ORIENTATION IN MULTIPLE LEXICAL TERMS
AND VERB PHRASES: A MODEL FOR
SPECIAL LANGUAGE COMBINANTS

Patricia C. Thomas, M. Phil., F.I.L.

Department of Linguistic and International Studies

© This dissertation is submitted
in fulfilment of the degree of Ph.D.
at the University of Surrey, U.K.
March 1995
Summary

The theme of this thesis is the 'orientation' of multiple lexical terms and special language verb phrases. Orientation is a necessary step for two main reasons:

- ascertaining the most logical placing of multiple lexical terms and special language verb phrases (combinants) in a dictionary;
- providing the most apposite terminological and terminographical background data for a multiple lexical term or phrase, these data being determined by the subject field of the text in which the term or phrase appears.

The research has drawn together aspects such as collocation and valency, and analyses of corpora have resulted in the development of a model for special language verb phrases in English and French which it is proposed can be applied and adapted to different specialised subject fields. Past research into special language verb phrases has been sparse and, in contrast to general language, it does not appear that a model pertaining to this construction has been developed previously. Of additional novelty is the application of the model to special language verb phrases in French, because it is hoped that the results will act as a precursor for a dictionary of verb collocations in that language. It is intended that the results of the research will benefit:

- learners of a foreign language who may become translators, to enable them to seek a term or phrase easily and efficiently;
- subject specialists who prepare papers in a language which is not their mother tongue;
- technical writers;
- pre-editors of texts for machine translation;
- terminographers who need guidelines for entering compound terms and phrases in: (i) printed dictionaries and (ii) computerised systems such as terminology data banks (term banks).

The results are supported by statistical data acquired from the compilation by the author of two special language corpora, one in English and the other in French, of restricted areas of virology and bacteriology.
Acknowledgements

My heartfelt gratitude is due to Professor Frank Knowles, Director of Studies, Language Studies Unit, Aston University, for the many long and inspiring discussions we have had during the preparation of this work, as well as for much practical help in numerous ways. A true mentor.

To Professor Carol Sanders, Head of French in the Department of Linguistic and International Studies at the University of Surrey, I extend my sincere appreciation for providing a valuable critique and much staunch support throughout the writing of the thesis.

To Dr. Peter Roe, Research Fellow, Language Studies Unit, Aston University, my gratitude for the free use of ATA (Aston Text Analyser), an invaluable tool. Thanks are also due to Dr. Agnes Kukulska-Hulme, Department of Modern Languages, Aston University, for helpful advice and the provision of textual material in electronic form. The use of COBUILD and texts from 'Le Monde' for comparative purposes is acknowledged.

I warmly thank the European Association for Lexicography (EURALEX) for the award of a Verbatim grant to help with my research.

I also acknowledge the following sources for the provision of the English and French texts which comprise the corpora:
- Drs. Henri Béjoint and Philippe Thoiron, Université de Lyon II
- European Court of Justice, Luxembourg
- Flammarion Médecine-Sciences, Paris
- International Association for Biological Standardization, Geneva
- Le Monde Economique, Paris
- Organisation Mondiale de la Santé, Geneva
- Organization for Economic Co-operation and Development, Paris

Finally, to my family, friends and colleagues for sustained encouragement and much practical support.
## Contents

### Page

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Summary</strong></td>
<td>2</td>
</tr>
<tr>
<td><strong>Acknowledgements</strong></td>
<td>4</td>
</tr>
<tr>
<td><strong>Contents</strong></td>
<td>5</td>
</tr>
<tr>
<td><strong>Glossary of terms used in the thesis</strong></td>
<td>16</td>
</tr>
<tr>
<td><strong>Introduction</strong></td>
<td>19</td>
</tr>
<tr>
<td><strong>Chapter 1</strong> <strong>Definitions of multiple lexical terms and phraseology, with an overview of related work</strong></td>
<td>21</td>
</tr>
<tr>
<td>1.1 Compilers of dictionaries</td>
<td>23</td>
</tr>
<tr>
<td>1.2 Users of dictionaries</td>
<td>24</td>
</tr>
<tr>
<td>1.3 'Orientation' or focal point of entry terms</td>
<td>24</td>
</tr>
<tr>
<td>1.4 Term bank record formats</td>
<td>26</td>
</tr>
<tr>
<td>1.5 Importance of subject classification</td>
<td>33</td>
</tr>
<tr>
<td>1.6 Size of subject field</td>
<td>36</td>
</tr>
<tr>
<td>1.7 Guidelines for terminologists</td>
<td>37</td>
</tr>
<tr>
<td>1.8 Definitions of phraseology by language professionals</td>
<td>38</td>
</tr>
<tr>
<td>1.9 Definitions of phraseology by linguists (general language)</td>
<td>39</td>
</tr>
<tr>
<td>1.10 The lexicographers' approach to phraseology (general language)</td>
<td>41</td>
</tr>
<tr>
<td>1.10.1 Collocation in LGP</td>
<td>41</td>
</tr>
<tr>
<td>1.10.2 Computational progress in lexicography</td>
<td>43</td>
</tr>
<tr>
<td>1.10.3 Relationships in dictionary definitions</td>
<td>44</td>
</tr>
<tr>
<td>1.10.4 Verb + complement</td>
<td>45</td>
</tr>
<tr>
<td>1.11 Identification of multiple lexical units and phrases</td>
<td>5</td>
</tr>
</tbody>
</table>
Chapter 1

1.12 Definition of phraseology by documentalists/information scientists (LGP and LSP)

1.13 The terminographers' approach: phraseology

1.13.1 The terminologist's approach: established criteria for phraseological units (phrasemes) in LSP

1.13.2 Additional observations on what constitutes LSP phrases

1.13.3 Internal disjuncture in LSP phrases

1.13.4 Definitions of and observations on LSP phraseology and verb phrases

1.13.5 Nomenclature of phraseology

1.14 Conclusions for special language phraseology

Chapter 2

The problems of the terminographer: identification and ordering of multiple lexical units and phrases in LSP; comparisons with LGP

2.1 The role of terminologists: identification of users

2.1.1 Importance of terminology training

2.1.2 Necessity of guidelines for terminographers

2.2 Formation and delimitation of LSP terms

2.2.1 Automatic recognition of compound terms

2.3 Lexical frequency as a method for identifying terms in a subject field: corpora analyses of LSP texts

2.3.1 Frequency of technical and sub-technical words
2.3.2 Comparison of frequency from parallel texts in English and French
2.3.3 Some syntactic characteristics of LSP
2.3.4 Verb and noun homonyms in computer retrieval

2.4 Problems in ordering headwords and identifying ‘focal points’ in terminography
2.4.1 The focal point in indexing methods
2.4.2 Results of questionnaire on the ordering of multiple lexical units in term banks
2.4.3 Search mechanisms in term banks
2.4.4 A comparison: ordering of headwords in LGP
2.4.5 Precedence in nominal groups in LGP
2.4.6 Summary of 2.4

2.5 Multi-word terms and phrases: comparison of LSP with LGP
2.5.1 Compounds
2.5.1.1 Compound nouns
2.5.1.2 Compound verbs
2.5.1.3 Compound adjectives
2.5.1.4 Compound adverbs
2.5.1.5 Adjective(s) + noun(s)
2.5.1.6 Adverb + adjective
2.5.1.7 Adjective + adverb
2.5.2 Phrasal and prepositional verbs
2.5.3 Idioms
2.5.4 Similes
2.5.5 Metaphors
2.5.6 Phrases
2.5.7 Conclusions to 2.5
Chapter 3  Grammatical and lexical collocations: 
orientation of their 'focal point'

3.1  Definitions of different types of collocation  120 
3.2  Grammatical and lexical collocations  122 
  3.2.1  Grammatical collocations  123 
  3.2.2  Lexical collocations  125 
  3.2.3  Conclusions on grammatical and lexical 
        collocations  126 
3.3  Substitution, 'free' collocations and disjuncture  126 
3.4  Base (node) and collocate  128 
  3.4.1  Collocatory 'span'  129 
  3.4.2  Entry at base or collocate?  130 
3.5  Comparisons of lexical collocations in LGP with 
     corresponding LSP constructions  133 
  3.5.1  Adjective + noun  135 
  3.5.2  'Fused' compounds  137 
  3.5.3  Noun + intransitive verb  139 
  3.5.4  Group nouns  140 
3.6  Verb + noun colligation  140 
3.7  Theoretical recommendations for placing 'collocations' in LSP  141 
  3.7.1  Compound nouns  142 
  3.7.2  Compound verbs (denominalised verbs/adverbs)  146 
  3.7.3  Compound adjectives  146 
  3.7.4  Compound adverbs  147 
  3.7.5  Adjective(s) + noun(s)  147 
  3.7.6  Adverb + adjective  148 
  3.7.7  Verb + adverb particle  149 
  3.7.8  Verb + preposition  149 
  3.7.9  Idioms  149
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.7.10</td>
<td>Verb + noun</td>
<td>149</td>
</tr>
<tr>
<td>3.8</td>
<td>Orientation and context</td>
<td>150</td>
</tr>
<tr>
<td>3.8.1</td>
<td>Collocation in corpora as an aid to orientation</td>
<td>152</td>
</tr>
<tr>
<td>3.8.2</td>
<td>The thesaurus as an aid to orientation</td>
<td>153</td>
</tr>
<tr>
<td>3.9</td>
<td>Mutual information and association ratio</td>
<td>158</td>
</tr>
<tr>
<td>3.10</td>
<td>Conclusions</td>
<td>158</td>
</tr>
</tbody>
</table>

**Chapter 4**

**Influence of a subject field on its verbs and verb phrase relationships: synergy of collocation and valency**

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1</td>
<td>Terminological versus lexical aspects of an LSP phrase</td>
<td>163</td>
</tr>
<tr>
<td>4.2</td>
<td>Influence of special subject fields on verbs</td>
<td>166</td>
</tr>
<tr>
<td>4.2.1</td>
<td>Verbs in verb phrases from 'overlapping' subject fields and 'quasi-LSP' verbs in English and French</td>
<td>171</td>
</tr>
<tr>
<td>4.3</td>
<td>Defining the limits of LSP verb phrases: the help of contrastive translation</td>
<td>175</td>
</tr>
<tr>
<td>4.3.1</td>
<td>Collocation in translation</td>
<td>180</td>
</tr>
<tr>
<td>4.4</td>
<td>Valency</td>
<td>181</td>
</tr>
<tr>
<td>4.5</td>
<td>Terminological relationships in LSP verbs and verb phrases</td>
<td>186</td>
</tr>
<tr>
<td>4.5.1</td>
<td>Semantic relationships in verbs</td>
<td>187</td>
</tr>
<tr>
<td>4.5.2</td>
<td>Syntactic relationships in LSP verbs and verb phrases</td>
<td>189</td>
</tr>
<tr>
<td>4.5.2.1</td>
<td>Brief recapitulation of verb properties</td>
<td>190</td>
</tr>
<tr>
<td>4.5.2.2</td>
<td>Verb phrases with a transitive verb</td>
<td>192</td>
</tr>
<tr>
<td>4.5.2.3</td>
<td>Verb phrases with an intransitive verb + preposition</td>
<td>192</td>
</tr>
<tr>
<td>4.5.2.4</td>
<td>Verb phrases with the verb in the passive (+/- preposition)</td>
<td>193</td>
</tr>
</tbody>
</table>
4.6 Terminological/encyclopaedic structure and valency: relationships leading to orientation 195
4.7 Combinants and valency: the psychological subject 199
4.8 Disjuncture 207
4.8.1 Syntactic disjuncture 208
4.8.2 Semantic disjuncture 209
4.9 Summary of the LSP status of verbs and verb phrases 210
4.9.1 Different verb + same noun object 212
4.9.2 Same verb + different noun object 212
4.9.3 Conclusions for comparison of LSP verbs in English and French 216
4.10 Criteria for frames for LSP verb phrases 220

Chapter 5 Symbiosis of collocation, valency and indexing: a terminological framework for indicating orientation in special language combinants 221

5.1 The corpora and text analyser 222
5.2 Problems of automatic identification of verbs in an LSP corpus 223
5.3 Some frequency statistics of LSP verbs 226
5.3.1 LSP verbs in the English corpus 226
5.3.2 Comparison with LSP verbs in the French corpus 227
5.4 Identification of LSP combinants by collocation (frequency >3) 228
5.4.1 Quasi-LSP verbs 228
5.4.2 Carrier or support verbs 235
5.5 Assessment of valency in combinants
5.5.1 Nominalisation, passivisation and the theme/rheme interchange
5.5.2 Valency differences between quasi-LSP and 'carrier' verbs
5.5.3 Valency differences between LSP and LGP
5.5.4 Intransitive and covertly reflexive LSP verbs
5.6 Passivisation and indexing as indicators of orientation
5.7 The LSP verb frame: comparison with a 'carrier' verb frame
5.8 French verbs in the biological sciences corpus
5.9 Adjuncts and disjuncts
5.10 Conclusions for constructing frames for LSP combinants

Chapter 6 Conclusions and areas for further research

6.1 Areas for future research

Figures

Figure 1.i Example of multilingual term bank record format (EURODICAUTOM)
Figure 1.ii Example of a record for the entry term measles virus
Figure 1.iii Terminological distinction of the term nucleus in different LSPs showing part/whole relationships
Figure 1.iv Terminological distinction of the term nucleus in different LSPs showing the relationship operator/agent
Figure 1.v  Poor example of subject index  48
Figure 1.vi  Example of unbroken references in a subject index  50
Figure 3.i  Orientation of measles vaccine indicated by examples of collocation in a text  153
Figure 3.ii  Example of virus classification showing measles virus  154
Figure 3.iii  Logical and ontological relationships in measles virus  155
Figure 3.iv  Intensional and extensional properties of viruses  155
Figure 3.v  Hypothetical thesaurus structure for measles virus  157

Tables

Table I.1  Types and attributes of categories of term bank data fields  33
Table II.1  Word tokens and frequency relative to total number of words  86
Table II.2  'Noise', 'stop' or 'transparent' words eliminated by the Canadian term bank 'Termium' for indexing purposes in English, French, German and Spanish  102
Table II.3  Example of bilingual multi-word concept from BATEM  103
Table II.4  Comparison of multiple lexical units in LSP and LGP  116
Table III.1  Comparison of the orientation of entry terms in LSP with lexical collocations in LGP  134
Table III.2  Comparison of proposed entry points for noun + noun compounds in LSP and LGP  138
Table IV.1  Elements influencing verbs and verb phrases
<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table IV.2 (a)</td>
<td>Comparison of a selection of LSP and quasi-LSP verbs from the biological sciences corpus in English with the LGP corpus, COBUILD</td>
</tr>
<tr>
<td>Table IV.2 (b)</td>
<td>Comparison of a selection of LSP and quasi-LSP verbs from the biological sciences corpus in French with an LGP corpus extracted from 'Le Monde'.</td>
</tr>
<tr>
<td>Table IV.3</td>
<td>Example of syntactic characteristics and terminological relationships of LSP verbs in a restricted subject field</td>
</tr>
<tr>
<td>Table IV.4</td>
<td>Frequency of some highly specialised English verbs in the biological sciences</td>
</tr>
<tr>
<td>Table IV.5</td>
<td>Frequency of some highly specialised French verbs in the biological sciences</td>
</tr>
<tr>
<td>Table IV.6 (a)</td>
<td>Frequency of <em>faire l'objet</em></td>
</tr>
<tr>
<td>Table IV.6 (b)</td>
<td>Frequency of <em>faire partie</em></td>
</tr>
<tr>
<td>Table IV.6 (c)</td>
<td>Frequency of <em>faire appel</em></td>
</tr>
<tr>
<td>Table IV.7</td>
<td>Analysis of the verb/noun homograph <em>fait</em></td>
</tr>
<tr>
<td>Table IV.8</td>
<td>Analysis of the noun/verb homograph <em>mise</em></td>
</tr>
<tr>
<td>Table IV.9</td>
<td>Verb + noun phrases in a restricted French LSP (law)</td>
</tr>
<tr>
<td>Table IV.10</td>
<td>English verbs in the LSPs of law and pharmaceuticals</td>
</tr>
<tr>
<td>Table IV.11</td>
<td>Contrast in LSP and LGP verbs in English and French</td>
</tr>
<tr>
<td>Table IV.12</td>
<td>Frequency ranges of LSP English verbs from the biological sciences</td>
</tr>
<tr>
<td>Table IV.13</td>
<td>Frequency ranges of LSP French verbs from the biological sciences</td>
</tr>
<tr>
<td>Table V.1</td>
<td>Frequency ranges of highly specialised LSP verbs in the biological sciences in English and French</td>
</tr>
<tr>
<td>Table V.2 (a)</td>
<td>Quasi-LSP verb <em>encode</em> + collocatory propositions occurring &gt;3 times</td>
</tr>
<tr>
<td>Table V.2 (b)</td>
<td>Corpus examples of the quasi-LSP verb <em>encode</em> + propositions occurring &gt;3 times</td>
</tr>
<tr>
<td>Table V.3 (a)</td>
<td>Quasi-LSP verb <em>express</em> + collocatory propositions occurring &gt;3 times</td>
</tr>
<tr>
<td>Table V.3 (b)</td>
<td>Corpus examples of the quasi-LSP verb <em>express</em> + propositions occurring &gt;3 times</td>
</tr>
<tr>
<td>Table V.4 (a)</td>
<td>Examples of 'carrier' verbs occurring &gt;3 times supporting the noun <em>protection</em></td>
</tr>
<tr>
<td>Table V.4 (b)</td>
<td>Corpus examples of 'carrier' verbs occurring &gt;3 times supporting the noun <em>protection</em></td>
</tr>
<tr>
<td>Table V.5</td>
<td>Theme/rheme interchangeability with nominalisation in the passive voice</td>
</tr>
<tr>
<td>Table V.6 (a)</td>
<td>Valency of English combinants with quasi-LSP verb <em>encode</em> and noun <em>enzyme</em></td>
</tr>
<tr>
<td>Table V.6 (b)</td>
<td>Valency of English combinants with quasi-LSP verb <em>express</em> and noun <em>sequences</em></td>
</tr>
<tr>
<td>Table V.6 (c)</td>
<td>Valency of combinants with <em>protection</em> + carrier verbs</td>
</tr>
<tr>
<td>Table V.7</td>
<td>Passivisation of quasi-LSP verbs, <em>encode</em> and <em>express</em> terminologically transitional and carrier verbs</td>
</tr>
<tr>
<td>Table V.8 (a)</td>
<td>Indexable form of passivised sentences</td>
</tr>
<tr>
<td>Table V.8 (b)</td>
<td>Indexable form of intransitive verbs and verb phrases with understood reflexive element</td>
</tr>
<tr>
<td>Table V.9</td>
<td>Proposed model of a verb frame for transitive quasi-LSP verb combinants</td>
</tr>
<tr>
<td>Table V.10</td>
<td>Proposed model of a verb frame for carrier verb</td>
</tr>
</tbody>
</table>
Table V.11  Examples of reflexive French verbs in the biological sciences  

Table V.12  Indexable form of French reflexive verbs  

Table V.13  Proposed model of a verb frame for reflexive French LSP verbs in the biological sciences  

Bibliography  

Background reading  

Annexes  

Annex 1  Recommendations for terminologists identifying the headword in multiple lexical units  

Annex 2  Questionnaire to world term banks  

Annex 3  74 term banks worldwide  

Annex 4  On-line search and retrieval mechanisms in term banks  

Annex 5  Frequency of '-isation' string in French biological sciences corpus  

Annex 6  Example of ATA concordance showing left-hand alphabetisation  

Annex 7  Example of Synoptic Profile of expresses using ATA  

Annex 8  Proposed empty model of a verb frame for transitive quasi-LSP verb combinants  

Index  

15
Glossary of terms used in the thesis

anaphoric reduction
Describes a term which has appeared in its complete form the first time it is used but which in the same text is subsequently abbreviated, either by acronyms or initialisms, or by an abbreviated form of the whole term, e.g. foot-and-mouth disease may be abbreviated to FMDV or referred to colloquially as foot-and-mouth. It can extend beyond sentence and paragraph boundaries. (Term coined by the author)

autonomous subject field
Speciality of the lexis of a subject field; a highly restricted field such as virology or computing is considered 'terminologically autonomous' because it contains a high number of nouns and verbs which belong solely to the subject field (e.g. to lyse in virology) as well as a number of terms which have some overlap with related fields. (Term coined by the author)

carrier verb
A verb with little semantic content, used to support the nominalised form of a verb, e.g. to provide protection. Nominalisation results in a series of static situations requiring a carrier verb to move the situation or text along

colligation
Used by some authors to refer to fixed verb phrases, but different from combinant (see below).

collocation
An attraction of two or more words for each other, not to be confused in LSP with multiple lexical unit (see below). There is more than one type, e.g. lexical and grammatical, and these are discussed in detail in Chapter 3.
**combinant**
Term used to cover a looser structure than collocation or colligation and permitting greater freedom in accepting the discontinuity frequently found in LSP verb phrases. (Term coined by the author)

**focal point**  *See orientation*

**multiple lexical unit**
May be a term or phrase in LSP; a clear distinction needs to be made between collocation as it is used in LGP, and the combination of lexical units which form an LSP term; a term representing a single concept and consisting of a multiple lexical unit is not considered to be a collocation, although in its formative stage, collocation may have been instrumental in bringing it to a definitive form, after which it no longer permits internal disjuncture. Internal disjuncture would result in the creation of a new term.

**orientation**
The 'focal point' of a multiple lexical term or LSP phrase by which it is entered by a terminographer to provide a lucid retrieval mechanism. Identifying the 'modificand' and 'modifier' of a multiple lexical term, or assessing the focal point of an LSP phrase through its collocatory pattern, are tasks which are of necessity context-dependent. It is important to note that the same term may have a different modificand and modifier depending on the orientation of its context, which will have been defined by profiling potential groups of users. It is claimed that such tasks take precedence over the more traditional requirements of lexicography and terminography which have hitherto tended to stand alone. Establishing orientation is a key part of the thesis and is analysed in detail in Chapters 3 and 5. (Term coined by the author)

**overlapping subject field**
Certain specialised subject fields such as law and economics depend by their nature on other subject fields. In these fields there are fewer LSP verbs *per*
se but a close degree of collocation is nevertheless encountered.

(Term coined by the author)

**psychological subject**
An agent which is the hidden or understood subject or object of a clause, e.g. *Tom was hit by the ball*, where the agent causing the ball to be propelled is omitted. In valency theory this is represented as an 'optional circonstant'. Such psychological subjects can extend beyond the sentence in which they occur. LSP verbs often have an understood reflexive implying a psychological subject.

**quasi-LSP verbs**
Verbs which (i) may have a difference in frequency when used purely in an LGP sense from that used in a specialised, technical sense or (ii) are influenced semantically by the accompanying noun or noun phrase with which they form a combinant.

**semantic valency**
The change in meaning of a number of LGP verbs which, when used in special subject fields and combined with subject-specific propositions, have a higher frequency than in LGP, and may acquire a special meaning in a particular subject field, e.g. *to type a virus*. They may alter their valency, e.g. *patients present with* (symptoms).

**support verb** See *carrier verb*

**terminology data bank (term bank)**
A computerised collection of terms from specialised subject fields containing support information provided by a terminologist to give specific knowledge about individual terms. By consulting a term bank, users can gain guidance on consistency of usage. Glossaries and similar works can be compiled from term banks, which may be monolingual or multilingual.
Introduction

Lexicographers have traditionally entered the words in their dictionaries in alphabetical order, providing a relatively rapid search method which is easily comprehended by the user. It is however a purely arbitrary form of ordering in which adjacent words generally bear no relationship to each other. In recent years there has been a move towards conceptualising entries by identifying the relationships between them and by ordering them thematically in an encyclopaedic, thesaurus-like structure. This conceptual, non-alphabetic approach might have rendered the search mechanism difficult had it not been for extensive developments in computational methods in three main areas: rapid word-search mechanisms, the ability to structure data conceptually, and vast storage capacity, all of which obviate the need for alphabetic ordering in an electronic publication. Furthermore, there are three main metalexicographical advantages: in addition to the storage of the contents of a dictionary, large amounts of current texts can also be stored, enabling the latest terms to be elicited and definitions compiled; the ease of up-dating existing contents; and rapid electronic publishing. All these factors have revolutionised the realm of lexicography in the past few decades.

Such sophisticated methods, however, do engender their own problems. Our current era has become and is continuing to become more specialised scientifically and technically. To cope with the expansion of knowledge, more dictionaries on specialised subject fields are needed to serve the needs of six principal groups of users: subject specialists in their own language; subject specialists whose mother tongue is not English and who need to present and publish their work in English which is currently, and likely to remain, by far the most widely used language in scientific fields; translators, largely as a result of the previous group; terminographers who provide support, not only for translators, but for classification and standardisation specialists; technical writers and, finally, teachers of English as a Foreign Language (EFL).
The main difficulty in ordering the terms and phrases of specialised subject fields is that a very large number in each domain will comprise a multiple lexical unit which, although an entity representing a single concept, nevertheless needs to be analysed according to the context in which it appears, and its use, to establish its 'orientation'. This is an important step so that terminographers may place a term or phrase in the most logical and appropriate manner to enable a user to gain rapid access to the information being sought.

The research reported in this thesis has aimed at structuring the need for 'orientation' by analysing the theoretical issues which are necessary for such a task to be viable and effective. In addition to nouns and adjectives as terms, verbs and verb phrases are also deemed to play an important role in terminology. Here also the problem is one of 'orientation' and order.

The assessment of verb phrases in this respect has proved an interesting study, particularly when comparing languages, in this case English and French, with additional examples from other languages. By verb phrase is meant the verb + noun collocation ('combinant' in LSP) which has a freedom of structure and components greater than that of idioms but which nevertheless displays an attraction between a particular verb and its collocate. Even more tantalising is the difference noted in the valency of some verbs used in general language from that found when the 'same' verbs are used in the language of specialised subject fields. Comparatively little research appears to have been done in this area, particularly in French. Tesnière in 1959 expounded valency and this was greatly enhanced by Mel'chuk in his Dictionnaire Combinatoire de la Langue Française (1984); however, further studies are sparse and there is a dearth of collocational dictionaries for this language. It is hoped that the research reported here, which provides frames for verbs in specialised subject fields of the biological sciences, will be instrumental in encouraging a growth of such dictionaries.
Chapter 1

Definitions of multiple lexical terms and phraseology, with an overview of related work

The problems addressed in this thesis are those which face compilers of general language dictionaries and specialised terminology works when multiple lexical terms and phrases must be placed in the most appropriate way to provide greatest ease of access for the user. Recommendations are proffered for terminology work in special subject fields in English and an assessment is made of their application to verb phrases in another European language, French. To cover multiple lexical terms and phrases, the global term 'phraseology' is used, representing as it does a wide range of word combinations. However, better defined parameters have emerged in the course of this study which are expounded in the thesis. The differences perceived between phraseology in general language, or language for general purposes (LGP), and special or sublanguages (languages for special purposes: LSP), have been identified and assessed. A detailed analysis of these differences is given in this chapter, which also provides a brief overview of how multiple lexical units and phrases are viewed and treated by different language users.

Before progressing, a discussion of the criteria which constitute LGP and LSP is needed. It is true that, whereas there are many clear-cut examples of subject fields which have their own lexis and restricted syntax, nevertheless there are a number of 'grey' areas where lexis in particular appears to vacillate between LGP and LSP. This is especially noticeable in certain technical and medical topics, which have become popular largely through the media and which differ from general language to a far lesser extent than 'true' LSP. It is probably
better to view LGP and LSP as opposite poles on a scale, the gradations of which are dependent on such features as lexis, semantics, syntax, morphology, style and register, and to understand that there are numerous instances of overlapping, such as the differences observed in register when medical practitioners talk to each other, in contrast to a doctor talking to a patient (see *inter alia* Sager, 1990: 103). In this thesis, LSP relates to discourse in specialised subjects which has a highly restricted use and is only understood by workers in the corresponding or closely related subject fields.

Since the thesis concerns dictionaries and allied compilations such as glossaries, terminologies and thesauri, it is of great importance to make an initial study of the level of knowledge and comprehension of the intended reader of these works. As indicated in the previous paragraph, an academic work may be destined for use either by academic peers or by 'lay' people and accordingly their expectations and purposes are likely to differ. Textual formulation will necessarily differ in the hands of a good writer who has to try and assess the different levels of reader comprehensibility.

It is also apposite at this point to move from the users to the compilers and to clarify the respective roles played by terminologists and terminographers (as opposed to lexicographers: see next section). Briefly, a terminologist studies the theory of terminology, with its conceptual basis, and proposes methods for arranging and displaying the (mainly) linguistic representation of knowledge of a particular subject field. A terminographer provides the practical input - the terms and their supportive descriptive data - based on the results of the terminologist's proposals for conceptual structures.

Chapter 1: 22
1.1 Compilers of dictionaries

From the point of view of compilers of dictionaries, it is generally accepted that lexicographers are those engaged in the production of general language dictionaries and glossaries, whereas terminographers are those involved in special languages or sublanguages, and these are the categories adopted in this thesis. The approach of lexicography is different from that of terminography; briefly, whereas lexicographers are interested in individual words, albeit within a lexical system which comprises metalexical elements such as etymology, terminographers deal with a subject system as a whole.

Because of the nature of specialised dictionaries and terminologies, the compiler is invariably a subject specialist who will ideally be aided by a terminologist so that the structure of the definitions in these works can follow an ordered pattern throughout. For example, in a dictionary of virology containing c. 4,000 terms in which the author was involved as terminographer, data elements were agreed with subject specialists for the description of a virus in the following order:

- name of virus (entry term)
- etymology
- synonyms, including acronyms and initialisms
- hierarchical position: family/group, genus or 'synonyms' of these for cross-referencing
- intrinsic characteristics (molecular weight, buoyant density etc.)
- extrinsic characteristics (epidemiology, host vectors etc.)
- terms within the description in small capitals for cross-referencing
- bibliographic reference(s)
1.2 Users of dictionaries

From the point of view of dictionary users, Riggs (1989: 89) states that at the functional level, 'lexicography has the primary aim of helping readers to interpret texts, whereas terminology aims to help writers to produce texts'. These are general statements and it should be reiterated that lexicography refers to general language and terminology to special languages. A thesaurus, for example, helps a writer to produce texts, but it might contain either general or specialised language. Riggs contrasts the two disciplines at the structural level in the following manner: 'lexicography follows a semasiological line, from words to their meanings, whereas terminology adopts an onomasiological model, proceeding from concepts (as defined by a text) to the terms that designate them'. Terminology undoubtedly proceeds from concepts to terms, but afterwards there may be a tendency to subside into semasiology; through time, terms may acquire new meanings, and so the process back and forth may be a continuous one. While bearing in mind the requirements of users, it is on the needs of terminographers that the current work concentrates.

1.3 ‘Orientation’ or focal point of entry terms

Concepts in LSP are often represented by a term or phrase comprising several orthographical words, giving rise to the question of which word should be selected as the ‘focal point’ for the term to which the terminologist needs to add supportive, informative data. It is also necessary to provide guidelines for the reader. Thus what I have called the ‘orientation’ of a multiple term, that is, the point at which it is entered, enables terminographers to take the needs of potential users into consideration. The words forming a term can be oriented in different ways and the aim is to adopt the method of ‘permutation’, as in indexing (see the
example in 1.11), to ascertain the primary semantic focus, i.e. in a conceptual rather than a linguistic sense, so that there is a link between encyclopaedic content and its linguistic realisation; this is a major focus of the thesis.

The first task of the terminologist is to define the limits of a subject field in consultation with a subject specialist and to identify the concepts it comprises. The next task is to devise the most efficient and comprehensive way of representing these concepts in, for example, a terminology data bank or 'term bank', and the following example serves to highlight the problems of orientation in this type of medium: when assessing which of the two words comprising the term *measles vaccine* takes precedence (in other words, which will be the 'modificand' and which the 'modifier'), the terminologist needs to ascertain the purpose and reason for which the terminology is intended, thus progressing from logic to a profile of the user. If it is to be for a product on paper derived from a term bank, such as a glossary, on the disease *measles*, the entry will be under *vaccine*; if, however, the glossary concerns vaccines for children, the emphasis will be on the disease. It is clear therefore that a terminologist will need to seek different data to provide the supporting material for the term, depending on its subject field and the needs of the intended user, as seen in the typical term bank records given in Figures 1.1 and 1.ii. Dykstra (1985: 54) gives the following example from indexing which shows orientation:

<table>
<thead>
<tr>
<th>Single-word term</th>
<th>Differential as focus</th>
<th>Differential as preceding adjective</th>
<th>Differential post-modifier</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Children</em></td>
<td><em>Children's literature</em></td>
<td><em>Games for children</em></td>
<td></td>
</tr>
<tr>
<td><em>Denmark</em></td>
<td><em>Danish pastries</em></td>
<td><em>Books on Denmark</em></td>
<td></td>
</tr>
</tbody>
</table>

Chapter 1: 25
Sager et al. (1980: 268) divide noun compounds into referent + nucleus, stating that the nucleus can be kept constant and be variously determined, e.g. chain wheel, planet wheel, or the determinant can be kept constant to create a subject or an operationally-related set of terms, e.g. space age, space capsule. This division however needs to be taken a step further to state how each is defined. I would argue that referent and nucleus change according to context as in the example above of measles vaccine. Sager (1990: 190-191) lists various options for the ordering of terms: strict alphabetical, keyword alphabetical, and permuted, or rotational, indexing; these search forms are simple and effective but do not take the orientation of the term into consideration.

1.4 Term bank record formats

Term banks are considered here because they are an important and widespread medium for representing terminology. Terminographers need to resolve the problem of representing terms in them; usually this is done in a global, 'cover-all' manner, but a more specific approach is needed if decisions for their orientation are to be made by terminographers and a knowledge base tailored to users' needs is to be envisaged, whereby users follow a series of optional pathways to obtain the information required. Each concept in a term bank is allocated certain types and numbers of individual information categories, known as 'fields', which need to be defined by the terminographer at the outset according to the subject field, so that each term and its attendant data may be stored in a consistent manner. The choice and number of fields are open-ended. The fields pertaining to each concept are grouped in a structure known as a 'record format'. With so many term record formats being developed more or less simultaneously for disparate needs in different term banks, it is inevitable that the fields deemed necessary to provide information in any subject field are usually adequate,
sometimes comprehensive, but are invariably widely divergent. A compatibility of record formats and techniques for the exchange of data such as Standardised General Mark-up Language (SGML) are clearly desirable.

Hvalkof (1985) has made a comprehensive comparison of six major term banks, discussing their record formats in detail. One of her aims has been to ascertain which data elements were interchangeable with other banks (Hvalkof, 1985: 113); she concluded that their capacity for exchange varies greatly and that harmonisation is desirable to improve exchangeability. As long ago as 1971, the mandatory and optional categories required for term banks (in this case relating to the Banque de Terminologie in Montréal and LEXAUTOM) were enumerated (Vinay and Kallio, 1971: 17-27). The main points emphasised were the word, the reference to a text and the subject field. Twelve fields were refined from earlier versions, two of which could be interchangeable:

1. entry term and automatically allocated number
2. grammar
3. source
4. geographical origin
5. year
6. author of the record (the terminologist)
7. translation
8. source
9. a) definition and source
    b) context and source
   ) interchangeable
10. usage
11. cross-references

Chapter 1: 27
Sager and his colleagues (Sager and McNaught, 1981; Sager and Price, 1983; Candelan and Sager, 1986) have discussed the content of record formats, the results of which have been largely adopted in Figure 1.ii.

To give an example of a completed record in a major term bank, Figure 1.i shows the term *measles* in three languages, taken from the European Commission's term bank, EURODICAUTOM. EURODICAUTOM's database is in French but queries are posed in English. The search mechanism has an automatic truncation function. Answers may correspond exactly to what has been requested, or contain several answers if the query is for a compound term; this can in some instances result in a vast number of replies or "hits" because it is not possible to 'tighten' the search. For example, it can be seen in the example below that, although *measles* was the requested term, the search has also produced *post-vaccinal measles* and in fact nineteen different 'hits' resulted. Although the *Code matière (CM)* can be specified, it is not necessary to specify this to retrieve a term. Furthermore, EURODICAUTOM does not take into consideration the 'orientation' of a compound term.
Figure 1.i Example of multilingual term bank record format (EURODICAUTOM).

Legend:  BE = bureau émetteur i.e. Bureau de Terminologie Brussels (BTB) or Luxembourg (BTL)
         TY = type i.e. sub-domain
         NI = numéro d'initiative i.e. term number
         CF = code de fiabilité (on a scale of 1 = poor, 5 = very good)
         CM = code matière i.e. domain
         VE = vedette i.e. term
         DF = définition
         EN = English; FR = French; DE = German (following ISO 639 (1988): Codes for Representing Names of Languages)

Figure 1.ii shows an early version of a completed record devised mainly by the author, this time for the term measles virus, in the term bank developed at the University of Surrey, based on DANTERM, the Danish Term Bank, and the one developed by Sager and his colleagues at UMIST.
Figure 1.ii Example of a record for the entry term measles virus entered in the Surrey term bank in 1985 and updated in 1990, with British English as the source language (SL). Not all record fields are necessarily completed. All 29 fields can be repeated for each additional language. My comments are italicized.

Chapter 1: 30
It was soon realised that the data collected for each concept could be divided into several categories, some mandatory and some optional. First, administrative data are necessary to identify each concept in the term bank; they have no connection with the term *per se* but include a term number, the date on which the term was recorded in the term bank, updated and so on.

The second category contains linguistic data which relate to the term itself (e.g. *measles virus*). Such data may be synonyms (e.g. *rubella virus*), that is they describe the same concept but differ in register; they may be acronyms or abbreviations, or even antonyms, which can be a useful ploy in translation, in conjunction with a negative; in other words, terms which denote the same, or diametrically opposed, concept.

The third category includes grammatical information, such as gender, agreement and number, as well as phonetic transcriptions, which are particularly useful for languages such as English (Judge and Thomas, 1988: 530).

The fourth category comprises data which show (a) the relationships between terms such as their position in a hierarchical structure (superordinate/subordinate, e.g. for a virus: family *Picornaviridae*, genus *Aphthovirus*) or (b) which may give indications of relationships such as intrinsic and extrinsic characteristics (Judge and Thomas, 1988: 528), for example 'Highly contagious and debilitating disease. Produces blisters on tongue and feet of cattle, pigs, sheep and goats' (Hull, Brown and Payne, 1989: 78). It is therefore in the text of the definitions or descriptions that the relationships between terms are mainly found, as already shown in the data elements given in the earlier section on 'Compilers of Dictionaries' (1.1). The
identification of such relationships can be a help when initial attempts are made to ascertain the orientation of a multiple lexical term. In later chapters of this thesis the relationships identified by the author and others in special language verb phrases will be discussed.

A final, rather ad hoc group of fields comprises those which increase our knowledge of a subject, such as representations of the concept as it appears in context (of prime use for translators; see Bucher et al., accepted for publication in 1995) and bibliographic references (also useful for translators, but particularly for subject specialists, teachers and students).

Table I.1 summarises the different types of data fields and the attributes in them which have been identified by the author. The descriptions 'stative' and 'dynamic' usually refer to the grammatical classification of verbs; however, in categorising data fields, the author has used 'stative' to represent a static, unchangeable data element, as given in the example in the third column, and 'dynamic' to represent a changeable element such as a synonym and its acronym, i.e. foot-and-mouth disease, FMDV. In the case of a dynamic element, the temporal aspect of synchrony and diachrony may also be introduced, e.g. the obsolete French equivalent cocotte for foot-and-mouth disease instead of the current fièvre aphteuse.

The results of incorporating a larger number of data fields in a term bank give greater specificity to the term, thus providing a necessary step in helping to support its orientation according to the needs of the user.
Table 1.1  Types and attributes of categories of term bank data fields

1.5  Importance of subject classification

A further help in orientation is the delineation of a subject field which needs to be considered carefully before embarking on any dictionary type of compilation, whether it be on paper or on a computer. A printed paper work is of necessity limited by publication restrictions, but computerised terminologies also need to have carefully delineated subject boundaries so that the same string appearing in more than one subject field can be clearly defined (Figs. 1.iii and 1.iv). In addition to considering the needs of term representation in the formats used in dictionaries, glossaries and term banks, where a term is isolated from its context and needs to be related to a subject-specific conceptual structure, the terminologist is wise to assess the textual material from which terms are derived, because text forms the 'megastructure' supporting the meaning of the term. Moreover, subject classification helps to clarify the exact meaning of homographs which, when appearing in more than one subject field, or even more than once in the same subject field, may vary semantically in each (Thomas,
1988: 4). The example given is of the term *nucleus* used in a number of different LSP domains; in nuclear physics, metallography, biology and astrophysics, it has a part/whole relationship (Figure 1.iii), while in meteorology, archaeology and again, metallography, it acts as operator/agent (Figure 1.iv). It is interesting that in the same field, metallography, *nucleus* has two distinct meanings which are represented differently in French, in Figure 1.iii (iv) being rendered as *coeur* and in Figure 1. iv (ii) as *germe*. It is surprising that *nucleus* is syntactically stable yet semantically ambiguous within such a narrow subject field, because the lexical root in English is the same. Nevertheless, in practically all cases, the importance of defining the subject field is paramount because it will usually resolve such ambiguities. In corroboration of this, Hoffmann (1987: 158), describing the LEXIS Term Bank, has stated that initially LEXIS translators consider the subject field to be of secondary importance; they tend to rely on intuition when seeking a target language equivalent, but later join terminologists in appreciating the need for well defined subject fields.

Various subject classifications have been developed, notably the Universal Decimal Classification (UDC) and that devised by Dr. Lenoch for EURODICAUTOM between 1971 and 1977 in Luxembourg. The latter comprises 46 principal divisions, each containing between 10 and 20 sub-divisions (Lenoch, 1981). This is a rather *ad hoc* system and, as far as has been ascertained, is used solely by EURODICAUTOM. It has been suggested that the C.E.C., which uses SYSTRAN machine translation for a considerable part of its translation work, has found it unnecessary to make use of the subject coding in EURODICAUTOM and estimates that a correct rendering results in approximately 90% of translation work (I. Pigott, 1985, personal communication).
1. **Part/whole relationship**

   For most subject areas it represents part of a whole,

   e.g. (i) **BIOLOGY:** Cell

   ![Diagram](attachment:image1.png)

   (ii) **ASTROPHYSICS:** Comet

   ![Diagram](attachment:image2.png)

   (iii) **NUCLEAR PHYSICS:** Atom

   ![Diagram](attachment:image3.png)

   (iv) **METALLOGRAPHY:** Case of metal

   ![Diagram](attachment:image4.png)

   Conclusion: Nucleus entails central part of core or a whole object - but:

   *(see Figure 1.iv)*

   
   **Figure 1.iii** Terminological distinction of the term *nucleus* in different LSPs showing part/whole relationships

   Chapter 1: 35
2. **Operator/agent** (nucleation).

There are instances where nucleus does not appear to form the 'part' in a part/whole relationship but seems to be regarded as the agent of a process, e.g.

(i) METEOROLOGY: nucleus = operator from water vapour

(ii) METALLOGRAPHY: nucleus = initial element of transformation  
    e.g. crystallization

(iii) ARCHAEOLOGY: stone, e.g. flint, from which tools may be made

*Figure 1.iv Terminological distinction of the term nucleus in different LSPs showing the relationship operator/agent.*

### 1.6 Size of subject field

It has been observed by the author of this thesis and corroborated by J. Sager (personal communication, 1985) that a good working number of terms in a subject field should not exceed 500 on average. The International Standards Organization has lowered its recommended number of concepts from 1,000 in 1969 (ISO R 919, 1969:8) to 200 in "International terminology standards - Preparation and Layout" (ISO 10241, 1992: 3) which states that "Experience has shown that if the number of concepts exceeds approximately 200, a subdivision of the project into a number of sub-projects becomes necessary." To give a working example, the Danish Term Bank, DANTERM, is using a system of classification developed in the Nordic countries which is conceptually based while at the same time being pragmatically applicable. The classification comprises an 'open' row of macro-units, i.e. more macro-units can be added as required. Each of the 18 macro-units is represented by a single letter, e.g. B = art and literature, H = medicine. Each macro-unit has a permitted maximum of 10,000
micro-units. Micro-units are sub-sets of macro-units and each can contain up to 500 terminological units. Micro-units are characterised by an alphanumeric code which includes the letter of the macro-unit followed by four digits, e.g. G7430 = microbiology -> measles virus. If a micro-unit exceeds 500 terminological units it should be divided and this step needs to be realised as early as possible to avoid reclassification of all the terminological units; this decision should be part of the project plan and is an area where recourse to a subject specialist is vital before embarking on any terminological work.

DANTERM's classification is composed of three parts:

- a general view of the macro-units
- a systematic list
- an alphabetic list of search terms

The general view of the macro-units shows the superordinate structure of the classification. For each macro-unit, a systematic arrangement of its micro-units is given with their alphanumeric code. To each micro-unit is allotted a number of keywords, or search terms, which describe the parameters of the micro-unit. Classification can be updated and extended and this is undertaken by a classification group. DANTERM's inclusion of a field in the term record not only for classification but also for application and project has already been mentioned (B. Nistrup-Madsén and K. Westerberg, 1987, personal communication). These fields could be an invaluable aid towards determining the orientation of a multiple lexical term.

1.7 Guidelines for terminologists

One of the problems I have sought to resolve in this thesis is that of providing a theoretical framework from which guidelines can be deduced by
terminologists who, as seen from the problems enumerated in the previous sections, need help with guiding their search for the supporting data for a term (see Chapter 3, Annex 1 and Thomas, 1993). A number of computational methods can help in this task. For example, by establishing word frequency lists through the use of concordancing techniques and by studying the different collocations which emerge, it is envisaged that some light may be shed on which word from a multiple lexical unit or phrase should be chosen as the focal point or 'orientation' for terminology work.

1.8 Definitions of phraseology by language professionals

Various groups of language professionals, which are not mutually exclusive, have been identified and a brief evaluation of their views on phraseology has been made to ascertain to what extent these views may have a bearing on terminology work. The groups assessed come under the categories of linguists (1.9), lexicographers (1.10), indexers (1.11), documentalists (1.12) and terminographers (1.13). Translators, clearly another group of language professionals, are included under the group of terminographers, because of their need to create terms and phrases, for example neologisms and paraphrases. Detailed definitions of "phraseology" are given in the chapter, particularly in the section on terminologists and terminographers (1.13), since one of the aims of this thesis is to provide guidance for terminographers. Grammatical constructions which come under the heading of phrases are discussed later in the thesis, with particular attention being paid to the verb + object/proposition construction in which collocation plays an important role, and to semantic considerations (Chapters 3 and 4).
While the phenomenon of multiple lexical units plays an important part in the thesis, it is work on phraseology that this section is seeking to identify, a task that has not proved easy. Practically all the work done by linguists in areas where western European languages are used relates to LGP. Adams, who has made a considerable contribution to work on word formation, comments that the construction of verb + complement + preposition, e.g. *to give rise to*, *to take advantage of*, are fixed in varying degrees; for example, modification of the noun in *to give rise to* would seem odd (Adams, 1973: 10). In addition, Adams (1973) and Bauer (1983) list many combinations of word-formation and these will be discussed further in Chapter 2 in an analysis of headwords. However, in their work these linguists study what constitutes lexicology generally, rather than phraseology in particular.

