Empirical Observation of Term Variations  
and Principles for their Description

Béatrice Daille, Benoît Habert, Christian Jacquemin, Jean Royauté

May 26, 2000

Contents

1 Introduction  .......................................................... 2
  1.1 Do terms vary? .................................................. 2
  1.2 A Symbolic Framework for the Study of Terminological Variation .... 4

2 The Most Common Types of English Two-word Terms 7
  2.1 Adjective Noun (A N) ........................................... 8
  2.2 Noun Noun (N₂ N₁) ............................................. 9
  2.3 Noun Preposition Noun (N₁ P N₂) ............................. 9

3 Observing and Representing Term Variants 9
  3.1 An Observation of Term Variants ................................ 9
  3.2 A Two-level Lexico-syntactic Description of Terms ................. 11
  3.3 Two Families of Grammatical Rules ............................ 15

4 Variations Rules: Coordinations and Permutations 16
  4.1 An Observation of Coordination and Permutation Variations ...... 16
    4.1.1 An Observation of Coordination Variations ............... 16
    4.1.2 An Observation of Permutation Variations ............... 17
  4.2 A typology of Coordination and Permutation Variations ......... 18
  4.3 Rules for Coordinations and Permutations ........................ 19

5 Production Rules: Insertion and Juxtaposition 20
  5.1 Insertions ....................................................... 20
    5.1.1 An Observation of Insertion Variations .................... 20
    5.1.2 A Typology of Insertions ................................ 21
    5.1.3 Rules for Substitutions .................................. 23
  5.2 Juxtaposition .................................................... 23
    5.2.1 Rules for Juxtapositions ................................. 25

6 "Doing" derivations within noun phrases to generate possible term variants 26
  6.1 An Observation of Complex Variants ............................ 26
    6.1.1 Complex insertions .................................... 27
    6.1.2 Permutations composed with insertions ..................... 28
    6.1.3 Coordination composed with insertions ..................... 29
  6.2 Composing Elementary Transformations .......................... 31
7 “Undoing” derivations within noun phrases to find possible term variants 32
7.1 Overall presentation ........................................................................................................ 32
7.2 Using the grammar as a simplification device .................................................................. 33
7.3 The simplification and matching process ........................................................................ 34
7.4 The need for other mechanisms than reverse derivation ................................................ 35
7.5 Related work ................................................................................................................... 35
7.6 Assessing and classifying potential variants .................................................................... 36

8 Conclusion ......................................................................................................................... 36

1 Introduction

1.1 Do terms vary?

Among the long linguistic tradition of research about idiomaticity, there is a general agreement on the scalar dimension of this phenomenon. The centrality of idiomaticity justifies the conception of formalisms intended for representing idioms and locations such as Lexicalised Tree Adjoining Grammars [Abelle and Schabes, 1989]. Between phrases following the general rules of the syntax and frozen sequences, allowing no variation at all, there is a whole range of possibilities. Moreover, totally frozen phrases seem rather exceptional.

However, surprisingly, research in terminology and in sub-languages develops a rather different analysis. Terms, which can be thought of as ‘domain idioms’, are supposed to be labels for precise concepts within a sub-language, and, as such, not prone to variation. As is often pointed out, standards organisation, in each field, try to “freeze” the domain terminology ## [Sager, 1990]. In addition, in information retrieval, for efficiency purposes, multiword terms are often represented as fixed sequences of characters, including spaces, dots and other separators, or as finite state automata corresponding to the list of known variants. The variants of multiword terms which are not explicitly listed are just ignored. As stemming is easier to account for than syntactic variation, most of the works on term retrieval have focused on single-word terms. Similarly, the trend towards statistical natural language processing focuses the attention on dominant collocations. The variants whose number is less than the chosen threshold are ignored as well. The ‘near misses’ which are implied by the existence of variants are supposed to be rather marginal, thus not threatening really the recall score of the information retrieval systems.

Describing terms as fixed sequences is obviously an idealised viewpoint. Terms are subject to different kinds of changes. # Given a list of controlled terms, the observation of technical corpora shows that variations can, be roughly into morphological and syntactical variations. Example (11) in table 1 is a morphological variant of epithelial cell and example (1) is a syntactic variant. In this study, we focus on syntactic variations which are based on dependency between head words and arguments. Syntactic variations can be further divided into variations which preserve concepts and variations which involve a shift of meaning. For example, a term and its permutation variant such as blood cell and cell of blood refer to the same concept while a term and its substitution variant such as blood cell and blood mononuclear cell refer to different concepts (the modified term denotes a more specific concept).#

A First Empirical Observation of Term Variants: Some “Epithelial Cells” from a Medical Corpus

In order to illustrate the diversity of # morpho-syntactic variation, we present first the variants of a medical term within PASCAL a 1.5 million-word medical corpus. The term chosen is Epithelial cell which is part of the indexing terminological lexicon of INIST (Institut de l’Information Scientifique et Technique). # This term is found in a corpus of medical abstracts under the 15 forms of Table 1.

