semi-automatic extraction of hypernymic contexts for terminology work. This work will focus on the linguistic patterns that underlie hypernymic contexts. Linguistic research of this type allows developers of knowledge extraction tools to better understand how patterns work and how tools can be designed to maximize the potential of the patterns.

Specific Objectives

Firstly, this thesis aims to discover hypernymic patterns in French as found in corpora for two domains: computers and genetics.

Secondly, the thesis carries out an intra-linguistic (within one language) comparison of French patterns for both domains. This comparison attempts to determine the relative frequency of the patterns, and whether the frequency of patterns and even the patterns themselves vary by domain. Thus, an attempt is made not only to discover *generic* patterns, but to identify any patterns which may be *domain-specific* for computers and genetics.

Thirdly, the thesis makes an inter-linguistic comparison between the French patterns identified and the existing set of English patterns established by Davidson (1998). In this comparison, I examine differences in the patterns, as well as general differing linguistic tendencies in the ways the patterns are expressed in context.

Lastly, this study proposes possible implications of the above findings for knowledge extraction tool development.

METHODOLOGY AND TOOLS

Because of the nature of my thesis, I needed to acquire background knowledge in two separate domains: terminology, and the contrastive linguistics of French and English. For terminology, I needed to understand basic terminology work methods, conceptual relations, patterns for knowledge extraction, corpus-building techniques, and corpus-analysis tools (Wordsmith and the Text Analyzer). For the contrastive linguistics aspect, I needed to understand the basic principles of this domain.

The methodology, discussed in detail in Chapter 2, involved three major steps:

1. **Corpus Building**
   - choosing two separate domains – genetics and computers – for the intra-linguistic comparison
   - building four separate corpora (one in English and one in French for both domains) to provide a source of hypernymic contexts
2. **Corpus Analysis**
   - discovering terms in two domains
   - extracting hypernymic statements using Key Word in Context (KWIC) concordances in French and, to a lesser extent, in English
   - identifying patterns in French which signal hypernymic contexts
   - identifying domain-specific patterns in French for computers and genetics
3. **Comparative Analyses of Linguistic Patterns**

The intra- and inter-linguistic comparisons involved, respectively:

- charting the frequencies of the French patterns for computers and genetics
- comparing the French and English patterns to identify stylistic differences

Motivation

As a research assistant for the COGNITERM project, I was able to work with linguistic patterns in both French and English, and contribute to tool development in English. I was interested in conducting further research in this area for French, and seeing if the same issues and problems arose in French as in English. In addition, I also wanted to incorporate an element of comparative linguistic work, since I had developed an interest in this field through the undergraduate course on French-English stylistics taught by Dr. Bossé-Andrieu at the School of Translation.

A review of the current body of literature on patterns (Chapter 1) revealed that far less work has been done in French than in English. As well, most studies have focussed on one particular domain, rather than examining how the patterns behaved in different domains. To date, no statistical analysis has been done to determine which patterns are the most productive. Because of these "gaps" in the research on patterns, this thesis will focus on compiling a list of hypernymic patterns in French, providing a statistical analysis of the use of these patterns in two domains, examining domain-specific patterns in these domains, and comparing the English and French patterns for differences. Programmers of knowledge extraction software can hopefully benefit from my findings on hypernymic patterns in the two languages and apply this knowledge towards tool development.

CHAPTER 1 - CONCEPTS IN TERMINOLOGY

1.0 Introduction

This chapter examines basic concepts in terminology that are relevant to this thesis.

1.1 The Changing Field of Terminology

> The perception of terminology as a tool also serves to explain its evolution as a regular field of study. Already in antiquity, terminology was recognised as essential to knowledge representation as the wheel is to land transport; both are continuously being developed further according to the raw materials available at the time and the function of the tool in the overall field.(...) The breakthrough for research in terminology, comparable to linking the wheel to a power drive, came relatively late and in the form of the computer, which for the first time permitted the manipulation of the large quantities of data required for any reliable experimental research (Sager, 1994a, 9).

According to Rondeau, as long as there has been general language, there have been special subject languages, "...qu'il s'agisse de la terminologie des philosophes grecs, de la langue des affaires des commerçants crétois, des vocables spécialisés de l'art militaire, etc." (1984, 1). Special subject languages, called "language for specific purposes" by Pearson and "sublanguages" by many researchers in natural language processing (Pearson, 1998, 7), are differentiated from general language, or language for general purposes, in that they are concerned with language as used in specialized domains of knowledge.

The Twentieth Century has been termed the "Information Age," an age characterized by an explosion of knowledge. The number of specialized domains is increasing exponentially as science and technology are advancing more and more rapidly, and with them, the numbers of terms they use. The field of terminology must keep pace with these advances as quickly as they are made. One of the ways to do this is by making use of the new technologies developed to locate examples of terms as used in context.

Traditional methods of gathering information manually from paper-based sources are not necessarily the most practical ways of managing knowledge. Recently, more efficient ways of processing knowledge have been explored for terminological purposes in the form of computer-aided knowledge extraction.

1.1.2 Knowledge Extraction: Manual and Automatic

Terminologists are constantly searching for terms used in context to provide textual support and form the basis of definitions on term records. Dubuc and Lauriston underscore the importance of the context for terminology work: " For the terminologist, the context is the key to the concept, which in turn constitutes the keystone of the terminologist's work" (1997, 82). In the past, terminologists have sought these contexts by reading through large amounts of specialized documentation and manually extracting those contexts which were most valuable to the terminologist. Today, however, large quantities of text are becoming available in electronic format. Sources of text already available in electronic format include CD-ROMs, and texts found on the Internet. As well, documentation can be converted to electronic format by scanning texts into a computer and applying optical character recognition.

The fact that documentation can be easily found in or converted to electronic format means that terminologists can quickly compile and store millions of words on the same subject. This collection of texts on the same subject is known as a specialized corpus.[1] An electronic corpus can be analyzed by a computer program, such as a concordancer. Concordancers allow the user to input a search term and view all its occurrences. Thus,

[1] Corpora will be discussed further in Section 2.2.

the computer program can reduce the amount of material the terminologist has to examine.

However, a sizeable corpus can easily contain several hundred or even thousand occurrences of a particular term. In such a case, it is necessary to filter the information even more to reduce the terminologist's work. It is the quality, rather than the quantity of the contexts which matters most.

1.2 Concept Analysis

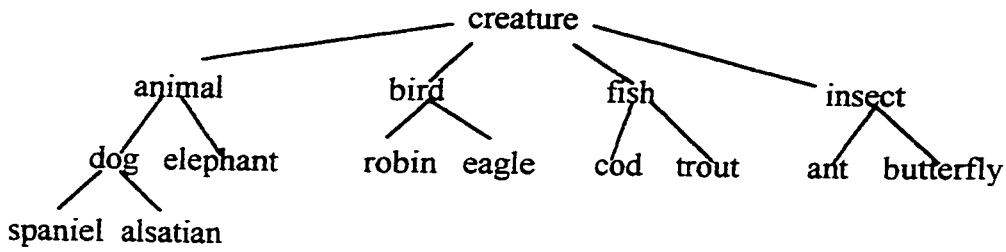
Concepts are the backbone of terminological work (Pearson, 1998, 10), and it is information about concepts that knowledge extraction tools seek to locate. Sager defines concepts as "elements of the structure of knowledge" (1990, 14) while Cole describes them as "the abstraction comprising the totality of the characteristics of any given object or group of objects" (1987, 78). Concept analysis requires understanding how a concept relates to other concepts in the domain, and thus, to the general structure of the domain.

1.2.1 Conceptual Relations

In any particular domain, there are many different relationships that exist between concepts. These include hyperonymy – the subject of the present thesis – meronymy (part-whole), functionality, synonymy, and causality. While the list of relations between concepts is much longer than this, certain relations are considered to be more important than others and are used more often for terminological purposes. The most important of these is the hypernymic relation.

1.3 The Hypernymic Relation[2]

Sager (1990) refers to this relation as the generic relationship, while Lyons (1968) terms it the relation of inclusion. Hyperonymy establishes an hierarchical order in that it relates the hypernym, or generic term, to a hyponym, or more specific one. The hyponyms inherit the characteristics of hypernym, but have at least one differentiating characteristic. Concepts linked to one another by this relation can be represented by a tree diagram, with the generic concepts at the top end and the more specific ones further down (Cruse, 1986, 136)[3]:



The lower on the diagram the concept is, the more characteristics it possesses. The sum of a concept's characteristics is known as its *intension*, and the set of objects referred to by a concept is called the *extension* (Meyer *et al*, 1997, 116). The highest term in an hierarchy, in this case *creature*, has the narrowest intension and the broadest extension. The lower the concept is in the hierarchy, the broader the intension becomes and the narrower the extension becomes (Lyons, 1968, 454).

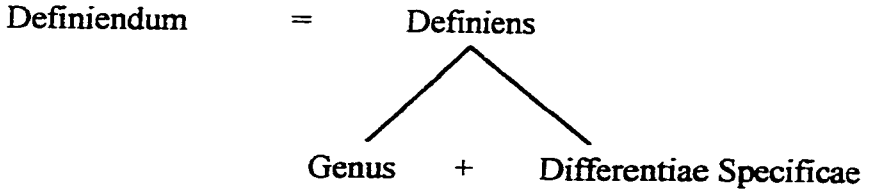
---

[2] Technically, in contexts which express hyperonymy, the search term is the subordinate concept (hyponym), and in contexts which express hyponymy, the search term is the superordinate concept. However, for reasons of space constraints, the term *hyperonymy* will indicate both hyperonymy and hyponymy in this thesis.

[3] This diagram is just an example to illustrate hyperonymy and is not meant to be a complete conceptual diagram. Also, the reader should note that scientists may classify concepts differently than laypeople, in that they would consider a bird to be in a different category than an animal.

1.3.1 Expressing the hypernymic relation

The hypernymic relation is typically expressed in a definitional structure. Picht and Draskau compare a definition to a mathematical equation (1985, 51). The traditional, or Aristotelian, definition can be said to be structured in the following manner:



(Davidson, 1998, 30)

OR

X = Y + differentiating characteristics

In this "equation," X represents the hyponym, or the term to be defined. Just as in mathematics, the equal sign in this formula means that the two sides of the equation "should be substitutable for each other in any context, without changing the meaning of the context" (Meyer, 1999, 4). When an hypernymic context is expressed in language, the equal sign is converted to certain verbs or verb phrases which overtly indicate an hypernymic relation. Examples of this in English include *is/are, consist(s) of, is/are defined as*, and *is/are called* (Pearson, 1996, 821) and *such as, and other*, and *known as* (Meyer *et al*, 1999, 257). Y is the hypernym, and the differentiating characteristics are those which differentiate Y from X's other hyponyms (i.e. from the co-hyponyms of Y). The following hypernymic context is expressed according to this formula:

RNA is a polymer that contains ribose rather than deoxiribose sugars.

1.3.2 Hypernymic contexts in naturally-occurring texts

We shall distinguish naturally-occurring texts from reference material such as dictionaries, glossaries and term records which have normally been created by lexicographers or terminologists. A naturally-occurring text can contain examples of hypernymic contexts if one of author's purposes is to provide information. While terminologists and lexicographers know what elements a definition should contain, authors of such texts sometimes alter the formula.

Since this study deals with hypernymic contexts as found in naturally-occurring texts, contexts will not be rejected if they are not "perfect" examples of the Aristotelian formula. In fact, Selinker *et al* point out that defining information is often implicit rather than explicit in real texts (1976, 40). This means that few definitions are explicitly signalled as such and preceded by "I define an X as..." (Davidson, 1998, 30). Since this is a subject which has been studied in much greater detail elsewhere (Pearson, 1998) we shall only briefly summarize the major tendencies in altering the formula as observed in the contexts extracted.

Flowerdew's (1992a) study of definitions in science lectures outlines several ways in which the structure of hypernymic contexts can vary. One of the most common variations found is that terms can be defined by their extension, or hyponyms. In this way, multiple co-hyponyms are given with no differentiating characteristics. The following hypernymic context exemplifies this:

> Les principaux périphériques d'entrée sont le clavier, la souris et le lecteur de CD-ROM.

In this example, the formula is reversed in that the hypernym (*périphérique d'entrée*) precedes the hyponyms. It contains a verbal phrase (*sont le*) which could be subsituted for the equal sign.

1.3.3 Inter-sentential hypernymic contexts

While most of the contexts found during the research stage of this study were limited to one sentence, several contexts were identified which were examples of inter-sentential hypernymic contexts. These contexts are those which express an hypernymic relation over more than one sentence. Pearson (1998) refers to such contexts as "formal definitions," whereas Meyer (1999) calls them "inter-sentential defining contexts." Research has shown that hypernymic contexts can span more than one sentence, especially when the hyponym occurs in the first sentence and is replaced by the demonstrative at the beginning of the second. Pearson provides the following example:

> (80) preference indicator. This is an indicator contained within the forward call indicators parameter field of ISUP, sent in the forward direction indicating whether or not the user... (1998, 156).

Inter-sentential hypernymic contexts were included in this study.

1.4 Patterns for Knowledge Extraction

Past research conducted by Lyons (1977), Cruse (1986), Winston *et al* (1987), Flowerdew (1992), Ahmad and Fulford (1992), Hearst (1992), Meyer (1994), Borillo (1996), Davidson (1998), and Pearson (1996, 1998) has shown that a given conceptual relation tends to be expressed by patterns. These patterns are essentially words, word

combinations or paralinguistic features of the text which frequently point to a certain conceptual relation. For instance, *is a* can denote hyperonymy, *consists of* points to a meronymic relation, and *is used for* indicates functionality. Bodson (1999, 6) underscores the importance for terminology in identifying such patterns:

> La méthode traditionnelle pour rechercher des éléments définitoires sur un terme consiste à interroger un concordancier sur ce terme afin de le repérer en contexte. Les informations trouvées sont de nature générale (...) En repérant les patrons de façon automatique ou semi-automatique, la recherche, entreprise par le terminologue, est grandement facilitée ; elle est plus précise et plus rapide. Concrètement, le terminologue peut retrouver dans un corpus uniquement les informations dont il a besoin pour une recherche qui va droit à l'essentiel.

Patterns that correspond to conceptual relations have been referred to in different ways by researchers. Davidson *et al.* (1998, 51) provide this succinct summary of the different terminology used to designate these patterns in English:

> These linguistic patterns have been designated by various terms in English, depending on the researcher: Lyons (1977) uses *formulae*; Cruse (1986) refers to *diagnostic frames* or *test frames;* Winston *et al* (1987) simply use *frames*; Flowerdew (1992) refers to *linguistic structures* that make certain definitional information salient (and breaks these down into *boosters* and *downgraders*); Pearson (1996) refers to *hinges*; Ahmad and Fulford (1992) call them *knowledge probes*.

To this I would add that Davidson *et al* themselves use *knowledge patterns*, as does Meyer (1999).

In the following examples of hypernymic contexts, there are certain patterns which signal to the reader that hyperonymy is being expressed.

> Les haut-parleurs sont des périphériques de sortie qui permettent de restituer du son à partir d'un courant électrique en provenance de l'unité centrale.

> RNA is a polymer that contains ribose rather than deoxiribose sugars.

Sont des links *haut-parleur* and *périphérique de sortie* in an hypernymic relationship in French, as *is a* links *RNA* and *polymer* in the same way in English. If it is found that *sont des* and *is a* will predictably and recurrently be used in hypernymic contexts, they can be termed hypernymic patterns.

Because these patterns have been found to consistently occur in contexts expressing conceptual relations, knowledge extraction software can be programmed with a set of patterns for a given relation. A search window[4] is specified by the programmer. This technology makes it quicker and easier to locate contexts showing a specific relation. Some studies which have specifically examined using patterns for knowledge extraction tools include Ahmad and Fulford (1992), Hearst (1992), Borillo (1996), Markowitz (1986), Bowden *et al* (1996), Condamines and Rebeyrolle (1998), and Davidson (1998).

1.5 Types of Patterns

Meyer *et al* (1999, 259) have proposed that there are three basic types of patterns: grammatical, lexical and paralinguistic.

- Grammatical patterns make use of the order of parts of speech which can denote a conceptual relation in certain instances. For example, the pattern NOUN + VERB has been found to indicate functionality (e.g. The memory (*noun*) stores (*verb*) information), and the pattern ADJECTIVE + NOUN can denote a property (e.g. readable (*adjective*) disk (*noun*)).

[4] The number of characters to the left or right of a pattern.

- Lexical patterns can be formed by a word or a group of words. Lexical patterns can include *is a/the* for hyperonymy, *contains* for meronymy, and *is used for* for functionality.
- Paralinguistic patterns can be composed of punctuation and elements of how a text is structured (Meyer, 1999, 11). For example, parentheses, commas setting off appositions, and colons have all been shown to indicate hyperonymy. Limited research has gone into determining the efficacy of certain paralinguistic patterns; for example, Davidson *et al* (1998) have briefly examined these patterns. They will be discussed in more detail in Sections 3.1.2 and 4.4.

Because grammatical patterns are not very pertinent to hyperonymy, this study will be restricted to the identification of lexical and paralinguistic patterns. For English, the set of patterns used for this study was compiled by Davidson (1998). For French, the list of patterns was developed through an analysis of hypernymic contexts in computers and genetics texts. These patterns will be listed in Section 3.1.1.

1.5.1 Generic vs. domain-specific patterns

In this thesis, generic patterns will be presented separately from those limited to one particular domain (see Chapter 3). Generic patterns are those which, intuitively, appear likely to indicate hyperonymy in a wide variety of domains. *Est un/sont des* is one such pattern which occurred in both the computers and genetics corpora:

Computers:

> La RAM **est une** mémoire volatile : on peut y lire et écrire à volonté mais toute coupure détruit son contenu.

Genetics:

Une enzyme **est une** protéine, donc une grande molécule.

One can foresee that this pattern could be used in any[5] domain.

In opposition to generic patterns are what are known as "domain-specific" patterns. It has been found (Meyer *et al*, 1999) that many patterns are generic to most domains (*e.g. is a*). However, it appears that many are domain-specific. As the name implies, domain-specific patterns are those which only denote a relation in one particular domain. Some English domain-specific hypernymic patterns which have already been identified include *flavour of* for computers[6] and *species of* for biology (Davidson *et al*, 1998, 54). This thesis examines the question of domain-specificity for computers and genetics and found that, indeed, there are a limited number of patterns specific to these domains. These findings will be presented in Section 3.3.

1.6 Suitability of Hyperonymy as a Relation for Study

> Another value of definition is its insistence on clarifying the writer's thoughts. Henri Bergson suggests we can never be sure of understanding something until we express it in words. We may carry the point one step further: verifiable knowledge of a concept depends on our ability to define it (Darian, 1982, 28).