Pilz (1981: 24) observes that while 'fixity' (Festigkeit/Fixiertheit/Stabilität) is a criterion of a phraseological term, strict morphosyntactic unchangeability only serves for a few standard examples. He quotes Häusermann (1977: 83) who rejects absolute fixity, and Wissemann (1961: 235-246) who expresses the need to distinguish between lexicalised or fixed and non-lexicalised or free grammatical elements. The term 'stability' is used to refer to semantic fixity only in the sense that the meaning of the phraseological entity remains fixed, even when the lexical structure may be changed through its context (e.g. by ellipsis or allusion). In this respect, Telija (1976) and colleagues from the former Soviet bloc, such as Eckert, refer to 'phraseologicity' to cover 'combinability', reproducibility, diachronic and contrastive phraseology, instead of 'stability' and rigidity. These authors note an increasing preoccupation with semantics in conjunction with a rapid rise in valency theory (see Pilz, 1981).
Valency and case have been included in this section because both cover combinations of words. Valency refers to the type and number of syntactic links between grammatical elements such as a verb and its complements and adjuncts, while case provides practical possibilities for computational linguists because of the way it relates syntax to semantics. The instigator of the valency theory in the late 1950s was Tesnière (1959), whose ideas on the topic were strictly verb-centred, with three types of 'actants' being distinguished: subject, direct object and indirect object, while relationships such as time and place were denoted by 'circonstants'. Over a decade later in Germany, Helbig and Schenkel (1973) considered the verb as the structural centre of the sentence, with actants, or complements, which are bound to the verb, being defined as either obligatory or optional, whereas adjuncts (or 'circonstants') are not bound to the verb. These observations have been taken into consideration and greatly expanded, particularly in Chapter 5. Somers (1987) explores the theories of valency and case in great depth, explaining their applicability to computational linguistics, artificial intelligence and in particular to machine translation. Furthermore, the set of universal cases expounded by Fillmore (1968) and the implied relationships which belong to different verb types e.g. stative, active, processual, will be analysed in relation to special languages in this thesis, and case-type frames established (Chapter 5). In addition, an assessment is made of the changes in dynamism and the inheritance factor when nominalisation of verbs occurs.

LSP verbs appear generally to have low valency, usually not greater than two (see also Kjaer's comments in 1.13.4). However, the type of actant appears to have an influence on the transition of verbs from LGP to LSP, depending on the special subject field; in legal terminology, for example, there is a high number of abstract nouns, as will be seen in Chapter 4. The hypothesis that there may be restricted numbers and types of actants in LSP is explored
further in Chapter 4.

In addition, it is interesting to consider whether the characteristics of a verb are inherited by their nominalised form and hence result in a change in valency. Certain verbs will inherit the semantic characteristics of the noun phrase from which they are conflated, e.g. *to fax = to send a fax, to hospitalise = to take into hospital for treatment* and in French, *s'aliter = to take to one's bed.*

In LSP, the very restriction in the meanings of terms, whether they be nouns or verbs, or one deriving from the other, provides less leeway for the presence of extraneous characteristics ('extension' in logic and terminology) than is found in general language; consider for example the verbs from virology *to fingerprint* (borrowed from police work), *to weld, to sediment, to centrifuge.*

1.10 The lexicographers' approach to phraseology (general language)

It has not proved possible to arrive at a global unified definition of phraseology as it is understood by lexicographers of general language. The approach seems to be intuitive to a large extent and covers a variety of lexical and syntactical, fixed and non-fixed word groups which fall outside more 'established' word combinations. Nevertheless, an exposé of what some lexicographers understand by phraseology in general language follows so that comparisons may be made with special language terminography.

1.10.1 Collocation in LGP

Aisenstadt (1979: 71) classifies what she terms 'non-idiomatic phrases' into 'free phrases' and 'restricted collocations' (RCs). She states that the latter have restricted commutability in their grammatical and semantic valency but have greater commutability than an idiom; they do not form one semantic unit.
(compare for example 'face the facts' with 'face the music'). Their meaning, unlike idioms, is made up of the sum of the meanings of their constituents. The definition and degree of restricted commutability, however, remain a problem and many RCs in LGP are either intuitive or colloquial, e.g. the construction V + Adv (to laugh heartily, loudly, merrily but not *to laugh tumultuously; to smile broadly, sweetly but not *to smile lusciously) and V + Deverbal Noun (to act as go-between). Others, especially the Adj + Noun construction (e.g. a 'black tie'), may be used more formally. Aisenstadt recommends that RCs be treated systematically in general dictionaries and accorded a special place in the manner of idioms (Aisenstadt, 1979: 74), but firmer criteria need to be established to ascertain what constitutes RCs in both LGP and LSP.

Cowie looks at phrases from the point of view of collocation studies and finds the phrase an insufficiently precise base unit, presumably because phrases do not have a standard form of syntax and can readily admit internal disjunctures (he therefore negates the notion of a fixed phrase). This lack of a standard form causes problems with strict collocation patterns. Cowie (1981: 225) states that: '"phraseology" - like "fixed phrase" - has the disadvantage of blurring a distinction which it is important to preserve in collocational studies, between lexical units of various kinds on the one hand and the more abstract clause and phrase structures in which they function on the other'. In collocation studies it is necessary to incorporate the notion of looking for 'near matches' when comparing almost identical phrases and to identify which parts of speech may be added when internal disjunctures occur.

Kunin (1984), in his English-Russian Phraseological Dictionary, divides English phraseology into three categories of set expressions: phraseological units or idioms, semi-idioms, and phraseomatic units. All three
sets share characteristics of semantic complexity, permanence of lexical composition, morphological and syntactic fixity and a refusal to follow the patterns of free-word combinations. Where idioms and semi-idioms vary, according to Kunin, is in the admission of occasional, more complex changes; these are occasional because they appear to be confined, in those few instances where they occur, to the use of the personal/possessive pronoun or to a living/non-living object (*to shut one's eyes to sb./sth.*) (Kunin, 1984: 14). In making these detailed analyses, Kunin identifies a number of types of set expressions which are syntactically disparate and which, although clearly of great use to a learner of English as a second language, are unlikely to be relevant to special languages in which such idiomatic constructions are rare.

1.10.2 Computational progress in lexicography

There have been radical developments in lexicographical practice and techniques in the past few years, among them Mel'chuk and Zholkovsky's Explanatory-Combinatorial Dictionary (ECD) of Modern Russian (1984) and the Collins COBUILD Dictionary of Sinclair (1987) and his colleagues. Mel'chuk and Zholkovsky aim to eliminate the vagueness found in general language dictionaries; their work is to a great extent concerned with 'productive' dictionaries to help with encoding; collocation and combination, particularly the semantic aspects, are important to provide the means whereby users can be helped to avoid producing false collocations. Sinclair and his team admit that, when studying concordances of general language, 'The lexicographer was still left facing the problem of identifying multi-word items in which the words did not appear consecutively in the text ....(e.g. *shake someone by the hand, shake someone's hand, shake hands with someone*, etc.)' (Krishnamurthy, 1987: 65). These workers state that 'Three basic types of phrases were isolated: fixed phrases(*....once in a while*), syntactic phrases (i.e. those which admit internal disjuncture - P.T.) (*....give somebody the pip*), and lexical phrases (*....a good deal/a great deal*).
was given on how to define phrases, at which element to place them, and where to place them within an entry. Polysemous phrases and phrases with pragmatic import were discussed. The distinction between phrases, compounds, and collocational or syntactic patterns was indicated.' (Krishnamurthy, 1987: 66). Comparison with these recommendations will be made with special languages in Chapter 3.

1.10.3 Relationships in dictionary definitions

Wierzbicka (1985: 52) questioned informants in a search for features relating to particular objects (e.g. dog). The object of the exercise was to achieve a distillation of prototypicality and to identify the linear order and spatial proximity of components found in them. She lists these as:

- category (artefact)
- purpose (relative to situation)
- material
- shape
- size

She also discusses the relationships found in definitions and it can be seen that these identify with those in terminology (Wierzbicka, 1985: 106-7):

- part/whole
- size
- structure
- function
- appearance
Her definitions deal with general language nouns only and not with phrases; however, where feasible, these relationships will be applied to the situation in LSP verb phrases later in this study (Chapter 4).

1.10.4 Verb + complement

Computational lexicographical research on the Longman's Dictionary of Contemporary English (LDOCE) grammar codes has assessed a verb's subcategorisation, in particular the range of alternative complement structures it can take (Boguraev and Briscoe, 1989: 106). Verbs which undergo the dative alternation are coded in LDOCE; these verbs are said to be ditransitive, such as 'give' and 'donate'. One tendency which emerged from the study shows that the more idiomatic the usage, the less likely will be the occurrence of the dative 'alternation' (Boguraev and Briscoe, 1989: 114-115). Compare for example:

She gave him the present (in the sense of 'donate')

She gave the present to him

with

The meal gave him indigestion (in the sense of 'cause to have')

*The meal gave indigestion to him

Consideration has also been given by these authors to the effect of Germanic v. Latinate roots on the dative alternation (Boguraev and Briscoe, 1989: 113), contrasting phonologically 'short' monosyllabic (and some bisyllabic) and 'long' (some bisyllabic with all polysyllabic) verbs (Boguraev and Briscoe, 1989: 114). Their results proved inconclusive but it is interesting to speculate whether these roots affect valency in LSP which often favours classical roots, e.g. in law. This factor has been deemed to be outside the remit of this study and has not been explored further but would be a topic for further research.
Identification of multiple lexical units and phrases by indexers of learned books (special language)

Most indexing of special language works involves the appropriate placing of nouns and compound noun terms. The work of the indexer of special subject fields has much in common with that of the terminologist in that it deals to a large extent with relationships. The concepts of the special subject field must be chosen from the text, and sub-headings of a hierarchical or part/whole relationship established. It is then relatively easy to identify further concepts and assign them to the appropriate place. The concepts are represented by terms which in various publications may be called keywords, index heads, entry points, lead terms or subject headings; keywords in a title or other context are known as keyword-in-context = KWIC or keyword-out-of context = KWOC. Highly restricted fields such as those in the pure sciences pose particular problems for indexing, firstly because of the need to identify their terms which are frequently multiple lexical units and secondly because the same terms occur with great frequency. These problems are exacerbated when such texts are not in machine-readable form, so that indexing has to be undertaken manually, a highly skilled task.

In a medical work such as the proceedings of a meeting, where all contributions focus on the same restricted subject, criteria must be established for the way in which terms occurring with high frequency should be represented in an index. There is a need to include terms which appear only occasionally, and the requirements of retrieval systems must be taken into account; sometimes these will make searches beyond keyword level and include terms from indexes, to provide comprehensive results. It is also important to include quasi-synonymous terms for which a reader might search but which do not appear in the text.

Chapter 1: 46
Monographs on highly specialised subjects are probably the most difficult of all publications for which indexes might be provided because the more specialised the work, the more complex and restricted the terminology. A restricted subject which has already been well researched will have many hundreds of occurrences of the same few terms. The author of this thesis was involved in editing the proceedings of a symposium on the control of pertussis, or whooping cough, in which the text revolved around terms such as pertussis and the same method of inactivation, and dealt mainly with the outcome of clinical trials, resulting in a highly restricted vocabulary. Readers belong to a particular discourse community and will therefore be engaged or interested in the subject; they are likely to be conversant with other workers in the field and their sphere of work. In other words, the thematic focus of the topic is well controlled and so restricted that the need for an index is obviated; it is as if a form of 'self-indexing' comes into force. It may be preferable not to produce a subject index at all rather than one which is unsatisfactory because it does not add anything for the restricted readership (Thomas, 1991). Furthermore, the practitioner is likely to be too busy to undertake such time-consuming and specialised work, and it is often left to people who are inexperienced in the subject field to undertake the task, sometimes with disastrous results. One example of poor indexing which I have encountered gave a one-page subject index, with typing errors, for 384 pages of text, and no sub-headings (Fig. 1.v).
SUBJECT INDEX

A.G.I.D. (Agar Gel Immuno Diffusion test), 52
A.I.D.S., 3, 50, 98, 107, 154, 169, 173
A.I.D.S. vaccines, 10, 11
Amino acid sequences of lentivirus, 26, 43
Anemia, 33
Animal models, 9
Arabian foals, 33
Avian erythroblastosis virus, 119, 133
Blood donors, 49
Bovine leukemia, 81, 91
Bovine immuno-deficiency-like virus, 97
C oncovirus, 101
C.I.D. (Combined Immuno Deficiency), 33
Caprine retrovirus, 201, 235
Cat plasmid, 44, 76, 59, 207
CELISA (Competitive Enzyme Linked Immuno Sorbent Assay), 52
Chimera, 164, 167
Coggins test, 52
Colostral antibodies, 82
Cytotoxicity, 100
D.I.A. (Dot-Immunobinding Assay), 247
Diagnostic tests, 54, 81, 195, 197, 247
Expression plasmids, 44, 59
E.O.F. (Epidemial Growth Factor), 134, 139
E.I.A.V., 19, 49
ELISA, 189
Feline retrovirus, 147, 157, 167, 173, 197
Flies, 49, 50
GP 45, 26, 27
GP 51, 81, 83
GP 70, 159, 164, 187
GP 90, 26, 27, 52, 53
GP 120, 28, 129
Goats retrovirus, 235
H.I.V., 3, 40, 42, 46, 91, 98, 103, 107, 123, 129,
154, 173, 213, 231
H.T.L.V. (Human T Lymphotropic Virus), 77,
104, 214
HeLa cells, 46
Hematopoietic cell, 133
Herpes virus, 250
Horse, 24, 27, 31, 50
Hybridization, 65, 67, 69, 78, 79, 159
Hyper variable region, 130

Ideal vaccine, 10
Insect cells, 13
ISCOM (immune Stimulating Complexes), 187
L.P.R.I.A. (Liquid Phase Radio Immuno Assay)
86
Lentivirinae, 4, 189
Lung disease, 229
Lymphoid interstitial pneumonia, 230
Maedi, 223
M.H.C., 6, 230
Milk cells, 223
Murine leukemia, 4, 123
Nucleotide sequence, 68
O.P.P.V. (Ovine Progressive Pneumonia Virus)
229
Oncogene, 139
O.R.Fs (Open Reding Frames), 105, 107
P 19, 217
P 24, 10, 103
P 26, 53, 103
P 27, 176, 197
Pathogen free, 53, 167, 174
Polymerase Chain Reaction (P.C.R.), 32, 145
158, 170
Pony, 20, 27, 51
Porcine retrovirus, 111
Recombinant vaccine, 11, 83, 161, 187
Replication (in macrophages), 7, 204
Retroviridae, 4, 97
Reverse transcriptase, 111
Rous sarcoma virus, 43, 125, 126
Simian retrovirus, 245
S.I.V., 10, 98, 103, 169
Spumavirinae, 4, 149, 189
Subunit vaccine, 81
Technology transfer, 14
Transactivator (TAT), 39, 59, 105, 207
Vaccination trial, 87, 185, 199
Vaccine, 5, 6, 10, 15
Visna, 207, 213, 223, 229, 241

Figure 1.v Example of poor subject index

Chapter 1: 48
However, a problem then occurs with retrieval systems which often rely on book indexes for compiling their data. BS 6529 (1991) and its related document ISO 5963 (1991) state that the purpose of identifying concepts, usually by statistical methods on texts such as frequency and collocation, is to provide a range of indexes, from the alphabetical to the mechanised storage of data elements for subsequent retrieval. Even in these systems, subject descriptors need to be humanly edited, although they may be compared against a thesaurus of accepted terms in a particular domain. In addition, the growth of information networks means that the same indexing data may be used by different groups of people, e.g. scientists and economists, and therefore it is important that other facets of a subject should not be overlooked. In addition, 'deprecated' terms could cause viable terms to be lost, e.g.

for *condenser* use *capacitor*;
*capacitor* used for *condenser*.

Problems in indexing arise when an untutored indexer attempts to make a subject index of a subject such as *pertussis*. It is easy for such a person to commit the mistake of including a heading followed by a solid block of a large number of unbroken references (*cf.* Fig. 1.vi). This is a phenomenon which may occur in a restricted domain where the same key terms appear with high frequency, and is to be avoided. That doyen of English indexers, G. Norman Knight, has stated that 'It is generally accepted that the number of unbroken references after a heading or sub-heading should not exceed five (or at the most, seven, and then rarely).' (Knight, 1979: 105). One method of dealing with high-frequency words is to reduce the index entries to a minimum of those that would not be sought under some other heading.
England: compared to France, 23, 55, 88, 89, 103; postwar attitude to Germany, 76; relations to America, 78; and European community, 80; and colonization, 81–82, 313-314, 341; industrialization compared to France, 121, 122, 123, 124, 130, 148, 150, 153, 154, 155, 361; and Common Market, 126, 348, 349, 352; its attitude to school, 255, 256, 257, 258; its attitude to marriage, 297; postwar condition, 319, 320; foreign policy mistakes before Second World War, 321–322; reaction to French Armistice (1940), 328; French army reaction to (1940), 329; guarantees restoration of French territory, 334; and de Gaulle, 336; and Saar, 338; and Near East colonies, 339; and Suez affair, 340–341; and depression of 1930’s, 363; France follows English foreign policy (1930’s), 364; mentioned, 343, 353

Figure 1.vi Example of unbroken references in a subject index
Permutation (see example later in this section) is used as a precursor to conceptual relationships and is therefore a prerequisite for indexing multiword terms. It is possible to index simply by words and not concepts but the result is rather crude. Once a set of categories, or generic concepts, has been established, the indexer can begin to assign more specific concepts to each. On a subsequent reading of the book, it will be easier to identify what is to be included in the index. The text then requires close analysis not only to identify named concepts, but to extract hidden concepts which are not named; in addition, an anticipation of the high or low levels in the user's perception of a subject is needed. The indexer must be able to formalise the structure of the subject field, to provide extra indexing 'keys' on the assumption that labels for these terms can be incorporated, i.e. by including quasi-synonymous terms. This is where the indexer needs to anticipate a user's search requirements by making a skilful assessment of these, and even deciding whether some terms may be taken for granted, a task no computer can be expected to do simply by identifying index entries. The skilful assessment of a comparatively broad subject field is in marked contrast to the type of restricted subject field found in a monograph, where self-indexing is invoked, because in the latter the number and variety of terms is low.

The final task of the indexer is to make a synthesis of all the concepts identified, with a term being assigned which is mutually exclusive, that is, the term represents only one concept. Terms which represent new concepts should be checked for acceptability in the following sources: (i) dictionaries and encyclopaedias; (ii) thesauri (cf. ISO 2788 (1986) and ISO 5964 (1985); (iii) library classification schemes.

These requirements impose certain constraints, e.g. the subject headings of existing classification schemes may not reflect the exact
representation of concepts in a publication. Booth (1987: 143) advocates in
particular the use of computerised thesauri for indexers because they are similar
in form and function to alphabetical subject categories used in libraries, such as
the Library of Congress subject headings (1986), and because of the help they
can give in subject familiarization, meaning and authenticity checks, vocabulary
control, and standardised vocabulary for subsequent use. She mentions that most
thesauri cover a restricted subject area, such as education in EUDISED, which is
multilingual, health sciences in DHSS-DATE, and broadcast news and current
affairs in BBC Data's thesaurus of terms. Crystal (1984: 3) proposes the help of
three main stages of inquiry in linguistics: observational, intuitional and
evaluative, which could be of help in indexing, and advocates establishing the
semantic properties of what he calls 'indexese', where intuition and experience
play a key role. These criteria could help terminographers in assessing how the
'focal point' of terms should be entered and how the supportive data should be
represented.

Linguistic (morphological) analyses are also involved in automating
the procedure of morphological clipping of prefixes and suffixes to expose the
stems of words and in thus reducing the number of words to be indexed. The BSI
Root Thesaurus in conjunction with the BSI Classaurus are invaluable aids in the
indexing of English texts.

In addition, and of particular relevance to this thesis, are the
classificatory relationships that have been identified in such systems as PRECIS
(PREserved Context Subject Index) (Austin and Butcher, 1969). PRECIS handles
compound or multi-faceted subjects only and removes redundant terms such as
nominalised verbs; for example in the string wood/sawing/saws generated by an
indexer, sawing would be omitted, whereas wood/turning/lathes would need to
retain *turning* because of its lexical difference from *lathes* (Austin and Butcher, 1969: 14). These authors have established four main types of relationships:

- **generic**
  - (concept-to-type)
  - *(flower -> orchid)*

- **attributive**
  - (thing-to-attribute)
  - *(measles -> vaccine)*

- **possessive**
  - (thing-to-part)
  - *(bicycle -> pedals)*

- **interactive**
  - (action or effect)
  - *(infestation - timber)*

The first three are akin to relationships in terminology, for example 'part/whole', while the fourth represents the role of the verb. Austin and Butcher state that the general rule of classifying decides which element is the principal concept, according to the document being classified (ibid. 8) (what I have called 'orientation' in terminology). The process of permutation in indexing shows how a multiword term may be oriented in different ways (ibid. 32):

**PLANTS**


**PHOTOGRAPHY**

Plant research - Films. Development.

**FILMS. Photography.**

Plant research - Development.

**DEVELOPMENT. Films. Photography.**

Plant research.

The dash denotes the subject of the next entry. In the fourth entry, 'development' is qualified by 'films' and 'photography' on the same line; the full stops distinguish it from equally valid subjects.
In addition, relationships can be added to individual terms by the construction of term phrases. By 'term phrase', some indexers refer to a combination of two or more terms even if both are nouns, e.g. computer programming (see for example Salton and McGill, 1983; the SMART system developed by Salton in the mid 1960s); they may also refer to term stems that do not appear when compared against a thesaurus of existing words in the subject field. It is important to note at this point that Salton et al.'s use of 'term phrase' is avoided in this thesis when describing single concepts realised as multiple lexical units, for which I am using 'term'. Moreover, a single concept represented by a term is not to be confused with 'collocation' which is being used in the thesis to refer to constructions comprising more than one concept, e.g. verb + noun (phrase) which I am referring to later in LSP as 'combinants'.

Austin and Butcher claim an interactive relationship, which is of particular interest to this thesis because the very word 'interactive', implies an action or effect which is wrought by a verb and/or nominalisation of a verb, linking a product or patient which is modified by the interaction. In an interactive relationship the rule is 'class at the passive system' as can be seen from the following examples from a virology text:

- foot-and-mouth disease particle / penetration / by proflavine
- capsid / disruption / by photodynamic dyes

where the 'process' is indicated in the nominalisation of the verb. When reconstituting the sentences from the indexing entry, it is almost intuitive to resolve them into a passive construction which has the tendency of 'emasculating' the subject or 'product' (a mechanism not always realised by translators), with
the result:

- the foot-and-mouth disease (FMDV) particle is penetrated by proflavine

and

- the capsid is disrupted by photodynamic dyes

more readily than

- proflavine penetrates the FMDV particle

and

- photodynamic dyes disrupt the capsid

More interesting, however, is the need to incorporate negatives in indexing; for example, in the sentence:

- the dyes do not penetrate enterovirus particles

the proposed entry would be:

- enterovirus particles / impenetration / by dyes

It is worth noting that in indexing, as in terminology (including phraseology), verbs tend to be the 'poor relations' in favour of nouns and compound noun terms. Verbs and verb phrases do not appear to figure in special language indexes but they are included in some indexes to works which may be said to use general language; thus in a book by Alasdair Milne (1988) on British broadcasting, the examples 'joins BBC' and 'leaves BBC' are included. They also feature at the second or subordinate level of indexing, e.g.
The technique of indexing is of great use in providing guidelines for the identification of lexicographic headwords in special languages and, through passivization, produces material useful for special language verb + object phrases (see Chapter 5).

1.12 Definition of phraseology by documentalists/information scientists (LGP and LSP)

In common with that of the previous section, the domain of information science did not originally refer to linguistics for help with term and phrase identification and retrieval; word frequency, which could be provided computationally, was an early aid, whereas linguistic help in the form of the recognition of anaphora, sentence length etc. has only been incorporated more recently. Although 'content-bearing' words are sought, the aim of these is to identify a particular document and assess its difference from other documents; documentalists are looking for a topic, whereas the provision of terminological background data sought by terminologists which require the identification of a pertinent headword within a compound term or phrase, is irrelevant to them.

Considerable work has been undertaken by information scientists using the statistical methods already mentioned, such as frequency, one aim of which is
to automate the process of term recognition. This process however is fraught with the problem of defining the full form of terms, their subsequent ellipsis or 'anaphoric reduction' (author's term) in texts and their possible syntactic variation (in that they may, for example, permit internal disjuncture) in different text types. Words occurring with a frequency range which is neither at the top nor at the bottom of the scale are those adopted by information scientists engaged in indexing for retrieval purposes, and criteria have been established for the optimal 'cut-off' points (for example, Zipf's Laws in Luhn 1958; Schultz 1968; Sparck-Jones and Kay, 1973). However, this method tends not to include verb phrases, since the indexing terms being sought are almost exclusively nouns. Moreover, the world of information retrieval has moved from document retrieval to fact retrieval from full texts. It is therefore desirable for the knowledge which is extracted to be formalised and coded into independent 'blocks', in other words, a categorisation akin to that of defining subject fields which is the lot of the terminologist. More recently, Sparck-Jones and Tait (1984) have observed a trend away from text processing (i.e. document retrieval from automatically generated abstracts) towards a semantic analysis of users' requirements (fact retrieval, with author-generated abstracts based on keywords to suit the user) in the form of case-like frames, based on verbs and allowing up to two arguments. This is the type of approach which I have adopted independently in Chapter 5.

1.13 The terminographers' approach: phraseology as a sub-category of terminology (special language)

It will be recalled that there is a difference between a lexicographer of general language, and a compiler of a special language dictionary who is usually a subject specialist, often with little lexicographical or terminological training. It is reasonable to assume that subject specialists are probably preoccupied with
their own subject and may not be very interested in terminography; they regard
the dictionary as a tool, without contemplating its format too deeply. For
instance, the way definitions are couched may be rather involved and their format
may lack consistency of structure. Moreover, many users of LSP dictionaries, in
particular translators, would doubtless prefer to find them written in a style
which is easily understandable, rather than in the style used by specialists,
because the result is greater ease and speed in the comprehension of the subject
field.

Another point to be considered, which helps in the process of
assimilating information, is the method of looking up terms, to assess whether
alphabetical ordering or conceptual (i.e. subject) structuring gives the best
results. With the development of different disciplines in the Western world from
mediaeval times onwards, conceptual ordering reflected the preoccupation of the
early scientists with taxonomy and classification, particularly in the eighteenth
and nineteenth centuries. Schoolmasters compiled monolingual glossaries adapted
from bilingual dictionaries as teaching aids for their students (Landau, 1989:
39). In some instances these glossaries preceded text books because work such as
metallurgy was practised in small workshops where the techniques were passed
verbally and practically from father to son, but not written. As communications
spread and text books began to be written, so the early glossaries often emerged
from notations and explanations in the margins and between the lines of text books
(Osselton, 1983: 14), sometimes appearing in as many as six or seven languages.
However, this multiplicity of languages could give rise to confusion; in
mineralogy, for example, much was written by the Greeks, Romans and Arabs, but
the exhaustion of raw materials due to over-mining of mineral seams engendered
linguistic problems; for example, mines in Cyprus producing zinc oxide, called
spodium, i.e. ash in Greek, were closed in the fourteenth century and spodium,
which can also refer to a fine powder obtained from various substances by
calcination (Shorter OED, 1959), was replaced *inter alia* by *ivory ash*. The
need thus arose for explanations and descriptions, and hence encyclopaedic
dictionaries were used as text books and as sources for translators to make
comparative studies of the existing terminologies of the day, i.e. the lists of
quasi-synonymous terms.

With regard to current lexicographical and terminographical
practices, Heltai observes that technical dictionaries give full descriptive and
explanatory coverage but do not provide space for the way the terms may be used
in a sentence (Heltai, 1988: 38). Furthermore, they carry the full forms of
terms but do not always include their abbreviated forms, entries which are of
vital importance to translators. Frawley (1988: 191) charges dictionaries with
being vague, particularly in not providing suitable contexts to make meaning
more specific; he endorses Moulin's comments that special language dictionaries
should include terms in their specialist contexts (Moulin, 1983: 150).
Moreover, there is clearly a need for subject specialists to be prescriptive, a
trend which is less and less apparent in present-day LGP dictionaries in which
the changing nature of language is reflected.

One dictionary which to a large extent overcomes these criticisms is
the *Longman Dictionary of Scientific Usage* (Godman and Payne, 1979) which,
despite its title, has a thesaurus-like structure. Terms and their definitions are
grouped according to their subject field, each of which is indicated by a two-letter
+ three-digit code, e.g.:

\[
\begin{align*}
\text{nucleus (BIO)} &= \text{CT 023 (Cytology)} \\
\text{nucleus (BIO)} &= \text{HB 079 (Nervous System)} \\
\text{nucleus (CH)} &= \text{DB 002 (Atomic Structure)}
\end{align*}
\]
The terms are listed in alphabetical order at the end, as with a printed thesaurus, and the code number attached to each directs them to the correct subject section in the book; these are listed alphabetically within their section. Related terms are indicated by directional arrows. Since the 'dictionary' also caters for those whose mother tongue is not English, collocations are sometimes given. This is a well-thought-out work with a layout which encourages easy retrieval.

Having evaluated the largely practical approaches of a number of language practitioners in both LGP and LSP, the theoretical basis of LSP phraseology is now examined in more detail. Phraseology is a notoriously ill-defined area of linguistic research which operates at a propositional level; its propositions must be linguistically and especially technologically possible; for example, one cannot forge an idea (H. Picht, 1991, personal communication) which would however be possible as an LGP metaphor. Metaphors are therefore rare in LSP, although certain LSPs adopt metaphorical usage, e.g. psychiatry, which has to link its ideas to the tangible world, for example The mind is a telephone exchange and, more recently, analogies made between the brain and the world of computing, such as inputting, outputting and so on. In this field the use of metaphors in a particular text invariably relates to one other specialised subject field, as the use of analogy with computing shows. A proposition is thus highly dependent on its subject field or on a related one. This is in contrast to general language where metaphorical usage is common, as in the example above and in to shelve an idea.

1.13.1 The terminologist's approach: established criteria for phraseological units (phrasemes) in LSP

The following criteria for LSP phrases (for use of this nomenclature,
see below) were discussed by leading LSP terminologists at a Phraseology Workshop in Vienna in November 1989 at which the author was present. Galinski (1990: 75) stated that 'there seem to be no clear-cut criteria for distinguishing between phraseological units (phrasemes) in LSP and LGP collocations'. However, since in this thesis LSP phrases are being analysed, attempts are made to compare them with LGP phrases and collocations (cf. Chapter 3). There are several terms used in the literature (e.g. phraseological units or phrasemes; see 1.13.5) for what I shall call an LSP phrase. To establish what is understood by the expression 'LSP phrase', an attempt is made to formulate its intrinsic features. Criteria stated by various authors, both during the workshop and much earlier, indicate that it should:

- have a minimum of two elements, one with object characteristics, the other with verb characteristics (Schlomann, 1928; Warner, 1966);
- include at least one linguistic/syntactic element (rather than both elements comprising symbols, graphics or formulae) (Galinski, 1990: 74);
- be a fixed multi-word expression exceeding the scope of a term, e.g. to acknowledge receipt of (Galinski, 1990: 76);
- be no longer than a sentence (Galinski, 1990: 74);
- contain at least one concept represented by a term, also known as terminological phraseology; discussion at the Phraseology Workshop, Vienna, 1989 (see also Galinski, 1990: 76)).
- contain propositions which should be professionally and technologically correct (H. Picht, 1991, personal communication).
1.13.2 Additional observations on what constitutes LSP phrases

The author of this thesis proposes that an LSP phrase should not include:
- so-called 'free' adjectives, e.g. desired result, reliable person, which belong to LGP, unlike attenuated vaccine, which forms a lexical combination specific to virology and related medical fields (the word 'collocation' is used here in connection with 'terms' for expediency; a more refined use is explained in subsequent chapters);
- so-called 'free' adverbs, e.g. satisfactorily resolved, attractively dressed, in contrast to an adverb which is subject-specific, e.g. genetically engineered, hermetically sealed.

It is interesting to note that in a term such as freeze-drying, the adverbial use of freeze (although it is not an adverb) is an adjunct of manner, e.g. to dry by freezing. This is clearly a syntactic evolution in word formation.

In special languages, adjectives are not considered free if they are juxtaposed with a noun (where they usually take the form of a past participle used adjectivally and thus form a term with a subject-specific meaning), as seen in attenuated vaccine. The same criterion of being 'free' applies to adverbs with their combinatory verb, as in the examples given earlier in this section. Word frequency studies are an invaluable aid in designating when these combinations occur in special languages, where they are likely to appear with greater frequency than their individual components.

However, the concept of what constitutes 'free' with respect to adjectives and adverbs is highly complex. Adams (1973: 57) gives as tests for a
'free' adjective in LGP:
  - can it be qualified?
  e.g. (very) wet day but not *very small talk
  - can it be inverted?
  e.g. the day is wet but not *the talk is small

Here, the example in small talk of a 'non-free' or fixed adjective forms part of a metaphor which, as we see in 2.5.5, is a grammatical form rarely encountered in LSP, and will therefore not be discussed in this thesis.

In LSP, however, such expressions as attenuated vaccine form a lexical combination which results in a term and is therefore not to be confused with collocation. However, the vaccine was attenuated is possible, leading to the conclusion that subject-specific adjectives are attributive and not predicative. Such adjectives, as stated previously, are frequently the past participle of the verb used adjectivally; together with adverbs in LSP, they can be deduced from collocational studies but will almost certainly be evident to an informed user or intelligent layman as forming part of a term. Criteria for defining LSP adjectives and adverbs are probably not relevant here although they would be a worthwhile subject for further terminological study. Tests for setting and sub-setting need to be devised for establishing the criteria but even these would need to differ; those for medicine for example would probably differ from mechanical engineering.

Additional criteria that an LSP phrase may allow or include:
  - fixed word combinations but not fixed word order; e.g. I acknowledge receipt of your letter; receipt of your letter is acknowledged;
  - internal disjuncture, e.g. receipt of your letter is hereby
propositions which are in current use for only a certain period of time, e.g. *to scarify with poxvirus*, a technique which is obsolete;
a specific meaning according to the text in which it appears, e.g. *to type a virus*.

1.13.3 Internal disjuncture in LSP phrases

Internal disjuncture occurs to varying extents in different languages. Although English generally has the structure of subject, verb, object, nevertheless a fairly free word order is possible, particularly with regard to the position of adverbs, certain adjectives, prepositions in phrasal verbs and so on. French, in contrast, is far more rigid, due to a considerable extent to the strict regulations imposed on the language by the Académie Française; examples in French are given in Chapter 4. Among other European languages, those which are highly morphological, such as Russian and Finnish, can adopt a free word order without the sense being impeded.

1.13.4 Definitions of and observations on LSP phraseology and verb phrases

In summarizing the previous sub-sections, it is clear that more detailed definitions of LSP phraseology are needed and several by other workers have been included for the purpose of comparison (Picht, 1987: 151):

"Fachsprachliche Phraseologie" ist eine fachsprachliche Disziplin, die einerseits die syntaktischen Bindungen fachsprachlicher Ausdrucksmittel, ihre Synonymie und Äquivalenz und andererseits die begrifflichen Beziehungen sowie deren Veränderungen jener fachsprachlichen Elemente untersucht, die zu
einer fachlich gültigen und sprachlich korrekten Aussage zusammengeführt werden können.'

("LSP phraseology" is an LSP discipline which on the one hand examines the syntactic links in LSP phrases, their synonymy and equivalence and on the other hand examines the conceptual relations as well as the (semantic) changes of those LSP elements which can be combined to form a technically valid and linguistically correct statement.)

Slight modifications have subsequently been provided by Arntz and Picht (1989: 34):

'"Fachsprachliche Wendung" oder kurz "Fachwendung" ist das Ergebnis der syntaktischen Verbindung von mindestens zwei fachsprachlichen Elementen zu einer Äusserung fachlichen Inhaltes, deren innere Kohärenz auf der begrifflichen Verknüpfbarkeit beruht.'

(An LSP phrase results from the syntactic linking of at least two LSP elements to form a phrase or expression which has an LSP content, the inner coherence of which is based on the ability of the elements to combine conceptually.)

Here the fact that mention is made of 'at least two LSP elements' relates to the frequent specialised use of various parts of speech in LSP, even though in other contexts those parts of speech may be considered as belonging to general language. Since this thesis is mainly concerned with LSP verb phrases, the examples will reflect this aim. For example, the French verb *entraîner*, when used technically as in *entraîner une machine*, is translated into English as *to drive a machine*. However, in general contexts it is used more as a carrier verb
to mean to obtain, to produce, to result in, to bring about. Once a verb changes semantically in this way, it becomes an 'indexable' element.

These are broad definitions and there does not seem to have been any research that has produced a more precise one. One question that arises is how the criteria that effect a semantic change from LGP to LSP should be established - by assessing the immediate context, a particular subject field, or register which might produce a combination of both general and specialised use? It is interesting to note that the reverse can happen with certain specialised terminologies, for example, from law where 'legal' verbs may be used in LGP, as in to appeal, which has been adopted by the special field of cricket. Different valencies may result; for example, in legal terminology, to appeal the case and in medical practitioners' terminology, patients presenting with .... (+ symptoms).

Syntactic differences may occur, as in the verb to replicate, which is transitive in LGP, but which has an intransitive surface structure with an underlying reflexive element in the LSP of virology, e.g. a mutant will not replicate under conditions in which the wild type replicates; They replicate in the ovaries of female parasitoids....., where 'itself' or 'themselves' are understood; and in to plead a mistake, where in general language the noun object does not have an article, e.g. to plead ignorance. Conversely, we say to dispense with in LGP but to dispense a prescription, drug, medicament in LSP. Moreover, the verb may change its type of actant depending on the subject; the verb express has substance in special languages, as in mathematics, a quantity is expressed in terms of another, and medically, breast milk is expressed, whereas in general language the object is abstract, e.g. to express an idea or to express reservations. In most instances, however, the differences are semantic, as in legal terminology. To stay proceedings appears in the archaic sense of to check, whereas to stay in LGP is, additionally, intransitive; in a similar manner the title (to the...
aristocracy) still obtains, a verb which is transitive in LGP.

It should be recalled that the focal point of terminology has traditionally been the concept, represented by a term, and that most terms are formed from nouns or from adjectives and nouns. Once terms are past the stage of being neologisms, they tend to be prescriptive and are essential tools for standardisation; they have a timeless, unchanging, 'static' quality. Furthermore, in the past, as already stated, verbs have been the 'poor relations' of terminology; they have been neglected because there are so few true LSP verbs such as to weld, to sediment, to polyadenylate, and little consideration has been accorded to them. However, it is emerging that verbs, which are often taken from LGP, play an important role in LSP phraseology where they may combine with a noun or noun phrase and change semantically when forming a proposition, e.g. to launder money, to type a virus. A proposition thus contains at least two concepts or elements which jointly result in a higher level of knowledge.

When terms are linked to verbs to form phrases, which may occur on only one occasion in a text, the terms are brought to life by the verb which designates relationships such as action or process. As seen from the example of entraîner, verbs are semantically and sometimes syntactically subject to change when moving from LGP to LSP (although there is a greater instance of the passive voice in LSP), and they therefore have a mobile, 'dynamic' quality; Budin (1990: 64-65) calls them 'processual concepts' and states that 'The main reason for this neglect of verbs in LSP collocations and their role in phraseological units may be that the relationship between the linguistic surface level and the conceptual level of the terminologies analysed has generally been, and at times still is, regarded as a one-to-one relationship (one concept = one word)'.

Chapter 1: 67
Although verbs have a processual, dynamic quality, nominalisation produces a series of static situations requiring a carrier verb to move the situation or text along. The value of nominalisation is that by halting the action, it allows a description to be made of a sequence of (static) events, for example, of scientific method in a laboratory. Although this movement from the dynamic to the discrete is a useful abstraction, it does not represent the reality of a moving scenario. However, with the addition of passivisation to avoid mention of themselves (the 'psychological subject'), specialists frequently adopt this ploy when reporting their work.

Furthermore, there do not seem to be any theories for explaining how and at what point a verb acquires special language status. Why should a verb be considered 'free' in some instances but not in others? Attempts have been made in this thesis to establish criteria for this phenomenon by categorising different types of verbs (Chapter 4). Kjaer (1990: 15) discusses the minimal and maximal combinability of verbs and concludes that, although attention should be paid by terminologists to the study of idiosyncratic word combinations, yet establishing an inventory of semantic features to predict combinability restrictions on such combinations is not viable because of the enormity of the task. It is possible however to envisage its success in a highly restricted subject field, as the results of corpus analysis in Chapter 5 on texts pertaining exclusively to virology and bacteriology indicate. It cannot be stressed too highly that a prerequisite for undertaking such a task would be a tightly defined subject classification.

This observation is borne out by Lehrberger and Bourbeau (1988: 25) who, in their work on machine translation, define a sentence as 'consisting of a predicate (verb or adjective), its arguments (one for intransitive verbs, two or three for transitive verbs), and possibly some circumstantial elements (sentence
Concerning the passive voice, these authors state that 'two examinations of context are ... required for dealing with passives: one to determine whether the verb is, in fact, being used in the passive, and another to identify the arguments of the verb (which may be complex noun phrases whose boundaries are not obvious).' (Lehrberger and Bourbeau, 1988: 194).

Commenting on 'the nature of admissible subjects and objects (arguments) of the verb, i.e. selectional restrictions', (Lehrberger and Bourbeau, 1988: 15), these authors also state: 'Within the field of machine translation it is common practice to use semantic criteria during syntactic analysis by checking for semantic compatibility between predicates and their arguments or between modifiers and heads' and that such compatibility 'can be used effectively in restricted semantic domains' (Lehrberger and Bourbeau, 1988: 49). These authors' use of the term 'semantic compatibility' is what I shall equate with the finding of collocational relationships (cf. Chapter 4).

Finally, a definition of a 'terminological phrase' given in ISO 1087 (1990) is that it is a 'phrase in which at least two elements (terms) are syntactically linked and thus form an utterance with subject-field-specific content; the internal coherence of the elements is based on their conceptual combinability.' Examples are *booking a flight, the tide ebbs and flows* or in Swedish, *avkunna en dom* (to record a verdict).

It can thus be seen that there is considerable fluidity in different workers' perception of LSP phraseology, which makes it a challenging area for study.
1.13.5 Nomenclature of phraseology

There are several different nomenclatures for LSP phrases that have appeared in the literature and which need to be explained:

(a) LSP phrase
- proposition containing a minimum of two elements (concepts), one with object characteristics, the other with verb characteristics (Schlomann, 1928; Warner, 1966)

(b) terminological phrase (De = Fachwendungen, fachsprachliche Wendungen)
- noun + verb clause (Kjaer, 1990: 4)

(c) phraseme, phraseological unit (De = Phraseologismen)
- fixed multi-word expression (includes idioms) (Galinski, 1990: 76)

(d) phraseological term
- multi-word term, the meaning of which is not deducible from the meaning of its components (including idioms), i.e. it has restricted commutability (quoted in Kjaer, 1990: 12)

It appears that (a) and (b) have the same meaning, while (c) and (d) are very similar, and in fact (a) and (b) could be incorporated into (c) and (d). The nomenclature 'LSP phrase' with the meaning given in (a) is being adopted in this thesis.

1.14 Conclusions for special language phraseology

The importance of the subject field and, more narrowly, the context within that field cannot be stressed too highly when identifying terms and phrases.
Yang (1986) *inter alia* uses frequency to identify technical terms; however, this must finally be a subjective task because the terminologist in conjunction with a subject specialist must decide what constitutes a term in the domain. Moreover, concordancing is used to show examples of a term or phrase in context, an invaluable aid for encoders of a language such as translators. Different text types reveal how different terms and phrases are used and the collocations that appear in a particular subject field.

Research is usually undertaken on either general texts or on subject-specific texts; what is interesting and needs to be studied is what happens when these areas, or two LSPs, overlap, as is the case with legal texts which by definition treat some other subject field. As has been observed, the behaviour of verbs can alter the number and type of actants (for example, *patients present with*) and this phenomenon is examined in Chapter 4 of this thesis. Semantic differences have already been noted (for example, the verb *entrainer* in French); there are also syntactic differences between LGP and LSP (for example, an LSP will have few personal pronouns; it will have a greater number of verbs in the passive voice than LGP, and these will be in the simple present, past and future tenses, and so on) (Sager, Dungworth and McDonald, 1980). In particular, a subject field will have its own lexical peculiarities. These observations help to shed light on the 'behaviour' of verbs and verb phrases in LSP and this aspect will be developed in subsequent chapters of the thesis.

In Chapter 2, the different types of multiple lexical units and phrases which have been identified in the light of LSP are discussed and a comparison made with LGP. In Chapter 3, collocation is assessed, both from the point of view of others who have worked in this area, and on the observations of the author who makes comparisons between LGP and LSP. Verb phrases are analysed in Chapter 4 to show how collocation and valency can give indications of conceptual
relationships, while in Chapter 5, a model is proposed for verb combinants in special languages, based on analyses from corpora in English and French drawn from current texts in the biological sciences. Conclusions concerning the feasibility of a global model for special subject fields are drawn in Chapter 6 and recommendations made for future research.
Chapter 2

The problems of the terminographer: identification and ordering of multiple lexical units and phrases in LSP; comparisons with LGP

The problems of terminographers are specifically addressed in this chapter, particularly their need to recognise the forms LSP terms may take (2.2), the computational methods they may use to identify terms in a specialised field (2.3), the problems of headwords and their 'orientation' (2.4) and finally a comparison of multi-word terms in LSP and LGP (2.5). These analyses lead naturally to semi-restricted word groups, i.e. phrases, in subsequent chapters of this thesis.

The computational methods are elucidated so that conclusions may be drawn to help with entering terms and phrases in two media:

- a printed dictionary, to provide consistent entering of multiple lexical units which comprise many of the terms and phrases in special languages;

- a computerised system such as an indexing system or a terminology data bank (term bank), to provide a 'focal point' in a multiple lexical term or phrase according to the subject field and the needs of the user. This ensures that the relationships between terms, in the case of indexing, and additionally the correct informative data, in the case of term banks, are included to support and constitute the entry.
The suggestions of other workers in this sphere, particularly those of terminographers working in current term banks, are included. Some practical solutions to the problem have already been published recently by the author (Thomas, 1992; 1993) and, because of their relevance to this chapter, a summary of proposed guidelines for terminographers is presented as Annex 1. This chapter contains a detailed analysis of multiple lexical units, including phrases. In the first instance, however, the requirements of terminologists are discussed.

