THE CAUSE RELATION IN BIOPHARMACEUTICAL CORPORA:

ENGLISH AND FRENCH PATTERNS FOR KNOWLEDGE EXTRACTION

by

Elizabeth Marshman

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 Faculty of Graduate and Postdoctoral Studies
of the University of Ottawa
in partial fulfillment of the requirements
for the degree of M.A. (Translation)

Table of Contents

List of Tables

List of Figures

Acknowledgements

I would like to say thank you to:

- the University of Ottawa, for its financial support through the Strategic Areas of Development and other scholarships
- Padmini Kalyanam, (recently of SCS) for convincing me I don't hate French.
- the faculty and staff of the STI, for all the help and support.
- Christine Hug and Silvia Pavel of the DTN for their help, patience and understanding.
- Josée Lacroix, Test Francophone.

Thanks are also due to two people who have helped enormously with this project: Caroline Barrière of SITE and Lynne Bowker of the STI. Their generosity with time and knowledge made it possible.

Thanks to my friends at the STI, for your great company, commiseration, and laughter, and for sharing some great times in your lives with me. It's been a privilege and a pleasure getting to know you. Congratulations for all you've achieved!

To my other friends, near and far, thanks for the comfort, support, chocolate, beer, laughs, optimism, and all the rest of it. You've been more fun than Ben, Jerry, Monty and Russell put together — and that's saying something. Particular thanks to Trish for giving me the courage to do this, by making it look easy. (Liar liar pants on fire.)

Thanks to my family, for listening patiently as I enthused about a topic that has all the intellectual mystery of an argument with a three-year-old. ('Why?' 'Because.' 'Because why?' 'Because I said so.') Thanks to my Dad, for having no idea what I was talking about, but being sure it was good anyway. I needed that faith! And to Dr. Joan Marshman — a.k.a. Mom — for knowing all too much about most of it, and helping me muddle my way through. Thanks for your tireless research, last-minute proofreading, and explaining pharmaceutical terms and concepts in small words I could understand.

Finally, I would like to express my sincere gratitude, respect and admiration to my advisor, Ingrid Meyer. To say that I could not have done it without you would be a

vast understatement; I don't think I'd even have tried. Your encouragement, advice, cheerleading, confidence, knowledge, wisdom, concision, humour and courage are unparalleled. Thank you for everything you've done for me. I wish you the very best.

Abstract

One of the most important aspects of a terminologist's work is extracting conceptual information about terms from texts. Because this task is so time-consuming, researchers are trying to develop tools which will extract conceptual information semi-automatically. Many of these tools are based on the use of linguistic indicators called *knowledge patterns*.

This thesis aims to identify some knowledge patterns in English and French which indicate the conceptual relation of *cause* and *effect*. This relation, though not as widely studied as those of *generic* to *specific* or *part* to *whole*, is critical in many subject fields including medicine and pharmaceuticals. For this reason, our research focuses on biopharmaceutical texts.

Our methodology involved building representative corpora in English and French, and then identifying possible knowledge patterns. The precision of the identified patterns was calculated in order to predict their possible effectiveness for semi-automatic knowledge extraction. We discuss some of the issues observed in the process of identifying these patterns, and those which might affect the subsequent implementation of the patterns for semi-automatic knowledge extraction. A brief interlinguistic comparison of the English and French patterns identified is also included.

Our research shows that the subject field of biopharmaceuticals contains many potentially productive knowledge patterns for the cause relation. However, there are also many issues which must be taken into account when identifying patterns and developing knowledge extraction tools.

Résumé

L'extraction de renseignements conceptuelles sur des termes constitue une des tâches les plus importantes et les plus difficiles du terminologue. Cette difficulté pourrait être surmontée par des outils d'extraction semi-automatique de connaissances exploitant des indicateurs lexicaux de relations conceptuelles (*marqueurs* ou *patrons*).

Le but principal de cette thèse est d'identifier certains marqueurs de la relation de cause à effet en anglais et en français. Cette relation, bien que moins étudiée que celle du générique au spécifique ou celle de la partie au tout, est très importante dans des domaines tels que la médecine ou la pharmacie. Pour cette raison nous avons choisi de travailler sur des textes dans le domaine des produits biopharmaceutiques.

À l'aide d'un corpus dans chacune des deux langues, nous avons identifié des patrons et calculé leur précision dans ces corpus, afin d'évaluer leur valeur potentielle pour l'extraction des connaissances. Nous présentons ici ces patrons, ainsi que les difficultés principales observées durant leur identification, et celles qui pourraient survenir lors de l'exploitation des patrons pour l'extraction semi-automatique de connaissances. Finalement, nous effectuons une brève comparaison interlinguistique des patrons.

Cette recherche a montré qu'il existe dans le domaine des produits biopharmaceutiques une gamme de patrons qui pourrait être utilisée pour l'extraction d'exemples de la relation de cause à effet. Cependant, les difficultés inhérentes à l'identification et l'exploitation de ces marqueurs doivent également être prises en compte.

Introduction

Computers have revolutionized terminology work over the last several years. They now assist in locating, storing, organizing and accessing information. Specialized texts in a wide range of subject fields are available on the Internet. In short, the information that is required to carry out terminological research is becoming more available and accessible.

There are two main foci of terminology work. The first, of course, is the terms used in a certain subject field (also called a *domain*). The second is the concepts that these terms designate. It is generally agreed that in order to do high-quality terminology work, terminologists must develop a good basic knowledge of the concepts in the subject field in which they are working.

One of the most important contributions to the accessibility of conceptual information has been the introduction of terminological corpora. These collections of machine-readable texts can supply terminologists with information about both terms and concepts. However, the usefulness of corpora is dependent on the terminologist's ability to access the information they contain.

While information about terms and their usage may be relatively easily accessed, information about concepts and how they relate to each other can be difficult to extract. In addition, in some cases useful information may be lost in masses of less helpful data. This problem is likely to worsen as larger quantities of data become available and are integrated into corpora. One of the challenges today is to develop more efficient ways of accessing information about concepts and the relations between them.

One method of extracting conceptual information is by using *knowledge patterns*, textual elements that are often associated with a certain relationship between concepts. Research has already been carried out on several conceptual relations, such as those of generic to specific, part to whole, and function. Examples of knowledge patterns indicating the generic-specific relation are *is a*/*est un* and *type of*/*type de*; examples of part-whole knowledge patterns include *contains*/*contient* and *part of*/*partie de* (Davidson 1998; Meyer *et al*. 1999; Marshman *et al*. 2002).

This thesis aims to identify some of the knowledge patterns associated with the relation of cause and effect. Intuitively, one might guess that lexical items such as *cause*/*cause* and *result*/*résultat* might indicate a cause-effect relationship.

Two corpora, one English and one French, were used to identify possible knowledge patterns for the relation. Once identified, these patterns were tested on the corpora in order to evaluate their efficiency in extracting pertinent contexts. In addition, the issues observed in identifying and using these patterns for knowledge extraction were discussed, and the patterns identified for English were briefly compared with those for French.

Objectives

General Objective

The general objective of this thesis is to add to the existing research in the area of pattern-based knowledge extraction, in particular for the cause relation. By identifying possible knowledge patterns in French and English and observing some of the issues which surround the identification and application of these patterns, this thesis can contribute to the development of knowledge extraction tools for this relation.

Specific Objectives

Major Objectives

The major objectives of this thesis are:

To identify lexical knowledge patterns for the cause-effect relation in French and English in the domain of biopharmaceuticals;

1. To study the patterns' efficiency in extracting knowledge from corpora, and to measure the patterns' precision in order to identify promising candidates for further research;
2. To identify and discuss some of the issues encountered during the research process and those that seem likely to affect future applications of these patterns.

Secondary Objective

Once the French and English patterns are identified, this thesis aims to present a very preliminary interlinguistic comparison of these patterns in order to draw some conclusions about the similarities and differences that are likely to be found between the two languages.

Methodology and Tools

The theoretical background of this project involved two main subjects, the use of knowledge patterns for computerized knowledge extraction, and the theory of the cause relation. The first subject involves elements of conceptual analysis, corpus building, corpus analysis, knowledge extraction, and knowledge patterns. The second aspect is an expansion of the conceptual relation component of conceptual analysis, involving conceptual relations in general and the cause relation specifically. The complexity of the relation made this task both challenging and interesting.

Our methodology consisted of four major steps, listed below.

1. Corpus building:
    a. choosing the domain;
    b. researching the domain;
    c. building two corpora (one for English and one for French) in this domain.
2. Corpus analysis:
    a. using terms identified in the domain research to generate key-word-in-context (KWIC) concordances in each corpus using the concordancing software WordSmith Tools[1];
    b. identifying contexts in the concordances which illustrated the cause relation;
    c. identifying lexical elements in these contexts that appeared to be associated with this relation;
    d. using these possible knowledge patterns to generate additional KWIC concordances;
    e. analyzing these concordances in order to evaluate the precision of the patterns for knowledge extraction.
3. Comparing the English and French patterns:
    a. comparing the grammatical category, frequency, and variability of the patterns in the two languages;
    b. discussing the cognates observed in the patterns of the two languages.
4. Describing issues of identifying and using patterns:

[1] Distributed by Oxford University Press, http://www.oup.com/elt/global/isbn/6890/

a. identifying and grouping together issues observed in the identification and use of the knowledge patterns;
b. describing the impact these issues are likely to have for future projects.

Motivation

The development of new ways of extracting information from corpora is to me one of the most interesting and important fields in terminology today. As the volumes of available information grow, so must the efficiency of knowledge extraction tools.

In discovering new knowledge patterns that may be integrated into such tools, researchers can take steps towards a new era of information retrieval, in which users will be able to search for specific kinds of information about concepts and how they are related.

The cause relation is interesting for this kind of research because it is both complex and pivotal in many fields. Unlike some more clear-cut hierarchical relations, such as that of generic to specific, cause has many different sub-types. In-depth study is required to develop ways of extracting information about the relation accurately and easily. The relation of cause to effect is critical in many technical domains, among them medicine and pharmaceuticals. There is a great demand for information about the causes of disease and the effects of treatment, to name only two examples.

Through this research we hope to add to the existing body of knowledge about the cause relation and the knowledge patterns associated with it, and moreover to do so in a field that is fast-growing, pertinent and fascinating.

1 Concepts in Terminology

1.1 Introduction

This chapter will provide an outline of the theoretical background of this research. The topics discussed are terminology, concept analysis, knowledge extraction and knowledge patterns. In addition, the chapter will provide an outline of previous research and of how this project differs from others.

1.2 Terminology

1.2.1 Defining terminology

Terminology as a field of work and study can be defined in numerous ways. For our purposes, we have drawn on Sager:

> We see terminology as a number of practices that have evolved around the creation of terms, their collection and explication and finally their presentation in various printed and electronic media. ...
>
> Terminology is the study of and the field of activity concerned with the collection, description, processing and presentation of terms, i.e., lexical items belonging to specialised areas of usage of one or more languages. In its objectives it is akin to lexicography which combines the double aim of generally collecting data about the lexicon of a language with providing an information, and sometimes even an advisory, service to language users. (1990:1-2)

and on Cole:

> ... [T]erminology is fundamentally practical in outlook. Its aim is not merely to study the semantic and syntactic characteristics of terms as an end in itself, but also to provide a theoretical structure which will permit the compilation of correct and verifiable [collections of terms] ... (1987:77)

Consistent with these two approaches, we will use *terminology* to refer to tasks which involve not only the *collection* and *standardization* of terms used in specialized

language, but also the *construction of theoretical structures of the concepts* within these specialized areas of knowledge, often referred to as *domains* (Meyer and Mackintosh 1996:260) or *subject fields* (Terminology and Standardization Directorate 2002)[2]. Thus there are two key elements in the field of terminology: *terms* and the *concepts* they designate.

1.2.2 Defining a term

Cole (1987:77) cites the following definition of 'term':

> a word (simple term) or multiword expression (complex term) that designates a specific concept within a given subject field. (Terminology Directorate 1983:59)

1.2.3 Concept analysis

As the pioneer of the field, Eugen Wüster (1981), described it, terminology is a field with an *onomasiological* approach. That is, a terminologist begins by studying concepts, and only then progresses to study the terms that are used to describe them. The concept is the heart of terminology work. Thus, the first step in any project is to study the concepts involved. This work is called concept analysis, and has been defined as follows:

> The process of discovering and representing the conceptual structures underlying the terms of a domain. This process is both the cornerstone and the most difficult aspect of terminography. (Meyer and Mackintosh 1996:261)

The importance of this analysis to terminology research was also noted by Sager:

> In terminological theory it is accepted that concepts should be ordered according to some conceptual classification scheme and presented in a systematic structure. In order to do this concepts are characterised by the relationships they form with neighbouring concepts. (1990:28)

Terminologists study both the attributes of concepts and the relations between them. *Attributes* are inherent characteristics or properties of concepts (e.g., height, width,

---

[2] In this thesis we will use the terms *domain* and *subject field* interchangeably.

colour). *Relations* are the ways in which concepts relate to each other within the conceptual network of the domain (e.g., generic to specific, part to whole, function, and of course, cause and effect) (Sager 1990:25; Meyer *et al*. 1997:102). In this project, our conceptual analysis concentrated on relations rather than attributes.

1.2.3.1 Conceptual relations

We have determined then that conceptual relations are the links between concepts in a specific domain, and that they are analyzed in the stage of conceptual analysis:

> To say that a term is used in a given subject field is to say that the concept which it designates forms part of a particular network of concepts. Concepts, like words, do not exist *in abstracto*. To arrive at an adequate understanding of a particular concept, the relations between it and the other concepts in a given field must be understood. Even concepts that seem, superficially, to correspond may be found to be differentiated when their relations to the other concepts in the field are taken into account. (Cole 1987:78)

Terminologists use the information acquired through conceptual analysis in a variety of ways. A thorough analysis allows terminologists to clearly delineate a concept and to differentiate it from others that are similar. It allows them to discover how the concept fits into the conceptual system of the field being studied. Once this is known, a concept can be properly defined (for example, through an Aristotelian definition, which consists of the generic plus the characteristics that differentiate a given concept from its generic and co-ordinate concepts) and the terms that are used to designate it can be reliably identified. Similar terms can be differentiated from one another, and synonyms identified and evaluated. Equivalency between terms in different languages can be established. Appropriate neologisms can be suggested, if needed. In short, a thorough analysis of concepts' attributes and relationships to one another allows terminologists to carry out high-quality, reliable terminology work.

1.2.3.1.1 Classification of conceptual relations

While conceptual relations are used to classify concepts within a domain, they themselves may also be classified. Much research has been done on this, including Lyons 1977, Cruse 1986, Ahmad and Fulford 1992, Garcia 1996/97, Nupponen 1999, and Barrière 2001/02.

However, despite this work and the importance of conceptual relations there is no commonly accepted, straightforward classification of relations:

> [T]here is little consensus as to the names of the various relations, let alone their forms. Indeed, attempts to come up with lists of relations rarely even identify the same set, notwithstanding the use of synonyms. (Bowden *et al*. 1996:157)

Relations have been categorized in various ways. One of these is hierarchical versus non-hierarchical relations. For example, the relations of hyperonymy (generic to specific) and meronymy (part to whole) are classified as hierarchical relations because they can be represented in a multi-level structure: a colour laser printer is a type of laser printer, which is a type of printer, which is a type of computer peripheral, etc.

However, not all relations are considered to be hierarchical. Some non-hierarchical relations include those of process/product, object/operation, activity/place (Sager 1990), producer/product, action/tool, and container/contents (Pavel and Nolet 2001). What is more, there is not always agreement as to which relations are hierarchical and which are not. For example, Sager (1990) and Wüster (1981) view the cause relation as non-hierarchical; Sager calls it complex, i.e., not conveniently "captured by straightforward generic and partitive structures" (34-5), whereas Barrière (2002) states that a hierarchical model is necessary to adequately represent the relation:

> [The process of semantic relation refinement] will be required to correctly characterize the complex cause relation. ... These factors argue for the use of a representational scheme which can accurately record both loose and precise indicators. We therefore

> conclude this section by emphasizing the importance of defining relations within a hierarchical framework which will allow groups of interrelated relations to be defined with varying degrees of ambiguity, and allow patterns applicable only to certain sub-types of a more general relation to be recorded. Based on our brief exploration on meronymy the use of such a model seems almost essential.[3]

For the purposes of this research we adopted Barrière's hierarchical analysis of the relation.

1.2.3.1.2 Adequacy of a relation's classification

Regardless of how relations are classified, the most important factor is whether the classification is adequate to account for and explain the majority of the occurrences of the relation in the domain in which it is applied:

> If a classification scheme is comprehensive, it should provide correct categories to most of all of the causal relations found in text. (Barrière 2002)

This research will provide data that can be used to assess the adequacy of the classification we adopted.

1.2.3.2 The cause relation

The cause relation has been of interest to scholars for hundreds of years. One of the earliest classifications is that of Hume (1739) (cited in Chaffin and Herrmann 1988:291). Since then the relation has been studied continuously in a philosophical context, but it was not until the 1970s that a new kind of analysis of the relation emerged, led by Leonard Talmy at the University of California at Berkeley[4]. Below are some examples of researchers who have studied the relation.

---

[3] We believe that in the first sentence Barrière is using "complex" in a more general sense than Sager's terminological use (to describe relations which are not easily classifiable hierarchically).

[4] Talmy's work seems not to build on the philosophical studies of the relation, but rather to introduce a new point of view. For this reason, previous analyses will not be discussed here.

1.2.3.2.1 Leonard Talmy

Perhaps the definitive study of the cause relation today is the *force dynamic* theory developed by Leonard Talmy, Associate Professor of Linguistics and Director of the Cognitive Science Program at the University of California at Berkeley. Talmy (1985, 1988, 2000) undertook his analysis of the cause relation through force dynamics because he felt there was a clear need for a more sophisticated understanding of the relation:

> A semantic category that has previously been neglected in linguistic study is that of "force dynamics" — how entities interact with respect to force. Included here is the exertion of force, resistance to such a force, the overcoming of a resistance, blockage of the expression of force, removal of such blockage, and the like.
>
> Though scarcely recognized before, force dynamics (FD) figures significantly in language structure. It is, first of all, a generalization over the traditional linguistic notion of 'causative': It analyzes 'causing' into finer primitives and sets it neutrally within a framework that also includes 'letting,' 'hindering,' 'helping,' and still further notions not normally considered in the same context. (1988:49-50)

Talmy began by defining the cause relation on the basis of an interaction of opposing forces. He used the terms *agonist* (*Ago*) and *antagonist* (*Ant*) to designate the two forces and examined the interaction between them in both static (steady-state) and dynamic (shifting) force-dynamic patterns. By steady-state force-dynamic patterns, he referred to those in which the relative strength of the forces remained the same; shifting force-dynamic patterns were those in which the relative strengths of the forces changed. Figure 1 below describes the effects of static interaction between an agonist and antagonist of different relative strengths. Since the forces are always opposing, it is their relative strengths that determine the outcome of the interaction, the effect. Figure 2 describes the shifting counterparts of the interactions in Figure 1. Both of these figures are taken from Barrière (2002), as for our purposes this work provides a simpler and therefore clearer portrait of the models described in Talmy (1988).

Case	Tendency	Relative Strength	Result
a	*Ago* toward rest	*Ago* weaker than *Ant*	*Ago* in movement
b	*Ago* toward rest	*Ago* stronger than *Ant*	*Ago* rests
c	*Ago* toward movement	*Ago* stronger than *Ant*	*Ago* in movement
d	*Ago* toward movement	*Ago* weaker than *Ant*	*Ago* rests

Figure 1: Talmy's force dynamics for steady-state relationship between forces (adapted from Barrière 2002)

To further explain these figures, we will use the examples given by Talmy (1988:55) to help clarify. The cases given above for steady-state relations are the following:

a. The ball kept rolling because of the wind blowing against it.
b. The shed kept standing despite the gale wind blowing against it.
c. The ball kept rolling despite the stiff grass.
d. The log kept lying on the incline because of the ridge there.

In *a*, the ball's tendency to remain still is overpowered by the wind's tendency to roll it along the grass, and it does roll. In *b*, the shed's tendency to stay standing is stronger than the wind's tendency to blow it down, and the shed remains standing. In *c*, the ball's tendency is different from in *a*; it is inclined to roll. Even the opposing force of the stiff grass is not able to keep it still. Finally, in *d*, the log that tends to roll (likely because of the incline) is being kept still by the ridge's tendency to keep it there.

Case	Tendency	Relative Strength	Result
e	*Ago* toward rest	*Ant* becomes stronger than *Ago*	*Ago* in movement
f	*Ago* toward movement	*Ant* becomes stronger than *Ago*	*Ago* rests
g	*Ago* toward movement	*Ant* becomes weaker than *Ago*	*Ago* in movement
h	*Ago* toward rest	*Ant* becomes weaker than *Ago*	*Ago* rests

Figure 2: Talmy's force dynamics model for shifting relationships between forces (adapted from Barrière 2002)

The cases of the shifting force-dynamic patterns are more complex because of their changing nature; rather than the constant forces in the examples above, the antagonists in this case become stronger or weaker. The net effects of the forces thus change. The situations outlined by Talmy (1988:57) correspond to the sentences below:

e. The ball's hitting it made the lamp topple from the table.
f. The water's dripping on it made the fire die down.
g. The plug's coming loose let the water flow from the tank.
h. The stirring rod's breaking let the particles settle.

In *e*, the lamp tends to stay in place, but the force towards its movement becomes stronger (when the ball hits it) and it moves. In *f*, the fire's tendency to burn is counteracted by the increasing force of the water dripping, and it dies down. In *g*, the obstruction to the water's tendency to flow from the tank is removed, and it is able to do so. Finally, in *h*, the stirring rod becomes unable to keep the particles in motion against their tendency to stay still, and they settle.

Thus we see eight possibilities for the interaction between opposing forces, which can be used to classify the majority of occurrences of the cause relation.

Talmy stressed that these models present very simplified interpretations of force interactions (1988:92), not describing many aspects (for example, the speed with which the result occurs). The effect described — the rest or motion of an entity — in fact represents a much larger spectrum of possibilities. This range is further described in the section on Barrière's classification.

Moreover, however scientific his approach, Talmy took care to note that the linguistic picture of the force dynamics involved in the cause relation are not those of pure physics (1988:92). In language, he stated, there are conceptions applied to objects (e.g., intrinsic tendencies toward motion or rest) that are not necessarily present or real in a physical sense. For example, to say that a ball has a tendency to roll (as above in sentence c) is not strictly true in a physical sense; it is subject to inertia like any other object. While it may roll more easily than a cube, for example, with no forces acting upon it a ball will stay still (as in sentence a above). However, in a linguistic context, rolling

(or bouncing, or being thrown) is simply what a ball does. Consider that we are much more likely to describe a ball rolling or bouncing than to describe it as staying still without mentioning a reason for its doing so.

Physical laws aside, the representation of the cause relation described by Talmy is intuitive and can be applied to linguistics. This was the task of other researchers: to develop and transpose the theoretical outline of the relation onto language from the perspective of knowledge extraction.

1.2.3.2.2 Daniela Garcia

One researcher who has done in-depth study of the cause relation is Daniela Garcia (1996, 1997). In the course of her work, she has developed both a hierarchy of the cause relation and a tool, COATIS, which helps to extract information on this relation from natural language texts in French.

Garcia's hierarchy is based on the work of Talmy. It divides the cause relation into six main sub-types: *création*, *maintien*, *empêchement*, *modification* (which has two sub-types, *facilitation* and *gêne*), *laisser-faire*, and *contribution* (with its own sub-type, *collaboration*) (Garcia 1996:99).

In her work with COATIS, Garcia concentrated on verbal indicators of the relation. However, some other patterns, including nouns and conjunctions, were also considered.

1.2.3.2.3 Anita Nupponen

Anita Nupponen (1994) also worked on developing a classification of the cause relation, which divides causes into causative agents ("substances, materials or other

elements that cause an effect" (39)), producing causes (an event, an action or a process (39)), an explanatory cause (a fact or a state (40)), and counteracting causes ("an agent, an event, a state or a fact that counteracts the causal process and prevents the effect" (40)). She also divides the effect into resulting states, resulting products, resulting events and complications (effects caused by the first effect) (40).

1.2.3.2.4 Caroline Barrière

Professor Caroline Barrière at the University of Ottawa has also prepared a classification of the relation, one that was adopted for this project. Barrière (2001, 2002) adapted Talmy's analysis and developed a hierarchy of the relation, dividing it into two main categories (existence and influence dependencies), each of which is further divided into more specific sub-categories. Each of these will be described below.

As the figures illustrate, the effect of the force interaction has been expanded to represent not only rest or motion on the part of an object, but also the existence or non-existence of an entity and the occurrence or non-occurrence of an event (existence dependency) and changes in the character of an event or entity (influence dependency.)

Figure 3 presents Barrière's hierarchical classification of the relation and the effects of the force interaction, in addition to some possible knowledge patterns. The left-hand column indicates the dependency; the column to its right indicates the sub-categories. The third column indicates the effect of the interaction between the opposing forces and the right-hand column lists some English knowledge patterns that may indicate this sub-category of the relation. *E* in the third column designates an entity or event, *~E* the non-existence of an entity or the non-occurrence of an event, E_1 an entity or event as

it is before the interaction of forces, and E_2 the entity or event after the interaction of the forces.

Existence dependency	Creation		~E → E	create generate produce
	Destruction		E → ~E	kill eliminate destroy
	Maintenance		E → E	allow keep maintain
	Prevention		~E → ~E	prevent discourage control
Influence dependency[5]	Modification		$E_1 <> E_2$	influence change modify
		Increase	$E_1 < E_2$	increase improve promote enhance
		Decrease	$E_1 > E_2$	reduce decrease shorten slow down deter discourage
	Preservation		$E_1 = E_2$	maintain keep retain

Figure 3: Barrière's classification of the cause relation (2002)

Readers will notice that occasionally in this table, the same pattern (e.g., discourage) is used to illustrate more than one sub-type of the cause relation (i.e., prevention and decrease). This sort of polysemy will be discussed further in Chapter 6. For the purposes of this research, each pattern was generally classified according to the sub-type of the relation that it most commonly indicated. Of course, this does lead to a somewhat limited picture of the relation, excluding much of its complexity.

[5] Barrière's representation of the influence dependency has been simplified for the purposes of this project. For her more technical representation, see Barrière 2002.

Below are descriptions of each of the sub-categories of the relation.

1.2.3.2.4.1 Existence dependency

As the term implies, this category designates cause relations that determine the *existence* or *non-existence* of an entity or the *occurrence* or *non-occurrence* of an event. The dependency is divided into four sub-types, creation, destruction, maintenance, and prevention.

1.2.3.2.4.1.1 Creation

This type of cause relation occurs when the interaction between the opposing forces brings into being an entity that did not previously exist or causes an event that was not previously occurring to take place. Barrière gave *create* and *produce* as examples of knowledge patterns indicating this relation.

1.2.3.2.4.1.2 Destruction

The opposite of creation, this relation is found when the interaction between opposing forces causes something that previously existed to cease to exist, or an event which was previously taking place to stop. Barrière gave *destroy* and *kill* as examples of this sub-type of the relation.

1.2.3.2.4.1.3 Maintenance

Maintenance designates a situation in which an entity or event existed or was occurring before the interaction of the opposing forces and continues to exist or occur thereafter. Barrière included patterns such as *maintain* in this category.

She also included patterns such as *allow* in this group. To clarify what might be a less intuitive classification than the others, we can think of 'allowing' as 'not-preventing' or as the polar opposite of 'preventing,' as 'creating' is the polar opposite of 'destroying.' From this point of view, the parallel is clearer.[6]

1.2.3.2.4.1.4 Prevention

In prevention, an entity or event did not exist or occur before the interaction between opposing forces, and continues not to exist or occur. Barrière's pattern examples for this sub-category include *prevent* and *avoid*.

1.2.3.2.4.2 Influence dependency

The influence dependency parallels the existence dependency, but with one major difference: in this kind of relation the interaction between opposing forces determines not the existence or non-existence of an entity or the occurrence or non-occurrence of an event, but rather some *characteristic* or *feature* of that entity or event. The characteristic in question can be modified, increased, decreased, or preserved.

[6] In future research, it would be interesting to examine alternate, more intuitive ways of classifying the relations indicated by such verbs as *enable* and *allow*. This issue will be further discussed in Chapter 6. However, for the purposes of this project it was decided to retain the established classification.

1.2.3.2.4.2.1 Modification

In Barrière's classification, modification is a sub-category that in turn includes those of increase and decrease. It groups together all types of the relation in which the interaction between forces causes a change of one kind or another in a characteristic or feature of an entity or event. However, the group is not limited to increase and decrease, but also constitutes its own sub-type of the influence dependency.

This sub-group includes cases in which the forces' interaction have an effect on a characteristic or feature of an event or entity, but in which the kind of modification may not be specified. An example is found in the patterns *influence* and *change*: without further explanation there is no way to know what form this influence may take or what kind of change occurs.

Thus, for the purposes of classifying the knowledge patterns in this project, modification was used as one of four sub-categories of the cause relation. Increase and decrease in turn were considered to be co-ordinate categories, although logically they do constitute specific types of modifications.

1.2.3.2.4.2.2 Increase

In this type of cause relation, the characteristic of the entity or event is intensified or increased by the interaction between the opposing forces. Barrière's examples for this type of relation include *increase* and *accelerate*.

1.2.3.2.4.2.3 Decrease

This is the mirror image of the increase sub-type: a characteristic of the entity or event is lessened or decreased by the interaction between the opposing forces. Examples given by Barrière include *decrease* and *inhibit*.

1.2.3.2.4.2.4 Preservation

This sub-type of the relation is analogous to the existence dependency's maintenance sub-category, since the characteristic of the entity or event exists before the interaction of the forces and continues to exist unchanged afterwards. However, perhaps less intuitively, it also includes those instances in which a characteristic or feature is *not* present and continues not to be present after the interaction between the forces. Barrière's examples of knowledge patterns for this sub-category include *maintain* and *keep*.

1.2.4 Suitability of the relation as a subject for research

We chose to study the cause relation for two reasons. First, in order to develop knowledge extraction tools that can extract as much knowledge as possible from texts, we must expand the coverage of conceptual relations. Second, a personal interest in the domain of pharmaceuticals led me to consider research that would be pertinent in this context.

Hamon and Nazarenko (2001:188) noted that the cause relation is "especially important in technical domains where diagnosis, repair, planning, [or] physical action are involved." These include many of the most important areas of research today; certainly the domains of medicine and pharmacy fit this description. Nupponen addressed this issue more specifically:

> In many subject fields causality is an important factor, and finding the causes and effects, and the relations between these, is essential. Examples might be medicine, law, physics, biology, etc. In medicine, for instance, the questions asked might include the following:
>
> - What caused this disease?
> - What are the complications of this disease?
> - What effects does this medicine have?
> - What side effects does this treatment have?
> - How can we prevent this disease?
>
> ...
>
> In terminological analysis this subject field knowledge of causal structures may be used to organise the concepts and terms as well as to define the concepts, etc. In order to make the terminologists' work easier, we need, however, some general information about how causality functions and how the concepts involved may be analysed and organised. (1994:36)

Thus we see that research on the cause relation is not only pertinent, but is in fact critical for developing effective knowledge extraction tools. Moreover, the domains of medicine and pharmacy constitute an excellent starting point for this sort of research.

Now that we have established the importance of research into the cause relation, we will describe how this kind of research is carried out.

1.3 Corpora

More and more terminological research is being carried out on corpora. Pearson (1999:22) cited Sinclair's (1994:2) definition of a corpus: "collection of pieces of language that are selected and ordered according to explicit linguistic criteria to be used as a sample of the language," and Atkins, Clear and Ostler's (1992:1): "a subset of an ETL (electronic text library) built according to explicit design criteria for a specific purpose."

Meyer and Mackintosh described the importance of corpora for terminology specifically:

> Terminographers do not usually rely on human experts alone. Rather, as a complementary — and often a *primary* — knowledge source, they use *texts* that are representative of the domain's specialized discourse. Indeed, it is generally accepted that the quality of a terminography project is directly related to the quality of the

> documentation on which it was based. In the words of Rogers and Ahmad (1994:840): "the use of corpora, in particular, electronic corpora, is equally well motivated, if not better motivated, in special-language work than in general language work." The reason why corpora, both paper-based and electronic, might be considered even "better motivated" in terminography than in lexicography has to do with the minimal role of introspection (Sinclair 1985) as a source of terminological evidence. (1996:264-5)

1.3.1 Issues in corpus-building

One important issue concerning corpora is that of balance. Meyer and Mackintosh (1996:267) discuss the fact that terminologists require evidence of the full range of terms used in the field. Because of this, they value access to texts representing various degrees of technicality (1996:270) and what are referred to by Pearson (1998:36) as *communicative settings*. However, the texts must not only include a broad range of terms, but also data on the attributes of and the relations between the concepts in this field.

The size of a corpus is another important aspect of corpus design. It is important that the corpus be large enough to provide reliable data on the concepts of the field in question:

> The chief advantages of corpus data ... are, of course, dependent upon the corpus itself being large enough and sufficiently well-balanced to be reliable. (Rundell and Stock 1992a:13)

However, as Meyer and Mackintosh state (1996:268):

> It is generally accepted that [a terminographical corpus] can be much smaller than a corpus for general-language study (Rogers and Ahmad 1994:849, Engwall 1994:51). With terminographical corpora, quantity is a much less important design issue than quality.