Hyperonymy is the relation that was chosen for analysis in this thesis. There are many reasons why this relation is a suitable choice. First and foremost, hyperonymy is the most fundamental conceptual relation. Since I am working principally on French patterns, and since only a limited amount of work has been done for French patterns, it

---

[5] Of course, this cannot be proven until all domains have been studied.
[6] Example: Windows NT is a flavour of Windows. The descriptive use of this word was most likely adopted from its use to designate a kind of ice cream.

makes sense to research hyperonymy first.

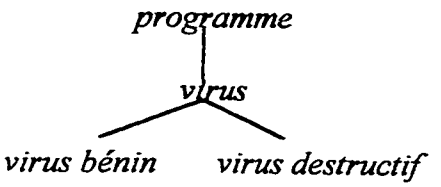
As well, while other relations can also be a good source of domain knowledge, I have decided to concentrate on the relation that is often used in the definition of terms. Flowerdew, who compiled a corpus of scientific lecture transcriptions, found that the most common form of definition occurring in his corpus was indeed the Aristotelian definition, which expresses an hypernymic relation (1992b, 166). Sager also states that hyperonymy is the most common relation (1990, 30). Thus, for reasons of frequency of occurrence alone, this relation is worthy of study.

In his article *Automatic Acquisition of Hyponyms from Large Text Corpora*, Hearst points out that the hypernymic relation is perhaps "particularly amenable" to an analysis using linguistic patterns to locate useful contexts because of its "naming" nature (1992, 542). He also suggests that this is an easy relation to understand, in that linguistic intuition can be used to understand hypernymic statements. Thus, someone with very limited domain knowledge can easily begin to grasp the hierarchical structure of a domain based on hypernymic statements. This idea is also expressed when Borillo states that "à partir de ces schémas purement linguistiques, la relation d'hyperonymie peut être repérée sans qu'il soit nécessaire d'injecter des connaissances concernant le domaine auquel renvoient les textes" (1996, 123). Consider the following statements, taken from the French computers corpus:

> Les VIRUS sont des programmes qui se dupliquent tout seul dans l'ordinateur sans que l'utilisateur s'en rende compte...
>
> Un virus bénin est un virus conçu pour ne rien endommager sur votre ordinateur.

Un virus destructif est un virus qui endommage les données stockées sur votre ordinateur, parfois de façon non intentionnelle.

Based on the information conveyed in these very basic statements, part of hierarchy for the term *virus* can be constructed:



1.7 How This Study Differs from Those Mentioned

The most obvious difference between this study and those mentioned is that none has been done from a comparative linguistics standpoint. This study differs in that it examines patterns inter- and intra-linguistically.

In addition, most of the research mentioned above was carried out on either French or English, with more work having been done on English patterns. Only Borillo (1996), Condamines and Rebeyrolle (1998), Davidson (1998), and Bodson (1999) have worked on patterns in French.

As well, not many studies have set out to observe which patterns are the most frequent[6], and most studies have not focussed on comparing how patterns differ by domain. Borillo (1996) conducted a study using scientific texts[7] to discover generic hypernymic patterns exclusively in French; Hearst (1992) produced a study of patterns

---

[6] Evens (1988) conducted a study on the frequency of hypernymic patterns in dictionary entries.

[7] Borillo specifies that these scientific domains include medicine, biology, geomorphology and economics (1996, 113).

for the same relation in English. Davidson's study, while bilingual, dealt with only one domain (composting) and did not explicitly compare linguistic information on the two languages.

Ahmad and Fulford (1992) studied the patterns in English for the relations of cause, hyponymy, meronymy and synonymy. These patterns were tested for accuracy using a knowledge extraction tool, whereas my study involves no testing and is purely linguistic, similar to the approach taken by Pearson (1998).

Markowitz (1986) also examined the hypernymic relation, but with very different goals from mine. She based her study solely on contexts found in the Webster's Seventh Collegiate Dictionary rather than naturally-occurring texts, with the goal of extracting information that an automatic lexicon builder would use to create lexical entries automatically.

Bowden *et al* (1996) examined non-domain-specific knowledge extraction using positive and negative triggering[8] of relations and combinatorial pattern-matching for extracting conceptual information in English. I have not looked at negative triggering, and he has not examined the statistics for the patterns themselves.

Laura Davidson's study is the one that most closely resembles mine, in that we both examined both English and French knowledge patterns. I, however, chose to concentrate solely on hyperonymy, while her study included hyperonymy, meronymy and functionality. Much of Davidson's study was geared to testing and refining the patterns as well as to reporting how productive or reliable they were. As well, because she took

---

[8] Patterns such as "define" are positive triggers, while "cannot be defined" is a negative trigger. Negative triggers reject contexts in which the original pattern is used in the wrong sense (1996, 151).

her examples from both texts from the Internet and TERMIUM (1996) records, she was restricted to a domain and terms which would be available in both sources, whereas I was able to choose two domains which are constantly being updated as new technology is invented: computers and genetics.

CHAPTER 2 - DETAILS OF THE METHODOLOGY

2.0 Introduction

This chapter presents the steps taken in carrying out research for both the intra- and inter-linguistic comparisons of hypernymic patterns. The major steps include choosing the domains, building the corpora, using Wordsmith, choosing terms, discovering patterns, and learning about knowledge extraction tools.

2.1 Choosing the Domains

In selecting the domains for study, three factors were considered. First, the domains should be a rich source of hypernymic contexts. This criterion was easily satisfied, as most fields have some sort of hierarchical structure. Second, the domains needed to be well represented on the Internet, since that was to serve as the principal source of texts. Third, the two domains had to be significantly different to allow an intra-linguistic comparison of patterns. This was very important, because if the fields were too similar, the variety of patterns used might not be as great.

The domains chosen for this study are computers (fundamental computer literacy concepts) and genetics. These choices reflect the type of domains that terminologists often study: both are constantly changing as new advances are being made. For instance, in computers, the development of the Internet required a whole new terminology to be developed in both English and French, and the GM (genetically modified) food developments in genetics have added terms such as *Bt corn* to our vocabulary. As both domains are widely covered in the news, related articles are widely available on the Internet.

2.2 Corpus Building

Once the domains had been selected, the next step was building the corpora that would be used as a source of hypernymic contexts.

2.2.1 Explanation of a corpus

A corpus is "a collection of naturally-occurring language texts... chosen to characterize a state or variety of a language" (Sinclair, 1991, 171). These days, the term *corpus* usually implies electronic format, which allows software such as concordancers to search corpus data. Corpora can be of a general or specialized nature. General corpora are comprised of non-specialized texts covering a variety of subjects. They can be used by lexicographers to see how words are used in general language (Rundell and Stock, 1992, 22). Specialized corpora tend to be smaller, since they are usually restricted to one domain. These can be used by terminologists to study terms in a specialized domain.

In addition to general or specific, a corpus can be either "closed" or "open," or, to use the terminology of Rundell and Stock, "static" or "dynamic" (1992c, 47). A closed corpus is one which is finite – that is, once it has been constructed, no new texts are added. An open corpus is one which has texts added to it or deleted from it as new information develops and old information becomes obsolete. While an open corpus offers the advantage that it is constantly being updated and can thus be used to monitor changes in information or in a language (Rundell and Stock, 1992c, 48), our own study is based on four closed corpora, as it was carried out over a short period of time (approximately one year).

2.2.2 Balancing and skewing

According to Davidson "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 writing style or mode of expression" (1998, 10). This is what is known as balancing a corpus.

This study aimed to observe hypernymic patterns as found in real texts. Therefore, it was important to select texts which would be a rich source of hypernymic contexts. For this reason, the corpus was deliberately skewed in favour of texts reflecting certain communicative settings[1] that generate many hypernymic contexts. Therefore, the corpora are intentionally somewhat skewed in favour of moderately technical texts, as opposed to highly technical ones.

2.2.3 The Internet as a source of texts

The Internet represents an immense repository of knowledge on a wide range of subject matter. It can be searched using various search engines such as Google, Alta Vista, Yahoo, Webcrawler, and Northern Light. Search engines will locate possible sites for a word or phrase entered by the user. It is up to the user to filter through these sites and decide which are relevant to the search query, a process which can be quite time consuming. Because of this, many major search engines have set up collections of hierarchically-organized links to many major fields of study. For instance, Yahoo contains a collection of links for biology, which includes the subcategory of genetics, which, in turn, includes such sub-domains as cloning, bioengineering, and gene therapy.

---

[1] Communicative settings will be discussed in Section 2.2.4.

This type of structure provides corpus-builders with a rough outline of the structure of the domain, and gives them an idea of the different subject matter the texts should present.

The Internet was the principal source of texts for all four corpora, although the French genetics corpus also includes texts from the 1989 - 1998 *Science & Vie* CD-Rom.[2] Using the Internet as a source has several advantages. These include ease of access from any computer with a modem, eliminating the need to scour multiple libraries. Also, it provides data in electronic format, which is very important in a study of this nature, where much time could be wasted manually converting the data. As well, the Internet often presents new research before the printed form appears, making it a convenient way to locate very up-to-date information. However, anyone is free to publish information on the web, and care must be taken to ensure that the information is credible.

Once I had decided on my domains and the source of text, I needed a starting point for locating information. The most efficient way to begin my search seemed to be to look up the subject in the Yahoo on-line publications links. As these are categorized by domain and subdomain, this provided an easy way to find links to sites for both of my domains. Most sites that I found contained a section on related links, which were investigated as well.

Once I had exhausted all the links from the Yahoo categories, I used Alta Vista to locate supplementary texts. I typed in key words and investigated the documents that matched my search. As much as possible, I tried to avoid saving the same text twice. This is very difficult, however, as many sites copy from one another.

---

[2] The CD-ROM contains full texts of articles appearing in the publication between 1989 and 1999.

2.2.4 Factors influencing the choice of texts

According to Cole, terminologists are looking for "documents which will provide the maximum number of terms employed clearly and correctly, accompanied by definitions or indications of the concepts that they designate" (1987, 80). Because my specific goal was to extract hypernymic contexts from these texts, the texts had to be of a somewhat explanatory nature.

According to Darian (1982, 29), the greater the gap in the knowledge level of the author and the audience, the greater the need to define. For instance, texts authored by experts for an audience with little or no background knowledge (non-specialists) of the subject would be more likely to contain hypernymic contexts than texts written by experts for other experts. This is because defining the basic concepts in a text targeting other experts would be superfluous. Pearson states that when experts in the same field communicate (expert-expert communication), "they tend ... to use a highly specialized jargon. It is assumed that the author and the reader share a common language and that when certain words or phrases are used, each understands what is meant" (1998, 36). Therefore, for the purposes of this study, texts written by experts for experts were rejected, as they are unlikely to contain many definitions.

In fact, the ideal text-type that I was looking for was that of a text written by experts for non-specialists. Pearson classifies this text-type into three "communicative settings":

- expert to initiates – initiates are persons who have some domain knowledge, but not as much as a real expert *e.g.* students of a particular domain
- relative expert to the uninitiated – a relative expert could be a journalist with a good level of domain knowledge, and the uninitiated could be adults with little domain

knowledge. Typical examples include popular science journals such as *Scientific American.*

- teacher-pupil – in this case "pupil" is reserved for people with no domain knowledge, but who need to acquire it for educational purposes (1998, 37-38).

Therefore, I tried to locate as many of these types of texts as possible. Some of the types of texts I included were those from popularized magazines, on-line newspapers, and class notes posted on the Internet for introductory university courses. The French genetics corpus also includes articles taken from the 1989 - 1998 CD-ROM version of *Science & Vie*. *Science & Vie* is a popularized scientific journal which targets a generally-educated readership rather than scientists. It contains explanatory articles dealing with general scientific issues.

2.2.5 Size of the corpora

To determine the size of the corpora, I made use of Wordsmith's wordcount feature. The French computers corpus is approximately 500 000 words, while the French genetics one is 575 000 words. For English, the computers corpus comprises 793 000 words, while for genetics there are 451 000.

Pearson suggests that a special-purpose corpus should contain approximately one million words; however, she qualifies this by saying that the "phenomena which [compilers] are investigating will appear with sufficient frequency in their smaller corpora to give them adequate results." For instance, she provides the example that Glendhill worked with a 500 000-word specialized corpus (1998, 56-57), which is approximately the same size as the French and English genetics corpora used. Lynne

Bowker (personal communication) also reports that she is able to obtain interesting findings with corpora in the 500 000-word range.

2.3 Tools for Extracting Information from the Corpus

There are many software programs which have been developed to extract the information contained in corpora. Most of these contain some form of concordancer. Some of the concordancers that are available on the market include Wordsmith, Multiconcord, and System Quirk. The concordancer selected for this study is Wordsmith, which is the concordancer used in courses taught at the School of Translation and Interpretation.

2.3.1 About Wordsmith

Wordsmith is a software program that has been developed by Mike Scott[3] and that is marketed by Oxford University Press. It can be used in a variety of ways to extract information from a corpus. It provides statistical information on the corpus, including the total word count, the number of different words used, and the frequency of individual words. It can also generate KWIC concordances, and allows a wide variety of sorting features. Wordsmith was used both in the term discovery and context extraction phases.

2.3.2 Dealing with accented characters in Wordsmith

When I began using Wordsmith on the French corpora, it became apparent that they were not ready to be used "as is." The files had been saved in HTML format. Some

[3] Information as well as a trial version of Wordsmith can be found on Mike Scott's web page at http://www.liv.ac.uk/~ms2928/wordsmith/screenshots.

HTML documents use SGML tags for accented characters and certain punctuation marks. When the documents are viewed by a web browser such as Netscape or Internet Explorer, the tags are automatically converted to the character they represent. For example, the SGML tag "&eacute" appears as "é" when read by the browser. However, Wordsmith requires users to write a tag file containing all the tags that are not translated and the characters they represent. Occurrences of tags not translated can easily be located by using the Wordlist feature and looking at the alphabetical statistics. These types of tags all begin with "&," which means they are grouped together. The tag file must be saved with a .tag extension, and activated in the "Translate Tags" section of the "Adjust Settings" tab. After I had drawn up and activated the tag file, Wordsmith was able to display the accented characters.

2.4 Discovering Terms in Computers and Genetics

One of the terminologist's tasks is to discover the terms (sometimes called the *nomenclature*) of a given domain (Cole, 1987, 79). While I did not seek to establish an exhaustive list of terms for computers and genetics, I needed to go about discovering terms in very much the same way as an actual terminologist would. Therefore, I began looking for terms and reading to acquire domain knowledge while I was building the corpora. While quickly scanning the texts, I took note of many of the terms which appeared in hypernymic contexts. I used many of the terms discovered in this way as search terms when I later ran concordances on the finished corpora. In addition, I used the Wordlist component of Wordsmith to identify possible terms. Wordlist essentially counts the words in a document and provides the user with various statistics on the

corpus. Some of the most important information it provides include the total number of words in the corpus, the number of different words in the corpus, an alphabetical list of all words counted, and a list of words in order of frequency of occurrence. I made use of the frequency of occurrence list in the selection of further terms for concordances, since terms in a domain often recur frequently.

2.4.1 Choice of Terms

I was not interested in *all* possible search terms, but rather in those which would be likely to yield hypernymic contexts. I therefore ran many KWIC concordances, retaining those terms that were obviously rich in hypernymic contexts.

2.4.2 Terms selected

Computers	Genetics
unité centrale	*transcription*
virus	*ARN*
ethernet	*OGM*
modem	*clonage*
périphérique	*transgénèse / transgène*
imprimante	*mitose*
bus	*séquence*
mémoire	*macrophage*
disque dur	*traduction*
processeur	*protéine*

I also kept contexts which showed my search term as part of a compound term. Here is one such example:

> L'ARN POLYMÉRASE est l'enzyme qui catalyse la synthèse d'ARN messager à partir d'une matrice d'ADN.

My original search term was *ARN*, but the Wordsmith concordancer shows all occurrences of the search term. I decided to include subtypes of my terms if they came up, since the hyponyms inherit the characteristics of the hypernyms. In this way, this context shows that *ARN polymérase*, like *ARN*, is also an enzyme.

In addition, when I noticed certain patterns (constitu*[4], désign*), I ran concordances on them to confirm that these were indeed patterns, and to find additional hypernymic contexts. This process yielded some additional terms, and for this reason, there are certain hypernymic contexts cited in this thesis for terms that are not on the above list.

2.5 Discovering Patterns in French

Discovering patterns in French involved examining all occurrences of a search term found by Wordsmith Concordancer. Below is a sample of a concordance search on the term *mitose*. Originally, in the KWIC display, only the line in which the term appeared was shown by the concordancer, but I have adjusted the settings of Wordsmith so that more of the context is shown:

---

[4] The asterisk represents what is known as a "wildcard search." It enables the user to search for all inflections of a term. This can be useful if the user is searching for occurrences of both singular and plural forms of a term, or, in this case, verbs used in any tense.