2.1 **The role of terminologists: identification of users**

Identification of the users of terminology and an assessment of their requirements is a prerequisite for terminology work. One type of 'learner' of the LSP will probably be a translator who is presumed to have native competence in the target language (TL), i.e. mother tongue. Picht and Draskau (1985: 20) state that in addition to a knowledge of the general language of both source and target language, 'two other vital elements come into play: a) professional competence in relation to the special field; [...] (b) linguistic proficiency in the LSP of both mother tongue and target language'. Since work in most scientific and many technical fields is conducted in English, it is likely that the translator will be a non-native speaker of English translating into his/her own language, the TL, in other words, a 'decoder' of the source language (SL).

Other users will be experts and students of a particular subject who can be discounted as learners of the terms because the field is already familiar to them. However, the converse occurs when experts write and give conference papers in a language, usually English, which is not their mother
tongue; these workers are 'encoders' of the TL. Many speakers of languages other than English are adept at creating nouns to express what they want to say, to the detriment of verbs; the tendency is clearly evident in speakers of French, German, Polish and Russian. This 'Nominalstil' is a form of rather pedantic precision which is found in more formal and official styles of writing and discourse, e.g. you must sign would appear in German as Unterschrift leisten. More examples are given in Chapter 4.2 which serve to illustrate and corroborate the contrastive translation studies of Vinay and Darbelnet (1958).

2.1.1 Importance of terminology training

For the terminologist and translator attempting to delimit terms in a specialised field with which they are not conversant, recourse to an expert in the subject is a prerequisite. This is borne out by the observations of the author of examples of work produced by three students who were set the task of eliciting terms from English texts on automotive engineering. None of the students had terminology training. Students 1 and 2 were students of translation and non-native English speakers; student 2 had some knowledge of the domain. Student 3 was a native English speaker and a student of linguistics but had no knowledge of the domain.

Student 1 found the term balanced operation as an example of an LSP term in the texts. This is incorrect because balanced is not part of an LSP term but is a 'free', LGP adjective. The student's finding confirmed a lack of familiarity with the domain and, because terminology training was also lacking, failure to differentiate between LGP and LSP in that balanced was not recognised as a free, LGP adjective; the student thought balanced related to 'weighing' or 'swinging'. Student 2 fared better in recognising more LSP terms, e.g. blipping; even though this is not solely an automotive
engineering term, the student recognised it as technical 'jargon' because he was familiar with this term being used of computers making a repeated noise. The native English speaker again had problems recognising LSP terms from the domain and included baseline conditions, because this term formed a sub-title of a list given in the text being perused, even though the term does not relate to the domain in question. The conclusion drawn from these examples takes Picht and Draskau's argument for the importance of terminology training a step further, in that familiarity with the domain through recourse to a subject specialist, with whom terms and background information can be discussed, is of equal importance to familiarity with the foreign language in determining LSP terms. The necessity of not stinting on the services of a subject specialist is also emphasised by Lerat (1995: 47) when he states:

'En fin de compte, le critère des critères est l'avis du spécialiste, dont on ne saurait raisonnablement tenter de faire l'économie.'

It also confirms that terminologists, like translators, should ideally only work in their mother tongue. In addition, since terminographers are often required to produce terminologies for emerging domains, for which standardised terms are often not available, they resort to descriptive terminography as a natural precursor to prescriptive terminology which results later through a consensus of opinion.

So far in this section, a number of theoretical and to some extent practical problems and requirements have been enumerated, but they do not specify precisely how terminologists should reach a decision on entering a term. In the rest of this chapter their requirements are elucidated and
attempts made by term bank operators, machine translators and indexers are assessed, as well as those of linguists of general language.

2.1.2 Necessity of guidelines for terminographers

The first task of the terminologist is to invoke and structure the important concepts in a subject field, and the second is to provide the lexical or symbolic form which will represent those concepts as succinctly as possible. It is desirable and indeed promulgated among technical writing advocates that the maxim of "one term, one concept" be followed as far as possible. What happens in practice, however, is that there are frequently too many terms in relation to the number of concepts identified for a particular field (this should not be confused with the overall number of terms which it is possible to lexicalise, the number of which is finite, whereas the number of concepts that exists is limitless). This plethora of terms is true particularly in an emerging domain where the naming process tends to be more arbitrary than when a domain becomes established and its lexis more prescribed.

As already expressed, guidelines are needed to help a terminologist (a) decide what constitutes the limits of a term in LSP and (b) choose the headword or focal point which will represent a multi-word term or phrase. Sager, Dungworth and McDonald (1980: 233) recognise the problems of identifying 'extended terminological units' which the special lexicographer or terminographer faces when 'making his decision about the unity of a term with reference to the knowledge structure of a discipline. He has a narrow set of criteria for weighing the evidence and his work may therefore be more prescriptive as a result. He establishes terminological units such as laterally-reversed, multiple-start screw thread which are not considered
lexicalised in general language (see next paragraph). [...]. The identification of extended terminological units causes difficulties' (my underlining). Sager et al. (1980: 268) distinguish between three major types of compounds:

(i) those where the nucleus refers to the object;
(ii) those where the nucleus refers to a property or disposition;
(iii) those where the nucleus refers to a process or operation, identified by 'of', 'by means of'.

The term 'lexicalisation' needs to be clarified and the basis for it explained. In machine translation lexicalisation is an empirical strategy for entering multiple terms, such as those given above by Sager et al., which may prove problematical to translate if the constituents were entered alone. The basis for lexicalisation may be tested by frequency; if a multiple lexical unit appears more often than the number of its individual components, it may be deemed a lexicalised term. However, it may appear only infrequently in a conceptual structure, or its presence be understood by the textual material with which it is surrounded; for example, a term comprising ten lexical units, albeit rare, such as \textit{opacity factor negative M serotype 18 group A isolate A992} is unlikely to appear more frequently than its individual components. These problems have been noted by the author \textit{vis-à-vis} (i) indexers who anticipate the perceptions of the reader in identifying potential search terms which however may not appear in the text of a book and (ii) the inclusion of 'anaphoric reduction' which includes the formation of acronyms and initialisms, e.g. AIDS, HIV, FMDV.

In 2.4 an appraisal has been made of the terminology literature to ascertain the criteria for what constitutes the limits of an LSP term; different forms of multiword terms have been identified in 2.5, while

Chapter 2: 78
guidelines for choosing a headword are, as stated, provided in Annex 1.

As far as the necessity for guidelines is concerned, it is clearly important to consider the task of the terminologist, for whom, as has been seen, the starting point is the identification of concepts in a subject field and the assignation of terms to those concepts. Pugh (1984) seeks conceptual motivations for the formation of special language compounds, stating that the naming process is based on concepts and on the conceptual classification of a subject field. She gives the following conditions of noun compounds and thus provides a working definition, that they must (i) result from a combination of two or more otherwise lexically autonomous elements where 'autonomous' covers both the paradigmatic and inflected but otherwise autonomous form, and (ii) they must exhibit grammatical and semantic cohesion, in other words, behave as a single word. This is where the help that can be given by collocation is important and this factor will be analysed in the next chapter.

2.2 Formation and delimitation of LSP terms

The formation of terms has been discussed at length by Felber (1984) and by Picht and Draskau (1985), *inter alia*. ISO 1087 (1990) states that the definition of a term may be 'any conventional symbol for a concept which consists of articulated sounds or of their written representation (= of letters). A term may be a word or a phrase.' Felber (1984: 181-2) gives the following requirements which should be met by terms. They should be:

- accurate;
- concise;

Chapter 2: 79
- easily spelt and pronounced;
- capable of allowing the formation of derivatives;
- linguistically correct.

The last two requirements occur particularly in the adoption of terms by one language from another. This often occurs in countries which are in the throes of technological development and whose language exponents adopt terms, largely from English but also from French, the latter in the case of the former French colonies in Africa and the West Indies. The terms need to be readily assimilable, morphologically, orthographically and phonetically. For example, in the Welsh language, even where Welsh language verbs appear, there is often in addition an adaptation of an English verb with the suffix o or io, e.g. riteiro = to retire, and drivo = to drive. Also in Welsh, English or Breton/French words may be adapted orthographically, e.g. iot = yacht, modur = motor, sinema = cinema, eglwys = église = church. In Japanese, a re-phoneticised form of English may appear. In some countries such as Czechoslovakia and Germany, an effort has been made not to borrow terms but to produce terms which are homogeneous with the native language, e.g. television in German = Fernsehen, a literal transcription from Classical roots.

In standards and similar documents, terms should in addition, if possible, be:

- monosemous and at the same time mononymous;
- a member of a term system.

If these requirements cannot all be met, careful examination is needed to

Chapter 2: 80
decide which should be given priority. However, it is necessary to define exactly the limits of an existing synchronous term in LSP, for example its semantics or register, so that a terminologist is clear about how many or how few words represent the actual concept. Not only is clarity of paramount importance, but also transparency of meaning, by which is meant the sum of its parts. Ideally, only one term should be assigned to a concept and the use of synonyms deprecated, to avoid ambiguity. The term should be as short as possible while at the same time clearly representing the concept it is describing. For example, the following shortening of a term in German could lead to ambiguity (Picht and Draskau, 1985: 116):

- **Maschine zum Hobeln** (machine/tool for planing)
- **Hobelmaschine** (machine/tool for planing)
- **Hobel** (plane/plane) (i) machine; (ii) tool; (iii) person.

LSP terminography can be either descriptive, where terms for existing concepts and their relationships gain currency in use, or prescriptive, where terms are designated by standardising bodies and terminology commissions (Felber, 1984: 189).

**2.2.1 Automatic recognition of compound terms**

Ananiadou (1988) has proposed methods for the automatic recognition of terms, concentrating in particular on the properties of neoclassical compounds. She has made use of four broad classes of technique: (i) frequency; (ii) stemming; (iii) probability and (iv) positional (syntactic). The basis of using frequency arises to a great extent from the indexing techniques of the 1960s and 1970s which have been analysed by such workers as Luhn, Schultz, van Rijsbergen, Sparck-Jones and Kay and others, and refined by Yang (1986), as discussed at the end of the previous chapter.
Stemming and morphological analyses are particularly useful techniques in computational linguistics, and in LSP the syntactic position and probability of terms is an additional pointer to automatic term recognition, which requires analysis at the phrase or sentence level; this is also necessary in machine translation. Despite providing a strong theoretical framework, Ananiadou still recognises the difficulties in performing 'statistical operations on a unit whose extent and composition cannot be reliably identified' (Ananiadou, 1988: 448), the perennial problem of types and lemmas.

2.3 Lexical frequency as a method for identifying terms in a subject field: corpora analyses of LSP texts

It is worthwhile showing how computational analysis such as that mentioned in the previous paragraph can help with the identification of terms within a particular subject field. As might be expected, the probability of the occurrence of a term can be corroborated by its frequency. A comparison has been made by the author between two highly restricted corpora of virology texts in English and French with the aims of ascertaining whether the most frequently occurring words are representative terms from a subject field (2.3.2) and whether they contribute to identifying that subject field. This is a useful technique for the automatic extraction of terms. In addition to observing the patterns in the combinations of words, it can be shown how the collocation between verb and object can help with the identification of LSP phrases. The results of the last aim are not shown here but are discussed in Chapter 4.
In the example given, the subject field is known beforehand. The texts have been chosen because I considered them to be as nearly parallel as possible in terminology, style and synchrony; in other words, they are thematically as well as lexically and semantically homogeneous. They are taken from major textbooks on virology (Topley and Wilson, 1984; Maurin, 1985) and letters and articles in 'Nature' (1985-87) and are therefore didactic at an advanced level of specialism, and informative, as in the 'Nature' articles where the readership may be classed as being experts in closely related fields, for example between a chemist developing a vaccine and a medical doctor requiring specific, current information on the subject. The corpora are deliberately small in this instance, to maintain the homogeneity and synchrony of the genre: the English corpus comprised c. 47,000 words, while the French corpus comprised c. 53,000 words. Small corpora have been deemed sufficient for the purpose of providing the examples which follow because it is possible to draw conclusions from a relatively small number of words when dealing with highly restricted subject fields whereas, due to the diversity of texts, this is not true of general language corpora. However, later in the thesis analyses of much larger corpora have been reported, each containing over half a million words of French and English texts relating to the biological sciences which I have compiled, to ascertain whether statistics can be corroborated in two different sized corpora and to provide examples of combinants which would not be viable from a small corpus.

The approach of using parallel texts for comparative purposes naturally requires that each analysis be undertaken monolingually, revealing an inevitable dilemma because it tends to countermand that of conceptual multilingual terminology, where the initial step is to structure the subject
field conceptually and to assign terms to those concepts in the required languages. An analogy may be made that terminology represents hierarchy or verticality, whereas textual analysis is linear or horizontal.

2.3.1 Frequency of technical and sub-technical words in English

The first analysis, to ascertain how frequency denotes a subject field, was made on the English corpus, on single lexical terms. The method can however be adapted to the identification of multiple lexical units which form terms, by observing combinations of words. The single lexical terms analysed were classified as 'technical' and 'sub-technical', as advised by Antoinette Renouf of the COBUILD project. The technical words chosen were those likely to have a high frequency in the virology corpus, while the sub-technical words are defined here as those which appear in both LGP and LSP. Technical words generally act as semantic 'carriers' and form part of a technical term comprising nouns or adjectives + nouns, e.g. *African green monkey kidney cells*. Sub-technical words in technical texts tend to indicate relationships such as cause or hierarchy between the technical words. Twelve 'key words' were chosen for this study in English in the technical category based on the author's knowledge of the subject. Seven verbs were chosen in the sub-technical category; these included noun/verb homonyms and one instance of the past participle used adjectivally. There was also one adjectival past participle in the technical group. A separate comparison was made between the base form *antibody* and *antibodies*, showing these words appeared with more or less equal frequency (81 and 98 respectively), while the forms *attack*, *attacked* and *attacks* appeared 1, 5 and 5 times respectively.

Concordances and word frequencies were run on the corpus. Sub-technical terms proved to be, as expected, less subject-specific; as can be
seen from Table II.1, only one verb from the sub-technical words occurs with a frequency comparable to that of the technical words. Function words were not excluded from the concordances.

Table II.1 illustrates the procedure for the identification of terms and shows the results of the author's research into the technical and sub-technical words, indicating their frequency relative to the total number of words (i.e. percentage). The high frequency of the words antibodies, disease, vaccine and virus confirms that a clear indication of the subject field can be apparent from this type of analysis and indicates the feasibility of machine recognition of terms in a particular field.

A comparison has been made between the earlier, smaller corpus of 47,000 words described earlier in 2.3 with the later, larger corpus of half a million words, to ascertain whether usage has changed over a period of seven years (1987 to 1994; sources of texts comprising the large corpora are given in Acknowledgements). Some highly subject-specific words such as virus and vaccine have remained remarkably stable but others, such as peptides, which might have been expected to increase in usage because of current research trends, have not. This unexpected result can be attributed to slight differences in the subject nature of the texts in the corpora.
<table>
<thead>
<tr>
<th>Selected word ('token')</th>
<th>No. of occurrences in virology corpus of 47,000 words</th>
<th>Percentage of total</th>
<th>No. of occurrences in virology corpus of 1/2 m. words</th>
<th>Percentage of total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Technical words</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>virus</td>
<td>486</td>
<td>1.034</td>
<td>4665</td>
<td>1.068</td>
</tr>
<tr>
<td>disease</td>
<td>178</td>
<td>.379</td>
<td>839</td>
<td>.192</td>
</tr>
<tr>
<td>vaccine</td>
<td>103</td>
<td>.219</td>
<td>973</td>
<td>.222</td>
</tr>
<tr>
<td>antibodies</td>
<td>98</td>
<td>.209</td>
<td>266</td>
<td>.061</td>
</tr>
<tr>
<td>tissue</td>
<td>45</td>
<td>.096</td>
<td>173</td>
<td>.039</td>
</tr>
<tr>
<td>peptides</td>
<td>41</td>
<td>.087</td>
<td>44</td>
<td>.010</td>
</tr>
<tr>
<td>species</td>
<td>41</td>
<td>.087</td>
<td>598</td>
<td>.136</td>
</tr>
<tr>
<td>immunity</td>
<td>35</td>
<td>.075</td>
<td>142</td>
<td>.033</td>
</tr>
<tr>
<td>synthetic</td>
<td>20</td>
<td>.043</td>
<td>50</td>
<td>.011</td>
</tr>
<tr>
<td>attenuated</td>
<td>17</td>
<td>.036</td>
<td>248</td>
<td>.057</td>
</tr>
<tr>
<td>culture</td>
<td>14</td>
<td>.030</td>
<td>280</td>
<td>.064</td>
</tr>
<tr>
<td>neurological</td>
<td>7</td>
<td>.015</td>
<td>11</td>
<td>.003</td>
</tr>
<tr>
<td><strong>Sub-technical words</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>cause</td>
<td>19</td>
<td>.040</td>
<td>188</td>
<td>.043</td>
</tr>
<tr>
<td>attacked</td>
<td>5</td>
<td>.011</td>
<td>1</td>
<td>.000</td>
</tr>
<tr>
<td>involves</td>
<td>4</td>
<td>.009</td>
<td>44</td>
<td>.010</td>
</tr>
<tr>
<td>reach</td>
<td>3</td>
<td>.006</td>
<td>8</td>
<td>.018</td>
</tr>
<tr>
<td>remain</td>
<td>2</td>
<td>.004</td>
<td>18</td>
<td>.041</td>
</tr>
<tr>
<td>reveal</td>
<td>2</td>
<td>.004</td>
<td>13</td>
<td>.029</td>
</tr>
<tr>
<td>yield</td>
<td>1</td>
<td>.002</td>
<td>20</td>
<td>.045</td>
</tr>
</tbody>
</table>

Table II.1 Word tokens and frequency relative to total number of words
2.3.2 Comparison of frequency from parallel texts in English and French

A comparison of concordances run by the author (Thomas, 1988) on both French and English corpora indicated, as might be expected, marked similarities of frequency in technical and sub-technical words in the two languages. The English word *vaccine* and the corresponding French *vaccin*, clearly technical, subject-specific words, appear perhaps not surprisingly as frequently as .22% in the English virology corpus and .16% in the French virology corpus. An example was taken of the technical word *culture* which has a frequency of 0.3% in both corpora. The word appears frequently as a component of a compound term, usually in English as *cell culture* or *tissue culture*; these terms occurred in 71.43% of instances and in no instance were the components interrupted. Further examples of terms taken from Singleton and Sainsbury (1987: 240-241) where *culture* is a conceptual 'atom' with many subordinate 'molecules' include *(an)*aerobic *culture*, *axenic culture*, *batch culture*, *broth culture*, *plate culture*, and many others. In the French corpus, 48.39% of instances showed the combination *culture cellulaire*, with 3.23% of instances of *culture de cellules* (Table II.2). It is clear from a comparison of these percentages that *culture cellulaire* is the terminological unit, whereas *culture de cellules* is used as a variation in discourse.

In conclusion, the technique of concordancing, with its word frequency and word order, gives clear indications of the terminology of the texts which in turn leads to an identification of the subject field. Furthermore, the syntax and collocatory patterns in the texts pave the way for studies on phraseology which are highlighted by concordancing.
2.3.3 Some syntactic characteristics of LSP

The differences in syntax between LGP and LSP have proved interesting. As mentioned, in the often rather formal style characteristic of LSP, there is recourse to a number of specific syntactic characteristics such as:

1. **Nominalisation of the verb**, e.g. *hospitalisation*. The verb is rendered static by nominalisation, as though the actants have been emasculated; the verb has lost its 'movement' or action of admitting someone to hospital. Indeed, the style of scientific discourse is used to distance people from the action, and this is reinforced by the way in which the act of nominalisation seems to infer an element of the passive voice modality of the verb from which it is derived, in addition to requiring that the semantic burden remains on the noun in a somewhat anodyne fashion. This contrasts with the style advocated by writers of technical manuals and text books on technical writing who advocate a more dynamic style in sentence construction. Verbs are seen to play the role of moving the text along, i.e. they have a 'processual' role.

2. **Verbalisation of terms.** The ease with which morphological endings can be added in English renders this a popular form to adopt, particularly in speech, e.g. *the vaccine is being field-tested; the power failure caused the program to be exited*.

3. **Adjectival use of the noun for ontological linkage**, e.g. *measles virus*.

4. **Adverb + adjective.** The fusion is more likely in LSP than in LGP,
providing a semantic integration of two concepts to form a third, e.g. *electrostatically controlled, dynamically operated*. The hyphen plays an interesting role in English; where the adverb + adjective construction immediately precedes the noun it describes or qualifies, the two words are usually hyphenated, e.g. *well-defined argument*. However, in sentences such as *this argument is well defined*, when written without a hyphen, the sentence seems to have a temporal aspect, i.e. *this argument has been well defined*, as described by the perfective in Russian. If the sentence is written *this argument is well-defined*, it seems to have more immediate currency. This phenomenon appears to apply to both LSP and LGP.

It is interesting to speculate at what point the integration of the adverb with the adjective has taken place. It may be reasonably assumed that there is a nominal form to which the adverb and adjective construction refers, for example *this device is electrostatically controlled/dynamically operated*, where the verb is semantically empty. The context is important here and would doubtless throw light on whether the *controlled* in *electrostatically controlled* is also empty; does an *[[electrostatically-controlled] device]* differ from an *electrostatic device*? It is clear that a *dynamic device* is not the same as a *[[dynamically-operated] device]*.

5. **Adverbial noun + noun.** In addition to the above categories, there is the adverbial use of the adjectival form of noun + verb which is a common construction in both LSP and LGP, e.g. *machine-assisted translation, typewritten document, hand-embroidered blouse, home-made cakes*. In these examples, the adverbial noun + verb concept acquires an adjectival use and combines with the second noun to form a third, integral concept which, as in the example above in (4.), produces a terminological unit in its own right. It
is interesting to note that *machine translation* could mistakenly be viewed as being the same as *machine-assisted translation*, whereas they are in fact two distinct concepts.

In scientific and technical texts, very often there are instances of what I am naming ‘anaphoric reduction’ which occurs when the complete form of a term appears in the first instance and is subsequently abbreviated, not solely by acronyms or initialisms but by an abbreviated form of the whole term, e.g. *foot-and-mouth disease virus* may be abbreviated to *FMDV* or referred to colloquially as *foot-and-mouth*. These anaphoric reductions provide operational surrogates for the terms with which experts are familiar but which could be a pitfall for the uninitiated reader, for example a translator or someone in another subject field; for example, *AI* can mean *artificial insemination* or *artificial intelligence*.

The above examples, with the exception of (1.), show the terminological practice of combining two or more semantic elements or concepts to form a third. English produces such facile forms of term production because of its lack of morphology (see 2.) which, however, could create problems in inflected languages such as German which need to retain their morphology and therefore resort to other techniques such as concatenation. It is important to note that, because examples such as those given above are integral terminological entities, each representing a concept, they are not considered to be collocations. This significant difference is expanded in Chapter 3.

2.3.4 Verb and noun homonyms in computer retrieval

Homonyms pose a well-known problem in computer handling, the noun/verb interchange being particularly prevalent in English 'journalese'.
Although Lehrberger and Bourbeau (1988:9) state that 'the computer could be programmed to look for the last noun when there is more than one, since that is the head noun in English' (an observation with which I have taken issue; see the example of *measles vaccine* in Chapter 1), they give examples of the problems that can arise from noun and verb homonyms, e.g.:

*The passenger flight arrival time changes every summer.*

and

*The passenger flight arrival time changes will be posted.*

In instances such as these, it is clear that a syntactic analysis needs to be carried out to resolve the problem.

### 2.4 Problems in ordering headwords and identifying 'focal points' in terminography

As discussed in the previous section, multiple lexical units as headwords or entry terms in special language are common phenomena, with inherent difficulties concerning which lexical unit to adopt as the point of entry for the term as a whole. This is not only a problem for the terminographer but also for the dictionary user and emphasises the imperative need for the user’s requirements to be adequately addressed by both terminographers and lexicographers. While it is true that computer retrieval of compound terms has the advantages of speed and storage, nevertheless, given that the average user will probably consult printed dictionaries for some years to come, rather than computerised pocket dictionaries, thesauri and term banks because of the higher cost of the latter group, it is still important to bear in mind the importance of ordering and
placing compound dictionary entries in both LSP and LGP.

As stated, it is clear that, from a computational point of view, rapid searches can be made through vast quantities of data and the actual lexical unit chosen as the entry term to represent a collocation or other multiple lexical entry in, for example, a term bank, is largely immaterial, because a computer's search facility can retrieve a requested lexical unit and hence a term very easily. What is important, however, is that (a) the terminologist building, for example, a term bank, needs to know how to determine which word in a multiple term is the optimum one to adopt as a headword or 'entry point' for the term, to enable the most precise data relating to that term to be chosen and inserted into the record format; (b) the user requires guidance when making a search for a term so that the required information appears at the first query, thus saving time, and (c) products derived from term banks such as glossaries on paper, which will invariably be in alphabetical order, need to be compiled in a consistent manner. In other words, it is highly desirable for a collocation or other multiword term or phrase to have one of its lexical units designated in the mind of the terminologist as its 'focal point' and this will depend on the 'orientation' of the term according to the needs of the user; there is a restricted number of lexical units in a multiword term whereas there is an infinite number of potential users. It is therefore necessary to establish a working number of user groups or user 'profiles'.

2.4.1 The focal point in indexing methods

An important area in which the identification of a focal concept within a multiple lexical unit comprising more than one concept occurs is that of indexing. In their work for the British National Biography, Austin and Butcher (1969) examined methods for adding subject data in the form of
descriptors to provide the basis for a pre-coordinated index following a conceptual analysis of a subject. The rules of 'classing', or classifying, are based on the recognition of relationships between entities (i.e. things, whether concrete or abstract) and their attributes (properties or activities), or between one entity and another. One element of a compound subject, or term, is designated the principal, or focal, concept and the other elements follow in a predetermined fashion, each being introduced by an appropriate relational operator, which is designated in terminology as a meronymic relationship, i.e. 'contains' or 'part of'.

'When more than one entity is discerned in a compound subject, the principal concept is decided by considering the relationship which exists between them. Two main relationships apply in this case, the possessive and the interactive .... (which) ensures that concepts are set down in a consistent, predetermined order suitable for a machine system.' (Austin and Butcher, 1969: 5, 6). The possessive (or generic relationship in terminology) is classed before using operators to introduce the part (the specific relationship in terminology), e.g. bicycles - wheels - spokes. In other words, this is a part/whole or meronymic relationship. In the interactive relationship, the rule is 'class at the passive system', where 'the product, the patient or the system which, from the viewpoint of the author, is most likely to be affected or modified by the interaction' (Austin and Butcher, 1969: 10). An example of passivisation, the timber was infested by wood-boring insects translates in indexing terms as infestation of timber by wood-boring insects; nominalisation has occurred from the passive form of the verb and this important point is expounded in depth in Chapter 5. Sometimes redundancy occurs, as in sawing/saws and ploughing/plough, and in such instances one term can be omitted, as explained in Chapter 1. Although the system was originally constructed for English, a number of multilingual versions have
been tested and, with necessary adaptations, these have been used successfully (Lamy-Rousseau, 1974; Supper, 1975; Austin, 1976; Austin and Sørensen, 1976; Lambert, 1976; Horsnell, 1977).

In his later work, Austin (1976: 9) comments on the need for 'differencing' compound terms in English; entry should be under the first word except where one word has more significance. This however poses the problem of how to decide which of several words is the most significant. Austin emphasises the need to differentiate between the 'focus', which will be a noun or noun phrase identifying the general class to which the concept belongs. In the example oak-veneered furniture, furniture is the focus. The differences are assigned to two levels: level 1 applies directly to the noun, e.g. veneered, while level 2, the indirect level, qualifies some other differential but not the focus per se, e.g. oak (Austin, 1976: 10).

The operators adopted by the NATO Classification Research Group in which Austin was also instrumental are as follows:

(0) Observation - the observer, his techniques and equipment
(1) Property, structure, material
(2) Sub-system
(3) Interaction within system
(4) Second system related to (5)
(5) Effect produced upon system - 'normal' events and system maintenance
(6) Second system related to (7)
(7) Effect produced upon system - metastatic change or 'detrimental' effects

Chapter 2: 94
The 'second system' will be translated into 'circonstants', an important feature of the verb frame developed in Chapter 5, where it is linked to valency. It can be explained at this stage by the preliminary systematic questions (in italics below) which an indexer should ask (Dykstra, 1985: 9); the representative comments following the questions are mine and relate Dykstra's questions to later work in the thesis:

- *Did anything happen?* = verb representing action;
- *If yes, to whom or what did it happen?* = object, or 'patient' according to Austin; beneficiary in Martin (1993: 227); my term = 'recipient of process';
- *Who or what did it?* = subject, sometimes psychological, or agent;
- *Where did it happen?* = mandatory actant or optional circonstant; the apparent discrepancy here is a crucial factor in the thesis (cf. also Halliday, 1985; Martin, 1993).

The NATO operators handle three types of relationship:

(a) the attributive: operator (1)
(b) the possessive: operators (2) and (3)
(c) the interactive: operators (4) to (7)

For example, 'Painting of office walls' would first be set down as:

Office (2) Walls (2 [and] 5) Painting
Dykstra (1985: 112) classifies the three relationships of PRECIS using slightly different nomenclature, represented by the following syntax:

1) the *predicative* relationship, expressible by "to be"
2) the *possessive* relationship, expressible by "to have"
3) the *interactive* relationship, expressible by "to do"

The indexing system devised by Austin and Butcher provides useful guidelines for terminologists in the analysis of main and subordinate concepts. Moreover, the process of automatic indexing relies on and reflects statistical data, which are also of use in the second type of computational system which requires an assessment of multiple lexical units, that is the terminology data bank or term bank.

### 2.4.2 Results of questionnaire on the ordering of multiple lexical units in term banks

To ascertain the criteria used by term banks on the entering of multiple lexical units and phrases, a questionnaire (Annex 2) was sent to 74 term banks worldwide (Annex 3), concerning *inter alia* the order in which multiple lexical units were inserted, in addition to the search terms. Replies were received from 30 term banks in which 25 gave information on their on-line search and retrieval mechanisms. Annex 4 shows the extent of these mechanisms, with some term banks using more than one method.

The survey provided the following 'rules' for ordering multiple lexical units:

- strictly alphabetical, from the left-hand word in European languages, including function words;
respect for the sequential order of words, e.g. escala de Beaufort;

by frequency or weight, a 'logique de pondération', in which the headword is a noun or verb, i.e. the word carrying the 'highest weight', and is never a preposition; in addition, terms are stripped of 'free' adjectives.

Phraseology and compounds are treated in some term banks as if they are single lexical units, with collocations being given under 'usage'. The Catalan BTERM-TERMCAT does not consider phraseology that permits internal disjunctures to be part of terminology but does however include fixed phrases, e.g. en línea, fora de tempo (A. Puiggené i Riera, 1990: personal communication). TERMIUM proposes creating a separate database for idiomatic phrases with a keyword approach (C. Leonhardt, 1990: personal communication); at present phrases are stored in one set of records in TERMIUM's Linguistic Data Base (1993).

Auger and his colleagues (Auger et al., 1988: 39), in their work with the Office de la langue française in Québec, comment on the problems in delimiting complex terms. After discarding extraneous discourse material, the remaining 'unités' require lexicalisation; after 'priming' by frequency analysis, the terminological units appear. Some terms may comprise several lexical units, e.g. séparateur magnétique à bandes croisées, yet they appear as a unit in texts and need to be included as a compound term in the terminology. However, Auger states that certain compound terms are in fact descriptive syntagmata which appear only in catalogues or lists and are not found in technical texts nor in specialist 'jargon'; terms of this type may designate complex concepts which can be reduced. Even though a long phrase

Chapter 2: 97
may represent a well-defined concept to a technician, if the phrase appears infrequently in a text, Auger advocates splitting it into segments, e.g. *vitesse d'avancement du front de taille* becomes *vitesse d'avancement* and *front de taille*, the presence of *du* being an indication that the whole term is not lexicalised. There is an important typological difference between French and English which is recognised here; French tends to be more discoursal, while English is more 'holistic'. This is borne out by Hoffmann who comments on the noun + genitive (noun phrase) or noun + genitive + prepositional phrase constructions in Romance languages, in which the genitive may take the definite or indefinite form, e.g. *traitement du/de signal* may also be in the plural, giving four entries (Hoffmann, 1987: 159). The comparatively recent trend in French in which prepositions are dropped in the specifications and manuals tendered by industrialists may lead to ambiguities, and has been commented on by Thomas and Judge (1989: 413).

While it is possible to search rapidly for any word, this involves unnecessary effort because it would doubtless result in too many 'hits' in a number of the searches made. It would seem necessary, where hard copy is to be produced for the use of translators and interpreters, that either the terminology is displayed in a conceptual structure or that some sort of alphabetisation be undertaken. Schulz (1980: 223), describing the policy of Siemens' TEAM term bank, states that terms are stored in their basic form, as in a dictionary: for example, nouns are usually in the singular, verbs in the infinitive. This also applies to complex terminology units, such as multi-word terms and compound names; these and phraseology units are recorded in their canonical word order, e.g. one in which the German adjective precedes the noun it modifies, e.g. *symbolische Adresse*, would
appear under *symbolische*. Although this format would be recognised by a subject specialist as a term, I would take issue on whether this would pose problems for the user who may not be fully conversant with the subject field but who would in all probability be familiar with the traditional lexicographical practice of looking for a term under the noun before searching for it under the adjective. If, however, *Adresse* were to appear without being qualified, then cross-referencing would have to be applied.

It may be considered by some that the position of the modifier before the noun which occurs on the left-hand side in some European languages such as English, German and Russian is less logical than that of, for example, the French language which has adjectives as post-modifiers. Polish is interesting in that where the juncture of the adjective + noun construction is semantically 'loose', the adjective is placed before the noun, whereas the adjective is in second position if the two words form a semantically 'tight' unit. English heraldic, 'royal' and legal language, with its Norman precursor, retains examples of this construction, e.g. *lion rampant, blood royal, treasure trove*.

It would appear that terminographers using computerised methods do not as yet feel a need to take the indexing methods of rotation, faceting, chaining and so on into consideration, but at present use a rather simpler form of cross-referencing of two of the elements of a term, which tends to be a rather subjective, unstructured approach; alternatively they rely on the computer's search mechanism to produce in some cases, as stated, too many 'hits'. The string searches used by information scientists reflect what is on the surface but this method does not take account of the underlying concepts invoked after hyperlemmatization in examples such as *revulcanisation*.
Eismann, on the other hand, advocates a 'strikt formales grammatikalisch-alphabetisches Prinzip':


"If the phraseological unit [phE = phraseologische Einheit] contains a noun, it will be explained under this entry, and examples will be given. If two or more nouns are present, the phrase will be entered under the first. If the phE has no nouns, but has an adjective, it will be entered under this. The order after this is verb, adverb, pronoun, numeral, preposition. Any phE will be entered as many times as it contains separate words, and a reference will be given to the location of this description".

This approach is probably the one a user would adopt more or less intuitively, at least in European languages, although there may be instances when a verb might be chosen before an adjective.

2.4.3 Search mechanisms in term banks

The requirements of the users, who were identified in 2.1, were taken into consideration by the different term banks when adopting methods for retrieving terms (see Annex 4). The pressing need was the retrieval of multiword terms and phrases, and the following methods were elucidated as a result of the enquiries:
on an individual word;
- under the specific term (Determinans, or modifier) rather than
the generic (Determinatum, or modificand) because the result
would be too many hits, e.g. valve resulted in more than 100
entries but air gulp valve produced only two (Schmitt, 1989:
156): one term bank advocates 'the rarer the better';
priority is given in the Ruhrgas term bank to (a) the complete
multi-word term and (b) terms containing more words than the
term entered for retrieval, i.e. those with internal disjuncture. A
truncation facility is available.
in the TEAM term bank, if a multiword term or phrase is not
found, the programme looks for individual components, e.g.

Question: Rückspulen zur Bandanfangsmarke
Wanted: English translation
De: Anfangsmarke, Rückspulen zur
En: rewind to BT marker

This is easier to accomplish in some languages than in others,
depending on factors such as inflection, concatenation, hyperlemmatization
and the delimiting of multi-word terms.

- on any combination in a multiple word group, eliminating
'noise', 'stop' or 'transparent' words, e.g. articles, conjunctions
and common prepositions: these are language-specific
(C. Leonhardt, personal communication 1990: see Table II.2);
English: a, an, and, at, but, by, for, from, in, nor, of, on, or, the, to, with
French: a, au, aux, avec, d', dans, de, des, donc, du, en, et, l', la, le, les, ni, ou, par, pour, sur, un, une
German: am, an, aus, bei, das, dem, den, der, des, die, für, im, in, mit, über, und, vom, von, zu, zum
Spanish: a, con, de, del, e, el, en, la, las, los, o, para, por, u, un, y

Table II.2 'Noise', 'stop' or 'transparent' words eliminated by the Canadian Term bank 'Termium' for indexing purposes in English, French, German and Spanish.

Another Canadian term bank, the minibank BATEM developed by Baudot and Clas at the University of Montréal, lists multiword terms by constituents which provide access. Multiword terms are restricted to noun phrase, verb phrase, adjectival phrase, adverbial phrase or a combination, hence prepositions are omitted. Put simply, these types of phrases are linked across the languages. The following table shows a bilingual example of how synonymous terms may be displayed:
Table II.3  Example of bilingual multi-word concept from BATEM
(reproduced from Baudot and Clas (1984: 51)).

<table>
<thead>
<tr>
<th>Language 1</th>
<th>Language 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. blade pitch reversal</td>
<td>A. inversion du pas d'hélice</td>
</tr>
<tr>
<td>1.1 blade</td>
<td>A.1 inversion</td>
</tr>
<tr>
<td>1.2 pitch</td>
<td>A.2 pas</td>
</tr>
<tr>
<td>1.3 reversal</td>
<td>A.3 hélice</td>
</tr>
<tr>
<td>1.4 blade pitch</td>
<td>A.4 pas de l'hélice</td>
</tr>
<tr>
<td>2. reversal of the propeller</td>
<td>B. mise en renverse de l'hélice</td>
</tr>
<tr>
<td>2.1 reversal</td>
<td>B.1 mise</td>
</tr>
<tr>
<td>2.2 propeller</td>
<td>B.2 renverse</td>
</tr>
<tr>
<td></td>
<td>B.3 hélice</td>
</tr>
<tr>
<td></td>
<td>B.4 mise en renverse</td>
</tr>
</tbody>
</table>

Legend: 1, 2 .... synonymous terms in language 1
A, B .... synonymous terms in language 2
1.1, 1.2 .... constituents pointing to 1, etc.
A.1, A.2 .... constituents pointing to A, etc.

2.4.4 A comparison: ordering of headwords in LGP

Given the strong tradition of lexicography in the U.K., an appraisal is made of work in this area so that comparisons can be made with LSP terminography. Cowie (1983: 103) remarks that: 'There are few features of dictionary organisation on which editorial policy differs as strikingly from one dictionary - or lexicographical tradition - to another as the treatment of
Opitz (1983: 173), on the alphabetic ordering of headwords, has stated that 'the lexicographer should take pains to make it a foolproof instrument. His first objective must be consistency, particularly in respect of the listing of compounds. A linguistically aware person may not see much difficulty in deciding which of the components of a compound should be the headword under which the entire expression is listed, but unsophisticated users do unless they can be sure of a certain principle that is applied throughout the dictionary and is at the same time simple enough to be acted upon by all users. To the extent that they are unaware of an established system of nomenclature, their first impulse will indeed be to presuppose the simplest and most consistent alphabetical order, which with Western languages begins with the item on the left-most side of the compound and ends at the right regardless of the semantic or syntactic value of its single components. Yet it is amazing how frequently these simple facts are disregarded by compilers of technical (as well as general) dictionaries who, while generally following the alphabetical principle, superimpose upon it an unrelated conceptual hierarchy within entries or when arranging headwords'.

Newmark (1991: 60) gives an example of the difficulty of searching for the headword within a compound in Collins-Robert F-E/E-F dictionary with *coup de tête* when, having looked through the many entries for *tête*, he was referred to *coup* and was then obliged to delve through two columns of *coup*. The author of this thesis has had a similar difficulty with *entrer en ligne de compte* which was finally found under *compte* in Harrap's *New Standard French and English Dictionary* Part 1 Vol. 2, ed. J.E. Mansion (1972), after searching under *ligne* and then *entrer*. Having to make
three, sometimes lengthy, searches in a heavy tome is tiring and time-consuming, and highlights the imperative need for the user's requirements to be adequately addressed by the lexicographer/terminographer, as well as emphasising the advantages of computer retrieval.

2.4.5 Precedence in nominal groups in LGP

Computational lexicography, that is, dictionary compilation relating to LGP, has been included because, despite the excellence of retrieval facilities, without linguistic input these are unlikely to be of very great help in the placing of headwords from nominal groups, which ultimately must be a semantic judgement. Sinclair has analysed massive corpora of current English LGP texts from which to elicit examples of words in use. In the case of nominal groups linked by *of*, he states that 'It is reasonable to expect the headword of a nominal group to be the principal reference point to the physical world' (Sinclair, 1991: 87). However, in a large number of cases, N2 is closer to a concrete physical object than N1:

\[ \text{the shapes of simple organisms} \]
\[ \text{the position of France} \]

In these examples, N2 would be the headword of the phrase. Often N1 is used for focus or support (Sinclair, 1991: 87, 89), an interesting observation which can be applied to special languages, with the addition of different relationships as indicated in Sager et al.'s categories listed in 2.1.2.

In the examples above, where the head or focal point of the noun phrase is at the end, it can be seen that English creates emphasis by its syntagmatic structure, and the same is true of French (*la position de la France*) and German (*die Stellung von Frankreich*). This is typical of
languages which follow a subject -> verb -> object pattern. However, the Latvian equivalent would translate as *France-of position* and in Turkish, which has the order subject -> object -> verb, the phrase would appear as *France-of position-its* (F. Knowles, 1993, personal communication).

Warren (1978) and Amsler (1987) have suggested that an analysis of the dictionary definition is clearly a step nearer to identifying the focal point of any compound term. However, the problem may arise of 'type-token', where whole texts are analysed and terms may take different forms either because of different registers within text types, or assume a different form in the same text, such as abbreviations, acronyms or initialisms, for which, as already mentioned, I have coined the term "anaphoric reduction". This point has also been noted by Ananiadou (1988: 467). Research on ascertaining how prevalent this tendency is and at what point in a text the phenomenon might occur would be of enormous benefit to translators.

2.4.6 Summary of 2.4

It can be seen from the above survey that simple word retrieval presents no problem in a computerised medium. However, from the terminographer's point of view, a subjective and semantic assessment is needed to give guidance on the placing of a term or phrase so that the correct supporting material can be provided. The analysis needed when choosing a suitable headword or entry term, together with the 'orientation' of the term or phrase as a whole is, as already mentioned, focussed in the guidelines for terminographers given in Annex 1.

While it is undisputed that computer-driven methods for the ordering of multiple lexical terms and phrases are highly efficient retrieval
mechanisms, term banks do not generally appear to use them to centre on verbs and verb phrases. The technique of frequency is widely used in lexicography to provide evidence backed by statistical data but in terminography, which records subject-specific terms and phrases, this method carries an inherent risk of overlooking those terms and phrases which appear only occasionally and which may be difficult to identify because of their internal disjuncture or other difference in form, but which should nevertheless be included. The identification of these terms and phrases is virtually impossible unless the corpus of text is exhaustive. In this respect, collocation is a highly useful guide for the identification of noun and verb phrases and this will be discussed in detail in Chapter 3.

2.5 Multi-word terms and phrases: comparison of LSP with LGP

As stated in the Introduction, different forms of compound terms have been identified, and it was for information on these different forms that term banks were requested to provide their data (see earlier in this chapter and Annexes 2 and 3). Most had not attempted to refine these points but were content to allow the computer to make random searches. It is therefore relevant to define what is understood by the designations given to the various forms of compound terms or word groups. In this chapter, the forms which have been analysed are:

2.5.1 Compounds
   2.5.1.1 Compound nouns
   2.5.1.2 Compound verbs
   2.5.1.3 Compound adjectives
   2.5.1.4 Compound adverbs
2.5.1.5 Adjective(s) + noun(s)
2.5.1.6 Adverb + adjective
2.5.1.7 Adjective + adverb

2.5.2 Phrasal and prepositional verbs

2.5.3 Idioms

2.5.4 Similes

2.5.5 Metaphors

2.5.6 Phrases (the verb + noun construction in particular is included in Table II.5 and analysed in depth in Chapter 4)

These designations vary slightly from those in the questionnaire sent to term banks (Annex 2). This is because information on the other forms mentioned (connotation; verbs and prepositional phrases; lexical syntagms; syntagms of discourse/phrasal units; extended terms) was not forthcoming and it may be assumed that these either proved irrelevant in the context of the LSP domains found in term banks, as would be the case with connotation and, for the most part, verbs, or else had not yet been addressed (my underlining) (e.g. phrases, phrasal units). It is only comparatively recently that the more advanced term banks such as DANTERM have begun to include LSP phrases, although at present these are in the guise of contextual examples. As seen earlier in this chapter (2.4.3), the Canadian banks are including phrases and this trend is being taken up by other term banks. Furthermore, some categories of compounds have been included in this section for the sake of completeness which were not specified in detail in the original request.

The examples given in the rest of the section are taken from both LGP and LSP in English for the purposes of comparison.
2.5.1 **Compounds**

Compounds represent a single 'subsetted' concept but their elements function independently in other circumstances. The terminologist, Felber (1984:180), emphasises the correspondence of a term with the definition of its concept: 'the term should be composed of the most significant characteristics of the concept to be designated. It should not contain elements which contradict the definition [...] but (should include) only those (concepts) needed to distinguish the concept in question from related concepts. The principles to be followed in forming a complex term are therefore those governing the formation of definitions and of concepts'. One guideline for LGP is given by Hausmann (1985: 120) who advocates the avoidance of free or unnecessary word combinations of the type new town, old house, which can be formed without help by a learner of the language.

2.5.1.1 **Compound nouns** Allan, as a linguist, (1986: 225) states that: 'Compounds are lexemes composed of two or more free forms, e.g. blackbird, air-conditioning, etc.' Noun compounds may be solid (LSP = red leaf; LGP = black bird), hyphenated (LSP = bee-paralysis; LGP = tape-recorder), or open (LSP = catalytic converter, LGP = picture frame). In English the progression is usually made from open via hyphenated to solid as a term gains currency, e.g. air crew in 1938 became air-crew in 1941 and aircrew at the end of the war (observed by N. Osselton). The trend in special language to describe a concept by a number of co-occurring words may be comprehensible to a specialist in the subject but poses problems for non-specialists and in particular to translators, e.g. in the term Douglas fir tussock moth nuclear polyhedrosis virus, it is difficult to know whether the word tussock refers to the fir or the moth without a
considerable knowledge of the subject. However, there is a propensity to hyphenate two nouns which are used adjectivally, e.g. *reverse-cycle heat-pump ventilation*, which leads to clarity, unlike the often indiscriminate hyphenation practised nowadays by even so-called serious newspapers, which appear to use unsophisticated hyphenation software packages, a practice which can lead to initial comprehension problems; however, it would seem that hyphenation in English is on the wane.

Allan does not include in his categorisation of compounds those nouns linked by a 'preposition' such as *suit of clothes, sister-in-law*; many examples of this construction can be classed as metaphors, e.g. *piece of cake, hole in one, time of day*. The use of prepositions to link two nouns to form a compound is prevalent in LSP, particularly *of; on and at* are rare, e.g. *bill of sale, payment on account, barrister at law*.