1.4 Knowledge Extraction

Once a corpus is built, it must be analyzed in order for the terminologist to extract the knowledge it contains. This knowledge may include terms, examples of usage and phraseology, attributes of concepts, and relations between concepts. Here we will focus on information about conceptual relations.

Such information is found in contexts that Meyer (2001) calls *knowledge-rich*.

1.4.1 Knowledge-rich contexts (KRCs)

Meyer defines knowledge-rich contexts as follows:

> By *knowledge-rich context,* we designate a context indicating at least one item of domain knowledge that could be useful for conceptual analysis. In other words, the context should indicate at least one conceptual characteristic, whether it be an attribute or a relation. (Meyer 2001:281)

KRCs can occur within a single sentence (*intra-sentential*) or can include two or more sentences (*extra-sentential*) (*Ibid.* 282).

Generally a terminologist will start by extracting contexts in which a certain term is used. However, how can he or she effectively gather information about a particular relation? One answer to this question is the use of knowledge patterns.

1.5 Knowledge patterns

Conceptual relations are, to some degree at least, intuitive. In support of this assertion, Chaffin and Herrmann (1988:289) cited the ability of people to intuitively evaluate the similarity of relations, to distinguish relations from one another and — most significantly for the purposes of this research — to express relations using every day words and expressions. They went on to state that: "While names of relations are few and little used, there are ways of *expressing* relations that are part of the average speaker's everyday vocabulary." (*Ibid.* 315)

Many different terms have been used for these ways of expressing relations: *formulae* (Lyons 1977), *diagnostic frames* and *test frames* (Cruse 1986), *knowledge probes* (Ahmad and Fulford 1992) *defining expositives* (Pearson 1996), and *triggers*

(Bowden *et al.* 1996). In this project we will use the term *knowledge pattern*, which is the term used by Meyer (2001) and others (Morgan 2000, Marshman *et al.* 2002).

What exactly is a knowledge pattern? Meyer defined the concept as follows:

> By [knowledge pattern], we loosely designate free (i.e. non-collocational) language combinations that frequently identify a particular conceptual relation or attribute. For example, patterns such as **X is a kind of Y**, **an X is a Y**, **As include Bs, Cs and Ds**, indicate generic-specific relations.... Such [knowledge patterns] may be reduced to a fairly limited set of words/phrases which occur in typical syntactic patterns. Ahmad/Rogers 1992 have further proposed that such contexts serve as **knowledge probes** – search patterns – to be used by a corpus analysis tool in assisting terminologists with knowledge acquisition. (1994:8)[7]

In French, this concept is designated by terms such as *patron* (L'Homme 2001), *indicateur* (Garcia 1996) and *marqueur* (Séguéla 1999).

Some examples of previously identified knowledge patterns for the relation of hyperonymy include *is a*, *type of*, and *kind of*. Examples of knowledge patterns indicating meronymy include *part of* and *contains* (Meyer *et al*. 1999; Marshman *et al*. 2002).

While some knowledge patterns are generic, that is, present in many different domains, others are domain-specific, limited to or much more frequent in a particular field. For example, for the relation of hyperonymy, *is a*, as in "Windows is an operating system," is considered to be generic; in contrast, *flavour*, as in the sentence "Arity is a flavour of Prolog," is domain-specific, being rarely used in this sense outside the field of computing (Meyer *et al*. 1999). In this project we did not differentiate between patterns which appeared likely to be generic and those which seemed likely to be domain-specific.

---

[7] Note that the term originally used in this context was *knowledge-rich context*. The terminology has evolved in the intervening years, with the term *knowledge pattern* being coined to differentiate this concept from that of contexts which provide useful information for terminologists. (Ingrid Meyer, personal communication).

1.5.1 Types of knowledge patterns

While all of the examples of knowledge patterns given above consist of words or combinations of words, these are not the only possible forms of knowledge patterns. In fact, there are three types identified by Meyer *et al*.:

> Knowledge patterns can be classified into three basic types: 1) *Lexical patterns* involve specific lexical items and can convey all types of relations. For example... HYPERONYMY patterns include *is the*, *is a*, *such as*, *and other* and *known as*. 2) *Grammatical patterns* apply to a small numbers of relations. The patterns NOUN + VERB (with some verbs excluded), for example, is highly productive for indicating the FUNCTION relation. 3) *Paralinguistic patterns* include punctuation, as well as various elements of the general structure of a text. The phrase "placenta previa (a placenta abnormally located in the lower part of the uterus)", for example, illustrates how parentheses may indicate HYPERONYMY. (1999:257-8)

In this project, we focused on identifying *lexical* knowledge patterns.

1.5.2 Using knowledge patterns for knowledge extraction

As mentioned above, corpora are repositories for immense amounts of knowledge. The challenge is to locate and extract this knowledge. One way of achieving this goal is by using computerized tools to search through corpora and extract pertinent, knowledge-rich contexts. Although ideally, this would be done completely automatically, the limits of current technology make a computer-*assisted* (semi-automatic) process a more realistic goal:

> The idea of automatically extracting KRCs on the basis of recurring linguistic and paralinguistic patterns is clearly attractive for terminography. Furthermore, early results (Ahmad/Fulford 1992, Bowden *et al* 1996, Condamines/Rebeyrolle 1998, Davidson *et al* 1998) appear promising enough to hope that conceptual sampling tools will eventually be part of every terminographer's workstation. However, we feel that conceptual sampling is not likely to be fully automated in the near future: rather, it will be a semi-automatic, terminographer-assisted technology. Terminographers will need to examine the output of sampling tools critically to eliminate errors (i.e., "noise"). Furthermore, even when the output is correct, they will want to bring their world knowledge to bear on it, in order to maximize its value. (Meyer *et al*. 1999:258)

Semi-automatic knowledge extraction tools could be programmed to search corpora using knowledge patterns that commonly indicate conceptual relations. The tool

would then extract the contexts containing the patterns — or the patterns in combination with a particular term specified by the user — and present them for evaluation. Many researchers are currently working on such knowledge extraction systems[8]. Of particular interest for this project is COATIS (Garcia 1996), since it deals with the cause relation in French.

1.5.3 Evaluating extracted contexts

Condamines and Rebeyrolle (2001:130) describe the evaluation of extracted contexts, grouping them into:

- bad contexts; i.e., contexts not showing the expected relationship or showing the [desired] relationship but within a specific or subjective point of view...

- good contexts; i.e., contexts showing the expected relationship with a generic point of view.

Classifying contexts as "good" and "bad" was a challenging aspect of our research. It will be discussed in detail in Chapter 2.

1.5.4 Evaluating system performance

A system's performance may be measured using two indicators, *precision* and *recall*:

> Precision is calculated as the number of extractions the system gets right divided by the number of extractions the system gives. Thus it answers the question "what fraction of the given answers were correct?". Precision says nothing about the coverage of a system, and so it is possible for a program to have a very high precision and yet extract very little of the available knowledge. The recall metric resolves this omission. Defined as the number of extractions the system gets right divided by the number of possible extractions, recall shows how comprehensively knowledge is vacuumed from the text. (Bowden *et al*. 1996:155)

---

[8] Knowledge extraction systems cited in Condamines and Rebeyrolle (2001:135) include Seek for hyperonymy (Jouis 1994), Mantex (Rousselot *et al*. 1996), Prométhée (Morin 1999), Caméléon (Séguéla 1999) and DocKMan (Skuce and Kavanagh 1999).

While this seems simple, the complex nature of knowledge patterns and the variations inherent in natural language texts cannot be ignored:

> In evaluating a knowledge-extraction system, one ultimately wants to know how good its recall and precision are. While these calculations may be straightforward in some types of information retrieval, they are highly complex in the case of KRCs. As mentioned above, there exists no consensus among linguists on how many relations there are, and on the precise nature of every relation. This causes problems in determining what constitutes a "hit". (Meyer 2001:298)

1.5.4.1 Recall

The pertinent contexts that are not found by the system are called *silences*. In order to determine the number of silences, it is necessary to examine the corpus by hand:

> Silences are KRCs that express a given relation in the corpus, but that are not found by the system. Normally, they are missed due to some lack in the existing knowledge patterns for that relation. Finding silences is a labour-intensive process, since it involves manually identifying all the valid KRCs for the relation in question. (Meyer 2001:294)

However, this task in itself can pose problems:

> There is a fundamental problem concerning the calculation of recall. This is the difficulty in assigning a figure to the number of relations present in a text. Preliminary reader trials using short texts (c. 2000 words) have shown that there is not often a consensus on what constitutes the relation set. (Bowden *et al*. 1996:157)

Given the labour-intensive nature and complexity of counting silences, we felt that for this project the best measure to use for the evaluating the productivity of the patterns was precision.

1.5.4.2 Precision

High precision is critical for a good knowledge extraction system; a knowledge extraction tool can only be considered efficient if a high percentage of the contexts identified are pertinent and only a small percentage are not[9]. Precision is measured only

[9] As seen above, these non-pertinent contexts are referred to as *noise.*

in regard to the specific corpus (Séguéla 1999:52). As discussed earlier, patterns may be far more efficient in some domains than in others.

Since precision involves evaluating the output generated by a pattern and not the corpus as a whole, it was much more straightforward to measure. However, this is not to say it is always simple:

> Even in cases of relatively uncontroversial relations, for example, HYPERONYMY, problems occur when the relation is not "perfectly" expressed in real text. For example, should we count as a hit for HYPERONYMY a case where it is not the immediate hypernym that is used...? Shall we accept examples [in which there is a] conceptually illogical hypernym... ? (Meyer 2001:298)

1.5.4.3 The interplay between recall and precision

As we examine the measures of recall and precision, we see that they are to a certain extent fundamentally opposed:

> The process of adding new patterns as a result of analyzing silences raises one significant problem: while every addition to the knowledge patterns should increase the number of hits compared with the preceding pass, it also has the potential to increase the noise. To use standard information retrieval terminology, any gain in *recall* tends to cause a reduction in *precision*. (Meyer 2001:294)

This issue will be discussed further in Chapter 6.

1.5.5 Possible applications of this research

The knowledge patterns discovered in this research were tested in a preliminary way, by measuring their precision in the corpora constructed for the project. This is the first evaluation of their possible usefulness, and the first step in discovering whether they could be useful for semi-automatic knowledge extraction.

Below is a brief outline of how this information may fit into the development of a knowledge extraction system.

The results of this research will help to determine whether the patterns should be further evaluated. Patterns that show very low precision in these corpora might be

excluded from further study. Those that seem to be precise, however, could then be tested for their precision in other corpora, as well as their recall potential. They could also be refined (e.g., by studying the search windows appropriate for use with multi-word patterns, or the patterns' average proximity to the terms to which they are linked). The patterns could also be more definitively classified by the sub-type of cause relation they most commonly indicate. They could then be programmed into a trial semi-automatic knowledge extraction tool.

Such a tool would extract knowledge from natural language texts by searching for knowledge patterns and identifying pertinent sentences and/or terms. It might also be programmed to classify these hits by the type of cause relation they represent, as reflected by the pattern used.

The tool would be tested by measuring its precision and recall on some sample corpora. Using the results of these tests, the researchers could refine the patterns, search windows, and pattern classification as needed. Then the testing cycle would begin again, until the system returns the best balance of precision and recall possible. The system would then be ready to be tested in a real-world situation by language workers such as terminologists.

1.6 Originality and usefulness of this research

We saw above that the cause relation has already been researched by scholars such as Talmy and Nupponen, and work has already been done by researchers such as Garcia and Barrière to identify knowledge patterns and develop knowledge extraction systems for the cause relation. How then is the present research project useful?

First, this research is being carried out in a new subject field. The field of biopharmaceuticals is an ideal one for researching this relation, and may provide insights into the relation and the associated knowledge patterns in this particular field. Since the subject field is discussed in thousands of documents every year, domain-specific data could be integrated into a tool that would be useful for health professionals and patients as well as terminologists.

Second, new patterns may be discovered with each project, and the precision of previously identified patterns can be re-evaluated in each corpus. The new data from this project can be added to data from other projects to give a more complete picture of patterns' usefulness.

Third, this research deals with both English and French. Most of the previous projects have dealt only with one language or the other (e.g., Garcia 1996, 1997 for French; Barrière 2001, 2002 for English). The bilingual approach used here is valuable because it can contribute to the development of bilingual knowledge extraction tools by providing data in two languages, classified according to a single system.

Fourth, this project will allow us to examine the issues of pattern identification and application that are relevant not only to research on knowledge patterns and knowledge extraction in general, but also to this field and this kind of bilingual approach in particular.

Taking into account all of these points, it seems clear that this project has something to add to the existing body of research in this area. The next chapter will detail how the project was carried out.

2 Methodology

2.1 Introduction

This chapter describes and explains the methodology used in this research. The explanation will be divided into five sections: choosing a classification of the cause relation; selecting the domain and sub-domains; building the corpora; choosing and using the concordancing tool; and the analysis itself. Much of the methodology is based on that used in similar projects carried out at the University of Ottawa (e.g., Morgan 2000).

2.2 Choosing a classification of the relation

Since the cause relation is so complex, it was important to choose an established classification of the relation that would be appropriate for the project. As we described in Chapter 1, we chose that of Barrière because it provided a clear, logical classification.

Once the framework for the project was established, with the aid of the classification of the cause relation, the design of the research project itself was developed. This included the choice of domain and sub-domains for the research.

2.3 Selecting the domain and sub-domains

The research was carried out on corpora built in the domain of biopharmaceuticals for a variety of reasons, including the field's pertinence to current scientific research and research needs, the availability of appropriate texts in both English and French, and appropriateness for terminological research on the cause relation. The sub-domains were chosen using the same criteria in order to more precisely

delineate the subject field. In the next section, a brief overview of the domain will be presented.

2.3.1 Biopharmaceuticals: A domain analysis

In order to carry out effective and accurate terminological research, a terminologist must first learn about and grow to understand the domain in question:

> The first step in any vocabulary research project is for the terminologist to gain some general knowledge of the subject field. (Cole 1987:79)

Below is a summary of the research done in the field of biopharmaceuticals in order to prepare for this project. It describes the field of biopharmaceuticals and how this field fits into the greater structure of biotechnology.

2.3.1.1 Biotechnology and biomedicine

Biotechnology is multidisciplinary, combining aspects of biology and technology. It groups together all of the areas in which biological systems, processes or organisms (i.e., microbes, plants or animals) are used to produce goods or services (National Biotechnology Advisory Committee 1984:2). Biotechnology is not a new field. In fact, it is one of the oldest resources known to humankind. Products such as bread, cheese and wine were all developed using biotechnology: they rely on yeast, bacteria and fermentation to produce the product. Nevertheless, the modern conception of *biotechnology* is rather more focused on the technological aspect. What is currently referred to as biotechnology usually involves scientific innovation and the use of relatively newly developed techniques to create or enhance products or services. These two kinds of biotechnology are classified as *old* and *new biotechnology* (*Ibid.*).

"New" biotechnology is used for many purposes, among them to produce energy, to manage waste, to grow crops and raise livestock, to process food, and to diagnose, treat and prevent disease. It is this last purpose, associated with health care, which is of interest in this thesis; the term widely associated with it is *biomedicine*. Biomedicine includes several sub-fields, such as gene therapy, gene mapping and — the one that concerns us here — *biopharmaceuticals.*

2.3.1.2 Biopharmaceuticals

Biopharmaceuticals are pharmaceuticals that are developed, produced or activated using biological processes, systems, or organisms. These products are used to prevent, diagnose and treat disease. Penicillin and the smallpox vaccine are good examples of well-known biopharmaceutical breakthroughs.

Antibiotics and vaccines are two of the major types of biopharmaceuticals. Others include regulatory proteins (such as insulin and human growth hormone), blood products (such as human serum albumin), monoclonal antibodies and antiviral compounds (such as Interferon). In addition to the creation of pharmaceuticals themselves, the field of biopharmaceuticals encompasses complementary resources, such as drug activation systems, diagnostic tools and catalysts.

2.3.1.2.1 Types of biopharmaceuticals

The sections below will describe some of the major biopharmaceuticals, grouped by pharmaceutical category.

2.3.1.2.1.1 Vaccines

Preventative vaccines stimulate the immune system to produce antibodies that protect the body from infection by a bacterium, a parasite or a virus. Therapeutic vaccines are used to treat existing conditions such as cancer.

2.3.1.2.1.2 Antibiotics

Most people are familiar with the biopharmaceutical origins of antibiotics because of their knowledge of Fleming's discovery of penicillin, which is produced by mould. However, most antibiotics are now produced synthetically.

2.3.1.2.1.3 Regulatory proteins

Regulatory proteins such as insulin and human growth hormone are produced by the body to regulate biological processes. When insufficient quantities of these proteins are produced, illnesses occur. Currently treatment for these diseases usually involves supplementing deficient proteins with biopharmaceutically-produced products.

2.3.1.2.1.4 Monoclonal antibodies (MAbs)

These antibodies are produced from cloned cells; they are useful in various contexts such as drug targeting.

2.3.1.2.1.5 Antivirals

Antivirals constitute one of the major foci of current research in biopharmaceuticals; this is hardly surprising given the enormous effect that HIV and AIDS have on the world today. Antivirals can prevent viruses from binding to cells, and also prevent viruses, once bonded, from replicating.

2.3.1.2.1.6 Blood products

Biopharmaceutical blood products include anti-hemophilic blood factors VIII and IX, thrombolytics, which dissolve clots, and human serum albumin, which is used in transfusion and blood substitutes.

2.3.1.2.1.7 Drug activation systems

Drug activation systems are typically pairs made up of a *prodrug* (an inactive form of a drug) and a *catalyst* (often an enzyme) that transforms the prodrug into the active drug form on contact. They may be used to treat cancers, for example, because they allow the toxic effects of the drug to be limited to tumour cells.

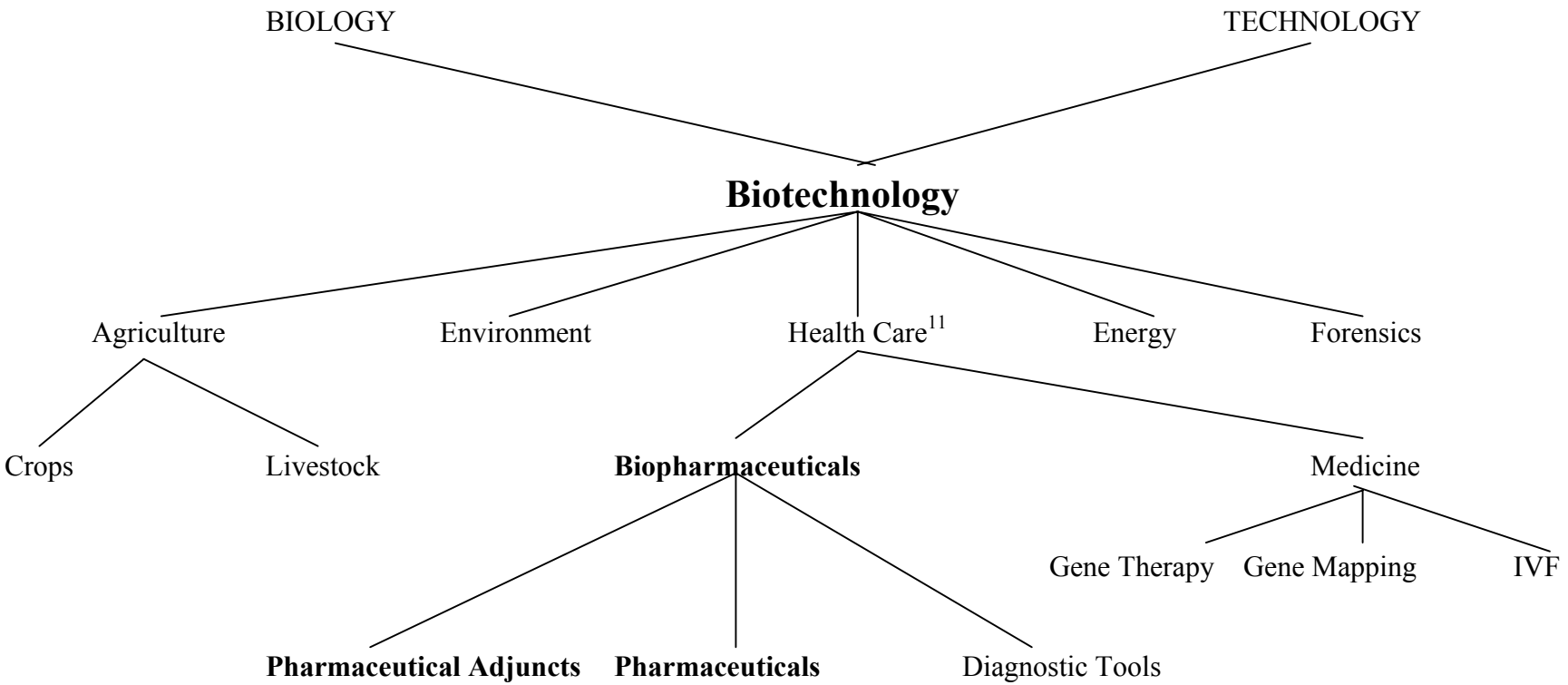
2.3.1.3 Concept Trees

Figure 4 and Figure 5 help to illustrate how the concepts in the fields of biotechnology and biopharmaceuticals are related. The list of concepts is far from exhaustive; rather it focuses on those that have been mentioned here, and provides only a few examples of each type of biopharmaceutical. The relations shown between the concepts are exclusively of hyperonymy[10]; while other relations are certainly present in the domain, they are less easily classified.

Our research focused on the sub-domains of vaccines, antivirals, regulatory proteins and activation systems. The choice of sub-domains will be discussed further in Chapter 2.

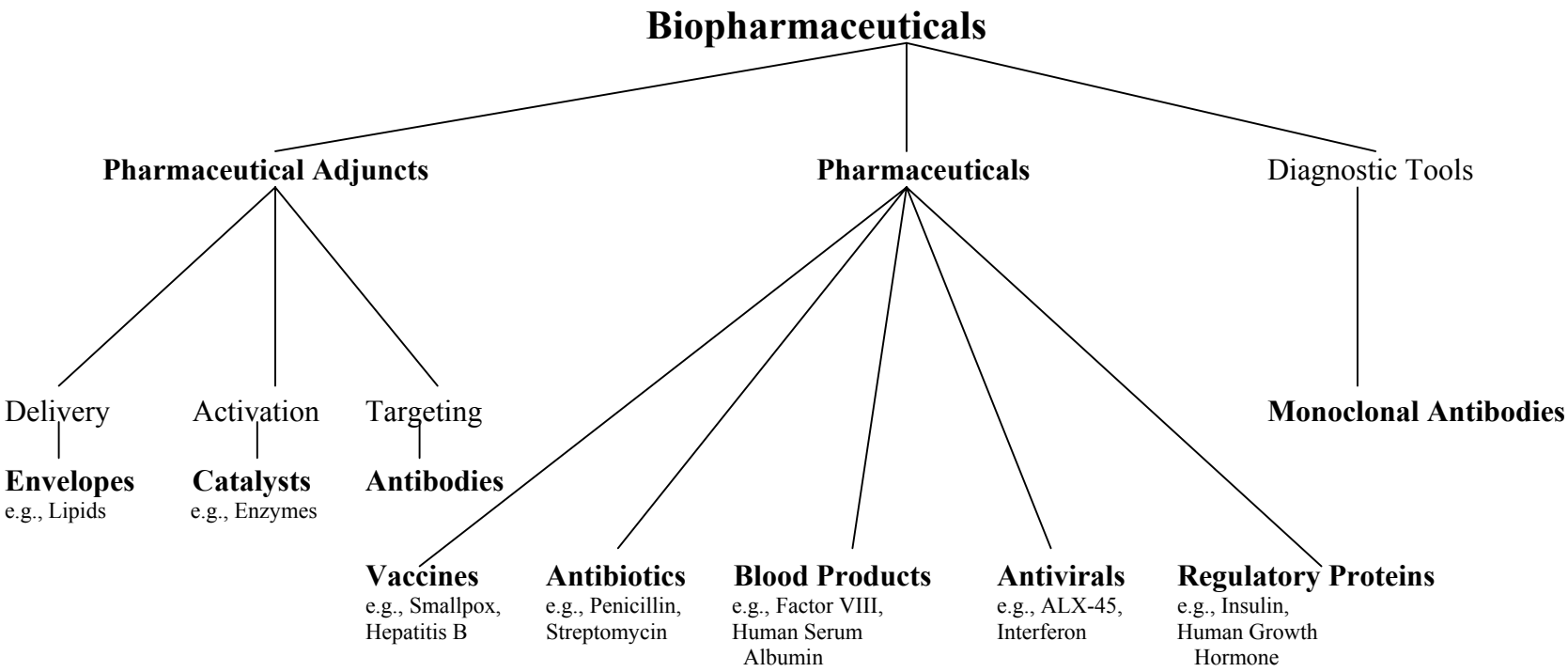
[10] An exception is the meronymic relation between a drug activation system and its parts: the prodrug and the catalyst.

Figure 4: Concept Tree — BIOTECHNOLOGY



[11] Sometimes called *biomedicine*.

Figure 5: Concept Tree — BIOPHARMACEUTICALS[12]



[12] Commonly used terms are indicated in bold. Where items (e.g., monoclonal antibodies) could be said to belong to more than one category, they are listed according to their most widely described function.

2.3.2 Current scientific research and research needs

Biopharmaceuticals are a hot topic in medicine today and will likely continue to be in the future. Many of the current advances in therapeutic and preventative pharmaceuticals are in some way based on biopharmaceutical developments. These include new antiviral drugs such as Tamiflu®, therapeutic vaccines such as those being studied for use in the treatment of cancer, and also the influenza vaccine, which is now being recognized as an important part of public health initiatives.

Given the enormous body of research carried out in this area, there is a need for a way to search the thousands of available articles efficiently in order to extract pertinent information. Researchers, physicians and patients need to obtain access to information that can be lost in a flood of similar or related data. Research projects such as this one, if successful, could contribute to developing a tool to facilitate this type of searching within documents, or even to locate documents themselves.

2.3.3 Availability of English and French texts

As discussed above, the availability of information is not at all problematic in this subject area. Rather, the opposite was true in some cases. Floods of information are available on the subject in both English and French. Given the international nature of the research and the impact biopharmaceutical developments and products have on health worldwide, information in many languages is plentiful.

The Internet was an excellent source of texts; in fact, it was the primary source for both the English- and French-language corpora. However, electronic databases and CD-ROMs were also rich sources of easily accessible information.

2.3.4 Appropriateness for cause-relation research

Although the concerns above were important in selecting the domain, the most important requirement was that texts in the field be a rich source of cause relations.

Since the focus of the field of biopharmaceuticals is the relation of cause and effect — drugs are designed and produced in certain ways in order to *cause* an *effect* when administered — the domain seemed very appropriate for the project.

2.3.5 Selection of the sub-domains

Due to the volume of information available, we had to delimit the field more closely to obtain a more structured and balanced corpus in a few coherent sub-domains.

The sub-domains chosen were activation systems, antivirals, regulatory proteins, and vaccines. However, a few general texts on the basics of biopharmaceuticals (with reference to one or more of these categories of drugs or drug activation systems) were also included.

These sub-domains were selected because they seemed to be particularly good sources of cause relations: catalysts in activation systems *cause* prodrugs to be transformed into active substances; antivirals *prevent* the bonding or reproduction of viruses; regulatory proteins *cause* processes to occur at certain speeds; and vaccines can *prevent* disease.

2.4 Building the corpora

Once the domain had been chosen several factors came into play in building the corpora, including the nature and size of the corpora and the sources and types of texts.

2.4.1 Nature of the Corpora

A terminological corpus — that is to say, a collection of specialized language texts, generally in electronic form — can vary widely in size. While some corpora can be a million words or more, in specialized domains they may be much smaller. Because of the very specialized subject matter of this research project and the density of the cause relation in the texts used, the English and French corpora used in this project were of approximately 224,000 and 225,500 words respectively. This was considered adequate (Ingrid Meyer and Lynne Bowker, personal communications).

The corpora were built within a one-year period, and were closed corpora, that is, corpora in which the contents remain fixed over time and are not updated as new texts become available and older texts become outdated. Given the time frame of the project, this kind of updating was not considered to be practical.

2.4.2 Sources of texts

As mentioned above, most of the texts used for the two corpora were taken from the Internet. However, some were also taken from electronic databases and CD-ROMs.

2.4.2.1 The Internet

In recent years, the Internet has become an invaluable resource for researchers. This research is no different; the texts available on the Web made up the majority of the corpora. As the numbers of non-English-speaking Internet users grows, more and more multilingual information is becoming available. This made locating French-language texts relatively easy, although admittedly not as easy as finding English texts.

In finding Internet-based texts, both search engines and directories proved to be extremely useful. The search engine Google[13] was very valuable for finding texts. Numerous hits were generally produced using combinations of terms from the subject field. Also useful were the Yahoo![14] directories, which classify Web sites hierarchically according to subject field.

Both Google and Yahoo! offer searching in different languages, which was critical for this bilingual project. While Yahoo offers a choice of localized sites for searching, Google offers the option of showing hits in only a selected language or languages. Both of these options were used during the corpus-building process.

2.4.2.2 Electronic databases

The University of Ottawa library offers Web-based access to several databases in electronic format. These were useful for the corpus-building project, as they offered a variety of specialized articles and collections of texts in one resource.

---

[13] www.google.com
[14] www.yahoo.com; http://fr.yahoo.com

In English, the Extended Academic ASAP abstracts and in French, the *Actualité Québec* articles from various Quebec newspapers and the magazine *L'Actualité* proved to be invaluable sources of texts.

2.4.2.3 CD-ROMs

Finally, the CD-ROMs for the magazine *Science et Vie*[15] and *Québec Science*[16] were also good sources of French-language texts. They complemented the mainly general information available in the databases of newspaper articles and the more specialized information that was available on the Internet.

2.4.3 Types of texts

In corpus-based research, the balance of the corpus is extremely important (Rundell and Stock 1992a). By *balance* we mean the types of texts included and the proportions in which they appear. The goal here was to create representative corpora by including texts from different sources and authors, in various degrees of specialization.

Sources of texts included manufacturers', associations', research groups', educational institutions' and governments' Web sites. These different types of sites were geared to different audiences and included a variety of information in a range of styles.

It was more challenging to be completely certain that the corpora contained texts by a variety of authors. There was fairly extensive reproduction of texts between Web sites, which made eliminating duplicate material in the corpora challenging. In addition, many Web sites do not credit the authors of the texts presented.

---

[15] Containing the full text of articles from 1989 to 1998.
[16] Containing the full text of articles from September 1989 to August 1997.

However, the sheer volume of the research being carried out and the diversity of the sources of the texts reflect a satisfying variety; the work of many authors is clearly included in the corpus.

Specialization was perhaps the most critical of the criteria to be considered in building the corpus. General texts, such as those found in newspapers (e.g., *The Ottawa Citizen, Le Soleil*) and in the consumer sections of biopharmaceutical manufacturers' Web sites (e.g., Lilly, Roche Canada) were included in each of the corpora. The largest proportion of texts in both corpora was composed of texts such as those found in popularized science periodicals (e.g., *Scientific American, Science et Vie*) and in the sections of biopharmaceutical manufacturers' Web sites intended for health professionals. Finally, the smallest proportion of each of the corpora was composed of the most technical texts, such as abstracts or full texts of articles in technical journals and theses (usually culled from the authors' websites or resources such as Medline[17]).

One interesting note on this subject is that technical texts were more useful for this research than they were considered to be for other projects (e.g., Morgan 2000; Marshman, Morgan and Meyer 2002). While those studying the relation of hyperonymy found that more explanatory texts were best for finding examples of the relation, the cause relation is more independent of the communicative setting (as described by Pearson 1998). While defining of terms and concepts is generally more frequent in less specialized, more didactic texts, the description of cause and effect relations is as likely to occur in a description of a ground-breaking experiment as in a popularized account of

[17] Medline citations and abstracts can be accessed through PubMed, a service of the National Library of Medicine at http://www4.ncbi.nlm.nih.gov/PubMed/.

how a drug works. Since cause and effect is often the focus of biopharmaceutical texts, it is more or less omnipresent.

Once the corpora were built, the next step was to analyze them. The next section will describe WordSmith Tools, the software used for this purpose.

2.5 The concordancer: WordSmith Tools

Created by Mike Scott (currently of the University of Liverpool), WordSmith Tools is a suite of corpus analysis tools.[18] It was chosen because of its ease of use, flexibility and features.

The first tool, Word List, provides statistical analysis of a corpus and word lists from the corpus in alphabetical order and order of frequency. This is the tool that was used to calculate the size of the corpora.

The second tool, Concord (as its name would indicate) is the concordancing feature. It permits searching by strings of characters, allowing wildcards and alternative forms, as well as context words. Once a concordance is generated, it also allows the concordance lines to be sorted and tagged, allowing great flexibility in analysis.

The third tool, Key Words — which can be used to identify unusually frequent lexical items in a given corpus— was not used for this project.

With the analysis tool and the corpora in place, the analysis could begin.

---

[18]More information and a trial version of the software are available on Mr. Scott's Web page at http://www.liv.ac.uk/~ms2928/ The versions used for this project were versions 2 and 3. Core functionality remained very similar between the two versions.