MITOSE: 12 entries (sort: File,File)

N Concordance

1 petit de la lionne et du tigre ne peuvent avoir de descendance. La règle 1, qui interdit la plupart des croisements interspécifiques, s'explique par le processus de la **mitose**. On désigne ainsi la division cellulaire qui est a l'origine de la croissance de l'individu, de la conception jusqu'a la mort: une cellule se divise pour donner deux

2 à la cytosine (G-C). C'est cette complémentarité des bases qui explique que le message d'un brin soit le calque de l'autre. Lors de la division cellulaire (**mitose**) quand les deux brins d'ADN se dupliquent et se séparent pour donner deux molécules filles, il est fréquent que des bases non complémentaires se mettent

3 era pourvue d'un stock normal de chromosomes (2n), par réunion d'un spermatozoïde et d'un ovule. La méiose étant beaucoup plus sophistiquée que la **mitose**, la lignée cellulaire qui se met en place a la puberté pour fabriquer les gamètes est tout a fait incapable d'en franchir les obstacles s'il y a eu croisement

4 paire, un chromosome venant du père et l'autre de la mère. Sur notre dessin, toujours par souci de lisibilité, l'oeuf ne porte que trois paires de chromosomes. 3. La **mitose** est un processus grâce auquel la nouvelle cellule va se diviser pour donner deux cellules filles, lesquelles, a leur tour vont se diviser pour en don

5 lle cellule va se diviser pour donner deux cellules filles, lesquelles, a leur tour vont se diviser pour en donner deux chacune. Et ainsi de suite. Schématiquement, la **mitose** se déroule de la manière suivante: 3 a. Chaque chromosome se dédouble et on en obtient deux identiques relies en x par leur centre. 3 b. Les paires de chromo

6 une des deux ébauches de cellule. 3 d. Enfin, les deux ébauches finissent de se séparer en deux cellules filles, contenant chacune 23 paires de chromosomes. La **mitose** est un processus qui se répète sans cesse, depuis la fécondation. Elle sert a multiplier les cellules: pour donner, a partir de l'oeuf, un embryon; pour

7 ividu complètement formé, pour la croissance, le développement, l'entretien et le renouvellement de ses tissus pendant toute sa vie, jusqu'a sa mort. 4. La **mitose**. Lorsque l'individu, d'homme ou femme, arrive à maturité sexuelle, le processus suivant intervient dans certaines de ses cellules, qui aboutira a la fabric

8 hromosomes (3 paires seulement sur le dessin), chaque chromosome se dédouble, comme précédemment et prend son aspect en X. 4 b. Comme dans la **mitose**, les paires de chromosomes dédoublés se regroupent vers le plan équatorial de la cellule. Au cours de cette

9 chromosomes, un doublon entier (les deux brins de l'X) a migré dans une des deux cellules filles, et l'autre dans la deuxième. C'est là la grande différence avec la **mitose**: les deux cellules filles n'ont, ci, que 23 chromosomes (quoique dédoublés) et non 23 paires. 4 d. Chacune de ces cellules filles se divise alors sel

10 es filles n'ont, ci, que 23 chromosomes (quoique dédoublés) et non 23 paires. 4 d. Chacune de ces cellules filles se divise alors selon le processus de la **mitose** : chaque double chromosome se scinde en ses deux chromosomes et on obtient quatre spermatozoïdes à 23 chromosomes. Le même p

11 ès la fécondation de la femme par l'homme, et la formation d'une cellule oeuf née de la rencontre d'un spermatozoïde avec l'ovule, s'enclenche le processus de la **mitose** qui permet la formation d'un embryon, puis d'un foetus et qui se répète en permanence pour assurer le développement et l'entretien des cellules de l'indi

12 enzymes arrêtent le processus: mules et mulets sont stériles. 4. Si l'on croise deux espaces plus lointaines, par exemple), les écarts sont trop grands; ni **mitose** ni méiose ne passent le contrôle des quatre enzymes. 5-6. En donnant congé aux quatre contrôleurs (ce qu'on sait faire aujourd'hui) rien ne s'oppose plus au m

Sentences 1, 2, 4, and 6 are clear examples of hypernymic contexts. Once it was determined which were hypernymic contexts, these could be evaluated to determine if there was a possible pattern signalling the relation. For instance, just by examining these four hypernymic contexts, it is possible to hypothesize that *désign** (context 1), parentheses containing the search term (context 2), and *est un* (contexts 4 and 6) are patterns found in hypernymic contexts. However, to be productive in a knowledge-

extraction tool, the pattern must be recurrent, meaning that it must repeatedly indicate the desired relation. This procedure was repeated for all the search terms in an attempt to locate as many hypernymic patterns as possible and chart their frequency.

2.6 Learning about Knowledge Extraction Tools

In this thesis, I examine the implications my findings might have for advanced knowledge extraction tools, in particular those which make use of a pre-programmed set of patterns to locate only the most useful contexts for terminologists. In order to gain a basic understanding of how such tools work and the types of issues that might be problematic for them, I acquired hands-on experience with a knowledge extraction tool called the Text Analyzer.

2.6.1 The Text Analyzer

The Text Analyzer (TA) is a component of DocKMan (Document-Based Knowledge Management) which is being developed by Doug Skuce *et al* in the LAKE lab (Language Analysis for Knowledge Engineering) at the University of Ottawa's School of Information Technology and Engineering (Kavanaugh *et al*, 1999). It is able to extract various types of information from documents. It can be used to perform operations on a corpus including concordances and question answering for certain relations (Kavanaugh *et al*, 1998, 3). DocKMan has been programmed with various English lexical and paralinguistic patterns which indicate certain relations. Users are required to specify which relation they are most interested in and input a search term. For each relation, DocKMan has been programmed to locate certain patterns occurring

within a specified distance to the left or right of the search term. The relations that DocKMan can presently handle, for English only, are hyperonymy, meronymy, functionality, causality, and synonymy.

As a research assistant for the COGNITERM Project directed by I. Meyer, I used the TA extensively and consulted with its programmer on various issues. I carried out research in English for hyperonymy, meronymy, and functionality. This gave me a solid idea of how the TA worked, and how it was refined. I discovered what kinds of changes had to be made to locate more "hits" (valid contexts found for the desired relation), and how to reduce the number of examples which constituted "noise" (contexts found which did not indicate the desired relation) (Meyer, Mackintosh, Barrière, Morgan, 1999, 258).

After learning how to use the TA for English, I was able to incorporate a small number of my thesis findings for French hypernymic patterns on an experimental basis. Certain patterns were incorporated into the TA, and the French computers and genetics corpora were added. Unlike the English corpora that have been incorporated in the Text Analyzer, the French corpora were not tagged for parts of speech. However, I was still able to conduct preliminary research on the French patterns, to get a basic sense of how productive they might be. It would be premature to report the results of these very simple experiments in this thesis. My purpose, rather, was to acquire the hands-on experience with a knowledge extraction tool that would allow me to fully understand the implications, for tools developers, of the linguistic analysis carried out in this thesis (and discussed in 3.5 and 4.3.2).

CHAPTER 3 – INTRA-LINGUISTIC COMPARISON

3.0 Introduction

This chapter presents the findings of the intra-linguistic comparison in French for the domains of computers and genetics. The research methodology has been outlined in Chapter 2, and includes a description of the corpora. This chapter presents the generic patterns identified and examples of their use (3.1), charts the frequency of patterns in two specialized domains (3.2), analyzes the domain-specific patterns discovered (3.3 and 3.4), and proposes possible implications of these findings for knowledge-extraction tools (3.5).

3.1 Examples of Hypernymic Patterns

Examples of hypernymic contexts for each pattern follow. The patterns are presented from the most to least frequent. For each pattern given, if applicable, the context from the computers corpus precedes the one taken from the genetics corpus. The patterns are in boldface type, the hyponyms underlined, and the hypernyms double-underlined.

3.1.1 Lexical patterns

a) type* (de / d')[1]

There are three variations of this pattern: *de type, type de,* or simply *type* in contexts where *de* has been replaced by *en*. In most cases, the hyponym occurs closest to *type*. For computers, *type* de* is the most common variation, with 51/72 hypernymic

[1] Where applicable, patterns are written the way they appeared on the list of patterns for the test version of the French Text Analyzer. Brackets indicate optionality, while * is a wildcard. (Personal communication with J. Kavanaugh, Oct. 20, 1999)

contexts containing *type*, while *de type* accounted for 21, and *type* with *en* replacing *de* had four occurrences. In genetics, the results were similar in that 15 were *type de*, 2 were *de type*, and none were *type* with *de* replaced by *en*. It is fitting that this is the most frequent pattern in both domain, as the hypernymic relationship in inherent in the meaning of *type*.

type* de (d'):

Il existe actuellement cinq **types de** bus : ISA, EISA, MCA, VLB et PCI.

Il y a trois **types d'**ARN : l'ARN de transfert, l'ARN ribosomaux et l'ARN messager.

type (+en):

Dans un deuxième temps, il devient vite indispensable de se munir d'une imprimante. Il **en** existe trois **types** sur le marché : les imprimantes "Apple Talk" pour Macintosh, "parallèles" et "série" pour les PC.

de (du) type:

Le socket 7 est un support pour les processeurs **de type** Pentium (Les pentium II sont sur un support de type Socket One).

Pour l'instant, l'équipe du Pr Shiloh a réussi à cloner une séquence génétique de 5 900 paires de base du gène ATM codant pour une protéine **du type** P13K.

b) est un(e) / sont des (de)

It is safe to say that this would be a reliable pattern for hyperonymy in any domain.

Les imprimantes **sont des** périphériques de sortie qui permettent de dessiner sur du papier des données en provenance de l'unité centrale.

Un organisme génétiquement modifié (OGM) **est un** organisme vivant dont on a modifié le patrimoine génétique en y insérant un ou plusieurs gènes issus d'un autre organisme vivant.

c) (s')appel*

This pattern can either be used in the pronominal or past participle form. In the latter case, the hyponym follows the pattern while the hypernym precedes it.

> Pendant cette exécution, les résultats de calculs intermédiaires seront stockés provisoirement dans une mémoire très rapide **appelée** SRAM (mémoire cache).
>
> L'ARN messager est utilisé par la cellule pour diriger la synthèse des protéines dans un processus **appelé** traduction.
>
> La période pendant laquelle la cellule est au repos **s'appelle** interphase.[2]

d) est le (la, l') / sont les (l')

> Le modem **est le** périphérique utilisé pour transférer des informations entre plusieurs ordinateurs (2 à la base) via les lignes téléphoniques.
>
> LA TRANSCRIPTION **est le** processus au cours duquel un ARN simple brin, appelé messager, est synthétisé à partir d'une matrice d'ADN.

e) comme

> Outre les périphériques indispensables on peut connecter à une unité centrale divers périphériques de sortie **comme** : Une imprimante Imprimante à aiguille, imprimante à jet d'encre, imprimante laser...
>
> Certains processus cellulaires, **comme** la transcription, génèrent un surenroulement de l'ADN et de tels processus peuvent également exercer des forces de traction sur la double hélice.

f) (se) nomm*

Like *(s')appel**, this pattern verb can be used either as a past participle (nommé*) or pronominally (se nomme*). Interestingly, while 5 of 6 examples in computers included the pronominal form, only the past participle was used in genetics.

---

[2] Context found in the analysis for Chapter 4. Cited to show how the pattern is used in the form of a pronominal verb.

> Mais dans la plupart des cas un bus implique généralement plus de deux dispositifs -- une version plus "moderne" du LocalBus **se nomme** AGP (Advanced Graphics Port)
>
> Les globules blancs **nommés** macrophages sont aussi des réservoirs, mais leur durée de vie est relativement brève.

g) tel(le(s)) (que, qu')

This pattern occurred most frequently without *qu** in the contexts extracted. In these cases, the pattern is *un / une + tel(lle)*, and it often occurs in intersentential hypernymic contexts, as in the second example given below.

> Différents connecteurs permettent d'alimenter la carte mère ainsi que les divers périphériques internes **tels** que lecteurs de disquettes, disques durs, lecteurs CD-ROM, etc.
>
> ... les ARN 16S des procaryotes et 18S des eucaryotes, a permis de construire un arbre universel du monde vivant (voir l'article de Christa Schleper dans ce numéro). Malheureusement, les virus ne possèdent pas un **tel** ARN dans leur génome, car ils utilisent les ribosomes des cellules hôtes pour se reproduire.

Another variation of this pattern could be *no article + tel(lle) + noun*, as in the example *des **périphériques**, tels les lecteurs.* No such examples occurred in the corpora.

h) sorte*

The preposition *de* can either precede or follow *sorte*. In addition, the genetics example may not be found owing to the number of characters between the pattern and the search term.[3]

> Il y a 2 **sortes** de mémoires La mémoire ROM (Read Only Memory) Contient les programmes du fabricant qui permettent de faire le lien entre les périphériques et le système d'exploitation. Ne s'efface pas lorsqu'on ferme le micro Les informations de cette mémoire ne peuvent qu'être lues. La mémoire RAM (Random Access Memory) Elle contient ; le SE (Système d'Exploitation), les logiciels et les données. Mémoire qui perd son contenu quand on éteint le micro.

[3] The number of characters between the search term and the pattern is large because the term is replaced by a pronoun. This will be discussed in Section 3.5.3.2.5.

Les amorces sont de courtes séquences d'ADN monobrin fabriquées en laboratoire. Elles sont de deux **sortes** : les unes correspondent à une extrémité du gène dont on veut obtenir des polycopies (ici la séquence de bases A-G-T-G-T-C de la sonde est complémentaire de la séquence T-C-A-C-A-G de l'extrémité de l'ADN monobrin) ; les autres à l'autre extrémité.

i) constitue*

This pattern could perhaps be seen as an alternative way of expressing *est un / sont des*, or *est le / sont les*. This will be discussed in more detail in Section 4.3.1.2.

Autres périphériques : les autres périphériques **constituent** les scanners et les lecteurs amovibles (zip, jazz...)

Ainsi, dans le cas du maïs, le gène bactérien responsable de la synthèse de la protéine détruisant la pyrale, **constitue** un tel transgène.

j) par exemple

... exprimée en nanosecondes et varie selon le type, l'age et la fonction de la mémoire désirée. **Par exemple**, pour de la mémoire vive, on compte actuellement entre 70 et 50ns, alors que par le passé, cette valeur pouvait atteindre

... en administrant par voie systémique des protéines recombinantes (des protéines produites dans des fermenteurs par des micro-organismes génétiquement modifiés, **par exemple** des bactéries...

k) forme* (de, d')

ROM signifie Read Only Memory (mémoire morte). Une ROM est une **forme d'**emmagasinage permanent.

Les chances de survie seraient encore meilleures si les cellules de moëlle osseuse à l'origine provenaient de la moëlle osseuse du patient lui-même, éliminant du même coup les risques de rejet – et créant donc une **forme de** clonage: le clonage de sa propre moëlle osseuse. Le "clonage thérapeutique", comme le terme commence à circuler.

l) désign*

*Désign** is similar to *constitue** in that it may be an alternative way of expressing *est un / sont des* or *est le / sont les*. It will also be discussed further in Section 4.3.1.2.

> FTP **désigne** donc à la fois un programme et un protocole, de sorte que vous utiliserez peut-être un autre programme que FTP mais régi par les règles FTP.[4]
>
> La transgénèse **désigne** le processus de transfert dans le patrimoine génétique d'un organisme vivant d'un gène qui lui est étranger.

m) il s'agi* de (d')

When followed by a noun phrase, *il s 'agi* de* can indicate hyperonymy.[5]

> Le virus ne mérite nullement l'aura de mystère qui l'entoure : **il s'agit d'**un banal logiciel, écrit avec un langage de programmation, à cela près qu'il n'a aucune utilité pratique et ne vise qu'à nuire.

n) catégorie*

> ARN Deuxième **catégorie** d'acide nucléique présente dans les cellules d'un organisme.

o) classe*

*Classe** was not found in any hypernymic contexts containing the search terms for computers. Since *classe** seemed as though it might be domain-specific to biological domains,[6] it was tested on the computers corpus. This reveals that it does occur in hypernymic contexts for the following terms: *système, fichier, sous-réseau,* and *adresse IP*.[7] This shows that this pattern is generic, though probably less frequent in computers than in genetics.

---

[4] Example found in concordance on the pattern itself done for the comparative linguistics portion of this thesis.

[5] Hyperonymy was expressed in 22 out of 100 and 45 out of 100 contexts from the genetics and computers corpora respectively. In all cases, a noun phrase followed the pattern.

[6] *Classe* is one of the levels of classifications in taxonomy.

[7] Examples: On distingue deux **classes** de système d'exploitation actuellement : les 16 bits et 32 bits.
Il existe le plus souvent deux **classes** de fichiers : les fichiers exécutables qui contiennent un programme et les fichiers de données.
On trouve trois **classes** principales d'adresse IP. Si le premier octet a une valeur de 0 à 127

Acides nucléiques : Divisés en deux **classes**, les ADN et les ARN, ces molécules portent en elles les instructions héréditaires permettant la transmission et le développement de la vie.

p) y compris

Although this was not a very productive pattern for the search terms, semantically it does denote hyperonymy, and it is worth including as a generic pattern. Concordances run on the pattern itself in both corpora yielded promising results, especially when the search term directly preceded or followed the pattern.[8]

Ils doivent pouvoir réparer les dommages dus aux virus (les supprimer en restaurant les données corrompues). Ils doivent pouvoir protéger l'ordinateur contre toute attaque du virus connu probable, **y compris** les macrovirus.

3.1.2 Paralinguistic patterns

There were five main paralinguistic patterns for hyperonymy noticed upon examining the data:

- **dashes**

  Ces derniers sont des éléments génétiques composés d'un seul type d'acide nucléique – **ADN ou ARN** – et ils sont incapables de se reproduire seuls.

- **commas setting off appositions**

  Cette division, **ou mitose**, se produit au terme d'un ensemble de processus que l'on appelle le cycle cellulaire.

- **hypernymic contexts following questions containing the search term**

---

c'est une adresse de classe A...

[8] Hyperonymy was expressed in 21 out of 35 and 10 out of 12 contexts retrieved from the genetics and computers corpora respectively.

Qu'est-ce qu'un OGM? Un organisme génétiquement modifié (OGM) est un organisme vivant dont on a modifié le patrimoine génétique en y insérant un ou plusieurs gènes issus d'un autre organisme vivant.

- **a colon preceding or following the search term**

La copie se fait, gène par gène, ou par groupe de gènes, grâce à un enzyme : l'ARN polymérase.

Modem (Modulateur-DEModulateur) : périphérique qui convertit les données numériques d'un ordinateur ou d'un terminal en données analogiques qui peuvent être envoyées via les lignes téléphoniques.

- **parentheses**

Enfin, les échanges d'informations s'effectuent entre les circuits de l'ordinateur par les bus **(bus de données pour les données et bus d'adresses pour les adresses)**...

Only parentheses have been programmed into the current version of the TA for English. .For this reason, I decided to examine only this paralinguistic pattern in detail. Therefore, this was the only paralinguistic pattern represented in the statistical analysis.

The data retrieved for this study have revealed that the information contained in the parentheses can be either an hyponym or an hypernym.

Parentheses containing hyponym(s):

Alors que l'IDE est réservé aux disques durs, SCSI est applicable à tous les périphériques (Cd-rom, imprimante, scanner, Hard-drive, etc...).

Lors de la division cellulaire (mitose)[9] quand les deux brins d'ADN se dupliquent et se séparent pour donner deux molécules filles, il est fréquent que des bases non complémentaires se mettent...

Parentheses containing hypernym:

... sur deux patientes atteintes de thalassémie, une forme d'anémie due à un manque de bêta-globine (une protéine entrant dans la composition des globules rouges).

[9] Though *mitose* may appear to be synonymous with *division cellulaire*, it is not, since *méiose* is another type of *division cellulaire.*

... l'Interim Licensing Authority, qui s'oppose aux objectifs « indésirables » comme la modification des gènes de l'embryon, le clonage **(reproduction par bouturage)** et le mélange de cellules animales et humaines.

3.2 Statistical Analysis of Generic Patterns Identified

In an introductory manual on statistics, Cohen (1954, 2) states that "statistics provides new ways of looking at data and new ways of manipulating data." This section will analyze the patterns from a statistical point of view[10] to see which are most frequent for the search terms. Looking at the patterns in this way will provide insight into which appear to be the most important in terms of retrieving hypernymic contexts. The statistics on pattern frequency will represent the general tendencies on which generic patterns are most widely used; however, this is not to say that the results would be identical if different search terms or different domains were used.