2.5.1.2 Compound verbs Personal observations show that compound verbs, or verbalised nouns (and see also 2.5.2 on phrasal and prepositional verbs), are prevalent in American English (LSP = *to mouth-pipette*; LGP = *to air-condition, to mass-produce*), i.e. N + N or N + V, and would not appear in a dictionary under their generic term *pipette, condition, produce*). This type of term formation is known as 'conversion' (Picht and Draskau, 1985: 111).

When the verbs in the above examples are considered independently, their valency structure differs from that of the compound form which loses a valency slot because one of the elements is already incorporated into the verb structure. This emphasises the point that valency applies to both nouns and verbs, and needs exploring semasiologically and
onomasiologically.

2.5.1.3 **Compound adjectives** Compound adjectives (LSP = laryngo-tracheo; LGP = hard-wearing, wide-ranging), are also found. In the LGP examples the adjective cannot stand alone without the adverb which is being used adjectivally.

2.5.1.4 **Compound adverbs** No instances of compound adverbs have been noted in LSP such as *down-under, far-off* in LGP. However, it is possible to add the American suffix *-wise* to a wide range of topics, e.g. *biology-wise* might be envisaged.

2.5.1.5 **Adjective(s) + noun(s)** This is a common form of term formation in LSP (Sager, Dungworth and McDonald, 1980: 234). Adjectival nouns appear in complex terms such as *African green monkey kidney cells*.

2.5.1.6 **Adverb + adjective** In LSP, the adjective is usually a past participle, e.g. *hermetically sealed, genetically engineered* and the colloquial *the biologically-inclined virologists, such as epidemiologists*.

2.5.1.7 **Adjective + adverb** In LSP this construction is rare and rather colloquial (personal observation), e.g. *quantum mechanically*.

2.5.2 **Phrasal and prepositional verbs**

These are also sometimes known as compound verbs, which may cause confusion with the examples in 2.5.1.2. They generally consist in English of a verb plus adverb particle, i.e. a phrasal verb (*turn off the light, see in the New Year*) or a verb + preposition, i.e. a prepositional verb (*look after the child*) (Quirk and Greenbaum, 1975: 349). The difference
between the examples is that in the first, the particle can also appear after the object (*turn the light off, see the New Year in*) but this transition cannot be made with the second example (*look the child after*). As far as machine recognition of phrasal verbs in context is concerned, a semantic problem may occur as in:

\[\text{He ran up a big bill: he ran it up}\]

and

\[\text{He ran up the street: he ran up it}\]

Examples of phrasal verbs in LSP will most probably be those which can add the preposition *up*, e.g. *to set up an experiment/to set an experiment up*. Phrasal verbs tend to be support or 'carrier' verbs.

2.5.3 **Idioms**

'An idiom is a language expression whose meaning is not determinable from the meanings [...] of its constituents' (Allan, 1986: 236), e.g. *hot-dog, pig-in-a-poke, marché aux puces*, that is, it is not transparent.

'An expression which functions as a single unit and whose meaning cannot be worked out from its separate parts' e.g. *He washed his hands of the matter* (*Longman Dictionary of Applied Linguistics*, 1985).

As far as this thesis is concerned, the definition of 'idiom' will correlate with those given above.

'A sequence of words which is semantically and often syntactically
restricted, so that they function as a single unit e.g. *It's raining cats and dogs*, where semantically the individual words do not produce the meaning of the idiom as a whole and syntactically the idiom is invariable, e.g. *It's raining a cat and a dog* (Crystal, 1985: 152). However, *hot-dogs are possible, although* *to buy pigs-in-a-poke and *wash his hand of the matter sound odd.*

Furthermore, the definitions of idiom are far from being accepted globally. Newmark (1990: 154) gives a definition of 'idiom' which appears to contradict those given above, stating that idioms have 'meanings which are deducible from their component words (e.g. in all conscience)' whereas phrases are 'transparent (e.g. a nice turn of phrase)'. It seems that there is room for confusion between Newmark's two definitions and also between the different versions to define idiom given above. Moreover, there seems to be some overlap in definitions to distinguish between idioms and phrases; probably the most notable difference is that an idiom is semantically based, whereas a phrase is more syntactically oriented.

Lehrberger (1982: 94) uses 'idiom' as a technical term for any multi-word expression in a machine translation (MT) dictionary but it should be emphasised that this use of 'idiom' is specific to MT and differs from its linguistic interpretations, even though these are not clearly defined, as seen earlier. Lehrberger gives the following criteria for entering a given LSP expression as an idiom in the parsing dictionary of these systems:

- the meaning of the expression is not predictable from the meanings of its components, e.g. *with respect to, nose gear, finger tight;*
- translation idioms: the corresponding expression in the target language (TL) is not predictable by the usual rules for associating structures in the TL with those in the source language (SL), e.g. aspect ratio (Fr: allongement), DC power (Fr: courant continu), buttock line (Fr: section longitudinale) and here Lehrberger would use the 'first match' approach;

- the expression occurs so often in the LSP that it 'feels like' a compound word: landing gear;

- the expression occurs very rarely and its parsing would require undesirable changes in strategies developed to handle the majority of cases in the LSP, e.g. right and left of center occurs only once in the corpus, while right and left occur quite often. It is extremely difficult in such cases to ascertain which of the units should be included in the database and how they should be represented; probably the best method would be to systematise them ergonomically as a result of users' queries.

2.5.4 Similes

'A simile is an expression in which something is compared to something else by the use of a function word such as like or as' e.g. Tom eats like a horse (Longman Dictionary of Applied Linguistics, 1985).

As may be seen from Table II.5, similes are not found in learned LSP works, probably because facts are being reported; similes may however appear in popular technical communications (cf. footnote to Table).

2.5.5 Metaphors

Very briefly, 'something is described by stating another thing with which it can be compared' e.g. His words stabbed at her heart (Longman

Chapter 2: 114
As with simile, metaphor is not a figure of speech found in learned scientific work dealing with tangible substances but it is found in the abstract areas of psychiatry, as with the example already mentioned, *the mind is a telephone exchange*, and in popular technical fields.

2.5.6 Phrases

Allan (1986: 230) states that phrases may admit internal disjuncture (e.g. *a black bird = a bird which is black; the man in the street = the man who is in the street*) , in contrast to compounds (2.5.1) which do not (e.g. *a *blackbird* = a species of bird; *the man-in-the-street* = the average person). In contrast to an idiom, the meaning of a phrase is transparent.

Table II.4 is a summary of the multiple lexical groups identified earlier in sections 2.5.1 - 2.5.6 and shows a comparison between LSP and LGP.

2.5.7 Conclusions to 2.5

The above categories of word groups fall into two main categories:
- restricted word groups or fixed phrases, e.g. idioms;
- semi-restricted word groups, i.e. those permitting internal disjuncture, ellipsis or what I have called 'anaphoric reduction' (and which therefore differentiates phrases from idioms).
### Multiple lexical units

<table>
<thead>
<tr>
<th>Multiple lexical units</th>
<th>LSP</th>
<th>LGP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Compound nouns (<em>cf.</em> 2.5.1.1)</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Compound verbs (i.e. verbalised nouns) (<em>cf.</em> 2.5.1.2)</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Compound adjectives (<em>cf.</em> 2.5.1.3)</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Compound adverbs (<em>cf.</em> 2.5.1.4)</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Adjective(s) + noun(s) (<em>cf.</em> 2.5.1.5)</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Adverb + adjective (<em>cf.</em> 2.5.1.6)</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Adjective + adverb (<em>cf.</em> 2.5.1.7)</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Verb + adverb particle (<em>cf.</em> 2.5.2) (phrasal verb)</td>
<td>?</td>
<td>✓</td>
</tr>
<tr>
<td>Verb + preposition (<em>cf.</em> 2.5.2) (prepositional verb)</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Idioms (<em>cf.</em> 2.5.3)</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Similes (<em>cf.</em> 2.5.4)</td>
<td>*</td>
<td>✓</td>
</tr>
<tr>
<td>Metaphors (<em>cf.</em> 2.5.5)</td>
<td>*</td>
<td>✓</td>
</tr>
<tr>
<td>Phrases (<em>cf.</em> 2.5.6)</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Verb + noun (<em>cf.</em> Chapter 4)</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

**Table II.4** Comparison of multiple lexical units in LSP and LGP

? May be used colloquially; unlikely to be found in formal written communications.

* Rare in LSP (although see 2.5.5 and the field of psychiatry); their popular use can be found in the media in the LSP of sports reporting, particularly football in the U.K., where the metaphors *over the moon* = delighted, *magic point* = penalty spot and similes such as *sick as a parrot* = very despondent, abound. In German football parlance, metaphors are popular, for example the
phrase Jemanden auf Sattel grün ziehen (lit. to pull s.o. to the green saddle) = to pull s.o. to the ground, Angstgegner = a bogey team (one that is difficult to beat) and Sonntagsschuf = a fluke (Gray, 1990). The reason for this colourful use of language is sociocultural; the broadcasts aim to amuse by appealing to as wide an audience as possible, humour being the common denominator, whereas scientific and technical usage is more formal and is in fact 'preaching to the converted', hence the lack of metaphors and similes (but, again in the media, advertisements aim to promote technical household objects, e.g. cars and cameras, by appealing to a possible non-technical user).

2.6 Conclusions to the identification and ordering of terms and phrases

In this chapter a number of the problems which face terminographers have been discussed and, it is hoped, resolved and a set of guidelines introduced (Annex 1). In Chapter 3 the combinatorial properties of the word groups in Table II.5 are analysed, leading to recommendations being given for terminologists regarding the placing of these groups. A more specific form of multiple lexical unit, verb phrases in both LGP and LSP, is analysed in Chapter 4 to ascertain whether the verb is always the 'focal point' of the phrase and to what degree the subject field influences its degree of speciality.
Chapter 3

Grammatical and lexical collocations:
orientation of their 'focal point'

Having identified the various word groups that combine to form terms and special language phrases, a more rigorous approach to their categorisation is considered necessary. Collocation is clearly an important linguistic phenomenon to incorporate in this study and the reason for assessing LGP collocations in the first instance has been to explore the different avenues which general language offers and to consider the way in which LSP may or may not be influenced by them. The two main points elaborated in this chapter are the differences between multiple lexical units and collocations; comparisons are made and the topics are subsequently treated separately. It cannot be stressed too strongly that a clear distinction needs to be made between collocation as it is used in LGP, and the combination of lexical units which form an LSP term; a term representing a single concept and consisting of a multiple lexical unit is not considered to be a collocation, although in its formative stage, collocation may have been instrumental in bringing it to a definitive form, after which it no longer permits internal disjuncture. Béjoint and Thoiron (1992: 515) recognize the problem and summarise it succinctly, stating that:

',.... les collocations ne sont pas des unités du lexique: tout ce qui est de l'ordre du mot - les mots composés par exemple - devrait être écarté. [...] les critères de différenciation entre [...] terme et "non-terme" ne sont pas toujours limpides.'
('.... collocations are not lexical units: everything pertaining to
the nature of a word - compounds for example - should be discounted. [...] 
criteria for differentiating between [...] term and "non-term" are not always
crystal clear.')

However, special language phrases which extend beyond the scope of the term
do reveal examples of collocation. Both forms are treated in this chapter.

Comparisons have been made in the chapter between so-called
'lexical collocations' in LGP and LSP with two aims: firstly, to help identify
their entry point in a printed work such as a dictionary and secondly, to
ascertain the 'orientation' of (LSP) terms comprising multiple lexical units
so that different types of terminology users, be they terminologists who
provide the most appropriate data pertaining to a term, or subject
specialists, information retrieval specialists, translators or others, may
acquire the data most appropriate to their needs (cf. 3.5.1).

Where comparisons are made between 'lexical collocations' in LGP
and LSP (cf. sections 3.3 and 3.5), the term 'collocation' subsumes, for ease
of reference, the terminological units analysed in these sections, particularly
in 3.5.1 and 3.5.2. However, wherever LSP is referred to alone, and
particularly in section 3.7, the terminological relationships are referred to
as they are more appropriate to the notion of terminological unity, in
contrast to collocation which implies a dual element consisting of two words
which are attracted to each other. An important step therefore has been to
analyse the criteria of different researchers when defining collocation and to
ascertain whether there is more than one type which relates to LSP.
Furthermore, it will be seen that collocation in LSP terminology is a

Chapter 3: 119
prerequisite of phraseology, whether or not a phrase admits internal disjuncturc.

Collocation in special languages is clearly a pointer which needs to be taken into consideration when studying phrases. To produce examples which will provide statistical support for LSP phrases, computational methods involving concordancing and word frequency have been undertaken by the author. Benson, Benson and Ilson (1986: xxvii) use 'collocation' in LGP to refer to a concept described by two nouns, e.g. jet engine. However, to reiterate my statement in the introductory paragraph of this chapter, and since the thesis refers to LSP, hence terminology, I regard a concept which is represented by a multiple lexical unit as a single terminological unit, e.g. foot-and-mouth disease; it is therefore not a collocation. On the other hand, any recurring but optional lexical juxtapositions are deemed to be collocations, e.g. foot-and-mouth disease of cattle, as opposed to that in other cloven-hooved animals which are also susceptible to the disease. A working distinction is therefore made between multiple lexical, i.e. terminological, units, and collocations that are formed from viable juxtapositions of words and are therefore phrases. However, as Béjoint and Thoiron (1992: 515) have pointed out, the distinctive line is indeed a fine one.

3.1 Definitions of different types of collocation

Work on the topic in LGP, in particular by Benson et al. and Hausmann (Benson et al., 1986; Hausmann, 1979; 1985), has formed a basis from which comparisons have been made by the author with regard to LSP. In sections 3.1 - 3.3, the term 'collocation' is used to encompass both 'collocation' and 'colligation', a term used by some authors to refer to fixed verb phrases, which in LSP I am naming 'combinants' and which are
discussed further in Chapters 4 and 5, where an indication of how to introduce order to the collocatory patterns found in combinants will be shown.

According to Gläser (1980), there is no generally accepted term under which composite lexical units such as idioms and indeed collocations can be subsumed, although Mitchell (1971: 57) has introduced the term 'composite element' to cover idioms, collocations and compounds; according to these authors, 'phraseology' is not suitable because it does not distinguish between compound lexical units and the phrase structures in which they function. In the first instance, the development of a phrase is arbitrary, as with a term, and it may only acquire a more stable form through usage. Given this difficulty and apparent vagueness, it is clear that a detection mechanism for phrases needs to be developed by means of statistical analysis and this has been undertaken for the LSPs studied.

More recently, Martin (1992: 157) has stated that collocations fail between idioms and free word groups, preferring the term 'restricted word groups'. He identifies two parameters in LSP, the communicative situation and the semantic domain, characterised by lexical, syntactic, semantic and pragmatic levels. He refers to the conceptual combinability of the attributive n. + n. or n. + adj. construction in LSP as 'collocations', albeit in inverted commas (see my first paragraph of this chapter). Both Martin (1992: 159) and Heid (1992: 528) refer to the following categories of combinations:
Having given a few definitions of collocation which are pertinent to the work in this thesis, a basis needed to be chosen from which to embark on an analysis of collocatory patterns in LSP. Because of their detailed studies on collocations in general language, the work of Benson et al. (1986; Benson, 1989) has been taken as a starting point. These authors divide collocations into two major categories and these are assessed in the following section.

3.2 Grammatical and lexical collocations

Benson divides collocations into two types: 'collocations may be grammatical or lexical and may be defined as "recurrent word combinations"' (Benson, 1989: 3). Within these categories of collocation, Benson et al. (1986) identified a number of different types in LGP and their observations have been used in the thesis as a basis for identifying and comparing collocations in LSP.

In their Introduction to The BBI Combinatory Dictionary of English, Benson et al. (1986: ix) describe a grammatical collocation as 'a phrase consisting of a dominant word (noun, adjective, verb) and a preposition or grammatical structure such as an infinitive or clause.' A lexical collocation however does not normally contain prepositions,
infinitives or clauses and typically consists of nouns, adjectives, verbs and adverbs (Benson et al. 1986: xxiv). Here the dominant word will always be the same part of speech whereas, as will be seen in Chapter 5, in the LSP verb + noun combinants studied, the dominant word will depend on the orientation of the phrase.

Benson et al. (1986: ix-xxviii) give a number of examples which differentiate between grammatical and lexical collocations. In the next two sections a brief summary of the categories is given with comments on the application of these authors’ conclusions to lexicographical practice.

3.2.1 Grammatical collocations

The following list gives eight types of grammatical collocation promulgated by Benson and his colleagues specifically for inclusion in their combinatory dictionary.

1. Noun + preposition (but not normally + of or by, since these are predictable), e.g. a witness to, the power within.

2. Noun + to + infinitive (e.g. a duty to do it), or noun + verb -ing (a pleasure doing it). They do not include noun + to + infinitive if in order can be inserted, e.g. I eat to live, nor do they include instances where the noun is preceded by a descriptive adjective (e.g. an interesting book to read). They also preclude instances where the infinitive can be replaced by a relative clause (e.g. a procedure to follow = a procedure that is to be followed) and colloquial phrases such as those often found in advertisements.

3. Noun + that clause, (e.g. he took an oath that he would do his duty),
including noun + that + present subjunctive (e.g. it was his wish that his estate be divided equally), but not if that is interchangeable with which and is not an impersonal clause + prep. + n, (e.g. it was pride that....)

4. Preposition + noun (e.g. by accident, on purpose)

5. Adjective + preposition (e.g. fond of, annoyed by/with)

6. Predicate adjective + to + infinitive (but not in order to) (e.g. it was good to see them)

7. Adjective + that (e.g. afraid that), including adjective + that + present subjunctive.

8. 19 verb patterns (pp. xiv - xxii), for example, those which allow the dative movement transformation, (e.g. he sent the book to his brother/he sent his brother the book).

After reflection it is not considered that the categorisation of these verb patterns will add any pertinent information to this thesis; nevertheless the fact that in the thesis verb constructions are being analysed requires that sub-section 8. above be borne in mind. Colligation is noted in some instances, although the number of actants is not mentioned. The latter fact is, however, important with respect to special language where a valency deviation from the general language 'norm' may be perceived, e.g. patients present with (symptoms), and the verb insert, a transitive verb in LGP, may in virology be used as vectors.... should insert with high efficiency and yet be stable. Valency will be studied in more detail in the next chapter and a model for special language verb phrases proposed in Chapter 5.

Chapter 3: 124
3.2.2 Lexical collocations

The following six categories of lexical collocation in LGP are also from Benson et al. (1986: ix - xxviii).

1. Verb (usually transitive) + noun denoting:
   (a) creation or action + noun/pronoun/prepositional phrase (e.g. reach a verdict);
   (b) eradication or nullification + noun (e.g. annul a marriage).
In English, the verb is frequently semantically 'empty' e.g. annulment of the marriage was carried out/effecte/d concluded. This results in a choice of verbs often being available to fill the role of collocation with a deverbalised noun; it is common for one part of a collocation to have greater variability than another, although interestingly it is not always the same part, as is seen in examples given later in this chapter.

2. (a) Adjective + noun (clichés, technical language: e.g. reckless ways, gay abandon, injured party);
   (b) Noun as adjective (attributive in LGP/ontological in LSP) + noun (e.g. jet engine).

3. Noun as subject + verb of action (e.g. snow falls/melts, accidents happen/occur).

4. (a) Group noun (e.g. a flock of sheep); (b) specific/generic (e.g. a bale of hay, a drop of rain).

5. Adverb + adjective (e.g. hardly awake, barely audible).

Chapter 3: 125
In LGP Benson et al. state that the copula is needed, e.g. *the boy is hardly awake*, not *the hardly awake boy*. This observation does not however seem to apply to all adverbs, e.g. *the musically-gifted child*. Possibly the difference is in the role of the adverb; where it is quantitative, e.g. *hardly, barely*, the copula is certainly required, but it is not necessary where the adverb is qualitative in nature.

6. 

| Verb + adverb | (e.g. laugh gaily, thank warmly) |

3.2.3 Conclusions on grammatical and lexical collocations

In comparing grammatical collocations between LGP and LSP I have observed that there would be no difference in the word used as a 'base' for an entry (the notion of 'base' is discussed in sections 3.4 and 3.5). Because of this observation, I have decided that it would not be worthwhile in the light of the thesis to pursue grammatical collocations in LSP as far as English is concerned, except for those which show a variation in the valency of verbs. Valency is discussed in detail in Chapter 4.

However, in section 3.5, I posit the hypothesis that it is mainly in lexical collocations that some differences can be noted between LSP and LGP. Conclusions are drawn and the types of collocation which are not deemed relevant to LSP will not be considered further in this thesis. Before proceeding solely to LSP collocations, however, other factors which affect collocations in general are assessed.

3.3 Substitution, 'free' collocations and disjuncture

Collocations, unlike idioms, can vary in one of their components

Chapter 3: 126
by the substitution of near synonyms or by the addition or insertion of other lexical units. These factors need to be taken into consideration because they may render the task of identifying the boundaries of a collocation more difficult. There does not appear to be an accepted theory that produces a clear-cut definition of the type and number of words constituting internal disjuncture. Cowie (1981: 224) states that:

'a collocation is by definition a composite unit which permits the substitutability of items for at least one of its constituent elements (the sense of the other element, or elements, remaining constant)'.

Thus, *run a business* and *wages freeze* are collocations because substitutes are possible in both cases. These are what Hausmann (1985) and Benson *et al.* (1986) call 'free collocations'; there is, however, a limited number of such collocations. Indeed, Hausmann (1979) assigns collocations to 'langue', i.e. the very system of an individual language. J. Sinclair (1994: to be published) expanded this point very succinctly when referring to recurring phrases and collocations in corpora; the 'vertical' patterns which emerge when viewing target words from corpora which are in alphabetical order represent 'langue', while the horizontal representation - the context - is the discourse or 'parole'.

Collocations are lexical and often intuitive. The arbitrary (as opposed to free) recurrent nature of collocations is clearly illustrated when comparisons are made with corresponding collocations in other languages; the following examples from Benson (1989: 3) are collocations in each of the languages given but whereas the nouns in each case exactly represent the concepts 'traffic' and 'oath' in English, the verbs exhibit semantic differences, e.g.
Although the resultant phrases have the same semantic representation, the choice of verbs is arbitrary even where languages are etymologically and culturally related. It therefore appears that one part of a collocation may show greater variation than the other; in these examples it is the verb. This phenomenon occurs particularly in the case of verbs which are semantically 'empty' such as cause, do, draw, make, set, take.

The dividing line between collocations and free combinations is not always clear, however; for example, blow-by-blow account but *blow-by-blow description? It is evident that statistical analyses can play a decisive role in clarifying the distinction between collocations and free combinations, particularly as collocations have a strong tendency to be native-speaker oriented and are often very difficult for learners of a language to grasp.

3.4 Base (node) and collocate

In addition to the elements of grammar and lexis which may form collocations, a study of the distance between the words in collocatory patterns is necessary, particularly as a prerequisite to running concordances which will provide statistical support for such patterns.
3.4.1 Collocatory 'span'

Martin, Al and van Sterkenburg (1983: 84) define collocation as follows:

'A "node" is the lexical item whose collocational pattern we are looking at. A "collocate" is any lexical item which co-occurs with the node within the specified co-text. A "span" is the co-text within which the collocates are said to occur. Span positions of collocates are numbered according to their distance from the node. In the idiomatic expression kick the bucket, e.g. the collocate bucket appears at span position +2 of the node kick. Defining the optimal span for a collocational study is always a matter of dispute. Theoretically, a node has an infinite range of influence which decreases with distance from the node. Statistical tests lead us to the conclusion that more than 95% of all relevant information can be obtained by examining collocates within a span of -5 and +5 (disregarding punctuation). A "significant" collocation is one in which the two items co-occur more often than could be predicted on the basis of their respective frequencies and the length of the text under consideration.'

The +2 and -5 spans referred to by these authors relate to the number of words between a node and its collocate, regardless of whether one appears after or before the other. The disregarding of punctuation is controversial and seems contrary to linguistic analysis; it seems more logical to make assessments within a clause or sentence. A problem arises with the examples from the LSP corpora that I have studied, because some terms may comprise five or more lexical units, for example, virology provides *lympho-proliferative disease of turkeys virus, multiple nucleocapsid*
nuclear polyhedrosis virus, and many others; it is therefore important to ascertain the optimum number of words either side of the term in its entirety which would provide information on collocation in special subject fields; the span therefore needs to be greater than five. It is clearly important to formulate criteria for establishing a boundary for excerpts from concordance studies, maybe by discarding certain constructions, so that the same information is gleaned from short concordanced excerpts as that found in flowing text, to ensure there is no distortion of 'evidence'. In addition, since valency is considered to be an integral part of this study, the span must ensure encapsulating pre- and post-positional actants, while at the same time encompassing the lengthy terms found in some LSPs.

In LGP on the other hand, Sinclair, Jones and Daley strongly suggested in 1970 that beyond four words from the node there were no statistical indications of the attractive power of the node. Recently, Sinclair and his group have experimented with environments of between one and five words on either side of the node to see whether there is an optimum setting (Sinclair, 1991: 106; Renouf and Sinclair, 1991: 128-143); Sinclair appears to corroborate his earlier findings of five words before the node and four after (J. Sinclair, 1994, personal communication). These analyses take the node as a single lexical unit. My own findings with regard to LSP corroborate the comment of Martin et al. mentioned earlier in this section; to work with such a small number of words either side of the node would certainly pose problems in a number of special languages.

3.4.2 Entry at base or collocate?

Martin et al.'s terms 'node' and 'collocate' appear in Hausmann (1985: 119-121) as 'base' (Basis) and 'collocator' (Kollokator) in lexical collocations. Hausmann gives the following criteria for the placing of
collocations; it is interesting to note that he does not, in company with Heid and Martin (*cf.* section 3.1), include the noun + noun 'collocation', but his categories are the same as theirs, although each lists the grammatical components in a different order (3.1):

- in verb + noun collocations, the noun is the base and the verb is the collocator;
- in adjective + noun collocations, the noun is the base and the adjective is the collocator;
- in adverb + verb collocations, the verb is the base and the adverb is the collocator;
- in adverb + adjective collocations, the adjective is the base and the adverb is the collocator.

Hausmann is explicit about the placing of collocations in dictionaries. Ideally there would be dictionaries which 'decode', that is, help users to comprehend texts, in which he advocates that collocations should be placed at the entries for collocators. However, in learners' dictionaries used for generating or 'encoding' texts, collocations should be placed at the entries for bases. Benson too (1989: 6) has assessed this approach as it applies to LGP (*cf.* 3.3) and quotes from Hausmann (1985: 119-123), stating that an advanced learner of a language such as a translator consulting a dictionary or term bank will normally know the base; for example, German speakers who want to express *Widerstand leisten* (i.e. base + collocator) in English will probably know the base, *resistance*, but under the entry should find the collocation *offer/put up resistance; leisten* is a multifunctional verb in German (e.g. *Bezahlung leisten, Unterschrift leisten*) which results in a number of possible English translations. Benson (1989: 7) criticises the
LGP dictionaries LDOCE and OALDOCE for their lack of lexical collocations; for example, LDOCE at the verb entry *draw* gives *draw blood, draw a crowd, draw a gun* but surprisingly does not give them at the appropriate noun entries, in other words, at the base. Kozlowska (1991) gives a hierarchy of parts of speech with nouns ranking first, verbs second, adjectives third and adverbs fourth as far as collocational methods are concerned; it is only useful for learners to know which modifiers qualify which modificand and not vice versa (reviewed by Knowles, 1993: 301).

LSP appears to differ insofar as the adjective + noun "collocation" forms part of an integral terminological unit, as has been shown; so also in many instances does the adverb + adjective and the attributive noun + noun constructions (cf. 3.2.2). Where the 'ranking' approach could be of use is in the verb + noun construction where the verb is from LGP and may be semantically 'empty'; this is less likely to be so where the verb is from LSP or has a special LSP meaning because in such instances the phrase is learnt as a quasi-idiom (e.g. *to debug a program*).

It can be seen from the verb examples given earlier in this section that the verbs are to a large extent semantically empty. While this also occurs in LSP, the number of true LSP verbs collocating with a term is more restricted, usually to one and occasionally to two; examples of this phenomenon are to be found in Chapter 4.

So far, the pragmatic basis for placing collocations has been discussed. However, although the methods for differentiation between base and collocator may be founded on grammatical criteria, semantic and intuitive aspects are also necessary. In the next section some theoretical hypotheses

Chapter 3: 132
are proposed which it is hoped will serve to reinforce the pragmatic basis of orientation in collocation. Furthermore, the variations in collocations support such lexical functions as 'synonymy', including abbreviations and acronyms, in addition to antonymy, ellipsis and paraphrasing.

3.5 Comparisons of lexical collocations in LGP with corresponding LSP constructions

It is clearly important to analyse which of the forms of collocation are found in special language and which are found in general language, to help experts in various fields (indexers, information scientists, journal editors, translators, terminologists, etc.) to decide on the orientation of a phrase and thus provide a lucid retrieval mechanism. In the following table, a comparison is made between the base words in so-called 'lexical collocations' in both LSP and LGP, and in the ensuing sub-sections the theoretical motivation for the decisions is discussed. The collocations in the table are based primarily on the categories of Benson et al. (1986) listed in section 3.2.2, in addition to those of Hausmann in 3.4.2.
Lexical collocations

1. (a) v.(usu. tr.) + n./pron./prep. phrase (3.6) n.(usu.) n.
   (denoting 'creation' or 'action')
   (b) v. + n. (3.6) n.(usu.) n.(usu.)
   (denoting 'eradication' or 'nullification')
2. (a) adj. + n. (3.5.1) adj.(as p.p.) n.(usu.)
   (b) n. as adj. (?attributive) + n. (3.5.2) 1st. n. (usu.) either
3. n. + v. of 'action' or 'process' (intr.) (3.5.3) v. n.
4. group n. (3.5.4) generic specific
5. adv. + adj. (3.5.6) adj. adj.
6. v. + adv. (3.5.7) v. v.
   (probably compound)

Table III.1 Comparison of the orientation of entry terms in LSP with lexical collocations in LGP

Legend: v. = verb; n. = noun; adj. = adjective; adv. = adverb;
p.p. = past participle; tr. = transitive; intr. = intransitive; usu. = usually.

Each of the four instances where differences appear between LGP and LSP, i.e. 2.(a), 2.(b), 3 and 4 in sections 3.5.1 - 4 respectively, will be considered more fully. What immediately becomes evident is a corroboration of the importance of the role played in LSP by the terminological entity formed by adjective(s) + noun(s) or noun(s) + noun(s). These are treated in depth in 3.5.1 and 3.5.2. Because the verb + noun construction (1.(a) and 1.(b) above) is a key feature of this thesis and forms the focus of the next chapter, the possible differences in placing this construction as an entry are
considered separately and only briefly in this chapter, in section 3.6. In subsequent chapters, a classification system for verb + noun phrases will be devised based on collocation and valency patterns, showing how these can be corroborated by statistical analyses.

3.5.1 Adjective + noun

In 2. (a) adj. + n., usually the noun is predominant in LGP, whereas in LSP, it is frequently the past participle used as an adjective which predominates because it is usually an integral part of the whole term and is not 'free'. In the LSP of virology, examples such as permissive cell, occluded virus, attenuated strain, inactivated vaccine, non-glycosylated protein, virulent strain, are each semantically fused, single concepts, and are therefore terms, even though they are not single words. These examples, taken from a corpus of virology texts comprising 40,000 words and from 'Virology: Directory and Dictionary of Animal, Bacterial and Plant Viruses' (Hull, Brown and Payne, 1989), show that the role of the adjective is to render the term more specific; it would therefore be considered the 'base' in LGP parlance and the noun its 'collocate' (cf. the introductory paragraphs to this chapter for the use of the appellation 'collocation' in this section and 3.5.2). However, in LGP, examples such as strong/weak tea, best regards, reckless abandon, formidable challenge, sweeping generalization etc., would all have the noun as the 'base' and the adjective as the 'collocate' (Benson et al., 1986: xxvi).

These clear differences between LSP and LGP are of interest to linguists, lexicographers and indexers because they give rise to considerations of the practical problems involved in entering or indexing such LSP terms. The resolution of these problems requires an investigation
into the theoretical foundations of the differences between collocatory patterns in LSP and LGP, and a further discussion of these is given in 3.7. Workers in the various fields listed at the beginning of 3.5 need to be able to identify terms by verifying their intension (the 's' in 'intension' is deliberate), thus enabling them to recognise a qualified term as a base unit and to look beyond a single word as the base form. The grammatical mechanism of the adj. + n. construction is the same in LGP and LSP, but the contrasts seen above give rise to the question of what it is that attracts words to one another to form a collocation; why, in fact, do collocations occur? Collocations appear to be synergistic patterns of words that have become established through frequency of use to produce metalexicographical mechanisms, the forms of which are highly variable. It is true that 'proof' of a number of collocations will respond to the question 'what are they?' by statistical analyses of their component parts. Moreover, the question 'how are they formed?' may be answered by certain grammatical patterns and, to a lesser extent, by the semantic attraction of their components. However, the question 'why do they form?' is more obscure. A simple answer to a number of them is doubtless that they are euphonious; they may show alliteration or be onomatopaeic, for example. But this does not explain the attraction in special languages where such considerations are rare. Clearly, in these instances, the juxtaposition of words in LSP occurs for specific intellectual reasons within a domain.

The needs of subject specialists differ from those of translators, while terminologists play an intermediate role. In the examples from virology at the beginning of this section, in which the adjective is used as the entry point, entry would be based on the specialist's encyclopaedic knowledge of the subject and its terminology; for example, the terms restriction
enzyme and non-permissive cell do not have corresponding entries at enzyme and cell, since to a subject specialist an explanation of these one-word terms would be unnecessary. It is likely therefore that few single word terms will be listed. Hence the approach is encyclopaedic rather than ontological; in the above examples, and with techniques such as Southern blot, the compound terms themselves are at the highest level of abstraction.

In practice, the entry of the term under the adjective in English helps to curtail the search process for the subject specialist but may prove problematical to a non-specialist such as a translator who is possibly not familiar with the compound term (cf. the problems of the student terminologists in Chapter 2, section 2.1.1). This confirms my observation that there is scope for more than one approach to the entering of special language terminology in English, depending on the user. This gives rise to an interesting operational process in English which would not be encountered in languages such as French where the noun precedes the adjective. From the lexicographical point of view it might be argued that the motivation is different, since French does not need the same operational process because the noun is the entry point in both LSP and LGP, and therefore the comparisons are not truly viable.

3.5.2 'Fused' compounds

In this section the linguists' approach is contrasted with that of terminographers. Example 2. (b) in Table III.1 comprises two nouns with the first being used as an adjective. Benson et al. (1986: xxvii) recommend entry at the second noun in instances such as jet engine, land reform, but propose that in cases of a 'fused' compound where the second noun does not have the same basic meaning as when it stands alone, e.g. sitting duck, stuffed shirt, double take, i.e. in the case of idioms, then the entry should be at the first noun. However, in virology, a term comprising two nouns,
although it cannot be said to form a 'fused' compound in the sense understood by Benson et al., nevertheless represents a single concept and would be entered at the first noun. In addition, it has already been shown that the noun chosen, as in measles vaccine (cf. 3.7.1), will depend on the context in which the term appears, i.e. its 'orientation', which in turn depends on the user. It seems that there is a far greater incidence of this phenomenon in English LSP than in LGP. The transition from LGP to LSP however follows a cline and it can be difficult to ascertain at which point the transition occurs. In addition, metaphorical usage must be considered, e.g. the dangerous virus of corruption. The following examples in Table III.2 help to illustrate this hypothesis, although it must be stressed that this is a perceived tendency, rather than a 'hard-and-fast rule'. The proposed entry point is underlined:

<table>
<thead>
<tr>
<th>LSP</th>
<th>LGP</th>
</tr>
</thead>
<tbody>
<tr>
<td>software program</td>
<td>concert programme</td>
</tr>
<tr>
<td>nonsense codon/mutant</td>
<td>nonsense verses</td>
</tr>
<tr>
<td>goose parvovirus</td>
<td>goose down</td>
</tr>
<tr>
<td>organ culture</td>
<td>?organ grinder</td>
</tr>
<tr>
<td>signal peptide/sequence</td>
<td>?distress signal</td>
</tr>
</tbody>
</table>

Table III.2 Comparison of proposed entry points (underlined) for noun + noun compounds in LSP and LGP

An interesting linguistic contrast has already been noted in French, where the first noun is the entry point in both LGP and LSP. Hence
in LGP we have *duvet d'oie* (goose down), *joueur d'orgue de Barbarie* (organ grinder), *signal de détresse* (distress signal). Although *le logiciel* would cover *software program*, in LSP the most common construction is to link two nouns with *de*, as in the above examples from LGP. This is particularly prevalent where the second noun is a proper noun or refers to a particular object, e.g. from virology, *maladie de Marbourg, rage des chauves-souris*, whereas more general terms, such as those relating to species, have an adjectival form, e.g. *variole porcine, rage selvatique, rage citadine*. (*Rage des rues* does exist, conjuring up a bizarre picture and presumably being a rather inapt borrowing from the English *street rabies.)*

3.5.3 Noun + intransitive verb

In Table III.1 (3.) (noun + intransitive verb), Benson and his colleagues assume that the noun will be the entry, e.g. *bees buzz, blood clots*, but this is surely a case where cross-referencing of the collocate is invaluable for translators (*cf.* Hausmann's advocacy of decoding dictionaries), as with *horses neigh, volcanoes erupt*; for example, horses do other things than neigh, yet this is specific to animals of this species and is important information for learners of the language. I feel, therefore, that the problem is a semantic one, the 'base' word of the collocation being the one I have underlined in the two examples given immediately above and which in my opinion should be the entry terms. It is likely in LSP to have the verb as the entry term when this is from LSP, for example, *viruses replicate, virus particles crystallise, viruses mutate, viruses transduce*. It is interesting that a verb borrowed from a different LSP may still be the entry term in the first LSP, e.g. *to fingerprint a virus*. LGP verbs may also have an LSP meaning in a special subject field, e.g. *vectors [...] should insert; to boot a computer; to disinfect software against viruses*; in these instances I would
propose entry under both noun and verb.

3.5.4 Group nouns

In Table III.1 (4.) (group nouns) the 'base' word would appear under the specific term, e.g. herd of buffalo, whereas in LSP the base would more likely be the generic term, e.g. family Retroviridae, sub-family Spumavirinae, phage group, polypeptide chains, antigenic properties of serotypes.

3.6 Verb + noun colligation

In scientific and technical language (i.e. LSP) there are verb + noun colligations which are on the dividing line between collocations and idioms, e.g. in computing, create a file, debug a program and in virology, confer immunity, replicate a virus, mount an immune response. It will be noted that the verbs are often from LGP, whereas the noun collocates are from LSP. Benson (1989: 5) advocates that these transitional and technical examples should be entered as collocations, indeed calling them 'transitional collocations' and giving as his example from LGP carry weight (be convincing). The verb + noun combination in LSP will be treated in greater detail, particularly in Chapter 4.

Some verbs are followed by a semantically limited number of direct objects, e.g. run a business, where the verb has the special meaning of 'direct' or 'manage', while others have only one or at most two collocates, e.g. debug a program, to debug a system. In the last example, the verb itself is from LSP, a much rarer occurrence than with nouns; virology provides for example to toxoid diphtheria and tetanus toxins.
Verbs that are followed by unspecified direct objects, such as buy, carry, hit, like, make, should not be entered as collocations except where they are part of an idiom, e.g. from LGP, buy a pig in a poke, hit the hay. These verbs were not found to form significant collocations in the LSPs studied.

3.7 Theoretical recommendations for placing 'collocations' in LSP

Following the comparisons between LSP and LGP, and the conclusions on different types of compounds discussed in sections 3.5 and 3.6, the theoretical aspects leading to recommendations for placing so-called 'collocations' in LSP are considered in greater depth in this section. It is anticipated that these recommendations will help terminologists draw conclusions for achieving consistency when inserting multiple lexical terms into, for example, a term bank record, in addition to providing the optimum data appropriate to each term according to its headword. This in turn will benefit users. As mentioned in the previous chapter, the guidelines form Annex 1.

In the light of the assessment of collocations expounded so far in this chapter, and incorporating the LSP categories identified in Table II.5, it can be seen that the collocations which appear in LSP (cf. the use of the term 'collocation' in LSP in the introduction to this chapter) are those found in:

- compound nouns (3.7.1)
- compound verbs (3.7.2)
- compound adjectives (3.7.3)
Examples will be given primarily from the domains of virology, automotive engineering, computing, and pharmaceutical/legal texts in French and English, to ascertain whether the guidelines of base and collocate adopted for LGP are applicable to LSP and to assess the impact of the theoretical considerations made.

Furthermore, an example of a thesaurus structure is provided in 3.8.2 to explain how collocatory patterns can help with retrieval by identifying orientation.

3.7.1 Compound nouns

It has already been noted that compound nouns are problematical for terminologists because they can be difficult to assess for their entry point. It is therefore vital that an intellectual appraisal be made of examples from LSP before there is any attempt to reach a decision. In the examples of English LSP terms in subsequent paragraphs of this section, two nouns are present, that is, two concepts which together form a third. Here the rationale is encyclopaedic rather than linguistic and is akin to the differences between terminological and lexicographical considerations. The relationships of terminology are evident, such as logical, for example generic, and
ontological, e.g. cause/effect, and these provide an initial guide to the creation and placing of compound terms. To these considerations must be added their orientation, depending on the context in which the term is to appear, and this should be the deciding factor (see Fig. 3.v). Hence the surrounding text which produces an encyclopaedic, thesaurus-like framework surrounding the term is of paramount importance. There is however a possible pitfall in this approach, because to provide an up-to-date thesaurus, the most recently published texts must be used; in a rapidly developing subject field, however, such publications may contain contentious material; only those appearing later will reflect a consensus of opinion.

In the first example studied, measles vaccine, measles may be considered as a noun used as a 'semi-free' adjective, as there are only a few virus diseases for which a vaccine is available. Moreover, if a vaccine were developed for other diseases, the same 'rule' would apply, as is already the case for AIDS for which no vaccine as yet exists but in anticipation of which the media have coined the term AIDS vaccine. This is corroborated in the recommendation given in ISO R 704, Principle 23 (1984): '.... in a complex term the determined constituent should be the genus of the concept.' In other words, determined constituent refers to the constituent being determined, i.e. the generic term, in this case vaccine.

Felber describes 'determination' (or 'specificity') as a phenomenon where 'a second concept is integrated as a characteristic into the intension of the first; thus the intension of the first concept is enlarged by at least one additional characteristic. The resulting concept is a species of the first concept.' (Felber, 1984: 124-125). The physical order of the concepts is immaterial. Clearly, vaccine has its own definition to which the addition of the disease which it counteracts has the effect of reducing its
sphere of application and of focussing on the disease, which is the specific
determining member. The 'imposition' of a disease on a vaccine can be viewed
as an enlargement of, and progression in, its terminological development,
giving an extensional, ontological relationship, characterised by
juxtaposition, where the noun used as an adjective is attributive. The
following clarifies the characteristics, both in content and the process, of a
vaccine:

Properties common to all vaccines:
- ability to protect against disease
- inclusion of the agent causing the disease

Different types of vaccines may:
- be made from inactivated, or "killed" virus
- be weakened or "attenuated" by passing in an unnatural
  host, e.g. yellow fever in mice
- contain sub-units such as toxins, e.g. diphtheria
- be engineered, e.g. hepatitis B made in yeast
- occur naturally, e.g. Marek's disease of turkeys which
  also immunises chickens

Thus it can be seen that the enlargement of the concept vaccine
also contributes to its specificity or 'narrowing-down' - what Felber calls
'determination'. These inherent, intensional characteristics of a vaccine
denote logical rather than ontological relationships.

A second example, brake booster, also comprises two nouns and in
this example the second word is used as an adjective of function, answering

Chapter 3: 144
the question 'what does it do?', rather than 'how does it restrict?'; in other words, it forms an ontological relationship. If the hypothesis is that booster should be the headword because it is the second noun, this would ignore its functional description and adjectival property. As seen above, measles vaccine also displays a functional description. These considerations enhance the scope of orientation and need to be studied before any attempt at ordering is undertaken; Felber analyses relationships but perhaps wisely does not seek to impose any conclusions; similarly, ISO R 704 expounds the relationships but does not explore the underlying ones in sufficient depth to solve questions that arise particularly between different languages, as the examples in the following paragraph show.

Translators face an additional problem because of the tendency in English towards ellipsis; multiple lexical units in French which form compounds usually include the use of a preposition indicating function or purpose, e.g. vaccin contre la rougeole (measles vaccine). There is clearly a difference in approach between French and English in the example containing contre, the French approach being clearer and more logical. Other instances where English tends to be elliptical, almost subliminal, are seen in The Government is mounting a terrorist campaign, not an anti-terrorist one, and in crème anti-rides which has been translated into English as wrinkle cream. While it is clear that the English examples should not be taken literally, it is conceivable that they might give rise to confusion, particularly to a non-native speaker of English.

In conclusion, computational analysis of the text surrounding a compound noun, and the frequency of terms and phrases, can help to give an indication of a term's orientation, but the final placing of the term may need
to depend on an expert's specialised knowledge of a subject.

3.7.2 Compound verbs (denominalised verbs/adverbs)

Technical experts frequently resort to abbreviation in speech. In automotive engineering, *to cat-con a car* is a concise but very colloquial way of saying *to fit a catalytic converter to a car* and it would not be possible to split the verb. Similarly, compound verbs formed from a verb + adverb combination cannot be split; they also appear as jargon in technical domains. Again, in automotive engineering, the examples of *to cadence-brake, to double-park* are found, where the first word has an adverbial function answering the question 'how?*, e.g. *to brake in cadence, *to park doubly.*

In virology too there is colloquial use, e.g. *to mouth-pipette, (to pipette by mouth)* which would be too informal to appear in written text. (The use of some nouns adjectivally in conjunction with a past participle, e.g. *vaccine-associated, vaccine-derived,* is frequent but the verbs *to vaccine-associate, to vaccine-derive* are not possible.) However, in computing, LSP verbs are formed in this way, e.g. *to screen-dump, to hand-shake, to remote-print,* the last three examples splitting the infinitive adverbially. Again, their use would be considered colloquial. The function of the grammatical components of compound verbs needs to be identified for the entry word to be deduced.

3.7.3 Compound adjectives

Compound adjectives such as *single-stranded, double-stranded* (DNA/RNA); *hydrogen-bonded* where the noun *hydrogen* is used adjectivally, mirror the noun + noun constructions in terms of their logical/ontological relationships. As in 3.7.1, the first example shows
adjectives denoting 'type of', i.e. the logical, generic/specific relationship, whereas the second shows an adjective of function, an ontological relation. It is interesting to note that no verbs can be derived from the first two examples but the verb to hydrogen-bond is possible. These observations open the door for continued research.

3.7.4 Compound adverbs

There appear to be very few of these in LSP where they might be considered colloquial, e.g. .... should be treated quantum-mechanically. The two components represent one concept.