2.6 The analysis

The analysis itself was carried out in four stages: 1) choosing the terms for and generating the initial series of concordances; 2) analyzing the initial concordances for possible patterns; 3) generating the second series of concordances using the possible patterns; and 4) analyzing the second series of concordances in order to identify how useful the patterns were for extracting examples of the relation.

2.6.1 Choosing the terms and generating the initial concordances

For this project, terms used to generate the first series of concordances were chosen from each of the four sub-domains. Nine to twelve terms were chosen in each language[19]. The terms (shown below) were identified during the domain research and analysis process and all occurred many times in their respective corpora. They referred to concepts likely to be involved in cause relations.

English terms	**French terms**
activate, activation	activer
	anticorps
catalyst, catalyse	catalyser
human growth hormone, hGH	hormone de croissance humaine, hGH
immune	immuniser
insulin	insuline
	prodrogue, promédicament
regulate	régulateur
reproduce	reproduire
	Turner
vaccine	vaccin
virus	virus

[19] The number of terms chosen varied according to the data retrieved from the initial concordances. Concordances were run until a satisfactory number and variety of possible patterns was obtained.

These terms, while listed in base forms, were truncated where necessary or complemented by other variants in order to obtain the most hits within the corpus. Thus for example *immune* became *immun**, in order to find occurrences of *immune*, *immunity*, *immunize* and *immunise*, and *prodrogue* and *promédicament* were represented as *prodrogue**/*promédicament** to find either of the terms in both singular and plural.

2.6.2 Analyzing the initial concordances

This part of the project involved two steps: sorting the concordance lines into those which reflected the cause relation and those that did not, and then extracting the lexical items which seemed to be used to express the cause relation. This was all done manually, and as such was relatively subjective. The sorting relied on a clear immediate context and a good understanding of the text and how the concepts interrelated. The extraction of possible patterns was much the same. Some possibilities that were not very clear or which were judged to be likely to produce excessive noise (e.g., *by*, *through*, *par*) were discarded immediately. The others were retained for testing in the second series of concordances.

2.6.3 Generating the second series of concordances

The second series of concordances was generated using the possible patterns observed in the first series. In order to produce the maximum result with the minimum noise, we used forms of the patterns that were intuitively considered to be inclusive without being impractically general. If a pattern's form was immediately observed to be either too restrictive or too noisy, modifications were

made. (For example, the original form of the pattern *stimul** was *stimulat**, but this was found to restrict forms such as *stimulus*, and thus was modified.) The concordances generated using the best form (or occasionally forms) of the pattern were then saved for analysis of the patterns' usefulness.

2.6.4 Analyzing the patterns' usefulness

Once the second series of concordances were generated, the concordance lines in each concordance were sorted into three groups: those which clearly reflected the cause relation (hits); those which reflected the relation less clearly (maybes); and those which did not reflect the relation at all, or which were useless for the purposes of knowledge extraction (noise).

When classifying the contexts extracted in the series of concordances generated using the possible patterns, it was essential to have fixed criteria for determining whether a given context was a hit, a 'maybe,' or an example of noise. When developing these criteria, several elements were considered: the search window within which the pattern and concepts occurred, the presence of the cause and effect concepts, the precision and validity of these concepts, the presence of other cause relations, and the strength of the causal link in the text.

2.6.4.1 Concordance window

For the purposes of this research, approximately four lines of text were shown in the concordance window. The size of this section of text can be seen below, in Figure 6.

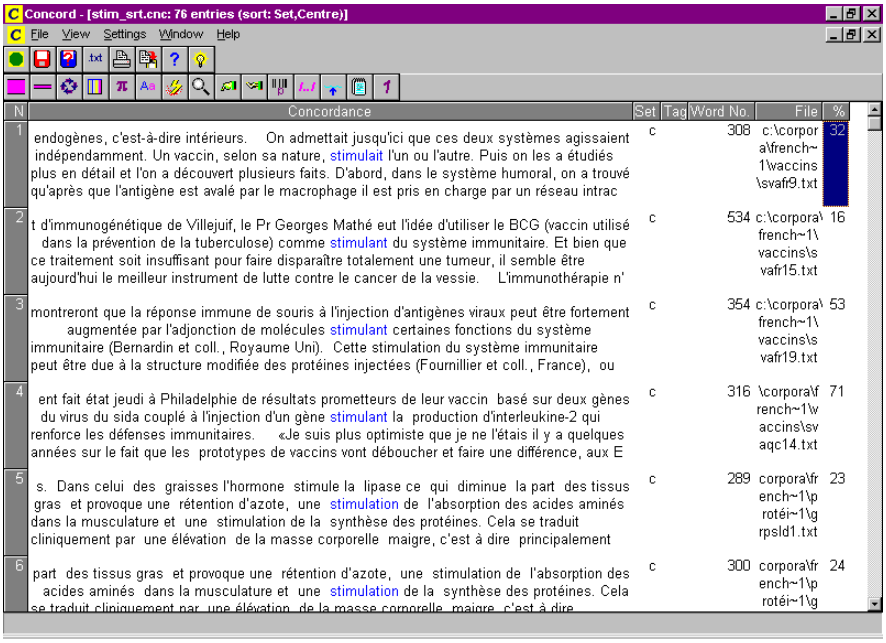


Figure 6: Size of the concordance window

This generally corresponded to one to three sentences of text, which was considered to be adequate when compared to the size of the contexts a semi-automatic tool might extract.

To be classified as a hit, the context was required to contain the pattern, the cause and the effect. If it appeared that one of the concepts was partially included in the window, expanding the window size by one or two lines was considered acceptable.

2.6.4.2 Presence of concepts acting as cause and effect

There were some problematic issues with causes or effects not being obviously present in the immediate context of the pattern[20]. While as Talmy observed (1985:52,61) in the context of theoretical research such a sentence is still a good example of the relation, the lack of a clearly stated concept would certainly affect the validity of the

[20] For a discussion of this issue, see section 6.3.2 in Chapter 6.

contexts for semi-automatic knowledge extraction, and thus had to be addressed when designing the criteria for the classification of hits and noise.

For this project, we decided that there should be a category in which less than perfect examples of the cause relation could be placed. This 'maybes' category was used for the contexts in which the cause relation was clearly present, but in which the concept acting as either the cause or the effect was either not clear, or was not present[21]. For example, here are some of the sentences classified as 'maybes' due to missing concepts[22]:

1) Abstract: "We show that recently **activated** T cells are susceptible to apoptosis when exposed to HIV gp120 in the presence of anti-gp120 antibody."
2) Thus, these mice represent a mode in which the **effects of** continuous interaction of hCD4/gp120 and anti-gp120 antibodies can be studied in the absence of HIV infection.
3) ... si vous devez être vaccine(é) avec un vaccin **atténué** ...
4) ... l'immunoprophylaxie au moyen du vaccin inactivé (virus **tué**) ...

2.6.4.3 Precision and validity of the concepts

When classifying the contexts with both the cause and effect present, it was also necessary to determine how precise those concepts should be required to be in order to be counted. For example, consider the following sentences:

5) A medicine now in testing **inhibits** the 3C protease, **inactivating** the enzyme and **stopping** production of the rhinovirus.
6) With little hope of a vaccine--because people do not **produce** lasting immunity to hepatitis C even if they fight off the initial infection-- Hoofnagle wagers that ...
7) Ce vaccin est fabriqué à partir de deux gènes du virus d'une souris atteinte par le Sida, dont on a **éliminé** les éléments pathogènes.
8) Administré à des personnes en santé, ce fragment donnait alors l'occasion au corps d'identifier le virus pour ensuite **l'éliminer** sans en être infecté.

These are clear examples of the cause relation, although the *medicine* is not named (5), it is the immune system and not *people* or *the body* as a whole which fights

[21] For a more detailed explanation, see section 2.6.4.

[22] The bold indicates knowledge patterns, single underlining a cause, and double underlining an effect.

off viruses (6, 8), and *on* is not a very specific agent for the action of eliminating pathogenic parts of a virus (7). While in isolation they might not convey much terminologically useful information about the concept, it was felt that this problem was linked to the texts more than the patterns and thus the contexts were counted as hits as long as they met the other criteria.

2.6.4.4 Multiple cause relations and chains of cause

For the purposes of this research, each occurrence of a pattern clearly accompanied by the cause and effect concepts was considered to be a hit. Thus if there was more than one cause relation within a context, each relation was counted separately.

By the term *chains of cause*, we refer to contexts in which one cause-effect relation was seen to lead to another, and so on. For example, sentences such as these:

9) C3H mice infected in the footpad with L. major resolve their lesions by **induction** of a TH1 immune response. The specific TH1 response **resulted from** the cytokine process **triggered** on antigen-*activated* CD4 targets by the release of IL-12 from natural killer cells.
10) Les ERO **altèrent** les propriétés des membranes cellulaires, dont la fluidité et la compartimentation. Cela **modifie** l'activité des récepteurs et **par conséquent** la fonction des messagers secondaires, en plus d'**entraîner** la fuite des composés intracellulaires tels la lactate déshydrogénase (LDH) et la CK.

reflect chains of cause. In sentence 9, the release of IL-12 from the natural killer cells causes a cytokine process to take place, leading in turn to a TH1 response and thus to the resolution of the lesions. Thus there are four events that take place in a sort of chain reaction of cause relations. In example 10, the chain of cause includes the altered properties of the cell membranes affecting the receptor activity and thus the transmission of cellular messages as well as the release of certain compounds. In these chains of

cause[23], only the most direct connection — and one hit —was counted for each pattern, (rather than separate hits for the immune response, the cytokine process and the release of IL-12 all causing the resolution of the lesions, nor for the alteration of the properties of cell membranes and the modification of receptor activity both leading to the alteration of the secondary messengers).

2.6.4.5 Strength of the causal link

In some contexts the causal link was stronger and clearer than in others[24]. We decided that the use of attenuating elements such as hedges, modals and negation was not a reflection on the validity of the pattern *per se*, but rather on the use of that pattern in context. If anything, the need to attenuate the pattern confirmed its power for indicating cause. Although, the value of the contexts for knowledge extraction would be decreased, the focus here is on discovering and evaluating the patterns. Thus, the contexts were counted as hits if they met the other criteria.

In many cases, domain knowledge and common sense was used to judge whether a cause relation was present. However, in cases where polysemous patterns or unclear phrasing were used and the context did not help to clarify, the concordance lines were classified in the 'maybes' category.

Once the contexts had been classified, statistics of the patterns' precision in the corpora were calculated. Chapters 3 and 4 describe the results of this research.

---

[23] For a discussion of chains of cause see section 6.2.4 in Chapter 6.
[24] See the sections of Chapter 6 on negation, hedges and modals, and lack of overt statements.

3 Results of the English Research

3.1 Introduction

Now that the process of building and analyzing the corpus has been described, some of the most productive causal patterns found in this analysis will be listed, accompanied by examples to illustrate their usage. The causal patterns will be classified using Barrière's structure, as presented in Chapter 1. There are two main groups with eight sub-groups: the existence dependency (incorporating creation, destruction, maintenance and prevention) and the influence dependency (including increase, decrease, modification and preservation). The patterns in each section will be presented in descending order of precision[25] in the corpus. The patterns that showed more than 50% precision in this corpus are included here. (For those that were less than 50% precise, see Appendix C.) In cases where the patterns were refined during the course of the research, each version of the pattern is listed separately. Statistics for each of the patterns are shown in Appendix A.

3.1.1 Presentation

In the representation of the patterns, a slash (/) represents an alternative, an asterisk (*) any character or string of characters (or neither), and the indicator */#R* or */#L* the number of words to the right or left of the main element that the accessory element of the pattern (usually a preposition) may appear. Optional elements are indicated in parentheses. Words which were excluded as possible forms of search terms including

[25] Calculated using the formula precision = hits / hits + noise (Meyer 1994).

wild cards are indicated by *not*. This option is used to prevent excessive noise from words that are similar to elements of a search term but would be unlikely to produce usable contexts. For example, the most useful forms of the pattern *confer** include *confer*, *confers*, *conferring* and *conferred*; other words which fit this pattern, such as *conference* and *conferences*, may occur often given the subject matter, but are very unlikely to locate any useful information about cause relations. In the examples provided, the pattern is indicated in bold, the cause in single underline, the effect in double underline, and the term used to generate the initial concordance — if present — in italics.

3.2 Patterns observed

The following are some of the most productive patterns observed in the corpus, presented in descending order of precision within each category.

3.2.1 Existence Dependency

3.2.1.1 Creation

*3.2.1.1.1 caus**

… HIV appears to **cause** AIDS after the onset of antiviral immunity.

In traditional vaccine manufacture, living vaccines or attenuated viruses are produced to stimulate the recipient's own *immunity* to the organism **causing** the infection.

B. anthracis is the **causative** agent of anthrax in animals and humans.

3.2.1.1.2 because

Because a single molecule of enzyme catalyzes the *activation* of many molecules of prodrug, a localized and high concentration of drug may be maintained at the tumor site.

The rapid growth in the number of TAPET organisms within solid tumors appears to occur **because** the organisms "feed" on components of DNA and proteins found in tumors -- while being protected within the tumor from the body's *immune* system.

*3.2.1.1.3 elicit**

Production of an individual cytokine by an effector cell is usually not constitutive but is rather **elicited** by various stimuli such as viral or bacterial components, protein antigens, mitogens, and other cytokines.

Animal insulin, however, differs slightly but significantly from human insulin and can **elicit** troublesome *immune* responses.

*3.2.1.1.4 ensur**

The delivery system **ensures** that enzyme production and prodrug *activation* is selectively localized within tumors…

When this protein is used as a vaccine, the foreign portion triggers helperT cells, which orchestrate an *immune* response that includes antibody-producing B cells. The other portion **ensures** that the resulting antibodies will recognise and inactivate CETP.

3.2.1.1.5 secondary to

Patients with preexisting tumors or with growth hormone deficiency **secondary to** an intracranial lesion should be examined routinely for progression or recurrence of the underlying disease process.

GH stimulates skeletal growth in pediatric patients with growth failure due to a lack of adequate secretion of endogenous GH or **secondary to** chronic renal insufficiency and in patients with Turner syndrome.

3.2.1.1.6 aris from*

However, potential massive market could **arise from** the increasing evidence that this growth hormone can increase muscle formation in normal individuals and now being exploited by many body freaks from WWF wrestlers to Salman Khan.

3.2.1.1.7 provok/provoc**

The accidents **provoked** by *vaccines* have today only an historical interest.

The plasmodium parasite was deliberately inoculated to the patients in order to **provoke** a high fever access capable of destroying the treponema of the syphilis.

*3.2.1.1.8 spur**

Attenuated live vaccines, usually viruses, do enter cells and make antigens that are displayed by the inoculated cells. They thus **spur** attack by killer T lymphocytes as well as by antibodies.

3.2.1.1.9 adverse drug event/adverse drug reaction**

Other **adverse drug events** that have been reported in growth hormone-treated patients include the following: 1) Metabolic: Infrequent, mild and transient peripheral or generalized edema. 2) Musculoskeletal: Rare carpal tunnel syndrome. …

3.2.1.1.10 culminat in*

Recognition of the MHC-peptide complex by the T cell receptor triggers a cascade of events that **culminate in** T cell response (secretion of cytokines or direct attack on the APC).

... the typical course of the disease: decades of slow liver damage often **culminating in** organ failure or cancer.

*3.2.1.1.11 incit**

The safety trials ask such questions as, are the plasmids toxic, and does DNA delivered as a drug **incite** an immune response against the body's own DNA?

3.2.1.1.12 prod/prods/prodding

In both cases, the *vaccines* **prod** the immune system to make antibodies able to bind to, and help eliminate, infectious viruses bearing those same proteins.

3.2.1.1.13 give/gives/gave/giving rise to

Those based on killed pathogens (such as the hepatitis A and the injected, or Salk, polio *vaccines*) or on antigens isolated from disease-causing agents (such as the hepatitis B subunit *vaccine*) cannot make their way into cells. They therefore **give rise to** primarily humoral responses and do not activate killer T cells.

3.2.1.1.14 due /5 *to*

For example, suspended protein powders can show *catalytic* activity for days in neat organic solvents at temperatures far above 100oC [34,35]. This is **due to** the fact that dehydrated proteins are "rigid" in organic solvents. Thus, even though they can be denatured by the solvents for thermodynamic reasons, **due to** this increased rigidity they are trapped…

The FDA has approved recombinant EPO—epoetin alfa—for the treatment of anemia **due to** chronic renal failure. … *Human growth hormone* (*hGH*) is used to counter growth failure in children that is **due to** a lack of *hGH* production by the body.

3.2.1.1.15 lead/led* /5R *to*

Beyond eliminating invaders, *activation* of the immune system against a specific pathogen **leads to** the creation of memory cells that can repel the same pathogens in the future.

Acute overdosage could **lead** initially **to** hypoglycemia and subsequently **to** hyperglycemia.

*3.2.1.1.16 attrib**

She **attributes** the *vaccine's* success to its ability to turn on the expression of the three major HIV proteins.

3.2.1.1.17 so that

The cells in such vaccines are irradiated **so that** they are not capable of further growth.

Recombinant DNA technology enables modifying microorganisms, animals, and plants **so that** they yield medically useful substances, particularly scarce human proteins...

*3.2.1.1.18 confer** **not** *conference/conferences*

Nevertheless, in a combined passive and active *vaccination*, the initial administration of specific Abs should **confer** an immediate helpful effect.

Vaccines **confer** protection by similarly inducing immune responses and the consequent formation of memory cells.

*3.2.1.1.19 induc**

In the in vitro system reported here, as little as 10 ng/ml (10-9 M) gp120 primed CD4+ T cells for *activation*-**induced** apoptosis ...

Vaccines confer protection by similarly **inducing** immune responses and the consequent formation of memory cells.

*3.2.1.1.20 *genic/*genesis/*genicity* **not** *transgenic, antigenic*

Enzymes of nonhuman origin that meet these needs are, however, likely to be highly **immunogenic**, a fact that makes repeated administration impossible.[26]

From this table, it can be noted that in all situations, CJD was transmitted with a low frequency, a common circumstance for the **iatrogenic** transmission of the disease.

*3.2.1.1.21 complication**

The spreading virus can overwhelm a person whose *immune* system does not respond strongly or quickly enough, leading to **complications** such as pneumonia.

3.2.1.1.22 consequence/consequent**

---

[26] Note that the pattern in this case is also part of the effect. Cases such as this one will be described in Chapter 6, section 6.5.1.2.

Also, apoptosis of immune T cells occurs as a **consequence** of the abnormal production of TNF-a or Fas ligand (57) from malignant tumors, which may diminish further the *immune* control of the tumor cells.

The rate of *insulin* absorption and **consequently** the onset of activity is known to be affected by the site of injection, exercise, and other variables.

*3.2.1.1.23 evok**

Plasmodium antigens have **evoked** significant cellular *immune* responses…

3.2.1.1.24 prime/primes /priming/primed

As is true of most genetic vaccines under study, standard types aim to **prime** the *immune* system to quash dangerous viruses, bacteria or parasites quickly, before the pathogens can gain a foothold in the body.

3.2.1.1.25 therefore

Untreated hypothyroidism prevents optimal response to Protropin. **Therefore**, patients should have periodic thyroid function tests and should be treated with thyroid hormone when indicated.

Humalog has a more rapid onset and a shorter duration of action than human regular *insulin*. **Therefore**, in patients with type 1 diabetes, Humalog should be used in regimens that include a longer-acting *insulin*.

3.2.1.1.26 secondary /5R *to*

See 3.2.1.1.5 *secondary to* above.

*3.2.1.1.27 prompt** not *promptly*

The scientists hope this second dose of DNA will **prompt** the *immune* system to produce large numbers of killer T cells to target cells invaded by HIV.

Yet simply putting DNA into skin or muscle cells and **prompting** those cells to display fragments of the encoded antigens should not have produced that outcome.

*3.2.1.1.28 reason** not *reasonable/reasonably/reasoned*

Chronic hepatitis C is the leading cause of chronic liver disease and the most common **reason for** liver transplant, according to WHO.[27]

The powerful immunogenicity of the MAP *vaccine* requires the timing of booster injections be carefully monitored for safety **reasons**.

*3.2.1.1.29 trigger**

---

[27] The same text continues, "Liver damage caused by HCV is the leading cause of liver transplantation in adults."

In Ab-treated C3H mice, the IL-12-**triggered** process was blocked as long as massive anti-IL-12 Abs of high affinity (as a result of repeated administration) were loading the lymphatic fluid…

The key idea is to use the mRNA or DNA specific to the disease state to **trigger** the *catalytic* release of a cytotoxic drug by promoting the association of a prodrug with a *catalyst* capable of releasing the drug.

3.2.1.1.30 thus

More-over, many biotech agents are identical to, or differ only slightly from, proteins the human body produces naturally; **thus**, biotech pharmaceuticals tend to have a lower potential for adverse reactions.

Muromonab-CD3 restrains *immune* response and **thus** increases the likelihood that the transplant will function.

3.2.1.1.31 account /5R *for*

Although the elderly tend to become sickest when they have the flu, children **account for** much of its spread.

The cell migration in part **accounts for** the overall interdependency of cytokines.

3.2.1.1.32 why

Why was this correlation established so late? This is a long story, but personally I think that the medical community was faced with a new problem …

*3.2.1.1.33 turn** /5R *on*

Cytokines may be produced continuously when regulatory feedback circuits are not functioning in E cells, as may occur in cancer- or virus-infected cells, when genetic changes vicariously **turn on** cytokine-promoter genes (Fig. 2).

*3.2.1.1.34 result** /5R *in/of/from*

The delivery system ensures that enzyme production and prodrug *activation* is selectively localized within tumors, **resulting in** local killing of tumor cells, whilst sparing other normal tissues from toxic effects.

The measurable increase in body length after administration of either Humatrope or *human growth hormone* of pituitary origin **results from** an effect on the growth plates of long bones.

*3.2.1.1.35 immuniz**

Similarly, ***immunization*** with PA plus LFn fused to a lymphocytic choriomeningitis virus (LCMV) nucleoprotein epitope protects against lethal LCMV infection.

To determine the functional relevance of the immunity generated with the targeted *vaccine*, BALB/c mice were **immunized** with 100 mg of DNA and challenged 9 weeks later with infectious virus.

*3.2.1.1.36 produc** **not** *product/products*

Immunotherapy (vaccine) Research is being conducted which follows both the classic and more recent approaches to vaccines, whereby PSMA is administered to **produce** or increase *immunity* to prostate cancer.

Immunization by this method **produced** M-specific antibodies and cytotoxic T lymphocyte (CTL) response, and acquired resistance against influenza virus challenge.

*3.2.1.1.37 effect** **/5R** *of*

By analysis of covariance * , the **effect** of growth hormone therapy was a mean height increase of 5.4 cm (p=0.001).

*3.2.1.1.38 *stimul**

IL-6 is involved in **stimulating** the production of acute phase proteins in the liver. … It is also involved in the **stimulation** of T cell *activation* and growth, and also aiding the proliferation of haemopoietic progenitor stem cells.

Production of an individual cytokine by an effector cell is usually not constitutive but is rather elicited by various **stimuli** such as viral or bacterial components, protein antigens, mitogens, and other cytokines.

… interleukins, which themselves have profound ***immunostimulatory*** effects throughout the immune system…

3.2.1.1.39 respond/respons** **/5R** *to*

The retention of total body potassium in **response to** growth hormone administration apparently results from cellular growth.

The side-effect profile of such vaccines should be excellent, with only the hallmarks of *immune* **responses** such as fever and redness at injection sites.

*3.2.1.1.40 effect**

… interleukins... themselves have profound *immunostimulatory* **effects** throughout the immune system …

In general IFN beta side **effects** are similar to those of IFN alpha, i.e. flu-like symptoms, fever, fatigue etc.

*3.2.1.1.41 aris**

Epivir-resistant mutants **arise** in at least 14% of those treated with the drug every year.

*3.2.1.1.42 forc**

The situation **forced** health care providers to give *immunization* priority to the elderly, children, and individuals with compromised *immune* systems.

3.2.1.2 Destruction

*3.2.1.2.1 lethal**

Similarly, *immunization* with PA plus LFn fused to a lymphocytic choriomeningitis virus (LCMV) nucleoprotein epitope protects against **lethal** LCMV infection.

Throughout the world there are millions of people who need regular intakes of *insulin* to overcome **lethal** effects of diabetes.

*3.2.1.2.2 deadl**

Staph can be **deadly** if it enters the bloodstream, causing pneumonia, encephalitis, liver abscesses and other severe problems.

The Ebola *virus*, first identified in 1976, causes Ebola hemorrhagic fever, one of the **deadliest** viral diseases known.

3.2.1.2.3 destroy/destruct**

While **destruction** of chronically infected cells may be a beneficial result of ADCC in infected patients, such antibodies could also direct the lysis of noninfected CD4-positive cells if they adsorbed gp120 in vivo and became susceptible to *immune* attack as demonstrated here.

DO NOT FREEZE since freezing **destroys** potency.

*3.2.1.2.4 fatal**

Currently, there is no approved vaccine for prevention or control of hepatitis C, a potentially **fatal** liver disease caused by the hepatitis C virus.

*3.2.1.2.5 kill**

Imagine a day when a cancer patient can have a blood or biopsy sample fed into a DNA diagnostics machine, which takes the disease-state DNA results and within hours comes up with a tailored drug/*catalyst* therapy. This treatment will **kill** the cancerous cells in the body and leave the others unharmed.

Globally, the B *virus* is by far the most common cause of liver disease, infecting about 350 million people and **killing** more than a million a year.

3.2.1.3 Maintenance

*3.2.1.3.1 enabl**

This uniquely selective metabolic *activation* **enables** significant reduction of the drug's effective doses with a concomitant reduction of toxicity and improved therapeutic index.

Growth hormone treatment, begun early enough, can **enable** the child to reach normal adult height.

*3.2.1.3.2 allow**

A gene encoding a humanized 38C2 should also find application in gene therapy strategies, **allowing** for the selective ablation of cells expressing it.

We strategize that in all these instances, additional anti-cytokine vaccination may **allow** a conventional vaccine to mount an adequate *immune* reaction against the antigenic aggressor.

3.2.1.3.3 essential for

Children with GHI do not produce enough of their own growth hormone, a protein **essential for** growth.

3.2.1.3.4 critical to/for

Including a mixture of CTL and Th epitopes may be **critical to** protection.

While other HIV-1 immunotherapies are being tested, epitope-based *vaccines* are a particularly promising approach because they are specifically designed to contend with the mutagenic nature of HIV and also to augment the cellular immune responses known to be **critical for** controlling this virus.

3.2.1.4 Prevention

*3.2.1.4.1 prophyla**

Epimmune is also pursuing development of therapeutic vaccines for hepatitis C, hepatitis B and HIV, and **prophylactic** vaccines for hepatitis C, HIV and malaria.

PMPA has shown unprecedented antiviral activity in the **prophylaxis** and treatment of simian *immunodeficiency* virus (SIV) infection, a primate model for AIDS.

*3.2.1.4.2 spar**

Many others were dubious at first, skeptical, for instance, that the immunity elicited would be strong enough to **spare** people **from** infection by a living pathogen.

Vaccines arguably constitute the greatest achievement of modern medicine. They have eradicated smallpox, pushed polio to the brink of extinction and **spared** countless people **from** typhus, tetanus, measles, hepatitis A, hepatitis B, rotavirus and other dangerous infections.

3.2.1.4.3 stop /10R from*

Immunity is achieved when such activity generates long-lasting "memory" cells--the sentries that stand ready to **stop** the pathogen **from** causing disease.

*3.2.1.4.4 preclud**

Unlike the derivation of insulin from animals, biotech production of the hormone virtually **precludes** contamination with other hormones.

Six had positive ITH-STs at day 83, **precluding** scheduled booster *immunization*.

*3.2.1.4.5 prevent**

More recently, the FDA approved the *immunosuppressant* daclizumab (Zenapax) for the **prevention** of kidney-transplant rejection.

Thus, HPV-associated cervical malignancies might be **prevented** or treated by induction of the appropriate virus-specific *immune* responses in patients.

*3.2.1.4.6 block**[28]

… *vaccine* can **block** the transformation of "good" cholesterol into "bad" cholesterol and so reduce fatty deposits on artery walls in rabbits, say researchers at a biotechnology company in Massachusetts.

*3.2.1.4.7 keep**

MPIF-1 can **keep** certain normal cells, including many *immunologically* important cells, from dividing and can thus protect them from anticancer drugs that target rapidly multiplying cells.

*3.2.1.4.8 stop**

A medicine now in testing inhibits the 3C protease, inactivating the *enzyme* and **stopping** production of the rhinovirus.

The *vaccine*, expected to be available for limited use by the fall, is expected to **stop** the spread of E. coli 0157:H7 bacteria …

3.2.2 Influence Dependency

3.2.2.1 Modification

*3.2.2.1.1 influenc**

Growth hormone of human pituitary origin **influences** the size of internal organs, including kidneys, and increases red cell mass.

... altering the cell membrane (eg, increasing cell fluidity) ... may **influence** virion maturation and release…

---

[28] Indicates preventing an event from occurring or entity from existing and preventing access to something.

3.2.2.1.2 marked by

Another MoAb, infliximab (cA2), appears effective against Crohn's disease, an *immune*-system disorder **marked by** intestinal inflammation.

*3.2.2.1.3 affect**

Interferon has profound *immunomodulatory* actions, some of which... would be expected to **affect** virus multiplication and spread.

It is not known whether Humatrope can cause fetal harm when administered to a pregnant woman or can **affect** *reproductive* capacity.

3.2.2.1.4 associated /5R *with*

Treatment with recombinant growth hormone eliminates the risk of disease **associated with** products derived from human tissue.

Nutropin growth hormone is approved for treating growth hormone deficiency (GHD), growth failure **associated with** chronic renal insufficiency (CRI) prior to kidney transplantation and short stature **associated with** Turner syndrome.

3.2.2.1.5 relationship /5R *between*

In adults, it is unknown whether there is any **relationship between** growth hormone replacement therapy and CNS tumor recurrence.

The **relationship**, if any, **between** leukemia and growth hormone therapy is uncertain.

3.2.2.1.6 seen /5R *with/in*

Treatment of children who lack adequate endogenous growth hormone secretion with Protropin resulted in an increase in growth rate and an increase in insulin-like growth factor-I levels similar to that **seen with** pituitary-derived *human growth hormone*.

Skeletal abnormalities including scoliosis are commonly **seen in** untreated *Turner* syndrome patients.

*3.2.2.1.7 role**

Concentrations of IGF-I, which may play a **role** in skeletal growth, are low in the serum of growth hormone-deficient pediatric patients.

Can HCiG, *immunoglobulin* (Ig) pool hyper specific for HCV, have a **role** in the control of HCV?

*3.2.2.1.8 alter** **not** *alternate/alternating/alternative*

Serum calcium is not significantly **altered** in patients treated with either *human growth hormone* of pituitary origin or Humatrope.

3.2.2.1.9 involv /5R in*

IL-6 is **involved in** stimulating the production of acute phase proteins in the liver.

Il-8 is a chemotactic factor that attracts neutrophils, basophils, and T cells, but not monocytes. It is also **involved in** neutrophil *activation*.

*3.2.2.1.10 neutraliz**

Thus, in developing offspring of anti-NGF-*immunized* female rats, maternal Abs of high affinity could **neutralize** NGFs acting on immature neurons (63) and hamper nerve tissue development.

Because HIV-infected people show some ability to **neutralize** the *virus* while non-functioning vaccines did not, Nunberg postulated that something about the process of active *virus*-cell binding and fusion is crucial for the body to mount a defense.

*3.2.2.1.11 control** **not** *controlled/controls*

Currently, there is no approved vaccine for prevention or **control** of hepatitis C, a potentially fatal liver disease caused by the hepatitis C virus.

When regular *insulin* is used to **control** postprandial blood glucose, adequate **control** is often not achieved because the amount of regular *insulin* needed to normalize postprandial glucose excursion often leads to late hypoglycemia.

3.2.2.2 Increase

*3.2.2.2.1 facilitat**

Linear growth is **facilitated** in part by increased cellular protein synthesis.

3.2.2.2.2 optimiz/optimis** **not** *optimistic*

Eventually, though, as vaccine makers learn how to **optimize** responses to genetic *immunization* (such as through the techniques described above), manufacturers may be able to achieve the needed effects by constructing genetic vaccines alone.

3.2.2.2.3 amplify/amplification*

Hence, increasing the number of *immunostimulatory* sequences in plasmids might well **amplify** the *immunogenicity* of the antigenic codes in a DNA vaccine.

The VEE vector can copy its own DNA once it enters a host cell, **amplifying** the amount of HIV protein produced and maximising the *immune* response.

*3.2.2.2.4 expand** **(transitive verb)**

Interleukin 14 can induce B-cell proliferation, inhibit *immunoglobulin* secretion and selectively **expand** certain B-cell populations.

*3.2.2.2.5 prolong** **not** *prolonged*

And the IgG mixer helps to **prolong** the powers of interleukin-2.

It's the change in solubility that **prolongs** absorption and extends its tail of action.

3.2.2.2.6 speed/sped* **(verb)**

Cangene plans to follow with Phase II/III clinical trials in two indications: one in children with abnormally small stature, and one looking at the potential use of this product to **speed** overall recovery in elderly patients who have sustained bone fractures.

Nevertheless, *immunologists* are honing ways to **speed** vaccine production, so that *immunization* can be carried out swiftly if a virulent epidemic starts abruptly.