The following is a list of generic patterns found in both the computers and genetics corpora and their frequency. This list is not meant to be exhaustive; it simply presents the hypernymic patterns that were found in contexts extracted for the search terms. Any inflections of a pattern are not counted separately; for instance, *est un* and *sont des* are considered the same pattern. The asterisks allow users to search for variations in patterns; for example *il s'agi* de (du, d')* and *type* de (du, d')* will locate instances of *il s'agit de, il s'agissait de,* and *type de, types de* respectively.

---

[10] Of course, given the limited size of this study, our analysis does not claim to be "statistically significant" in the technical sense of the term. Our goal was simply to get a general impression of frequency.

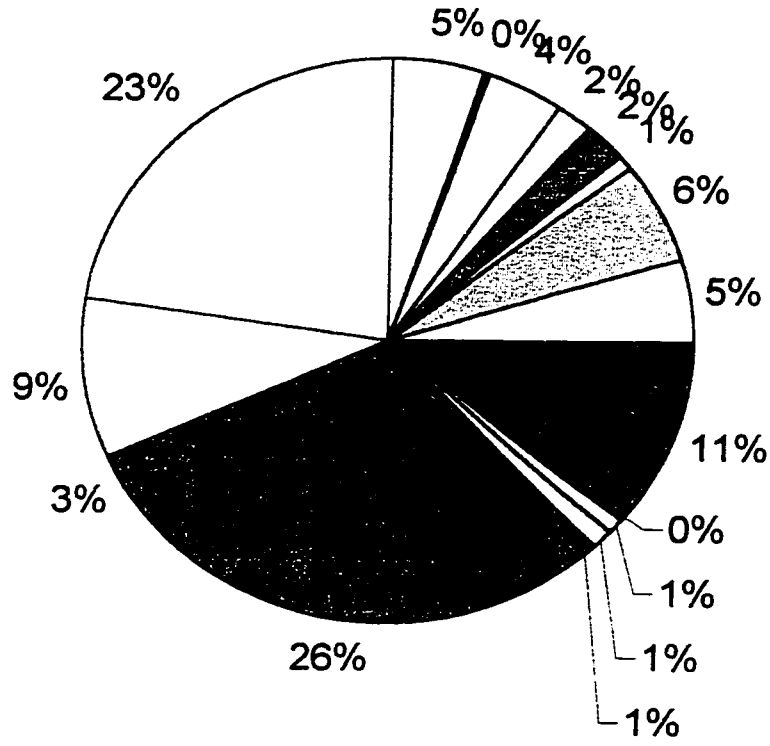
For the statistical analysis, only contexts containing the original list of search terms were included. Hypernymic contexts found for other terms over the course of analysis were not included. Every occurrence of the patterns listed below was counted, as long as the pattern denoted an hypernymic relation for one of the search terms. This means that hypernymic contexts which included two different patterns, such as the following, were counted twice:

> La mémoire VRAM (Video RAM) **est un type de** mémoire qui a l'intéressante propriété de pouvoir être accédée en lecture et en écriture de façon simultanée, car contrairement à la DRAM elle possède deux chemins d'accès.

This example was counted once for *est un** and again for *type**, because these have been identified as two distinct patterns. However, it was more commonly the case that only one pattern was used per context.

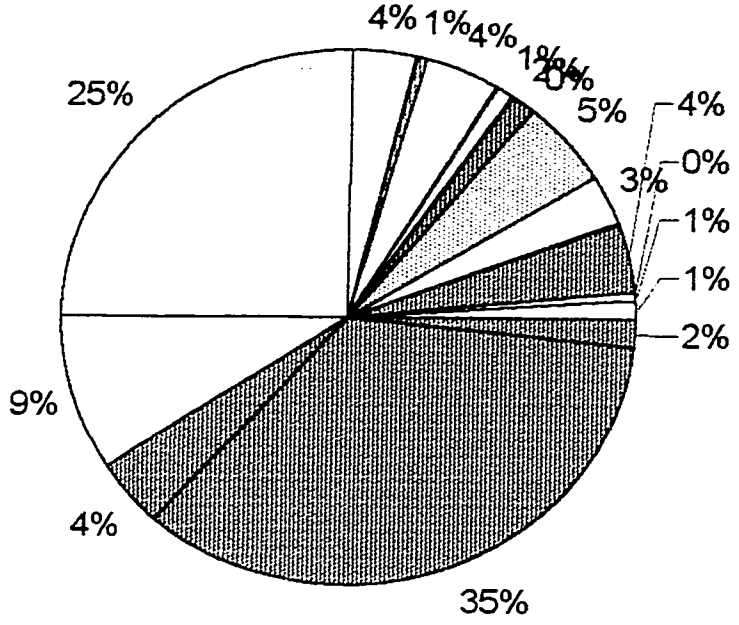
Pattern:	Frequency in Genetics:	Frequency in Computers:	Total:
type* (de, d')	17	67	84
est un(e) / sont des (de)	24	48	72
(s')appel*	26	8	34
est le (la l') / sont les (le)	11	18	29
comme	8	10	18
(parentheses)	9	7	16
(se) nomm*	9	6	15
tel(le(s))(que)	5	8	13
sorte* de (d')	1	8	9
constitue*	4	3	7
par exemple	4	2	6
forme (de, d')	2	2	4
désign*	3	0	3
il s'agi* de (d', du)	0	3	3
catégorie*	1	0	1
classe* (de, d', du)	1	0	1
y compris	0	1	1
Total:	125	191	316

Total Percentage in Both Domains



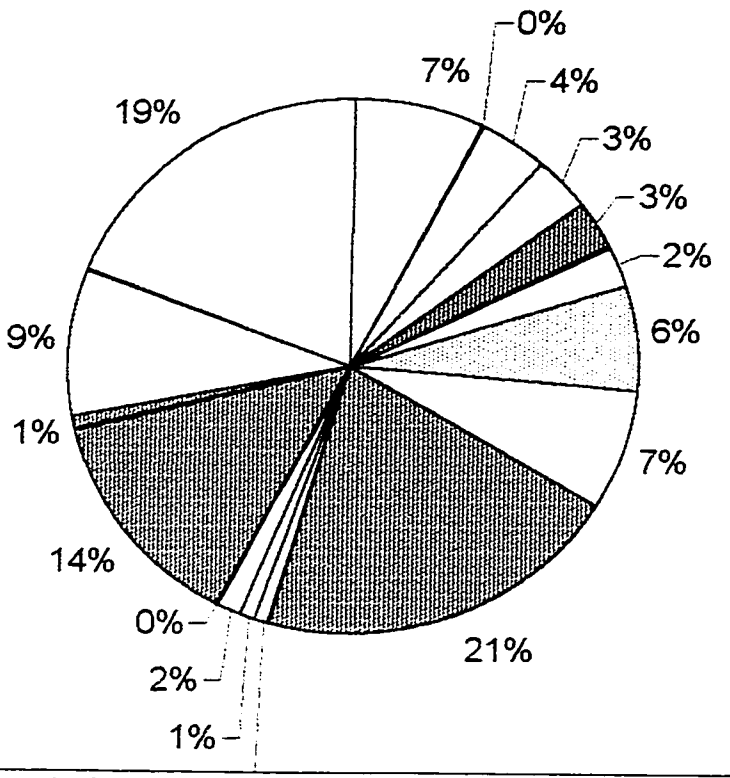
parentheses (5%)	y compris (0%)	tel* (qu*) (4%)
par exemple (2%)	constitue* (2%)	désign* (1%)
comme (6%)	nomm* (5%)	(s')appel* (11%)
classe* (0%)	catégorie* (1%)	forme* (d*) (1%)
il s'agi* de (d') (1%)	type* (26%)	sorte* de (d') (3%)
est le (la, l') / sont les (le) (9%)	est un (une) / sont des (de, d') (23%)	

Percentage Breakdown of Patterns for Computers



parentheses (4%)	y compris (1%)	tel* (qu*) (4%)	par exemple (1%)
constitue* (2%)	désign* (0%)	comme (5%)	nomm* (3%)
(s')appel* (4%)	classe* (0%)	catégorie* (1%)	forme* (de, d') (1%)
il s'agi* de (d') (2%)	type* (35%)	sorte* de (d') (4%)	est le (la, l') / sont les (le) (9%)
est un (une) / sont des (de) (25%)			

Percentage Breakdown of Patterns for Genetics



parentheses (7%)	y compris (0%)	tel* (qu*) (4%)	par exemple (3%)
constitue* (3%)	désign* (3%)	comme (6%)	nomm* (7%)
(s')appel* (21%)	classe* (1%)	catégorie* (1%)	forme* (de, d') (2%)
il s'agi* de (d') (0%)	type* (14%)	sorte* de (d') (1%)	est le (la, l') / sont les (le) (9%)
est un (une) / sont des (de) (19%)			

3.2.1 Analysis

As the chart indicates, a total of 316 instances of hypernymic patterns were identified, with 191 in computers and 125 in genetics. The pie graph for the total percentages of patterns in both domains reveals that overall, the five most productive patterns were *type* (de, d'), est un(e) / sont des (de), (s')appel*, est le (la, l') / sont les (le),* and *comme*, representing 26%, 23%, 11%, 9%, and 6% of hypernymic patterns respectively. There were six patterns that each represented less than 1% of the total hypernymic patterns used: *y compris, classe* (de), catégorie*, forme* (de), il s'agi* de (d')* and *désign**.

Looking at the two domains comparatively, one of the first things that can be noticed is that the patterns are more evenly distributed in genetics than they are in computers. For computers, 2 patterns represented 60% of hypernymic patterns used (*type** (35%) and *est un(e) / sont des (de)* (25%)), while the two most frequently used patterns in genetics were *(s')appel** and *est un(e) / sont des (d')* at 21% and 19% respectively, for a total of 40%. Though 40% is still a large figure for two patterns, it is considerably smaller than 60%. Interestingly, while the percentage of hypernymic contexts containing the second-most productive pattern in both domains, *est un(e) / sont des (de),* is relatively close in computers and genetics (25% and 21%), the most productive pattern for genetics (*(s')appel** at 21%) represented only 4% of hypernymic patterns in computers, and *type*(de, d')*, while accounting for 35% of hypernymic patterns in computers, occurs in less than half of that amount in genetics (14%).

In computers, only 4 of 17 patterns yielded results greater than or equal to 5%, while in genetics, there were 7. This means that for these particular search terms in the

domains of computers and genetics, 13 and 10 patterns respectively represented hypernymic patterns which were used in less than 5% of the total.

3.3 Domain-specific Patterns

A limited number of domain-specific patterns were identified for both computers and genetics. The following domain-specific patterns were initially found in the course of examining hypernymic contexts for the preceding work. Subsequently, I ran concordances on the patterns themselves (i.e. as opposed to on a particular search term) to verify if they consistently indicated hyperonymy. Therefore, many of the contexts extracted for this section do not include the original search terms used as a control factor when identifying generic patterns. For each pattern, only the first 100 contexts[11] were viewed, and the number of contexts in which hyperonymy was shown was noted.

Like generic patterns, domain-specific patterns are not always used to indicate hyperonymy (i.e. they produce noise as well). The sections which follow detail the frequency of patterns for the first 100 contexts, examples of the hypernymic contexts containing these patterns, and examples of possible sources of noise identified for the patterns.

3.3.1 Examples

In the following examples, the pattern is in bold, the hyponym underlined, and the hypernym double-underlined. Only two examples per pattern are given.

[11] For the pattern *génération**, only 43 contexts were found.

3.3.1.1 Computers

a) format*

*Format** can indicate hypernymic relationships between concepts such as files, images, extension cards, motherboards, buses, and memory. *Format* is defined as an "*organisation formelle des données, résultant d'un formatage,*" (Rey, 1991) and denotes hyperonymy in the following contexts:

> Les slots d'extensions sont les accès du BUS avec l'extérieur.
> Le Bus ISA (IndustrieStandartArchitecture) : Le **format** le plus répandu est le bus ISA , doté d'une largeur de 8 ou 16 bits , il est capable de faire transiter les données à ...

> La wavetable restitue des fichiers de son au **format** MIDI. Des timbres de base (64 sons de percussion et 128 instruments de musique) sont stockés dans une mémoire ...

In the above examples, *bus ISA* is a type of *bus*, and *MIDI* is a type of *fichier de son*.

b) version*

*Version** is used to distinguish a wide variety of types of concepts in the domain of computers, including types of software, operating systems, and memory.

> Les langages possibles sont JavaScript (une **version** simplifiée de Java) et VBScript (une **version** simplifiée de VB).

> Le langage HTML est utilisé sur le WWW depuis 1990. La **version** actuellement en vigueur est *HTML 2.0*.

In these examples, *JavaScript* is a type of *Java*, and *HTML 2.0* is a subtype of *HTML*.

c) génération*

Like *version**, this pattern is used to denote hyperonymy for various concepts such as the processor, software and hardware.

> **Générations** de microprocesseurs 286, 386, 486, Pentium, Pentium MMX, Pentium II, et les Power PC 603, 604 ...

Acheter un disque dur de type "big foot" (y compris la nouvelle **génération** de big foot UDMA), même s'ils sont moins chers.

3.3.1.2 Genetics

It is difficult to say that there are actually patterns which are specific only to genetics, since the subdomains of biology overlap so much. Even *souche** and *lignée**, which seem as though they would be more genetics-specific than *famille** or *espèce** *de (d')*, can be used as patterns in microbiology or immunology to refer to the same types of concepts. Therefore, by "domain-specific patterns in genetics," I mean patterns which are frequently used in genetics and the fields related to it, but which would not be used in domains of a completely different nature (e.g. computers, engineering).

a) famille*

This pattern is a term which is used in biological taxonomy, and is situated fairly high in this hierarchy: Kingdom, Phylum, Class, Order, **Family**, Genus, Species. Concepts such as viruses, genes and proteins can be classified into families.

Si les gènes de la **famille** daf semblent s'être spécialisés dans la défense des cellules contre les agressions de toutes sortes, les gènes du type clock apparaissent eux comme des sabliers de l'organisme.

Or, le Pr Shiloh a constaté qu'une portion de la protéine codée par le gène ATM est très semblable aux protéines de la **famille** des phosphatidylinositol 3-kinases (PI3K) que l'on trouve chez les mammifères et chez les levures.

b) souche*

This pattern is used mainly to define subtypes of bacteria, viruses, or organisms such as mice. The name of the strain of bacteria will determine if certain genes are mutated; in the case of viruses, it can indicate how virulent the strain is. A particular strain of mouse has usually had some of characteristics of the genome altered for experimental purposes, and the name of the strain will indicate to a geneticist what those characteristics are.

> ... en 1991 par une équipe anglo-américaine ont consisté à introduire le gène de la dystrophine directement dans le muscle d'une **souche** de souris (**souche** mdx) déficiente en dystrophine.
>
> L'enzyme alpha-amylase, par exemple, produite par une **souche** de bactérie - bacillus subtilis - génétiquement modifiée peut d'ores et déjà être utilisée en France ...

c) lignée*

Lignée is used to discern the origin of tissue culture cells. The first example shows that *Caco-2* is a type of *cellule épithéliale*, while the second identifies six subtypes of lymphocytes:

> Grâce à une co-culture de lymphocytes issus de plaques de Peyer et d'une **lignée** de cellules épithéliales particulière (Caco-2), l'équipe d'Eric PRINGAULT a pu obtenir une différenciation en cellules.
>
> Ainsi, pour faire « mûrir » la **lignée** lymphocytaire[12] (qui donne : prélymphocytes, prélymphocytes B et T, lymphocytes B et T, plasmocytes), il faut une cytokine, interleukin 7 (flèche marron foncé), à laquelle les autres lignées semblent insensibles.

d) espèce* de (d')

Like *famille**, *espèce* de (d')* is used in the classification of living things. It is the lowest term in the classification hierarchy, excluding subspecies.[13] Perhaps one of

[12] In this example, the hypernym – *lymphocyte* – is expressed as the adjective, *lymphocytaire*.
[13] No concordance entries were found for *sous-espèce*.

the reasons why *espèce* de (d')* is not a very productive pattern for genetics is that many types of concepts dealt with in genetics (i.e. proteins, enzymes, genes) are not classified by their species. However, many of the organisms used in genetics research are indeed classified by their species.

> Les différentes souches de Saccharomyces cerevisiae (l'une des quelque 600 **espèces de** levures existant dans la nature) utilisées dans la fabrication du pain, mais aussi du vin et de la bière, représentent...
>
> Il s'agissait précisément d'éléments P, dont l'origine fut ensuite attribuée à une autre **espèce de** drosophile, présente uniquement en Amérique centrale : Drosophila willistoni.

3.4 Frequency of Domain-specific Patterns

3.4.1 Computers

Pattern	Frequency in first 100 contexts
format*	25
version*	18
génération*	14 (total number of contexts found by Wordsmith was 43)
Total	57

3.4.2 Genetics

Pattern	Frequency in first 100 contexts
famille*	28
souche*	17
lignée*	12
espèce* de (d')	10
Total	67

3.4.3 Analysis

For computers, a total of 57 instances of domain-specific hypernymic patterns were found for the three patterns identified. If the number of hypernymic contexts extracted as compared to the number of contexts examined is considered, *version** produced the poorest results of the three domain-specific patterns for this domain, with

18 hypernymic contexts. However, this number may not be a true reflection of this pattern's potential, for reasons which will be discussed in Section 3.5.3.1.3.1 under possible sources of noise. *Génération** was the only domain-specific pattern in this study for which Wordsmith found fewer than 100 occurrences when the concordance was run. Since it is used in 14 hypernymic contexts, it accounts for close to 25%[14] of the total contexts found with domain-specific patterns. This is the smallest percentage of the three domain-specific patterns in computers. However, considering that 14 contexts out of 43 showed hyperonymy, *génération** returned hypernymic contexts approximately 33% of the time, without any restrictions or additional control factors. This makes *génération** the most promising domain-specific pattern for computers. *Format** indicated hyperonymy in 25 out of 100 contexts, accounting for 44% of total hypernymic contexts for computers containing domain-specific patterns. This was the largest percentage of the domain-specific patterns for this domain.

For genetics, four domain-specific patterns were located, and 67 occurrences of these patterns in hypernymic contexts were extracted. *Famille** accounts for the greatest proportion (42%) of these contexts, with 28 occurrences. *Espèce* de (d')* is the least promising pattern with only 10 (15%) occurrences, while *souche** (17 occurrences for 25%) and *lignée** (12 for 18%) are more productive.