3.7.5 Adjective(s) + noun(s)

Since an adjective serves to increase the specificity of the noun it is describing and precedes it in English, it is interesting to speculate whether the recognition process of the term by a subject specialist is expedited by the positioning of the adjective before the noun or after it, as in French, which seems to be the logical, 'top-down' approach, from the generic to the specific. It is interesting to note that Russian follows the English form, with the adjective preceding the noun, but in technical language, Russian is more stylised and resembles French word order, e.g. theory of information rather than information theory. Clearly, a multiple term in virology such as 

[[[African] green monkey] [kidney cell]] would not be recognised by African, which is semantically too broad, and even the completion of the first term, African green monkey, would be misleading. The reader or listener has to await the completion of the whole term with kidney cell to understand the intention of the writer or speaker, unless the subject of the discourse focusses narrowly on African green monkey kidney cells, whereupon it is feasible to suppose that ellipsis might take place. The
process of placing adjectives and adjectival phrases before the object of discourse is not conducive to comprehension, unless the term as a whole can be recognised early by the specificity of the adjective. A similar problem would occur with this example in French where, although the reverse word order occurs, the term *cellule* and even *cellule rénale* are too broad in scope for immediate comprehension of the whole term. However, in *[[apple stem][grooving][virus]]*, the use of the gerundive *grooving* draws the process of comprehension on to the causative agent, e.g. *virus*. In these examples, the adjectives and adjectival phrases are descriptive and ontological. Similar examples can be found in everyday use, e.g. *washing machine*.

As far as entering these terms is concerned, because they are complex multiple lexical units representing a single concept, it is probably the first word in each case that would be the entry point. It is in examples such as *[[dual-bed][catalytic converter]]*, where a specific, logical relationship occurs, that an entry might also be found under *catalytic converter*. There are, I think, three reasons for the conclusion relating to the second example: (i) we are looking at a logical, generic/specific relationship; (ii) the comparative 'newness' of catalytic converters and (iii) the low number of different types. This is in contrast to the first two examples from virology, where an ontological relationship is apparent and there are in addition a large number of viruses and kidney cells.

3.7.6 Adverb + adjective

In this construction the adjective is invariably a past participle used adjectivally, e.g. *hermetically sealed*; *orally administered*, and the adverb plays the role of making the adjective more specific. The construction
3.7.7 **Verb + adverb particle**

This category and the one following are likely to be of more interest to linguists than to terminographers. No examples of that peculiar phenomenon of English, the phrasal verb, have been found to occur in the written LSPs studied. The prevalent, frequent use of *up* may be heard in discourse in all walks of life (e.g. *to pot up a plant/to pot a plant up; to grow up a virus/to grow a virus up*) where the preposition is redundant and not considered good literary style.

3.7.8 **Verb + preposition**

There can be little doubt that the search word will be the verb. Examples are *ease up* (on the accelerator), *replicate in, bud through, cloned into, code for, extruded through, actuated by, transmitted to*.

3.7.9 **Idioms**

Terminologists need to know how to recognise when a verb + noun construction in special language forms a new concept; because the two comprise a whole and change their meaning if split, the idiom should be entered under both noun and verb, and cross-referenced. Examples: *to debug a program, to boot a computer*.

3.7.10 **Verb + noun**

As very few verbs *per se* exist in LSP, it is most probably the noun which will decide the entry word, e.g. from virology we have *to confer immunity, to replicate a virus, to mount an immune response* and from computing science, *to create a file*. Where there are examples of true LSP
verbs, the collocation should be entered under the verb also. However, if the verb is highly restricted to its subject field, e.g. *to debug a program*, *to initialize a disc*, it is probable that entry will be under the verb. This construction is studied in detail in Chapter 5.

3.8 Orientation and context

From examples of collocation elicited from the corpora and other sources, it is hypothesised that a study of the examples will provide a contributory factor in formulating guidelines for recognising a compound term or LSP phrase, and for assessing the headword in these. Furthermore, it is suggested that the orientation of a term or phrase may be derived from collocatory patterns in the text in which they appear. It will be recalled that I have deliberately chosen the word 'orientation' to help identify the content, determined semantically, of a context in which a term appears (cf. *nucleus* in Chapter 1). Although a headword is clearly needed for a printed work such as a dictionary or an index, and will alter according to the intended readership, however, in term banks and other computer systems, where a basic search for a word is an easy operation, it is on the background information provided by a terminographer that the orientation of a term needs to be focussed and this depends on identifying various groups of user profiles. I am claiming that orientation will in fact take precedence over the more traditional requirements of lexicology and terminology and believe statistical studies of collocatory patterns will serve to substantiate this claim. This aspect does not appear to have been addressed previously, and the following reasons are postulated:

- It would not appear to apply to multiple lexical units in general
language because of the infinite variety of semantic content of the contexts in which compounds occur;

- it has not, as far as I have been able to ascertain, been a prerequisite in terminology work where the concept alone has been the prime consideration, even though it is part of a subject field;

- the concept has been regarded in the light of the particular (con)text in which it is situated without consideration being given to its occurrence in a different context (cf. the example of nucleus in Chapter 1).

To corroborate the notion of orientation, and again using the example of foot-and-mouth disease of cattle, if the term is divided in two places, foot-and-mouth disease and disease of cattle, two scenarios can be assumed. In the first, i.e. foot-and-mouth disease of cattle, the generic term foot-and-mouth disease indicates the semantic content of the whole text; in other words it is assumed that the text is about foot-and-mouth disease which can attack any cloven-hoofed animal. The specific term foot-and-mouth disease of cattle denotes the host species which the disease attacks. Scenario 2, foot-and-mouth disease of cattle, gives the generic term disease of cattle, so we can assume that the text concerns bovine diseases in general. The specificity of foot-and-mouth denotes a particular disease. It is interesting to note that where generic and specific terms are lexically related, the specific is invariably a longer version of the generic form, as in the example above. In addition, it must be borne in mind that terms evolve and may not assume their final accepted lexical form for many years.
These observations however beg the question of how the orientation of terms can be discerned from their context and it is hoped that the thesaurus structure shown in 3.8.2 will demonstrate the feasibility of this approach. Initially it is true that a semantic assessment has to be made, but it is important to ascertain how such an assessment can be corroborated statistically. For instance, in the first example given in the previous paragraph, other relationships such as (quasi)-synonymous terms need to be sought in the text, e.g. from scenario 1, the names of other cloven-hooved animals, and from scenario 2, the use of the terms sickness, symptoms of disease and so on.

3.8.1 Collocation in corpora as an aid to orientation.

How therefore can a user, whether terminologist or information retrieval specialist, be guided firstly in the search for the whole term and secondly for the orientation of the context in which the term occurs? It is undeniable that corpora are a useful aid in terminology. A study of corpora shows that the juxtaposition of lexical units and their higher than average occurrence in a given text provide a strong guideline for the formation of multiple lexical terms (see Doyle, 1963 for an early exposition and also 'association ratio' in 3.9); it has already been stated that one-word terms occur less frequently. Furthermore, the subject field of the context can be recognised by the frequency of the terms identified. More specifically, the orientation of terms is irrefutably conditioned by the context of the document containing the terms and this becomes apparent when studying the collocational patterns which appear. The following figure shows in a simplified manner how the orientation of a multiple lexical term, in this case measles vaccine, can be deduced by the collocatory patterns found in two extracts from texts. It is important to note that the collocations are not necessarily terms, e.g. prevent.
Having recognised the possibility of identifying the orientation of a term through collocation, I consider further elucidation to be worthwhile, by providing a hypothetical entry in a thesaurus which, if the thesaurus were given in toto, could provide a browsing facility for both terminologist and retrieval expert. However, because of the constraints of the thesis, a single entry is given as an example (Fig. 3.v).

3.8.2 The thesaurus as an aid to orientation.

I have constructed the thesaurus entry by taking advantage of the terminological relationships revealed by the collocatory patterns produced from the corpora. As examples with measles vaccine and measles virus

<table>
<thead>
<tr>
<th>Term</th>
<th>Context</th>
<th>Possible collocations</th>
</tr>
</thead>
<tbody>
<tr>
<td>measles</td>
<td>diseases of children</td>
<td>febrile illness</td>
</tr>
<tr>
<td></td>
<td></td>
<td>rash</td>
</tr>
<tr>
<td></td>
<td></td>
<td>presents with (v.i.)</td>
</tr>
<tr>
<td>vaccine</td>
<td>paediatric medicine</td>
<td>prevent*</td>
</tr>
<tr>
<td></td>
<td></td>
<td>inoculat*</td>
</tr>
<tr>
<td></td>
<td></td>
<td>attenuated</td>
</tr>
</tbody>
</table>

Figure 3.1 Orientation of measles vaccine indicated by examples of collocation in a text. * = wild card
have been used previously in this thesis, the theme is continued in the
thesaurus entry, with *measles virus* and its genus, *Morbillivirus*,
providing an explanatory study. Firstly a brief explanation of the
classification of the virus helps to make the thesaurus entry more readily
comprehensible (Figure 3.ii):

---

**Family:** *Paramyxoviridae*

**Genus:**

- *Paramyxovirus*
- *Morbillivirus*
- *Pneumovirus*

**Type species**

- Newcastle disease virus
- Measles virus
- Respiratory syncytial virus

**Other Finch paramyxovirus**

- Canine distemper virus
- Bovine respiratory syncytial virus

**members**

- Mumps virus
- Rinderpest virus
- Pneumonia of mice virus

**Nariva virus**

- Peste-des-petits-ruminants (PPR) virus

---

**Figure 3.ii** Example of virus classification showing *measles virus*.

Although it may appear contrary to the approach of a general
language lexicographer, it should be recalled that the term *virus* is unlikely
to appear on its own in a special language dictionary or thesaurus, since
experts in the field do not need to have such a fundamental term explained;
hence the starting term is often a multiple lexical unit. To facilitate
structural 'maps' and the 'look-up' strategy, the first step is to provide
terminological relationships for the thesaurus; logical and ontological
properties have been assessed and those considered relevant for *measles virus* are marked with a tick in the figure below; examples of these
relationships pertaining to *measles virus* are given in Fig. 3.v:
Logical      Ontological

✓ generic/specific
✓ similar
✓ overlapping
x exclusive on same level of abstraction
✓ negative
✓ spatial or temporal (consecutive process) contiguity
✓ cause/effect
x whole/part

Figure 3.iii  Logical and ontological relationships in measles virus

The next step is to assess the intensional and extensional properties of viruses, to ascertain which of these could be incorporated into the logical/ontological relationships.

Intensional properties of viruses                                 Extensional properties of viruses
Virions (size, shape etc.)                                       Physical stability
Genome                                                          Biological (host, transmission, vector, geographic distribution)
Proteins
Replication

Figure 3.iv  Intensional and extensional properties of viruses

It is worthy of note, although perhaps hardly surprising, that the logical and ontological relationships are found in the extensional properties of the virus. For the purposes of the thesaurus entry which follows, the
intensional properties will be disregarded. The thesaurus structure in Figure 3.5 is therefore a synthesis of the logical and ontological (extensional) properties of *measles virus*. The words and phrases in bold denote collocations that can be used to provide relationships from texts stored in corpora; these relationships can in turn be used to build further thesaurus entries. Although the sample entry is for a hypothetical thesaurus, it is interesting to note that the difference from a traditional thesaurus entry is in the important role that collocation can play; when linked with terms, both logical and ontological relationships are denoted. An additional enhancement to a traditional thesaurus which derives from assessing collocation in corpora has already been mentioned in the introduction to this chapter and in 3.4.1: a 'significant', i.e. >2, collocatory pattern might give an intimation of a 'term in the making', i.e. the pre-terminological status of a term, as has been noted in the example of *AIDS vaccine*.

The spatial and temporal ontological relationships combine to become the *epidemiology* of the virus which could form a separate entry in a thesaurus. As viruses can be assessed from the viewpoint of their epidemiology, a different approach is provided from that of the taxonomy of virus families; in other words, they can be classed as enteric, respiratory, arthropod-borne, oncogenic (close contact) or congenital. Some respiratory viruses, of which *measles* is one, are allocated to the virus family *Paramyxoviridae*, among others; hence there is a risk of circularity which, however, if exhaustive, should provide a 'fail-safe' mechanism during the construction of the thesaurus for assuring the inclusion of all possibilities.
logicaL

generic/specific (two levels)

member of
(virus family) BT1 Paramyxoviridae
(virus genus) BT2 Morbillivirus

similar and overlapping RT1 canine distemper
(same level of abstraction: RT2 bovine rinderpest
antigenically related RT3 ovine peste-des-petits ruminants
(e.g. measles vaccine can confer RT3 ovine peste-des-petits ruminants (PPR)
immunity to dogs against distemper)
susceptible RT4 monkeys

negative (lack of) stability:
sensitive to, killed by ANT1 formaldehyde
ANT2 lipid solvents
ANT3 oxidising agents
ANT4 non-ionic detergents
ANT5 heat
ANT6 vaccination, vaccine

ontological

epidemiology:
(i) spatial contiguity SPA1 airborne transmission
SPA 2 geographic distribution
(ii) temporal contiguity TEM1 incubation period

cause/effect CAU1 acute febrile illness in children
causes CAU2 cough
associated with CAU3 coryza
chronic CAU4 conjunctivitis
CAU5 spots on buccal mucosa
spreading to CAU6 rash on head and neck
occurs CAU7 sometimes encephalitis
progressive, degenerative CAU8 sometimes fatal, death

Figure 3.5 Hypothetical thesaurus structure for measles virus showing how collocation can provide logical and ontological relationships

Legend: BT = broader term; RT = related term; ANT = antonymous term; SPA = spatial; TEM = temporal; CAU = cause/effect.

Chapter 3: 157
3.9 Mutual information and association ratio

The above assessments are nevertheless still semantic, assuming a knowledge of the subject matter of the texts and hence of its terminology. However, when the texts are analysed for collocatory patterns, if the number of lexical items most frequently collocating with the generic term appears more frequently than is decreed by probability, pointers emerge from the context to give indications of the orientation of the term and its collocate. Here it is useful to invoke Fano's (1961: 28) theory of 'mutual information' (MI), which states that if two points or words, x and y, have probabilities \( P(x) \) and \( P(y) \), then their mutual information, \( I(x,y) \), can be defined as:

\[
I(x,y) = \log_2 \frac{P(x,y)}{P(x)P(y)}
\]

MI compares the probability of x and y occurring (joint probability) with the probabilities of x and y occurring separately (chance). If there is a true association between x and y, the joint probability will be greater than chance, \( P(x) \) and \( P(y) \) and, as a result, \( I(x,y) > 0 \). This is akin to the 'association ratio' of Church and Hanks (1989: 77) who differentiate their method from mutual information in that in association ratio, words can appear in either forward or backward order.

3.10 Conclusions

To summarise, the importance of collocation is already well known to translators and it is hoped that in this chapter I have established its importance for terminologists who are, for example, building terms into a term bank, by providing theoretical guidelines which lead to practical
solutions for the orientation of such terms (Annex 1). These collocation studies have led me to attempt to ascertain whether 'patterns' emerge in verb + noun collocations in special fields. In the next chapter I shall analyse the relationships between terminological and lexical collocations in LSP verb phrases with the aim of providing a framework for verbs in special languages. An interesting aspect which has come to light is that of grammatical collocations which deviate from the general language 'norm' in their valency when used in LSP, such as in the example of confer in 3.6 and 3.7.10, and this aspect will also be studied further in the next chapter.
Chapter 4

Influence of a subject field on its verbs and verb phrase relationships: synergy of collocation and valency

In the previous chapter it was shown how collocation from text can be incorporated into a thesaurus-like structure to provide logical and ontological relationships between terms, in addition to analysing how their orientation can be determined. The study of collocation is now taken a step further by applying it to special language verb phrases comprising the verb + noun construction, with or without disjuncture, for which I have coined the term 'combinant'. The reason for adopting a term which has not previously been used in Linguistics is that it appears to provide a more apt description of the attraction already in existence between a verb and a noun than the terms 'collocation' and 'colligation', which have an active, rather deliberate sense of words being 'placed with' or 'bound together'. The term 'combinant' is being used to cover a looser structure and permit greater freedom in accepting the discontinuity frequently found in LSP verb phrases, e.g. adjectives and adverbs may be 'free' or 'fixed', the latter forming part of a term.

This chapter has two main aims. Firstly, to ascertain to how much the subject field influences the 'LSP-ness' of a verb from LGP; this can to a great extent be corroborated by the higher frequency of verbs pertaining to a subject field when compared with an LGP corpus (in English, the COBUILD corpus and in French, a corpus from 'Le Monde'). To effect the analyses and for comparative purposes between English and French, two large LSP corpora containing texts on the biological sciences were compiled, mainly by the author, each containing half a million words of English and half a million words of
French. Smaller corpora on overlapping subject fields (law and pharmaceuticals) from the 1980s have also been compiled by the author. The second aim, and the most important facet of the current investigation, is to show that the orientation of combinant verb phrases can be ascertained after analyses of collocation, i.e. whether the verb or its following noun phrase takes terminological precedence, i.e. orientation. To this end, I have assessed other linguistic areas, and valency, with its emphasis on the verb, is clearly a valuable aid to the studies on collocation. The pivotal role of verbs in a sentence and the terminological relationships formed with their object/proposition have therefore been analysed.

Establishing the orientation of a verb phrase provides a pragmatic tool for solving the problem of which component of the phrase should be sought in a dictionary; guidelines would clearly be of invaluable help to translators in particular. The conclusions drawn from these studies have led me to construct frameworks for verbs in special languages which incorporate the relationships identified in LSP combinants, hence leading to the orientation, or identification of the entry word in an LSP verb phrase (Chapter 5). Orientation can thus affect the format of dictionaries, information thesauri and the terminological work of translators.

It has become evident during this research that verbs *per se* and verb phrases in LSP are integrated to such an extent that in some sections of this chapter, a degree of overlap in their treatment has proved necessary. In the chapter I am thus seeking to draw together the elements which (i) influence the verb in LSP combinants, (ii) help to identify and define the collocatory combinants which proponents of a special subject field 'automate' through usage, the meaning of which may however be opaque to non-specialists (e.g. *to type a virus*), and (iii) give indications of the terminological relationships identified

Chapter 4: 161
in them. These elements are summarised in Table IV.1. It should be reiterated that 'autonomy' in relation to a subject field indicates the degree of speciality of the lexis; a highly specialised field will have restricted lexical autonomy, where words appear only in that field; less specialised fields will have lexical items that are used in more than one field, and finally words from LGP will be frequent in less technical spheres. Furthermore, differences in valency between LSP and LGP (patients present with ....) also support the notion of autonomy in a special subject field.

<table>
<thead>
<tr>
<th>Terminological representation</th>
<th>'LSP-ness'</th>
<th>Grammatical representation</th>
<th>Colocation</th>
<th>Valency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Verb <em>per se</em> (intensional characteristics) e.g. <em>it gels</em></td>
<td>Degree of autonomy varies according to the subject field</td>
<td>Active Passive (in)transitive Reflexive Mood, Voice Tense etc.</td>
<td>Not applicable</td>
<td>Applicable only for verbs with mono valency</td>
</tr>
<tr>
<td>Verb + object = proposition e.g. <em>to levy taxes</em></td>
<td>Type of object = (in)animate, abstract/concrete etc.</td>
<td>Verb + n.(phr.) /prepositional phrase, [(in)direct object]</td>
<td>Restricted</td>
<td>Usually two or three</td>
</tr>
<tr>
<td>Internal disjuncture (i.d.) e.g. <em>to confer (little) protection</em></td>
<td>Restricted LSP = little i.d.; less-restricted LSP = i.d. more likely</td>
<td>Adverb(ial phrase), 'free' adjectival phrase, clause, negative</td>
<td>Probable in LGP; in LSP may form part of new term</td>
<td>Possible; circumstances likely to be restricted in number</td>
</tr>
</tbody>
</table>

Table IV.1 Elements influencing verbs and verb phrases in LSP.

In this chapter I have contrasted the terminological aspects of an LSP phrase with lexical aspects (4.1). In 4.2, the influence of the subject field on the verbs it contains is assessed; semantic changes can occur when certain verbs pass from LGP to LSP and vice versa, either diachronically or when their
environment changes. The degree of specialisation of verbs in different subject fields, and in French and English according to their differing ‘behaviour’ in these languages, are evaluated in an attempt to show how they acquire LSP status. The problems of defining the limits of a verb phrase in LSP are highlighted in 4.3, with comments on the contrastive nature of translating phrases and the role played by collocation in helping translators. Collocation is then linked with valency (4.4) to help point to terminological relationships in LSP verb phrases (4.5). From this, orientation can be assessed through linking terminological/encyclopaedic structures with valency (4.6) and the importance of including a valency slot for the psychological subject is discussed (4.7). The problem of internal disjuncture in phrases is addressed in section 4.8. Finally, I have drawn conclusions from all these aspects and a summary of the LSP status of verbs is given in 4.9. A combination of the foregoing provides criteria for building frames for LSP verb phrases (4.10). I believe I have been able to provide a theoretical basis for categorising LSP verb combinants.

4.1 Terminological versus lexical aspects of an LSP phrase

It should be borne in mind that the terminological or conceptual aspect of an LSP phrase is separate from the purely lexical aspect, which provides the surface structure. Compare for example to draw a conclusion, the metaphoric to draw the line, to (with)draw cash from the bank and to (?with)draw a gun, where the semantic characteristics of the verb are transferred to the proposition or object (in these examples merely implying a certain movement); the verb acts as a carrier in the process of semantic weakening. The propositional logic in verb phrases draws the meaning away from the lexical surface level to the underlying concepts. It is interesting that the above examples are from LGP, because the notion of semantic weakening
seems somewhat counter-intuitive when considering LSP, where one expects terms, whether nouns or verbs, to be conceptually, and therefore semantically, rich. However, it has already been noted that verbs in LSP are often 'emasculated' by the term accompanying them. Moreover, given the preponderance of verbs in the passive voice, particularly in the sciences, the effect of rising tension as the sentence progresses towards the verb gives it additional semantic weight and action, e.g. *the fall in infectivity* *can be greatly inhibited.* The process of nominalisation, however, tends to lead to stasis and helps to obviate the overt mention of a subject. This corresponds to the scientist's viewpoint which does not focus on the person who is performing the work (the psychological subject [cf. 4.7] but emphasises the methods and results of the work undertaken (the *fait accompli*, or static part).

As stated, from a stylistic viewpoint, the semantic emphasis in English moves towards the end of the sentence, thus holding the interest of the listener or reader. However, the stylistic mechanism varies from language to language: in German, for example, the verb in sentence-final position has the same effect, whereas in other morphological languages such as Latin and Russian, word order is largely immaterial. Uninflected English relies on element order which can however be fluid, needing recourse to structuring the syntax to establish the theme. Meaning in English can therefore be conveyed through element order or by using a large number of function words.

From the terminological point of view, the internal structure of an LSP combinant can be divided into two parts conceptually: first, concepts with verb attributes, the semantic characteristics of which may be transferred to become concepts in the second part, i.e. the object, with its propositional characteristics. It is on this semantic variability, in conjunction with the
subject field, that the orientation of the combinant hinges.

There are in addition certain restrictions that apply to special language verb phrases. Syntagmatic relationships cover phrases which form a syntactic unit and are therefore subject to restrictions in the grammatical components they include. They also have semantic and lexical restrictions. Syntagmatic relationships may therefore have the following selectional restrictions:

- **syntactic restrictions**: a proposition which completes a verb phrase will be a noun or noun phrase; in the case of intransitive verbs, it will be a preposition + noun phrase. As already noted, valency may differ between verbs in LGP and LSP, e.g. *Most MuLV strains express inefficiently in cells in early developmental stages...* where *express* has 'themselves' understood and would probably include it if the verb were used in an LGP sense. The difference would require a variation in the valency slots for LGP and LSP use.

- **semasiological restrictions**: the sentence or phrase must make professional sense, hence there is semasiological, i.e. selectional, restriction with respect to the noun or noun phrase following the verb. The subject field influences and restricts the choice of noun or noun phrase; the more specific a verb is to its subject field, the fewer will be the number and type of its actants (see 4.9.1 and Tables IV.9 and 10), e.g. *porter atteinte à l'honneur de/au principe de; to lyse cells.*

- **lexical 'term-naming' restrictions** which are manifest as collocational restrictions in a particular field, e.g.:
  - *it binds fibronectin*
  - *it binds the cellular receptor*
In general language these syntagmatic lexical restrictions are more likely to be individual lexemes than those occurring in special languages where terms often comprise several lexical units. Lexical restrictions may form encyclopaedically but from the point of view of collocation they may be compatible by dint of habitual co-occurrence.

**onomasiological restrictions** are assessed in the terminological sense propounded by Riggs (1989: 89), i.e. starting with concepts and progressing to the terms designating them. These restrictions are difficult to assess because the number of concepts is infinite. Neologisms in both nouns and verbs arise to describe the growing number of concepts, as do neosemanticisms, where existing words acquire new meanings. In fact, initially the concept of 'onomasiological restriction' might appear to be an oxymoron because the idea of 'restriction' implies that an attempt is being made to define an unknown and unlimited terminology. Within this context, the onomasiological approach, however, allows scope for description; instead of prescriptive, standardising terms, concepts can be described by a number of terms which, synchronously, are finite. This approach could be of use when assessing the nouns and noun phrases that complete a combinant.

4.2 Influence of special subject fields on verbs

The quest for orientation leads to the first aim of this chapter which is to evaluate through semantic assessment the extent to which the subject field influences the degree of speciality, or 'LSP-ness', of a verb, and there has been remarkably little research in this area. The semantic change wrought in verb
concepts due to the influence of their subject field is a key factor in the conceptual combinability of a verb with a proposition, e.g. to drive someone mad (LGP), to drive a 'bus ('low-grade' technical LSP), to drive a hard bargain (LSP, or figurative in LGP). An LSP phrase is naturally influenced semantically by the subject field in which it occurs. More specifically, the verbs in LSP phrases may undergo a semantic change. Draskau (1986: 25, 26) categorises verbs in LSP as (a) those with little or no LSP content; (b) those which are terms in themselves; (c) those in the 'twilight zone' which often collocate with an LSP term and take on a special meaning, such as the examples of to draw in 4.1. Enlarging on this classification to include different types of LSP subject field, i.e. variation in the restricted nature of their terminology, I have noted the following points concerning the behaviour of verbs:

1. In a highly restricted subject field, such as virology or computing, which may be considered 'terminologically autonomous', that is, subject fields which barely overlap with others, there is a high number of verbs which belong solely to their subject field (e.g. to lyse in virology), or which occasionally have been borrowed from another LSP (e.g. to fingerprint from police detective work to virology: Thomas, 1983: 190). In English in particular, compound verbs may be formed from a surface combination of noun + verb, e.g. to hydrogen-bond (underlying structure: to bond with hydrogen) or verb + verb, e.g. to stir-fry, or verb + verb where one verb has an adverbial function, e.g. to freeze-dry, i.e. to dry from the frozen state (and not to freeze by drying; in other words, it is not a verb + adjective construction). It is important to note that these verbs have a minimal valency pattern because one of the valency elements becomes absorbed into the compound construction. Moreover, in their subject field, they have restricted semantic collocability with their objects and in addition the resultant phrase does not permit internal disjuncture so readily
as do phrases in less narrowly restricted subject fields. The number of their actants is limited, often to one, e.g. to debug a program, to debug a system (with a software tool/with a colleague = optional circonstants). The question of valency is discussed in greater depth in sections 4.4, 4.6 and 4.7. Where highly subject-specific verbs form part of a combinant, orientation is likely to centre on the verb. Furthermore, the restricted grammar of special subject fields and in the telegraphic style of newspaper titles cannot be reflected in a dictionary.

2. In specialised subject fields such as law and economics, which by their nature depend on other subject fields (what I am naming 'overlapping' subject fields), there are fewer LSP verbs per se, but nevertheless a comparatively 'close' degree of collocation is also encountered, where the verb has few actants; indeed, the fewer the actants, the more tied they are to their verb, e.g. to terminate a contract, appears idiomatic, almost cliché-like, even though in this case the expression is ambiguous and could mean either at its end or prematurely. It is moreover noteworthy that the verb terminate is semantically wider and less specific than the noun termination which invariably refers to abortion (98% of occurrences in COBUILD). A formal effect occurs in discourse due to the frequent use of words of Latin etymology, thus distinguishing such amalgamations of relatively restricted subject fields from those which have less rigid collocational patterns and which tend to fall into the categories of the verbs mentioned in (3.) below. The verbs are often from LGP but the phrases of which they are a part belong, through semantic and collocational restriction, to the subject fields. It is interesting to note that in the main, in many special languages which have a long and highly regarded professional standing, such as law and the medical sciences, the etymology of the verbs derives from Greek and Latin; in other words, the forms which have been
imposed by dominant societies with an existing strong social structure, whereas everyday speech in English tends to favour the Anglo-Saxon version of the verb, e.g. verify/check.

3. In 'popular' technical subject fields which directly affect the man-in-the-street, e.g. automotive engineering, audio and televisual communications, photography and household appliances (the majority of households in the Western world have a telephone, car, television, camera and washing machine), the verbs which occur are from general language, couched in what may be termed 'in-house jargon' with a restricted grammar, often with the imperative mood of the verb, e.g. set timer, load film, respond to button input. The phrases have restricted collocability, but the verbs are seldom peculiar to the subject field. The criteria for identifying such quasi-specialised verbs are (i) the difference in their frequency when used purely in an LGP sense from that used in a technical sense and (ii) the semantic nature of the accompanying noun or noun phrase which forms a combinant.

A number of verbs in categories (2.) and (3.) above may be termed 'support' or 'empty' verbs, that is, verbs that are necessary for syntactic reasons but have little semantic content. They are often to be found with deverbal nouns which have their semantic dynamism 'frozen' and represent stasis, e.g. to conclude = to reach a conclusion, and are therefore transitive. The orientation of a verb + noun phrase construction can usually be deduced by assessing which part of the phrase has the greater degree of specialisation and, as will be seen in Chapter 5, by invoking the use of frequency.

The use of nominalisation is an important factor to consider in relation to studies on verbs because it has the effect of arresting the process implicit in the verb, thus calling a halt to the action and allowing the resulting
series of static situations to be examined, rather in the manner of a series of frames for cartoon films. Nominalisation is a mechanism particularly favoured in French in both LSP and LGP and can lead to vagueness in a sentence. Gowers (1954) quotes G.M. Young as stating that 'an excessive reliance on the noun at the expense of the verb will, in the end, detach the mind of the writer from the realities of here and now, from when and how and in what mood the thing was done, and insensibly induce a habit of abstraction, generalisation and vagueness'. Certainly the increasing trend towards nominalisation in scientific writing (Halliday, 1988), with its resultant reliance on support or 'carrier' verbs, leads to stasis, as well as to an avoidance of naming the agent, particularly when this is human. Abstract nouns are particularly notable in French legal terminology, e.g. mesures, nullité, where the accompanying verb is 'emasculated' (prendre des mesures; frapper une clause de nullité); Roman law has clearly managed to cover all eventualities by 'freezing' the events being examined through nominalisation, the better to be able to assess them. In the French biological sciences corpus, nominalisation in the form of the -isation string represents 0.25% of the total French corpus (Annex 5). A random look at the verbs accompanying some of the nouns produced the following examples of carrier verbs:

- la banalisation doit se constituer
- entamer (embark upon) la concrétisation
- parvenir à l'harmonisation
- aborder (tackle) l'informatisation
- utiliser la désinsectisation
- s'agit d'immunisation

The above nominalisations are likely to be comprehensible to a
speaker of English yet this form does not appear to be becoming part of current English usage, although as Halliday has noted, it is becoming more prevalent in scientific writing (Halliday, 1988).

4.2.1 Verbs in verb phrases from 'overlapping' subject fields and 'quasi-LSP' verbs in English and French

To provide evidence of the above categories, assessments were made of two corpora of considerably different size. The smaller contain parallel French and English texts comprising approximately 20,000 words in each language in overlapping subject fields, law and commerce, and law and pharmaceuticals, resulting from Proceedings of the European Court of Justice on pharmaceutical topics. The larger corpora of the biological sciences have already been described at the beginning of the chapter. It was thought it might be possible to ascertain whether the results from the smaller would in any respect be corroborated by the larger and thus whether, in a restricted subject field, a small corpus would provide conclusive evidence. For most analyses, clearly the larger the corpus the better, to provide conclusive evidence and brook no argument.

Concordances were run firstly on the small corpora. An interesting point has emerged in all the corpora studied: it is noticeable how few in number are the different verbs in these types of subject fields. In the English corpus these comprised approximately 8% of the total number of different words, a figure which is corroborated in the larger corpus of half a million words. (For the type of analyses in Chapter 5, a large corpus is necessary to provide a sufficient number of examples.)

My aim is to corroborate the hypothesis that 'autonomous' LSP
subject fields have a number of true LSP verbs but that many LGP verbs, when used in these subject fields and combined with subject-specific propositions, have a higher frequency than in LGP, and may acquire a special meaning in a particular subject field. In addition, they sometimes change their valency pattern, which I am calling 'semantic valency', e.g. *to dispense a prescription*. A comparison has been made between words, or 'types', from the large English corpus of the biological sciences, selected by the author because of their frequent use in the biological sciences; they include some noun/verb homographs out of interest, e.g. *cause, constructs, isolate* (the last two are always nouns in the LSP texts). The scientific 'types' have a considerably higher frequency in LSP than in LGP, as can be seen from Table IV.2, and this places them in the categories of LSP and 'quasi-LSP' verbs. The marked difference in frequency supports the hypothesis of the subject field influencing the 'LSP-ness' of the verb, i.e. the greater the degree of specialisation of the text, the more it will be characterised by highly restricted vocabulary.

Table IV.2 (b) is a comparison of French verbs from the biological sciences corpus and an LGP corpus of texts from 'Le Monde'. The verbs have been extracted from Tables IV.5 and IV.12 and interestingly show less variation in frequency between LSP and LGP than the English corpora, although not all were contained in the LGP corpus. This is probably because the French biological sciences corpus is rather wider in the scope of its content, since it contains medical material which can often be found in an LGP corpus of 'educated layman' standard, such as 'Le Monde'. It will be noted that the verbs have not been lemmatized, to reflect a true picture of the way in which they appear in the corpora.
<table>
<thead>
<tr>
<th>Frequency per %</th>
<th>Type</th>
<th>Frequency per %</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>LSP</td>
<td>COBUILD</td>
<td>LSP</td>
<td>COBUILD</td>
</tr>
<tr>
<td>0.59</td>
<td>0.01</td>
<td>activate</td>
<td>1.09</td>
</tr>
<tr>
<td>5.67</td>
<td>0.01</td>
<td>attenuated</td>
<td>1.30</td>
</tr>
<tr>
<td>0.77</td>
<td>0.02</td>
<td>binds</td>
<td>1.87</td>
</tr>
<tr>
<td>0.80</td>
<td>0.56</td>
<td>bound</td>
<td>4.30</td>
</tr>
<tr>
<td>4.30</td>
<td>1.50</td>
<td>cause</td>
<td>1.46</td>
</tr>
<tr>
<td>1.05</td>
<td>0.02</td>
<td>characterised</td>
<td>3.02</td>
</tr>
<tr>
<td>1.44</td>
<td>0.02</td>
<td>characterized</td>
<td>6.43</td>
</tr>
<tr>
<td>0.54</td>
<td>0</td>
<td>cleaved</td>
<td>5.31</td>
</tr>
<tr>
<td>0.57</td>
<td>0.04</td>
<td>coded</td>
<td>1.07</td>
</tr>
<tr>
<td>0.59</td>
<td>0.05</td>
<td>codes</td>
<td>1.83</td>
</tr>
<tr>
<td>2.26</td>
<td>0.01</td>
<td>coding</td>
<td>0.98</td>
</tr>
<tr>
<td>0.75</td>
<td>0.02</td>
<td>confer</td>
<td>1.53</td>
</tr>
<tr>
<td>1.76</td>
<td>0</td>
<td>conserved</td>
<td>1.03</td>
</tr>
<tr>
<td>1.37</td>
<td>0.01</td>
<td>constructs</td>
<td>0.36</td>
</tr>
<tr>
<td>0.59</td>
<td>0.02</td>
<td>cultured</td>
<td>2.12</td>
</tr>
<tr>
<td>0.82</td>
<td>0</td>
<td>deleted</td>
<td>0.57</td>
</tr>
<tr>
<td>0.36</td>
<td>0.04</td>
<td>digest</td>
<td>1.28</td>
</tr>
</tbody>
</table>

Table IV.2 (a) Comparison of a selection of LSP and quasi-LSP verbs from the biological sciences corpus in English with the LGP corpus, COBUILD.
Table IV.2 (b)  Comparison of a selection of LSP and quasi-LSP verbs from the biological sciences corpus in French with an LGP corpus extracted from 'Le Monde'.

The verb entraîner serves as an example here of a quasi-LSP verb: when used in a technical sense, that is, assuming an autonomous subject field, entraîner has a restricted meaning, but when it is used in a legal sense (an overlapping subject field), it has the meaning of obtain, produce, result in, a less specific action than when it is used in the technical sense and translated into English, because in this language a different verb is needed (which, incidentally, also denotes action): entraîner une machine is translated in English by to drive a machine. However, in the large French corpus of the biological sciences, the meaning reflects in all instances that of the 'legal' sense. This is an example of the influence of the subject field on the proposition with its combinatory verb; moreover, the example bears witness to the value of translating a phrase into another language to illustrate the contrast. As is seen in the next section, the translation of phrases often results in the use of

Chapter 4: 174
4.3 Defining the limits of LSP verb phrases: the help of contrastive translation

Since translation concerns the transfer of as many characteristics as possible from the source language to the target language in the most effective way, often by using different parts of speech, it is beneficial to make a comparison with another language to help with the recognition of an LSP phrase, and not simply a verb phrase, because the limits of a phrase are often more easily identifiable when considered from an interlingual point of view. The examples which follow serve to illustrate some of the problems in assessing phraseology multilingually and in determining the limits of a phrase due to differences in the parts of speech used to express the same phrase. This point has been made *inter alia* by Picht (1990: 32) who provides a good example of contrastive translation: in German, *eine Flüssigkeit hinaufpumpen* corresponds in Spanish to *elevar un líquido por bomba*, showing how the phrase undergoes a linguistic re-deployment; the action of the verb *pumpen* is represented here by the technical adverbial phrase *por bomba* (although there is a verb 'to pump' in Spanish), while the direction implicit in the verb action *elevar* is represented by the preposition *hinauf*. This is probably a question of style and could be corroborated by assessing whether there is a Spanish tendency to translate in this manner. However, whether or not the action is present in the verb does not detract from the fact that the overall characteristics are the same, a useful guide to the identification of the limits of a phrase. This is essentially a 'chassé-croisé' in Latin languages.

However, as already stated, it is not always simple to define the
limits of a verb phrase, even when assessing, from a contrastive viewpoint, slightly larger phrases which include the subject as well as the verb. The following examples are taken from the Swedish EG-term bank and illustrate the type of problem which can occur. Phrases may start with determiner + noun (the members shall resign as a body; these stages may be curtailed), not necessarily in subject position, or they may start with a personal pronoun (they shall abstain from any measure which could). This last phrase, however, does not form a syntagmatic unit and it appears that the criteria for the choice of words which end a phrase may sometimes in English be somewhat arbitrary; moreover, although the collocation abstain from is shown, there is no indication of what collocates with measure, such as adopting. The reason for this, I think, is because of the source language from which the translations are made, which in this example appears to be French. Thus the clause in French may illustrate a construction or collocation which sits less happily in one of the other languages into which it is translated. In the example just given, the French ils s'abstiennent de toutes mesures susceptibles de produces prepositional collocations of s'abstenir de (v. + prep.) and susceptible(s) de (adj. + prep.), constructions which are useful for the translator. The result in Swedish is similar to that in English: de skall avhålla sig från alla åtgarder ägnade att; this phrase appears almost as inconclusive as the English but at least gives the prepositional collocations avhålla från and ägnade att. The examples certainly show a contrast in the parts of speech used in the different languages and, if the object is to see how terms ‘behave’ in context, how relevant is it to define the limits of phrases in different languages?

Hartmann has viewed the translation process more broadly, at the level of whole texts, stating that ‘Translation of discourse is only possible if we

1 The EG-term bank has been developed by the Tekniska Nomenklaturcentralen (TNC) in Stockholm and contains E.U. terms and phrases from the Treaty of Rome and the E.C. White Paper of 1985.
know what the equivalent structures are in the language into which we want to translate. And this knowledge is gained .... from contrastive textology.' (Hartmann, 1980: 67).

Returning to the level of the phrase, probably the number of collocations in a phrase is far more useful for the translator, and for a native speaker who needs to have recourse to an LSP which is possibly unfamiliar. However, some measure of systematic analysis is necessary so that a translator or subject specialist can derive maximum benefit from a search for a correct phrase and not waste time with too many 'hits' in response to queries posed.

EURODICAUTOM functions as a multilingual dictionary, giving individual examples. Glossaries are the next step towards providing phrases but deviate from terminological representation because they are unlikely to include verb collocations.

Since the choice of words constituting a phrase may differ from language to language, the problem remains of what constitutes a phrase and how its constituent parts may be defined. Nattinger and DeCarrico (1989: 118) explain it at the syntactic rather than the semantic level thus: 'Lexical phrases are multi-word lexical phenomena that exist somewhere between the traditional poles of lexicon and syntax. They are similar to lexicon in being treated as units, yet most of them consist of more than one word, and many of them can at the same time be derived from the regular rules of syntax, just like other sentences. These phrases are patterned sequences, usually consisting of a syntactic frame that contains slots for various fillers, and run the gamut from completely fixed, unvarying phrases to phrases that are highly variable'.
Nattinger and DeCarrico's definition applies at a monolingual level and to general language in English but also takes into account the arbitrary nature of corresponding multilingual phrases, where the way in which concepts are represented linguistically may vary widely from language to language (Nattinger and DeCarrico, 1989: 129); for example, the 'chunking' of complexes in Anglo-Saxon languages which are based on oral language having primacy over written language. Although Nattinger and DeCarrico mention a syntactic frame with slots for various fillers, they do not appear to give guidance on the valency, or number of fillers (mostly actants) that are likely or possible, and this I hope I have been able to quantify (Chapter 5).

The examples given in this section would appear to need some sort of 'tightening procedure' to limit the phrases so that they appear syntagmatically complete in each language. The limit of a verb phrase has to be determined by the proposition it contains, regardless of language and surface structure. A proposition may be animate or inanimate, concrete or abstract, depending on the subject field: as has been noted, law has a large number of abstract nouns. In general, however, such lexical functions are less necessary than in general language because the type of function is likely to be known to the user. Phrases from highly restricted subject fields permit internal disjuncture less readily than do those from a subject field such as law which never stands alone but always embraces any number of different subject fields, cf. the phrases taken from legal texts in English and French which, in the examples in Tables IV.8, IV.9 and IV.10 (section 4.9.2), are used with the LSP of pharmaceuticals.

Comparative translation studies provide considerable insight into the transfer of meaning between different parts of speech. As the examples have shown, however, such comparative studies may give rise to problems in defining the limits of a verb phrase, and there are a number of syntactic and
semantic features which therefore need to be considered:

**Syntactic**

1. What do LSP verb phrases consist of grammatically?
2. Does their syntax give any indication of relationships in verb phrases?
3. What is the number of their actants?

**Semantic**

4. What is the type of their actants?
5. What are the relationships in verbs and verb phrases?
6. Will the construction of verb frames be feasible?

Two examples of verb phrases taken from the English corpus incorporate nominalisation and illustrate valency differences when the nominalisations are 'reverbalised'. Underlined are three instances of nominalisation, one of which can become the carrier verb *use*, with a second carrier verb *reduce*. The first phrase is active and divalent:

*Use of codon wobbling to reduce the recombination frequency of vectors*

However, after reverbalisation, one monovalent and two divalent structures result:

**Monovalent**

*When the codon wobbles*

**Divalent**

*it is used to reduce the frequency*

**Divalent**

*(frequency) with which the vector recombines*
The second phrase shows divalency in:

*Tn 916 undergoes excision, recircularization and insertion*

and becomes monovalent when passivized:

*Tn 916 is excised, recircularized and inserted*

These valency differentials are explored in this chapter and it is hoped that answers have been provided, particularly in sections 4.4, 4.5, 4.6 and 4.7, and in Chapter 5.

4.3.1 Collocation in translation

The attributes of collocation as a valuable tool for the indication of orientation have been discussed extensively in the previous chapter. Collocation is mentioned briefly here (a) because of its importance as a help to translators and (b) because of the interesting light it throws on valency, with which it is linked in the next section. The following example illustrates the type of problem that may cause confusion to a translator:

(a) i.o. + verb in 3rd person only

\[\ldots\] manqué aux obligations qui leur incombent\[\ldots\]

(b) v. + i.o.

\[\ldots\] coopération mutuelle qui incombent aux états membres\[\ldots\]

The problem arises because *incomber* can have the following translations in English:

(a) \[\ldots\] obligations which devolve upon them \[\ldots\]
\[\ldots\] obligations which fall on them \[\ldots\]

but:

*\[\ldots\] observations which are incumbent on them \[\ldots\]*

*\[\ldots\] observations which bequeath them \[\ldots\]*
Conversely:

(b) [...] cooperation which is incumbent on [...] 

 [...] cooperation which behaves [...] 

but: *[...] cooperation which ?devolves upon [...] 

*[...] cooperation which ?falls on [...] 

The textual collocation needs to be given for the correct verb to be chosen; alternatively, some indication of a contrastive nature should be given if cooperation is followed by of + responsibility of a person/people so that a more colloquial translation results. Indeed, Jackson has stated that 'collocational studies, though they are concerned with meaning, are not usually located within semantics; they are referred to rather as 'lexical studies'.'' (1988: 244); in other words, collocations have a referential function within their text, whereas terms are extra-textual, i.e. encyclopaedic. Not only is the identification of collocations, which can be corroborated by examining frequency statistics, an invaluable aid to translators, but it is also an important step in helping to define combinants per se and in determining their orientation as an aid to placing them in a dictionary.

4.4 Valency

A brief overview of the main proponents of valency has already been given in Chapter 1. This section addresses valency in greater detail and subsequently links it with collocation to assess whether these aspects can combine to confirm the identification of relationships in verb phrases, which in turn could lead to ascertaining the orientation of an LSP verb and the production of frames for these. Of particular value would be the identification of instances where the valency of a verb differs between LSP and LGP.

Chapter 4: 181
Valency theory considers the verb to be the pivot of a phrase or sentence. Fellbaum (1990: 278) states that, because fewer verbs exist than nouns, verbs have a wider role to play than nouns and are as a result more polysemous. This observation is corroborated by the work of Draskau (1986: 28) who, when comparing the same verbs in both general and special language, ascertained that in one of the scientific fields (veterinary), the verb to excite had a valency of one or two, with at least one 'concrete' or inanimate actant, whereas the same verb has greater valency in general language, where its actants are abstract on the surface, with a deeper structure invoking a human actant. Draskau gives the examples Vibrations will excite the gear casing - an LSP use - and Her beauty excited comment i.e. Her beauty made (people) (to) comment, in its general language sense (Draskau, 1986: 28, 29). In addition, she gives an example of the verb in an LSP text where it is however used in its LGP meaning: The adrenalin injections excited the rats, in which the second actant is animate. It is interesting to note the author's findings from the LSP of law, which contrast with those of Draskau (1986: 28); in the examples in Tables IV.8 and IV.10 taken from law texts, all the nouns which form the actants of the verb are abstract and it is therefore clear that the subject field influences the type of actant of the verb.