*3.2.2.2.7 strengthen**

Separate work suggests that the plasmid DNA surrounding antigenic genes is more than a mere gene-delivery vehicle; it **strengthens** the *immune* response evoked by the antigens.

3.2.2.2.8 maximiz/maximis**

The VEE vector can copy its own DNA once it enters a host cell, amplifying the amount of HIV protein produced and **maximising** the *immune* response.

*3.2.2.2.9 augment**

Moreover, the presence of infectious material in the SNC of patients apparently healthy **augmented** the number of infected glands that were, probably, introduced into the pool used for manufacturing *hGH*.

While other HIV-1 immunotherapies are being tested, epitope-based vaccines are a particularly promising approach because they are specifically designed to contend with the mutagenic nature of HIV and also to **augment** the cellular *immune* responses known to be critical for controlling this virus.

*3.2.2.2.10 lengthen**

Adding polyethylene glycol (PEG) to therapeutic proteins increases their stability in the body and **lengthens** the time they stay in the bloodstream, thus decreasing the number of injections needed.

*3.2.2.2.11 alert** **(verb)**

Epimmune *vaccines* under development are composed of 'epitopes', protein fragments that act like chemical 'flags' to **alert** and activate the immune system against foreign invaders or diseased cells.

*3.2.2.2.12 bolster**

So far human tests are examining vaccines designed to prevent various infections ... to **bolster** the impaired *immunity* of patients already infected with HIV and to treat a number of cancers...

Our hope is that through a combined strategy of HAART and vaccine therapy, the *immune* system can be **bolstered** to permanently reduce viral replication and potentially eliminate infected cells…

*3.2.2.2.13 potentiat**

Antibody 38C2 demonstrated long-lived catalytic activity in vivo and was shown to selectively *activate* prodrugs and **potentiate** killing of colon and prostate cancer cell lines when applied at therapeutically relevant concentrations in culture.

*3.2.2.2.14 orchestrat**

When this protein is used as a vaccine, the foreign portion triggers helperT cells, which **orchestrate** an *immune* response that includes antibody-producing B cells.

3.2.2.2.15 favour/favor** **not** *favourite/favorite/favourably/favorably*

Interleukin 12 is produced mainly by monocytes and macrophages. It acts in a opposing manner to IL-10 as it **favours** TH1 responses and stimulates interferon g production.

The resulting induction of undesirable cytokines by these stromal cells may locally impair the cytokine network in the tumor and **favor** the establishment of *immune* suppression and neoangiogenesis, features characterizing the cancer cell's microenvironment.

*3.2.2.2.16 contribut**

This work suggests that deletion of CD4+ T cells upon *activation* may **contribute to** the progressive depletion of CD4+ T cells in AIDS.

We speculated that the *immune* system may participate in the pathogenesis of AIDS and play a **contributory** role in its own destruction.

*3.2.2.2.17 increas** **not** *increasingly*

An optimized combination of prodrugs that can be activated by antibody 38C2 may **increase** the sensitivity further.

As expected, the addition of the polar linker to campothecin **increased** its solubility in water.

*3.2.2.2.18 increas**

See 3.2.2.2.17 above.

*3.2.2.2.19 accelerat**

A recombinant version of one of the enzymes that **accelerate** this conversion can contribute to the treatment of heart attacks, strokes, and pulmonary emboli.

Onyvax is developing 105AD7 which it is believed will elicit an *immune* response against DAF (Decay **Accelerating** Factor, CD55) a molecule over-expressed in colorectal, ovarian and pancreatic cancers.

*3.2.2.2.20 enhanc**

The rapid growth in the number of TAPET organisms within solid tumors appears to occur because the organisms "feed" on components of DNA and proteins found in tumors -- while being protected within the tumor from the body's *immune* system. This appears to **enhance** their ability to inhibit tumor growth and allow the continuous delivery of anticancer drugs within the tumor cells themselves.

… a nucleic acid vaccine using a novel non-viral promoter induces a vigorous *immune* response in mice to the Herpes Simplex Virus II (HSV-2). The response is further **enhanced** by the inclusion of a Vitamin D3 adjuvant.

*3.2.2.2.21 help** **not** *helper*

Growth hormone can **help** patients with these conditions potentially grow at a normal rate.

Although cancer is not an infectious disease, much evidence indicates that harnessing the body's *immune* defenses may **help** combat it.

*3.2.2.2.22 activat**

Antibody 38C2 ***activated*** procamptothecin in a concentration-dependent manner, revealing the full toxic effect of the drug.

CA4P is a prodrug that is ***activated*** in the body by a naturally occurring phosphatase enzyme.

*3.2.2.2.23 expand**

Interleukin 14 can induce B-cell proliferation, inhibit *immunoglobulin* secretion and selectively **expand** certain B-cell populations.

*3.2.2.2.24 promot**

Excessive glucocorticoid therapy will inhibit the growth **promoting** effect of *human growth hormone*.

… Emisphere's compounds have been shown to **promote** significant absorption of recombinant *human growth hormone* (r*hGH*) from the gastrointestinal tract.

3.2.2.2.25 speed/sped*

See 3.2.2.2.6 above.

*3.2.2.2.26 encourag** **not** *encouraging/encouragingly*

They are also working on injection-free vaccines, to improve acceptance and to **encourage** *immunization* of children.

*3.2.2.2.27 improv**

Recombinant *hGH* has greatly **improved** the long-term treatment of children whose bodies do not produce enough *hGH*.

The greatest **improvement** in adult height was observed in patients who received early growth hormone treatment and estrogen after age 14 years.

*3.2.2.2.28 prolong**

See 3.2.2.2.5 above.

*3.2.2.2.29 cataly**

Many endogenous enzymes **catalyze** these types of reactions.

We demonstrate that generic drug-masking groups may be selectively removed by sequential retro-aldolretro-Michael reactions **catalyzed** by antibody 38C2. This reaction cascade is not **catalyzed** by any known natural enzyme.

*3.2.2.2.30 help**

See 3.2.2.2.21 above.

3.2.2.3 Decrease

*3.2.2.3.1 protect** **/5R** *against*

In 1996, an NIAID-funded efficacy trial found the *vaccine* to be 85 percent **protective** against influenza when tested in 92 healthy adults directly exposed to the influenza virus.

3.2.2.3.2 minimiz/minimis**

In the future, such "rationally" designed genetic vaccines are likely to provide new *immune* therapies for cancer and powerful ways to prevent or **minimize** any number of devilish infections that elude human control today.

*3.2.2.3.3 dimin**

Also, apoptosis of immune T cells occurs as a consequence of the abnormal production of TNF-a or Fas ligand (57) from malignant tumors, which may **diminish** further the *immune* control of the tumor cells.

Now it appears stress can even **diminish** the efficacy of *vaccinations*.

*3.2.2.3.4 hamper**

Although the limited clinical studies of these compounds have indicated considerable potential in cancer therapy, the restricted supplies have severely **hampered** conclusive experimentation.

Discovery of conserved, *immunologically* relevant cross-clade regions of the HIV-1 genome has been **hampered** by a lack of powerful tools that would enable researchers to mine existing large HIV-1 sequence databases for vaccine components.

*3.2.2.3.5 lessen**

Non-narcotic analgesics and bedtime administration of Infergen may be used to prevent or **lessen** some of these symptoms.

Also, in late November the FDA reported that during an outbreak, daily doses of Tamiflu, a Hoffman–La Roche pill prescribed to **lessen** flu symptoms in adults, could prevent an individual from getting the illness almost as well as a vaccine.

*3.2.2.3.6 restrain**

Muromonab-CD3 **restrains** *immune* response and thus increases the likelihood that the transplant will function.

*3.2.2.3.7 alleviat**

Growth hormones have also been shown to **alleviate** wasting …

*3.2.2.3.8 endanger**

If left untreated diabetes mellitus can damage the kidneys, eyes, heart, and limbs, and can **endanger** pregnancy.

For all intramuscular injections, the needle should be long enough to reach the muscle mass and prevent *vaccine* from seeping into subcutaneous tissue, but not so long as to **endanger** underlying neurovascular structures or bone.

*3.2.2.3.9 curb**

Novel and powerful *vaccines* hold some promise in therapeutic efforts to **curb** existing and emerging disease threats such as cancer, autoimmune diseases and infectious diseases.

*3.2.2.3.10 dampen**

Concerns about cross-clade efficacy initially **dampened** enthusiasm for *vaccine* trials of these *vaccines* in developing countries.

3.2.2.3.11 slow /5R down*

Insulin glargine differs from human *insulin* in three amino acids, **slowing down** its release and reducing its solubility in the blood.

3.2.2.3.12 *slow** (verb)

Adding PEG to recombinant ADA enables effective weekly infusions, as PEG **slows** the breakdown of ADA in the body.

*3.2.2.3.13 *suppress**

This drug is primarily targeted at the treatment of rheumatoid arthritis and inflammatory bowel diseases, and Centecor believes it may also prove useful as an ***immuno*suppressant** in transplant patients.

The FDA has approved a recombinant variant of G-CSF, filgrastim, for controlling infections in patients on anticancer drugs that **suppress** *immune* responses….

*3.2.2.3.14 inhibit**

In accordance with its mechanism, *catalysis* was completely **inhibited** by 2,4-pentanedione …

Potential in vivo problems like rapid clearance or **inhibition** of the *catalytic* activity of 38C2 by covalently binding diketones or other potential inhibitors were not observed.

*3.2.2.3.15 impair**

Long-term animal studies for carcinogenicity and **impairment** of fertility with this *human growth hormone* (Humatrope) have not been performed.

The resulting induction of undesirable cytokines by these stromal cells may locally **impair** the cytokine network in the tumor ...

3.2.2.4 Preservation

See Appendix C.

3.3 Conclusions

As indicated above, a large number of possible patterns indicating a cause relation have been observed. Many of them appear to be very reliable in this context, with precision levels of between 90% and 100%. This makes these patterns very good candidates for further research in the area of semi-automatic text analysis.

However, there is a striking lack of precise patterns for the influence dependency of preservation. More research is needed to determine if this phenomenon is linked to this particular research project or corpus, or if it is indicative of the field or the sub-relation itself.

The patterns that were observed in this corpus are also interesting because many of them appear to be domain-linked. For example, many of the patterns identified for destruction deal with death (e.g., *lethal*, *fatal*, *deadly*, *kill*). Others seem likely to be particular to medical fields (e.g., *prophylaxis*, *adverse drug reaction*). It would be interesting to test these patterns on corpora in non-medical fields, to determine if this is in fact the case, and if so how close the domain link is.

In carrying out this research, we were also struck by how often the cause relation was indicated by polysemous grammatical words such as prepositions (e.g., *after*, *by*). This phenomenon creates a challenge for semi-automatic knowledge extraction, since these kinds of patterns are likely to produce large amounts of noise. (Because of this noise, none of these kinds of patterns met the threshold of 50% precision, and so do not appear in this chapter.) In addition, it may not always be clear whether these polysemous elements really indicate a cause relation. Given the importance of precision in many scientific texts, it seemed unusual for this kind of expression to be so prevalent. Closer study of the contexts in which these kinds of patterns are used could provide interesting insights into when this kind of "vagueness" is permitted in scientific texts. These issues will be discussed further in Chapter 6, sections 6.2.5 and 6.3.3.

In the next chapter, parallel research in the French corpus will be described.

4 Results of the French Research

4.1 Introduction

To complement the patterns described in Chapter 3, parallel research was done on the French corpus. The results are presented below. The format and conventions of the presentation are the same as those in the previous chapter, and statistics for each pattern are provided in Appendix B. Again, only those patterns that showed 50% or higher precision in this corpus are shown. For the less precise patterns, see Appendix D.

4.2 Existence Dependency

4.2.1 Creation

4.2.1.1 *provoc*/provoq**

Chez des patients avec maladie hépatique sévère et cirrhose décompensée, l'interféron est contre-indiqué car il peut **provoquer** une insuffisance hépatique fatale.

De plus, ce traitement **provoque** des effets secondaires souvent très gênants.

4.2.1.2 *parce qu**

Les diabétiques ne peuvent assimiler correctement le sucre (glucides) **parce que** leur pancréas ne sécrète pas correctement une hormone qui active l'utilisation du glucose dans l'organisme.

Parce qu'une déficience des récepteurs LDL cause une réduction du retrait et une production accrue de LDL, de petites diminutions dans le nombre ou dans l'activité des récepteurs LDL entraînent une augmentation disproportionnée des niveaux sanguins de LDL.

4.2.1.3 *grâce à/au*

Au cours du premier semestre de 2001, nous avons connu plusieurs succès importants au niveau de la découverte de médicaments, de nos programmes cliniques et de nos activités stratégiques, **grâce à** l'acquisition d'Aurora Biosciences.

Mais c'est **grâce au** génie génétique que la production de *HGH* pourra passer de la pénurie à l'abondance.

4.2.1.4 *abouti* à/au*

L'oxygène moléculaire doit être activé pour manifester sa toxicité soit par une activation photodynamique qui **aboutit à** l'oxygène singulet ($^1O_2$), soit par une activation réductrice avec formation séquentielle de l'anion superoxyde ($O_2^{\cdot}$), du H_2O_2 et du radical hydroxyle (OH)...

... les études épidémiologiques montrent que seule une stratégie de *vaccination* universelle portant à la fois sur les adolescents et sur les nourrissons, associée à une *vaccination* des adultes à risque, est susceptible d'**aboutir à** une réduction du nombre de cas d'hépatite B dans nos pays.

4.2.1.5 *suscit**

En effet, depuis maintenant une vingtaine d'années, *l'immunothérapie* active **suscite** un intérêt croissant parmi les chercheurs.

Ils **suscitent** aussi une augmentation des triglycérides et du cholestérol sanguin, une résistance à *l'insuline* suivie de diabète et parfois des maladies cardiovasculaires.

4.2.1.6 *de façon à*

Il s'agit de prélever des cellules du patient, de leur ajouter des gènes qui inhibent la prolifération du *virus*, puis de les réimplanter **de façon à** protéger l'organisme.

Le canary-pox est manipulé **de façon à** exprimer un antigène du *virus* du sida.

4.2.1.7 *incit* à*

Doit-on alors administrer des substances qui activeraient volontairement les cellules au repos, les **incitant à** produire des particules de VIH, à exposer des antigènes viraux à leur surface et à déclencher ainsi une réaction immunitaire destructrice pendant qu'une multithérapie empêcherait le *virus* d'infecter d'autres cellules?

Une équipe de chercheurs a réussi à modifier génétiquement la bactérie de la salmonelle pour en faire un nouveau moyen contraceptif: un *vaccin* absorbable par voie buccale qui **incite** le système immunitaire **à** repousser le sperme avant la conception.

4.2.1.8 **génèr*/*génicité*

De plus, elle devrait stimuler la production d'époxystérols qui atténuent l'expression de l'HMGCoA réductase, ce qui **génère**, en synergie, un mode régulateur négatif. Contrairement aux statines, ces agents ne **généreraient** pas une surexpression de la réductase HMGCoA hépatique et ne réduiraient pas les niveaux d'ubiquinone et de dolichol...

4.2.1.9 *entraîn**

L'activité mutagène in vitro de la lamivudine n'ayant pas été confirmée par les tests in vivo, la lamivudine ne devrait pas **entraîner** de risque génotoxique chez les patients sous traitement.

L'administration d'une suspension orale de charbon activé en poudre ou de colestyramine **entraîne** une augmentation rapide de la clairance plasmatique de l'A771726 (voir section surdosage); il en résulte une réduction de la demi-vie d'élimination à 24 heures.

4.2.1.10 *car*

La modulation des canaux sodiques est probablement le principal mécanisme de l'activité antiépileptique, **car** il est partagé par plusieurs antiépileptiques autres que la phénytoïne.

Le *vaccin* antigrippal ne peut causer la grippe, **car** il ne contient pas de virus vivant.

4.2.1.11 *dû/due à/au*

Contrairement à la lovastatine, 75 % de l'activité de la pravastatine est **due à** la moléculemère.

Par ailleurs on a rapporté récemment que l'activité antioxydante de l'atorvastatine était **due à** un de ses métabolites principaux (l'ohydroxy) ...

De façon générale, il semblerait que plusieurs atteintes tissulaires **dues à** un stress oxydatif … soient liées à une défectuosité dans la coordination entre l'activité oxydative augmentée...

4.2.1.12 *par conséquent*

Les ERO altèrent les propriétés des membranes cellulaires, dont la fluidité et la compartimentation. Cela modifie l'activité des récepteurs et **par conséquent** la fonction des messagers secondaires, en plus d'entraîner la fuite des composés intracellulaires tels la lactate déshydrogénase (LDH) et la CK.

Les études animales ont indiqué que le léflunomide ou ses métabolites passent dans le lait maternel. **Par conséquent**, les femmes ne doivent pas allaiter lorsqu'elles sont traitées par le léflunomide.

4.2.1.13 *confèr*/confér** not *conférence**

Les *vaccins* contre l'hépatite B (Engerix B, Recombivax HB, Twinrix) sont sécuritaires, immunogènes et **confèrent** un haut degré de protection aux personnes *vaccinées*.

… les anticorps que l'enfant a reçus de sa mère perçoivent alors le *vaccin* comme s'il était l'agent pathogène et le détruisent comme ils le feraient pour la maladie elle-même, si bien que l'immunité **conférée** par le *vaccin* est perdue.

4.2.1.14 *à l'origine d**

La transmission de la grippe entre les travailleurs de la santé atteints d'une infection clinique ou infraclinique et leurs patients vulnérables est **à l'origine d**'une morbidité et d'une mortalité importantes.

Une mutation ponctuelle dans le site catalytique YMDD de la polymérase virale est **à l'origine de** la résistance du VHB au 3TC chez 10 à 15 % des patients traités.

4.2.1.15 *déclench**

La première administration avait **déclenché** une synthèse accrue des enzymes d'oxydation, cytochromes P-450, responsables de l'hydroxylation accélérée du phénobarbital lors d'une administration ultérieure.

Ces derniers **déclenchent** à leur tour l'activation d'autres lymphocytes, les B, qui sécrètent alors dans la lymphe et le sang des anticorps, les immunoglobulines.

4.2.1.16 *engendr**

Propagation : il y a d'abord formation d'un radical lipoperoxyde (L· + O2 ® LOO·) qui, en réagissant avec un autre LH, **engendre** simultanément un hydroxylipoperoxyde (LOOH), assurant ainsi la propagation du processus.

Systèmes *enzymatiques* liés au GSH : les superoxydes **engendrent** du peroxyde d'hydrogène (H2O2), composé non radicalaire et diffusable dont la toxicité semble liée à sa conversion en HO· au contact d'ions métalliques.

4.2.1.17 *responsable* d*/pour*

L'activité locale dans les tissus extrahépatiques pourrait être **responsable de** plusieurs effets des statines (entre autres, des effets préventifs de l'athérosclérose ou des effets sur les lipides membraneux). Elle pourrait également être **responsable des** baisses de niveaux de certains dérivés d'intermédiaires dans la synthèse de CH…

À long terme (10 - 20 ans), quelque 60% des patients développent une infection chronique active, **responsable de** maladies hépatiques graves telles que la cirrhose (20% des patients infectés de façon chronique), l'insuffisance hépatique et le cancer du foie.

4.2.1.18 *puisqu**

Puisque ces agents ne sont pas absorbés, il n'y a pas d'effets systémiques graves.

Si les mycoplasmes sont partiellement responsables, les antibiotiques devraient constituer un bon traitement de la maladie, **puisqu**'ils affaiblissent les bactéries.

4.2.1.19 *stimul**

Insuline de séquence humaine, moins immunogène que les *insulines* bovines ou porcines. Se lie à la sous-unité cellulaire de son récepteur et **stimule** l'activité tyrosine kinase de la sous unité bêta par autophosphorylation.

Ce vaccin serait le premier à utiliser des gènes du *virus* du Sida pour **stimuler** l'action des lymphocytes cytotoxiques (CTL).

4.2.1.20 *donc*

Les diabétiques ne peuvent assimiler correctement le sucre (glucides) parce que leur pancréas ne sécrète pas correctement une hormone qui active l'utilisation du glucose dans l'organisme. C'est l'*insuline*. Le patient doit **donc** surveiller constamment son taux de glucide dans le sang et s'injecter régulièrement de l'*insuline*.

Il en résulte des enzymes sur lesquelles les antirétroviraux ne peuvent plus agir. Le *virus* peut **donc** se multiplier librement.

4.2.1.21 *(se) faire* + infinitive verb[29]

… les lymphocytes T CD4 stimulent d'autres cellules, les lymphocytes T CD8 cytotoxiques, et leur **font détruire** les cellules infectées, qui libèrent de nouvelles particules virales…

[29] For certain patterns the concordances were sorted manually to enable us to evaluate the precision of pattern forms belonging to a given grammatical category, e.g., in this case, infinitive verbs.

Les employeurs et leurs employés devraient songer à **se faire *vacciner***, car il a été établi que la *vaccination* annuelle contre la grippe des travailleurs adultes en bonne santé contribuait à réduire l'absentéisme associé à des maladies respiratoires et à d'autres troubles...

Par exemple, le bacille de Koch **fait baisser** la sécrétion d'interféron gamma par les lymphocytes dits « T auxiliaires »...

4.2.1.22 *se tradui* en/par*

… la lovastatine et la pravastatine … peuvent réduire les LDL de 20 à 35 %, une réduction qui peut **se traduire par** des baisses de 30 à 35 % d'incidence de maladies coronariennes.

Dans le plasma, un effet antioxydant **se traduit** principalement **par** une inhibition de l'oxydation des LDL.

4.2.1.23 *caus**

Parce qu'une déficience des récepteurs LDL **cause** une réduction du retrait et une production accrue de LDL, de petites diminutions dans le nombre ou dans l'activité des récepteurs LDL entraînent une augmentation disproportionnée des niveaux sanguins de LDL.

Cette mutation a peut-être été **causée** par l'utilisation massive d'antibiotiques pendant les 15 dernières années.

Les patients doivent interrompre ou cesser les traitements *antiviraux*, soit à **cause** de leur toxicité, soit à **cause** de la résistance du virus.

4.2.1.24 *rend** (+ adverb) (+ object) + adjective

C'est son ARN, caché au coeur de la particule virale, qui **rend le VIH infectieux**.

Selon l'hypothèse du Dr Montagnier, cette nouvelle souche de mycoplasmes (venue des États-Unis) aurait pu, en entrant en contact avec le *virus* du sida africain, activer ce dernier et **le rendre** plus **pathogène**.

4.2.1.25 *indui*/induct**

Les traitements actuellement disponibles, vidarabine monophosphate et interféron alpha, permettent d'annuler la réplication et d'**induire** une rémission de l'hépatopathie chez seulement 20 à 40 % des patients.

4.2.1.26 *imput* à*

Pour **imputer** à un *vaccin* des effets secondaires inattendus, on fait appel à deux critères : l'imputabilité intrinsèque et l'imputabilité extrinsèque.

4.2.1.27 *(re)lié* à/au*

L'action des inhibiteurs de l'HMGCoA réductase est **reliée à** la dose.

Cette résistance est **liée à** des mutations des gènes codant pour les deux *enzymes* virales impliquées dans le mécanisme d'action de l'ACV...

4.2.1.28 *conséquence**

Le déficit en *hormone de croissance* conduit au "nanisme" hypophysaire dont les **conséquences** familiales et sociales aboutissent souvent à une véritable exclusion.

4.2.1.29 *(en) raison (d*)*

Avec un tel régime les LDL et le cholestérol plasmatique ne sont pas diminués sous l'action des statines (**en raison de** l'effet suppresseur additionnel de l'activité des récepteurs du foie).

Les effets bénéfiques de la cyclosporine sont lents à se manifester et ne sont généralement pas définitifs ; on observe des rechutes au bout d'un certain temps. La **raison** en est probablement le mode d'action de la molécule, qui empêche l'activation de nouveaux lymphocytes T, mais n'a pas d'effet sur ceux qui sont déjà activés.

4.2.1.30 *amen*/amèn**

L'opération de concentration en cause **amène** à examiner de manière plus approfondie les activités de recherche des parties dans le domaine de la thérapie génique HS-TK pour le traitement des tumeurs cérébrales et autres.

Un tel *vaccin* empêcherait l'embryon de se fixer au placenta et pourrait même **amener** le système immunitaire de la mère à détruire l'ovule dès qu'il est fertilisé.

4.2.1.31 *attribu* (à/au)*

... plusieurs effets nocifs qui sont **attribués aux** statines pourraient être reliés à leur activité dans les tissus extrahépatiques.

4.2.1.32 *condui*/conduct* (à/au)*

Dans ces cas, le VHB peut **conduire à** des cirrhoses et à des cancers du foie mortels.

L'inhibition de cette protéase virale **conduit à** la production de *virus* non infectieux.

4.2.1.33 *réact*/réag**

Doit-on alors administrer des substances qui activeraient volontairement les cellules au repos, les incitant à produire des particules de VIH, à exposer des antigènes viraux à leur surface et à déclencher ainsi une **réaction** immunitaire destructrice pendant qu'une multithérapie empêcherait le *virus* d'infecter d'autres cellules?

Cette étude vise à quantifier la **réaction** d'hypersensibilité aux trois *vaccins*, les produits entiers monovalents et les produits sous-unitaires monovalents.

4.2.1.34 *effet**not *en effet*

Le NO effectue son **effet** vasodilatateur en activant la cyclase guanylate...

D'abord conçu pour traiter la maladie de Parkinson, l'amantadine, dont on a noté les **effets** antiviraux sur les virus de type A seulement, cause toutefois des **effets** secondaires (étourdissements, confusion et autres **effets** pouvant aller jusqu'aux convulsions) qui limitent sérieusement son utilisation.

4.2.1.35 *crée*/créa** not *créatinine*

Ces souches diffèrent par leur pouvoir pathogène : les souches présentant un phénotype TKD ont un pouvoir pathogène atténué du fait de leur faible aptitude à **créer** une infection latente...

Cela expliquerait pourquoi le *virus* existait depuis longtemps en Afrique sans **créer** d'épidémie, et aussi pourquoi l'épidémie a éclaté simultanément en Afrique et aux États-Unis.

4.2.1.36 *manifestation**

L'hormone de croissance est également indiquée chez les filles présentant un syndrome de Turner, affection génétique presque aussi fréquente que la trisomie 21 dont l'une des **manifestations** est un retard statural.

Le terme "hépatite virale" est utilisé pour décrire l'atteinte hépatique causée par un nombre limité de *virus* hépatotropes, c.-à-d. infectieux pour le foie lui-même et dont la **manifestation** clinique principale est une hépatite.

4.2.1.37 *étiologi**

Les **étiologies** des déficits en *hormone de croissance* sont diverses [2] : - tumeur de la région sus-hypophysaire (craniopharyngiome) ; - irradiation dans le cadre de traitement de tumeur crânio-faciale, de neuroblastome de la fosse postérieure, leucémie aiguë ; - traumatique ; - malformative (interruption de tige pituitaire, malformation du névraxe associée) ; - ou idiopathique.

Le diagnostic est aisément évoqué lorsque l'on a la notion d'un traitement par *hormone de croissance* extractive et que l'on a écarté, en particulier par l'imagerie neuroradiologique, les **étiologies** classiques à cet âge responsables de troubles neurologiques rapidement évolutifs (tumeur de la fosse postérieure, leucodystrophie, récidive du processus tumoral initial, radionécrose, sclérose en plaques).

4.2.2 Destruction

4.2.2.1 *nocif*/nocive**

... plusieurs effets **nocifs** qui sont attribués aux statines pourraient être reliés à leur activité dans les tissus extrahépatiques.

4.2.2.2 *interfér*/interfèr** not *interféron*

... l'inhibition de l'OSC n'**interfère** pas avec leurs synthèses respectives...

4.2.2.3 *mortel**

Dans ces cas, le VHB peut conduire à des cirrhoses et à des cancers du foie **mortels**.

Le diabète non soigné est **mortel**.

4.2.2.4 *supprim*/*suppress**

… la synthèse de CH est **supprimée** par une inhibition de l'*enzyme* hydroxyméthylglutarylcoenzyme A (HMGCoA) réductase, qui est l'*enzyme* déterminant la vitesse de réaction…

Lorsque le CH est présent dans le régime alimentaire, l'ajout d'acides gras (à longues chaînes) augmente l'effet **suppresseur** des stérols alimentaires et entraîne une **suppression** encore plus grande de l'activité des récepteurs du foie.

4.2.2.5 *annul**

Les traitements actuellement disponibles, vidarabine monophosphate et interféron alpha, permettent d'**annuler** la réplication et d'induire une rémission de l'hépatopathie chez seulement 20 à 40 % des patients.

4.2.2.6 *tue*/tué**

On entend par *promédicament* un précurseur de substance active qui, en liaison avec le gène enzymatique, a pour effet de **tuer** les cellules.

... les antibiotiques diminuent l'effet **tueur** du *virus* sur des cellules infectées.

4.2.2.7 *élimin**

… simultanément, l'organisme produit aussi des anticorps, qui se fixent sur les particules virales libérées par les cellules et facilitent leur **élimination**. Malgré cette intense activité, le système immunitaire n'**élimine** pas tous les virus.

Ce procédé, utilisé à l'Institut Pasteur, élimine tout microbe des préparations d'hormone d'extraction.

4.2.2.8 *détrui*/destruc**

A priori, ce type de molécule complexe, ou " *promédicament* " dans la mesure où il s'active au contact de sa cible, peut être décliné sur tous les modes et combiner toutes sortes d'anticancéreux ; non seulement les cytostatiques, mais aussi les cytotoxiques, qui **détruisent** la tumeur.

Les cellules T tueuses, au contraire, se spécialisent dans la **destruction** directe de tel ou tel type d'antigène (tumeur, greffon).

4.2.3 Maintenance

4.2.3.1 *sine qua non*

Autre inconvénient majeur : E. coli n'effectue pas les modifications post-traductionnelles des protéines (en particulier la glycosylation, la carboxylation, etc.), qui constituent souvent une condition **sine qua non** d'activité de la protéine.

... en France, l'industrie pharmaceutique a insuffisamment intégré les nouveaux outils de la biologie moléculaire, condition **sine qua non** du développement des activités biotechnologiques.

4.2.3.2 *permet*/permi**

Le séquençage du gène codant pour la TK ou pour l'ADN polymérase **permet** l'identification du support génétique de la résistance.

L'activité *antivirale* de ces analogues contre un modèle organotypique de l'infection au VPH a été évaluée dans les laboratoires de l'université Penn State, **permettant** d'identifier plusieurs composés extrêmement actifs dont le profil d'activité est actuellement évalué par des essais traditionnels.

4.2.3.3 *indispensable* (pour/à/au)*

Un titrage préalable du *virus* est **indispensable pour** obtenir des plages isolées.

Ce sont des antiprotéases, substances qui bloquent une *enzyme* **indispensable** à la fabrication du virus.

4.2.3.4 *nécess**

Dans le lait, le fromage ou la crème glacée, la molécule moyenne de lactose ne dispose tout simplement pas de l'énergie **nécessaire** pour subir la réaction de décomposition, ce qui fait qu'il faudrait attendre bien longtemps avant que cette réaction ne se produise...

La mesure in vitro de la sensibilité des HSV aux *antiviraux* est réalisable en routine par des techniques simples comme le test colorimétrique au rouge neutre, mais elle **nécessite** l'isolement préalable du *virus* sur cellules.

4.2.3.5 *essentiel** not *essentiellement*

L'hormone de croissance est **essentielle** pour la protéine de synthèse, l'action hormonale (spécialement les hormones pour la thyroïde et le sexe), une récupération après l'exercice et peut être critique pour le processus de la guérison lui-même, spécialement après plusieurs blessures physiologiques.

4.2.4 Prevention

4.2.4.1 *empéch*/empêch**

... ces composés agissent lors de la phase précoce de multiplication virale en **empêchant** l'acidification de l'intérieur de la particule virale…

... une multithérapie **empêcherait** le virus d'infecter d'autres cellules...

4.2.4.2 *bloqu*/blocage**

... le cocktail de médicaments administrés n'est pas assez puissant pour **bloquer** la réplication virale.

De façon similaire, l'acivicine **bloque** la croissance des cellules du lignage B.

4.2.4.3 *enray**

L'objectif d'un *vaccin* thérapeutique ... est de stimuler le système immunitaire du sujet infecté afin de lui permettre de ralentir par lui-même, voire d'**enrayer** la progression de l'infection virale et donc le développement de la maladie hépatique.

On doit maintenant vérifier si cette protéine ainsi «déguisée» permet d'**enrayer** la capacité du virus à se reproduire dans de vraies cellules infectées, prélevées chez des individus séropositifs.

4.2.4.4 *évit**

D'autre part, les deux glycoprotéines de surface des particules virales, également sialylées, sont soumises à l'activité sialidasique de leur propre *enzyme*, ce qui **évite** leur agrégation les unes aux autres.

Les " *promédicaments* ". Des médicaments qui s'activent au contact de la cible pour **éviter** des effets désastreux.

4.2.4.5 *préven**

… le cortisol, une hormone humaine naturelle, et ses dérivés synthétiques, les glucocorticoïdes (en particulier la Prednisolone), ont une puissante activité anti-apoptotique sur les lymphocytes CD4 activés et infectés par le VIH. Cette activité **préventive** de la mort des CD4 est couplée à une inhibition partielle de leur activation.

Les " *promédicaments* ". ... Indications : tumeurs solides et **prévention** des métastases.