3.5 Implications for Tool Development

In this section, various factors relevant to tool development will be discussed. These include general factors concerning performance (precision and recall), and how performance can be improved through pattern restrictions. Possible sources of noise for

[14] All percentages are rounded up or down to a whole number.

the patterns individually (where applicable) and the patterns collectively will be addressed.

3.5.1 Precision vs. recall

Meyer *et al.* describe automatic knowledge extraction as a constant trade-off between precision[15] and recall[16] (1999, 259), and Davidson points out that the improvement of one generally comes at the expense of the other (1998, 79). Recall refers to the percentage of hits retrieved out of the total number of potential hits in the corpus. Therefore, calculating recall involves examining a KWIC concordance for all occurrences of the search term, locating hypernymic contexts, and then seeing what percentage of potential hits was retrieved by the tool. Precision refers solely to the accuracy of the results of the concordance, that is, the number of valid contexts as compared to the noise, or contexts which do not express the desired relation. Ideally, the terminologist wants to locate all hypernymic contexts in the corpus for the search term (i.e. 100% recall) and no noise (i.e. 100% precision). Improving recall and precision is a motivating factor behind research on patterns. In the sections which follow, I will propose ways to increase precision and discuss certain factors which inhibit recall.

3.5.2 Pattern restrictions

Sometimes, hypernymic patterns can indicate things other than the hypernymic relation. This can create noise. One way to reduce the amount of noise in a knowledge extraction tool is to apply certain restrictions to each pattern. According to Meyer,

---

[15] The number of hits divided by the hits plus the noise (Meyer *et al*, 1999, 258).
[16] The number of retrieved contexts divided by the hits plus the misses. (*ibid.*)

"Pattern restrictions should ensure that the pattern does not generate too much noise, but on the other hand, it must not exclude valid contexts" (1999, 11). Therefore, before restrictions are added, it must be considered whether or not doing so will create too many additional misses. It is preferable to have to manually filter out a few instances of noise than to miss too many valid contexts. By studying extracted contexts containing hypernymic patterns, one can often discover possible restrictions for the patterns that can be placed on the patterns to eliminate certain instances of noise. Some restrictions can be applied to patterns if the corpus has been tagged.[17] Tagging means that additional information has been added to the text for analytical purposes. The most common type of information added is the part of speech. This can be done automatically by a computerized tagger with good results. The following example provides a case where part-of-speech tagging can aid in eliminating noise:

> Des taux de transferts élevés ne sont pas nécessairement une garantie par le seul fait que vous obtenez de belles valeurs lors de vos téléchargements **tel que** semble le démontrer votre programme de surveillance des activités modems.

While we had previously shown that ***tel(le(s))(que)*** corresponds to hypernymic contexts, this data shows that when followed directly by a verb, ***tel(le(s))(que)*** can retrieve noise. However, with corpora tagged for parts of speech, it is possible for programmers to add the restriction that the tool will not retrieve examples in which ***tel(le(s))(que)*** is followed by a verb.

3.5.3 Issues affecting recall

---

[17] Only certain restrictions require tagging. For examples of restrictions that do not require tagging, see Section 3.5.3.2.1.

Both Davidson *et al* (1998) and Meyer *et al* (1999) describe certain "issues" which complicate pattern-based knowledge extraction. These problematic issues stem from the fact that both patterns and relations are complex in nature. These issues may affect only certain individual patterns, such as the polysemy of certain generic patterns and domain-specific patterns, or can affect all patterns, or certain "types" of patterns. These issues can include multi-word patterns which have been "broken up," variability of patterns, underspecified hypernyms, and anaphoric reference. The sections which follow will briefly examine pattern-specific (where applicable) and generic "issues" which may constitute possible sources of noise for the French patterns we have identified.

3.5.3.1 Pattern-specific sources of noise

3.6.3.1.1 Lexical patterns:

a) est un(e)/sont des (de)

This pattern can generate noise when the search term is not the subject of the verb. This can be seen in the following examples:

> Décrypter l'état de l'opinion publique européenne vis-à-vis des *OGM* **est un** exercice périlleux, source de vives controverses, tellement complexes sont les réactions des Européens.

There is no simple way to avoid retrieving such examples. What would be needed is a full parse of the sentence, showing that *décrypter*, rather than *OGM*, is the subject of *est un*.

b) (s')appel*

While *(s')appel** is used as a pattern for hyperonymy, it can also indicate another relation, synonymy:

> Appelé CPU (Central Processing Unit), le processeur est une unité centrale de traitement des données qui gère et contrôle les différentes opérations arithmétiques...
>
> Les ports d'entrée-sortie sont les endroits sur l'ordinateur où d'autres machiches (appelés aussi périphériques) peuvent être reliées à l'ordinateur, permettant ainsi son emploi divers.[18]

In the second example, noise could be limited by stipulating that *aussi* may not directly follow the pattern, and the second might not be retrieved at all if the search window were set at a low number of characters. However, these examples would at least be "good noise," in that they express another conceptual relation.

c) est le (la, l')/sont les (le)

One potential problem with this pattern is that it has been found in contexts which provide a metaphorical comparison between two objects. This type of usage was particularly prominent in the computers corpus. While these examples may be an effective way of describing a device in such a way that the reader can conceptualize it, they do not provide true hypernyms, and, strictly speaking, would be examples of noise if located by knowledge extraction tools:

L'unité centrale (UC) C'est le chef d'orchestre du système informatique.

d) comme

Upon examining the data, it is apparent that *comme* would produce examples of noise if programmed into a knowledge extraction tool:

---

[18] Although none of the designated search terms appear in this example, it was found upon examining the larger context for one of the search terms. It has been retained as an example of possible noise in which the pattern denotes synonymy rather than hyperonymy.

> Cette nouvelle ne réduit toutefois pas à néant les perspectives de clonage, **comme** certains l'ont affirmé fort témérairement.
>
> **Comme** dans la mitose, les paires de chromosomes dédoublés se regroupent vers le plan équatorial de la cellule.

While the first example is probably unavoidable, the second could be eliminated if a restriction is programmed that would eliminate contexts in which a preposition comes between *comme* and the search term. This would eliminate comparative examples like this one. Alternately, the search window could be set at a very low number of characters.

e) tel(le(s))(que, qu')

Tel can be used in a simile, in which case the contexts would be examples of noise, as seen in the following example:

> On sait aussi qu'il existe dans nos chromosomes d'énormes quantités de séquences répétées. Mais de séquences qui s'allongent, **tel** le nez de Pinnochio, on n'avait pas encore entendu parler.

Looking at this example, it might seem as if one possible solution to filter out these types of examples would be to greatly restrict the search window for *tel* (qu*)* so that the search term would occur directly beside the pattern. While this solution would work for many examples, it might also create additional misses. For example, contexts in which the search term is a term in an enumeration, and is separated from the pattern by other items in the enumeration would be missed (see example under *tel* (qu*)* for *disque dur*, Section 3.1.1(g)). As mentioned earlier, it is preferable to allow a few additional examples of noise (decreased precision) in order to find a greater number of hits. Therefore, greatly restricting the search window would probably not be worthwhile in this case.

f) sorte*

Only the noun *sorte** can yield hypernymic contexts, while the verb form of this pattern generates noise:

L'industrie fut en paix jusqu'à ce qu'Intel **sorte** son processeur 80386 de 32 bits.

While adding the preposition *de* after *sorte* would eliminate such examples of noise, it would also eliminate valid contexts such as the example given earlier for *séquence de bases* in which the *de* precedes the pattern (see Section 3.1.1(h)). Instead, if working with tagged corpora, restrictions should be placed on the pattern so that only occurrences of the noun *sorte* will be retrieved.

g) constitue*

*Constitue** can be found in contexts denoting another relation, meronymy. In fact, Davidson includes it as a meronymic pattern (1998, 90), found mostly in phraseology such as *constituent de, est constitué de*. However, even when used in the simple present, the verb often conveys meronymy rather than hyperonymy:

> Les gènes dits "structuraux" **constituent** près de 99 p. cent du génome.
>
> Le temps avait quand même fait son œuvre, détruisant l'ADN nucléaire, c'est-à-dire l'ADN qui **constitue** les chromosomes et sur lequel se trouve inscrit l'ensemble du patrimoine génétique de tout individu.

In the same way that Davidson proposes that *constitue** is problematic as a pattern for meronymy owing to its polysemy (1998, 91), the same is true for this pattern in terms of hyperonymy. This is to be considered before adding it as a pattern to knowledge extraction tools, especially given the low percentage of contexts in which it is used (2% overall). On the other hand, at least the "noise" will be largely good noise, in the sense that another important conceptual relation (meronymy) is detected.

h) par exemple

This pattern often denotes hyperonymy when it occurs within parentheses, another hypernymic pattern (discussed in Section 3.1.2). However, several occurrences in which it appears between commas show that it does not denote hyperonymy under these circumstances:

> C'est le cas, **par exemple**, lorsqu'un transgène ne s'exprime pas dans les parties comestibles de la plante.

In another example of potential noise found for *par exemple*, it seems to be used to denote a property:

> Ainsi le bus PCI sera **par exemple** cadencé à 37,5 MHz avec un bus système à 75 MHz, au lieu des 33 MHz "conventionnels" obtenu avec la fréquence du bus système de 66 MHz.

In fact, since *par exemple* is used so often in ways that do not denote hyperonymy, it may not be worth including as a pattern in a knowledge extraction tool, given its low overall frequency (2%), and the fact that it often occurs in conjunction with another hypernymic pattern, parentheses.

i) forme* (de, d')

This pattern is often used to indicate the shape of something rather than hyperonymy:

> La famille des rudivirus correspond à des virus **en forme de** tige rigide, sans enveloppe, constitués par une double hélice d'ADN génomique à laquelle une protéine est associée.[19]

These examples of noise can be eliminated by programming the tool to reject contexts in which the word preceding *forme* is *en* or *sous*.

j) il s'agi* de (d')

[19] Context found in concordance results for *famille* when tested as a domain-specific pattern.

When followed by a verb, this pattern does not seem to denote hyperonymy, as the following examples demonstrates:

> C'est là une percée scientifique considérable, que les biologistes croyaient irréalisable. Le principe du clonage est pourtant simple. **Il s'agit d'**inciter une cellule à se multiplier pour reconstruire un individu complet.

This could be eliminated by the restriction that the tool will not retrieve contexts in which verbs directly follow this pattern.

3.5.3.1.2 Paralinguistic patterns

a) parentheses ( )

Though parentheses often do indicate hyperonymy, they can also be used in other ways. These other ways can include:

- the full version of acronyms or abbreviations

  La copie obtenu est un ARN **(acide ribonucléique)**, autre macromolécule légèrement différente de l'ADN.

- acronyms or abbreviations

  Un organisme génétiquement modifié **(OGM)** est un organisme vivant dont on a modifié le patrimoine génétique en y insérant un ou plusieurs gènes issus d'un autre organisme vivant.

- references to other articles or parts of the article for further explanation

  ...chez lesquelles la télomérase était inactivée par suppression d'un de ses éléments essentiels, le composant ARN **(voir l'encadré "La télomérase, une enzyme complexe")**.

- words, letters or numbers which are hyperlinks[20] to other parts of the text or web pages which provide more information. In the following example, the letters (d) and

[20] The original HTML file was viewed to verify that these were hyperlinks in the document.

(e) are hyperlinks to a glossary entry where *ADN double-brin* and *ADN cellulaire* are defined.

Ensuite, l'enzyme virale nommée transcriptase inverse recopie le matériel génétique du virus, présent sous forme d'ARN, en ADN double-brin (**d**), qu'une autre enzyme virale, l'intégrase, incorpore dans l'ADN cellulaire (**e**).

- synonyms

  La mémoire morte (**ROM, Read Only Memory**) C'est une mémoire permanente. figée en usine.

- clarifications

  Ces programmes sont ensuite compilés (**traduits en une suite d'instructions élémentaires qui font partie du langage propre du processeur**).

3.5.3.1.3 Domain-specific patterns

3.5.3.1.3.1 Computers

a) format*

A possible source of noise for this pattern is that it is sometimes used in a generic sense that does not denote hyperonymy:

> L'information vidéo tirée d'un CD est envoyée à la carte vidéo (circuit électronique) qui convertit cette information en un **format** que l'écran de l'ordinateur peut accepter. L'information audio est également tirée du CD et est envoyée à la carte de ...

b) version*

The results for this pattern may have been skewed by the fact that *version* occurs frequently in certain commands and site information found in web pages. For instance, several text-only versions of PowerPoint slides for an introductory university computers

class were included in the corpus. On each of these, the command "Afficher la version graphique" was found. As well, web pages created using Claris include information such as: "This file created 01/12/97 08:48 by Claris Home Page version 2.0." These contexts would probably not affect the output of a knowledge extraction tool such as the TA, since a search term would be specified, as would the distance to the left or right of *version*. In fact, the entire premise behind such knowledge extraction tools is to eliminate contexts such as these.

c) génération*

This pattern is polysemous, in that while the sense used in many examples denotes hyperonymy, the sense used in the example below indicates something being drawn up or created, and would be an example of noise if retrieved:

> Le circuit de **génération** d'adresses convertit les adresses virtuelles reçues du compilateur en adresses utiles au circuit de support...

If the search term occurred close enough to *génération** used in this way, these types of examples could probably not be eliminated by restrictions placed on the patterns.

3.5.3.1.3.2 Genetics

a) famille*

In genetics, *famille** is also used in a very general-language way: to refer to people who are related. Since genes are passed down in families, it is obvious that genetics would often study how certain genes are passed down. Such contexts extracted with the pattern *famille** do not indicate hyperonymy, and would be examples of noise:

> Les généticiens ont étudié en détail des **familles** dans lesquelles plusieurs cas de retard mental avaient été répertoriés.

> Le premier élément nécessaire pour établir une localisation primaire est la disponibilité de l'ADN provenant de **familles de** personnes atteintes de cette maladie.

Of course, since a search term would be input into an advanced knowledge extraction tool, the first example would probably not be returned, since it does not contain a term that would yield hypernymic contexts (i.e. a terminologist would be unlikely to search for types of *généticiens*). However, if the user wanted to know types of *ADN*, the second context would probably be located, and would constitute noise.

b) souche*

While many contexts containing the pattern *souche** include both hypernym and hyponym, other examples of contexts in which *souche** is used contain only the hypernym, without listing the specific names of the hyponyms:

> Les différentes **souches** de Saccharomyces cerevisiae (l'une des quelque 600 espèces de levures existant dans la nature)...

In this example, the pattern *souche** does not indicate hyperonymy, in that it does not name any strains of *Saccharomyces cerevisiae*.

c) lignée*

Like *famille**, *lignée** has an alternate general-language meaning that may generate noise:

> ... était une petite fille de la reine Victoria, et l'étude des mitochondries d'un cousin ou cousine, nièce ou neveu de la **lignée** féminine de cette reine aurait résolu la question.

d) espèce* de (d')

The *species* is used to differentiate between two types belonging to the same *genus*. Often, the *genus* (*genre* in French) is left implicit. The following example demonstrates this, as it does not say to which genus the *âne* and the *jument* belong:

> ... appartenant à une autre espèce ; et qu'un lapin, par exemple, incorpore des gènes de lièvre. Lorsque le croisement entre espèces différentes - l'hybridation - est possible (comme entre un âne et une jument, par exemple), les descendants (des mulets)...

Examples such as this do not give an hypernym, but rather two co-hyponyms.

3.5.3.2 Generic sources of noise

3.5.3.2.1 Restrictions on certain words

While examining the data, it became apparent that when certain words occur close to the patterns, the contexts indicate a meronymic rather than an hypernymic relation. These words include *élément, composant,* and *partie*. In the same way, the word *phase* can point to temporal parts (Meyer *et al*, 1999, 262). However, these words do not stand alone; in fact, they are often preceded or followed by hypernymic patterns. This has been found to be the case for *est un ' sont des* and *tel**:

> Tous les autres ***éléments***, qu'ils soient intégrés ou non à l'unité centrale (comme le lecteur de disquettes, par exemple) appartiennent à la catégorie de périphériques.

> ... micro-ordinateur, ce qui permet à la machine de fonctionner. Ce boîtier intègre un ensemble de ***composants***, tels le processeur, la carte mère et la mémoire, chargés de traiter, traduire, enregistrer, stocker, déplacer et transmettre...

> La mitose est la ***phase*** la plus remarquable du cycle cellulaire, elle est subdivisée en quatre étapes : prophase, métaphase, anaphase, télophase.

These words would probably generate the same type of noise when used in conjunction with other hypernymic patterns, and thus knowledge extraction tools should be programmed to reject contexts in which these words occur close to the pattern when searching for hypernymic contexts.

3.5.3.2.2 Multi-word patterns that are "broken up"

Sometimes, multi-word patterns are interspersed with other elements. For hyperonymy, the most commonly affected patterns include *est le / sont les* and *est un / sont des*:

> ... besoin d'un support physique ou logique (disque dur, programme hôte, fichier...) pour se propager ; un ver **est** donc **un** virus réseau.
>
> Le microprocesseur n'**est** rien d'autre qu'**un** super circuit intégré, c'est à dire un ensemble de transistors, de diodes, de résistances et de conducteurs imprimés sur une minuscule pastille de silicium qu'on appelle une puce (chip).

These hypernymic contexts would not be retrieved by knowledge extraction tools, unless they allow for the possibility of patterns being broken up.[21]

3.5.3.2.3 Variability of patterns

Certain patterns can be expressed in varying ways. The placement of *de* relative to *type* (either before or after) is one example. This can be solved by programming both *de (du) type** and *type* de (d')*. Of course, this may decrease recall, as *type* often indicates hyperonymy when *de* + search term has been replaced by *en*. However, the alternative – inputting the pattern as *type* – would most likely decrease precision (i.e.

[21] The TA, for example, does not allow this, as, according to the programmer, it introduces another level of programming complexity.

create too much noise). Testing in knowledge extraction tools will show whether this is an acceptable solution. Other patterns, however, present other variations and require different solutions, if there is a possible solution:

Le premier périphérique d'un PC, c'est son moniteur.

Perhaps a more common way of expressing this context would be *Le moniteur est le premier périphérique d'un PC*, which contains the pattern *est le*. In this case, it is unlikely this context would be located with a knowledge-extraction tool, because the linguistic structure indicating hyperonymy is *c'est son* in this case.[22]

3.5.3.2.4 Underspecified hypernyms

According to ISO 1087 (4), the most immediate hypernym should be given in a terminological definition. Underspecified hypernyms are those which are not the most immediate. Reasons why writers sometimes use underspecified hypernyms include the fact that the most immediate one has already been given earlier in the text, or that it is too obvious (Flowerdew, 1992a, 205). A small number of examples was found in which the hypernym given may be too general to be useful from a terminological point of view. In the following examples, the pattern is in bold, the hyponym underlined, and the underspecified hypernym double-underlined:

> Le lecteur de CD-Rom **est un** appareil permettant de lire des données stockées sur des disques compacts, CD audio (ceux qui ne contiennent que du son) ou CD-Rom.
>
> La transgénèse **est une** opération de grande précision, effectuée dans des conditions de contrôle optimal, comme l'explique....