In addition to the influence of the subject field on the type of actant, the number of actants has also been assessed, particularly in relation to the syntactic aspects discussed in section 4.5. Moreover, the role played by valency is considered to be an integral component in the assessment of relationships in verb phrases. The majority of English verbs range from having zero valency to tetravalency. In the following examples of valency, s= subject; v = verb; o = object; io = indirect object; and prep. = preposition.

Chapter 4: 182
1. **Zero valency: it + verb**

   Usually verbs used in meteorology, except in a figurative, metaphoric sense, e.g. *it's snowing cherry blossom*, and verbs such as *smell*, where zero valency indicating *bad* is understood. In Italian and Spanish, the semantically empty subject pronoun is not needed.

2. **Monovalency: subject + intransitive verb (denotes action, state)**
   - animate subject performing bodily activity (*dream, sneeze, yawn*)
   - inanimate subject (*the tide ebbs, the water evaporates*)

3. **Divalency: s + v + d.o./prep. phrase [can take passive form]**

   *(I read the book; the book was read by me)*;

   s + v + prep. + adverbial phrase of manner/instrument *(I spy with my little eye)*;

   s + v + prep. + (adverbial phrase of time or place) *[can substitute then, there] (I saw the children at school/this morning)*;

   s + v + predicate [no passive]

   *(Mary is an intelligent person)*

   A large number of commonly used verbs come into this category.

4. **Trivalency: s + v + d.o. + i.o.;**

   s + v + i.o. + d.o.;

   s + v + d.o. + prep. *(to/for)* phrase

   Typically verbs such as *give, write* etc.

5. **Tetravalency:**

   s + v + d.o. + prep. phrase + prep. phrase;

   Chapter 4: 183
s + v + i.o. + d.o. + prep. phrase

  he claimed a large sum from the insurance company for
  the damage;

and from Allerton (1982: 116)

  The firm charged/paid Oliver a large sum for the job.

Occasionally greater valency can occur, as in John built his brother a shed from corrugated iron with a hammer in the garden yesterday.

Apart from zero or a valent verbs, all the above categories must have a subject which is the agent or instrument of the action of the verb.

Allerton has treated the subject of valency in English verbs in great depth and suggested 30 categories for these (1982: 145-147). He describes verb 'elaborators' (1982: 33) as elements which are dependent on a verb, such as noun phrases and adjectival phrases, e.g. he painted the town red; these are what Tesnière calls 'actants', while Allerton's 'adverbial elements' of time and place are designated 'circonstants' by Tesnière, and these may occur with any verb. This is an extension of the idea of the government of verbs in which the verb has the prime role and is invariable, while elements dependent on it are variable. Allerton (1982: 142) applies valency to any lexical category and refers to a 'lexical core' or 'governor' which appears to equate with 'base' in collocation, while his lexical 'specifiers' or 'dependents' or 'complements' would identify with the 'collocate' or 'collocator'. This linking of collocation with valency, which can be substantiated through analyses of frequency, can shed light on terminological relationships (4.5).

It has already been noted that there may be a difference in valency in
special language (cf. the example of patients present with ....[symptoms] and those of Draskau given above). Allerton has also noted that a different type of actant can occur in special language, with a variation in valency, quoting she coughs in general use, but she coughs blood in a medical sense (1982: 52) and makes a similar analogy with drink and took (the vaccine took) (1982: 71). Similarly, she swallowed the toffee (LGP: transitive verb) but I can't swallow (LSP of medicine: intransitive verb). It is interesting to reveal such a difference in valency patterns in LSP and further examples have been sought in the texts analysed to ascertain to what extent such patterns can be substantiated and whether similar patterns occur in a related language. Sometimes there may be a problem in identifying whether an 'actant' or a 'circonstant' is implied; a preposition often introduces a circonstant but this is not necessarily so, e.g. to hospitalise (someone)/by ambulance.

To summarise, the chief factors which need to be taken into consideration when assessing valency are summarised as follows:

Type of actants
1. whether actants are influenced semantically by the degree of LSP-ness of the verb;
2. whether they are influenced by the subject field;
3. how they, in conjunction with their verb, point to terminological relationships.

Number of actants
4. whether they behave syntactically in the same way in LSP and LGP;
5. whether there are circumstances where grammatical differences occur in LSP and LGP;
6. whether there are fewer actants in LSP than LGP given the
greater incidence of passive verbs to convey meaning in LSP;

7. how extensive the differences are in the number of actants for a verb in LSP;

8. whether they are mandatory or optional;

9. whether there is a difference in their type of actants.

The last point is particularly relevant to English, given its flexibility in including or omitting words; for example, the object may be suppressed if reflexive use is understood (Allerton, 1982: 136); this sometimes occurs in LSP, as was seen in the example given earlier in this chapter:

Most MuLV strains express inefficiently in developmental cells in early stages....

and in:

....and may form an aphipathic helix which could insert into the cell membrane in a manner....

where 'insert' could be written 'become inserted' or 'itself' be added.

4.5 Terminological relationships in LSP verbs and verb phrases

Verbs and their associated noun propositions, or actants, form semantic relationships with each other which can denote action, cause, duration, function, location, process and state. A link is made with section 4.5.2 which enumerates the syntactic aspects of verb phrases in addition to identifying terminological relationships. In some instances more than one relationship may be deduced from the same construction, as can be seen in some of the examples in

Chapter 4: 186
4.5.2. It seems logical that the semantic aspects between a verb and its proposition will prove fruitful in supplying relationships which in turn can lead to an analysis of their syntactic forms. Firstly an assessment of the semantic relationships found in verbs per se is made; this is for the sake of completion because the more restricted the subject matter, the fewer the number of special language verbs which fulfil the following categories, and indeed for the sake of clarity it has been deemed more expedient to provide examples from general language. However, when analysing verb phrases, syntactic relationships are readily recognised in the LSPs studied and examples of these are given in 4.5.2.

4.5.1 Semantic relationships in verbs

Semantic relationships are paradigmatic and may also indicate terminological relationships such as part/whole (meronymy).

(i) Hyponymy, e.g. Billy is munching/eating an apple. (Munching is a kind of eating, usually involving swallowing; both involve the same action).

(ii) Similarity, or quasi-synonymy, is a special case of hyponymy often found in English. Here it refers to 'semantic' similarity, for example, in the different etymologies of verbs, e.g. close/shut a door, although conversation and correspondence (as in This correspondence must now close - ed.) collocate with close but not shut.

(iii) Antonymy, where not just an opposite but often a change of state is implied, e.g. to be born, to die. This category includes many de-adjectival verbs, e.g. to whiten, to blacken.

(iv) Taxonomy, which can be a form of similarity as in (ii), but also
indicates a hierarchy, as in classification and terminological systems, e.g. *to move* would be the superordinate of *to walk, to drive* etc.

(v) Meronymy, another terminological relationship in which the part/whole relation in respect to verbs involves inclusion, e.g. *chewing* is a part of *eating* which does not involve *swallowing* and, because it is a part but not the whole of the eating process, in addition has a temporal aspect. *Chewing* would be referred to by Cruse (1986: 162-163) as an optional or facultative meronym of *eating* (meat is chewed, but not ice-cream); similarly, *chewing tobacco* is a separate action from the unlikely act of *eating* it, although *chewing meat* is usually a prerequisite to *swallowing* it.

(vi) Troponymy, a category designated by Fellbaum and Miller (1990: 568-569), indicates manner, e.g. *to barter* is to exchange goods rather than *to buy* which is a special form of barter, to exchange money for goods. The end result is therefore different, which distinguishes it from hyponymy in (i). The indication of 'manner' is not made by the addition of an adverb or adverbal phrase.

(vii) Eponymy, which includes denominal verbs from proper names such as *appertise* (after Appert, who in 1795 invented the preservation of foodstuffs by heat treatment in sealed containers), *bowdlerise, boycott, christen, degauss, galvanise, macadamise, mercerise, mesmerise, pasteurise*, in addition to trade names, e.g. *hoover, sellotape, xerox*.

These relationships are more likely to refer to LGP because of its wider variation; verbs in LSP have a more restricted use. In addition, the relationships deal with the verb *per se* but do not take account of words in the
surrounding text of the sentence. To take these into consideration, syntagmatic relationships need to be analysed and these may become apparent from studies of collocatory patterns.

It is not immediately clear how or even whether these 'autonomous verb' relationships can contribute to the notion of providing verb frames for special languages. However, it appears that the paradigmatic, terminological relationships in verbs per se can be linked to the semantic type of actants found in valency, and this is manifest by collocation, while a syntactic assessment of verb phrases elicits syntagmatic relationships such as:

- synchrony = temporal (when?, within?)
- troponymy = manner (how? - adv.)
- location = place (where? - v + prep. in, on, under, within)
- material (out of, of, made of)
- purpose/function (for, to, so that)
- process (verbs - can, maintain)
- agent (the person or thing doing the action)
- patient (the person or thing receiving the action or process)

In the next section an indication of verb phrase relationships (sections 4.5.2.2 - 4.5.2.4) is given from a syntactic standpoint.

4.5.2    Syntactic relationships in LSP verbs and verb phrases

The verb patterns of Benson et al. (1986: xiv-xxiii) have been assessed to ascertain whether relationships emerged from the collocations found in the nineteen English verb patterns they have identified. 'Purpose' is mentioned, and certain semantic groups are featured such as verbs which
involve passing objects from 'agent' to 'patient', and verbs of utterance. Benson et al., however, restrict their categories to grammatical patterns and do not seek to elicit relationships.

Pilz has identified the following categories of verb phraseolexemes, with examples in German (1981: 61-62):

1. verbs only          baden gehen (+ slang meaning)
2. + adj. element      rot sehen
3. + adv. element      jmdm (tüchtig) einheizen
4. + n. phr.           Trübsal blasen
5. + n. + prep. phr.   alle Hebel in Bewegung setzen
6. + adv.              seine Augen überall haben
7. + prep. obj.        in Kraft treten
8. + comparison        erscheinen/kommen wie gerufen
9. + relative clause   wissen, woher der Wind weht
10. special class: to be - phraseological lexeme: kurz angebunden sein.

These categories need to be taken into account when considering the structure of combinants. Collocations for example may form particular patterns according to the syntax of the combinant and observations of syntagmatic relationships discussed in 4.5.1 may be linked with paradigmatic (semantic) relationships.

4.5.2.1 Brief recapitulation of verb properties It seems appropriate at this stage to give a brief recapitulation of the wide grammatical characteristics of verbs, which apply to both special and general languages. Verbs may be *inter alia*:
1. (di-)transitive, intransitive, reflexive;
2. finite or non-finite;
3. indicative, imperative or subjunctive in mood;
4. modal or auxiliary;
5. active or passive in voice;
6. past, present or future in tense;
7. impersonal;
8. habitual or progressive in aspect;
9. phrasal/prepositional.

It will be seen how these characteristics are key points in providing the number of actants of a verb.

It is unlikely for reflexive verbs to be encountered in English LSP, whereas because of the structure of the language, this form appears in Swedish special language. In French too the reflexive is more widely used, particularly in LSP, e.g. *ces inclusions finissent par se rejoindre; ces plages se nécrosent en leur centre et, en se décollant du verre* [...] (Maurin, 1985: 250, 251). The corresponding structure in English would be the passive voice, using a nominalised verb, e.g. *This door is alarmed* (noted at Gatwick Airport). The register of different types of text also plays a role: the imperative mood is not likely to be found in learned works but it occurs frequently in instruction manuals. Modality is rare. Both active and passive voices are encountered, but there are few verbs in the future and conditional tenses, since scientific texts are usually surveying developments in the field and stating what the author has recently discovered to advance knowledge, with perhaps a few hypotheses for the future. Impersonal verbs are comparatively rare, and phrasal verbs do not appear, although prepositional verbs are common (see also 2.3.2).
The above observations are corroborated by the examples which follow. They have been taken from small concordances of virology texts containing approximately 53,000 words of French and 47,000 words of English, chosen to be as nearly parallel as possible in content, genre and style; they are not translations. The small concordanced corpora contain texts from the mid 1980s and have been kept separate from the large corpora which contain texts from the 1990s. The syntactic characteristics of the verbs have been linked to the terminological relationships in 4.5 (indicated to the left of each example).

4.5.2.2 Verb phrases with a transitive verb indicate action (A) or function (F), e.g.

A, F:  
*to induce the uptake of virus, polyhedrin occludes virus particles;*

F:  
*[..] an enzyme [which is] used to label nucleic acids [..]*

A:  
*[..] Sir Henry Dale, who *ampouled* in a dry and stable form what became the First International Standard (IS) for Insulin;*

A:  
*to toxoid diphtherla and tetanus toxins;*

A:  
*to freeze-dry/to vacuum-dry a product;*

A:  
*to dispense a prescription;*

A:  
*précrire le médicament;*

F:  
*protéger la souris contre l'infection rougeoleuse;*

A:  
*[..] lorsque les enfants *contractent* la rougeole.

4.5.2.3 Verb phrases with an intransitive verb + preposition show a wider variety of relationships in that they can be temporal (T) or locative (L) indicators of an action (A), function (F), process (P) or state (S), e.g.
\text{T, P} \quad [...] \text{different lipids will go (implied present) into gel phase and phase-separate at different levels of hydration;}

\text{A, L, P} \quad [...] \text{sugars can hydrogen-bond to phospholipid groups [...];}

\text{P} \quad \text{the module denatures into [...] strands;}

\text{L} \quad \text{the virus localises in the salivary glands;}

\text{L} \quad \text{Ces spicules apparaissent sur les microphotographes;}

\text{S} \quad \text{Les relations entre ces trois virus reposent sur les points suivants;}

\text{S} \quad \text{Un effet protecteur croisé existe entre les trois infections.}

In addition, they indicate possible differences in valency; for example, in \text{the virus localises in the salivary glands}, does the place where the virus localises provide sufficient evidence for scientific colleagues or do they have other expectations? The answer would result either in an optional spatial circonstant giving a monovalent verb (\text{the virus localises}), or in a mandatory actant providing divalency, as in the original clause. The result is judgemental rather than dogmatic; one needs to know whether extra slots are required, and of what type, whether of time or place, for example. Valency becomes an encyclopaedic matter and it is interesting to speculate what the statistical implications might be. It is furthermore interesting to note the incidence of compound verbs in the examples above.

4.5.2.4 Verb phrases with the verb in the passive (+/- preposition) usually indicate a process (P) but can less frequently denote action (A) or state (S) e.g.

\text{P} \quad \text{polyhedrin is synthesised late in virus infection;}

\text{P} \quad \text{granulins are highly conserved;}

\text{P} \quad \text{a polyprotein which is cleaved to give [...] protein;}

Chapter 4: 193
All cytokines were ampouled in microgram quantities [...];
l'atténuation a été obtenue par passages en série;
quand les cellules sont fixées au méthanol.

In addition to grammatical forms indicating relationships, true LSP verbs are specific to a subject field where they are highly restricted semantically. The following table briefly summarises the syntactic characteristics and terminological relationships found in the LSP verb phrases above, from a highly restricted subject field, although the function and action characteristics are not clear-cut and can be interchangeable (for example, the verb procéder when passive nevertheless denotes action).

<table>
<thead>
<tr>
<th>Grammatical form of LSP verb</th>
<th>No. of actants</th>
<th>Tense</th>
<th>Aspect</th>
<th>Terminological relationship(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transitive</td>
<td>&lt;3</td>
<td>Usually present or past Habitual or progressive with a degree</td>
<td>action, function</td>
<td></td>
</tr>
<tr>
<td>Intransitive</td>
<td>&lt;3</td>
<td>+ preposition of overlap;</td>
<td>action, locative, process, state, temporal</td>
<td></td>
</tr>
<tr>
<td>Passive</td>
<td>2</td>
<td></td>
<td></td>
<td>process, state,</td>
</tr>
<tr>
<td>Reflexive</td>
<td>&lt;2</td>
<td></td>
<td></td>
<td>process</td>
</tr>
</tbody>
</table>

Table IV.3 Example of syntactic characteristics and terminological relationships of LSP verbs in a restricted subject field.

Chapter 4: 194
4.6 Terminological/encyclopaedic structure and valency: relationships leading to orientation

Verbs have hitherto tended to be rather neglected in special language terminology which, as has been seen in previous chapters, concentrates for the most part on terms comprising one or more noun(s) or a combination of adjective(s) and noun(s), and on the conceptual relationships between terms so that hierarchical, thesaurus-like structures can be constructed. This neglect of verbs is doubtless due to the importance accorded to taxonomy and classification which has always obtained within the scientific community. It has been equally interesting to posit the sort of terminological relationships which might be found in verb phrases; because verb relationships are quite different from those which pertain to terms, an attempt has been made to assess both the pragmatic and the theoretical knowledge of the relationships which verbs acquire when they are associated with other words in a phrase or sentence (section 4.5). By assessing these relationships and grouping them, it is posited that more logical verb frames will result.

In the encyclopaedic structure of Austin's indexing work discussed in Chapter 2 (Austin and Butcher, 1969), we see that information retrieval is concerned with a natural semantic order, which is akin to orientation; in other words, the greatest semantic impact comes first in an indexed phrase and lessens as the phrase progresses. For example, in the following indexed term:

*infestation* (PROCESS)

*of timber* ('PATIENT' or RECIPIENT OF PROCESS)

*by wood-boring insects* (AGENT)
the nominalised process is *infestation*, the recipient of the process is *timber*, and the agent, *wood-boring insects*. If this example is considered from the viewpoint of mandatory and optional valency, the 'agent' slot is optional because the emphasis is on the damage and what is affected by it. The true order of importance or orientation of the term is passivised (*the timber has been infested*).

Since both nouns and verbs have valency patterns, it is interesting to ascertain whether a verb has the same valency as its corresponding nominalised form, e.g. *to conclude, to reach a conclusion* (it is usual that the object of a semantically empty verb is a nominalisation found in its first valency slot).

The question of differentials in valency can occur in a denominal verb, as the following example shows.

*The scientist field-tested the vaccine in cattle*

has two mandatory 'actants' or slots:

- *scientist* (actor)
- *vaccine* (recipient of process)

with two optional 'circonstants':

- *in cattle*
- *field* (spatial adverbial phrase).

However, passivisation destroys the grammatical subject:

*the vaccine was field-tested in cattle by the scientist*
resulting in one mandatory 'actant': the vaccine (recipient of process)
and three optional 'circonstants': in cattle
by the scientist (actor)
field (spatial adverbial phrase)

A valency dictionary would show vaccine as being followed by in.

This example neatly corroborates the observation that the 'recipient of process' in first position is the mandatory actant of a passivised phrase and is the prime indicator of orientation. This important conclusion is extended in Chapter 5, section 5.6, corroborated further by corpus examples and demonstrated in Table V.7.

Various changes in actants and circonstants present themselves depending on the way the sentence is constructed, as the following examples show (m.a. = mandatory actant; o.c. = optional circonstant):

1. The scientist (m.a.) field-tested the vaccine (m.a.).

2. The scientist (m.a.) field-tested the vaccine (m.a.) in cattle (o.c.)

2. (a) The scientist (m.a.) field-tested [the vaccine in cattle (m.a.)].

3. The vaccine (m.a.) was field-tested.

4. The vaccine (m.a.) was field-tested in cattle (o.c.).

5. The vaccine (m.a.) was field-tested by the scientist (o.c.).

Chapter 4: 197
6. The vaccine (m.a.) was field-tested in cattle by the scientist (o.c.).

6. (a) The vaccine (m.a.) was field-tested by the scientist in cattle (o.c.).

Nominalisation of the verb *field-test* results in:

1. The field-testing of the vaccine (m.a.) in cattle (o.c.) by the scientist (o.c.).

However, ellipsis from a previous sentence might result in:

1. The field-testing by the scientist (m.a.).

2. The field-testing of the vaccine (m.a.).

3. The field-testing in cattle (m.a.).

Conversely,

1. Vaccine-testing was carried out in cattle (o.c.).

2. Vaccine-testing was carried out by the scientist (o.c.).

3. Vaccine-testing was carried out [by the scientist in cattle (o.c.)].

4. Vaccine-testing was carried out in the field (o.c.).

Differentials with regard to thematisation and orientation are clearly evident.
4.7 Combinants and valency: the psychological subject

Having drawn an analogy between terminological/encyclopaedic structure and valency, I am now considering optional circonstants which may be 'missing', psychological subjects. It could be postulated that a simple sentence in English and other European languages will be divided intuitively by native speakers into noun phrase (NP) on the one hand + verb and the rest of the sentence on the other, as in the type of phrase-structure grammar (PSG) propounded by Chomsky (1957). There may, however, be differences in a person's perception of and familiarity with (a) a grammatical subject and object, (b) a logical subject and object (representing the semantics of a clause, i.e. who does what etc.) and (c) a psychological subject and object where one of these might be hidden or understood. Such psychological subjects can outlive the sentence 'carrying' them; they are supra-sentential, providing a background, a meta-concept. Halliday (1985: 33) expands the point of the thematic psychological subject and in addition differentiates between the logical subject and the grammatical subject, the latter referring to a subject plus its predicate and the former to the relations between things.

In LSP, however, the aim is to make clear statements rather than leave matters in semi-obscurity, which could result in misunderstanding. Ideally, psychological subjects are to be avoided. This is a matter of functional stylistics; in LSP the aim is to convey facts by communicating information, rather than using language for purely aesthetic purposes. In French scientific reporting, particularly that written in the past, the style tended to be more literary and even rather 'precious' when compared with English scientific reporting; even as recently as thirty years ago, the style persisted:

{....} la Vaccination, grâce à laquelle l'homme se modifie et se met à

Chapter 4: 199
l'abri des contagions, doit être débarrassée de sa magie et pensée comme une thérapeutique médicale complexe. (Dagognet, 1964: 5)

A comparable introduction to vaccination in English is notable for its factual approach, although the possibility of the text being taken from different genres, e.g. popular essay versus technical report, should not be overlooked:

Vaccines have been a major success story for human and animal health in the twentieth century. We have seen smallpox eradicated and the control in the developed countries of nine major human infections: diphtheria, measles, mumps, poliomyelitis, rubella, tetanus, tuberculosis, whooping cough and yellow fever. (Corpus example, 1994)

Moreover, in English scientific reporting, scientists traditionally do not figure in the text in which they are reporting their conclusions. However, they occasionally refer to their 'lab.' to distinguish their work from that of another, but it is what they have discovered that assumes prime importance. They therefore assume the 'external' role of 'circonstants' or psychological subjects and would be represented as optional valency slots. An example from French, protéger (qn.) contre le paludisme (avec un vaccin) also corroborates the necessity for optional valency slots. It is therefore particularly important to keep in mind the notion of psychological subjects when constructing a verb frame.

From the large biological sciences English corpus, a small sample of highly subject-specific verbs has been identified 'visually' from the large corpus, based on the author's twenty years' experience in editing scientific texts in these fields and subsequent corroboration by subject specialists (Table
IV.4). The verbs were chosen at random, the aim being to make a comparison with a selection of some highly restricted French verbs from the large French corpus. The frequency of the majority of English verbs usually exceeds by far that of the LSPs occurring in parallel French texts due largely, as stated, to the greater use of nominalisation in French. However, in Table IV.5, French verbs have been chosen which have a higher frequency than their English counterpart and this can be accounted for by the variation in textual content which in the French corpus contains slightly more medical material than the English.

<table>
<thead>
<tr>
<th>En. freq.</th>
<th>Fr. freq.</th>
<th>English type</th>
<th>En. freq.</th>
<th>Fr. freq.</th>
<th>English type</th>
</tr>
</thead>
<tbody>
<tr>
<td>91</td>
<td>6</td>
<td>bind</td>
<td>7</td>
<td>1</td>
<td>immunised</td>
</tr>
<tr>
<td>34</td>
<td>10</td>
<td>binds</td>
<td>20</td>
<td>2</td>
<td>inactivate</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>clone</td>
<td>3</td>
<td>2</td>
<td>inoculate</td>
</tr>
<tr>
<td>54</td>
<td>3</td>
<td>cloned</td>
<td>8</td>
<td>1</td>
<td>insert (s'insérer)</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>denature</td>
<td>6</td>
<td>1</td>
<td>propagate</td>
</tr>
<tr>
<td>2</td>
<td>-</td>
<td>denatured</td>
<td>5</td>
<td>1</td>
<td>secrete</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>denatures</td>
<td>12</td>
<td>-</td>
<td>secreted</td>
</tr>
<tr>
<td>7</td>
<td>-</td>
<td>immunise</td>
<td>29</td>
<td>1</td>
<td>sequenced</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>9</td>
<td>6</td>
<td>vaccinate</td>
</tr>
</tbody>
</table>

Table IV.4 Frequency of some highly specialised English verbs in the biological sciences. The number of occurrences of each type is shown and the verbs are reproduced in the form in which they appear in the corpus; the -ed suffix is only included when it acts as a past participle but not when it is used adjectivally.
Table IV.5 Frequency of some highly specialised French verbs in the biological sciences. The number of occurrences of each type is shown and the verbs are reproduced in the form in which they appear in the corpus. It is notable that many French verbs appear in the reflexive form.

It was originally postulated that in French LSPs, verbs would be desemanticized more than in English because of the greater tendency towards nominalisation in French, so that the meaning is concentrated in the noun, resulting in a greater number of 'carrier' verbs. In the biological sciences, it is noticeable that the total of highly specialised French verbs pertaining to the subject field is considerably lower than the number of English LSP verbs (Tables IV.12, 13), as is the frequency at the highest level of occurrence. Given that the texts comprising the corpora are recent and as far as possible analogous, a possible reason for the low number of LSP verbs is the originality of the subjects being treated, because highly innovative research in these
subject fields is invariably reported in English in the first instance, to gain scientific currency and credibility (the Annales of the Institut Pasteur in Paris have for the past few years been published in English as well as in French). It is therefore possible that corresponding verbs do not yet exist in French, a tendency I have noted with nouns (Thomas, 1983: 85) and this, in conjunction with the preference for nominalisation, could account for the lower occurrence of LSP verbs in this language.

However, in some compensation for the above apparent deficit, carrier verbs such as faire are prominent, as the following Tables IV.6 (a), (b), and (c) show. The frequently occurring instances of the phrases faire + the propositions l’objet, partie and appel were analysed because of their high frequency, using the Synoptic Profile facility of the Aston Text Analyser (ATA) which provides span positions and which is described in detail in Chapter 5. The tables include different grammatical parts of faire. The phrases were drawn from the corpus comprising half a million words of French texts pertaining to the biological sciences. Because fait is a verb/noun homograph, it is treated separately in Table IV.7.
Table IV.6 (a) Frequency of faire l'objet

The phrase represents c. 10% of all occurrences of the verb in the corpus, a high proportion.

<table>
<thead>
<tr>
<th>Freq.</th>
<th>% %</th>
<th>Type</th>
<th>Freq. of l'objet in span +1</th>
<th>Freq. of l'objet in span +2</th>
</tr>
</thead>
<tbody>
<tr>
<td>420</td>
<td>8.2</td>
<td>faire</td>
<td>44</td>
<td>0</td>
</tr>
<tr>
<td>104</td>
<td>2.0</td>
<td>font</td>
<td>10</td>
<td>7</td>
</tr>
<tr>
<td>77</td>
<td>1.5</td>
<td>faisant</td>
<td>7</td>
<td>0</td>
</tr>
</tbody>
</table>

Table IV.6 (b) Frequency of faire partie

<table>
<thead>
<tr>
<th>Freq.</th>
<th>% %</th>
<th>Type</th>
<th>Freq. of partie in span +1</th>
<th>Freq. of partie in span +2</th>
</tr>
</thead>
<tbody>
<tr>
<td>420</td>
<td>8.2</td>
<td>faire</td>
<td>13</td>
<td>0</td>
</tr>
<tr>
<td>104</td>
<td>2.0</td>
<td>font</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>77</td>
<td>1.5</td>
<td>faisant</td>
<td>3</td>
<td>1</td>
</tr>
</tbody>
</table>
Table IV.7 shows how a verb/noun homograph may be distinguished in a corpus by analysing the words preceding the 'type', in this case *fait*, at span positions -2 and -1. A verb may be preceded by a negative, an auxiliary verb, a reflexive pronoun or a relative pronoun. Equally valuable is an assessment of the words in span position type +1. The total frequency for *fait* is 573 which represents 11.1%% of the total corpus. The distinguishing of verb/noun homographs is discussed more fully in Chapter 5.

Another 'carrier' or 'support' verb which occurs frequently in French is *mettre*. As might be expected, the noun/verb homograph *mise* appears more frequently as a noun, as Table IV.8 shows. The verb pointers in span position -1 have a very low frequency.
<table>
<thead>
<tr>
<th>Freq. in span -2</th>
<th>Freq. in span -1</th>
<th>Type fait</th>
<th>Freq. in span +1</th>
<th>Freq. in span +2</th>
</tr>
</thead>
<tbody>
<tr>
<td>18 qui</td>
<td>48 a</td>
<td>573 fait</td>
<td>47 l'objet</td>
<td>37 de</td>
</tr>
<tr>
<td>11 a</td>
<td>48 ont</td>
<td></td>
<td>36 de</td>
<td>24 d'une</td>
</tr>
<tr>
<td>8 n'a</td>
<td>25 se</td>
<td></td>
<td>10 appel</td>
<td>12 des</td>
</tr>
<tr>
<td>6 n'ont</td>
<td>17 ne</td>
<td></td>
<td>10 pas</td>
<td></td>
</tr>
<tr>
<td>14 avait</td>
<td></td>
<td></td>
<td>8 partie</td>
<td></td>
</tr>
<tr>
<td>13 qui</td>
<td></td>
<td></td>
<td>6 part</td>
<td></td>
</tr>
<tr>
<td>9 est</td>
<td></td>
<td></td>
<td>4 guère</td>
<td></td>
</tr>
<tr>
<td>8 l'a</td>
<td></td>
<td></td>
<td>4 qu'il</td>
<td></td>
</tr>
<tr>
<td>8 pas</td>
<td></td>
<td></td>
<td>3 aucun</td>
<td></td>
</tr>
<tr>
<td>7 avoir</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6 ayant</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6 rien</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5 elle</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5 jamais</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4 il</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4 s'est</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table IV.7 Analysis of the verb/noun homograph *fait*. Differentiation has been made by assessing the syntactic characteristics of verbs at span positions -2, -1 and +1.
Table IV.8 Analysis of the noun/verb homograph *mise* showing how an assessment of the span positions post- and pre-ceding the type make possible the automatic recognition of the different grammatical parts of speech.

4.8 Disjuncture

So far combinants have been considered in a fixed form, but the feature that distinguishes phrases from idioms, and makes them a difficult but interesting topic to study, is the flexibility of their construction which can admit disjuncture. The following are examples of combinants which will permit disjuncture (underlined in the examples), although the number and type of
disjuncts is semantically limited by the subject field, as well as being
syntactically limited:

- déceler l'erreur commise/présumée;
- avoir pour effet d'entraîner un avantage injustifié sur [....]/
  d'entraîner un tel avantage

The above adjectives demonstrate the fine line between being 'free',
when they would be considered as disjunctive or adjunctive, but yet are
undeniably influenced by the subject field in which they appear. Were they to
form a term with the noun accompanying them, they would not be considered as
disjuncts. The area of 'freedom' of adjectives and the application of emphasis in
special languages requires further research.

4.8.1 Syntactic disjuncture

The above examples of disjuncture are all adjectives, limited in each
example to two; numbers could substitute for numbers tel. Disjuncts may also
be adjectival phrases, adverbs or adverbial phrases, or prepositional phrases.
A lexical phrase may be a complement with one verb and an adjunct with another
verb, e.g. to live in London; to look for him in London. Provision needs to be
made for a defined number of adjuncts or disjuncts when constructing, for
example, an LSP verb phrase frame with slots and fillers.

Benson et al. (1986: xxiv) mention prepositions but their
comments relate particularly to English where it is often feasible for
prepositions to be omitted, whereas in Romance languages, for example, their
omission may result in the wrong preposition being re-admitted when the term
is written in full, with possible repercussions for safety. For example, the
abbreviated form *Armat. atelier préfa.*, which omits any preposition, might result in *atelier préfabriqué pour armatures* or *atelier pour armatures préfabriquées* or yet again *armatures pour atelier préfabriqué* (Thomas and Judge, 1989: 413). Another point to consider is whether LSP terms in French containing prepositions are fixed and therefore de-grammaticalised; in other words, are these terms sufficiently flexible to permit internal disjunction, e.g. *mise en renverse (?lente) de l’hélice* and would such an interposition change the terminological status of the term? It would appear that the example given, with the inclusion of the adverbial use of *lente*, indicates a temporal terminological relationship, resulting in a different concept. The number of such interposed occurrences in syntagmatic structures could be compared against the rate of the basic terminological unit, to assess the type and frequency of disjunction which in turn could give an intimation of the stage at which a future term shows signs of pre-terminological status. In French and other Romance languages, the preponderance of prepositions in compound terms is important collateral information (cf. Auger's interpretation in the previous chapter), whereas in English the tendency is to juxtapose lexical units (concepts), while German favours concatenation to provide a different concept.

4.8.2 Semantic disjuncture

Furthermore, the differences may be semantic rather than syntactic, and this is a far more difficult problem to address. A study of the translations in English, French and Swedish in the Treaty of Rome has been made by the author (1993, unpublished observations), and the following examples are taken from the Treaty. In the phrase:

*the institution whose failure to act has been declared contrary to this Treaty,*
the French translation is given as:

*l'institution dont l'abstention a été déclarée contraire au présent Traité*

and in Swedish as:

*den institution vars underlåtenhet att handla förklarats stride mot detta fördrag,*

where *l'abstention* and *underlåtenhet* do not necessarily mean *failure to act* because both Swedish and French terms imply that a conscious decision has been taken, whereas *failure to act* does not and in fact has a rather negative aspect.

The work of Vinay and Darbelnet (1969) explains many instances of differences in extension (the 's' is deliberate) between languages, different styles, everyday versus technical words, figurative and concrete terms, and 'lacunae' occurring in one or many language(s), e.g. cricket terms which are peculiar to the British English of the U.K. and some former and current Commonwealth countries.

It is probably impossible - certainly it is beyond the remit of this thesis - to envisage portraying semantic disjuncts in the form of conceptual innuendoes in a frame; for practical purposes, provision for syntactic patterns and the placing of disjuncts provides a workable solution.

4.9 **Summary of the LSP status of verbs and verb phrases**

The way in which a verb acquires LSP status clearly depends largely
on the autonomy of the subject field. Restricted subjects have a number of verbs which only occur in that subject field. The more LSP subject fields overlap, the greater will be the use of LGP verbs which combine with an LSP noun to form a verb + noun phrase pertaining to those subject fields, e.g. to *stay proceedings*. In such fields, the verb will in all probability have a different meaning in LGP from that which it has if it occurs in LSP (cf. the example of *entrainer*). *Picht* (1987: 151-152) poses the question of which factors effect semantic change: is it (a) context; (b) subject field; or (c) (Mitspieler) valency? From the examples given it appears that it is the subject field which has the greatest influence on semantic change; context can vary within a subject field (cf. the example of *nucleus* in Chapter 1 of this thesis (Thomas, 1988: 4), while valency analyses have shown variation according to the subject field in terms of number and in the type of actant which, as mentioned, in legal terminology is almost entirely abstract. The degree to which verbs acquire LSP status is therefore determined by the propositions with which they combine and which are determined by their subject field.

Since verbs are seen to play a key role in several forms of LSP phraseology and have, as stated earlier, been rather neglected in terminology which is strongly noun-based, it is interesting to assess their value in the LSP phraseology being studied and in particular the way they influence and are influenced by nouns. As has been shown, the relationships formed within a verb phrase differ from those hitherto associated with terminology, i.e. the hierarchical, meronymic relations between concepts which are manifest as nouns and adjectives. To re-iterate, verbs and verb phrases denote such attributes as action, function, or process; alternatively they may be stative. It has also been noted that disjuncture is provided by adjectival, adverbial and
prepositional phrases. Another question which this chapter has sought to address is how to limit an LSP verb phrase for the purposes of building a frame, and these elements are brought together by a synthesis of collocation, valency, semantic, syntactic and contrastive features and disjuncts. The following examples seek to amalgamate these aspects and show certain key factors which have emerged and which can be used in the construction of the framework, thus pointing to the orientation of the combinant.

4.9.1 Different verb + same noun object.

The examples underlined below, taken from the 'legal'/pharmaceutical corpus, show different verbs from LGP which are transitive and denote action. They combine in each case with the same noun. The number of nouns which can fulfil this combinatorial role is semantically restricted according to the subject field; likewise, if the object is referred to by different verbs, it is interesting to postulate which verb would be considered prototypical:

- conduire à (/procurer) un avantage injustifié;
- avoir identifié/constaté un prétendu avantage injustifié/non justifié;
- rejeter/refuser le recours;
- exécuter/rédiger une ordonnance
- to produce/manufacture a vaccine lot
- to cause/lead to dehydration

4.9.2 Same verb + different noun object.

More commonly found are verbs which come apparently from LGP but which have restricted combinability with their objects in LSP. The large number of examples reinforces the notion that the verbs are to a certain degree
restricted semantically in these subject fields (the quasi-LSP verbs) because the same verb is seen to combine with a restricted type of proposition to form different LSP phrases, although each proposition may be one of a large group; cf. the examples of *genes*, *enzymes* etc. in Chapter 5. The second type of verb comprises those which act as carriers for their combinant noun or noun phrase. An attempt has been made to classify the verbs into LGP and/or LSP in English and French and it is clear how few subject-specific LSP verbs exist, pointing to differentials in the role of verbs between LSP and LGP. This point has been expanded in Chapter 5. Table IV.9 gives French verb + noun phrases from the corpus of law texts, covering commercial recompense.

In summary, most of the verbs in the examples above are transitive and denote action rather than function; the intransitive verb *procéder* also denotes action. Those verbs which differ belong to commercial terminology and indicate 'state', e.g. *porter une marque* (Table IV.11). The verbs *to bear* and *témoigner* denote 'function', while interestingly, *infringe* and *enfreindre* show 'state' in the past and present, and 'action' in the future; these factors would need to be incorporated in any framework devised for LSP verbs. The combinability of the verbs with a limited number of propositions indicates that they have a restricted collocatory role in the subject field.

The same categorisation has been made for English verbs in Table IV.10.
<table>
<thead>
<tr>
<th>LGP</th>
<th>LSP</th>
</tr>
</thead>
<tbody>
<tr>
<td>engager</td>
<td>une action en recouvrement des [....]</td>
</tr>
<tr>
<td></td>
<td>la procédure de [....]</td>
</tr>
<tr>
<td></td>
<td>l'encontre de chacun [....]</td>
</tr>
<tr>
<td>entraîner</td>
<td>un tel avantage</td>
</tr>
<tr>
<td></td>
<td>l'obligation</td>
</tr>
<tr>
<td>porter</td>
<td>au principe de sécurité</td>
</tr>
<tr>
<td>atteinte</td>
<td>à l'honneur de qu.</td>
</tr>
<tr>
<td></td>
<td>aux intérêts essentiels</td>
</tr>
<tr>
<td>constituer</td>
<td>un groupement d'intérêts</td>
</tr>
<tr>
<td></td>
<td>un avantage injustifié/nonjustifié/</td>
</tr>
<tr>
<td></td>
<td>légitime</td>
</tr>
<tr>
<td></td>
<td>des impositions</td>
</tr>
<tr>
<td></td>
<td>la base des pratiques</td>
</tr>
<tr>
<td>étudier</td>
<td>les MCM applicables</td>
</tr>
<tr>
<td></td>
<td>la perception/paiement de montants compensoir</td>
</tr>
<tr>
<td>Instituer</td>
<td>un lien, un régime, une procédure</td>
</tr>
<tr>
<td>imposer</td>
<td>des droits, des charges</td>
</tr>
<tr>
<td>introduire</td>
<td>les deux présents [....], des déclarations</td>
</tr>
<tr>
<td>payer</td>
<td>des intérêts, de (tels) droits</td>
</tr>
<tr>
<td>procéder</td>
<td>au recouvrement a posteriori de</td>
</tr>
<tr>
<td></td>
<td>montants compensatoires des droits</td>
</tr>
<tr>
<td></td>
<td>non perçus</td>
</tr>
</tbody>
</table>

Table IV.9 Verb + noun phrases in a restricted French LSP (law).

Note the paucity of LSP verbs.
<table>
<thead>
<tr>
<th>LGP</th>
<th>LSP</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>to stay</em> (trans.)</td>
<td><em>proceedings</em></td>
</tr>
<tr>
<td><em>to take</em></td>
<td><em>disciplinary action (against)</em></td>
</tr>
<tr>
<td><em>to constitute</em> (in the sense of create)</td>
<td><em>rules/measures</em></td>
</tr>
<tr>
<td><em>this constitutes</em> (i.e. involves)</td>
<td><em>misconduct</em></td>
</tr>
<tr>
<td></td>
<td><em>restrictions on imports</em></td>
</tr>
<tr>
<td></td>
<td><em>a restriction on freedom</em></td>
</tr>
<tr>
<td></td>
<td><em>a justification of the rules</em></td>
</tr>
<tr>
<td></td>
<td><em>a barrier to trade/a trade barrier</em></td>
</tr>
<tr>
<td></td>
<td><em>a breach of the peace</em></td>
</tr>
<tr>
<td></td>
<td><em>a serious threat to public health</em></td>
</tr>
<tr>
<td><em>to justify</em></td>
<td><em>a rule/a decision/restrictions</em></td>
</tr>
<tr>
<td><em>to impose</em></td>
<td><em>restrictions on trade</em></td>
</tr>
<tr>
<td><em>to impede</em></td>
<td><em>trade</em></td>
</tr>
<tr>
<td><em>to hinder</em></td>
<td><em>trade</em></td>
</tr>
<tr>
<td><em>to grant</em></td>
<td><em>a licence/permission</em></td>
</tr>
<tr>
<td></td>
<td><em>to prescribe</em> a drug/a medicinal product*</td>
</tr>
<tr>
<td></td>
<td><em>to dispense</em> medicine/a (medicinal) product*</td>
</tr>
<tr>
<td></td>
<td><em>a therapeutically equivalent product</em></td>
</tr>
</tbody>
</table>

*Table IV.10  English verbs in the LSPs of law and pharmaceuticals*

All the above verbs are transitive and all denote action, not function. Verbs collocating in the legal field in English are almost entirely LGP verbs, and
all verbs forming part of a phrase in this field have abstract noun propositions. The last three examples quoted have 'concrete' noun propositions but are from the LSPs of commerce and pharmaceuticals. Moreover, two of these last three examples also show LSP verbs.

In Table IV.11, a few contrastive phrases in English and French serve to show a tendency, in the corpora studied, for more true LSP 'legal' verbs to exist in French than in English. In the large English and French corpora of the biological sciences, the number of verbs closely identified with the subject field was >110 in English (Table IV.12), whereas in French there were between 80 and 90 (Table IV.13). These figures are in sharp contrast to the specialised verbs found in legal terminology. Where more than one suffix appears alongside the same stem, each form occurs the same number of times. Alternative spellings (e.g. -ise, -ize) are listed separately. In Chapter 5, more statistical data are given.

4.9.3 Conclusions for comparison of LSP verbs in English and French

The conclusions to be drawn from the above examples are that verbs in English forming part of an LSP phrase derive mostly from LGP, whereas in French legal terminology, for example, there is a tendency for more LSP verbs to occur, although this does not necessarily follow with other LSPs in French. However, the underlying propositions, or concepts, are the same in both languages and it is therefore at the surface level that these French verbs have a higher degree of 'LSP-ness' than in English; on the linguistic surface the English verbs appear to belong to LGP but the basic concept does not alter. Differences are to be found at the semantic level where verbs appear in both LGP and LSP; in the examples given, the LGP meaning of these verbs is far more concrete than the abstract LSP meaning. Object nouns are abstract, and
different abstract object nouns acting as propositions to the verbs are fairly restricted, numbering between three and five. This low restriction means that it is possible to predict the number of propositions and to construct frames for these verbs on the 'slot and filler' principle, while at the same time allowing for disjuncture in phrases.