4.3 Influence Dependency

4.3.1 Modification

4.3.1.1 *influenc**

Le niveau d'activité de ces récepteurs est grandement **influencé** par la balance nette du cholestérol dans le foie et par le type d'acides gras qui atteignent les hépatocytes en provenance de la diète.

Les grandes variations dans l'activité de ces *enzymes* peuvent **influencer** la réponse individuelle à la lovastatine.

4.3.1.2 *neutralis**

En fait, notre idée est de se servir de la protéine Vpr comme cheval de Troie pour introduire un agent *antiviral* dans le virus pour le **neutraliser**.

L'interféron joue ce rôle, ainsi que des molécules qui **neutralisent** des protéines de régulation du *virus* (appelées TAT et REV).

4.3.1.3 *(s')impliq* dans*

... il y a un lien entre la synthèse du cholestérol et l'activité de la 7ahydroxylase (*l'enzyme* **impliqué dans** la formation d'acides biliaires) dans le foie.

Ces anti-alpha-glucosidases se comportent comme des *antiviraux* du fait que la glucosidase est **impliquée dans** le repliement et le transport dans le réticulum endoplasmique des glycoprotéines de l'enveloppe virale.

4.3.1.4 *(jouer) rôle** (+ adjective)

Prolifération des cellules des muscles lisses : … Cependant, on a rapporté qu'elle jouait un **rôle protecteur** contre la rupture des plaques ...

Acide ascorbique (COH) : joue un **rôle** très **important** en assurant la régénération de l'atocophérol en se transformant en un radical très peu réactif (CO·) ...

4.3.1.5 *dépend* d**

Ainsi l'équilibre net du cholestérol et les niveaux plasmatiques des LDL **dépend** grandement **des** événements survenant dans le foie.

« Chez le foetus, le contrôle de la croissance est surtout sous la **dépendance des** facteurs nutritionnels et de l'*insuline*, » explique le professeur Paul Czernichow.

4.3.1.6 **modul* not *modules***

... l'anticorps bloque la transmission d'un signal (peut-être de la « cascade ») qui va **moduler** la production …

Virocell développe des agents thérapeutiques dits "**immunomodulateurs**". Ce sont des substances qui activent et renforcent le système *immunitaire*.

4.3.1.7 *agi** not *s'agi**

Les trithérapies consistent à bloquer la reproduction du virus. Elles font appel à deux classes d'antirétroviraux qui **agissent** à deux endroits du cycle de réplication du VIH ...

Il en résulte des *enzymes* sur lesquelles les antirétroviraux ne peuvent plus **agir**.

4.3.1.8 **dépendant***

L'effet sur la relaxation **endothéliumdépendante** a également été rapporté chez le lapin hypercholestérolémique ...

4.3.1.9 *régul** not *régulier(s)/régulière(s)/régulièrement*

Insuline de séquence humaine, moins immunogène que les *insulines* bovines ou porcines. … Stimule l'activité intrinsèque de transporteurs et **régule** leur synthèse.

Cela réduit également la transcription des gènes **régularisés** par les stérols.

4.3.1.10 *normalis**

Le sujet assiste à la **normalisation** des analyses d'enzymes hépatiques six mois de suite au moins après l'arrêt de son traitement.

4.3.2 Increase

4.3.2.1 *amplifi**

Les radicaux thyil (RS·), issus de l'action de HO· sur les thiols (SH) cystéiniques des protéines contribuent à **amplifier** le processus de peroxydation lipidique.

4.3.2.2 *maximis**

Pour **maximiser** l'hypercholestérolémie, nous avons opté pour un régime contenant du triacylglycérol et du cholestérol.

4.3.2.3 *contribu**

Plusieurs autres facteurs **contribuent** au développement d'une lésion, tels le dommage à l'endothélium…

La réabsorption accrue de sels biliaires tend à contrecarrer l'effet bénéfique de la thérapie aux statines et **contribue** à la prolongation du temps nécessaire à l'obtention de l'état d'équilibre.

4.3.2.4 *favoris**

De fait, une augmentation du contenu en cholestérol des plaquettes **favorise** l'activation des plaquettes chez les patients hyperholestémiques.

Et si le système médico-administratif en place leur refuse de la *HGH* pour leur enfant, ces parents n'auront aucun scrupule à se la procurer clandestinement. Déterminés à obtenir de leur médecin ce traitement, ils **favoriseront** ainsi le développement d'un marché noir médical de la *HGH*, comme il en existe un dans le monde du sport.

4.3.2.5 *renforc**

Virocell développe des agents thérapeutiques dits "*immunomodulateurs*". Ce sont des substances qui activent et **renforcent** le système immunitaire.

Cette thérapie vise a **renforcer** la réponse du système *immunitaire* contre le HIV chez des personnes déjà infectées par le virus du Sida.

4.3.2.6 *aid* (à/au)*

Les possibilités offertes par la biothérapie sont prometteuses: stimuler nos défenses *immunitaires* pour les **aider à** éliminer les cellules cancéreuses ou corriger les anomalies génétiques responsables des tumeurs.

On appelle substrat la molécule qu'un *enzyme* **aide à** réagir.

4.3.2.7 *catalys**

Cette *enzyme* **catalyse** l'estérification intracellulaire de CH (ce qui en facilite le stockage).

Le catabolisme des chylomicrons est ***catalysé*** par la lipase lipoprotéique.

4.3.2.8 *accél**

Cette activation est **accélérée** en présence de métaux de transition.

Les réducteurs tels l'acide ascorbique, O2·, ou la cystéine **accélèrent** la lipoperoxydation par les métaux de transition.

4.3.2.9 *facilit**

… simultanément, l'organisme produit aussi des anticorps, qui se fixent sur les particules virales libérées par les cellules et **facilitent** leur élimination.

Les *catalyseurs* **facilitent** la réaction en permettant à une série de molécules réactantes de se transformer en produits, pour ensuite aider d'autres molécules à subir la même réaction.

4.3.2.10 *dilat**

L'endothélium régule la tonicité vasculaire des muscles lisses, l'adhésion et l'agrégation des plaquettes, la coagulation locale et la croissance vasculaire... Il sécrète des facteurs **dilatateurs** et constricteurs.

4.3.2.11 *augment**

… le stockage de CH est **augmenté** par l'activation de l'acétylCoA cholestérol acyltransférase (ACAT).

De plus, la température élevée **augmente** l'efficacité de plusieurs globules blancs et de certaines substances *antivirales*.

4.3.2.12 *amélior**

De même, chez la souris, l'administration systémique ou par aérosol d'une combinaison d'amantadine ou de rimantadine et de ribavirine **améliore** l'activité *antivirale* et augmente le taux de survie des animaux.

Le traitement substitutif par *hormone de croissance* est le seul recours pour **améliorer** la taille des enfants ayant un déficit somatotrope avec un bon rapport efficacité/tolérance.

4.3.2.13 *optimis** not *optimiste*/optimisme*

La mise à profit de ces connaissances scientifiques permet d'envisager une synthèse ciblée, afin d'**optimiser** l'efficacité thérapeutique tout en diminuant les effets latéraux non désirés.

4.3.2.14 *aggrav**

Le chercheur pense qu'une infection mycoplasmique peut activer le virus et **aggraver** la maladie en facilitant la production du *virus*.

Les données épidémiologiques disponibles infirment l'existence d'un lien causal entre la *vaccination* contre l'hépatite B et la survenue ou l'**aggravation** d'une SEP.

4.3.2.15 *hauss**

Les fibrates augmentent l'activité de la lipase lipoprotéique, **haussant** ainsi le catabolisme des VLDL et favorisant le transfert de CH au HDL.

4.3.3 Decrease

4.3.3.1 *endommag**

Une étude chez le chien... a démontré que, contrairement à la simvastatine et à la lovastatine, la pravastatine n'**endommageait** pas la respiration mitochondriale du myocarde au cours d'une ischémie.

Or, des anticorps ont bien été identifiés dans 50 % des cancers les plus fréquents, mais la plupart d'entre eux ne sont pas suffisamment spécifiques pour qu'on soit sur de ne pas déposer ces détonateurs biologiques au hasard dans l'organisme et d'**endommager** des cellules saines.

4.3.3.2 *constrict**

L'endothélium régule la tonicité vasculaire des muscles lisses, l'adhésion et l'agrégation des plaquettes, la coagulation locale et la croissance vasculaire.... Il sécrète des facteurs dilatateurs et **constricteurs**.

4.3.3.3 *délétère**

Malheureusement, la multiplication des obstacles à la réplication d'un *virus* a également un effet **délétère** sur les rendements de production de ce *virus*.

4.3.3.4 *ralenti**

L'hormone de croissance augmente les masses protéiques au niveau musculaire, des organes et du sang tout en **ralentissant** la destruction des protéines et en activant leur production.

D'origine pathologique : une atteinte hépatique sévère peut **ralentir** l'élimination de certains médicaments en raison d'une diminution de l'activité de certains enzymes qui les métabolisent.

4.3.3.5 *diminu**

... Les anticorps antibeta-2, connus pour **diminuer** la sécrétion d'interleukine 2, un messager cellulaire impliqué dans l'activation cellulaire...

… le retrait de LDL est **diminué** par une baisse de production de récepteurs LDL…

4.3.3.6 *intérromp**

Une fois là, le médicament **interrompt** la croissance de la chaîne d'ADN, empêchant donc le *virus* d'achever sa réplication.

Le Dr Samuel Broder, directeur d'oncologie clinique au National Cancer Institute, souligne que « c'est la première fois qu'un produit fabriqué par ingénierie génétique est utilisé pour tenter d'**interrompre** le cycle viral du HIV ».

4.3.3.7 *abaiss*/baiss**

L'administration de charbon activé (poudre mise en suspension) par voie orale ou par sonde nasogastrique (50 g toutes les 6 heures pendant 24 heures) s'est avérée **abaisser** les taux plasmatiques du métabolite actif A771726 de 37% en 24 heures et de 48% en 48 heures.

Il y a cinq ans, la mortalité atteignait 95 %. La découverte d'un agent *antiviral* l'a abaissée **à 30 %.**

4.3.3.8 *inhib**

L'amantadine et la rimantadine sont spécifiques des *virus* grippaux de type A et ne sont pas actives vis-à-vis des *virus* de type B, bien qu'à des concentrations élevées (> 10 mg/ml) ils **inhibent** également de façon non spécifique la multiplication des *virus* grippaux de type B et C ainsi que d'autres virus tels que le *virus* respiratoire syncytial (VRS), les *virus* parainfluenza 1, 2 et 3 ou encore les *virus* de la dengue…

… l'amantadine et la rimantadine agissent également en **inhibant** l'action stabilisatrice qu'exerce la protéine M2 en limitant l'abaissement de pH qui aurait pour effet un changement conformationnel prématuré de la HA et la production de particules virales non infectieuses.

4.3.3.9 *retard**

Dans chacun des deux groupes, huit malades prenaient de l'AZT (zidovudine, ou azidothymidine), un *antiviral* qui **retarde** l'apparition des symptômes du sida.

L'hormone de croissance est indiquée dans le traitement à long terme des enfants présentant une insuffisance de sécrétion de *l'hormone de croissance* par l'hypophyse, ce qui entraîne un **retard** statural.

4.3.3.10 *perturb**

Ces micro-organismes sont terriblement rusés. Dès qu'ils pénètrent dans l'organisme, les plus roués s'efforcent de **perturber** les communications entre les cellules *immunitaires*.

4.3.3.11 *inactiv**

Une fois liée au récepteur, la séquence peptidique peut l'activer, dans le cas ou la séquence est identique au peptide naturel (agoniste) ou l'**inactiver** en cas de séquence modifiée par rapport à l'original (antagoniste).

Les inhibiteurs de protéases peuvent cependant **inactiver** certaines protéases des cellules de la personne atteinte et ainsi perturber des fonctions biologiques importantes.

4.3.3.12 *rédui*/réduc**

... il a été établi que la *vaccination* annuelle contre la grippe des travailleurs adultes en bonne santé contribuait à **réduire** l'absentéisme associé à des maladies respiratoires et à d'autres troubles...

Un *vaccin*, même imparfait, **réduirait** le nombre des mérozoïtes et donc l'ampleur des dégâts.

4.3.4 Preservation

4.3.4.1 *mainten*/maintien** not *maintenant*

L'effet combiné de ces deux actions est de **maintenir** de faibles taux sanguins de LDL.

L'importance de l'activité antioxydante de l'ubiquinone réside dans sa capacité à **maintenir** un minimum d'activité suite à un stress.

4.4 Conclusions

As in English, a large number of possible patterns indicating a cause relation have been observed. Many of these patterns show good precision, in the 90% to 100% range, making them very good candidates for further research in the area of semi-automatic text analysis. As in the English results, there were few patterns discovered for the influence dependency of preservation. This should be considered when evaluating the possibilities for future research on this phenomenon.

In French, patterns for destruction that referred to death were observed (e.g., *tuer*, *mortel*), but there were more neutral patterns as well (e.g., *nocif*, *supprimer*, *annuler*).

The use of prepositions to indicate the cause relation, however, seemed to be very similar in French. Patterns such as *après* and *par* appeared to be very common, and led to the same kinds of difficulties described in Chapters 3 and 6.

These and other similarities and differences between the patterns in English and French would be an interesting subject for study. Chapter 5 will briefly introduce some possible areas for future work by outlining the results found in this project.

5 Comparison of English and French Patterns

5.1 Introduction

Collecting possible patterns for use in semi-automatic knowledge extraction and observing the issues in pattern identification and use were the primary foci of this thesis. However, the originality and pertinence of the work lies in large part in the bilingual nature of the extraction.

This chapter will take a very preliminary look at how the patterns identified in English and French are similar and different. The analysis will focus first on the patterns' form, and then on their frequency. The issue of cognates (words sharing a common root in both languages) and how their usage compares will also be addressed.

5.2 Form

Similarities and differences noted between the English and French patterns included those of grammatical category, word order, voice, emphasis, and agreement.

5.2.1 Grammatical category

In order to evaluate the predominant grammatical categories of the patterns observed, a simple count of the patterns studied for each of the sub-categories of the relation was carried out. In cases in which the patterns were likely to take more than one grammatical form, both or all of the possibilities were included in the analysis.

Given that the cause relation focuses on the interaction between two forces and the effect of this interaction on entities or events, it is not surprising that the vast majority

of the patterns observed were verbs. Of the patterns retained in this analysis, approximately 70% in English and 60% in French were verbs.

The remaining major grammatical categories observed were those of nouns or noun phrases, which accounted for 15% of the patterns in English and 25% in French; adjectives, which accounted for 10% of the patterns in English and 11% in French; and conjunctions, with 4% of the English and 6% of the French patterns. Adverbs or adverb phrases were also noted in English, but accounted for only 1% of the patterns.

Prepositions such as *by*, *from*, *with*, *de* and *avec* were also commonly observed. However, they were not very productive as patterns for semi-automatic extraction, given the noise associated with them in both languages. As such, this group was not retained for the purposes of this project and thus no statistics are available.

The major differences observed between the languages were found in the categories of verbs and nouns. There appears to be a clear difference of 10% of the patterns in the proportion of verbs (70% in English and 60% in French) and nouns (15% in English, 25% in French). In fact, this kind of difference is consistent with the observations of such researchers as Vinay and Darbelnet (1977:102-104), who observed that "Le rôle prépondérant du substantif en français a été constaté maintes fois, aussi bien par les hommes de lettres que par les linguistes." They quote André Chevrillon (1921:222) who stated that "le français traduit surtout des formes, états arrêtés, les coupures imposées au réel par l'analyse." Thus the difference in numbers between verb to noun forms is not unexpected.

5.2.2 Word order

As observed in Marshman, Morgan and Meyer (2002) and Morgan (2000:87), as well as by many other researchers (including Demanuelli 1995:130), word order in French can be much more flexible than in English. For example, with the pattern *result** /1L or /1R *in/of/from* in English, the preposition occurred after *result** in every case. However, with the pattern *résult** /1L or /1R *en/de* in the French corpus, there were several cases of *en*[30] preceding *résult**, as in the sentence

> 11) Lorsque l'ischémie est prolongée et que le dommage cellulaire devient irréversible, il **en résulte** un infarctus.

This can lead to difficulties in identifying possible patterns and in using them in semi-automatic tools, as well as in extracting the correct concepts from a sentence and linking them with the correct roles in the relationship.

5.2.3 Voice

Morgan (2000:90), citing Vinay and Darbelnet (1995:136), observed that pronominal forms in French may represent the middle voice, in which the subject is also the object of the action expressed by the verb. Morgan mentioned that this middle, pronominal voice is often translated by the passive voice in English, and also that this phenomenon makes it necessary to include pronominal forms and the variations that may be associated with them when choosing pattern forms for knowledge extraction.

In this corpus, some patterns were observed in pronominal forms. These included *s'allonger* and *s'impliquer* (*dans*). However, for the purposes of this project, the pronominal forms were less significant than they were for Morgan. This can be simply

[30] Note that in these cases *en* is a pronoun, not a preposition, although the surface form — and thus the pattern used for extraction in this project — is identical.

explained by referring back to the definition of the middle voice: cases in which the subject and the object of an action are the same. While this did occur in the corpus, for example in the sentence:

12) L'ADN viral cesse alors de **s'allonger**, et le virus ne peut plus se multiplier.
13)

the occurrences were fairly rare, since the focus was on the causative interaction between two concepts and not the isolated actions of one concept alone. However, given the possibility of contexts such as the one above, it was important to allow for possible pronominal forms in the proposed patterns.

Issues relating to the passive voice (see Chapter 6, section 6.2.2) were found in both languages.

5.2.4 Emphasis

Placing emphasis on a particular sentence element in French often involves changing word order, which in turn leads to some changes in sentence structure. One method of adding emphasis, mentioned by Vinay and Darbelnet (1977:210) is by using the structure "*c'est... qui*", as seen in this example from the corpus:

14) Lorsqu'on reçoit un vaccin contre une maladie virale, **c'est** ce mécanisme **qu'**on cherche à stimuler.
15)

This can make automatic extraction of the concepts in the sentence more difficult, as there can be unusual word orders and ambiguous structures present. For example, it is unusual to have the pattern (in this case, *stimuler*) occur at the end of the sentence, after both of the concepts it connects. In addition, if a system were trained to ignore relative clauses introduced by a pronoun — since they often interrupt the main idea in a sentence — examples such as the one above might also be excluded.

In comparison, the English counterpart of this method of placing emphasis, the "*it was... that*" structure, is much less commonly used. Vinay and Darbelnet attribute this to the wider freedom in accentuation available in English (1977:210).

5.2.5 Agreement

As we will see in Chapter 6, the patterns in French were almost inevitably more variable than those in English, in part because in French there are more agreements made (e.g., gender for nouns and adjectives, number for adjectives) or more varied forms used (e.g., conjugation of verbs). This phenomenon has already been noted in Morgan (2000:93) and Marshman, Morgan and Meyer (2002).

5.3 Frequency

The second major aspect of the interlinguistic comparison is that of frequency of patterns. An analysis of the pattern frequency data presented in Appendices A and B was carried out, taking into account the total number of occurrences in the corpus and the relative frequencies of the various patterns. Since it is difficult to designate the threshold of occurrences below which a lexical item is considered to be a fortuitous combination as opposed to a pattern, only the highest numbers of occurrences were compared.

The total numbers of occurrences of patterns in the corpora were largely parallel between the two languages, both if the patterns were divided by category and if the occurrences of all of the types of patterns were grouped together. There were similar maximums and similar relative frequencies between the top three most frequent patterns as well. For more details, see Appendix E.

5.4 Cognates

The final, and most specific, aspect of the interlinguistic comparison to be discussed here is that of cognates. It is obvious when looking at the data that many similar forms in English and French are present in the observed patterns. Appendix F lists cognates in English and French, accompanied by their frequency within the applicable corpus and their precision (not including the 'maybes').

If we consider similar values to be those that differ by fewer than ten occurrences and/or 10% precision, we can break down the statistics as follows. Approximately one-sixth of the cognates, including *incit** (*à*), *confer**/*confér**, *consequence**/*conséquence**, *maximiz**/*maximis**, *accelerat**/*accél**, *inactivat**/*inactiv**, and *influenc** showed both similar numbers of occurrences and precision. Approximately one-quarter of the cognates, including *immuniz**/*immunis**, *alter**/*altèr** and *maintain**/*mainten** showed similar numbers of occurrences in the corpora. Approximately one-third of the cognates showed similar precision in the corpus. Among these were *provok**/*provoc**, *generat**/*génèr**, *favour**/*favoris**, *reduc**/*rédui**, and *role**/*rôle**. The other cognates, more than half, did not show close similarities in either occurrences or precision.

What this indicates is that while there are many common forms found in knowledge patterns indicating the cause relation in the corpora, commonality of form rarely guarantees common usefulness for the purposes of knowledge extraction. While it is certainly interesting to investigate the cognates of previously identified patterns in another language, the patterns must be evaluated within their own language for their usefulness to be discovered.

Many factors contribute to this phenomenon. However, in this context only linguistic tolerance for repetition, differences in frequency of usage, and variation of form and usage between languages will be discussed.

Differences in frequency of usage are likely to be closely linked with linguistic tolerance for repetition, among other factors. As Quillard (1994; 2001) noted, French texts are likely to contain a larger variety of synonyms used to express a particular concept. Thus the individual patterns identified are likely to occur fewer times in a given corpus than their English counterparts. Consider as examples the quintessential causal pattern, *caus**, as well as the patterns *induc**/*indui** and *produc**/*produi**. While *cause* occurs 212 times in the English corpus, it occurs only 150 times in French. The same holds true of *induc**, with 207 occurrences in English and only 110 in French. *Produc** occurred 306 times in English and only 61 times in French. It seems likely that these significant differences can be explained in part by the use of such synonyms as *entraîner* and *inciter*. This can influence the usefulness of patterns, as it may be necessary to identify a larger number of patterns in order to achieve the same recall.

However, the differences in frequency are not all attributable to this tendency. In fact, there are cases in which patterns occur much more frequently in the French corpus than in the English. These patterns include *gén**/*gèn**, which occurred 397 times in the French corpus and only 172 times in the English, and *responsable* d*/pour*, which occurred more than twice as many times in French (77) as *responsible* /5R *for* did in English (30). Some other contributing factors to these variations in frequency may be collocations, fixed expressions or differences in register commonly used. Further study would be necessary to determine what role, if any, these factors play.

The final factor to be discussed here is differences of form between languages, including variations in word roots and relationships of elements in multi-word patterns.

The languages permit varying degrees of specificity in pattern forms. In French, there is often a nominal and a verbal form of words with common roots, such as *produire* and *production*, *réduire* and *réduction*. The roots in conjugated verbs also vary, as in *obten**/*obtien** and *mainten**/*maintien**. This means that greater care must be taken in selecting the pattern forms than would be the case in English with its less variable forms such as *produce*/*production* and *reduce*/*reduction*.

In another type of problem, in one language a pattern may be associated with fewer prepositions (for example, *résult** *en/de*) than in the other (*result** *in/of/from*). As well, there may simply be more forms used of a cognate in one language than in another. Consider *consequen** in English, which corresponds to both *conséquence** and *par conséquent* in French.

From this very preliminary analysis, we see some of the benefits and the pitfalls of using cognates of established patterns in one language to develop patterns in another, and some of the reasons for these phenomena.

5.5 Conclusions

It is clear from the above discussions that many factors come into play in differentiating the identification, analysis and use of patterns in English and in French. All of these must be considered at the stage of compiling a list of patterns as well as in programming a tool to use them for knowledge extraction.

In general, it seems that there are more similarities than differences in the English and French patterns present in this corpus. However, this is not to say that the differences

are negligible. They must be considered at every stage of a project in this area: the design of the project, identification of possible knowledge patterns, testing of these patterns, design of a knowledge extraction tool, and integration of the patterns into it.

Many other interlinguistic differences (e.g., commonly occurring syntactic patterns; the number, range, types, and density of patterns found; the distance of the patterns from the terms to which they refer; the use of chains of cause; the frequency and types of anaphora; and the use of hedges and modals) should also be considered throughout such a project. This is only a very preliminary analysis, and serves only to suggest some of the interesting avenues for further study, rather than leading us to draw any firm conclusions.

6 Issues in Pattern Identification and Use

6.1 Introduction

This chapter will discuss some of the difficulties and issues observed in the course of this research. They are categorized into five groups: pattern-related issues, text-related issues, language-related issues, relation-related issues, and human-related issues. Each will be described and most will be illustrated with examples from the corpora.

6.2 Pattern-related issues

These issues include difficulties posed by the form and usage of the patterns[31], and will be divided into issues of polysemy, variation, negation, non-contiguity, multiplicity, noise, and unpredictability of placement in sentences.

6.2.1 Polysemy

The phenomenon of polysemy in knowledge patterns is analogous to polysemy in general language words; it occurs when a linguistic unit can have more than one meaning. It can create the same ambiguities with patterns as with general language words, but with the added difficulty of deciding whether the two or more senses of the pattern may indicate different sub-types of the cause relation, or whether one or more of the senses is entirely unrelated to cause.

---

[31] In this chapter, patterns will be designated by neutral forms, rather than by the coded forms used for generating the concordances. For example, in English the uninflected forms of verbs, and in French masculine singular forms of adjectives and infinitive forms of verbs will be used.

For example, the patterns *stimulate* in English and *stimuler* and *inciter à* in French can express two different sub-types of the cause relation, either creation or increase. Consider the following sentences taken from the corpus:

16) PAF can amplify the inflammatory response. Lexipafant works by blocking the PAF receptor and thus stopping PAF from **stimulating** cells. (increase)
17) All these active ingredients are antigens: substances that can **stimulate** the immune system to produce specific antibodies. (creation)
18) L'objectif est de **stimuler** les défenses naturelles de l'organisme contre le virus. (increase)
19) ... résultats prometteurs de leur vaccin basé sur deux gènes du virus du sida couplé à l'injection d'un gène **stimulant** la production d'interleukine-2 qui renforce les défenses immunitaires. (creation)
20) Aujourd'hui encore, des publicités à l'usage des médecins **incitent à** la vaccination des bébés, celle par exemple des laboratoires SmithKline Beecham. (increase)
21) Une équipe de chercheurs a réussi à modifier génétiquement la bactérie de la salmonelle pour en faire un nouveau moyen contraceptif: un vaccin absorbable par voie buccale qui **incite** le système immunitaire à repousser le sperme avant la conception. (creation)

Another polysemous pattern in English is *inhibit*, which can occur in expressions of decrease or prevention :

22) Potential in vivo problems like rapid clearance or **inhibition** of the *catalytic* activity of 38C2 by covalently binding diketones or other potential inhibitors were not observed. (decrease)
23) In accordance with its mechanism, *catalysis* was completely **inhibited** by 2,4-pentanedione … (prevention)[32]

This type of polysemy will pose challenges for developing tools to automatically provide a detailed analysis of the cause relation in texts.

However, in a more problematic type of polysemy a pattern may indicate cause in some cases and not in others. This is the case, for example, in this sentence, in which the pattern *responsible for*, usually indicating creation, does not indicate cause at all:

24) Each company is **responsible for** its own selling expenses.

[32] Note here that although it is *completely* that identifies this context as an example of prevention rather than decrease, the context as it would be extracted by a semi-automatic system — i.e., the end result — nevertheless indicates prevention.

Similar problems can occur in French with the pattern *cause*, which may also occur in expressions such as *remettre en cause*, as in the following sentence:

> 25) Malheureusement des publications récentes semblent **remettre en cause** ces résultats.

Clearly, these sentences are not meant to express that companies cause their selling expenses to exist or that the publications cause the results. Rather the pattern is being used in another, unrelated sense.

This kind of polysemy poses difficulties in the development of automatic knowledge extraction tools. It is difficult to consistently exclude the non-cause senses of these patterns from results; some patterns may simply not be usable if they present too many alternative senses. Of course, this brings up the eternal compromise of automatic knowledge extraction tools: as noise is decreased, so is recall.

6.2.2 Variation

Variability of pattern forms was another type of pattern-related difficulty encountered during this research. While in many cases WordSmith Tools' wildcard and alternative features could compensate for pattern variations, others were less predictable or led to excessive noise if the search string used was adjusted to include them.

Among the patterns that showed problematic variability were *cause* and *ease* in English. Variations on the pattern *cause* included not only *cause*, *causes*, *causing* and *caused*, but also *causative* and *causation*. As these forms are less common, they could easily be excluded accidentally when choosing the pattern form (e.g., by using a form such as *cause*/causing*) and thus require extra care and forethought on the part of the researcher. For *ease*, the verb forms *ease*, *eases* and *easing* were promising, but

variations such as the noun *ease*, *easy*, *easier*, *easiest*, and *easily*, as well as the domain-related abbreviations *EASD* and *EASL*[33], were not generally helpful.

The situation was even more difficult in French, as the variations in verb conjugation and agreement are wider than in English, and the use of diacritics can also introduce variation. Some examples of highly variable patterns in French are *générer* and *réguler*. *Générer* occurred in multiple forms, such as *génère*, *génèrent*, *générant*, *-gène*, and *–gènes*, but forms such as *général*, *génétique*, *génie*, and *génome* were excluded because they often produced noise. Forms of the pattern *réguler* included *régule*, *régulant*, *régulait*, and *régulation*, but forms such as *régulier*, *régulière*, *réguliers* and *régulières*, as well as *régulièrement* were not productive and were therefore excluded.

There were also issues of variation in the number, type and order of pattern elements. For example, the pattern *result* can occur with or without the prepositions *in* and *from*. Obviously, the preposition used, if any, can completely change the sense of the statement, as seen in these examples:

26) ... TAPET carrying a cancer-fighting drug can **result in** much better antitumor activity and regression of tumors.

27) ... a significantly faster increase in serum insulin levels and a shorter plasma half-life **resulted from** an injection of Humalog when compared to Humulin R.

One example of changes in both type of pattern elements and pattern order is *résulter en*; it is possible for the pronoun *en* to occur before *résulter*, as in

28) L'administration d'une suspension orale de charbon activé en poudre ou de colestyramine entraîne une augmentation rapide de la clairance plasmatique de l'A771726 (voir section surdosage); il **en résulte** une réduction de la demi-vie d'élimination à 24 heures.

[33] These acronyms stand for the European Association for the Study of Diabetes and the European Association for the Study of the Liver, respectively.

Finally, there were differences attributable to the passive and active voices. Often, as in example 14, when a pattern occurred in the passive the agent responsible for the action was not clearly specified, reducing the context's value for knowledge extraction.

> 29) Les oncogènes sont, en effet, des gènes qui, à l'état normal, participent au développement régulier de la cellule, mais qui, lorsqu'ils sont mutés ou **modifiés**, peuvent provoquer un cancer.

It is apparent that accounting for all possible variations of a pattern is extremely difficult. It requires experimentation in order to determine what forms should be included and which excluded. This task is even more challenging in French.

6.2.3 Non-contiguity

This problem, which arises when the elements of complex patterns are interrupted by other textual elements such as adjectives, adverbs, and noun, adjective or adverb phrases, is frequent in both languages. Some examples of this problem include:

> 30) Acute overdosage could **lead** initially **to** hypoglycemia and subsequently **to** hyperglycemia.
> 31) L'activité prooxydante de l'ubiquinone est **reliée**, en sus de la localisation du SQH dans la membrane, **à** la disponibilité des partenaires possibles pour les réactions secondaires.

Non-contiguity can make the automatic identification and extraction of knowledge much more difficult: the larger the search window, the more opportunity there is for contexts to be mistakenly be identified as pertinent (for example, if the system identifies elements from separate clauses as belonging together), but valid examples may not be identified if the search window is too small or the pattern form used too restrictive.[34]

[34] For the purposes of this project, search windows for multi-word patterns were generally adapted as much as possible to the individual patterns. The first trials generally used a 5-word window on both sides of the main pattern element, after which the windows were adjusted up or down according to the results produced.

6.2.4 Multiplicity of patterns

Two main variations on this problem were observed: 1) the use of two or more possible patterns to indicate the same relation, and 2) chains of cause, in which multiple cause and effect relationships occur in the same sentence and are indicated by separate patterns.[35]

An example of the phenomenon of multiple patterns indicating the same relation, either between the same concepts or different concepts, is found in the following sentences:

32) Chronic hepatitis C is the leading **cause** of chronic liver disease and the most common **reason for** liver transplant, according to WHO.[36]
33) When this protein is used as a vaccine, the foreign portion **triggers** helper T cells, which **orchestrate** an ***immune*** **response** that includes antibody-**producing** B cells. The other portion **ensures** that the **resulting** antibodies will recognise and **inactivate** CETP.
34) **Parce qu**'une déficience des récepteurs LDL **cause** une **réduction** du retrait et une **production** accrue de LDL, de petites **diminutions** dans le nombre ou dans l'**activité** des récepteurs LDL **entraînent** une **augmentation** disproportionnée des niveaux sanguins de LDL.
35) L'étude des fonctions de reproduction n'a montré aucun **effet délétère** (notamment **tératogène**).

Examples of chains of cause include the following:

36) For example, suspended protein powders can show *catalytic* activity for days in neat organic solvents at temperatures far above 100oC [34,35]. This is **due to** the fact that dehydrated proteins are "rigid" in organic solvents. **Thus**, even though they can be denatured by the solvents for thermodynamic **reasons**, **due to** this increased rigidity they are trapped...
37) C3H mice infected in the footpad with L. major resolve their lesions by **induction** of a TH1 immune response. The specific TH1 response **resulted from** the cytokine process **triggered** on antigen-*activated* CD4 targets by the release of IL-12 from natural killer cells.
38) De façon générale, il semblerait que plusieurs atteintes tissulaires **dues à** un stress oxydatif ... soient **liées à** une défectuosité dans la coordination entre l'**activité** oxydative **augmentée**, qui **diminue** les niveaux tissulaires d'ubiquinone, et la biosynthèse de l'ubiquinone.