[22] Again, the TA would not be able to locate such a context.

The most immediate hypernym for the first example would probably be *périphérique d'entrée*; however, this context does provide the differentiating characteristics of a *lecteur de CD-Rom*, while the second one does not. *Opération* is probably too conceptually general to be useful. In addition, the second example does not list any differentiating characteristics and is probably useless for terminological purposes.

3.5.3.2.5 Anaphoric reference

Meyer points out that a pattern-based approach to knowledge extraction relies on finding patterns located within a specific distance to the left or right of a search term. However, in naturally-occurring texts such as those used for this analysis, the search term "is not repeated over and over again, but rather, replaced by pronouns, the generic term, term variants, etc." (1999, 16). This is called anaphoric reference. Most examples of anaphoric reference found by researchers are for relations other than hyperonymy, such as meronymy and functionality. One possible reason for this is that hyperonymy, being the most important relation, is generally expressed first, and thus occurs close to the search term. However, of course, this is not always the case, as can be seen in the following examples:

> La transgénèse par voie pollinique consiste à déposer directement le transgène sur les stigmates de la fleur **c'est un** procédé en cours d'étude...[23]
>
> Il en existe 2 **sortes** : les scanners à main avec lesquels on passe sur l'image et les scanners à plat qui s'utilisent comme une photocopieuse.

The first example gives the function first, then the hypernym. The search term is replaced with the demonstrative pronoun *ce*. Because of the large number of characters

[23] There was no punctuation between *fleur* and *c'est* in the original text.

between the search term and the pattern, the knowledge extraction tool may miss this context. In the second example, the term *scanner* is replaced by the pronoun *en*.

A much more detailed study is required to attempt to find contexts containing anaphoric reference. Perhaps if the pronouns etc. were tagged in such a way as to show the antecedents, examples of this type could be found.

3.7 Chapter Summary

This chapter has presented the results of research on hypernymic patterns in French in computers and genetics. First, a list of the patterns detected was given, along with examples. Second, the frequency of the patterns was examined, and it was determined that the most productive generic patterns are *type* (de d')*, *est un(e) / sont des (de)*, and *(s')appel**. Third, the domain-specific patterns *version**, *génération**, and *format** were identified for computers, and *famille**, *espèce* de (d')*, *souche** and *lignée** were found for genetics. Finally, we addressed the implications of our patterns for the development of knowledge extraction tools by examining a number of issues that complicate the process, and that should be considered by programmers intending to use the patterns in their toolset.

CHAPTER 4 – INTER-LINGUISTIC COMPARISON

4.0 Introduction

This chapter presents a comparison of French and English hypernymic patterns. The French contexts were found during the analyses described in the previous chapter, and the English data were extracted from concordances run on certain search terms and on the English patterns themselves. First, the English set of patterns used to extract the hypernymic contexts is presented, then, any differences noticed in the English and French linguistic patterns. General differences in linguistic tendencies that may affect knowledge extraction tools will be examined, followed by a comparison of paralinguistic patterns in the two languages.

4.1 The Field of Contrastive Linguistics

> Pourquoi la linguistique contrastive? Tout d'abord parce que la comparaison de deux langues met à jour des phénomènes qui concernent le langage en général. Elle permet à la fois de discerner ce qui a une portée commune et d'en établir les limites. Elle constitue aussi un moyen de cerner de plus près les représentations qui caractérisent les deux langues envisagées (Guillemin-Flescher, 1992, 1).

To analyze the differences between the French and English hypernymic statements I had extracted, I needed the theoretical tools to do so. This prompted a brief study of the literature on English and French contrastive linguistics.

Contrastive linguistics obviously involves comparing two languages to identify the differences. Vinay and Darbelnet cite the use of pronominal verbs in French or the preference for the passive voice in English as examples of differences that become apparent when the two languages are compared (1995, 17). For example, certain

pronominal verbs are translated by the passive in English, and the passive in English is often translated by the active in French (Vinay and Darbelnet, 1995, 138-39):

Le saumon **se mange** froid. (pronominal)	Salmon **is eaten** cold. (passive)
You **are wanted** on the phone. (passive)	On vous **demande** au téléphone. (active)

In 1958, a seminal contrastive study of French and English appeared in the form of Vinay and Darbelnet's *Stylistique comparée du français et de l'anglais*.[1] Since the publication of this work, many other authors, among them Guillemin-Flescher (1981, 1992), Ballard (1984, 1993), Chuquet and Paillard (1987), Grellet (1993), Quillard (1994), Fergusson (1995), Salkoff (1999), and Bossé-Andrieu (2000) have conducted research in various aspects of this field of study.

According to Vinay and Darbelnet, the discipline of contrastive linguistics (which they call comparative stylistics) is extremely vast "since it relies primarily on the knowledge of two linguistic structures: two lexicons, two morphologies; but also, and perhaps above all, because it relies on two particular viewpoints of life which inform these languages or which result from them..." Because of the large scope of this discipline, motivation for studying contrastive linguistics can differ. It has applications in bilingual writing (Bossé-Andrieu, 2000), translation pedagogy (Vinay and Darbelnet, 1958; Chuquet and Paillard, 1987; Guillemin-Flescher, 1981, 1992; Grellet, 1993; Fergusson 1995), and the development of machine translation systems (Salkoff, 1999). However, most of the contrastive literature on English-French deals with a comparison of these tendencies as reflected in human translation, which is undoubtedly because

[1] Sager's translation of this work (1995) is quoted in this thesis.

translators are the group most likely to be interested in this type of comparison. This is demonstrated by the fact that most bilingual stylistics manuals (Vinay and Darbelnet, 1958, and Chuquet and Paillard, 1987, for example) are meant to be used for translation pedagogy.

To the best of our knowledge, no French-English contrastive studies have focussed particularly on terminology. However, as we attempt to show below, terminologists can also apply the principles of this domain to knowledge extraction. This is because the majority of research on linguistic patterns has been conducted in English (see Section 1.4). Therefore, studying and identifying differences between patterns in the two languages, as well as differing general linguistic tendencies, can help tool developers implement the existing techniques in knowledge extraction for English in French-oriented tools.

4.2 English Set of Patterns

Since the principal focus of this thesis is French patterns, the set of English hypernymic patterns used is the one drawn up by Davidson (1998). Davidson originally discovered her patterns in TERMIUM records from the 1996 CD-ROM. A total of 21 patterns were found, and these were subsequently added to the Text Analyzer for testing and further refinements (1998, 61). Davidson's patterns were tested on my English corpora. A small number of patterns which did not denote hyperonymy in either English corpus was rejected. These include *following groups, is classified as / are classified as, descriptive of, occur*... (20 characters)... as, constitute**, and *a general term for*.

As well, concordances were run on the most obvious French translations of Davidson's patterns to see if they were used in the French corpora. In total, 13 patterns were briefly examined for English, to see if their corresponding French patterns were different, or if inputting these patterns into a knowledge extraction tool could entail differences in tool programming for the two languages. A list of the patterns occurring in the genetics or computers corpus[2] follows, along with one example of how each pattern is used. In each of the examples, the pattern is in bold, the hyponym underlined, and the hypernym in double-underlined.

4.2.1 Examples

such as	Genes related to human cancer **such as** the human metalothionen gene were introduced into Canola among a number of crops and into Poplar trees.
refer*...to	Cancer **refers to** the abnormal growth of cells, which result in tumors.
is defined as	Extended Integrated Drive Electronics (sometimes called Enhanced IDE) EIDE **is defined as** an improved version of IDE/AT Attachment, with faster data rates, 32-bit transactions, and (in some drives) DMA...
includ*	... the sale, cultivation, growth, harvest and/or consumption of genetically modified crops. Such crops, referred to as GMOs, **include**: Round-Up Ready Soybeans, Bt Corn and Round-Up Ready Cotton, among others.
type*	Another very interesting **type** of RNA is called a ribozyme, which is an RNA that has catalytic activity.
is a / is an	A catalyst **is a** molecule which increases the rate of a reaction but is not the substrate or product of that reaction.

[2] In this chapter, the primary source used is the genetics corpus. Since an intra-linguistic comparison between the two domains is no longer being carried out, the computers corpus was used in English only when there were no occurrences of a pattern in the genetics corpus. However, the French examples come from both corpora.

is the / are the	As soy **is the** most common genetically engineered food, the researchers say their findings indicate that genetically modified food could...
is any / are any	A gene **is any** given segment along the DNA that encodes instructions that allow a cell to produce a specific product...
is called / are called	Another very interesting type of RNA **is called** a ribozyme, which is an RNA that has catalytic activity.
and other	Consequently, Bt crops (**and other** GE foods) are generally considered "substantially equivalent" to their non-GE counterparts...
term used	The nonpenetrant carrier is a loose **term used** to describe an individual who carries the abnormal gene, but does not express the disease or the trait.
among	Of course, there are different effectors, or proteins, that direct transcription. Primary **among** these is the RNA polymerase holoenzyme...
(parentheses)	Specific DNA sequences called "promoters" control the extent of copying (transcription) of genes.

4.3 Differences Identified

In 4.3.1 and 4.3.2, specific differences are differentiated from general differences. By specific differences, I mean differences which apply to specific patterns, whereas general differences can apply to several patterns or refer to general linguistic differences which must be taken into account when programming knowledge extraction tools.

4.3.1 Specific differences

The types of specific differences identified between the two sets of patterns include patterns with no direct literal equivalent in the other language, patterns whose recall potential differs greatly between the two languages, and different potential for pattern variation. These patterns will be discussed below, and examples will be provided to illustrate the differences.

4.3.1.1 Patterns with no obvious literal equivalent

Two patterns in French and one in English were found which had no direct equivalent in the other language: *il s'agi* de (d')*, *dont* and *refer* ... to*.

a) il s'agi* de (d')

This expression is described as one that can be used for many uses, and which is constantly being used in written French (Quillard, 1994, 324-325).

> Le virus ne mérite nullement l'aura de mystère qui l'entoure : **il s'agit d'**un banal logiciel, écrit avec un langage de programmation courant, à cela près qu'il n'a aucune utilité pratique et ne vise qu'à nuire.[3]

> Peu à peu, les molécules contenant l'information génétique ont été précisément identifiées : **il s'agit des** acides nucléiques, - l'acide désoxyribonucléique (ADN) et l'acide ribonucléique (ARN) -, contenus par chaque cellule.

As well, because it is an impersonal construction, this pattern tends to occur with either a colon, semi-colon, dash or period separating it from the search term.[4] In this way, its use is very similar to cases of anaphoric reference, and the search window for this pattern should therefore perhaps be larger than for other patterns.

---

[3] Hyponym in italics, hypernym underlined, pattern in bold.
[4] In a concordance run on the pattern, 17 hypernymic contexts containing this pattern were examined. Fourteen of these contexts contained the punctuation described above between the hypernym and hyponym.

b) dont

This is another pattern that has no direct literal equivalent in English. Fergusson lists it as a problematic word to translate into English, given the different meanings it can have: "Un premier déblayage du terrain consiste à identifier les trois fonctions de ce terme, qui peut être pronom possessif (...), introduction du complément d'objet indirect (...), ou, enfin, pronom à valeur partitive (« plusieurs pays, dont la France ou le Royaume-Uni ») (1995, XIV). It is this last usage which can yield hypernymic contexts:

> ... intégrant au génome de la cellule hôte sous forme d'ADN, les rétrovirus, **dont** le VIH, sont en effet, au départ, des virus à ARN.

However, as this pattern is polysemous, it will also return contexts such as the following, which would be examples of noise:

> ... début de la partie la plus importante de la recherche, qui consiste à identifier précisément la mutation sur l'ADN et le gène **dont** l'expression est défectueuse...
>
> Voici les espèces **dont** on envisage de séquencer le génome.

c) refer* ... to

Davidson lists *refer**... (20 characters) ... *to* as an hypernymic pattern for English. It produced hypernymic contexts in 54 out of 100 contexts, including the following:

> Cancer **refers to** the abnormal growth of cells, which result in tumors.

However, this pattern is not without problems, as the following example demonstrates:

> These antigenic areas are found to protrude out of the surface of the virus. **Refer** also **to** the above diagram, showing the depth of the canyon...

Les faux amis ou les pièges du vocabulaire anglais (Koessler and Derocquigny, 1961, 311), written especially for translators, lists *refer to* as an item that has an "emploi plus étendu et plus usuel en anglais." Therefore, this pattern has no obvious literal equivalent. Possible French patterns expressing the same idea could include *est le, est un,* and *désign**, among others.

4.3.1.2 Patterns whose recall potential differs greatly in the two languages

The use of certain patterns is more restricted in one language than in the other. In French, *constitue** and *désign** denote hyperonymy more reliably than the English counterparts, *constitute** and *designate**, while *includ** denotes hyperonymy more often in English.[5]

Perhaps these first two patterns are used more frequently to denote hyperonymy in French because the general writing tendencies differ between French and English. It is widely accepted that English is more accepting of *verbes passe-partout* (such as *to be* or *to have*) than French (Quillard, 1994, 323). As well, according to Quillard, repetition in French is generally less frequent than it is in English. She proposes that English is more repetitive than French "puisqu'il ne dispose pas d'un système de genres lexicaux" and because "il préfère les phrases courtes et coordonnées aux phrases complexes" (1994, 318-19). Thus, perhaps the more frequent use of *désign** and *constitue** in French is a result of attempts by writers to either avoid the verb *être* altogether, or to replace it with another word that conveys a similar meaning. In *Étude de certaines différences dans l'organisation collective des textes pragmatiques*, Quillard studies this aspect of the

[5] The relative frequency of these patterns will be discussed below.

differing stylistics of French and English. She summarizes these attitudes towards this type of verb in the following passage:

> De toute évidence, les anglophones n'éprouvent pas la même répugnance que les francophones à se servir de verbes dits passe-partout, tels que : avoir, être, faire, dire, etc., sans doute parce qu'ils n'ont pas été tenus de se livrer
>
> à moultes *(sic)* reprises au savant et ô combien périlleux exercice qui consiste à tenter de les remplacer par des termes plus précis, plus vivants ou plus colorés, comme disent joliment les manuels (1994, 323).

Of course, this is not to say that French never or even rarely uses the patterns *est un/ sont des*, and *est le/sont les* to convey hyperonymy, as was demonstrated in the previous chapter.

Quillard lists several of the common methods of avoiding many of these verbs, some of which are relevant to this thesis. She states that the phrase *il s'agit de*, which has already been identified as a pattern with no direct English equivalent, is often used to avoid the copular *to be*. Quillard has also identified another instance when it is desirable in French to avoid *to be*. She points out that, in English, when the verb is followed by a noun – or, in other words, when the pattern *is a /is the* is used – that the verbs *constituer* or *représenter* are often used in French (1994, 325). *Désign** may also be used in this way.

a) constitue*

*Constitue** was identified in the last chapter as a pattern in French. Although it was not used very often for the search terms, a concordance on the pattern itself yielded 13 hypernymic contexts out of 50 contexts viewed:

> Autres périphériques : les autres périphériques **constituent** les scanners et les lecteurs amovibles (zip, jazz...)

On considère en effet qu'un envoi supérieur à 20 messages **constitue** un spam.

Ainsi, dans le cas du maïs, le gène bactérien responsable de la synthèse de la protéine détruisant la pyrale, **constitue** un tel transgène.

Chaque gène correspond à un caractère héréditaire particulier et **constitue** donc une unité d'information génétique.

English equivalent

Although Davidson listed *constitute** as an English hypernymic pattern, I could find no clear-cut examples in either of my corpora where hyperonymy was shown. Most often, it returned hypernyms which were too conceptually broad to be useful:

Vaccines arguably **constitute** the greatest achievement of modern medicine.

Might a genetic flaw **constitute** a "pre-existing condition" that would be excluded from insurance coverage?

These examples do not give true hypernyms, and are not useful from a terminological standpoint.

However, as discussed in last chapter, *constitue** is not only used to discuss hyperonymy, as it is a polysemous word. I found that it was often used to indicate the composition of something (i.e. meronymy) in both computing and genetics. In English as well, *constitute** is a pattern for meronymy. In fact, it seemed to denote meronymy more often than hyperonymy. In a concordance run on the genetics corpus, of the 17 contexts, 9 were hits for meronymy, and none were true hits for hyperonymy. For this reason, *constitute** is probably not a good hypernymic pattern to include in knowledge extraction

tools for English, while it may be more reliable in French, as 13 out of 50 contexts containing the pattern denote this relation.

b) désign*

The verb *désigner*, like *constituer*, was not used frequently for the search terms. However, it can and does function as an hypernymic pattern, and was used in this way in 18 out of 50 contexts viewed in a concordance run on the pattern:

> FTP **désigne** donc à la fois un programme et un protocole, de sorte que vous utiliserez peut-être un autre programme que FTP mais régi par les règles FTP.
>
> La transgénèse **désigne** le processus de transfert dans le patrimoine génétique d'un organisme vivant d'un gène qui lui est étranger.
>
> La règle 1, qui interdit la plupart des croisements interspécifiques, s'explique par le processus de la mitose. On **désigne** ainsi la division cellulaire qui est à l'origine de la croissance de l'individu, de la conception jusqu'à la mort : une cellule se divise pour donner deux nouvelles cellules qui se diviseront à leur tour et ainsi de suite.
>
> Il existe des morceaux d'ADN non codant jusqu'à l'intérieur des gènes. Ces séquences internes sont **désignées** sous le terme d' "introns".
>
> ... le transfert de gènes s'est déroulé correctement, la bactérie produit alors la protéine désirée, que l'on **désigne** du nom de protéine recombinante.

English equivalent

Davidson did not identify *designate** as an English hypernymic pattern in her study of TERMIUM records for the domain of composting. It does appear in rare examples (2) in the genetics corpus, including the following:

> The fundamental structural unit of chromatin is an assemblage, called the nucleosome, composed of five types of histones (**designated** H1, H2A, H2B, H3, and H4) and DNA.