<table>
<thead>
<tr>
<th>LGP</th>
<th>LSP</th>
</tr>
</thead>
<tbody>
<tr>
<td>to appeal</td>
<td>to the Court of Appeal</td>
</tr>
<tr>
<td>saisir</td>
<td>la Court of Appeal (sic.)</td>
</tr>
<tr>
<td>to bear</td>
<td>witness</td>
</tr>
<tr>
<td>témoigner (de)</td>
<td></td>
</tr>
<tr>
<td>to bear</td>
<td>a brand name</td>
</tr>
<tr>
<td>porter</td>
<td>une marque</td>
</tr>
<tr>
<td>bearing</td>
<td>a trade mark</td>
</tr>
<tr>
<td>revêtu</td>
<td>d'une marque</td>
</tr>
<tr>
<td>to adopt</td>
<td>rules of ethics</td>
</tr>
<tr>
<td>to lay down</td>
<td>rules of ethics</td>
</tr>
<tr>
<td>adopter</td>
<td>les règles de déontologie</td>
</tr>
<tr>
<td>édicter</td>
<td>les règles de déontologie</td>
</tr>
<tr>
<td>to infringe</td>
<td>the law</td>
</tr>
<tr>
<td>enfreindre</td>
<td>la loi</td>
</tr>
<tr>
<td>to stay</td>
<td>the proceedings</td>
</tr>
<tr>
<td>surseoir</td>
<td>à un jugement</td>
</tr>
<tr>
<td>to pronounce</td>
<td>judgment</td>
</tr>
<tr>
<td>statuer</td>
<td>un jugement</td>
</tr>
<tr>
<td>to impose</td>
<td>(disciplinary) sanctions</td>
</tr>
<tr>
<td>prononcer</td>
<td>des sanctions disciplinaires</td>
</tr>
</tbody>
</table>

Table IV.11 Contrast in LSP and LGP verbs in English and French
<table>
<thead>
<tr>
<th>&gt;200</th>
<th>60-41</th>
<th>40-21</th>
<th>20-14</th>
<th>13-6</th>
<th>5-1</th>
</tr>
</thead>
<tbody>
<tr>
<td>infect</td>
<td>281</td>
<td>58</td>
<td>36</td>
<td>20</td>
<td>13</td>
</tr>
<tr>
<td>attenuated</td>
<td>248</td>
<td>57</td>
<td>36</td>
<td>20</td>
<td>13</td>
</tr>
<tr>
<td>infected</td>
<td>231</td>
<td>57</td>
<td>35</td>
<td>20</td>
<td>13</td>
</tr>
<tr>
<td>mediated</td>
<td>56</td>
<td>34</td>
<td>20</td>
<td>13</td>
<td>13</td>
</tr>
<tr>
<td>confered</td>
<td>52</td>
<td>33</td>
<td>18</td>
<td>12</td>
<td>12</td>
</tr>
<tr>
<td>vaccinated</td>
<td>33</td>
<td>18</td>
<td>12</td>
<td>12</td>
<td>12</td>
</tr>
<tr>
<td>integrated</td>
<td>27</td>
<td>17</td>
<td>10</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>activated</td>
<td>27</td>
<td>17</td>
<td>10</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>classified</td>
<td>26</td>
<td>17</td>
<td>10</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>code</td>
<td>25</td>
<td>15</td>
<td>9</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>labelled</td>
<td>25</td>
<td>15</td>
<td>9</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>package</td>
<td>24</td>
<td>15</td>
<td>9</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>incubation</td>
<td>22</td>
<td>15</td>
<td>9</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>phosphorylated</td>
<td>22</td>
<td>14</td>
<td>8</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>cleaves</td>
<td>21</td>
<td>14</td>
<td>8</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>occluded</td>
<td>21</td>
<td>14</td>
<td>8</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>adsorb</td>
<td>13</td>
<td></td>
<td></td>
<td></td>
<td>5</td>
</tr>
<tr>
<td>activate</td>
<td>20</td>
<td></td>
<td></td>
<td></td>
<td>5</td>
</tr>
<tr>
<td>mediate</td>
<td>20</td>
<td></td>
<td></td>
<td></td>
<td>5</td>
</tr>
<tr>
<td>yield</td>
<td>20</td>
<td></td>
<td></td>
<td></td>
<td>5</td>
</tr>
<tr>
<td>passed</td>
<td>18</td>
<td></td>
<td></td>
<td></td>
<td>5</td>
</tr>
<tr>
<td>confered</td>
<td>18</td>
<td></td>
<td></td>
<td></td>
<td>5</td>
</tr>
<tr>
<td>screened</td>
<td>18</td>
<td></td>
<td></td>
<td></td>
<td>5</td>
</tr>
<tr>
<td>react</td>
<td>17</td>
<td></td>
<td></td>
<td></td>
<td>5</td>
</tr>
<tr>
<td>synthesised</td>
<td>17</td>
<td></td>
<td></td>
<td></td>
<td>5</td>
</tr>
<tr>
<td>contaminated</td>
<td>16</td>
<td></td>
<td></td>
<td></td>
<td>5</td>
</tr>
<tr>
<td>initiate</td>
<td>16</td>
<td></td>
<td></td>
<td></td>
<td>5</td>
</tr>
<tr>
<td>labeled</td>
<td>16</td>
<td></td>
<td></td>
<td></td>
<td>5</td>
</tr>
<tr>
<td>mapped</td>
<td>16</td>
<td></td>
<td></td>
<td></td>
<td>5</td>
</tr>
<tr>
<td>digests</td>
<td>15</td>
<td></td>
<td></td>
<td></td>
<td>5</td>
</tr>
<tr>
<td>disrupted</td>
<td>15</td>
<td></td>
<td></td>
<td></td>
<td>5</td>
</tr>
<tr>
<td>disseminated</td>
<td>15</td>
<td></td>
<td></td>
<td></td>
<td>5</td>
</tr>
<tr>
<td>enhanced</td>
<td>15</td>
<td></td>
<td></td>
<td></td>
<td>5</td>
</tr>
<tr>
<td>evolved</td>
<td>15</td>
<td></td>
<td></td>
<td></td>
<td>5</td>
</tr>
<tr>
<td>synthesised</td>
<td>15</td>
<td></td>
<td></td>
<td></td>
<td>5</td>
</tr>
<tr>
<td>translated</td>
<td>15</td>
<td></td>
<td></td>
<td></td>
<td>5</td>
</tr>
<tr>
<td>encapsulated</td>
<td>14</td>
<td></td>
<td></td>
<td></td>
<td>5</td>
</tr>
<tr>
<td>hybridised</td>
<td>14</td>
<td></td>
<td></td>
<td></td>
<td>5</td>
</tr>
<tr>
<td>segmented</td>
<td>14</td>
<td></td>
<td></td>
<td></td>
<td>5</td>
</tr>
<tr>
<td>cleave</td>
<td>14</td>
<td></td>
<td></td>
<td></td>
<td>5</td>
</tr>
<tr>
<td>synthesis</td>
<td>14</td>
<td></td>
<td></td>
<td></td>
<td>5</td>
</tr>
<tr>
<td>co-inoculated</td>
<td>14</td>
<td></td>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>co-transcribed</td>
<td>14</td>
<td></td>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>cross-reacts</td>
<td>14</td>
<td></td>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>denature(s)</td>
<td>14</td>
<td></td>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>electroporated</td>
<td>14</td>
<td></td>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>force</td>
<td>14</td>
<td></td>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>homogenised</td>
<td>14</td>
<td></td>
<td></td>
<td></td>
<td>2</td>
</tr>
</tbody>
</table>

Table IV.12 Frequency ranges of LSP English verbs from the biological sciences (total number of types = 111)
<table>
<thead>
<tr>
<th>&gt;81</th>
<th>40-21</th>
<th>20-9</th>
<th>8-3</th>
<th>2</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>95 contaminé</td>
<td>40 lié(e/s/s)</td>
<td>20 combattre</td>
<td>8 dosés</td>
<td>2 atténuées</td>
<td>1 activer</td>
</tr>
<tr>
<td>85 atteints</td>
<td>39 contaminée</td>
<td>20 diagnostiquer</td>
<td>8 induire</td>
<td>2 clonés</td>
<td>1 apparenté</td>
</tr>
<tr>
<td></td>
<td>39 provoquer</td>
<td>19 hydrolysé(e/s)</td>
<td>8 prescrits</td>
<td>2 coloniser</td>
<td>1 caractérisé(e/s)</td>
</tr>
<tr>
<td>60-41</td>
<td>38 dérivés</td>
<td>19 infecté</td>
<td>7 aménés</td>
<td>2 cultivé</td>
<td>1 contracte</td>
</tr>
<tr>
<td></td>
<td>35 prélevés</td>
<td>18 atteinte</td>
<td>7 greffer</td>
<td>2 dosé</td>
<td>1 décontaminé</td>
</tr>
<tr>
<td></td>
<td>31 dérivée</td>
<td>18 prédire</td>
<td>6 hydrolyse</td>
<td>2 éradiquer</td>
<td>1 dégrade</td>
</tr>
<tr>
<td>54 atteint</td>
<td>30 contrôlées</td>
<td>17 caractérisé</td>
<td>6 infecter</td>
<td>2 excise</td>
<td>1 diagnostiqués</td>
</tr>
<tr>
<td>54 code</td>
<td>29 atteintes</td>
<td>17 transfusés</td>
<td>6 isoler</td>
<td>2 excrêten(t)</td>
<td>1 greffé(es)</td>
</tr>
<tr>
<td>50 lutter</td>
<td>29 soumissions</td>
<td>15 transfusées</td>
<td>6 prescrire</td>
<td>2 hydrolysée</td>
<td>1 hospitalisés</td>
</tr>
<tr>
<td>49 contaminés</td>
<td>29 transmis</td>
<td>15 transmises</td>
<td>6 vacciner</td>
<td>2 inactiver</td>
<td>1 hydrolysée</td>
</tr>
<tr>
<td>43 chauffés</td>
<td>27 extraits</td>
<td>14 atténué</td>
<td>5 marquée</td>
<td>2 infecte</td>
<td>1 inactiv(ait)(é)es</td>
</tr>
<tr>
<td>43 protégés</td>
<td>27 liées</td>
<td>14 caractérisé</td>
<td>4 absorbé(e/er)</td>
<td>2 infesté</td>
<td>1 inhib(ait)(e)</td>
</tr>
<tr>
<td>42 infectées</td>
<td>27 rechercher</td>
<td>14 infectée</td>
<td>4 (se) caractérisé</td>
<td>2 ingéré(e/s)</td>
<td>1 inject</td>
</tr>
<tr>
<td></td>
<td>26 contrôler</td>
<td>14 ingérées</td>
<td>4 marqué</td>
<td>2 inhiber/ait</td>
<td>1 insérée</td>
</tr>
<tr>
<td></td>
<td>26 exposées</td>
<td>13 contaminé</td>
<td>4 métabolisé(e)(s)</td>
<td>2 injecté(e/s)</td>
<td>1 jugulées</td>
</tr>
<tr>
<td>(N.B. There were no LSP verbs between the 80 - 61 range)</td>
<td>25 infectés</td>
<td>13 radiomarqué</td>
<td>4 filtrer</td>
<td>2 insérer</td>
<td>1 (se) métabolisent</td>
</tr>
<tr>
<td></td>
<td>24 absorbé</td>
<td>12 codant (pour)</td>
<td>4 induire</td>
<td>2 irradies</td>
<td>1 oxygénér</td>
</tr>
<tr>
<td></td>
<td>24 prélevé</td>
<td>12 doser</td>
<td>4 synthétiser</td>
<td>2 métabolisé(e)</td>
<td>1 prescritvent</td>
</tr>
<tr>
<td></td>
<td>22 éliminer</td>
<td>12 ingéré(e)</td>
<td>3 aggraver</td>
<td>2 oxydés</td>
<td>1 s'enchaînent</td>
</tr>
<tr>
<td></td>
<td>22 fixé</td>
<td>12 stérisilis</td>
<td>3 atténué</td>
<td>2 procréer</td>
<td>1 s'intoxiquer</td>
</tr>
<tr>
<td></td>
<td>22 transfuse</td>
<td>11 emballé</td>
<td>3 caractérisé</td>
<td>2 protégées</td>
<td>1 se propageant</td>
</tr>
<tr>
<td></td>
<td>21 atteindre</td>
<td>11 imprégnées</td>
<td>3 désinfectés</td>
<td>2 provoquait</td>
<td>1 se réimplanter</td>
</tr>
<tr>
<td></td>
<td>21 transfert</td>
<td>11 infestées</td>
<td>3 échantillonn(e/r)/es</td>
<td>2 réagit</td>
<td>1 séquenc(e/r)(é)s</td>
</tr>
<tr>
<td></td>
<td>10 éradiquée</td>
<td>3 élimine</td>
<td>3 excrétées</td>
<td>2 s'hydrolysent</td>
<td>1 sévisait</td>
</tr>
<tr>
<td></td>
<td>10 excrété</td>
<td>3 excrétées</td>
<td>3 infestent</td>
<td>2 stérilisé(e/s)</td>
<td>1 toléré</td>
</tr>
<tr>
<td></td>
<td>10 isolés</td>
<td>3 infecté</td>
<td>3 testé</td>
<td>2 synthétisée</td>
<td>1 transfusé</td>
</tr>
<tr>
<td></td>
<td>10 prescrit</td>
<td>3 inhibé</td>
<td>3 tolérent</td>
<td>2 uriner</td>
<td>1 transplante</td>
</tr>
<tr>
<td></td>
<td>9 absorbée</td>
<td>3 se lier</td>
<td>3 sévit</td>
<td>2 vaccinés</td>
<td>1 tumore</td>
</tr>
<tr>
<td></td>
<td>9 contaminant</td>
<td>3 (se) transmet(tent)</td>
<td>2 vaccinés</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table IV.13 Frequency ranges of LSP French verbs from the biological sciences (total number of types = 84)
4.10 Criteria for frames for LSP verb phrases

The aim of drawing together all the foregoing aspects and providing relationships has been to create frames which can be applied to certain combinant verb phrases in special subject fields and this has been done in Chapter 5, supported by statistical data. The orientation of the combinants can therefore be deduced.

The method used was to run concordances on the two corpora which comprised texts on virology and bacteriology in English and French. Collocation was invoked principally to elicit combinants of verb + proposition occurring three or more times, in addition to other collocatory patterns. Three was an arbitrary number chosen to ascertain whether it would provide a viable starting point for analysis, which proved to be the case. The collocation patterns in the concordances provided examples of valency, and the number and type of actants were recorded and compared with the degree of speciality of the verb. Disjuncture was also noted and incorporated into the frame.

The corpora have been analysed for verbs (a) occurring only in their subject field, e.g. *vaccinate* (see comparisons of frequency in this chapter and in Chapter 5); (b) those having a specialised use within that field, e.g. *type* and (c) those having a 'carrier' role supporting a proposition pertaining to the subject field. The following factors have been assessed:

(i) syntactic variation and disjuncture;
(ii) semantic variation;
(iii) valency restriction;
(v) whether different tenses provide different relationships.
(vii) the past participle: from passivisation to orientation.

In Chapter 5 the verb frames are assessed for the degree of help they can give in indicating the main theme of this thesis, the orientation of verb phrases.
Chapter 5

Symbiosis of collocation, valency and indexing: a terminological framework for indicating orientation in special language combinants

The approach to constructing a frame for LSP verbs and phrases differs from that of LGP verbs; a verb in LSP does not need to have its semantic field defined because this is implicit in the context of its restricted subject domain. For the same reason, the notion of incorporating lexical functions can for the most part be discarded. In restricted LSPs we can conceive of a closed sphere of knowledge which to some extent is supra to cultural, social and many other elements which, however, need to be taken into consideration when dealing with general language. The unifying feature in LSP is one of communication within a discourse community of specific knowledge which, where it is graphical or of non-linguistic sound, e.g. music, should present no problems to a specialist in the field, given the senses of sight and hearing and, in some cases, touch, for example, in medical examinations. The category of communication which requires systematic analysis within its subject field is language. Because of the restricted nature of special fields, the frames which have been constructed refer to verbs as they occur within their special, 'autonomous' subject field, rather than to any outside experience or knowledge. This is admittedly an idealistic viewpoint but it provides a working base from which a frame can be constructed.
5.1 The corpora and text analyser

The two large corpora that have been compiled, one in English (American and British) and one in French, each contain approximately half a million words. The texts which form the corpora are highly restricted and homogeneous, comprising textbooks and proceedings of symposia on specific subjects within the areas of virology and bacteriology, from leading publishers in these fields and large international organisations. Their primary value is that they are not prototypical, chimaeric examples of texts but are current records of recent research being carried out in these autonomous fields and their discourse may therefore be said to be 'ecotypical' (term used in discussion with Dr. P. Roe and Professor F. Knowles) - i.e. the difference between 'langue' and 'parole(s)'.

Text analyses have been performed using the Aston Text Analyser (ATA) which is currently being developed by Dr. Peter Roe and his colleagues in the Language Studies Unit at Aston University, Birmingham, U.K., in collaboration with commercial partners MS Technology A/S Copenhagen. ATA can analyse very large text files at great speed and handle character sets in most EC languages; the exception in the version I used was Greek accents.

Frequency lists can be compiled and the relative frequency of the 'type' (search word) out of 10,000 is given, expressed as %%. Because one of the main attributes of ATA is to analyse special language texts, a comparison can be made against the COBUILD reference list of English general language. Function words can be omitted if desired. Concordances can be run, showing text of four words either side of the search word, with the possibility of displaying the concordance with either left-hand or right-hand
alphabetisation in relation to the search word (Annex 6). The corroboration of hypotheses through the power of statistics such as these is vital for information retrieval, particularly as the search word is not decontextualised. Later versions of ATA will be able to provide alphabetisation of words in positions other than -1 and +1. In addition to the concordanced lists, ATA can provide lists containing given strings of characters. A highlighted line in the concordance list can be used to invoke a window containing several lines of text, including the highlighted concordance line. Words containing strings can be saved to a file, as can highlighted text showing words in context.

A particularly useful tool is the Synoptic Profile. This shows the search word represented as an asterisk with its number of occurrences. In the four columns to the left and four to the right are words which appear in span positions - 4 to - 1 and + 1 to + 4 respectively, in descending order of frequency. A condensed version of the Synoptic Profile can be seen in Tables V.2 (a), 3 (a) and 4 (a) and Annex 7; in these examples the search word is written in capital letters in the tables and bold type in the annex, rather than being represented by an asterisk. The Synoptic Profile has been of paramount importance to the research done in this thesis by permitting the identification of combinants and in enabling them to be studied in context beyond the sentence level.

5.2 Problems of automatic identification of verbs in an LSP corpus

The study of verbs has been an interesting one. Since the majority of the verbs in the scientific texts which form the corpora appear in the third person of the present, imperfect, conditional and perfect tenses and in the infinitive, it has not proved too difficult to gain an initial overview.
of verbs in the French corpus through a morphological search for those tenses which have the endings -ait, -alent, -er, -(o)ir and re. It is not yet possible to search on accented characters in ATA; in the case of the French past participle -é(é), the search has been on the auxiliary verb where feasible.

No attempt has been made at lemmatization which would incur a linguistic imbalance (Sinclair, 1991:8). This is particularly true in LSP with its restricted use of tenses, a fact which has implications for teaching and technical writing. It will have been noted that the examples in Tables IV.12 and IV.13 have not been lemmatized.

However, I have found that searching for verbs in the English corpus has required greater ingenuity and is an area which would benefit from further research. The following difficulties were encountered:

1. The lack of morphology in English makes the automatic recognition of verbs difficult.
2. Where inflections occur, the result is a large number of homographs, e.g.
   (i) -ed. The past participle which is often also used adjectivally. Even when a word ending in -ed follows are, it may be an adjective and not the passive voice, i.e. represent stasis rather than action, e.g. Virions are unenveloped and icosahedral. (This point is of course not peculiar to English.)
   (ii) -s can denote the noun plural and the third person present of a verb in the singular; the latter predominates in the texts studied.
   (iii) -ing can denote the present participle and the
gerundive.

(i) -ate can denote noun/verb homographs, e.g. *conjugate*, which is always a noun in the LSP texts.

(ii) adjective/verb homographs, e.g. *live*, which in the texts studied always appears as an adjective.

To solve the problem, it was necessary to adopt a rather heuristic approach and look for parts of speech which are associated solely with verbs. In this respect the Synoptic Profile of ATA proved an invaluable help in searching for:

(i) negatives (*not, never*, but not *hardly, scarcely, barely* which did not occur in the corpus);

(ii) adverbs ending in *-ly*, usually preceding the verb, but not those denoting a certain vagueness, i.e. *possibly, probably*. * Likely* is found more in American English. Care however needs to be taken with *-ly* adverbs preceding a past participle in LSP where the past participle is used adjectivally and the adverb combines with it to form part of a term, e.g. *A recently developed, orally administered vaccine*;

(iii) a subsequent preposition, e.g. *results in, transmitted by/in, conserved among*;

(iv) a preceding relative pronoun (*that/which results in*): * which* does not generally occur as an interrogative adjective in LSP;

(v) a subsequent article, adjective or noun, or a combination of two or more of these;

(vi) some verb/noun homographs such as *causes* where the noun can be distinguished if *causes* is followed by *of*;

(vii) a homograph which, when followed by a full stop, is likely to be a noun;
(viii) a preceding personal pronoun;
(ix) a subsequent conjunction plus a past participle denoting action rather than stasis, although the difference may be difficult to assess, e.g. GV granulins are highly conserved and contain related sequences, and may need corroboration from a subject specialist;
(x) words ending in -ed which are past participles if followed by a present participle, e.g. fragments were generated using....;
(xi) words ending in -ed which are past participles if followed by an article or similar function word, e.g. characterized a/as/the Lactococcus.
(xii) an -ed word at the end of a sentence is invariably, but not by any means exclusively, a past participle.

5.3 Some frequency statistics of LSP verbs

A reasoned 'guesstimate' of verbs of all types with a frequency <7 based on the figures for >8 and >9 indicated that the total percentage of all verbs was 8.24%, a figure which corroborates my finding from a much smaller corpus, as noted in the previous chapter. The first notable feature was the paucity of LSP verbs in both corpora.

5.3.1 LSP verbs in the English corpus

LSP verbs of all frequencies were extracted from the text. Including those verbs which had a considerable frequency difference between LSP and LGP (the quasi-LSP verbs, based on comparison with the COBUILD corpus), the number of different types was 111 and their proportion of the corpus was approximately 0.02% of the total number of words.
5.3.2 Comparison with LSP verbs in the French corpus

As can be seen from Tables IV.12 and IV.13, the low frequency of LSP verbs reinforces the importance of exploring the lower reaches of the corpora to extract verbs from a subject field; in the English corpus, 28 of the LSP verbs identified occurred only once, with 24 in the French corpus. It needs to be borne in mind that these specialised subject fields are 'en plein essor'; research in them is advancing rapidly and they are likely to contain many neologisms. To summarise, Table V.1 shows that the greatest number of LSP verbs in all morphological forms in both English and French occur with low frequency:

<table>
<thead>
<tr>
<th>Frequency</th>
<th>Number of English verbs (all forms)</th>
<th>Number of French verbs (all forms)</th>
</tr>
</thead>
<tbody>
<tr>
<td>&gt;200</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>200 - 101</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>100 - 81</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>80 - 61</td>
<td>9</td>
<td>0</td>
</tr>
<tr>
<td>60 - 41</td>
<td>14</td>
<td>8</td>
</tr>
<tr>
<td>40 - 21</td>
<td>28</td>
<td>23</td>
</tr>
<tr>
<td>20 - 16</td>
<td>22</td>
<td>8</td>
</tr>
<tr>
<td>15 - 11</td>
<td>19</td>
<td>15</td>
</tr>
<tr>
<td>10 - 6</td>
<td>21</td>
<td>18</td>
</tr>
<tr>
<td>5 - 1</td>
<td>140</td>
<td>77</td>
</tr>
</tbody>
</table>

Table V.1 Frequency ranges of highly specialised LSP verbs in the biological sciences in English and French.
Not only do many of these verbs occur with low frequency, but compound verbs also occur in English, particularly in the lower reaches of the corpus. This style tends to be favoured particularly by authors using American English and corroborates the trait already noted to form denominalised verbs. The low frequency can be explained by the novelty of the work presented, since some of these verbs will be neologisms coined by the authors.

5.4 Identification of LSP combinants by collocation (frequency >3)

As stated at the end of Chapter 4, examples of collocatory patterns of verb + proposition occurring more than three times were chosen for analysis. The method used was to scan the Synoptic Profile of ATA for nouns in span positions +1, +2, +3, +4 and occasionally greater. Where the combined total of a given word or 'type' exceeded three, regardless of span position and number (singular or plural), these were extracted for analysis. Because the true LSP verbs had predictable propositions and were few in number, it proved more fruitful for this study to examine verbs designated as having quasi-LSP status.

5.4.1 Quasi-LSP verbs

Examples are now given, firstly for the verb, *encode*, which occurs 35 times in the corpus, i.e. a frequency of 0.77‰. It is noteworthy that it did not occur in the COBUILD reference list. Five combinant propositions occurred more than three times: *enzyme(s), gene(s), pheromone, protein(s) and sugar(s)*, and include the nomenclature

---

1 ‰‰ = frequency >1/10,000

Chapter 5: 228
of individual enzymes, genes and so on, although it should be noted that occurrences of these biochemical entities listed by individual names in the corpus as a whole in some instances is quite high (see later in this section), certainly higher than that of the verb part of the combinant. The need for individual nomenclature to be included in the combinant reinforces the importance stated in Chapter 2 for the terminologist to have an in-depth knowledge of the subject field; in this example there are three indicative factors: (i) the suffix -ase - which indicates an enzyme; (ii) GTF plus a letter which signifies a gene; (iii) prg2 which is a pheromone receptor. Table V.2 (a) shows the span positions of the five propositions with their number of occurrences and Table V.2 (b) gives examples of the combinants. It is interesting that even where encode appears at the end of a sentence, the proposition These phosphorylated sugars appears in the following sentence, referring to phosphorylated sugar in the previous sentence, thus showing that the concepts invoked may link with each other supra-sententially.

The second example analysed, express, has also been designated a quasi-LSP verb after comparison with COBUILD, where the frequency is 0.47%%, in contrast to 1.87%% for LSP occurrences. It should be pointed out that all the instances of express in the LSP corpus are verbs. In Tables V.3 (a) and (b) the analysis of the verb express is shown, with six combinant propositions occurring more than three times.

The results from these examples and those of other verbs studied show that most of the nouns forming part of the propositions appear in span position +3, closely followed by span positions +2 and +4, depending on the noun positions in the multiple lexical units which comprise the term. The verbs have a dynamic quality indicating 'action' and 'process' whereas their noun combinant represents stasis.
<table>
<thead>
<tr>
<th>(35) ENCODE</th>
<th>Span + 1</th>
<th>Span +2</th>
<th>Span +3</th>
<th>Span +4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1) enzymes</td>
<td>(3) proteins</td>
<td>(4) proteins</td>
<td>(2) proteins</td>
</tr>
<tr>
<td></td>
<td>(1) GTF-I</td>
<td>(1) enzyme</td>
<td>(2) sugar</td>
<td>(1) GTF-S</td>
</tr>
<tr>
<td></td>
<td>(1) pheromone</td>
<td>(1) GTF-SI</td>
<td>(1) endonuclease</td>
<td>(1) II mt 1</td>
</tr>
<tr>
<td></td>
<td>(1) protein</td>
<td>(1) PTS</td>
<td>(1) enzyme</td>
<td>(1) permeases</td>
</tr>
<tr>
<td></td>
<td>(1) ScrFI</td>
<td>(1) genes</td>
<td>(1) pheromone</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(1) prgZ</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(1) protein</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(1) sugars</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table V.2 (a)  Quasi-LSP verb encode + collocatory propositions occurring >3 times
Bracketed figures indicate the number of occurrences of each type.
## Corpus examples of *encode* + *enzyme(s)*

| *encode* enzymes for sugar metabolism | (+1) |
| *encode* an enzyme endowed with the third motif | (+2) |
| *encode* the PTS enzyme II mt1 | (+2, +3, +4) |
| *encode* the ScrF1 endonuclease with the third motif | (+2, +3) |
| *encode* protein dependent peptide permeases | (+4) |

## Corpus examples of *encode* + *protein(s)*

| *encode* protein-dependent peptide permeases | (+1) |
| *encode* exported proteins | (+2) |
| *encode* 'killer' proteins | (+2) |
| *encode* surface proteins | (+2) |
| *encode* an IgG-binding protein | (+3) |
| *encode* different surface proteins | (+3) |
| *encode* DNA binding proteins | (+3) |
| *encode* functionally different proteins | (+3) |
| *encode* homologous antirestriction proteins | (+3) |
| *encode* numerous different surface proteins | (+4) |
| *encode* the cytoplasmic ATP-binding proteins | (+4) |

## Corpus examples of *encode* + *genes*

| *encode* GTF-I, GTF-SI, and GTF-S | (+1, +2, +4) |
| *encode* all the genes necessary | (+3) |

## Corpus examples of *encode* + *pheromone*

| *encode* pheromone receptor (prgZ) | (+1, +3) |
| *encode* antibiotic resistance and *pheromone* | (+4) |

## Corpus examples of *encode* + *sugar(s)*

| *encode* enzymes for *sugar* metabolism | (+3) |
| *encode* the various *sugar* transport systems | (+3) |
| *encode* These phosphorylated *sugars* include | (+3) |

---

**Table V.2 (b)** Corpus examples of the quasi-LSP verb *encode* + propositions occurring >3 times

Brackets indicate span positions following the verb.
<table>
<thead>
<tr>
<th>(83) EXPRESS</th>
<th>Span + 1</th>
<th>Span + 2</th>
<th>Span + 3</th>
<th>Span + 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>(2) polysaccharide</td>
<td>(4) sequences</td>
<td>(3) phenotype</td>
<td>(1) genes</td>
<td></td>
</tr>
<tr>
<td>(2) protein</td>
<td>(2) protein</td>
<td>(2) gene</td>
<td>(1) polysaccharide</td>
<td></td>
</tr>
<tr>
<td>(1) gene</td>
<td>(2) proteins</td>
<td>(1) cell</td>
<td>(1) protein</td>
<td></td>
</tr>
<tr>
<td>(1) cells</td>
<td>(1) cells</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(1) genes</td>
<td>(1) genes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(1) genes</td>
<td>(1) polysaccharides</td>
<td>(1) protein</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(1) proteins</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(1) sequences</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table V.3 (a)  
Quasi-LSP verb express + collocatory propositions occurring >3 times  
Bracketed figures indicate the number of occurrences of each type.
<table>
<thead>
<tr>
<th>Corpus examples of <em>express + cell(s)</em></th>
<th>Corpus examples of <em>express + protein(s)</em></th>
</tr>
</thead>
<tbody>
<tr>
<td><em>express</em> packaging <em>cells</em>. A major</td>
<td><em>express</em> protein <em>F</em> even under</td>
</tr>
<tr>
<td><em>express</em> inefficiently in <em>cells</em> in</td>
<td><em>express</em> protein <em>F</em> normally (compare ....)</td>
</tr>
<tr>
<td><em>express</em> on their <em>cell</em> surface</td>
<td><em>express</em> either <em>protein</em> <em>F</em> or</td>
</tr>
<tr>
<td>(†2)</td>
<td>(†1)</td>
</tr>
<tr>
<td>(†3)</td>
<td>(†1)</td>
</tr>
<tr>
<td>(†3)</td>
<td>(†2)</td>
</tr>
<tr>
<td>(†3)</td>
<td>(†2)</td>
</tr>
<tr>
<td><em>express</em> heterologous <em>protein</em>-specifying genes</td>
<td><em>express</em> heterologous <em>protein</em>-specifying genes for</td>
</tr>
<tr>
<td>(†2)</td>
<td>(†2)</td>
</tr>
<tr>
<td>(†3)</td>
<td>(†2)</td>
</tr>
<tr>
<td>(†3)</td>
<td>(†2)</td>
</tr>
<tr>
<td><em>express</em> the IFN- <em>gene</em>.</td>
<td><em>express</em> these <em>proteins</em> in <em>L.</em></td>
</tr>
<tr>
<td>(†3)</td>
<td>(†2)</td>
</tr>
<tr>
<td><em>express</em> the reporter <em>gene</em>.</td>
<td><em>express</em> multiple M-like <em>proteins</em>:</td>
</tr>
<tr>
<td>(†3)</td>
<td>(†3)</td>
</tr>
<tr>
<td><em>express</em> viral and recombinant <em>genes</em>.</td>
<td><em>express</em> the <em>gE</em> <em>protein</em>, and</td>
</tr>
<tr>
<td>(†4)</td>
<td>(†3)</td>
</tr>
<tr>
<td><em>express</em> <em>polysaccharide(s)</em></td>
<td><em>express</em> a plasmid encoded <em>protein</em></td>
</tr>
<tr>
<td>(†1)</td>
<td>(†4)</td>
</tr>
<tr>
<td>(†1)</td>
<td>(†4)</td>
</tr>
<tr>
<td>(†1)</td>
<td>(†4)</td>
</tr>
<tr>
<td>(†3)</td>
<td>(†4)</td>
</tr>
<tr>
<td>(†4)</td>
<td>(†4)</td>
</tr>
<tr>
<td>(†4)</td>
<td>(†4)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Corpus examples of <em>express + polysaccharide(s)</em></th>
<th>(reported to) <em>express polysaccharide</em> were examined under</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>express</em> polysaccharide. Microbial polysaccharides have</td>
<td>(†1)</td>
</tr>
<tr>
<td>(†1)</td>
<td></td>
</tr>
<tr>
<td><em>express</em> polysaccharide. Microbial polysaccharides have</td>
<td>(†1)</td>
</tr>
<tr>
<td>(†3)</td>
<td></td>
</tr>
<tr>
<td><em>express</em> two distinctly different <em>polysaccharide</em> phenotypes</td>
<td>(†4)</td>
</tr>
<tr>
<td>(†4)</td>
<td></td>
</tr>
<tr>
<td><em>express</em> retroviral <em>sequences</em>. One reason</td>
<td></td>
</tr>
<tr>
<td>(†2)</td>
<td></td>
</tr>
<tr>
<td><em>express</em> such <em>sequences</em> when placed</td>
<td></td>
</tr>
<tr>
<td>(†2)</td>
<td></td>
</tr>
<tr>
<td><em>express</em> these <em>sequences</em>. Polytropic viruses</td>
<td></td>
</tr>
<tr>
<td>(†2)</td>
<td></td>
</tr>
<tr>
<td><em>express</em> VL30 <em>sequences</em> at high</td>
<td></td>
</tr>
<tr>
<td>(†2)</td>
<td></td>
</tr>
<tr>
<td><em>express</em> VL30 retroviral-related <em>sequences</em></td>
<td></td>
</tr>
<tr>
<td>(†3)</td>
<td></td>
</tr>
<tr>
<td><em>express</em> the ropy <em>phenotype</em>. Commercial</td>
<td></td>
</tr>
<tr>
<td>(†3)</td>
<td></td>
</tr>
<tr>
<td><em>express</em> the ropy <em>phenotype</em> expressed</td>
<td></td>
</tr>
<tr>
<td>(†3)</td>
<td></td>
</tr>
<tr>
<td><em>express</em> the ropy <em>phenotype</em>. This</td>
<td></td>
</tr>
<tr>
<td>(†3)</td>
<td></td>
</tr>
<tr>
<td><em>express</em> two distinctly different polysaccharide <em>phenotypes</em></td>
<td>(†5)</td>
</tr>
<tr>
<td>(†5)</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table V.3 (b)</th>
<th>Corpus examples of the quasi-LSP verb <em>express + propositions</em> occurring &gt;3 times</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Brackets indicate span positions following the verb.</td>
</tr>
</tbody>
</table>
To complete the analysis, a search was made on each of the noun propositions in the above combinants, to ascertain whether they might form collocations with verbs significant to the subject field other than *encode* and *express*. As mentioned, the number of occurrences of the nouns varied considerably, some being quite high, and does not take individual nomenclature into consideration:

(878) *cell(s)*, (365) *enzyme(s)*, (2391) *gene(s)*, (187) *phenotype(s)*, (189) *pheromone(s)*, (203) *polysaccharide(s)*, (1959) *protein(s)*, (703) *sequence(s)* and (99) *sugar(s)*;  
[N.B. *gene(s)* and *protein(s)* appear with both *encode* and *express*].

The same criterion of >3 occurrences continued to be used as a baseline, this time for verbs. There was a significant number of LGP verbs which could be said to form 'free' collocations, that is, the collocations could be readily produced by a competent LGP speaker. These included forms of the verbs *to be*, *contain*, *demonstrate*, *determine*, *enter*, *find*, *have*, *involve*, *possess*, *reveal*, *show*, *use*, and these have therefore been omitted from this analysis.

Only two further combinants of significance to the subject field emerged:

12 instances of *digested with* + (name of) enzyme;  
11 instances of *bind (to)* + (name of) protein or other biochemical entity.

As with *encode* and *express*, their frequency is lower than that of the overall number of the nouns with which they form a combinant.
5.4.2 Carrier or support verbs

Verbs which act in support of their accompanying noun were studied, using the noun protection as an example. As has been stated already, such support or carrier verbs tend to become emasculated; in this example they are seen to confirm the static quality of protection, rather than acting in a dynamic way; Banks (1994: 5) calls this a non-dynamic, relational process. This is in contrast to the quasi-LSP verbs which, as mentioned above, denote action and process. The sentences emphasise the rhematic aspect, where the theme may be said to be the human agents which are not mentioned (see next section). Again the criterion of >3 occurrences was used and an example is given in Table V.4 (a) of the verbs which support the noun protection, showing nine instances of provide and elicit, five of confer and three of induce, with their span positions before the noun. Other words deemed relevant are given in span position +1. As can be seen from the table, most of the 'carrier' verbs are in span positions -1 and -2. Table V.4 (b) shows the verbs as they appear in corpus examples.

We now have a selection of examples of combinants typical of the subject field to provide a basis for further analyses.

5.5 Assessment of valency in combinants

The next step in formulating an LSP verb frame has been to apply valency theory to the basic combinant structure, without in any way attempting to re-evaluate the theory of valency grammar. Because of its descriptive techniques, this method of analysing verbs focusses first on verbs and then on those elements which depend on nouns and also on nominalisation. Although valency is regarded as a syntactic structure, nevertheless the semantic restrictions which it invokes give rise to conceptualisation and the realisation of terminological and encyclopaedic relationships. It has proved a
<table>
<thead>
<tr>
<th>Span -4</th>
<th>Span -3</th>
<th>Span -2</th>
<th>Span -1</th>
<th>(132) PROTECTION</th>
<th>Span +1</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) elicited</td>
<td>(2) provide</td>
<td>(3) confer</td>
<td>(4) elicit</td>
<td>(4 6) against</td>
<td></td>
</tr>
<tr>
<td>(2) provided</td>
<td>(3) provide</td>
<td>(2) conferred</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(1) elicit</td>
<td>(2) confers</td>
<td>(2) provided</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(1) elicited</td>
<td>(2) eliciting</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(1) induced</td>
<td>(1) induces</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(1) provides</td>
<td>(1) inducing</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(1) providing</td>
<td></td>
</tr>
</tbody>
</table>

Table V.4 (a) Examples of 'carrier' verbs occurring >3 times supporting the noun protection
Bracketed figures indicate the number of occurrences of each type.
Corpus examples of carrier verb *confer* + *protection*

*confer protection* in chickens (1)
*confer protection* against smallpox (1)
*confer no protection* against challenge with virulent (2)
*confer passive protection*. Protein VP4 (2)
*confer variable protection* against virulent TGEV (2)

Corpus examples of carrier verb *elicit* + *protection*

*elicit protection*. However, the recent work (1)
*elicit protection*. Live vaccines were superior (1)
*elicit protection* against intravenous infection (1)
*elicit protection* against pneumococci (1)
*elicit cross-protection* against four out of (1)
*elicit cross-protection* against more than (1)
*elicit moderate protection* in volunteers (2)
*elicit a protection* equal to (2)
*elicit a highly significant protection* (4)

Table V.4 (b) Corpus examples of 'carrier' verbs occurring > 3 times supporting the noun *protection*

Bracketed figures indicate the span positions of verbs preceding the noun.
useful aid in assessing the terminological relationships in combinants, as well as providing some noteworthy differences between LSP and LGP (5.5.2)³. Despite the restrictive nature of LSP, the identification of actants and circonstants is not only important for deriving the orientation of combinants but also has implications for teaching and for technical writing. Furthermore, since valency does not operate supra-sententially, it is ideal for application within clause and sentence boundaries. In the only instances where it may be said (grammatically) to extend beyond a sentence boundary, it is linked by incorporating demonstrative pronouns, as in: HIS/3T3 cells. These express ....sequences, and the example of sugars given above.

5.5.1 Nominalisation, passivisation and the theme/rheme interchange

A potential compatibility problem appears to arise between valency and functional sentence perspective (FSP) in the case of carrier verbs. This is due to possible interchangeability in the position of the 'recipients of process' and the agent (cf. the third example in Table V.8 (a)) where two valency structures may collaborate by being embedded one within the other in the linear flow of text. It is possible to envisage a variance between the valency structure and the deeper notion of theme/rheme, which would also be true in French, although it does not necessarily apply in all languages.

The basis for this observation is a combination of nominalisation and passivisation which, with the incorporation of a carrier verb, is responsible for a theme/rheme interchange, as the following examples show.

³ I hope in the future to explore comparisons of terminological relationships between different languages further, based on the lexical functions propounded by Mel'chuk and his colleagues (1984).
There is no difference in the position of theme and rheme in the two examples in the active voice, even when a carrier verb is present.

<table>
<thead>
<tr>
<th>Theme</th>
<th>Rheme</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Active voice</strong></td>
<td></td>
</tr>
<tr>
<td>Immunisation</td>
<td>protected the chickens against disease</td>
</tr>
<tr>
<td>Immunisation</td>
<td>provided protection against disease in chicken</td>
</tr>
<tr>
<td><strong>Passive voice</strong></td>
<td></td>
</tr>
<tr>
<td>The chickens</td>
<td>were protected against disease by immunisation</td>
</tr>
<tr>
<td>Protection</td>
<td>was provided by immunisation against diseases in chickens.</td>
</tr>
</tbody>
</table>

Table V.5  Theme/rheme interchangeability with nominalisation in the passive voice

In the first passive example there is a theme/rheme interchange but the rheme contains the same verb as in the active voice. However, with the nominalisation of the verb, the nominalisation becomes the theme and the original theme becomes part of the rheme. The necessity of incorporating a carrier verb to effect this theme/rheme interchange lends status to these verbs in that they alter the information structure of the sentence, living up to their dependency grammar reputation as governors of the sentence. As has been seen earlier in the thesis, valency structure can alter and a systematised study has been made to provide mandatory and optional valency slots; differentials become apparent according to whether a verb is a true LSP one, where it is normally monovalent, or whether it is passivised or a carrier verb (cf. the verb frames in Tables V.9, 10 and 13). These observations have wide-ranging implications for technical writers and...
academic authors who should be aware of the need for phrase ordering when trying to achieve emphasis as it may affect valency structure and the natural sequence of events.

In a different but related way, the optionality of the prepositional phrase in sub-valency position 1 to the noun in valency position 2 (cf. the sentences in Table V.6 (c) from the corpus) can be resolved on the grounds of semantic specificity. This is a very important point which merits further work to try and establish criteria for the mandatory versus optional characteristics of such apparently peripheral phrases.

In addition to the theme/rheme problem discussed above, the idea of theme in these subject fields frequently refers to the undocumented human agent. This psychological subject could be incorporated into a verb frame, probably in practical terms as an optional slot so that occasional references, for example to someone's work or laboratory may be included.

The method for incorporating valency has been to use ATA to provide sentences from the corpus which include examples of combinants.

5.5.2 Valency differences between quasi-LSP and 'carrier' verbs

In Tables V.6 (a) and (b) respectively, the quasi-LSP verbs encode and express are assessed for the valency of their noun combinants, which are always in valency position 2, while the agent is in valency position 1. For comparison, the same test has been done in Table V.6 (c) with the carrier verbs which combine with protection. Protection in valency position 2 results in the 'patient' or recipient of the process being sub-valent to the actant protection; it is therefore in sub-valency position.
<table>
<thead>
<tr>
<th>Valency position 1 (Mandatory actant)</th>
<th>Verb</th>
<th>Valency position 2 (Mandatory actant)</th>
<th>Sub-valency position 1 to (an) enzyme(s) (Optional circonstant)</th>
</tr>
</thead>
<tbody>
<tr>
<td>[_three genes] encode</td>
<td>[enzymes]</td>
<td></td>
<td>[for sugar metabolism]</td>
</tr>
<tr>
<td>[genetic loci (that)] encode</td>
<td>[protein-dependent peptide permeases]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>[mt1D genes (that)] encode</td>
<td>[the PTS Enzyme IImt1]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>[Neither] was shown to encode</td>
<td>[the Scr FI endonuclease]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>[Synthetase] has been found to encode</td>
<td>[an enzyme (endowed with the third motif)]</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table V.6 (a) Valency of English combinants with quasi-LSP verb *encode* and noun *enzyme* 
[...] indicate mandatory valency slots; {...} indicate optional slots; (....) indicate disjuncture 
Punctuation has been included to clarify ends of clauses or sentences.
<table>
<thead>
<tr>
<th>Valency</th>
<th>Verb</th>
<th>Valency</th>
<th>Sub-valency</th>
</tr>
</thead>
<tbody>
<tr>
<td>position 1</td>
<td>do not express</td>
<td>position 2</td>
<td>(Mandatory actant)</td>
</tr>
<tr>
<td>(Mandatory</td>
<td></td>
<td>(Mandatory</td>
<td>(Optional circonstant)</td>
</tr>
<tr>
<td>actant)</td>
<td></td>
<td>actant)</td>
<td></td>
</tr>
</tbody>
</table>

[cell lines (that)]

[cells (from many species)]

[not all murine cells]

[NIH/3T3 cells. (These)]

[packaging cells]

<table>
<thead>
<tr>
<th>[retroviral sequences.]</th>
<th>[(such) sequences]</th>
<th>[when placed in culture.]</th>
<th>[(these) sequences]</th>
<th>[that are packaged into the virions.]</th>
<th>[VL30 retroviral-related sequences]</th>
<th>[VL30 sequences]</th>
<th>[at high frequency.]</th>
</tr>
</thead>
</table>

Table V.6 (b) Valency of English combinants with quasi-LSP verb *express* and noun *sequences*.

[...] indicate mandatory valency slots; {....} indicate optional slots; (...) indicate disjuncture.

Punctuation has been included to clarify ends of clauses or sentences.
<table>
<thead>
<tr>
<th>Valency position 1 (Mandatory actant)</th>
<th>Verb</th>
<th>Valency position 2 (Mandatory actant)</th>
<th>Sub-valency position 1 to protection (Mandatory or optional*)</th>
</tr>
</thead>
<tbody>
<tr>
<td>[immunisation]</td>
<td>provided</td>
<td>[(complete) protection]</td>
<td>[against (moderate to severe) diarrhoea.]</td>
</tr>
<tr>
<td>[an avirulent strain of cowpox virus as a vaccine (which)]</td>
<td>can confer</td>
<td>[protection]</td>
<td>[against smallpox.]</td>
</tr>
<tr>
<td>[fowlpox virus recombinant]</td>
<td>confers</td>
<td>[protection]</td>
<td>[in chickens.]</td>
</tr>
<tr>
<td>[PspA]</td>
<td>is capable of eliciting</td>
<td>[protection]</td>
<td>[against pneumococci (of more than one capsular type.)]</td>
</tr>
<tr>
<td>[the strain]</td>
<td>induced</td>
<td>[(significant) protection]</td>
<td>[without shedding (in goats.)]</td>
</tr>
</tbody>
</table>

Table V.6 (c): Valency of combinants with protection + carrier verbs

[...] indicate mandatory valency slots; {....} indicate optional slots; (....) indicate disjuncture. Punctuation has been included to clarify ends of clauses or sentences.

* May be a mandatory actant or an optional circonstant on the grounds of semantic specificity; examples 1 and 5 are optional, whereas 2, 3 and 4 are mandatory.
1. There are two instances of an optional 'circonstant' in sub-valency position 2, while the other three examples are deemed to be mandatory.

In general, the criterion for identifying optional circonstants is their ability to form a separate clause which is not mandatory to the sentence, by applying the 'backformation' test (the 'Zurückführungstest' of Helbig and Schenkel (1973: 37-38), *inter alia*). I would take this criterion a step further because this test is a useful guide, rather than providing a conclusive argument, as the sentences in Table V.6 (c) show. Using the backformation test on all the prepositional phrases in sub-valency position 1 would indicate that these are optional. However, at the semantic level, examples 2, 3, and 4 in valency position 1 are so specific to their subject field that inclusion of the sub-valent prepositional phrase is crucial for the sentence to make sense, suggesting that a terminological, semantic assessment is necessary to decide the nomenclature of the slot that is to be provided (i.e. whether it is to be mandatory or optional).

This observation corroborates the importance accorded by Brinkmann (1971) to the nominative on which he bases his valency hierarchy; the 'Mitspieler' of a verb can attract and control its own Mitspieler. Helbig and Schenkel (1973: 23-24) enlarge on this, stating that 'the noun in subject position can have control in a syntactic sense only over optional, but not mandatory, components in a sentence. In the case of deverbalised nouns, the valency of their underlying verb is abstracted'. A verb must have a valency, even if it is zero; however, when its semantic content is represented by a noun, the surface valency structure may alter. Nouns may have a syntactic optional valency only when they are the nominalisation of a verb. The proof of noun valency is not at the syntactic
level but at the semantic level. In summary, there are several important differences in the valency pattern between quasi-LSP verbs and 'carrier' verbs with the nominalised form of the verb.