---

[35] In this section, due to the potential confusion caused by the many intersecting cause relations, the patterns will be indicated in bold but no underlining will be used to identify the related concepts.

[36] This text continues, "Liver damage **caused** by HCV is the leading **cause** of liver transplantation in adults."

39) Les ERO **altèrent** les propriétés des membranes cellulaires, dont la fluidité et la compartimentation. Cela **modifie** l'activité des récepteurs et **par conséquent** la fonction des messagers secondaires, en plus d'**entraîner** la fuite des composés intracellulaires tels la lactate déshydrogénase (LDH) et la CK.

Regardless of whether the patterns indicate one relation or a chain of cause, however, this phenomenon can lead to repetition in automatic extraction, since the same sentence is identified and extracted more than once.

6.2.5 Noise

As described in Chapter 1, *noise* is a term used to refer to contexts extracted by a knowledge extraction tool that do not really reflect the desired relationship. Some patterns produce more noise than others because of their form or meaning.

In these corpora, many indicators of cause relations were prepositions. Some examples are *with*, *by*, *in*, *after*, *après*, and *avec*:

40) **After** thorough agitation, the vaccine is a slightly opaque, white suspension.
41) Protein encapsulation **by** the double-emulsion/solvent-evaporation (w/o/w) technique.
42) L'immunité **après** l'administration du vaccin inactivé dure généralement < 1 an.
43) **Avec** la prise de CH et de triacylglycérol saturé, il y a une augmentation marquée des taux de LDL ...

These are closed-class words with many subtle nuances of meaning. Thus it was difficult to determine precisely in which sense the word was being used. Given this polysemy and the frequent occurrence of these words, these patterns were not generally precise enough to be used for the purpose of identifying cause relations.

A more specific sub-type of the above problem is linked to the patterns *after* and *après*, but also other patterns such as *following* and *à la suite de*. These were not very promising for this research, because of the dangers of the post-hoc fallacy (the assumption that because an event occurred *after* another, then it must have occurred *because of* that other event). As examples 25 and 27 show, this line can be difficult to

draw. There seems to be a dual meaning to the statements: certainly the cloudiness and the immunity occur after the agitation and the vaccination, but in these cases it is evident that the actions are the *causes* as well. In other cases the relationship may not be so clear.

In addition to prepositions, other grammatical categories of linguistic patterns produce too much noise to be easily used for knowledge extraction. Among these are prefixes and suffixes. While some affixes, such as *immuno-* and *-genic* produce relatively precise results, others, such as *pro-* and *anti-* produce excessive amounts of noise. While these affixes are certainly used to express clear cause relations in some contexts, they are too common to be precise for this purpose, also occurring for example in words such as *propose, prolifération*, *antigen*, and *antidote*.

There are many other knowledge patterns that, independent of their grammatical category, produce too much noise to be used for semi-automatic knowledge extraction. Among these were the patterns *support*, *limit*, *émerger de* and *relatif à*. While these patterns can be used to extract causal relations such as those found in the following sentences:

44) Only antivascular targeting activity can destroy existing blood vessels **supporting** tumor growth.
45) Their antiviral activity may therefore be **limited** by an inefficient metabolism...
46) Les activités prooxydantes de SQH· **émergeant de** la fonction antioxydante de QH2.
47) Les patients atteints de troubles endocriniens, y compris ceux **relatifs à** un déficit en hormone de croissance, présentent un risque accru d'épiphysiolyse.

they may also produce noise such as that in the following examples:

48) **Supported** by an $850,000 grant from the National Institute of Mental Health (NIMH), AIDScience.com is designed to serve AIDS prevention and vaccine researchers.
49) EPTTCO **Limited**, a private UK biotechnology company today announced a multi-year agreement...
50) Ce fut le cas en 1957 et en 1968 avec l'**émergence des** virus A(H2N2) et A(H3N2) respectivement, ou encore lors de la réintroduction des virus ...
51) La meilleure preuve de cette « coévolution», c'est qu'il existe dans les zones endémiques des individus **relativement** résistants **au** Plasmodium.

One more sub-type of noise remains to be discussed: domain-related noise. This occurs when lexical items specific to the subject field are similar to generic patterns used and can be mistakenly identified as pattern words. This is the case with terms such as the *Centre for Disease Control and Prevention*, *benzyl alcohol preserved*, and *reductase*, as well as the previously mentioned examples *EASD* and *EASL*. Other noisy English patterns were *result* and *control*. While these intuitively seem to be very promising patterns, in this corpus they generated an enormous amount of noise. The examples below:

52) The **results** of studies of three cancer cell lines at the two different time points are summarized in Table 1.
53) Seven months later, these vaccinated monkeys and a **control** group were exposed to a highly virulent hybrid of simian and human immunodeficiency virus.

illustrate domain-related noise, designating the observations from a scientific study and a group of subjects used as a basis of comparison during clinical trials. Given the domain, these senses were quite prevalent.

While it is certainly possible in some cases to exclude a portion of this noise by modifying the patterns used, in the case of many pattern candidates there appeared to be too many noise-producing variations for precise results to be obtained without immense investment of time.

6.2.6 Unpredictability of placement

In semi-automatic knowledge extraction, the more precisely the location of the pattern relative to the concepts involved in the relation can be stated, the more useful the tool will be. However, in natural language texts, the immense variation possible in sentence structure can complicate calculating a generic formula or algorithm for predicting the relative placement of the pattern and concepts. An enormous amount of

work is required to determine the likely placement of the elements, and even then the accuracy is likely to be limited.

Consider these contexts, which show what might intuitively be identified as common sentence structures:

54) Under physiological conditions, *activated* effector cells **trigger** a transient but composite reaction.
55) S'il demeure indispensable de rappeler que les personnes à risques doivent **prévenir** la grippe plutôt que la guérir en se faisant vacciner...
56) Dernière piste : modifier des cellules de manière que, lorsque le *virus* se présente, elles produisent une grande quantité d'interféron, une protéine naturelle de **défense** contre les *virus*.

Now consider the following sentences, using the same patterns:

57) In Ab-treated C3H mice, the IL-12-**triggered** process was blocked as long as massive anti-IL-12 Abs of high affinity (as a result of repeated administration) were loading the lymphatic fluid…
58) Dans ces cas, le VHB peut conduire à des cirrhoses et à des cancers du foie mortels. Certes, il existe déjà un *vaccin* préparé à partir de plasma. Il a été autorisé en 1982. **Préventif**, il est toutefois onéreux (150$) et peu connu.
59) Les protéines de l'enveloppe ont des régions hypervariables qui permettent au virus d'échapper facilement aux anticorps, **défenseurs** de l'organisme.

While a change in voice or the form of a pattern (for example, its grammatical category, as in examples 43 and 44) can be a predictor of more unusual distribution of the pattern and concepts in a sentence, it will require a substantial amount of research to identify these general tendencies with any precision. Even if this work was done, semi-automatic systems would likely still encounter problems with contexts such as example 43, in which the pattern occurs in a completely different sentence from either the cause or the effect.

6.3 Text-related Issues

These issues are related not to the patterns themselves, but to the texts used in the corpora. However, there are clearly links between pattern- and text-related issues that

cannot be completely ignored. Some issues could be classified in either or both of the categories.

The issues to be discussed in this section include misleading attachment, anaphora, lack of overt statements, hedges and modals, negation, non-contiguity of pattern and concepts, complexity of concepts, writing problems, and repetition.

6.3.1 Misleading attachment

In this section two types of attachment problems will be discussed: the use of general or vague words to designate cause or effect and reference to concepts in contexts that differ substantially from the usual or most precise interpretations.

Consider the following contexts:

60) The situation **forced** health care providers to give *immunization* priority to the elderly, children, and individuals with compromised *immune* systems.
61) … simultanément, l'organisme produit aussi des anticorps, qui se fixent sur les particules virales libérées par les cellules et facilitent leur élimination. Malgré cette intense activité, le système immunitaire n'**élimine** pas tous les virus.
62) S'il demeure indispensable de rappeler que les personnes à risques doivent **prévenir** la grippe plutôt que la guérir en se faisant vacciner...
63) ... leur pancréas ne sécrète pas correctement une hormone qui **active** l'utilisation du glucose dans l'organisme. C'est *l'insuline*.

In example 45, *situation* could perhaps be more precisely stated as *the shortage of vaccine*. In example 46, it is obvious from the previous sentence that it is the *antibodies* of the *immune system* that eliminate *viral particles*; in 47 it is the *vaccine*, and not the *people* who directly prevent the flu. The words used are relatively general and as such do not convey much information about the concepts.

Consider now patterns associated with terms that are not, scientifically speaking, accurate. Take for example the following contexts:

64) ...it has been shown that HIV can be safely **inactivated**, there are no private companies developing a preventative whole-killed vaccine at this time.

65) Une immunisation par des vaccins atténués ne doit être réalisée qu'après au moins 6 mois d'arrêt du traitement par Arava ...

In this case, it is not the *vaccine* that is whole, killed, attenuated or inactivated, but in fact the *virus* used to produce the vaccine. While these are common terms in the field, used to differentiate vaccines made with attenuated viruses from those which are made from killed or genetically modified viruses, they nevertheless could be misleading when the contexts are extracted by a semi-automatic system, which does not have the benefit of common sense or the ability to make inferences as humans do. Users would be required to examine information with a critical eye and a good working knowledge of the subject field in order to correctly identify such cases.

6.3.2 Anaphora

This is one of the most commonly identified issues in automatic and semi-automatic knowledge extraction (Pearson 1996, Morgan 2000, Meyer 2001, Marshman, Morgan and Meyer 2002). This term is used to describe cases in which concepts are missing from the direct context in which they are discussed (generally the sentence). The phenomenon is frequent, as of course it makes for a very boring text to have a term repeated again and again. Note, for example, that in the above sentences, we used *issue*, *term*, and *phenomenon* to refer to the concept 'anaphora.' Had we not, the repetition would be striking and would make for an unpleasant reading experience.

Cases of anaphora include those in which the term denoting a concept is replaced by a pronoun or a generic term, or even cases in which the concept is not overtly referred to at all, as seen below:

66) IL-6 is involved in **stimulating** the production of acute phase proteins in the liver. … It is also involved in the **stimulation** of T cell *activation* and growth, and also aiding the proliferation of haemopoietic progenitor stem cells.

67) Habituellement, les nucléosides non naturels n'ont aucune activité biologique. « Pendant la phase de développement, nous avions voulu les éliminer du mélange, raconte Gervais Dionne, car nous croyions qu'ils étaient **responsables de** la toxicité de notre produit. »

68) Les ERO altèrent les propriétés des membranes cellulaires, dont la fluidité et la compartimentation. Cela **modifie** l'activité des récepteurs et par conséquent la fonction des messagers secondaires, en plus d'entraîner la fuite des composés intracellulaires tels la lactate déshydrogénase (LDH) et la CK.

69) Moi, je suis un porteur chronique. Je ne ressens aucun **symptôme** pour le moment sauf que mon foie est en train de s'abîmer lentement.

In these examples it is necessary to consult the larger (extra-sentential) context in order to identify the concept referred to by *it* (51), *les/ils* (52) and *cela* (53). In example 54, the cause of the liver damage (in this case, hepatitis) is not overtly mentioned in context even using a pronoun or other lexical indicator. It is left up to the larger context to indicate what is meant.

It is commonly accepted that repetition of the same lexical elements is even less acceptable in French than it is in English (Quillard 2001). Therefore, anaphora may occur more frequently in French. The number of good examples found in the French corpus compared to the English corpus seems to support this hypothesis.

Unfortunately, anaphora is a very difficult issue to deal with in automatic or semi-automatic knowledge extraction. A sophisticated parser may be required to deal with the phenomenon, and even then many cases are unsolvable without human intervention. In poorly written texts, even human readers may not be able to correctly or certainly identify the concept to which the author is referring.

6.3.3 Lack of overt statements

The distinctiveness of the text types included in the corpora unavoidably affects the patterns observed and their usage. Many of the texts used were either researchers' first-hand accounts of their observations (journal articles and abstracts, theses) or reports

of this type of account (*Scientific American* and *Science et vie* articles). Because of this, many of the texts adhered to the norms that are generally imposed on scientific texts.

Supporting statements with hard data and/or references is one of these norms. If this support is not available because of the scope of the research or the stage it has reached, authors may make statements of observation, rather than assertion. This is often reflected in the pattern used to indicate cause relations. Consider the following sentences:

70) Treatment with recombinant growth hormone eliminates the risk of disease **associated with** products derived from human tissue.
71) Skeletal abnormalities including scoliosis are commonly **seen in** untreated *Turner* syndrome patients.
72) Dans 53 cas, le diagnostic de maladie de Creutzfeldt-Jakob iatrogène **après** traitement par hormone de croissance extractive a été retenu.
73) La perte de poids **associée au** VIH est la première et jusqu'à présent la seule altération de métabolisme pour laquelle les effets de l'*hormone de croissance* aient été approfondis et pour lesquels elle est officiellement recommandée comme thérapie.

In these examples, the intention is clearly to express a causal relation. However, the patterns used (*associated with*, *seen in*, *après*, and *associé avec*) are all at best ambiguous. The post-hoc fallacy risked by interpreting *après* as an indicator of cause has already been discussed. However in example 57 *iatrogène* indicates that the treatment caused the infection. The patterns *associated with*, *seen in*, and *associé avec* could be interpreted in different ways, although the only certainty is co-occurrence; the events occur together but not necessarily in a causal relation. In addition, even if one assumes that a cause relation is present, the statements do not confirm which event is the cause and which the effect. The reader is required to use his or her own knowledge and judgment to answer that question.

Similar ambiguities occur with patterns such as *relationship between* and *lié à*:

74) The **relationship**, if any, **between** leukemia and growth hormone therapy is uncertain.

75) De façon générale, il semblerait que plusieurs atteintes tissulaires dues à un stress oxydatif... soient **liées à** une défectuosité dans la coordination entre l'activité oxydative augmentée...

While with domain knowledge and common sense it is possible to guess with some accuracy which of the concepts linked by the pattern is cause and which is effect, the pattern itself does not make this clear.

While it would take an in-depth comparison of the frequency of occurrence of these patterns in specialized corpora in this and other domains — and perhaps even in a general language corpus — to confirm, it seems likely that the conventions of scientific writing play a role in the popularity of these types of implied statements of cause. They likely also contribute to the phenomenon to be discussed in the next section, the use of hedges and modals.

6.3.4 Hedges and modals

These two methods of attenuating the certainty of statements, as mentioned above, are likely linked to the text types used in the corpora and to the domain.

The first method is hedging, used to limit the extension or generality of a statement. The second, modal verbs, achieve the same effect using a more restricted grammatical category.

Consider the following contexts in which hedges are used:

76) … HIV appears to **cause** AIDS after the onset of antiviral immunity.
77) The **relationship**, if any, **between** leukemia and growth hormone therapy is uncertain
78) There has been no evidence to date of Humatrope-induced **mutagenicity**.
79) It is not known whether Humatrope can **cause** fetal harm when administered to a pregnant woman…
80) Ils **suscitent** aussi une augmentation des triglycérides et du cholestérol sanguin, une résistance à *l'insuline* suivie de diabète et parfois des maladies cardiovasculaires.
81) C'est le quatrième scientifique de Pasteur à être mort d'un cancer de type très rare chez le jeune adulte. On s'est demande [sic] si le lymphome lymphoblastique qui l'a

tue était, ou non, **imputable à** la manipulation de bactéries ou de *virus* modifies génétiquement, ou de substances utilisées dans le cadre de mutations génétiques.

82) De façon générale, il semblerait que plusieurs atteintes tissulaires dues à un stress oxydatif (Beyer et Ernster, 1990) soient **liées à** une défectuosité dans la coordination entre l'activité oxydative augmentée, qui diminue les niveaux tissulaires d'ubiquinone, et la biosynthèse de l'ubiquinone.

83) La libération de *l'insuline* topique et non-invasive, en utilisant les Transfersomes® comme vecteurs appropriés, suggére [sic] une potentialité thérapeutique. Il est possible que ceci **mène à** une régulation métabolique ajustée et plus précise, et améliore la qualité de vie.

The items *appears to* (61), *if any*, *uncertain* (62), *there has been no evidence to date* (63), *it is not known whether* (64), *parfois* (65), *on s'est demandé si... ou non* (66), *de façon générale il semblerait...* (67), *suggère* and *il est possible* (68) all attenuate the statements or implications made in the contexts.

The modals *can* (69), *may* (70), and *pouvoir* (71) all reduce the effectiveness of the statements made in the following contexts:

84) Growth hormone can **help** patients with these conditions potentially grow at a normal rate.

85) Including a mixture of CTL and Th epitopes may be **critical to** protection.

86) Un tel *vaccin* empêcherait l'embryon de se fixer au placenta et pourrait même **amener** le système immunitaire de la mère à détruire l'ovule dès qu'il est fertilisé.

87) ... plusieurs effets nocifs qui sont attribués aux statines pourraient être **reliés à** leur activité dans les tissus extrahépatiques.

One type of hedge that is often used is to attribute a statement to another source, as seen in example 67. While in scientific literature citing others' research is respected and even required, it also does distance the author from the statement and indicates that there is a possibility of error not only in the scientific observations but also in the author's interpretation of those observations. Moreover, the use of *attributed* or *attribué* can introduce unnamed sources, as in example 72, giving even more reason for questioning the statement.

Hedges and modals may both appear in a single sentence, attenuating the statement or implication even more than they would individually (as in example 69, which contains both *can* and *potentially*).

In French there is yet another possibility of attenuation, that of using the conditional tense — and in some cases the subjunctive — to imply or accentuate doubt, as seen in examples 71 and 72.

When these techniques are used, the knowledge extracted from the contexts must be carefully evaluated.

6.3.5 Negation

As seen above in example 64, some statements in contexts may be made in negated form. This brings into question their validity for knowledge extraction. Can such contexts still be considered to express the cause relation? Consider these examples:

88) A vector is a harmless, non-disease-**causing** *virus* (such as canarypox or cowpox) or bacteria (like weakened salmonella)...
89) Although binding to the IGF-1 receptor is higher than for regular human insulin (´1.5), it is significantly less than that of IGF-1 itself (more than 1 000 times less) and does not **promote** cell growth in biological assays to any greater extent than human *insulin*.
90) ... la lamivudine ne devrait pas **entraîner** de risque génotoxique chez les patients sous traitement.
91) Il n'y a pas d'évidence que *l'hormone de croissance* **cause** le cancer.

In these examples, *non*- (73), *not* (74), and *ne... pas* (75, 76) directly affect to the cause relation and could cause users of an information extraction tool to question the validity of the information obtained from these contexts. However, in some cases it is equally valid to understand what is *not* caused than what is. In addition, in cases such as examples 73 and 74 the negated pattern in fact affirms a cause relation: they show that viruses generally cause disease and both human and recombinant insulin promote cell growth. For these reasons contexts in which negation was found were counted as valid hits for this research.

6.3.6 Non-contiguity of patterns and concepts

This issue is comparable to the pattern-related issue of non-contiguity described above, except that in this case it is the pattern and one or more of the concepts to which it refers which are separated. It poses similar problems for automatic recognition of patterns and the concepts to which they relate. A system may either not be able to identify the concepts associated with the pattern, or may incorrectly identify concepts as being part of a causal relation when in fact they are part of a clause or phrase that has been inserted between a pattern and the real cause or effect. Consider the following examples:

92) Gilead's expertise has resulted ... products in development to treat diseases **caused** by human *immunodeficiency* virus, hepatitis B virus, herpes simplex virus, human papillomavirus and influenza virus.[37]
93) Aujourd'hui encore, des publicités à l'usage des médecins **incitent à** la *vaccination* des bébés, celle par exemple des laboratoires SmithKline Beecham.
94) La perte de poids chez les personnes infectées par le VIH ou atteintes du SIDA **provient**, non seulement **de** la perte de graisse, mais surtout **de** la diminution de la masse corporelle maigre.

In these examples, the elements separating the causes from their effects could make it very difficult for a semi-automatic knowledge extractor to identify them. In examples 77 and 79, this is at least in part due to more than one cause or effect being listed. However, this is not the only source of the problem. Consider the inaccuracy of a semi-automatic extraction system identifying *doctors*, rather than *advertisements*, as the encouraging force behind the vaccinations in example 78, because it was looking for the plural noun immediately preceding the pattern.

There are also more unique cases, such as this one in which not only is the cause separated from the pattern and effect, but the effect is also interrupted by the pattern:

[37] Although in cases like this one it is likely that a knowledge extraction system would find one or more of the causes of the diseases, the causes at the end of the list (e.g., *human papilloma virus*, *influenza virus*) might not be identified if they appeared too far away from the pattern.

95) Les diabétiques ne peuvent assimiler correctement le sucre (glucides) parce que leur pancréas ne sécrète pas correctement une hormone qui active l'utilisation du glucose dans l'organisme. C'est l'*insuline*. Le patient doit **donc** surveiller constamment son taux de glucide dans le sang et s'injecter régulièrement de l'*insuline*.

It would be extremely difficult for an automatic system to correctly identify complete cause and effect in this case.

In some cases, the cause or effect in a relation is expressed so far away from the other elements that it does not appear at all in the context as extracted by the tool, resulting in a problem very similar to anaphora. Consider these examples:

96) Many others were dubious at first, skeptical, for instance, that the *immunity* **elicited** would be strong enough to spare people from infection by a living pathogen.
97) Without **prophylaxis**, infants born to women whose sera are positive for both the hepatitis B surface antigen and the e antigen have an 85-90% likelihood of being infected...
98) Toutes les études ont mis en évidence des bénéfices d'un traitement *antiviral* précoce. Une étude de **prophylaxie** dans une institution pour personnes âgées n'a pas permis de tirer de conclusions car les effectifs des sujets inclus dans l'étude étaient trop faibles.
99) Et des protéines humaines normales **d'activation** (SP1, TFIID, etc.) ont des sites d'atterrissage bien à elles sur le HIV 1. Ces voies **d'activation** sont redondantes et coopèrent les unes avec les autres : c'est le réseau de communication de la cellule.

In general, non-contiguity is one of the most difficult problems an automatic system will face. Lacking human reasoning skills, the computer relies solely on algorithms and parsing for locating concepts, and variation from the norm can interfere with the analysis.

6.3.7 Complexity of concepts

Much of current research into automatic knowledge extraction (Davidson 1998, Morgan 2000, Rebeyrolle 2000, Marshman, Morgan and Meyer 2002) has focused on the relations of hyperonymy and hyponymy, as well as meronymy. These are excellent relations to use for developing computer tools because in general the concepts linked by them tend to be relatively simple. With the cause relation, however, things are often

much more complex. While there are certainly some very simple expressions of cause, such as "HIV causes AIDS," there are also many complex ones, such as the following:

100) **Because** *human growth hormone* may induce a state of insulin resistance, patients should be observed for evidence of glucose intolerance.

101) Untreated hypothyroidism prevents optimal response to Protropin. **Therefore**, patients should have periodic thyroid function tests and should be treated with thyroid hormone when indicated.

102) L'administration d'une suspension orale de charbon activé en poudre ou de colestyramine **entraîne** une augmentation rapide de la clairance plasmatique de l'A771726.

103) Les employeurs et leurs employés devraient songer à se faire *vacciner*, **car** il a été établi que la vaccination annuelle contre la grippe des travailleurs adultes en bonne santé contribuait à réduire l'absentéisme associé à des maladies respiratoires et à d'autres troubles

Given the complexity of these types of concepts, it is unlikely that a semi-automatic extraction system could hope to identify them completely and correctly. While there are some commonalities, for example, the frequency of complex concepts (in these cases, usually the entire clause) being linked by conjunctions such as *because*, *therefore*, *thus*, *par conséquent*, and *car*, other examples such as 87, containing *entraîner*, would be unexpected. Moreover, many simpler concepts are also linked by conjunctions.

6.3.8 Writing problems

Several different aspects of writing may lead to difficulties in corpus-based research using patterns. Often the most obvious are spelling and grammar mistakes:

104) Le traitement par l'*hormone de croissance* recombinante chez les patients de petite taille **suite à** un déficit en *hormone de croissance* est nécessaire et indispensable afin d'accélerer [sic] la vitesse de croissance de ces patients de façpn [sic] à optimiser leur croissance...

105) La libération de *l'insuline* topique et non-invasive, en utilisant les Transfersomes® comme vecteurs appropriés, suggére [sic] une potentialité thérapeutique.

106) On s'est demande [sic] si le lymphome lymphoblastique qui l'a tue [sic] était, ou non, **imputable à** la manipulation de bactéries ou de *virus* modifies génétiquement...

These may call the reliability of a source into question. While spelling and grammar are certainly not indicators of scientific brilliance, they do indicate the level of care taken

when preparing the document. In addition, when spelling mistakes occur in patterns, the contexts may not be identified.

Another problem is the occasional non-standard use of a pattern in a certain context. For example, in the sentence

> 107) Le traitement substitutif par hormone de croissance est le seul recours pour **améliorer** la taille des enfants ayant un déficit somatotrope avec un bon rapport efficacité/tolérance.

améliorer, which is usually qualitative, is used in a quantitative context.

Another issue is the different ways in which authors view relations. Consider the different interpretations found in these sentences:

> 108) Les différents traitements contre les infections opportunistes **provoquées** par l'effondrement du système immunitaire étaient initialement prescrits...
>
> 109) Plusieurs paramètres ont été étudiés en permanence, en particulier l'apparition d'infections opportunistes - celles qui **profitent de** la déficience immunitaire du malade pour se déclarer...

In the first sentence it is the weakness of the immune system that causes the infections, while in the other (the more widely accepted version) this weakness simply allows the infections to take hold. While the end result — illness — is the same, the type of cause relation is quite different (existence dependency of creation versus influence dependency of increase).

Finally in this section, there is an undeniable tendency among authors — even in what are usually considered to be "dry" scientific texts — to be creative. While not a writing problem *per se*, this phenomenon does pose challenges for knowledge extraction.

The use of tropes, metaphor and simile in scientific writing has been well documented (e.g., Halloran and Bradford 1984). The popularity of imagery is due in part to the necessity of explaining complex scientific concepts to the layperson in popularized scientific writing. Consider these examples:

110) The rapid growth in the number of TAPET organisms within solid tumors appears to occur **because** the organisms "feed" on components of DNA and proteins found in tumors...

111) *Immunity* is achieved when such activity generates long-lasting "memory" cells--the sentries that stand ready to **stop** the pathogen **from** causing disease.

112) Or, des anticorps ont bien été identifiés dans 50 % des cancers les plus fréquents, mais la plupart d'entre eux ne sont pas suffisamment spécifiques pour qu'on soit sur de ne pas déposer ces détonateurs biologiques au hasard dans l'organisme et d'**endommager** des cellules saines.

113) Ces micro-organismes sont terriblement rusés. Dès qu'ils pénètrent dans l'organisme, les plus roués s'efforcent de **perturber** les communications entre les cellules *immunitaires*.

Since of course a machine cannot pass judgment on whether an item is metaphorical or not (although in the English examples given here, the quotation marks could give it a hint), it could very likely extract from the corpus the information that organisms' feeding on DNA causes rapid growth; sentries prevent disease; biological detonators damage cells; and sneaky, malicious micro-organisms disrupt cell communications. This information is not very useful for technical purposes.

6.3.9 Repetition

It is generally accepted that there is a considerable duplication of information on the Internet. In long documents, or documents which are divided between several pages, sentences and paragraphs are often repeated word-for-word, or nearly so. Portions of text — sometimes even entire texts — are reproduced on several pages, and pages from one site can be linked to another in order to facilitate access to information. This poses a challenge for corpus-builders, since it is difficult to be entirely certain that the same texts are not included more than once. While this is not an issue that generally affects the

quality of information retrieved by a tool, it does certainly affect the validity of statistical measures of numbers of occurrences of elements.[38]

In this analysis of the corpora, care was taken to note the origin of similar sentences, and to exclude any duplicate texts that were found. However, repetitions of sentences within a single text were included, as they often occurred in summaries or conclusions. In addition, if texts shared very similar or identical passages but appeared to be significantly different in other sections, both were generally retained.

6.4 Language-related issues

These issues are related expressly to an aspect of one particular language or to the interaction between the languages. They include the avoidance of repetition in French and false cognates (*faux-amis*) which can lead to confusion in interpretation.

6.4.1 Avoidance of repetition in French

As discussed in Chapter 5 and above in the section on anaphora, it is generally accepted that there is even more of a tendency in French than in English to avoid repetition of lexical elements where possible. This increased variation can lead to the use of hypernyms (for example, *hormone* instead of *insuline*), generic terms (for example, *réaction*, *méthode*, *phénomène*), or pronouns (for example, *il*, *cela*) instead of more exact terms. All of these substitutions can increase the difficulty and decrease the accuracy of semi-automatic knowledge extraction.

---

[38] One solution to this problem would be for the tools' programmers to include a function that would eliminate repetition in the output. This would likely involve a relatively modest investment of time and financial resources.

6.4.2 False cognates

Some of the possible patterns identified in this research in some cases are considered to be false cognates (*faux-amis*). As a result, the validity of contexts extracted using these patterns can be difficult to interpret. Two examples of this type of pattern are *contrôler*, which may be translated as *to control* or *to monitor*, and *suite à* or *suite de*, which could be translated as either *because of* or *after*. Below are some examples of contexts in which these patterns do indicate cause:

114) ... les virus qui sont devenus résistants au zanamivir **suite à** une mutation de type E119G dans le site catalytique de leur neuraminidase ...

115) Une personne dans chacun des deux groupes décéda au cours du traitement, pas à cause de celui-ci, mais **à la suite de** l'infection par le cytomégalovirus (CMV), fréquemment associé au sida.

116) L'effet physiologique de l'*hormone de croissance* dans l'organisme humain est de **contrôler** une croissance normale et linéaire.

However in the cases below they do not:

117) Une fois la cellule infectée, l'ARN viral est transcrit en ADN (acide désoxyribonucléique), grâce à une enzyme, la transcriptase inverse, présente dans le virus. **A la suite de** quoi, l'ADN d'origine viral s'intègre aux chromosomes humains de la cellule hôte, qui se met à fabriquer les divers composants du virus.

118) Le calcium sanguin sera **contrôlé** régulièrement pendant le traitement.

Examples 104 to 106 are some of the most difficult examples to analyze, as it is difficult to tell whether or not the cause relation is present:

119) **À la suite des** « ratés » de l'an dernier, Santé Canada attend ces résultats avant de donner le O.K. à la distribution des 12,2 millions de doses...

120) **À la suite** d'une infection aiguë par le VHB, certains sujets sont incapables de développer une réponse immunitaire qui leur permette d'éliminer le virus...

121) Même en cas d'urgence, il ne faut pas dépasser 150 mg d'EP/min. Il est recommandé d'utiliser un système **contrôlant** le débit de la perfusion.

While it is certainly possible to make an educated guess about the meaning these patterns have in the contexts, either interpretation is plausible.

6.5 Relation-related issues

These are issues that are, if not specific to, at least closely linked to the cause relation. They include issues in defining the cause relation, including classifying various sub-types of cause and differentiating the relation from others.

6.5.1 Defining and delimiting the cause relation

6.5.1.1 Classifying the relation

Given the complexity of the cause relation, it was necessary to settle on a definition and classification of the relation that would cover as many of the occurrences found in the corpora as possible. As described in Chapter 2, the classification of the relation used for this project was taken from Barrière (2001, 2002). It worked very well, but nevertheless some aspects of the research suggested that a review would be helpful.

Among these aspects was the classification of permission (e.g., *allow*, *permit*, *enable*) under the maintenance sub-classification of the existence dependency, which did not seem very intuitive. These patterns might be better grouped under a more precise heading such as "necessary conditions."

6.5.1.2 Re-classifying the relation

Some of the classifications could be further sub-divided. For example, within some of the influence dependency categories the effect is often inherent in the pattern. Consider these examples:

➔ *to strengthen* or *renforcer* is to cause to be stronger

➔ *to slow down* or *ralentir* is to cause to be slower

➔ *to neutralize* or *neutraliser* is to cause to be neutral

Many other examples in the corpus, such as *lethal* and *fatal*, follow this pattern. They are often indicated in Chapters 3 and 4 by the double-underlining of the pattern in the examples. These kinds of intrinsic cause relations were clearly described in Talmy (1985:61,76). It would thus be interesting to be able to group them together in order to more precisely classify the relation.

Another issue for consideration is the point of view taken when classifying patterns of inherent effect. Should examples such as *lethal* and *fatal* be classified as creation or destruction, and on what grounds? They indicate the causing of death or the destruction of life, depending on the point of view.

Unfortunately, this sort of review is far beyond the scope of this project and would be better addressed in more in-depth research.

6.5.1.3 Differentiating the relation from function

Another issue in defining the cause relation was differentiating it from function. It is very common in the biopharmaceutical corpora for the two relations to co-exist. After all, biopharmaceuticals are generally used to *cause* or *prevent* something. This is their *function*. Take for example these sentences:

122) Nasal Spray Vaccine **Prevents** Both the Flu and Flu-related Earaches

123) Les traitements antiviraux tentent de **bloquer** la multiplication du virus en agissant sur les processus clés de sa réplication.