It is more frequently used in French in this manner, which is perhaps owing to the less frequent use of animate verbs with inanimate objects in English.[6]

As well, in the English computers corpus, *designate** was never used for hyperonymy, but rather for functionality:

> When the PC was first developed, designers had to decide how many bytes would be **designated** for addressing particular memory locations within the system, including hard drive memory.

c) includ*

In English, Davidson listed *includ** as one of her patterns. A concordance run on this pattern on the genetics corpus revealed that of the first 100 contexts, 37 were hits for hyperonymy, while 8 denoted meronymy. In the first example, hyperonymy is shown, and in the second, meronymy:

> ... the sale, cultivation, growth, harvest and/or consumption of genetically modified crops. Such crops, referred to as GMOs, **include**: Round-Up Ready Soybeans, Bt Corn and Round-Up Ready Cotton, among others.

> Human genes vary widely in length, often extending over thousands of bases, but only about 10% of the genome is known to **include** the protein-coding sequences (exons) of genes. Interspersed within many genes are intron sequences, which have no ...

The results were even better for *including*, which generated 60 hypernymic contexts and only 2 meronymic ones per the first 100 contexts.

French equivalents

[6] Personal communication with Jacqueline Bossé-Andrieu, July, 2000.

According the Robert-Collins (1993, 384), possible French translations of this verb include *comprendre, compter, englober, embrasser,* and *inclure*, in this order. Concordances on these words revealed that, as could be expected, only *comprendre* and *inclure* yielded hypernymic contexts. In cases where the number of concordance entries exceeded 100, only the first 100 contexts were viewed.

> Les virus des hépatites peuvent être classés en deux groupes. (...)* TRANSMISSION PARENTERALE Ce deuxième groupe **comprend** les virus B, C, D à transmission parentérale (sang, sperme, éventuellement salive*) et caractérisés par le risque de...

It is interesting to note that the types of noise can be quite different in English and French. For instance, since *comprendre* can also indicate meronymy, contexts such as the following would be retrieved in French, whereas the English pattern *includ** would be unlikely to generate such noise:

> Mais des problèmes persistent : les plasmides sont souvent instables et **comprennent** des gènes de résistance aux antibiotiques.

> Comme toutes les protéines, la tubuline est produite à partir d'un gène. Mais, une fois synthétisée, elle se modifie par adjonction d'une chaîne latérale, non codée génétiquement, qui peut **comprendre** jusqu'à 34 acides aminés.

In addition, since the verb *comprendre* also means to understand, examples found in which it is used in this sense would be examples of true noise. For instance, if a terminologist were looking for different kinds of *cellules*, this example would be found.

> ... les forces mécaniques sont transmises à l'intérieur des cellules, via des trajets moléculaires spécifiques désormais, nous **comprenons** mieux comment les *cellules* détectent les stimulations mécaniques qui régulent le développement des tissus.

For *inclu** in French, only one example in each corpus showed an hypernymic relation. However, while the computers example is a valid context for the domain, the genetics example found has nothing to do with genetics and is useless for a teminologist:

> Plusieurs PC sur le marché aujourd'hui, **incluant** plusieurs blocs-notes, sont livrés avec le bus USB. De plus, de nombreux périphériques USB innovateurs ont fait leur...
>
> ... Richard Shaddick, de la GRC, indique qu'il y a des spécialistes formés par l'EMIUBC dans toutes les villes du pays, **incluant** Montréal.

However, despite this one example in both corpora, *inclu** was used principally to indicate a meronymic relation, as the following example demonstrates:

> L'homéobox commande la synthèse d'une séquence protéinique de soixante acides aminés, **incluse** dans la protéine codée par l'homéogène au complet (3).

Of the possible translations listed in the dictionary, only *compren** repeatedly yielded valid contexts for hyperonymy in genetics. In this way, the translation that looks the most like the English, *inclu**, is actually a better pattern for meronymy than for hyperonymy.

4.3.1.3 Different frequency of pattern variation

In the last chapter, *type* was identified as a pattern for hyperonymy. However, using solely the word *type* as a pattern may generate noise. Since this pattern denotes hyperonymy most often when it is preceded or followed by *de*[7], both *type* de (d')* and *de (d') type* should be programmed into knowledge extraction tools. In concordances run on

[7] In some instances, *de* was replaced by *en* (see 3.1.1(a)). If only *type* d** and *de type* were programmed, contexts in which *en* is used in place of *de* + *noun* would be missed.

the English corpora, only one example was found in which *of a / the type* was used, rather than *type of*:

> Several labs are investigating whether the gene for this protein, which is **of a type** known as a lectin, could be added to crops such as rice to make them resistant to sap-sucking insects.

Therefore, this variation in French in which the preposition precedes the noun *type* would be much more productive than its English equivalent.

4.3.2 General differences

As mentioned earlier, general differences include general linguistic phenomena which may need to be taken into account when programming knowledge extraction tools. This is in opposition to the specific differences detailed in Section 4.3.1, which deal with differences in the individual patterns. The general differences examined in this section include word order, the pronominal voice in French, emphasis, and agreement.

4.3.2.1 Word order

Chassigneux remarks "J'ai toujours été frappé par les différences d'ordre des mots entre l'anglais et le français" (1991, 71). The phenomenon of differences in word order was noticed frequently while examining hypernymic contexts. English has a marked preference for what Chuquet and Paillard (1987) refer to as *l'ordre canonique*. This term describes the standard sentence structure of *Subject + Verb + Object*[8]. Guillemin-Flescher refers to this as the *schéma canonique* and points out that not all elements need be present in a sentence (1981, 415). Salkoff refers to this word order as a sequence

[8] In sentences containing state-of-being verbs, it is not an object that follows the verb, but rather a subjective completion.

called an assertion string[9] (1999, 12), and points out that, like in English, this is the "major sentence structure of French" (1999, 25). However, according to Demanuelli, French "privilègie les agencements plus souples, sinon plus lâches, et la mise en relief par la biais de l'ordre des mots, autrement dit de l'inversion totale ou partielle des imbrications" (1995, 130). He gives the following example (taken from Guillemin-Flescher, 1981, 139) to show how commas are omitted in the English translation because of more rigid syntax rules, or, as he states, "le refus de l'anglais de bouleverser l'ordre canonique et de séparer le verbe de son complément" (1995, 128):

> J'ai planté pour elle, dans le jardin, sous ta chambre, un **prunier d'avoines**, et je ne veux pas qu'on y touche, si ce n'est pour lui faire plus tard des compotes, que je garderai dans l'armoire, à son intention, quand elle viendra. (Mme Bovary)
>
> I have planted a **wild plum tree** for her in the garden underneath your window and I won't have anyone touch it except perhaps later on to make jam which I shall keep in the store-cupboard for her when she pays me a visit.

This difference in syntax order is examined by Salkoff to discern the implications for machine translation systems: "Furthermore, other types of French inverted center strings contain certain inversions of the subject, verb and object that cannot be inverted in the same way in English translations..." (1999, 125) He proposes that Machine Translation systems should always translate the French structure *Object + Verb + Subject* by *Subject + Verb + Object* in English "to simplify the translation process" (1999, 27). He gives the example that the English translations of the following French structures are improved by re-establishing the habitual *Subject + Verb + Object* order:

---

[9] Salkoff defines a string as "a sequence of grammatical categories which constitute a syntactic structure of the grammar." (1999, 12)

> celui qu'avaient proposé Watson & Crick → ??the one which had proposed Watson & Crick → the one which Watson & Crick had proposed
>
> Frappant aussi est son utilisation de couleurs → ??Striking also is his use of the colors → His use of color is also striking[10] (1999, 27-28).

4.3.2.1.1 Implications for knowledge extraction

a) Hyperonymy/Hyponymy distinction

Many advanced knowledge extraction tools, including the Text Analyzer, differentiate between hyperonymy and hyponymy. For a given search term, users must specify whether they wish to retrieve hyponyms or the hypernym. In English, if the user wishes to locate the hypernym of the search term, the tool looks for contexts in which patterns occur to the right of the search term, as in the following example:

> RNA **is a** polymer that contains ribose rather than deoxiribose sugars.

This is done to avoid locating hyponyms. In French, however, because of the inversion of the Subject and Verb, contexts such as the following would be missed, since the program would identify these as hyponymic, rather than hypernymic contexts:

> Or, il s'avère que leur outil de choix, l'ADN des organites cellulaires que **sont les** mitochondries, pourrait muter beaucoup plus vite qu'on ne le pensait.

Thus, if the user wished to locate contexts showing the hypernym of *mitochondries*, this context would not be located, since the program would flag *mitochondries* as the hypernym of *organites cellulaires*, rather than the other way around. According to Fergusson, this type of inversion is quite frequent in French (1995, XV). Therefore, it is

[10] In the second example, Salkoff did not include the improved translation; he merely suggested that the translation would be improved if the Subject + Verb + Object order were re-established. I provided the translation.

necessary to find a means of locating these contexts. A possible solution may be to specify that if *que* precedes certain pattern verbs then the search term will be found a certain number of characters to the right of the pattern for hypernymic contexts, and to the left for hyponymic contexts.

b) Altered patterns

In French, disturbing the order can often slightly alter a pattern. This must be taken into account when programming knowledge extraction tools. In certain instances the order of the Subject and Verb was inverted:

> Le gène de prédisposition à cette maladie héréditairement multifactorielle qu'**est la** SEP n'a pas seulement profité du commerce des trafiquants d'esclaves vers le sud.
>
> En tant que tels, ils sont aussi présents chez cet eurcaryote unicellulaire qu'**est la** levure. Celle-ci est alors un merveilleux outil pour approcher la fonction de ces gènes chez l'homme.
>
> Dans le jargon informatique, ce terme, s'opposant à celui de "matériel", fait référence à la construction d'une machine abstraite, composée d'instructions logiquement reliées et rédigées dans un langage compréhensible par la machine concrète qu'**est l'**ordinateur.

In these examples, the Subjects – *SEP*, *levure*, and *ordinateur* – would normally precede the Verb in a construction such as:

La SEP **est une/la** maladie héréditairement multifactorielle...

However, as this hypernymic statement is embedded (*enchassement*) within a larger statement (i.e. there are two clauses), the Subject and Verb are inverted in the embedded clause and the relative *que* is introduced between the Subject and Verb.

Also, because the pattern *est le* is preceded by *que*, the final vowel of *que* is dropped (*élision*) and *qu'est* is formed so that two vowels will not be side by side. This, in turn, alters the pattern. Therefore, the alternate form of the pattern - *Y qu'est le X* where *Y* is the hypernym and *X* is the hyponym - should be added to retrieve additional contexts.

4.3.2.2 The pronominal voice in French

According to Sager's translation of Vinay and Darbelnet, in addition to the active, middle[11] and passive voices which both languages share, French grammar also includes a pronominal voice "because in French the reflexive pronoun has a wider range of functions than it has in English" (1995, 136). Pronominal forms are broken down into two main categories with a small number of subcategories:

- reflexive pronominal forms
  - true reflexive pronominal forms (*e.g.* Il s'est tué)
  - reciprocal pronominal forms (*e.g.* Elles se téléphonent tous les matins.)
- non (or false) reflexive pronominal forms
  - inherent pronominal verbs (no transitive counterparts expressing the same meaning - *e.g.* Il se gargarisa à l'eau et au sel.)
  - neutral pronominal forms (no object for the action of the verb - *e.g.* La tour se détachait sur le fond de verdure.)

[11] According to Sager, "The middle voice occurs in sentences where the subject is at the same time the object of the action indicated by the verb" e.g. The capillaries are contracted under the influence of the cold (1995, 136).

- middle pronominal verbs (indicate normal occurrence of events - *e.g.* Le saumon se mange froid.) (1995, 137-8).

Middle pronominal verb forms of patterns have proven to express hyperonymy, as can be seen in the following examples:

> La période pendant laquelle la cellule est au repos **s'appelle** interphase.
>
> Le terme inondation (spam, en anglais) **se définit par** la diffusion massive de plusieurs copies d'un même message dans le but de joindre des personnes qui ne choisiraient pas normalement d'en recevoir.

To express this idea, English often uses a passive construction (Grellet, 1993, 71), as in the patterns *is called* and *is defined as*.

> Another very interesting type of RNA **is called** a ribozyme, which is an RNA that has catalytic activity.
>
> Extended Integrated Drive Electronics (sometimes called Enhanced IDE) EIDE **is defined as** an improved version of IDE/AT Attachment, with faster data rates, 32-bit transactions, and (in some drives) DMA...

Of course, French can also use a passive structure, as exemplified in the following context:

> Ce type de clonage, réalisé à partir de cellules différenciées, **est appelé** clonage somatique, par opposition au clonage embryonnaire, réalisé à partir de cellules indifférenciées issues d'embryons.

4.3.2.2.1 Implications for knowledge extraction

a) Altered patterns

The fact that there is a pronominal voice in French means that certain patterns which contain verbs have more variations in French than in English. While in both languages a knowledge extraction tool must be able to look for plural and singular forms, past and present tenses, in French, the pronominal form of the verbs must also be included as a variation of the pattern if the verb begins with a vowel. In this study, for example, the verb *appeler* was identified as a pattern. Therefore, both *appel** and *s'appel** would have to be included as patterns in an attempt to locate as many contexts as possible.

4.3.2.3 Emphasis

Vinay and Darbelnet point out that "English leaves many emphatic stresses implied, and expects readers to re-establish the emphasis through their understanding..." (1995, 220). While there are various means of emphasizing certain elements of a sentence in English, such as underlining, italics, and capitals, French sometimes makes use of syntactic repetition with what Vinay and Darbelnet refer to as "pre- and postpositioned pronouns." Cadiot also examines this feature, stating that "Pour mémoire, je rappelle qu'en français, on dispose de deux marqueurs qui ont bien l'air d'être les meilleurs agents possibles de toute mise en relief : la forme du clivage (*c'est*, très lié au pronom "neutre" "*ce*" / "*ça*" / "*c'*") et le verbe "*avoir*"" (1991, 22). Examples have been found in which the pronoun *ce* is used:

Le premier périphérique d'un PC, **c'est** le moniteur.

> Le microprocesseur **c'est le** « moteur » du micro-ordinateur : il traite et fait circuler les données.[12]

4.3.2.3.1 Implications for knowledge extraction

a) Altered pattern

The above examples contain a variation of the pattern *est le*. However, because of the postpositioned pronoun *ce* + the verb *être* which becomes *c'est*, examples such as this might not be found by knowledge extraction tools. Depending on how the tool is programmed, it may read *c'est* as one word, and therefore not recognize the pattern *est un*. One solution that would ensure that contexts in which this syntactic repetition occurs are located is to program *c'est un* and *c'est le* as additional variations of these patterns.

4.3.2.4 Agreement

In both English and French, the subject and the verb must agree in number. Therefore, for English, the Text Analyzer is programmed to locate both plural and singular inflections of verbs which are part of patterns. In French, however, articles and adjectives must also agree in number and gender with the words they modify.

4.3.2.4.1 Implications for knowledge extraction

a) More variations of patterns

In the same way that the Text Analyzer looks for the variation *is an* for the pattern *is a* to account for the fact that *an* rather than *a* precedes a word beginning with vowel in English, knowledge extraction tools must be programmed to take into account all possible variations of articles and verbs. This means that there are more variations of

---

[12] Because this example is a metaphor comparing the *microprocesseur* to a *moteur*, it is not a true hypernymic context. However, it has been included to demonstrate that this structure does occur in French.

patterns to input for French.[13] The following French patterns are affected by agreement in gender and number: *tel (que), est le,* and *est un.*

tel (que) → *tel (que), tel (qu'), tels (que), telle (que), telle (qu'), telles (que)*

est le → *est le, est la, est l', sont les, sont le, sont l'*

est un → *est un, est une, sont des, sont de*

Sont le and *sont l'* may not seem as if they would need to be included, but when the hypernymic context only includes a list of hyponyms and no differentiating characteristics, these variations are often used:

> Les principaux périphériques d'entrée **sont le** *clavier, la souris* et *le lecteur de CD-ROM.*
>
> Parmi ces périphériques les plus courants **sont : *l'imprimante...***[14]

Sont de needs to be added to account for cases in which the adjective precedes the noun:

> Les protéines, qui imposent la structure des cellules et plusieurs de leurs fonctions, **sont de** longues chaînes d'acides aminés.

Of course, when the past participles of verbs like *appeler* and *nommer* are used to denote hyperonymy, they must also agree in gender. However, because verbs are input with a wildcard so as to account for all tenses, such contexts would still be found and no additional variations of the pattern need be input.

---

[13] While verb patterns are programmed with a wildcard search, the TA's programmer input the separate variations for *est le* and *est un* (Oct. 20, 1999, personal communication with J. Kavanaugh). This is probably because the knowledge extraction tool will pick up all combinations in which two words begin with these letters, including cases in which adjectives beginning with *l* follow the verb *être.*

[14] Although this example has a colon within the pattern that would interfere with the ability of a computer to locate this context, one can imagine other similar hypernymic contexts in which there would be no colon.

4.4 Paralinguistic Patterns

Davidson identified parentheses as a paralinguistic pattern for English, and I found that they also functioned this way in French. Paralinguistic patterns were briefly examined in the last chapter. Commas, dashes, colons, parentheses and phrases following questions containing the search term were all identified as such. Only parentheses were examined in detail as an hypernymic pattern in the last Section, because the other paralinguistic patterns had been too difficult to incorporate into the TA in a way that would produce satisfactory results (i.e. not too much noise).[15] Since this chapter is dedicated to a comparative analysis of French and English patterns, I decided to verify if it were not perhaps more worthwhile to include these paralinguistic patterns in French.

Vinay and Darbelnet provide a brief section on "graphic presentation as part of the large field of comparative stylistics" (1995, 243). They focus on the different uses of certain punctuation marks and on how paragraph structure differs, while Chuquet and Paillard (1987) deal with the differences in the uses and frequency of certain punctuation between the two languages. This section will discuss if and how punctuation expresses hyperonymy in the two languages, and if certain patterns might be worth including in one language and not in the other. A brief analysis was conducted to compare the frequency of hypernymic patterns expressed by various paralinguistic features. In addition, any implications of these findings for tool development were noted.

[15] Personal communication with J. Kavanaugh, chief programmer for the Text Analyzer.

4.4.1 Commas

The comma appears to be used more frequently in French than it is in English (Demanuelli, 1995, 127).[16] According to Chuquet and Paillard, "On constate une plus grande densité des virgules en français, qui est liée à la tendance du français à juxtaposer et à pratiquer l'antéposition et l'incise, alors que l'anglais a plutôt recours à l'intégration et à la coordination" (1987, 419). While this may seem to have no bearing on hypernymic patterns, the reader will remember that appositions set off by commas were listed as an hypernymic pattern for French. While this pattern also occurs in English, perhaps the increased frequency of use in French warrants further study to see if it is worth including as a pattern in French. Indeed, a concordance run on five search terms with a comma following directly before or after (example: *RNA,* and , *RNA*) revealed that in French hyperonymy is often denoted solely by this paralinguistic means more often than in English:

English Term	**Frequency**	**French Term**	**Frequency**
RNA	1/20	*ARN*	1/14
protein	2/40	*protéine*	8/48
sequence	0/13	*séquence*	1/8
mitosis	0/3	*mitose*	1/6
translation	0/7	*traduction*	0/1
Total:	3/83	**Total:**	11/77

[16] In a comparative study, Demanuelli found that the French source texts contained 140 commas, while the English translations contained only 105.