A comparison of a verb and its corresponding noun, e.g. protect/protection, shows a difference in valency. The verb protect in the active has the same valency pattern as the quasi-LSP verbs encode and express, while in the passive it has one mandatory and one optional valency slot. However, the noun protection has one mandatory and two optional slots regardless of its carrier verb. As we have seen, nominalisation of a verb therefore alters in valency and is in addition a mechanism for avoiding the use of a primary agent, particularly if the agent is human. Halliday (1988) has stated that in the development of scientific writing over the years, authors have increasingly had recourse to nominalisation, leading Banks to conclude that 'the nominal groups which result will naturally tend to be linked by relational process verbs' (Banks, 1994: 5), i.e. verbs in the passive and, invariably, carrier verbs. It would be interesting to study the degree of security of valency mechanisms when nominalisation comes into focus; for example, is it the actant or the circonstant that becomes more marginal? Carrier verbs tend to be collocational rather than denominal, but is there a way of quantifying the extent to which they become effete? A further worthwhile study would be to evaluate to what degree there is inheritance from the verb in nominalisation.

5.5.3 Valency differences between LSP and LGP

A difference in valency can also occur between LSP and LGP, as has already been noted in the examples patients present with [+ symptoms] and the vaccine took. The first example has an understood reflexive,
Patients present (themselves) with [+ symptoms], whereas the second example is intransitive (*the vaccine took itself). An example from the corpus, *Streptococci initially present on the outer surface of the skin* also indicates an understood reflexive element. The mandatory actant proposition *Streptococci* is found in valency position 1, while there is a mandatory spatial adverbial phrase in valency position 2. In the next section it is shown how this example can be converted into a form suitable for indexing purposes. Although the format of the original concept of a verb + noun combinant is not generally adhered to in these examples, nevertheless intransitive and reflexive verbs form important groups and need to be included when constructing verb frames; in other words, we are now considering the noun + verb construction. There were few instances of a valency change observed in the English corpus; none was immediately apparent in the French corpus, although examples have been noted in other spheres (Pavel, 1993: 71):

- *itérér une équation* (transitive);
- *l’agrégat cristallise* (and not *se cristallise*);
- *entre des données* (into a computer).

(These examples have been taken from text and are thus not in the form of lexicographical entries)

It would be interesting to ascertain, given any verb with its set of valency arguments, what the statistics are of unoccupied valency slots. This would be possible using corpus analysis but has not been undertaken at present since it is not crucial to the current issue of orientation.

5.5.4 Intransitive and covertly reflexive LSP verbs

A search in the corpus for the intransitive verb, *replicates*, resulted in 16 examples of *The virus replicates* and a sample of these
follows:

*The virus replicates in certain fish cell lines*
*The virus replicates in the nucleoplasm of the host.*
*The virus replicates in the peripheral region of the host cell*
*The virus replicates in vitro in cell lines derived from the homologous host.*

.....a lytic cycle where the virus replicates.

*The live attenuated virus replicates only in the gut epithelial cells.*

*This group of viruses replicates to high titres in the respiratory tract*

The example of *The virus replicates* could also be considered to contain an understood reflexive construction (*the virus replicates itself*) and would appear in its reflexive form in French and Swedish. In the following section the verb is passivised to give a probable indexable entry. Whether it is passivised or not, the following results:

One mandatory actant: (agent) *virus*
One optional circonstant: (spatial adverbial phrase) *in certain fish cell lines etc.*

5.6 *Passivisation and indexing as indicators of orientation*

Having studied the valency of the combinants and used this to compare the active forms of the verbs in Tables V.6 (a), (b) and (c), I now anticipate the indexing process as a pointer to orientation by passivising the verbs to form complete, semantically correct, statements. In each case the result shows one mandatory actant with two optional 'circonstants' (Table V.7). In addition, the verb *protect*, as the counterpart to *protection*, has been included in the study.

Chapter 5: 247
### Table V.7: Passivisation of quasi-LSP verbs

<table>
<thead>
<tr>
<th>Terminological relationship</th>
<th>Optional of constant in valency position 1</th>
<th>Optional of constant in valency position 2</th>
<th>Optional of constant in valency position 3</th>
<th>Action, process</th>
<th>Recipient of process (patient)</th>
<th>Recipient of Process</th>
<th>Recipient of Process</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chapter 5: 248</td>
<td>Enzymes</td>
<td>by three genes</td>
<td>by cell lines</td>
<td>by all cell lines</td>
<td>encode</td>
<td>Process, state</td>
<td>against diarrhoea in chicken</td>
</tr>
<tr>
<td></td>
<td>Protein-dependent peptide permease</td>
<td>by three genes</td>
<td>by cell lines</td>
<td>by all cell lines</td>
<td>express</td>
<td>Process, state</td>
<td>against diarrhoea in chicken</td>
</tr>
<tr>
<td></td>
<td>SrfI endonuclease</td>
<td>was shown to be encoded</td>
<td>were expressed</td>
<td>were not expressed</td>
<td>process</td>
<td>Process, state</td>
<td>against diarrhoea in chicken</td>
</tr>
<tr>
<td></td>
<td>Retroviral sequences</td>
<td>was protected against</td>
<td>was protected against</td>
<td>was provided</td>
<td>state</td>
<td>Process, state</td>
<td>against diarrhoea in chicken</td>
</tr>
<tr>
<td></td>
<td>Sequences</td>
<td>by low-pox virus</td>
<td>by PspA</td>
<td>conferred</td>
<td>process</td>
<td>Process, state</td>
<td>against diarrhoea in chicken</td>
</tr>
<tr>
<td></td>
<td>Recipient of process (patient)</td>
<td></td>
<td></td>
<td></td>
<td>process</td>
<td>Process, state</td>
<td>against diarrhoea in chicken</td>
</tr>
<tr>
<td></td>
<td>Recipient of process (patient)</td>
<td></td>
<td></td>
<td></td>
<td>state</td>
<td>Process, state</td>
<td>against diarrhoea in chicken</td>
</tr>
<tr>
<td></td>
<td>Protection</td>
<td></td>
<td></td>
<td></td>
<td>is conferred</td>
<td>Process, state</td>
<td>against diarrhoea in chicken</td>
</tr>
<tr>
<td></td>
<td>Protection</td>
<td></td>
<td></td>
<td></td>
<td>elicited</td>
<td>Process, state</td>
<td>against diarrhoea in chicken</td>
</tr>
</tbody>
</table>

Examples of (i) the quasi-LSP verbs encode and express, terminologically transitional and carrier verbs. (ii) terminologically transitional (action -> process -> state) LGP verb protect, and (iii) the carrier verbs conferred, elicited and provided, show a difference in the position of the Recipient of Process ("Patient").
While the agent in all cases has proved stable, both in its position and its introduction with the preposition by, several interesting results emerged:

- the 'process' evidenced in the verbs encoded, expressed, protected was not present in the 'carrier' verbs, which denoted only 'state';
- the 'recipient of the process' with these verbs, after passivisation, was in valency position 1.

However, with the 'carrier' verbs,

- the 'process' was in valency position 1;
- the recipient of the process, or 'patient', was in valency position 3.

If the sentences in Table V.7 are now condensed and transformed into potential indexable entries (Table V.8 (a); cf. Austin and Butcher's work (1969) and the discussion in Chapter 2 of this thesis), the results show that nominalisation of the quasi-LSP verbs and protect brings them into first position, i.e. in indexing terms, the semantically important one which for our purposes denotes orientation:
and the same would be true of other examples. The order of the optional
constants is immaterial for the sentence to make sense; however, when
indexing, the recipient of the process is semantically closer to the process
than the agent by which it is achieved. From the point of view of providing
slots in a frame, the first two examples in the table present no problem;
however, the third could have a number of valency slots but linguistic choices
need to be made and it is thought that adequate provision has been made for
these in the frames which follow.

It is important to note that the same conversion cannot be made
with the 'carrier' verbs.

Furthermore, if the sentences containing the understood
reflexive and intransitive verbs in the previous section are passivised, the
following results:
To return to orientation at the surface or sentence level, which is the source of our potential framework for a terminographic entry, and also for compiling a dictionary of collocations, combinants for all verbs appear to have one mandatory actant, with one or more optional 'circonstant'(s) (Table V.7). However, moving away from the sentence level to passivisation, it is, perhaps not surprisingly, the mandatory actant (in the case of quasi-LSP verbs, their nominalised form) in valency position 1 which provides the orientation of the combinant.

In contrast, however, when the carrier verbs are assessed at sentence level, the orientation focusses on the noun in valency position 2. This observation reflects that of Danlos and Samvelian (1992: 21) who, in relation to machine translation, suggest starting from the predicative noun when these have 'support' verbs, since the noun is the most informative element in a sentence.

Verbs with an understood reflexive have the same 'agent' and 'recipient of process' reflected in their subject, as do intransitive verbs. In these constructions, it must be recalled that we are dealing with a noun +
verb combinant. After passivisation and indexing, orientation would be indicated by the verb which is therefore the base, while the attendant subject/reflexive becomes the collocatory part of the phrase (see Table V.13 for reflexive French verbs).

5.7 The LSP verb frame: comparison with a 'carrier' verb frame

Annex 8 shows a proposed blank model of a verb frame for transitive LSP verbs combinants, while in Table V.9 it is completed. In addition to its linguistic element, the frame incorporates encyclopaedic, terminological relationships. Examples of sentences and clauses as they appear in the corpus, i.e. in natural word order, as in Table V.6, can be analysed through assessment of collocatory patterns to identify combinants. It is envisaged that these combinants might be tagged according to the slots and headings in Tables V.9 and V.10; the combinants could then be assessed for their orientation to give base and collocate, as well as providing relationships relevant to the terminologist.

Table V.10 contains a similar verb frame for carrier verbs. The important differences which have emerged between LSP verbs and carrier verbs are:

(i) the valency position of the relationships which stem from indexing, i.e. agent, process, recipient of process (patient);
(ii) the terminological relationships of the verbs (action, process, state);
(iii) the relative positions of base and collocate, i.e. orientation.
<table>
<thead>
<tr>
<th>SLOT 2</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Corpus example</td>
<td>Valency position 1</td>
<td>Actant</td>
<td>Agent</td>
<td>Span position -1 to (&gt; -7 (depending on length of m.l.u.)</td>
<td>Grammatical representation: subject Optional</td>
</tr>
<tr>
<td></td>
<td>Three</td>
<td>genes</td>
<td></td>
<td></td>
<td>articles</td>
</tr>
<tr>
<td></td>
<td>encode</td>
<td></td>
<td></td>
<td></td>
<td>numbers</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>negatives</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>pronouns</td>
</tr>
</tbody>
</table>

**SLOT 1**

Quasi-LSP verb: encode

**BASE OF COMBINANT**

<table>
<thead>
<tr>
<th>Relationship</th>
<th>Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>action</td>
<td>transitive</td>
</tr>
<tr>
<td>process</td>
<td>infinitive</td>
</tr>
<tr>
<td></td>
<td>active voice</td>
</tr>
</tbody>
</table>

**SLOT 3**

Valency position 2 | Actant | Recipient of process (patient) | Span position +1 to (> + 7 (depending on length of m.l.u.) | Grammatical representation: direct object Optional | Mandatory |
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>enzymes</td>
<td></td>
<td></td>
<td></td>
<td>articles</td>
<td>noun + (noun)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>numbers</td>
<td>adjective + noun</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>negatives</td>
<td>adverb + adjective + noun</td>
</tr>
</tbody>
</table>

**SLOT 4**

Valency position 3 | Circconstant | Relationship Purpose | Grammatical representation: adverbial phrase prepositional phrase |
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>for sugar metabolism</td>
<td>optional</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Table V.9** Proposed model of a verb frame for transitive quasi-LSP verb combinants

(m.l.u. = multiple lexical units)
<table>
<thead>
<tr>
<th>Slot 2</th>
<th>Slot 1</th>
<th>Slot 3</th>
<th>Slot 4</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Agent</strong></td>
<td><strong>Verb</strong></td>
<td><strong>Actant</strong></td>
<td><strong>Purpose/Location</strong></td>
</tr>
<tr>
<td><strong>Valency</strong></td>
<td><strong>Valency</strong></td>
<td><strong>Valency</strong></td>
<td><strong>Valency</strong></td>
</tr>
<tr>
<td><strong>Corpus example</strong></td>
<td><strong>Verb example</strong></td>
<td><strong>Verb example</strong></td>
<td><strong>Verb example</strong></td>
</tr>
<tr>
<td>Fowipox virus</td>
<td>contains</td>
<td>protection</td>
<td>in chickens</td>
</tr>
</tbody>
</table>

**Chapter 5: 254**

**Table V. 10 Proposed model of a verb frame for carrier verb combinants**

(m.l.u. = multi lexical units)
<table>
<thead>
<tr>
<th>Valency position 1 (Mandatory actant)</th>
<th>Agent and recipient of process</th>
<th>Verb</th>
<th>Process</th>
<th>Adverbial phrase (Optional circonstant)</th>
<th>Prepositional phrase (Optional circonstant)</th>
</tr>
</thead>
<tbody>
<tr>
<td>le gène marqueur...</td>
<td>glooms</td>
<td>s'y exprimait</td>
<td>s'aggravant</td>
<td>&quot;avec une efficacité tout à fait surprenant&quot;,</td>
<td>_</td>
</tr>
<tr>
<td>essoufflement</td>
<td>symptômes neurologiques: céphalée</td>
<td>allant en s'aggravant</td>
<td>s'aggravant</td>
<td>_</td>
<td></td>
</tr>
<tr>
<td>tuméfaction unilatérale d'installation brutale et déficits centraux, déments, céphalées</td>
<td>_</td>
<td>_</td>
<td>_</td>
<td>_</td>
<td></td>
</tr>
</tbody>
</table>

Chapter 5: 255
5.8 French verbs in the biological sciences corpus

Many of the French transitive specialised verbs can be treated in the same way as their English counterparts. Where, however, English uses the passive voice, in French the reflexive form is often preferred and in the French corpus of half a million words, approximately 1,500 were reflexive. This large group merits particular attention because in these verbs the 'agent' and 'recipient of the process' are the same, the latter being represented by the reflexive pronoun. Semantically, reflexivity brings the idea of 'change'. In Table V.11 examples are given and Table V.12 shows how the clauses would appear in 'indexable' form. French does not have the linguistic mechanism to indicate that a term is made up of a series of units as in English, e.g. cephalus aggravation, and has to rely on the use of de and prepositions.

<table>
<thead>
<tr>
<th>Process</th>
<th>Agent and Recipient of Process</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expression</td>
<td>du gène marqueur</td>
</tr>
<tr>
<td>Agglomération</td>
<td>de flocons</td>
</tr>
<tr>
<td>Aggravation</td>
<td>d'essoufflement</td>
</tr>
<tr>
<td>Aggravation</td>
<td>de céphalée</td>
</tr>
<tr>
<td>Aggravation</td>
<td>d'une tuméfaction</td>
</tr>
<tr>
<td>Aggravation</td>
<td>de démence, de céphalées</td>
</tr>
</tbody>
</table>

Table V.12 Indexable form of French reflexive verbs

As in English, the verb frame needs to be constructed with
acknowledgement to a possible supra-sentential psychological subject which, if it is indicated, will have been elaborated in preceding sentences of the text. The presence of a psychological subject, often referring to the subject specialist, incurs a problem of double polarity between subject and object and can give rise to a diversity of valency; sometimes reflexivity in French is used where English would use the passive, while at others it has a purely pronominal role. Pronominal verbs are conjugated with the same personal pronoun as the subject, e.g. *les symptômes se produisent*. Where there are two pronouns, the latter is the indirect object complement, e.g. *il se vante*. However, when the action reflects directly on the subject, the verb is said to be reflexive pronominal, as in *il se lève*, whereas with reciprocal pronominal verbs, there is more than one subject, each of which acts upon the other(s), as in *le virus et le vaccin s'aggravent*. Certain verbs have only a pronominal form, such as *se repentir*, where the pronoun is never separated from the verb. The verbs which adopt the pronominal form to express the passive are those which appear with greatest frequency in the corpora, e.g. *les médicaments se prennent dans l'eau*, where a psychological human subject is necessary for the phrase to make sense.

In Table V.13 a frame is proposed for reflexive verbs in the French biological sciences. Because of their reflexivity, the mandatory actant in valency position 1, being both 'agent' and 'recipient of process', becomes the collocatory part of the combinant.

5.9 **Adjuncts and disjuncts**

This section is included for the sake of completeness, because it must be borne in mind that the source of this work is flowing text and therefore adjuncts will be present. Adjuncts are defined by Quirk *et al.*
Table V.13 Proposed model of a verb frame for reflexive French LSP verbs in the biological sciences

(m.l.u. = multiple lexical units)
adjuncts in that they cannot appear first in
a negative sentence but can be the focus of a question or clause negation (ibid. 208). It was clear that in most instances of the verb + proposition phrase, disjuncts in English would take the form of 'free' adjectives, most of which act in a quantifying capacity to their accompanying proposition: all, complete, cross-, different, full, good, long-lasting, moderate, no, numerous, other, passive, significant, variable, various, and numbers, or adverbial phrases: (functionally, mainly, only, partially, at competence, in vitro). In the examples of encode + protein(s) in Table V.2 (b), the noun surface combines adjectivally with protein to form a term, whereas the free adjective different occurs on three occasions but cannot be said to form a term. As already stated, the question of how 'free' such adjectives are warrants further research.

5.10 Conclusions for constructing frames for LSP combinants

The value of linking two areas which are of prime importance in lexicography and terminography, namely, corpus analysis for eliciting collocations, and valency theory, has undoubtedly proved fruitful in identifying terminological relationships in the different categories of verbs and verb combinants studied. Moreover, the mandatory and optional characteristics of 'actants' and 'circonstants' which have become apparent, are vital factors to incorporate into a frame for combinants. Two important differences have been noted in:

- quasi-LSP verbs and carrier verbs, the latter with their

---

4 Ranganathan (1957: 159-160) recognised what he calls 'peripheral facets' of 'place' and 'time', placing them in that order.
- the valency of certain verbs which 'behave' differently in LSP and LGP.

The application of indexing principles, which has also led to the identification of relationships, has been invaluable in the ultimate quest for orientation. Frequency can also be a useful pointer for assessing orientation, on the indexing principle of 'the rarer the better'. However, care needs to be taken with this approach because context has to be the deciding factor, as has been seen with the example of *measles vaccine*. The final assessment must be a subjective one on the part of the terminologist, corroborated by the informed decision of a subject specialist.

It is hoped that these frames for special language combinants will prove instrumental in providing a basic construction from which terminologies incorporating collocation can be compiled in a systematic manner based on theoretical principles which can be corroborated statistically. By studying a special language, it has proved more straightforward to produce a workable model than it would have been for general language, because of the highly restricted nature of the lexis, syntax and to a great extent semantic content of the texts analysed. It should however be possible to adapt the model to less rigid subject fields by the methods used.
Chapter 6

Conclusions and areas for further research

I have sought to explain the concept of what I am calling orientation and to emphasize its importance as a primary factor in the area of terminology research which covers multiple lexical terms, and verb phrases, or combinants. Establishing the orientation of terms and phrases has been made possible by assessing various linguistic mechanisms which have been deemed relevant for the purpose and which have produced conclusive results. The mechanisms have been incorporated in what has seemed a logical continuum, from the compilation of special language corpora comprising authentic current contextual material in different languages, through corpus analysis which has provided frequency, to the extraction of collocation patterns. These analyses have led naturally to a study of valency which has brought to light a number of valency inheritance mechanisms when parts of speech change roles, e.g. timber infestation (by wood-boring insects); infestation of timber (by wood-boring insects); wood-boring insects' infestation of timber, as well as a certain conflict between mandatory actants and optional circonstants, particularly in relation to carrier verbs. Furthermore, writers need to order the elements of their utterances with regard to thematisation and orientation. Servicing the writer's needs is complex because certain options may be removed or constrained. Subsequently, the process of passivisation, followed by nominalisation as a precursor to indexing, has resulted in the goal of orientation. A bridge has therefore been created which links information science with LSP linguistics and hence between the citation form of a term and discourse.

The process of providing orientation for terms and phrases in a
subject field can be applied to compilations such as books, where an index needs to be created, or to term banks for the purpose of 'terminologisation'. In these cases, it is vital to identify the aim of these products and their intended readership or user before embarking on the process of providing specialised subject material for orientation. A study of particular discourse communities, as identified by Swales (1990: 24-27), is of great value in making these assessments.

As I have shown, frequency can provide a first-stage pointer to orientation, on the indexing principle of 'the rarer the better', but it would be irresponsible to use this as a global criterion and great care needs to be taken with this approach. I have shown that orientation depends on context, which is reflected in the needs of the intended readership or user; hence frequency will reflect the context. However, in a short work such as a monograph there may be insufficient material to provide significant pointers to orientation and therefore assessments by subject specialists and terminologists are crucial.

Lerat (1994: 46) states that we are still a long way from automatically extracting a dictionary for a specific use from a corpus of texts and documents, for various linguistic reasons: words are often polysemous, even within the same domain (cf. the example of nucleus in Chapter 1 of the thesis); Lerat quotes the example of droit in French which can mean either tax or prerogative and may even have no terminological significance, as in à bon droit. Texts are often heterogeneous (a notice for customers differs in many respects from a feasibility study), even when an attempt has been made at homogeneity within the same organisation. It is well known that a comparison between terminology and style in different languages is vital for accurate translating work. Some language differentials have been explored in this work, particularly in translation practice. A new hypothesis which I
intend to examine further is the evidence that technical terminology in French is more discoursal than in English which for the most part takes the form of lexical juxtaposition. There are, however, some interesting examples related to indexing techniques which show a difference within English, e.g. Management of Libraries in Canada is more discoursal than Canadian Library Management which is more akin to terminology.

It is hoped that the method propounded in this thesis for constructing frames for verb combinants in special languages will be of benefit to terminologists and to compilers of special subject dictionaries, who are usually subject specialists with scant lexicographical knowledge, and that these methods will also pave the way for the compilation of dictionaries of collocation. Although an understanding of a subject field is likely to be difficult for non-subject specialists, nevertheless it is probably more fruitful to provide a basis for dictionaries of collocation by starting with the restricted lexis and grammatical structure found in these fields and to draw conclusions from the patterns which emerge, rather than attempting to start by compiling general language collocatory dictionaries because of the broad range of the contextual material. Dictionaries of collocations are likely to become more prevalent in French not only because few are available but because they are undeniably useful for encoders of the language. This is particularly important at the present time as currently any scientific meeting organised in France by French people must include French as one of its languages.

In addition to the importance of orientation to terminology, the different stylistic mechanisms which have come to light, such as the nuances in meaning engendered by word order and the apparent conflict between mandatory actants and optional circonstants, have a considerable impact on
the important area of technical writing. This has been explored in depth by Martin (1993: 221-280) who has also recognised the conflict between theme and rheme.

6.1 Areas for future research

The work I have carried out has brought to light a number of areas in which further research would be of benefit linguistically, lexicographically and terminographically. These can be broadly grouped into the following categories:

Adjectives and adverbs

1. It would be interesting to know what criteria are required for adjectives to pass from LGP to LSP status, e.g. *an attractively dressed woman* \(\rightarrow\) *a musically gifted child* \(\rightarrow\) *an orally administered drug*. The last example may be perceived as a term, but does the second example have a similarly fixed status? In other words, how and at what point do the words become integrated? It seems that there would be some sort of cline, but resolving the process of how this occurs would be worthy of study.

2. The past participle used adjectivally is a frequent construction in LSP and an assessment of its use in this way, in comparison with its function as a past participle, with or without an auxiliary verb, would reveal some interesting pointers to the structure of LSP.

3. Compound adjectives and the adverb + past participle construction, as in *freeze-dried* and the examples in point 1 above, appears to be more common in LSP than in LGP. It would be interesting to discover to what extent these constructions are 'verbalised' (in the sense of becoming a verb) in the formal expression of special subject fields, or whether they belong
more commonly to spoken language.

Verbs

1. The retrieval of verbs from the corpora was a tedious task, particularly in English, and applied research on search mechanisms for verbs to provide automatic tagging would be of great benefit.

2. A considerable amount of work, both theoretical and statistical, has been carried out in this thesis to ascertain the 'LSP-ness' of verbs and a number of points have been resolved. Further criteria, for example, semantic ones, for defining whether and how a verb belongs to its subject field would be valuable.

3. Carrier verbs in attendance on a deverbalised noun and the changing role of nominalisation in scientific writing (cf. Halliday, 1988; Martin, 1993) are intriguing major areas for further study.

4. Allied to the previous point, it would be worthwhile evaluating to what degree there is inheritance from the verb in nominalisation, by exploring valency mechanisms, which are frequently affected by changes in grammatical roles.

5. The type and number of valency differences between LSP and LGP verbs has been remarked upon and several examples have been given in this thesis. It is interesting to speculate on how and why this phenomenon should occur. Furthermore, given any set of valency arguments, it would be worthwhile ascertaining the statistics for unoccupied valency slots. Corpus analysis would be the only feasible way to make such an assessment. In addition, the hidden reflexive, which can alter valency structure, e.g.
patients present (themselves) with, merits further work.

Special subject fields

1. It would be interesting to define the nature of the 'ecotype' of a subject field, together with the typicality and cohesion within a discourse community and its and degree of autonomy of the discourse community in which it thrives.

2. In the instances where special subject fields overlap, as is invariably the case with law and economics which never stand alone, it would be worthwhile ascertaining whether there is a natural precedence of one over another.

3. It would be useful - but maybe impossible - to be able to assess and quantify how small a corpus would be feasible to provide conclusive statistical evidence in restricted LSPs, to avoid lengthy and sometimes abortive searches for suitable material in machine-tractable form. Kittredge (1987: 63) states however that size or complexity are less of a qualifying factor for a sublanguage than its adherence to systematic usage. 'It is in fact the degree of systematicity which will determine how appropriate a sublanguage is (in Kittredge's case) for automatic translation.'.

Linguistic research

1. An analysis of the use of what I call 'anaphoric reduction', by which I mean the abbreviation of terms and their acronyms in the course of a text, would be of enormous benefit to translators and interpreters. This is an important textual consideration and is one of the principal areas which I am interested in pursuing. I would ascertain how prevalent the tendency is and
try to establish whether it is possible to assess the approximate point in a
text where this phenomenon begins to occur so that at a certain distance into a
text, translators and particularly interpreters would be alerted to the
possibility of the phenomenon.

2. An assessment of whether and to what degree Germanic versus
Latinate roots alter in valency could provide some interesting results

3. The typological difference between French and English, where
French tends to be more discoursal while English is terminologically
'holistic', would be interesting, to ascertain inter alia the proportion of
each.

4. The theme/rheme interchange which provokes valency differentials
would be an important basis for work on classifying the 'mandatoriness' of
actants and the optionality of circonstants, together with their degree of
'hierarchical' importance with regard to their position in a sentence, or
'marginality'.

5. The mechanisms whereby one word is attracted to another or to a
group of words to form collocations would be invaluable, as would an
assessment of whether the criteria for these factors vary between LSP and
LGP. For example, alliteration and euphony are likely to be instrumental in
forming collocations in LGP but intuitively are less likely to occur in most
LSPs.

6. Finally, I should like to make a more detailed comparison of the
terminological relationships in different languages, based on the lexical
functions described by Mel'chuk and his colleagues (1984: 6-7).
Bibliography


Bibliography: 268


Bibliography: 269
Model.' Report No. 86/1. CCL/UMIST.


Bibliography: 270


Bibliography: 271


Bibliography: 272

ISBN 83-01-10443-0.

Bibliography: 273
Russky Yazyk.

Lambert, G. 'The application of PRECIS in French.' (PRECIS in a Multilingual Context, Part 4). Libri, 26, 302-323.


Lehrberger, J. (1982). 'Automatic Translation and the Concept of Sublanguage.'


Bibliography: 274


Bibliography: 276


Bibliography: 277
University of Leeds, April 1994 (to be published).


**Background reading**


Bibliography: 280
Annex 1

Recommendations for terminologists identifying the headword in multiple lexical units

The numbered sections refer to the text in Chapter 3.

**Compound nouns (3.7.1)**

When two or more nouns combine to form a concept and are subsequently represented by a term, the generic, non-logical/ontological noun is used to denote the place of entry (the headword is underlined in each example), e.g. *brake booster*; in English, if the insertion of *of* after inversion of the nouns points to the attributive use of the second noun, the first noun should be used to denote the entry, e.g. *drug pusher*. In lexicography, adjectival attributive use, or in terminography, specific rather than generic use (i.e. ontological, function) are not usually recommended for the point of entry. However, the reason for which the work is to be used and the needs of the user are paramount.

Recommendation: It appears that the way a noun is used in a term consisting of compound nouns is vital to where it would be placed as an entry term and depends primarily on the needs of the user.

**Compound verbs (3.7.2)**

With verbalised nouns formed from a noun + verb, an adjective + verb or a noun + noun, the first word is used as the entry term (e.g. to aircondition, to cat-con). However, if the first word is used adverbially, it is suggested that the second word should be the point of entry for the term (e.g. to cadence-brake, to double-park, to mouth-pipette).
Recommendation: The function of the grammatical components of compound verbs needs to be identified so that the entry word can be deduced.

**Compound adjectives (3.7.3)**

Recommendation: It is suggested that, where both parts of the compound adjective together form a single concept with a logical, generic/specific relationship, e.g. *single-stranded*, entry would be at the first word of the compound. *Hydrogen-bonded*, on the other hand, is an ontological relationship formed from two distinct concepts. Here I suggest that assessment of the entry point depends on the context in which the term appears.

**Compound adverbs (3.7.4)**

Recommendation: The whole term represents one concept and would be indexed as a unit under *quantum*, e.g. *quantum mechanically* (colloquial use).

**Adjective(s) + noun(s) (3.7.5)**

Recommendation: The specificity of the adjective(s) in relation to the noun(s) described tend to give the adjective precedence as the point of entry for the term, e.g. *catalytic converter*.

**Adverb + adjective (3.7.6)**

Recommendation: Since it is the adverb which renders the term specific, it is pertinent to place the collocation under the adverb e.g. *hermetically sealed*.

**Verb + adverb particle (3.7.7)**

Recommendation: Infrequent in LSP; the verb takes

Annex 1: 282
precedence.

Verb + preposition (3.7.8)

Recommendation: In a verb + preposition construction, the verb will figure as the point of entry, e.g. to replicate in.

Idioms (3.7.9)

Both noun and verb should be entered and cross-referenced, e.g. to debug a program would have a cross-reference to debug and program.

Recommendation for verb + noun (3.7.10)

In a verb + noun collocation it is the LSP noun which will figure as the headword for the entry, e.g. to confer immunity, except in the few instances where the verb belongs to LSP, e.g. to debug a program. 'True' LSP verbs need to be distinguished through consultation with a subject specialist.
Dear

I am working on my Ph.D thesis which will assess how organisers of term banks enter and resolve problems concerning entry terms, or head words, containing more than one word. It would be most helpful to me if you could supply the following information:

1) Date of inception of your term bank.
2) The languages covered.
3) The domains included.
4) The number of terms in each language/domain.
5) How you deal with phraseology or multi-word terms, both as entry terms and in the text of the definitions, and which may at various times and in different domains be referred to as any of the following:
   - compound/complex nouns/terms;
   - phrasal verbs;
   - collocation;
   - connotation;
   - metaphors;
   - idioms;
   - verbs and prepositional phrases;
   - lexical syntagms;
   - syntagms of discourse/phrasal unit;
   - extended terms.
6) What data base system is used?
7) What retrieval methods are used and how do they cope with queries/searches for multiple word groups, i.e. which word in a phrase is used to search on and what are the criteria for the choice?
8) Any bibliographical references you may feel relevant to this study.
9) Example of your term record format.

I very much appreciate your assistance in this work and look forward to hearing from you.

Yours sincerely,

(Mrs.) P. Thomas.
Annex 3

List of 74 Term Banks worldwide taken from International Term Bank (1989) list with date of reply

<table>
<thead>
<tr>
<th>International</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1. FAO</td>
<td>11th Dec. '89</td>
<td></td>
</tr>
<tr>
<td>2. World Bank</td>
<td>19th Dec. '89</td>
<td></td>
</tr>
<tr>
<td>3. ILO</td>
<td>5th Dec. '89</td>
<td></td>
</tr>
<tr>
<td>4. IMF</td>
<td>15th Dec. '89</td>
<td></td>
</tr>
<tr>
<td>5. MITRAD</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. UATI</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7. UNO</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8. WMO</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Multinational</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>9. Euro-Arabian Terminology Data Bank</td>
<td></td>
<td>Defunct</td>
</tr>
<tr>
<td>10. FARABI</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Europe</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>11. EURODICAUTOM</td>
<td>13th Dec. '89</td>
<td></td>
</tr>
<tr>
<td>12. OECD</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>United Kingdom</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>13. UMIST</td>
<td></td>
<td></td>
</tr>
<tr>
<td>14. PRECIS REN</td>
<td>9th Jan. '90</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Ireland</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>15. Irish Terminology Data Bank (planned)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>FRG</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>16. FASTERM</td>
<td>29th Jan. '90</td>
<td></td>
</tr>
<tr>
<td>17. MADSTA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>18. LEDA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>19. LEXIS</td>
<td>12th Dec. '90</td>
<td></td>
</tr>
<tr>
<td>20. Ruhrgas AG</td>
<td>1st Dec. '90</td>
<td></td>
</tr>
<tr>
<td>21. TEAM</td>
<td>15th Jan. '90</td>
<td></td>
</tr>
<tr>
<td>22. TERM</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Germany</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>23. DMF</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Austria</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>24. AÖKÖVEN</td>
<td>22nd Dec. '90</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Czechoslovakia</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>25. Czech terminological data bank</td>
<td>10th Jan. '90</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Poland</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>26. CINIM</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>France</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>27. ALEXIS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>28. CEDATA/TERM</td>
<td></td>
<td></td>
</tr>
<tr>
<td>29. CEDATETERM</td>
<td></td>
<td></td>
</tr>
<tr>
<td>30. MIDIST</td>
<td>12th Jan. '90</td>
<td></td>
</tr>
<tr>
<td>31. LINGUIS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>32. NORMATERM</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Spain</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>33. TERMACAT/TERM</td>
<td>15th Jan. '90</td>
<td></td>
</tr>
<tr>
<td>34. EUSKALTERM</td>
<td></td>
<td></td>
</tr>
<tr>
<td>35. EITA</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>USSR</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>36. MULTLEX</td>
<td></td>
<td></td>
</tr>
<tr>
<td>37. SUVTERM</td>
<td>27th Apr. '90</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Finland</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>39. NOKIA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>39. TEPA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>40. TELENOKIA</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Annex 3: 285
Annex 4

On-line search and retrieval mechanisms in term banks

<table>
<thead>
<tr>
<th>Mechanism</th>
<th>Number of term banks using method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Keywords (KW)</td>
<td>6</td>
</tr>
<tr>
<td>KW + relationships</td>
<td>1</td>
</tr>
<tr>
<td>KW + subject field</td>
<td>1</td>
</tr>
<tr>
<td>KW + pre- and post-coordinated terms</td>
<td>1</td>
</tr>
<tr>
<td>Abbreviations</td>
<td>1</td>
</tr>
<tr>
<td>Boolean operators (and/or/not)</td>
<td>1</td>
</tr>
<tr>
<td>First word of term</td>
<td>1</td>
</tr>
<tr>
<td>Each word of term/part terms</td>
<td>3</td>
</tr>
<tr>
<td>Term minus 'noise' words</td>
<td>1</td>
</tr>
<tr>
<td>Truncation/wild card</td>
<td>2</td>
</tr>
<tr>
<td>Cross-references</td>
<td>1</td>
</tr>
<tr>
<td>Rarest word/'weightiest' word of term (to 7 letters)</td>
<td>2</td>
</tr>
<tr>
<td>All fields</td>
<td>1</td>
</tr>
<tr>
<td>All fields minus definition and bibliographic references</td>
<td>1</td>
</tr>
<tr>
<td>Free text</td>
<td>1</td>
</tr>
</tbody>
</table>

The above are not mutually exclusive and there is some overlap.

Not all term banks questioned replied to this request.
### Annex 5: Frequency of '-isation' string in French biological sciences corpus

<table>
<thead>
<tr>
<th>Rank</th>
<th>Frequency</th>
<th>String</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>192</td>
<td>l'actualisation</td>
</tr>
<tr>
<td>2</td>
<td>145</td>
<td>l'actualisation</td>
</tr>
<tr>
<td>3</td>
<td>135</td>
<td>l'actualisation</td>
</tr>
<tr>
<td>4</td>
<td>128</td>
<td>l'actualisation</td>
</tr>
<tr>
<td>5</td>
<td>123</td>
<td>l'actualisation</td>
</tr>
</tbody>
</table>

... (Continued with similar entries)
Annex 6

Example of ATA concordance showing left-hand alphabetisation

- rinderpest double recombinant vaccinia a virus in areas of the
- replication of the challenge virus. No anamnestic response at
- virus of humans, distemper virus of dogs, and pestes
- closely related to measles virus of humans, distemper virus
- the entire period of virus replication. Thorough examination revealed
- countries (Metzgar, 1985). Rinderpest virus The causative parvovirus of
- peste des petits ruminants virus of goats and sheep.
- of the vesicular stomatitis virus (Mackett et al., 1985).
- the field, and the virus does not exist in
- facilitate penetration of the virus by membrane fusion (F).
- F proteins of the these two surface proteins
- control. For a vaccinia virus control, two animals were
- with v50, a vaccinia virus recombinant expressing the glycoprotein
- more highly attenuated vaccinia virus recombinant vaccine which expresses
- the experiment for vaccinia virus antibody as indicated by
- lyophilised form of vaccinia virus is heat-stable, easily produced,
- M genes of vaccinia virus were insertionally inactivated by
- low level of vaccinia pathogenicity and provides a
- with those of vaccinia virus. Despite this wide host
- the transmissibility of vaccinia virus recombinants from vaccinated to
- wide host range, vaccinia virus does not cause any
- al., 1988). Recombinant vaccinia virus vaccines for rinderpest Vaccines
- a Wyeth strain vaccinia virus double recombinant was developed
- region of the vaccinia genome. Although any non-essential
- units) of the vaccinia recombinant vaccine were administered
- for rinderpest. The vaccinia virus recombinant vaccine would seem
- note that the vaccinia virus recombinant vaccine for rinderpest
- different viruses. The vaccinia virus system has a number
- Cattle vaccinated with vaccinia virus recombinants (both groups) were

N.B. It is possible with one key stroke to change to right-hand alphabetisation
## Annex 7

### Example of Synoptic Profile of *expresses* using ATA

<table>
<thead>
<tr>
<th>Span -3</th>
<th>Span -2</th>
<th>Span -1</th>
<th>Type</th>
<th>Span +1</th>
<th>Span +2</th>
<th>Span +3</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 Pir-46</td>
<td>2 JRS4</td>
<td>4 and</td>
<td>expresses</td>
<td>8 a</td>
<td>2 F</td>
<td>2 2</td>
</tr>
<tr>
<td>2 a</td>
<td>2 cells</td>
<td>4 that</td>
<td>expresses</td>
<td>5 the</td>
<td>2 and</td>
<td>2 F</td>
</tr>
<tr>
<td>2 generally</td>
<td>2 contain</td>
<td>4 which</td>
<td>expresses</td>
<td>2 it</td>
<td>2 functional</td>
<td>2 Pheromone</td>
</tr>
<tr>
<td>2 perhaps</td>
<td>2 encodes</td>
<td>2 constitutionally</td>
<td>expresses</td>
<td>2 pheromone</td>
<td>2 genes</td>
<td>2 Staphylococcus</td>
</tr>
<tr>
<td>2 shown</td>
<td>2 more</td>
<td>2 efficiently</td>
<td>expresses</td>
<td>2 protein</td>
<td>2 inhibitor</td>
<td>2 amino</td>
</tr>
<tr>
<td>2 the</td>
<td>2 respectively</td>
<td>2 importantly</td>
<td>expresses</td>
<td>2 recombinant</td>
<td>2 multi-functional</td>
<td>2 protein</td>
</tr>
<tr>
<td>2 vaccinia</td>
<td>2 single</td>
<td>2 pLRO49</td>
<td>expresses</td>
<td>1 CR2</td>
<td>2 stable</td>
<td>2 under</td>
</tr>
<tr>
<td>2 which</td>
<td>2 vector</td>
<td>2 plr</td>
<td>expresses</td>
<td>1 authentic</td>
<td>1 Thirty-three</td>
<td>2 urease</td>
</tr>
<tr>
<td>1 MVA</td>
<td>1 Raji</td>
<td>1 OECD</td>
<td>expresses</td>
<td>1 both</td>
<td>1 Cells</td>
<td>1 As</td>
</tr>
<tr>
<td>1 Recombinant</td>
<td>1 The</td>
<td>1 presumably</td>
<td>expresses</td>
<td>1 its</td>
<td>1 PVP</td>
<td>1 H</td>
</tr>
<tr>
<td>1 bodies</td>
<td>1 VV</td>
<td>1 simultaneously</td>
<td>expresses</td>
<td>0</td>
<td>1 fusion</td>
<td>1 HA</td>
</tr>
<tr>
<td>1 line</td>
<td>1 and</td>
<td>0</td>
<td>expresses</td>
<td>0</td>
<td>1 haemagglutinin</td>
<td>1 Proceedings</td>
</tr>
<tr>
<td>1 recombinant</td>
<td>1 developed</td>
<td>0</td>
<td>expresses</td>
<td>0</td>
<td>1 immunising</td>
<td>1 antigen</td>
</tr>
<tr>
<td>1 selection</td>
<td>1 recombinant</td>
<td>0</td>
<td>expresses</td>
<td>0</td>
<td>1 influenza</td>
<td>1 glycoprotein</td>
</tr>
<tr>
<td>1 tested</td>
<td>1 that</td>
<td>0</td>
<td>expresses</td>
<td>0</td>
<td>1 recombinant</td>
<td>1 hemagglutinin</td>
</tr>
<tr>
<td>1 virus</td>
<td>1 vaccine</td>
<td>0</td>
<td>expresses</td>
<td>0</td>
<td>1 sincere</td>
<td>1 thanks</td>
</tr>
<tr>
<td>1 was</td>
<td>1 virus</td>
<td>0</td>
<td>expresses</td>
<td>0</td>
<td>1 the</td>
<td>1 were</td>
</tr>
</tbody>
</table>

*Note: The table above represents a synoptic profile of the expression of recombinant 1 developed to express recombinant 1 influenza glycoprotein tested that was virus, using ATA.*
Annex 8 Proposed empty model of a verb frame for transitive quasi-LSP verb combinants. Because valency theory is verb-based, Slot 2 precedes Slot 1 so that text in English, for example, follows its natural order.
homonym 84, 90-1
hyperlemmatization 99, 101
hyponymy 187-8
idiom 20, 42-3, 70, 112-6, 121, 127, 129, 137, 140-1, 149, 207
indexer 38, 46-54, 76, 78, 95, 133, 135
index head 46
indexing term 57, 93, 249
information retrieval 57, 195, 223
information retrieval specialists 119, 152
(see also documentalists)
initialism 16, 23, 78, 90, 106
intension 136, 143-4, 155-6
International Standards Organization (ISO) 29, 36, 49, 51, 79, 143, 145
keyword 26, 37, 46, 57, 97, 286
Kollokator (see collocator)
KWIC (keyword-in-context) 46
KWOC (keyword-out-of-context) 46
Language for General Purposes (LGP) - adjectives 63, 75, 125 - collocation 42, 61, 118, 125, 140 - phraseology 61 - verbs 18, 40, 67, 71, 139, ch. 4, 5, 265 Language for Special Purposes (LSP) - combiners 161, 164, 228, 259 - phraseology 60, 70, 211 - term 17, 73, 77-81, 118-119, 134, 135, 142, 167, 209 - terminography (see under termography) - verb phrase (see under verb phrase) - verbs 18, 40, 67, 132, 146, 149, ch. 4, 5 Latin 46, 164, 168, 175, 267 Latvian 106 LDOCE (see Longman Dictionary of Contemporary English) legal terminology 40, 66, 170, 211, 216 lexicomorphization 224 (see also hyperlemmatization) LEXAUTOM 27 lexicalisation 78, 97 lexicographer 19, 22-3, 41, 43, 57, 77, 91, 104, 135, 154 lexicography 17, 19, 23-4, 43, 103-7, 259 lexis 16, 21, 22, 77, 128, 162, 260, 263 LEXIS (term bank) 34 LGP (see Language for General Purposes) linguists 39-40, 77, 135, 137, 149 Longman Dictionary of Applied Linguistics 112-4 Longman Dictionary of Contemporary English (LDOCE) 45, 132 Longman Dictionary of Scientific Usage 59 LSP (see Language for Special Purposes) machine-assisted translation (MAT, MT) (see dictionaries)
meronymy 187 (see also relationship: part/whole)

Index: 292
synchrony 32, 83, 189
Syncopic Profile 203, 225, 228, 289
SYSTRAN 34

target language (TL) 34, 74, 114, 175
TEAM 98, 101
term
- complex, compound 28-9, 56, ch. 2, 137, 143, 150, 209
- deprecated 30, 49, 81
- generic 110, 140, 143, 151, 158
- multiword 24, 51-3, 78, 92, 100-2, 106
- standardised 52, 76
- sub-technical 84
term bank(s) (and see under individual names) 25, 36, 73-6, 91-2, 96-108, 131, 150, 158, 261
- Canadian 97, 102
- Danish 29, 36-7, 108
- European 28, 29, 34, 176, 177
- Surrey 30
term record (format) 26, 37, 141, 284
terminology 17, 23, 41, 57, 76, 81, 91, 103, 107, 259, 281
terminological unit 37, 77-8, 87, 89, 97, 119-20, 132, 209
terminology data bank (see term bank)
TERMIUM 97, 102

themes/rheme 238-40, 260, 267
thesaurus 22-4, 49, 51-4, 60, 142-3, 152-160
- computerised entry 52, 91
troponymy 188-189
truncation 28, 101, 286
Turkish 106

Universal Decimal Classification (UDC) 30, 34

valency (see also Mitspieler)
- a- (see zero valency)
- di- 179, 183, 193
- mono- 179, 183, 193, 239
- sub- 240, 244
- tetra- 182-3
- theory 40, 181, 235, 259
- tri- 183
- zero 183, 184
verb
- auxiliary 205, 224, 264
- carrier 65, 68, 170, 179, 203, ch.5
- combinatory 62, 174
- de-adjectival 187
- denominational 146, 186, 228
- impersonal 191
- intransitive 66, 68, 134, 139, ch. 4, 246, 250-1
- nominalised 41, 52, 191, 195-6, 228, 245, 251
- passive 54, 67, 69, 71, 89, 93, ch. 4, 5
- phrasal 64, 111-2, 116, 149, 284
- prepositional 110-2, 116, 176, 191
- reflexive 66, ch. 4, 5, 246, 256-7, 268
- support (see carrier)
- transitive 66, 68, 125, 134, ch. 4, 5
voice
- passive (see passive verb)

Welsh 60

word formation 39, 62

(This index has been compiled with the help of ATA)