The term Epithelial cell has no occurrence under its base-form over the whole corpus which is exceptional. Out of the 15 variant forms, only Epithelial tumor cell is retrieved more than once (three if one accepts that the spelling variant tumor for tumour may be detected automatically). It is therefore difficult to refer to the thirteen other variant forms as candidate terms on the basis of a single attestation in the corpus.
| (1) | epithelial tumor cells |
| (2) | epithelial tumour cells |
| (3) | epithelial cancer cell |
| (4) | epithelial respiratory cell |
| (5) | epithelial cultured cell |
| (6) | epithelial ) cell |
| (7) | epithelial and carcinoma cells |
| (8) | epithelial and myoepithelial cells |
| (9) | epithelial and gial cells |
| (10) | cells expressed epithelial |
| (11) | epithelium specific cell |
| (12) | cell infiltrate of the epithelium |
| (13) | cell associated with breast epithelium |
| (14) | cell carcinoma contained dysplastic epithelium |
| (15) | cells were identical tononpigmented ciliary epithelium |

Table 1: Some variants of “epithelial cell” from a medical corpus

In the fifteen examples above, there are four large variation families: syntactic variations of insertion, of permutation and of coordination as well as morpho-derivational variations with different degrees of acceptability.

- Examples (1) to (5) concern the insertion variation. The following words: tumor, tumour, cancer, respiratory, cultured as well as the closing parenthesis are inserted elements.

- The coordination variation is illustrated by examples (6) to (8). Forms which are coordinated to Epithelial cell are: carcinoma cell (7), myoepithelial cell (8) and gial cell (9).

- Example (10) shows how the word of a term can permute around a single “pivot element” (for example the verb expressed). Most of the permutations take place around prepositions. We will only study these permutations in the remainder of the paper.

- Examples (10) to (15) are morpho-derivational variations. The adjective epithelial can be derived from the noun epithelium which in a classical morpho-derivational scheme links epithelial cell to the two forms (a) epithelium cell and (b) cell(s) of the epithelium. The derivational scheme (a) is composed with an adjectival modification (11): the adjective specific is the inserted modifier. The derivational scheme (b) which involves a permutation is illustrated by examples (12) to (15). Permutation only takes place around the preposition of with example (12). # For examples 13 to 15, the following verbal sequences are the “pivot elements” of permutation: associated with (13), contains (14) and were (15).

These 15 variant forms of the term Epithelial cell show the complexity of morpho-syntactic phenomena involved in terminological variation. This paper shows that technical terms do vary, more than expected or predicted. It uses large technical corpora to demonstrate the amount of term variation which occur in sub-languages, and to # classify it, on syntactic grounds. It distinguishes variations yielding candidate terms such as example (1) from variations which are not alike to produce new term such as coordination (9).

Section 2 is a presentation of the different term structures which are observed in English terminologies. Sections 3, 4 and 5 present variations of terms occurring in a English corpora and show empirically the different families of variations which are observed. The variants are separated into two families: variants which do not modify the string of the original term and variants which break term strings through word insertions or permutations of word order. The latter family plays a key role in the identification of term variant occurrences and must be accounted for in NLP systems aiming to extracting or retrieving terms from a corpus. In order to introduce the kernel of rules that will be shared by such systems, ## a greater stress is laid on variations which alter the string of the original terms. They are presented as transformations
rules operating on terms presented in parts 4.3 and 5.1.3. The actual use of such rules in a purpose of generating the different possibilities of variants is demonstrated in section 6. Section 7 proposes a method to relate noun phrases retrieved by a parser with controlled terms, so as to propose new variants of terms.

1.2 A Symbolic Framework for the Study of Terminological Variation

Some tools intended for compound and term extraction such as *INTEX* presented in [Silberstein, 1994] have no concern for the problem of variation. However, Section 1.1 has convinced us that variation is a real problem for an exhaustive identification of term occurrences. This study aims at giving a good picture of the phenomenon through its observation in corpora and at providing the reader with a formal toolbox. The tools consist of rules which can serve as a basis for the implementation of a system aiming at analysing term variants within texts.