Commas were used more frequently with these particular search terms in English than in French.[17] In terms of the percentage of contexts viewed which expressed hyperonymy, in English not even 4% of the contexts denoted hyperonymy, while in French this number was 14%.

4.4.2 Dashes

> Très utilisé en anglais, parfois jusqu'à l'excès dans la langue écrite familière (lettres, par exemple), le tiret correspond à différents signes de ponctuation en français selon le cas : des virgules ou des parenthèses lorsqu'il s'agit d'une incise, une virgule ou un point-virgule lorsqu'il s'agit d'un élément postposé... (Chuquet et Paillard, 1987, 420).

Tests were carried out on the same five search terms first with a dash as a "context word." The search window specified was 12 words[18] to the left or right of the search term. The findings confirm Chuquet and Paillard's assertion that the dash is more widely used in English than in French.

English Term	**Frequency**	**French Term**	**Frequency**
RNA	1/26	*ARN*	1/3
protein	4/34	*protéine*	2/25
sequence	2/11	*séquence*	2/6
mitosis	1/1	*mitose*	0
translation	0/9	*traduction*	0
Total:	8/81	**Total:**	5/34

[17] However, the comma is used 4724 times in the French corpus, and 1658 times in the English one.

[18] Wordsmith's search window is called a "horizon" and is set in terms of words rather than characters.

Based on these limited results, it seems as though this pattern would generate too much noise to warrant including it as a pattern in either language, with 10% precision in English and 15% in French. Additional concordances were run to see if precision could increased by reducing size of the search window. The results were the same for both languages: only 3 contexts were retrieved, and only 1 of these was a hit for hyperonymy in each language. This way, too few hits are retrieved.

4.4.3 Colons

Chuquet and Paillard (1987) state that colons are used much more frequently in French than in English. In translations of French texts into English, they noted several equivalents which often replace colons, including semi-colons, periods, and dashes, as well as prepositional or verbal phrases. In addition, Grellet points out that colons are used in French to begin an enumeration of examples (1993, 24), which can often list co-hyponyms. The uses of the colon in English are presented by Chuquet and Paillard (taken from the *Advanced Learner's Dictionary*). The two uses are :

1. "After a main clause when the following statement illustrates or explains the content of that clause."
2. "Used before a long list, and often introduced by phrases such as: *such as, for example* [...] *as follows*" (1987, 420).

The second usage would be likely to yield hypernymic contexts.

In the following chart, the search window was one word to the right of the search term in English, and two words in French.

English Term	Frequency	French Term	Frequency
RNA	0/2	*ARN*	0/3
protein	0/1	*protéine*	3/7
sequence	0/1	*séquence*	0/2
mitosis	0	*mitose*	1/4
translation	0	*traduction*	1/2
Total:	0/4	**Total:**	5/18

Based on these results, it is not worth including the colon as a paralinguistic pattern for English. However, perhaps including a colon as a pattern with a very limited search window may be useful in French, since this punctuation mark is more frequent, and has a precision of 28% for these search terms.

4.4.4 Parentheses

Parentheses were identified in both English and French as hypernymic patterns. No testing concordances were run on this pattern in either language, since it was included in the statistical analysis in the last chapter for French, and since Davidson had already examined it from this point of view in English. More interesting for this section is how the information contained in the parentheses *differs* in the two languages. As was discussed in the last chapter, the information contained in the parentheses is not always an hypernym or hyponym(s). It can include acronyms or abbreviations, acronyms written out in full, synonyms, clarifications, and references to other parts of the text. In French,

however, parentheses would produce an additional type of noise. Especially in the computers corpus, the English equivalent of the French term was often given, as in:

> Le *Microprocesseur* **(CPU)** est le circuit principal de la carte mère.

Sometimes, the parentheses included the English term followed by *en anglais*, as is the case in the following example:

> Le cédérom (CD ROM qui signifie **en anglais** Compact Disk Read Only Memory) est une mémoire de masse (ou mémoire de stockage) qui ne permet que la lecture de données (on ne peut pas enregistrer de données sur un cédérom.)

In this case, tools for French could be programmed to reject these contexts for unilingual terminology work in French. However, this would not avoid retrieval of cases where only the simple English equivalent is contained in the parentheses.[19] The implications of this for knowledge extraction include a higher rate of noise for this pattern in French (since the reverse is not likely to occur in English[20]), but the additional retrieval of bilingual terminology for the terminologist.

> Les rétrotransposons sont divisés en deux sous-classes d'éléments, selon qu'ils possèdent ou non une longue séquence terminale répétée de nucléotides **(LTR : long terminal repeat).**

4.5 Chapter Summary

This chapter has presented the findings of a brief comparison of French and English hypernymic patterns and certain possible issues that tool developers may have to take into account when applying to French the methods developed for English. The most significant findings were the following:

---

[19] Unless, of course, the system included some kind of language recognition module, which is not at all typical for knowledge extraction tools, and is not the case for the Text Analyzer.

[20] In both the French and English genetics corpus, the scientific Latin term was sometimes contained in the parentheses as well.

Pattern-specific differences:

- certain patterns have no direct equivalent in the other language (i.e. *il s'agi* de (d'), dont, refer*... to*)
- the recall potential of certain patterns and their direct equivalent in the other language (*constitue*, désign*, includ**) differs greatly

General linguistic differences:

- more flexible syntax rules in French (i.e. altering the *ordre canonique* of Subject + Verb + Object is more frequent) require additional programming
- pronominal verb forms of patterns require additional programming
- increased use of pre-positioned pronouns (namely *ce*) in French may require changes in programming for certain patterns beginning with vowels (*est un(e), est le (la, l')*)
- French rules of agreement require additional programming
- a limited study of paralinguistic patterns in French and English suggested that colons may be worth including for hyperonymy in French

CHAPTER 5 – CONCLUSIONS AND FURTHER RESEARCH

The general objective of this thesis was to add to the existing body of research on patterns for knowledge extraction. This general objective was accomplished through four specific objectives. First, a list of hypernymic patterns was compiled in French (3.1). Since these patterns were discovered in two different domains, the second objective of carrying out an intra-linguistic comparison of these patterns was accomplished. A modest amount of statistical analysis was conducted to get an impression of which patterns are the most important for tool development in the two domains (3.2), and several domain-specific patterns, including *version**, *génération** and *format** for computers, and *famille**, *espèce** *de (d')*, *souche** and *lignée** for genetics, were identified and examined for frequency in both domains (3.3 and 3.4). Any foreseeable implications for tools development were suggested. The third objective of carrying out an inter-linguistic comparison of French and English patterns was accomplished in Chapter 4. In 4.3.1.1, patterns which were unique to one of the languages were pointed out (*ils 'agi** *de (d')*, *dont*, and *refer** *to*), as were patterns whose recall potential varied greatly between the two (*constitue**, *désign**, and *includ**) (4.3.1.2). The rest of Chapter 4 dealt with various linguistic features of French and English, such as word order (4.3.2.1), the pronominal voice in French (4.3.2.2), emphasis (4.3.2.3), agreement (4.3.2.4), and the differences between paralinguistic patterns in the two languages (4.4). Implications for tool development were proposed where applicable.

Developing knowledge extraction tools has two basic components: linguistic and programming. This thesis focussed on the linguistic aspect, given my linguistic training in translation and terminology. The results of this research are now ready to be handed

over to programmers, who can apply these findings to the development of knowledge extraction tools. In this way, I hope to have achieved my principal goal of making a modest contribution to the research on patterns for knowledge extraction. What should follow is the testing of linguistic patterns in actual knowledge extraction tools, with particular attention to the elements pointed out in the discussions on implications for tool development. There may be further refinements that can be made to patterns that will be discovered when the patterns are tested.

Further Research

As stated at the beginning of this thesis, empirical, terminology-oriented research on linguistic patterns has been relatively limited thus far. This work has revealed many other possible areas for further research that will help refine knowledge extraction tools and techniques:

1. Discovering additional linguistic patterns for hyperonymy in different text-types and different domains. Because hyperonymy can be expressed using so many linguistic structures, there are undoubtedly other generic and domain-specific patterns that can be added to this list. This research will equip knowledge extraction tools to better handle a variety of domains.
2. Discovering linguistic patterns for relations other than hyperonymy. I concentrated on this relation alone in this thesis, but knowledge extraction tools should be capable of locating contexts for other major relations such as meronymy, functionality, causality, and synonymy.

3. Looking at the differences in English and French patterns for other relations, and discovering if differing linguistic tendencies may affect recall for these relations as well.
4. Researching paralinguistic and grammatical patterns with a view to testing them in knowledge extraction tools.
5. Conducting patterns research in languages other than English or French.

BIBLIOGRAPHY

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.

BALLARD, M. (1984). (Ed.) *La traduction : de la théorie à la didactique.* Lille: Presses universitaires de Lille.

BODSON, C. (1999). "Acquisition d'informations sur les unités terminologiques à partir de corpus spécialisés : typologisation des patrons définitoires." L'examen de synthèse (Doctorat). Département de linguistique et de traduction, Université de Montréal.

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.

BOSSÉ-ANDRIEU, J. (sous presse). "Au-delà des genres : décalages stylistiques entre l'anglais et le français." *TECHNOSTYLE*. May 2000.

BOWDEN, P., *et al.* "Extracting Conceptual Knowledge from Text Using Explicit Relation Markers." *Advances in Knowledge Acquisition: 9th European Knowledge Acquisition Workshop, EKAW 96.* Nottingham: Springer.

CADIOT, P. (1991). "La mise en relief, un bilan linguistique à propos de la traduction anglaise des premières pages de *Mort à Crédit* de L.-F. Céline." *Palimpsestes* no. 5 (La mise en relief) Vol. 1, 19-36.

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.

CHASSIGNEUX, A. (1991). "Avant la charrue, les bœufs. La mise en relief du sujet et/ou du prédicat dans la traduction de textes économiques." *Palimpsestes* no. 5 (La mise en relief), Vol. 1, 71-75.

CHAURAND, J. and F. MAZIÈRE. (1988). "La définition." Eds. Jacques Chaurand and Francine Mazière. *Actes du Colloque La Définition*, organisé par le Centre d'études du lexique de l'Université Paris-Nord, Paris, 18-19 novembre 1988. Larousse: Paris.

CHUQUET, H. and M. PAILLARD. (1987). *Approche linguistique des problèmes de la traduction anglais ↔ français.* Paris: Ophrys.

COHEN, L. (1954). *Statistical Methods for Social Scientists.* New York: Prentice-Hall.

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

CONDAMINES, A. and J. REBEYROLLE. (1998). "CTKB: A Corpus-based approach to a Terminological Knowledge Base. Computerm '98: First Workshop on Computational Terminology, Montreal, COLING-ACL '98. 29-35.

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

DARIAN, S. (1982). "The role of definitions in scientific and technical writing: forms, functions, and properties." *Pragmatics and LSP*, Proceedings of the 3rd European Symposium on LSP, Copenhagen, August 1981. Eds. Jørgen Høedt, Lita Lundquist, Heribert Picht, Jacques Quistgaard. 1982. Copenhagen: Erhvervsøkonomisk Fortay.

DAVIDSON, L. (1998). "Knowledge Extraction Technology for Terminology." Master's Thesis. School of Translation and Interpretation, University of Ottawa.

DAVIDSON, L., *et al.* (1998). "Semi-automatic Extraction of Knowledge-Rich Contexts from Corpora." *Computerm '98: First Workshop on Computational Terminology*, Proceedings of the workshop, COLING-ACL '98, Montreal, Canada, August 1998. eds. D. Bourigault, C. Jacquemin, and M.-C. L'Homme. Montreal: Université de Montréal. 50-56.

DEMANUELLI, C. (1995). "La virgule en question." *Relations discursives et traduction.* Lille: Presses universitaires de Lille. 121-140.

DUBUC, R., and A. LAURISTON. (1997). "Terms and Contexts." *Handbook of Terminology Management.* Eds. Sue Ellen Wright and Gerhard Budin. Amsterdam/Philadelphia: John Benjamins Publishing Company. 80-87.

ECK, K. (1993). "Bringing Aristotle into the 20th Century." Master's Thesis. School of Translation and Interpretation. University of Ottawa.

EVENS, M. (1988). "The Dictionary and the Thesaurus can be combined." *Relational Models of the Lexicon: Representing Knowledge in Semantic Networks*. Cambridge: Cambridge University Press. 75-96.

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

FERGUSSON, B. (1995). *Thème anglais, Filière LEA.* Paris: Presses universitaires de France.

FLOWERDEW, J. (1992a) "Definitions in Science Lectures." *Applied Linguistics*, Vol. 13, No. 2.

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

FULFORD, H., and K. AHMAD. (1992). *A Collage of Semantic Relations,* MULTILEX Project EC ESPRIT II No. 5304, Guildford: University of Surrey.

GRELLET, F. (1993). *Initiation au thème anglais : The Mirrored Image.* Paris: Hachette Livre.

GUILLEMIN-FLESCHER, J. (1981). *Syntaxe comparée du français et de l'anglais. Problèmes de traduction.* Paris: Orphrys.

GUILLEMIN-FLESCHER, J. (1992). (Ed.) *Linguistique contrastive et traduction.* Paris: Ophrys.

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

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.

KAVANAUGH, J. *et al.* (1998). "The Text Analyzer" LAKE Lab, Dept. of Computer Science, University of Ottawa.

KAVANAUGH, J. *et al.* (1999). *The Text Analyzer.* http://www.site.uottawa.ca/~kavanaugh/LakeLab/TA/index.html June 1999 (last updated June 24, 1999).

KOCOUREK, R. (1991). *La langue française de la technique et de la science. Vers une linguistique de la langue savante.* Wiesbaden: Brandstetter Verlag.

KOESSLER, M., and J. DEROCQUIGNY. (1961). *Les faux amis ou Les pièges du vocabulaire anglais (Conseils aux traducteurs)*. Paris: Librairie Vuibert.

LOFFLER-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*. Cambridge: Cambridge University Press.

MARKOWITZ, J., *et al.* (1986). "Semantically Significant Patterns in Dictionary Definitions." *Proceedings of the 24th Annual Meeting of the Association for Computational Linguistics*. 112-119.

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 Valiquette. Ottawa: Public Works and Government Services Canada. 6-10.

MEYER, I. (in press). "Extracting Knowledge-rich Contexts for Terminography: A Conceptual and Methodological Framework." *Recent Advances in Computational Terminology*. Eds. D. Bourigault, C. Jacquemin, M.-C. L'Homme.

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

MEYER, I., and K. MACKINTOSH. (1996). "The Corpus From a Terminographer's Viewpoint." *International Journal of Corpus Linguistics*, Vol. 1, No. 2.

MEYER, I., K. ECK, and D. SKUCE. (1997). "Systematic Representation of Concepts in a Knowledge-based System." *Handbook of Terminology Management*. Vol. 1. Eds. S.E. Wright and G. Budin. Amsterdam/Philadelphia: John Benjamins Publishing Company. 98-118.

MEYER, I., D. SKUCE, and J. KAVANAUGH. (1997). "Bases de connaissances et bases

textuelles sur le Web : Le système IKARUS." *Actes des V^es journées scientifiques AUPELF/UREF* (Tunis, Sept. 1997).

MEYER, I., K. MACKINTOSH, C. BARRIÈRE, and T. MORGAN. (1999). "Conceptual Sampling for Terminological Corpus Analysis." Proceedings of the Fifth International Congress on Terminology and Knowledge Engineering (TKE99), Innsbruck, Austria, August 1999.

OOI, V. (1998). "Corpus evidence and lexicon-based language modelling." *Computer Corpus Lexicography.*

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.

PEARSON, J. (1998). *Terms in Context.* Amsterdam/Philadelphia: John Benjamins Publishing Company.

PERELMAN, C., and L. OLBRECHTS-TYTECA. (1969). *The New Rhetoric: A Treatise on Argumentation.* Trans. Wilkinson, J. and P. Weaver. Notre Dame: University of Notre Dame Press.

PICHT, H., and J. DRASKAU. (1985). *Terminology: An Introduction.* Guilford : University of Surrey.

QUILLARD, G. (1994). "Étude de certaines différences dans l'organisation collective des textes pragmatiques anglais et français." *Babel*, 43:4. 313-330.

QUILLARD, G. and G. AKHRAS. (1996). "And/et. Analyse distributionnelle de la conjonction copulative en français et en anglais." *META.* 41 (3). 459-470.

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

RONDEAU, G. (1984). *Introduction à la terminologie.* (2^e ed.). Boucherville : Gaëtan Morin, éditeur.

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.

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

SAGER, J. (1994a). "Terminology: Custodian of knowledge and means of knowledge transfer." *Terminology 1:1.* Eds. H.B. Sonneveld and K. Loening. Amsterdam: John Benjamins Publishing Company. 7-15.

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

SAGER, J. (1994c). "What's wrong with terminology work and terminology science?" *Terminology 1:2.* Amsterdam: John Benjamins Publishing Company. 375-381.

SALKOFF, M. (1999). *A French-English Grammar: A Contrastive Grammar on Translational Principles.* Amsterdam/Philadelphia: John Benjamins Publishing Company.

SELINKER, L., T. TRIMBLE, and L. TRIMBLE. (1976). "On Reading English for Science and Technology: Presuppositional Rhetorical Information in the Discourse." *Teaching English for Science and Technology.* ed. J.C. Richards. Singapore: Singapore University Press. 37-67.

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

STRUNK, W. and E.B. WHITE. (1979). *The Elements of Style.* New York: MacMillan.

VINAY, J.-P. and J. DARBELNET. (1958). *Stylistique comparée du français et de l'anglais.* Paris: Didier.

VINAY, J.-P. and J. DARBELNET. (1995). *Comparative Stylistics of French and English.* Trans. J.C. Sager and M.-J. Hamel, Amsterdam/Paris: John Benjamins Publishing Company.

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

ISO. (1990). *ISO 1087 Terminology – Vocabulary.*

Collins Robert Unabridged. (1993). Eds. B. Atkins *et al.* Scarborough/Paris:

HarperCollins*Publishers*/Dictionnaires Le Robert.