The function of the vaccine is to prevent the flu. The function of the antivirals is to prevent the multiplication of the virus.

The relations are nevertheless not inseparable. Consider the example used by Meyer *et al.* (1997:102) to illustrate the function relation:

124) The function of a fax machine is to send and receive faxes.

It is, however, not accurate to say that:

125) *Fax machines cause faxes to be sent and received.

It is not the machine, but the human agent who causes the actions to occur.

To use more domain-specific examples, consider the following sentences:

126) An early example of the utility of bioinformatics is cathepsin K, an enzyme that might turn out to be an important target for treating osteoporosis, a crippling disease **caused** by the breakdown of bone.

127) ... le virus de l'immunodéficience humaine (VIH) qui **cause** le sida...

We clearly see that there is no function present in these relations, simply cause. The breakdown of bone and the virus are not *intended* to do anything; they simply *do*. This is the clearest way to differentiate between function and cause.

Using this guideline, intended effects of medications generally involve both cause and function; side effects generally involve merely cause. In this research all cases of the cause relation that were observed were used, regardless of whether the function relation was present. The relations were considered to be co-occurring, not mutually exclusive.

6.6 Technology-related Issues

This section will discuss issues related to technology as it exists today and the limits that this technology imposes on the application of this research in natural language.

The belief that automatic knowledge extraction will never be perfect is widely held. It is impossible to teach a computer to faultlessly interpret all of the nuances and variations of natural language, especially when in some cases this is even a problem for humans. Moreover, the inverse variation between precision and recall must be considered: the higher the precision users demand from a tool, the lower the recall of that tool will be. If users restrict the hits they want to see to only the most precise ones,

unusual structures and uses will not be recognized. This may or may not be acceptable to the user, depending on his or her needs.

In addition, the requirements of humans and machines in absorbing information from texts are completely different. As mentioned above in the section on anaphora, humans generally prefer variation and minimal repetition in the texts they read. Computers, on the other hand, would work much better with texts if the same concepts were always referred to in the same words and the same form. Computers would also work better with very simple sentences, while many human readers would find this style tiresome.

This is a fundamental challenge of natural language processing: to make something acceptable for a human, accessible to a machine. These limitations will remain a simple fact of life, at least until there is a revolution in information technology.

6.7 Human-related issues

This section will include descriptions of the human element in the writing of texts and the analysis of the corpora in this project.

6.7.1 Author-related issues

The current inability of computers to handle more complex natural language is, as mentioned above in the section on writing-related issues, due in part to the undeniably creative human spirit. Regardless of the "seriousness" or "dryness" of a topic, authors generally find ways to introduce creative elements in sentence structure and vocabulary.[39]

---

[39] There are of course exceptions to this rule, such as writers producing documents in controlled language.

6.7.2 Researcher-related issues

6.7.2.1 Knowledge-related issues

The general capacities and knowledge of the researcher certainly play a role in this type of research. Given the wide range of knowledge needed in this type of project, any researcher's training may be lacking in one area or another. First, knowledge of terminological theory and research is required to design and carry out the project. Second, knowledge of the subject field is necessary for understanding the texts at the stages of corpus building and corpus analysis. Third, linguistic competence in both French and English is also required at these stages.

Fortunately, in the case of the first and second needs, experts in the fields of terminology and conceptual relations (Ingrid Meyer, Lynne Bowker, Caroline Barrière) and in the field of pharmaceuticals (Joan Marshman), as well as publications in both fields, could be consulted. This helped to provide a stronger theoretical basis for the project and better judgment in analyzing the corpora.

In the case of the third requirement, in most cases it is unavoidable for a person's first language to be stronger than their second (or third, or fourth). This motivates the recommendation that translators work towards their mother tongue where possible. It is certainly possible that because this researcher's first language is English, the analysis in this language could be more precise and accurate than the analysis in French. However, consulting French speakers when in doubt helps to minimize this problem. The ideal solution would of course be to involve a second researcher in each stage of the project. However, given the scope of the project, it was simply not practical to do so. Moreover, even with another researcher differences in competence are possible.

6.7.2.2 Design and application of criteria

Having a single researcher carry out all of the analysis has an impact not only on issues of linguistic competence, but also on issues of the research process. These issues involve subjectivity and consistency. Given that there are nuances of meaning and unclear phrasing in the texts used for the corpora, it was often solely a matter of opinion and judgment whether to classify a particular context as a hit, a 'maybe' or an example of noise. It is impossible to reduce this sort of judgment to totally objective criteria that cover all possibilities. The classification thus relied heavily on one person's opinion.

Moreover, even once the criteria were designed, it was often difficult to apply them consistently. There were innumerable subtle variations that required interpreting the criteria. In addition, given the sheer volume of the concordances examined — and the time period within which this examination was carried out — the criteria may have evolved imperceptibly over time. For example, comparisons between contexts or concordances could have affected the evaluations.

Finally, as humans are not infallible, they have a tendency to overlook the familiar and focus on the different (Rundell and Stock 1992a:13), and to become fatigued or even distracted after a certain period of time. This is why computers make such good tools for repetitive tasks, and in fact is one of the reasons why researchers are searching to develop computer-assisted tools such as knowledge extraction systems.

These factors, while not totally avoidable, were minimized where possible by the efforts to design and apply precise criteria, and by minimizing factors such as fatigue that could interfere with the application of these criteria.

6.8 Conclusions

In this chapter, several different types of issues observed during this research project were described and illustrated. They did pose difficulties for the research and should be taken into account for the evaluation of this project and for future projects.

However, despite the difficulties, we believe that the patterns observed in Chapters 3 and 4 remain valid ones for future study. This conclusion and possible future research will be discussed further in Chapter 7.

7 Conclusions and Further Research

This chapter will outline the conclusions reached through this research and present some suggestions for future work.

7.1 English knowledge patterns

The analysis of the English corpus allowed us to identify potential knowledge patterns indicating the various types of cause. Many of these patterns showed high precision in this corpus, and are promising subjects for future research.

7.2 French knowledge patterns

The analysis of the French corpus produced similar results. Knowledge patterns for each of the types of cause were identified, many of which showed high precision in the corpus and should also be interesting subjects for future research.

7.3 Issues of pattern identification and use

Chapter 6 categorized and discussed some of the issues observed in identifying the patterns, and others that are likely to become apparent in applying them. These issues were linked to the patterns themselves, the texts in which they occurred, the language, the cause relation, and human factors.

7.4 Interlinguistic comparison

The preliminary comparison between the English and French patterns in Chapter 5 allowed us to take a brief look at how the patterns in these two languages may be similar and different. Many similarities were observed.

7.5 Summary

This thesis has met its major objectives in contributing to the body of knowledge about knowledge patterns in general and those associated with the cause relation in particular, examining the precision of the patterns identified for the French and English corpora, and discussing issues observed in the identification of the patterns. The secondary objective, a brief comparison of the patterns in each language, was also met, although there is much work still to do in this area.

The research has produced results that can now be tested in other domains, refined, and applied by programmers in order to develop trial knowledge extraction systems. Once integrated into a system, the patterns can be further tested and refined in order to increase their efficiency for knowledge extraction. Furthermore, the issues of pattern identification and application observed and the interlinguistic comparison of the patterns will assist other researchers in similar projects. Finally, this work will contribute to research on the cause relation by illustrating some of the specific types of the relation and providing data on which to base a revised and expanded classification.

7.6 Suggestions for further research

In this section, we suggest some research projects that would build on or complement this research. They will be divided into two sections: research on knowledge patterns and research on the relation.

7.6.1 Research on knowledge patterns

Now that a selection of promising patterns has been identified, there are several possibilities for further research.

Since the patterns were evaluated solely on the basis of *precision*, there is still a need to evaluate the patterns' *recall* potential. This should ideally be done on the same corpora. Subsequently, the recall and precision of the patterns should be evaluated on other corpora in different domains, in order to gather more data on their effectiveness for knowledge extraction. This testing on other corpora will also allow researchers to identify patterns that are domain-specific.

Another step in this research would be refining the forms of the patterns used. It is almost certain that greater precision or recall could be achieved by modifying the forms used for many of the patterns.

Thirdly, there remains much work to be done on optimizing the search windows to be used for each pattern. As the size of the search window can have enormous impact on the recall and precision of knowledge extraction systems, it will be necessary to determine where the patterns are likely to occur in relation to the terms, in order to identify occurrences of the relation as completely and precisely as possible.

Fourthly, identifying where the patterns and associated concepts occur relative to each other could enable knowledge extraction systems to automatically extract the concepts that are linked in the relation, in order to assist the user even more.

Finally, as this list of patterns constitutes only the beginning, further research should be done to identify patterns — lexical, grammatical or paralinguistic — which have not yet been observed in this domain, as well as those used in other domains. (For example, given the lack of precise knowledge patterns found for the influence dependency of preservation, more research should be done on this sub-category of the relation in order to determine whether this is characteristic of the domain or of the sub-

relation itself, or if it is simply linked to this particular corpus.) Once identified, the patterns discovered should be studied as indicated above.

7.6.2 Research on the relation

Another development that would be extremely useful for knowledge extraction would be the automatic classification of contexts according to the type of sub-relation each one represents. This could allow users to automatically extract very precise information about causal relations. In order to achieve this precise classification, however, each of the patterns must first be associated with the sub-relation it most frequently indicates in a given structure. For some this is straightforward, but for the more polysemous patterns it will constitute a challenge.

One necessary step towards this goal would be to revisit the classification of the cause relation as developed by Talmy, Garcia, Nupponen, Barrière and others. As the adequacy of a classification is determined by its ability to deal with the vast majority of the occurrences found in corpora (Barrière 2002), the adequacy of each of the systems should be evaluated. The ideal classification would likely merge elements of several of the proposed classifications in order to create an adequate — and hopefully intuitive — model on which to base future work.

Works cited

Linguistics

AHMAD, K. and H. FULFORD. (1992). "Knowledge Processing: 4. Semantic Relations and their Use in Elaborating Terminology" (Computing Sciences Report CS-92-07). Guildford: University of Surrey.

AHMAD, K. and M. ROGERS. (1992). "Terminology Management: A Corpus-Based Approach." In *Translating and the Computer 14: Quality, Standards and the Implementation of Technology*. London: ASLIB.

ATKINS, A., J. CLEAR and N. OSTLER. (1992). "Corpus Design Criteria." In *Literary and Linguistic Computing*, 7(1). Oxford: Oxford University Press. 1-16.

BARRIÈRE, C. (2001). "Causal Links to Semi-Technical Texts: Discovery, Classification and Representation." *Terminology,* 7(2).

BARRIÈRE, C. (2002). "Hierarchical Refinement and Representation of the Causal Relation." *Terminology*, 8(1).

BOWDEN, P.R., P. HALSTEAD, and T.G. ROSE. (1996). "Extracting Conceptual Knowledge from Text Using Explicit Relation Markers." *Advances in Knowledge Acquisition*, Eds. N. Shadbolt, K. O'Hara and G. Schreiber. Proceedings of the 9th European Knowledge Acquisition Workshop, EKAW'96, Nottingham, U.K., May 1996. 147-162.

CHAFFIN, R. and D.J. HERRMAN. (1988). "The nature of semantic relations: a comparison of two approaches." *Relational Models of the Lexicon*, Ed. M. Evens. Cambridge/New York: Cambridge University Press, 289-334.

CHEVRILLON, A. (1921). *Trois études de littérature anglaise*. Paris : Plon. p. 222.

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

CONDAMINES, A. and REBEYROLLE J. (1998). "CTKB: A Corpus-based Approach to a Terminological Knowledge Base." In *Computerm '98: First Workshop on Computational Terminology*. Eds. D. Bourigault, C. Jacquemin, and M.-C. L'Homme. Proceedings of the workshop, COLING-ACL '98, Montreal, Canada, August 1998. Montreal: Université de Montréal, 29-35.

CONDAMINES, A. and J. REBEYROLLE. (2001). "Searching for and identifying conceptual relationships via a corpus-based approach to a Terminological Knowledge Base (CKTB): Method and Results." *Recent Advances in Computational Terminology*. Eds. D. Bourigault, C. Jacquemin and M.-C. L'Homme. Amsterdam/Philadelphia: John Benjamins. 127-148.

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

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

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

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

ENGWALL, G. (1994). "Not Chance but Choice: Criteria in Corpus Creation." *Computational Approaches to the Lexicon*. Eds. B.T.S. Atkins and A. Zampolli. Oxford: Oxford University Press. 49-82.

GARCIA, D. (1996). « COATIS, un outil d'aide à l'acquisition des connaissances causales exprimées dans les textes. » *Acte du Colloque Linguistique et Informatique de Montréal, CLIM'9*6. 97-103.

GARCIA, D. (1997). « Structuration du lexique de la causalité et réalisation d'un outil d'aide au repérage de l'action dans les textes. » *Actes des deuxièmes rencontres — Terminologie et Intelligence Artificielle, TIA '97*. 7-26.

HALLORAN M. and A.N. BRADFORD. (1984). "Figures of Speech in the Rhetoric of Science and Technology." *Essays on Classical Rhetoric and Modern Discourse*. Eds. R.J. Connors *et al*. Carbondale and Edwardsville: Southern Illinois University Press. 179-192.

HAMON, T. and A. NAZARENKO. (2001). "Detection of synonymy links between terms." *Recent Advances in Computational Terminology.* Eds. D. Bourigault, C. Jacquemin and M.-C. L'Homme. Amsterdam/Philadelphia: John Benjamins.

HUME, D. (1739/1965). *A Treatise of Human Nature.* Ed. L.S. Selby Bigge. Oxford: Clarendon Press.

JING, H. and E. TZOUKERMANN. (2001). "Determining semantic equivalence of terms in information retrieval." *Recent Advances in Computational Terminology.* Eds. D.

Bourigault, C. Jacquemin and M.-C. L'Homme. Amsterdam/Philadelphia: John Benjamins.

JOUIS, C. (1994). "Contextual approach: SEEK, a linguistic and computational tool for use in knowledge acquisition." In *Proceedings of the first European Conference Cognitive Science in Industry*, Luxembourg, September 28-30, 1994. 259-274.

KAVANAGH, J. (1995). "The Text Analyzer: A Tool for Extracting Knowledge from Text." Unpublished MSc Thesis. Department of Computer Science, University of Ottawa, Ottawa, Canada. [http://www.csi.uottawa.ca/~kavanagh/Thesis/thesisAbstract.html].

L'HOMME, M.-C. (2001). « Nouvelles technologies et recherche terminologique : techniques d'extraction des données terminologiques et leur impact sur le travail du terminographe. » Proceedings of the symposium "The Impact of New Technology on Terminology Management," Glendon College, York University, Toronto, Ontario, August 18, 2001.

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

MARSHMAN, E.,T. MORGAN and I. MEYER. (2002). "French patterns for expressing concept relations." *Terminology,* 8(1). 1-29.

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

MEYER, I. and K. MACKINTOSH. (1996). "The Corpus from a Terminographer's Viewpoint." *International Journal of Corpus Linguistics,* 1(2). Amsterdam/Philadelphia: John Benjamins. p. 4.

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

MEYER, I., K. MACKINTOSH, C. BARRIÈRE and T. MORGAN. (1999). "Conceptual Sampling for Terminographical Corpus Analysis." *Proceedings of the Fifth International Congress on Terminology and Knowledge Engineering (TKE '99).* 256-267.

MEYER, I. (2001). "Extracting Knowledge-Rich Contexts for Terminography: A Conceptual and Methodological Framework." *Recent Advances in Computational Terminology*. Eds. D. Bourigault, C. Jacquemin and M.-C. L'Homme. Amsterdam/Philadelphia: John Benjamins. 279-302.

MEYER, I. (2002). Personal communications.

MORGAN, T. (2000). "A Comparative Study of Hypernymic Patterns for Knowledge Extraction." Master's Thesis. School of Translation and Interpretation, University of Ottawa.

MORIN, E. (1999). « Acquisition de patrons lexico-syntaxiques caractéristiques d'une relation sémantique. » *Traitement automatique des langues (TAL)*, 40(1). Paris: Université de Paris VII. 143-166.

NUPPONEN, A. (1994). "Causal Relations in Terminological Knowledge Representation." *Terminology Science and Research*, 5(1). 36-44.

PAVEL, S. and D. NOLET. (2001). *The Handbook of Terminology*. Adapted into English by C. LEONHARDT. Ottawa: Minister of Public Works and Government Services Canada.

PEARSON, J. (1996). "The Expression of Definitions in Specialised Texts: A Corpus-based Analysis." *Euralex '96 Proceedings, Part II.* Papers submitted to the Seventh Euralex International Congress on Lexicography in Göteborg, Sweden. Eds. M. Gellerstam et al. Göteborg: Göteborg University, Department of Swedish. 759-769.

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

PEARSON, J. **(**1999). « Comment accéder aux éléments définitoires dans les textes spécialisés ? » *Terminologies nouvelles*, 19(1). 21-28.

QUILLARD, G. (1997). « Étude de certaines différences dans l'organisation collective des textes pragmatiques anglais et français. » *Babel*, 43(4). 313-330.

QUILLARD, G. (2001). « To be or not to be / exister ou ne pas exister. L'autocensure de d'être et avoir. » Presented at the 2001 Congress of the Canadian Association for Translation Studies, Laval University, Laval, Quebec, May 26-28, 2001.

REBEYROLLE, J. (2000). « Forme et fonction de la définition en discours. » Thèse de doctorat. Université de Toulouse II.

ROGERS, M. and K. AHMAD. (1994). "Computerised Terminology for Translators: the role of text." *Applications and Implications of Current LSP Research*, vol. II. Eds. M. Brecke, Ø. Andersen, T. Dahl and J. Myking. 840-860.

ROUSSELOT, F., P. FRATH, and R. OUESLATI. (1996). "Extracting Concepts and Relations from Corpora." In *Proceedings of the 12th European Conference on Artificial Intelligence ECAI'96.*

RUNDELL, M. and P. STOCK. (1992a). "The Corpus Revolution." English Today, 30. 9-14.

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

SÉGUÉLA, P. (1999). « Adaptation semi-automatique d'une base de marqueurs de relations sémantiques sur des corpus spécialisés. » *Terminologies nouvelles,* 19(1). 52-60.

SINCLAIR, J. (1985). "Lexicographic Evidence." *Dictionaries, Lexicography and Language Learning.* Ed. R. Ilson. Oxford: Pergamon Press. 81-94.

SINCLAIR, J. (1994). "Corpus Typology: A Framework for Classification." EAGLES document. 1-18. *Studies in Anglistics*. Eds. G. Melchers and B. Warren. (1995). Stockholm: Almquist and Wiksell International. 17-34.

SKUCE, D. and KAVANAGH, J. (1999). "A Document-Oriented Knowledge Management System." *Proceedings of the Fifth International Congress on Terminology and Knowledge Engineering (TKE '99)*. 320-329.

TALMY, L. (1985). "Lexicalization patterns: semantic structure in lexical forms." *Language Typology and Syntactic Description*, *Vol. III: Grammatical Categories and Lexicon*. Ed. T. Shopen. Cambridge: Cambridge University Press.

TALMY, L. (1988). "Force Dynamics in Language and Cognition." *Cognitive Science,* 12. 49-100.

TALMY, L. (2000). *Toward a Cognitive Semantics*. Vol. 1 and 2. Cambridge, MA: MIT Press.

TERMINOLOGY AND STANDARDIZATION DIRECTORATE, Translation Bureau, Public Works and Government Services Canada. (2002). *TERMIUM® Plus*. [http://www.termiumplus.translationbureau.gc.ca/site/english/welcome.html]. Visited : April 23, 2002.

TERMINOLOGY DIRECTORATE, Translation Bureau, Secretary of State of Canada. (1983). *Vocabulary of Terminology*, New Edition. Ottawa: Secretary of State, Translation Bureau.

VINAY, J.-P. and J. DARBELNET. (1977). *Stylistique comparée du français et de l'anglais*, Nouvelle édition revue et corrigée. Laval : Beauchemin.

VINAY, J.-P. and J. DARBELNET. (1995). *Comparative Stylistics of French and English.* Trans. and edited by J.C. SAGER and M.-J. HAMEL. Amsterdam/Paris : John Benjamins.

WÜSTER, E. (1981). "L'étude scientifique générale de la terminologie, zone frontalière entre la linguistique, la logique, l'ontologie, l'informatique et les sciences des choses." Trans. Bureau des traductions, Secrétariat de l'État du Canada. *Textes choisis de terminologie.* Eds. G. RONDEAU and H. FELBER. Laval: GIRSTERM.

Biopharmaceuticals

NATIONAL BIOTECHNOLOGY ADVISORY COMMITTEE. (1984). *Brief to the Commission of Inquiry on the Pharmaceutical Industry*. Ottawa: October 9, 1984. Appendix 2-7.

Works Consulted

Linguistics

AUSSENAC-GILLES, N. (1999). "Géditerm : un logiciel pour gérer des bases de connaissances terminologiques." *Terminologies nouvelles,* 19(1). 111-123.

BARRIÈRE, C. (1996). "Including certainty terms into a knowledge base of conceptual graphs," *Actes du Colloque Linguistique et Informatique de Montréal*, CLIM'96. 184-191.

CABRÉ, M.-T., J. MOREL AND C. TEBÉ. (1996). "Las relaciones conceptuales de tipo causal: un caso práctico". *Actas del V Simposio Iberamericano de terminologie RITerm.* Mexico City, November 3-8, 1996. [http://www.unilat.org/dtil/MEXICO/cabremt.html]. Last visited July 18, 2002.

CIMINO, J.J. (2001). "Knowledge-based terminology management in medicine." *Recent Advances in Computational Terminology*, Eds. D. Bourigault, C. Jacquemin and M.-C. L'Homme. Amsterdam/Philadelphia: John Benjamins. 111-126.

DUBUC, R. (1992). *Manuel pratique de terminologie.* Brossard (Quebec): Linguatech.

FLOWERDEW, J. (1992a). "Definitions in Science Lectures." *Applied Linguistics*, 13(2).

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

IRIS, M.A., B.E. LITOWITZ and M. EVENS. (1988). "Problems of the part-whole relation." *Relational Models of the Lexicon*. Ed. M. Evens. Cambridge/New York: Cambridge University Press. 261-288.

LEVIN, B. (1993). *English Verb Classes and Alternations, A Preliminary Investigation*. Chicago: University of Chicago Press.

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

RONDEAU, G. (1984). *Introduction à la terminologie*. 2e édition. Chicoutimi (Quebec): Gaëtan Morin.

RUNDELL, M. and P. STOCK. (1992b). "The Corpus Revolution." English Today, 31. 21-32.

RUNDELL, M. and P. STOCK. (1992c). "The Corpus Revolution." English Today, 32. 45-51.

SAGER, J. (1994). "Terminology: Custodian of knowledge and means of knowledge transfer." *Terminology,* 1(1). 7-15.

SOWA, J.F. (1984). *Conceptual Structures: Information Processing in Mind and Machine*. Reading, MA: Addison-Wesley.

Biopharmaceuticals

BAINS, W. (1993). *Biotechnology from A to Z*. Oxford: Oxford University Press.

BROWN, C.M., I. CAMPBELL and F.G. PRIEST. (1987). *Introduction to Biotechnology*. Basic Microbiology Series, vol. 10. Oxford: Blackwell Scientific Publications. 118-128.

CHAMBERLAIN, A. "Biotech firms told to confront consumer fears." *Toronto Star*, August 31, 1993.

DIVISION OF BIOPHARMACEUTICS AND PHARMACOKINETICS, Department of Pharmacy, Biocenter. Helsinki, Finland. [http://www.biocenter.helsinki.fi/ml/farml/biopharmacy.html]. Visited November 11, 2000.

GENENTECH INC. "Press Release 1989: Genentech Expands CD4 Clinical Trials For AIDS." [http://www.gene.com/news/1989/19890606_000000.html]. Visited November 13, 2000.

INSTITUTE FOR APPLIED BIOMEDICINE. "Non-Technical Summary." [http://www.appliedbiomed.org/drug/summary.html]. Visited November 12, 2000.

MARSHMAN, J.A. (2000). Personal Communications. November 13-14, 2000.

MORRIS, B. (1995). *Biotechnology*. Melbourne: Cambridge University Press. 21-35.

PAPP, L. "Protein find may help fight bowel disease." *Toronto Star*, July 23, 1996.

PAPP, L. "Hormone holds hope for repairing intestine." *Toronto Star*, June 23, 1997.

"Playing Tricks with living organisms." *Toronto Star*, October 18, 1992.

REID, J. "Research firm develops drug said to forestall AIDS deaths." *Toronto Star*, May 7, 1992.

SLOAN-KETTERING INSTITUTE. "Targeted-dye trace cancer imaging during prostatectomy." [http://www.ski.edu/project_summary.cfm?LAB=205&PROJECT=358&STARTROW=2&#POS_6]. Visited November 12, 2000.

SMITH, M. "Biotech firm testing safety of anti-HIV drug." *Toronto Star*, February 3, 1994.

Appendix A: Statistics for English Patterns

Organization of Table 1

The figures in Table 1 represent two calculations of precision: one in which only the 'hits' (the clearest, most complete examples of cause relations) were counted, and the other in which the 'maybes' (the less certain examples) were also included.

The table lists the patterns that showed more than 50% precision in this corpus. (For those that were less than 50% precise, see Appendix C.) In cases where the patterns were refined during the course of the research, each version of the pattern is listed separately.

Table 1: Statistics for Causal Patterns in the English Biopharmaceutical Corpus

Pattern	Total Occur-rences	Hits	Maybes	Noise	Precision	Precision (%) incl. maybes
Existence Dependency						
Creation						
*caus**	212	212	0	0	100	100
because	188	188	0	0	100	100
*elicit**	48	48	0	0	100	100
*ensur**	15	15	0	0	100	100
secondary to	7	7	0	0	100	100
aris from*	4	4	0	0	100	100
provok/provoc**	3	3	0	0	100	100
*spur**	3	3	0	0	100	100
adverse drug event/adverse drug reaction**	3	3	0	0	100	100
culminat in*	2	2	0	0	100	100
*incit**	1	1	0	0	100	100
prod/prods/prodding	1	1	0	0	100	100
give/gives/gave/ giving rise to	1	1	0	0	100	100
due /5R *to*	87	86	0	1	99	99
lead/led* /5R *to*	77	75	0	2	97	97
*attrib**	12	11	0	1	92	92
so that	23	21	2	0	91	100
*confer** not	10	9	0	1	90	90

Pattern	Total Occur-rences	Hits	Maybes	Noise	Precision	Precision (%) incl. maybes
conference/conferences						
*induc**	207	184	19	4	89	98
**genic/*genesis/*genicity* not *transgenic, antigenic*	172	149	20	3	87	97
*complication**	24	21	0	3	87	87
consequence/consequent**	26	22	0	4	85	85
*evok**	6	5	0	1	83	83
prime/primes/priming/primed	17	14	0	3	82	82
therefore	76	61	13	2	80	97
secondary /5R *to*	9	7	0	2	78	78
*prompt** not *promptly*	11	8	0	3	73	73
*reason** not *reasonable/ reasonably/reasoned*	25	18	0	7	72	72
*trigger**	69	49	20	0	71	100
thus	83	58	7	18	70	78
*account** /5R *for*	15	10	0	5	67	67
why	9	6	0	3	67	67
*turn** /5R *on*	6	4	0	2	67	67
*result** /5R *in/of/from*	299	196	9	94	66	68
*immuniz**	130	82	48	0	63	100
*produc** not *product/products*	306	189	115	2	62	99
*effect** /5R *of*	175	102	65	8	58	95
stimul	181	96	72	13	53	93
respond/respons** /5R *to*	613	314	275	24	51	96
*effect**	605	311	266	28	51	95
*aris**	14	7	0	7	50	50
*forc**	4	2	0	2	50	50
Destruction						
*lethal**	15	15	0	0	100	100
*deadl**	14	14	0	0	100	100
destroy/destruct**	61	51	10	0	84	100
*fatal**	6	5	1	0	83	100
*kill**	107	83	24	0	78	100
Maintenance						
*enabl**	63	59	3	1	94	98
*allow**	87	62	12	13	71	85
essential for	6	4	0	2	67	67
critical to/for	12	7	0	5	58	58
Prevention						
*prophyla**	26	26	0	0	100	100
*spar**	5	5	0	0	100	100

Pattern	Total Occur-rences	Hits	Maybes	Noise	Precision	Precision (%) incl. maybes
*stop** /10R *from*	5	5	0	0	100	100
*preclud**	2	2	0	0	100	100
*prevent**	192	158	33	1	82	99
*block** [40]	54	37	5	12	69	78
*keep**	11	7	0	4	64	64
*stop**	21	12	0	9	57	57
Influence Dependency						
Modification						
*influenc**	13	13	0	0	100	100
marked by	7	7	0	0	100	100
*affect**	60	57	3	0	95	100
associated /5R *with*	125	106	11	8	85	94
relationship /5R *between*	12	9	0	3	75	75
seen /5 *with/in*	42	30	1	11	71	74
*role**	39	27	3	9	69	77
*alter** not *alternate/ alternating/alternative*	41	27	14	0	66	100
*involv** /5R *in*	45	29	3	13	64	71
*neutraliz**	43	27	15	1	63	98
*control** not *controlled/ controls*	162	86	14	62	53	62
Increase						
*facilitat**	25	25	0	0	100	100
optimiz/optimis** not *optimistic*	21	21	0	0	100	100
amplify/amplification*	17	17	0	0	100	100
expand (transitive verb)	14	14	0	0	100	100
*prolong** not *prolonged*	8	8	0	0	100	100
speed/sped* (verb)	6	6	0	0	100	100
*strengthen**	5	5	0	0	100	100
maximiz/maximis**	4	4	0	0	100	100
*augment**	4	4	0	0	100	100
*lengthen**	4	4	0	0	100	100
*alert** (verb)	2	2	0	0	100	100
*bolster**	2	2	0	0	100	100
*potentiat**	1	1	0	0	100	100
*orchestrat**	1	1	0	0	100	100
favour/favor** not	6	5	0	1	83	83

[40] Indicates preventing an event from occurring or entity from existing and preventing access to something.

Pattern	Total Occur-rences	Hits	Maybes	Noise	Precision	Precision (%) incl. maybes
favo(u)rable /favo(u)rably						
*contribut**	28	22	0	6	79	79
*increas** not *increasingly*	299	233	48	18	78	94
*increas**	309	234	48	27	76	91
*accelerat**	17	13	0	4	76	76
*enhanc**	98	74	22	2	75	98
*help** not *helper*	63	46	0	17	73	73
*activat**	276	198	78	0	72	100
*expand**	21	15	0	6	71	71
*promot**	55	37	14	4	67	98
speed/sped*	11	6	0	5	55	55
*encourag** not *encouraging/encouragingly*	11	6	5	0	54	100
*improv**	161	85	70	6	53	96
*prolong**	17	9	0	8	53	53
*cataly**	94	49	43	2	52	98
*help**	92	46	0	46	50	50
Decrease						
*protect** /5R *against*	52	52	0	0	100	100
minimiz/minimis**	17	17	0	0	100	100
*dimin**	11	11	0	0	100	100
*hamper**	9	9	0	0	100	100
*lessen**	4	4	0	0	100	100
*restrain**	4	4	0	0	100	100
*alleviat**	3	3	0	0	100	100
*endanger**	3	3	0	0	100	100
*curb**	2	2	0	0	100	100
*dampen**	1	1	0	0	100	100
*slow** /5R *down*	1	1	0	0	100	100
*slow** (verb)	7	6	0	1	86	86
suppress	45	36	9	0	80	100
*inhibit**	198	141	48	9	71	95
*impair**	43	29	14	0	67	100

Preservation						
See Appendix C						

Appendix B: Statistics for French Patterns

The presentation of Table 2 is identical to that used in Table 1.

Table 2: Statistics for Causal Patterns in the French Biopharmaceutical Corpus

Pattern	Occur-rences	Hits	Maybes	Noise	Precision	Precision incl. Maybes
Existence Dependency						
Creation						
provoc/provoq**	92	92	0	0	100	100
*parce qu**	50	50	0	0	100	100
grâce à/au	30	30	0	0	100	100
abouti à/au*	22	22	0	0	100	100
*suscit**	13	13	0	0	100	100
de façon à	12	12	0	0	100	100
incit à*	7	7	0	0	100	100
**génèr*/*génicité*	3	3	0	0	100	100
*entraîn**	126	124	2	0	98	100
car	99	96	2	1	97	99
dû/due à/au	27	26	0	1	96	96
par conséquent	17	16	1	0	94	100
confèr/confér** not *conférence**	14	13	1	0	93	100
*à l'origine d**	26	24	0	2	92	92
*déclench**	48	44	4	0	91	100
*engendr**	11	10	1	0	91	100
pourquoi	66	58	6	2	88	97
responsable d*/pour*	77	67	0	10	87	87
*puisqu**	76	66	4	6	87	92
*stimul**	76	60	16	0	86	100
donc	222	189	19	14	85	94
(se) faire + infinitive verb[41]	95	79	15	1	83	99
se tradui/traduct* en/par*	29	24	4	1	83	97
*caus**	150	122	14	14	81	91
*rend** (+adverb) (+	53	43	9	1	81	98

[41] For certain patterns the concordances were sorted manually to enable us to evaluate the precision of pattern forms belonging to a given grammatical category, e.g., in this case, infinitive verbs.

Pattern	Occur-rences	Hits	Maybes	Noise	Precision	Precision incl. Maybes
object) + adjective						
indui/induct**	110	88	22	0	80	100
imput à*	10	8	2	0	80	100
(re)lié à/au*	119	93	9	17	78	86
*conséquence**	26	20	5	1	77	96
(en) raison (d)*	134	100	10	24	75	82
amen/amèn**	6	4	0	2	66	66
attribu (à/au)*	34	21	5	8	62	74
condui/conduct* (à/au)*	73	43	0	30	59	59
réact/réag**	237	126	98	13	53	95
*effet** not *en effet*	588	304	270	14	52	98
crée/créa** not *créatinine*	28	14	8	6	50	79
*manifestation**	16	8	8	0	50	100
*étiologi**	8	4	2	2	50	75
Destruction						
nocif/nocive**	26	26	0	0	100	100
interfér/interfèr** not *interféron*	5	5	0	0	100	100
*mortel**	25	21	0	4	84	84
supprim/*suppress**	41	34	7	0	83	100
*annul**	5	4	0	1	80	80
tue/tué**	62	44	18	0	71	100
*élimin**	123	82	41	0	67	100
détrui/destruc**	95	60	35	0	63	100
Maintenance						
sine qua non	2	2	0	0	100	100
permet/permi**	310	301	7	2	97	99
indispensable (pour/à/au)*	23	16	7	0	70	100
*nécess**	163	91	59	13	56	92
*essentiel** not *essentiellement*	36	19	12	5	53	86
Prevention						
empéch/empêch**	66	62	3	1	94	98
bloqu/blocage**	74	67	7	0	90	100
*enray**	10	9	1	0	90	100
*évit**	40	30	9	1	75	98
*préven**	132	76	53	3	58	98
Influence Dependency						
Modification						
*influenc**	17	17	0	0	100	100

Pattern	Occur-rences	Hits	Maybes	Noise	Precision	Precision incl. Maybes
*neutralis**	19	18	1	0	95	100
(s')impliq dans*	32	30	2	0	94	100
*(jouer) rôle** (+adjective)	90	70	19	1	78	99
dépend d**	54	38	13	3	70	94
modul not *modules*	17	12	5	0	70	100
*agi** not *s'agi**	89	59	19	11	66	88
dépendant	55	34	17	4	62	93
*régul** not *régulier(s)/régulière(s)/régulièrement*	37	23	13	1	62	81
*normalis**	12	6	6	0	50	100
Increase						
*amplifi**	6	6	0	0	100	100
*maximis**	3	3	0	0	100	100
*contribu**	41	39	2	0	95	100
*favoris**	26	24	1	1	92	96
*renforc**	16	14	2	0	88	100
aid (à/au)*	46	36	2	8	78	83
*catalys**	32	24	2	6	75	81
*accél**	24	18	6	0	75	100
*facilit**	13	9	0	4	69	69
*dilat**	3	2	1	0	66	100
*augment**	225	145	48	32	64	86
*amélior**	106	64	42	0	60	100
*optimis** not *optimiste*/optimisme*	14	8	4	2	57	86
*aggrav**	15	8	7	0	53	100
*hauss**	4	2	1	1	50	75
Decrease						
*endommag**	8	8	0	0	100	100
*constrict**	1	1	0	0	100	100
*délétère**	6	5	1	0	83	100
*ralenti**	25	20	5	0	80	100
*diminu**	201	151	37	13	76	94
*intérromp**	15	11	4	0	73	100
abaiss/baiss**	98	60	37	1	61	99
*inhib**	277	166	111	0	60	100
*retard**	50	30	18	2	60	96
*perturb**	25	13	11	1	52	96
*inactiv**	43	22	18	3	51	93
rédui/réduc**	250	125	66	59	50	76

Pattern	Occur-rences	Hits	Maybes	Noise	Precision	Precision incl. Maybes
Preservation						
mainten/maintien** not *maintenant*	40	24	16	0	60	100

Appendix C: Less Precise English Patterns

Introduction

The following are English patterns that showed less than 50% precision in this corpus. They are presented as in Chapter 3 and Appendix A.

Table 3: Statistics for Less Precise Causal Patterns in the English Biopharmaceutical Corpus

Pattern	Total Occur-rences	Hits	Maybes	Noise	Precision	Precision (%) incl. maybes
Existence Dependency						
Creation						
*yield**	31	14	3	14	45	55
so	97	41	10	46	42	53
responsible /5R *for*	30	12	0	18	40	40
effect /5R *of*	200	77	103	20	39	90
*leav**	10	3	0	7	30	30
*obtain**	60	14	0	46	23	23
*generat**	115	21	61	33	18	71
*accomplish**	13	2	0	11	15	15
*initiat**	75	10	0	65	13	13
*direct** not *directly/direction/director*	86	6	53	27	7	69
Destruction						
None						
Maintenance						
None						
Prevention						
*avoid**	32	12	0	20	37	37
*halt**	8	3	0	5	37	37
Influence Dependency						
Modification						
*relat** not *relatively*	146	62	10	74	42	49
*regulat**	136	56	34	46	41	66
*control**	250	93	31	126	37	50
Increase						
*alert**	5	2	0	3	40	40
*exten**	46	11	12	23	24	50
*encourag**	29	7	22	0	24	100

*boost**	62	14	38	10	23	84
*eas** not *easy/easier/easiest/ easily/east/eastern/easd/easl*	9	2	0	7	22	22
*prepar**	89	4	0	85	4	4
*support**	74	3	16	55	4	26
Decrease						
*protect**	153	73	79	1	48	99
*decreas**	136	63	59	14	46	90
*inactivat**	42	18	24	0	43	100
*reduc**	163	70	85	8	43	95
*damag**	31	13	15	3	42	90
*slow** not *slowly*	25	8	0	18	32	32
*limit**	126	16	4	106	13	16
Preservation						
maintain/mainten**	43	11	8	24	26	44
*preserv** not *preservative**	36	7	0	29	19	19
*preserv**	61	3	4	54	5	11

Patterns observed

Existence dependency

Creation

*yield**

Intravenous administration of this drug **yields** a 100% cure rate of bulky, human tumors in SCID mice.

In steps not shown, DNA vaccines also **yield** memory helper T cells needed to support the defensive activities of other memory cells.

so

"Kaposi's sarcoma is also a highly vascularized tumor, **so** it makes sense to try agents that target the vessel system at the disease site," said Siemann.

responsible for

The *vaccine*, expected to be available for limited use by the fall, is expected to stop the spread of E. coli 0157:H7 bacteria, which is also responsible for outbreaks of hamburger disease…

effect* /5R *of

Excessive glucocorticoid therapy will inhibit the growth promoting **effect of** *human growth hormone*.

The quicker glucose-lowering **effect of** Humalog is related to the more rapid absorption rate from subcutaneous tissue.

leav*

Such stimulation **leaves** the immune system prepared to destroy bacteria and viruses whose antigens correspond to the antibodies it has learned to produce.

obtain*

Studies with hepatitis B *vaccine* derived from plasma have shown that a lower response rate (81%) to *vaccine* may be **obtained** if the *vaccine* is administered as a buttock injection.

generat*

Immunization with transition-state analogs **generates** antibodies that are generally highly specific for their designed substrates.

An extensive review of iatrogenic accidents **generated** by *immuno*-prophylactic products can be found in the book published by Sir Graham Wilson in 1966 ("The Hazard of Immunization").

To determine the functional relevance of the *immunity* **generated** with the targeted vaccine, BALB/c mice were *immunized* with 100 mg of DNA and challenged 9 weeks later with infectious virus.

accomplish*

First, selective prodrug activation requires the *catalysis* of a reaction that must not be **accomplished** by endogenous enzymes in the blood or normal tissue of the patient.

initiat*

Theoretically, a MoAb designed for a particular antigen on cancer cells can **initiate** an *immune* response that would destroy cancer cells without harming normal cells.

direct* **not** ***directly/direction/director***

Onyvax is an early stage biotechnology company focussed on the research and development of vaccines and other therapies which will **direct** the *immune* system to attack cancer cells.

Destruction

None.

Maintenance

None.

Prevention

avoid

Humulin is synthetic human *insulin* which **avoids** many of the problems associated with extracted animal insulin used until its invention.

*halt**

… a neutralizing *immune* response… prevents the autoimmune disease cascade by specifically blocking the presentation of self-antigen by disease-linked MHC molecules to autoreactive T cells that cause rheumatoid arthritis, potentially **halting** the progression of the disease.

Influence Dependency

Modification

relat not relatively*

It is uncertain whether this increased risk is **related to** the pathology of *growth hormone* deficiency itself, growth hormone therapy, or other associated treatments such as radiation therapy for intracranial tumors.

No serious *vaccine*-**related** adverse events were reported.

*regulat**

Growth hormone is a naturally occurring human protein produced by the pituitary gland at the base of the brain. It **regulates** certain aspects of metabolism and is the primary hormone promoting normal growth of bones and tissues.

Two recombinant growth factors (cytokines that ***regulate*** cell division) are undergoing major clinical trials…

*control**

See Chapter 3, section 3.2.2.1.11.

Increase

*alert**

See Chapter 3, section 3.2.2.2.11.

*extend**

It's the change in solubility that prolongs absorption and **extends** its tail of action.

Would finding a way to **extend** plasmid survival lead to stronger *immunity*, or would it backfire and encourage attacks against unvaccinated, healthy tissue?

encourag*

See Chapter 3, section 3.2.2.2.26.

boost*

Interferons are a family of proteins that activate various *immune* functions in the body and exhibit anti-viral and anti-proliferative activities, in addition to **boosting** natural killer cells (the body's surveillance or patrol cells) activity and other effects.

… the *vaccine* can "**boost** the body's own *immune* system" to prevent progressive damage from spinal cord injury.

prepar*

All three of these proteins, in turn, induce specific *immune* responses, thus better **preparing** the body to recognize and attack the virus.

support*

In steps not shown, DNA *vaccines* also yield memory helper T cells needed to **support** the defensive activities of other memory cells.

Decrease

protect*

Researchers from the University of California system and the National Cancer Institute report in the August 2001 issue of the Journal of Virology a poliovirus-based *vaccine* that **protects** monkeys from a highly virulent strain of simian immunodeficiency virus (SIV).

decreas*

In growth hormone deficient patients, long-term growth hormone administration often **decreases** body fat.

Protropin therapy may **decrease** glucose tolerance. Administration of Protropin to normal adults and patients who lacked adequate secretion of endogenous growth hormone resulted in increases in mean serum fasting and postprandial insulin levels.

inactivat*

The other portion ensures that the resulting antibodies will recognise and **inactivate** CETP.

Genelabs Technologies, Inc. is a biopharmaceutical company engaged in the discovery and development of a new class of pharmaceutical products that selectively *regulate* gene expression or **inactivate** pathogens by binding directly to DNA or RNA, the fundamental material of genes.

reduc*

A genetically engineered vaccine that was tested in kidney dialysis patients, who have a high incidence of staph infections in the bloodstream was found to **reduce** infections by 57 percent.

… data from a Phase I/II study demonstrating that the antiretroviral PMPA significantly **reduced** human *immunodeficiency* virus (HIV) RNA levels by greater than 90% after eight doses, with no significant side effects.

Six months of twice-weekly injections **reduced** the virus to undetectable levels in 40 percent of the Taiwanese subjects who received it, an effect that lasted at least 18 months.

damag*

If left untreated diabetes mellitus can **damage** the kidneys, eyes, heart, and limbs, and can endanger pregnancy.

slow* **not s*lowly***

See Chapter 3, section 3.2.2.3.12.

eas* **not** ***easy/easier/easiest/easily/east/eastern/easd/easl***

The additional ingredient also allows patients to reduce weekly injections from three to one, he says, which may **ease** the flulike side effects considerably.

limit*

Because it responds only to specific antigens, the MoAb component **limits** the toxic effects of the *immunotoxin* to target (e.g., tumor) cells.

Preservation

maintain*/mainten*

For the treatment of patients with diabetes mellitus who require *insulin* for the **maintenance** of normal glucose homeostasis.

Shake well before use. Thorough agitation at the time of administration is necessary to **maintain** suspension of the *vaccine*.

preserv* **not** ***preservative****

Five functional classes of adjuvant are described. These exploit mechanisms which a) create an antigen depot, b) **preserve** antigen conformation, c) direct antigen to specific *immune* cells, d) induce mucosal responses …

preserv*

See section 1.2.2.4.2 above.

Appendix D: Less Precise French Patterns

Introduction

This appendix lists the French patterns that showed less than 50% precision in this corpus. The format of the presentation is identical to that used in Chapter 4 and Appendix B.

Table 4: Statistics for Less Precise Causal Patterns in the French Biopharmaceutical Corpus

French Causal Patterns						
Pattern	**Occur-rences**	**Hits**	**Maybes**	**Noise**	**Precision**	**Precision incl. Maybes**
Existence Dependency						
Creation						
*suite à/au/d**	61	26	12	23	43	63
*résult** /5L or /5R *en*	44	18	0	26	41	41
produi/product**	611	246	170	195	40	68
*action**	143	55	78	10	38	93
gèn*/*gén not *gène/gènes/général*/ génétique*/antigène*/ antigéniq*/génie/ génique*/génom**	397	141	11	245	35	38
obten/obtien**	169	60	17	92	35	46
*immunis**	124	43	71	10	35	90
men/mèn* à*	31	11	1	19	35	39
*symptôm**	109	31	78	0	28	100
*résult**	250	59	52	139	24	44
proven/provien* d**	37	7	2	28	19	24
Destruction						
None						
Maintenance						
*requis**	13	4	9	0	31	100
Prevention						
protég/protèg*/ protection* (contre)*	142	60	79	3	42	98
défend/défense**	64	27	35	2	42	97

prophyla	59	16	43	0	27	100
Influence Dependency						
Modification						
*agi**	151	62	19	70	41	54
associ à/au*	214	82	21	63	38	48
*modifi**	142	51	91	0	36	100
altèr/altér**	31	11	20	0	35	100
*contrôl**	74	21	13	40	28	46
Increase						
*(s')allong**	7	3	3	1	43	86
*activ**	393	147	207	39	37	90
particip (à)*	43	11	23	9	26	79
Decrease						
intérromp/interrupt**	23	11	11	1	48	96
*atténu**	32	12	20	0	37	100
*limit**	99	27	22	50	27	49
dépr	40	4	31	5	10	88
Preservation						
*conserv**	46	15	26	4	33	89

Existence Dependency

Creation

suite à/au

Ainsi l'induction des récepteurs LDL **suite à** l'activité des statines entraîne une augmentation de l'activité de la 7ahydroxylase.

On a rapporté des baisses de la synthèse hépatique de CH **suite à** un traitement avec des statines.

7.6.2.1 *résult* (en)*

L'administration d'une suspension orale de charbon activé en poudre ou de colestyramine entraîne une augmentation rapide de la clairance plasmatique de l'A771726 (voir section surdosage); il **en résulte** une réduction de la demi-vie d'élimination à 24 heures.

On a également rapporté qu'un traitement avec des statines pouvait **résulter en** une baisse d'activité plasmatique de la CETP mais pas de la LCAT.

*produi**

Les formes ionisées des métaux de transition participent activement à cette phase de propagation. LOO· peut être directement **produit** par l'action de l'oxygène singulet sur les lipides insaturés. LO· est beaucoup plus réactif que LOO· et perpétue la peroxydation lipidique.

Les *enzymes* **produisent** des centaines de réactions chimiques essentielles à notre survie.

*action**

... l'ubiquinone n'aurait pas besoin de l'atocophérol pour exercer son **action** antioxydante, contrairement à ce qui avait été rapporté précédemment...

Puis différentes catégories de récepteurs à la vasopressine ont été identifiés : * V1 (vasculaires), sur lesquels la vasopressine a une **action** vasoconstrictrice...

gèn*/*gén* not *gène/gènes/général*/génétique*/antigène*/ *antigéniq*/génie/génique*/génom

Ces souches diffèrent par leur pouvoir **pathogène** : les souches présentant un phénotype TKD ont un pouvoir **pathogène** atténué du fait de leur faible aptitude à créer une infection latente [25], alors que les souches TKalt, beaucoup plus rares, conservent leur **pathogénicité**.

Dans 53 cas, le diagnostic de maladie de Creutzfeldt-Jakob **iatrogène** après traitement par *hormone de croissance* extractive a été retenu.

7.6.2.2 *obten*/obtien**

Le but de cette étude était de démontrer la faisabilité de libérer *l'insuline* au sein du circuit sanguin pour **obtenir** une réponse biologique.

Cependant, il a été observé que la réponse biologique au Transfersulin® a été plus reproductible que celle **obtenue** avec l'injection de *l'insuline* à action prolongée.

*immunis**

De nombreux chercheurs pensent qu'il peut être un excellent vecteur, capable d'introduire dans les cellules des particules **immunisantes** contre d'autres maladies.

Est-ce à dire que l'**immunisation** avec de l'ADN nu sera le *vaccin* universel de demain ?

mene/mèn* à*

L'oxydation de l'ADN peut **mener à** de multiples lésions telles que coupure de chaîne, formation de dimères cyclobutaniques, oxydation des bases puriques et pyrimidiques, et pontages des protéines nucléiques.

La libération de *l'insuline* topique et non-invasive, en utilisant les Transfersomes® comme vecteurs appropriés, suggére [sic] une potentialité thérapeutique. Il est possible que ceci **mène à** une régulation métabolique ajustée et plus précise, et améliore la qualité de vie.

*symptôm**

L'amantadine et la rimantadine ont également fait la preuve de leur efficacité thérapeutique, chez l'enfant, l'adulte et les personnes âgées, vis-à-vis de différents sous-types de *virus* de grippe A [3, 15]. Administrés dans les 48 h (si possible 24 h) après le début des **symptômes**, elles permettent de réduire, d'un à deux jours, de façon spécifique et significative, la sévérité et la durée des **symptômes** tels que la fièvre ou les signes respiratoires, et permettent globalement un retour plus rapide à une activité normale.

Moi, je suis un porteur chronique. Je ne ressens aucun **symptôme** pour le moment sauf que mon foie est en train de s'abîmer lentement.

*résult**

See section 1.2.1.2 above.

proven/provien* d**

La perte de poids chez les personnes infectées par le VIH ou atteintes du SIDA **provient**, non seulement **de** la perte de graisse, mais surtout **de** la diminution de la masse corporelle maigre.

On a montré que cette perte d'activité **provenait** essentiellement **d**'une métabolisation plus rapide du phénobarbital.

Maintenance

*requis**

La quantité d'énergie **requise** pour qu'une réaction donnée ait lieu s'appelle énergie d'activation. Bien que la décomposition du lactose soit possible, elle ne se produira que si les molécules du lactose possèdent l'énergie **requise**.

En *vaccinant* le tiers de la population contre l'influenza avec un vaccin trivalent, selon lui, on utilise les mêmes capacités qui sont **requises** pour fabriquer le triple des *vaccins* monovalents qui devraient servir en cas de pandémie.

Prevention

protég/protèg*/protection* (contre)*

Un *vaccin* est une préparation pouvant apporter à un individu une **protection** immunitaire **contre** une maladie infectieuse.

Un *vaccin* préventif a pour objectif de **protéger** la population contre l'infection, avant l'apparition des symptômes, en permettant à la réponse immunitaire, d'abord de reconnaître le virus lorsque le sujet est infecté, puis de l'éliminer définitivement.

défend/défense**

Les protéines de l'enveloppe ont des régions hypervariables qui permettent au virus d'échapper facilement aux anticorps, **défenseurs** de l'organisme.

Dernière piste : modifier des cellules de manière que, lorsque le *virus* se présente, elles produisent une grande quantité d'interféron, une protéine naturelle de **défense** contre les *virus*.

prophyla

... seule l'amantadine a été approuvée par la Food and Drug Administration en 1966, pour l'indication de traitement **prophylactique** vis-à-vis des *virus* grippaux A(H2N2) et seulement en 1976 vis-à-vis de tous les sous-types de *virus* de grippe A.

Il existe au Canada deux mesures qui permettent de réduire les effets de la grippe : l'**immunoprophylaxie** au moyen du *vaccin* inactivé (virus tué) et la **chimioprophylaxie** ou le traitement par des médicaments *antiviraux* spécifiques contre la grippe (amantadine et inhibiteurs de la neuraminidase).

Influence Dependency

Modification

*agi**

Les trithérapies consistent à bloquer la reproduction du virus. Elles font appel à deux classes d'antirétroviraux qui **agissent** à deux endroits du cycle de réplication du VIH...

Il en résulte des *enzymes* sur lesquelles les antirétroviraux ne peuvent plus **agir**.

associ à/au*

Cependant, cette mutation est **associée à** une très forte baisse d'activité *enzymatique* (80 %) et à une médiocre réplication en culture cellulaire, signes d'une perte de vitalité virale.

La perte de poids **associée au** VIH est la première et jusqu'à présent la seule altération de métabolisme pour laquelle les effets de l'*hormone de croissance* aient été approfondis et pour lesquels elle est officiellement recommandée comme thérapie.

*modifi**

Les ERO altèrent les propriétés des membranes cellulaires, dont la fluidité et la compartimentation. Cela **modifie** l'activité des récepteurs et par conséquent la fonction des messagers secondaires...

L'activité des *enzymes* peut être **modifiée** par la prise de certains médicaments.

altèr/altér**

Les ERO **altèrent** les propriétés des membranes cellulaires, dont la fluidité et la compartimentation.

Cependant, il a été rapporté quelques cas d'isolats cliniques d'HSV résistants à l'ACV et sensibles au penciclovir, la résistance de ces *virus* ayant été attribuée soit à une TK **altérée** [14], soit à une **altération** de l'ADN polymérase.

*contrôl**

L'effet physiologique de l'*hormone de croissance* dans l'organisme humain est de **contrôler** une croissance normale et linéaire.

La mise en place, au début des années 70, de France-Hypophyse, organe central de distribution qui **contrôle** le « marché » de l'*hormone de croissance*, a beaucoup contribué à débloquer la situation.

Increase

*(s')allong**

Mais des molécules *antivirales*, comme l'AZT (1), permettent d'**allonger** l'espérance de vie des malades.

L'ADN viral cesse alors de s'**allonger**, et le *virus* ne peut plus se multiplier.

*activ**

L'idée est donc que le composé, sorte de *prodrogue*, ne sera **activé** que par les cellules productrices de PSA.

Il faudra donc rechercher ou même créer de tels couples (prodrogue inactive + *enzyme* **activateur**) pour élargir l'éventail des cancers vulnérables à cette technique.

particip à*

Cette arginine est impliquée, comme deux autres arginines, dans la liaison de l'*enzyme* avec le substrat et également dans un changement conformationnel **participant** directement **à** l'activité catalytique.

Les colibacilles sont en général des constituants inoffensifs de la flore intestinale et **participent** même **à** la digestion.

Decrease

*intérromp**

Une fois là, le médicament **interrompt** la croissance de la chaîne d'ADN, empêchant donc le *virus* d'achever sa réplication.

... « c'est la première fois qu'un produit fabriqué par ingénierie génétique est utilisé pour tenter d'**interrompre** le cycle viral du HIV ».

*atténu**

Tel que discuté à la section 3.1.8i, les statines pourraient **atténuer** la prolifération des cellules du muscle lisse (SMC).

De plus, elle devrait stimuler la production d'époxystérols qui **atténuent** l'expression de l'HMGCoA réductase, ce qui génère, en synergie, un mode régulateur négatif.

*limit**

Plusieurs autres *antiviraux* sont utilisés dans le cadre d'essais cliniques mais leur efficacité est **limitée** par l'apparition de phénomènes de résistance.

Ces *antiviraux* peuvent donc être utiles dans le traitement d'infections à HSV résistants à l'ACV, mais leur toxicité **limite** leur utilisation.

dépr

D'autre part, la cyclosporine étant un ***immunodépresseur***, elle ne peut être administrée pendant un temps indéfini, car elle provoque une déficience *immunitaire* dont pourraient profiter des agents pathogènes opportunistes.

Toutefois, chez des sujets **immunodéprimés** traités par l'amantadine ou la rimantadine, une excrétion prolongée de *virus* résistant a été observée pendant plusieurs semaines, voire plusieurs mois.

Preservation

conserv*

Dans le but de supprimer l'activité *immunosuppressive* et de ne **conserver** que l'affinité à la cyclophiline, en corrélation avec l'activité anti-virale, une modification appropriée de résidus de la CsA impliqués dans l'interaction à la calcineurine a été réalisée.

Le *virus* mutant n'exprime pas l'Ag HBe, mais **conserve** sa capacité de réplication et sa pathogénicité.

Appendix E: Frequency Analysis of the Patterns in English and French

Table 5 summarizes the frequencies of the most common patterns in each of the sub-categories of cause.

ENGLISH		FRENCH	
Pattern	**Total Occur-rences**	**Pattern**	**Total Occur-rences**
Existence Dependency			
Creation			
respond/respons** /5R *to*	613	*produi*/product**	611
*effect**	605	*effet** not *en effet*	588
*produc** not *product/products*	306	**gèn*/*gén** not *gène/gènes/général*/ génétique*/antigène*/ antigéniq*/génie/ génique*/génom**	397
TOTAL	1524		1596
Destruction			
*kill**	107	*élimin**	123
destroy/destruct**	61	*détrui*/destruc**	95
*lethal**	15	*tue*/tué**	62
TOTAL	183		280
Maintenance			
*allow**	87	*permet*/permi**	310
*enabl**	63	*nécess**	163
critical to/for	12	*essentiel** not *essentiellement*	36
TOTAL	162		509
Prevention			
*prevent**	192	*protég*/protèg*/ protection* (contre)*	142
block	54	*préven**	132
*avoid**	32	*bloqu*/blocage**	74
TOTAL	278		348

Influence Dependency			
Modification			
*control**	250	*associ* à/au*	214
*relat** not *relatively*	146	*agi**	151
*regulat**	136	*modifi**	142
TOTAL	532		507
Increase			
*increas**	309	*activ**	393
*activat**	276	*augment**	225
*improv**	161	*amélior**	106
TOTAL	746		724

Table 5: Frequency analysis of most common patterns

The most frequent patterns indicating creation in English occurred 613 times (*response to*), 605 times (*effect*), and 306 times (*produce*). The most common French creation patterns occurred 611 (*produire*), 588 (*effet*) and 397 times (*générer*).

The total occurrences recorded for the most frequent English patterns indicating destruction were 107 (*kill*), 61 (*destroy*), and 15 (*lethal*). In French the total occurrences were 123 (*éliminer*), 95 (*détruire*) and 62 (*tuer*).

The most frequent English patterns indicating maintenance occurred 87 times (*allow*), 63 times (*enable*) and 12 times (*critical to/for*). The corresponding French patterns occurred 310 times (*permettre*), 163 times (*nécessaire*) and 36 times (*essentiel*).

The total occurrences of the English patterns indicating prevention peaked at 192 (*prevent*), followed by 54 (*block*) and 32 (*avoid*). The highest number of occurrences in French was 142 (*protéger*), followed by 132 (*prévenir*) and 74 (*bloquer*).

In the influence dependency, the most common English patterns for modification occurred 250 times (*control*), 146 times (*related to*) and 136 times (*regulate*); their

French counterparts occurred 214 times (*associé à*), 151 times (*agir*) and 142 times (*modifier*).

The most common patterns indicating increase in English occurred a maximum of 309 times (*increase*), followed by 276 (*activate*) and 161 (*improve*). In French the most common pattern for increase occurred 393 times (*activer*), 225 times (*augmenter*) and 106 times (*améliorer*).

The most common English patterns for decrease occurred 198 (*inhibit*), 163 times (*reduce*) and 153 times (*protect*). The most common French patterns for this sub-type of the relation occurred with a frequency of 277 (*inhiber*), 250 (*réduire*) and 201 (*diminuer*).

Finally, the frequencies for the patterns indicating preservation were 61 (*preserve*) and 43 (*maintain*) in English, and 46 (*conserver*) and 40 (*maintenir*) in French.

From these statistics we can see that most of the sub-types of the relation show parallels in the highest numbers of occurrences. In fact, the total occurrences are strikingly similar. There are similar maximums and similar relative frequencies between the top three most frequent patterns as well.

The exceptions to this rule are found in the sub-types of maintenance, prevention and decrease. In the maintenance category of the existence dependency, the number of occurrences are much higher in French (301, 163 and 36 compared to 87, 63 and 12). The numbers for the prevention category are more similar, although there are considerably more occurrences of the top English pattern (192) than the top French pattern (142), while the second most common French pattern occurs far more often than its English counterpart (132 as opposed to 54). Finally, in the decrease sub-type of the influence dependency, there is a much wider separation between the first and second most common

French terms (391 and 225) than between the two most common English terms (305 and 276).

Table 6 summarizes the total occurrences of all of the patterns in each sub-category.

	ENGLISH			FRENCH		
	Chapter	**App**	**TOTAL**	**Chapter**	**App.**	**TOTAL**
Creation	3779	767	4546	2830	1976	4806
Destruction	203	0	203	382	0	382
Maintenance	168	0	168	534	13	547
Prevention	316	40	356	322	235	557
Modification	589	532	1121	422	612	1034
Increase	1672	314	1986	574	443	1017
Decrease	400	676	1076	999	194	1193
Preservation	0	140	140	40	46	86
TOTAL	7127	2469	9596	6103	3519	9622
TOTAL ALL	**9596**			**9622**		

Table 6: Frequency analysis of all patterns

When the total number of occurrences of all of the patterns together are compared, there are relatively minor differences in numbers for the categories of creation, destruction, modification and preservation. The categories of prevention and decrease show slightly larger differences. However, the largest difference is found in the category of increase, in which approximately 700 more occurrences were found in French than in English. Another large difference is found in the category of maintenance, in which approximately 400 more occurrences were found in French than in English.

However, regardless of these localized differences, the parallelism in the frequencies of occurrences is striking.

Appendix F: Cognates observed in the Research

Table 7: Frequency and precision of cognates in English and French

ENGLISH			FRENCH		
Pattern	Total Occurrences	Precision	Pattern	Total Occurrences	Precision
Existence Dependency					
Creation					
*caus**	212	100	*caus**	150	81
provok/provoc**	3	100	*provoc*/provoq**	92	100
*incit**	1	100	*incit* à*	7	100
due /5R *to*	87	99	*dû/due à/au*	27	96
*attrib**	12	92	*attribu* (à/au)*	34	62
*confer** not *conference/conferences*	10	90	*confèr*/confér** not *conférence**	14	93
*induc**	207	89	*indui*/induct**	110	80
**genic/*genesis/*genicity* not *transgenic, antigenic*	172	87	**gèn*/*gén** not *gène/gènes/général*/ génétique*/antigène*/antigéniq*/génie/ génique*/génom**	397	35
			**génèr*/génicité*	3	100
consequence/consequent**	26	85	*conséquence**	26	77
			par conséquent	17	94
*reason** not *reasonable/ reasonably/reasoned*	25	72	*(en) raison (d*)*	134	75
*result** /5R *in/of/from*	299	66	*résult** /5R or /5L *en*	44	41
immuniz/immunis**	130	63	*immunis**	124	35
*produc** not *product/products*	306	62	*produi*/product**	611	40
stimul	181	53	*stimul**	76	86
*effect * /5R *of*	175	58	*effet** not *en effet*	588	52
*effect**	605	51			
responsible /5R *for*	30	40	*responsable* d*/pour*	77	87
*obtain**	60	23	*obten*/obtien**	169	64

Destruction					
destroy/destruct**	61	84	*détrui*/destruc**	95	63
Maintenance					
essential for	6	67	*essentiel** not *essentiellement*	36	53
Prevention					
*prophyla**	26	100	**prophyla**	59	27
*prevent**	192	82	*préven**	132	58
*block**	54	69	*bloqu*/blocage**	74	90
Influence Dependency					
Modification					
*influenc**	13	100	*influenc**	17	100
associated /5R *with*	125	85	*associ* à/au*	214	38
*role**	39	69	*rôle**	90	78
*alter** not *alternate/ alternating/alternative*	41	66	*altèr*/altér**	31	35
*neutraliz**	43	63	*neutralis**	19	95
*regulat**	146	42	*régul* not régulier(s)/régulière(s) /régulièrement*	37	62
*control**	250	37	*contrôl**	74	28
Increase					
*facilitat**	25	100	*facilit**	13	69
*amplif**	17	100	*amplifi**	6	100
*prolong** not *prolonged*	8	100	*(s')allong**	7	43
*prolong**	17	53			
maximiz/maximis**	4	100	*maximis**	3	100
*augment**	4	100	*augment**	225	64
favour/favor** not *favo(u)rable /favo(u)rably*	6	83	*favoris**	26	92
*contribut**	28	79	*contribu**	41	95
*accelerat**	17	76	*accél**	24	75
*activat**	276	72	*activ**	393	37
*cataly**	94	52	*catalys**	32	75
Decrease					
*dimin**	11	100	*diminu**	201	76
*inhibit**	198	71	*inhib**	277	60
*inactivat**	42	43	*inactiv**	43	51

*reduc**	163	43	*rédui*/réduc**	250	50
*damag**	31	42	*endommag**	8	100
*limit**	126	13	*limit**	99	27
Preservation					
maintain/mainten**	43	26	*mainten*/maintien** not *maintenant*	40	60
*preserv** not *preservative**	36	19	*conserv**	46	33