As a starting point for this study, let us give a workable definition of variants: a variant of a term is an utterance which is semantically and conceptually related to an original term. Section 3 shows that language offers a large variety of resources for expressing conceptually related constructions. We leave apart the problem of inflectional modifications of terms and focus on the variations which cannot be described through this sole consideration. A first classification of variations separates, on the one side, variants where content words and their dependency relations are preserved such as examples (1)-(9) from Table 1 and, on the other side, variations such as (11)-(15) which involve a derivational transformation of at least one of the words without modifying the thematic relations. These variations include, of course, nominalisation, adjectivisation, but also acronymisation, abbreviations…

The different kinds of variations play a key role in automatic terminology processing. In a purpose of acquisition, the statistical approaches to collocate extraction such as [Sinadji, 1993] or [Church and Hanks, 1989] use probabilistic measures of the strength of the lexical dependency between two words. They assume that two words of a collocation co-occur within a window of arbitrary size more frequently than they would do just by chance. In such studies no assumption is made about the kind of syntactic links that hold between the words that co-occur within a limited textual span. This ignorance is damaging to a precise study of collocates in corpora. Is it really acceptable to consider lead to a higher nitrate as an occurrence of lead nitrate and phase and reduced with time as a variant of phase time? Only a part of such collocates are good candidate terms. Applications aiming at improving the results of a statistical filtering must use a prior or a posterior syntactic and/or semantic filtering of correct sequences [Daille, 1994]. Besides the lack of precision of pure statistical methods, the use of a restricted window for the observations prevent such method to extract variants spanning over large sequences of words such as multiple coordinates.

Due to the necessity of a precise and exhaustive description, terms and their variants observed in corpora will be represented in a symbolic framework throughout this paper. This description can be applied to NLP analysis purposes such as automatic indexing, text categorisation or text (pre-)processing and to generation purposes such as machine translation or automatic abstracting.

Thus, any symbolic approach to term extraction from corpora has to account for variation and a great care must be taken about the distinction between correct and spurious variants. A sequence of words including the components of a term is a potentially correct variant (i.e. a non-incorrect variant) if it corresponds to one of the acceptable syntactic trees describing the local variations of a term. The structural rules proposed here for the description of variation are intended for corpus parsing. They would fail at producing correct variants in a generation task. A positive description of term variation would be a longer study where, in particular, semantic information would have to be accounted for. For example, a *surgical exploration and closure* is a correct (and attested) variant of *surgical closure*. But *surgical shock and closure* is not acceptable because both terms do not refer to the same semantic categories: *surgical shock* is a surgical "pathology" when *surgical closure* is a surgical act. Positive rules for generation would necessitate a minimal semantic information which could be used for checking that the constructed variants are correct. Such an information is extracted from thesauri with semantic features or can be acquired from observation of variants encountered in large corpora. The amount of knowledge required is vast and is a topic of study on its own.

On the one hand, filters selecting correct variants can be intuitively devised and confirmed through a heuristic tuning by the observation of different types of variations in a large corpus. On the other hand, they can also be linguistically motivated through a syntactical description of term variants and a translation
of these correct trees into sequences of words or part-of-speech categories. In this study, we have chosen to approach the problem of variation from these two opposite directions: experimental observation and linguistic grounding. Given a window including the words composing a term, is there a correct syntactic variation of the term matching this sequence?

**Terms as Complex Lexical Entries**

As will be outlined in section 3.2, terms can be felicitously considered as complex lexical entries with a syntactic structure and represented as lexicalised phrases, that is to say phrases whose leaves are lexical items. The interaction between the local syntax and terms results from a fine-grained description of the transformations performed on terms by syntactic rules. These rules take as input one or more terms (for example cell metabolism and cell proliferation) and yield as output a new structure # which includes all the words of the input term(s) (for example cell metabolism and proliferation), see Figure 1 for examples. The transformations are local and contextual. # In order to simplify the presentation, the span of the variation is bordered by two words of one of the input terms. As we are mainly concerned about identification tasks, we only give the rules associated to the transformations which alter the string of at least one of the input terms. In the preceding transformation, the string of the base-term cell proliferation is modified by an insertion. Hence, the occurrence cell metabolism and proliferation cannot be related to cell proliferation only through stemming and string matching but requires the model to accept # two inserted words. #

![Figure 1: Three examples of transformation rules](image)

- **Base-Term**
  - `internal carotid`
  - `external carotid`
  - `tumor cell`
  - `flow measurement`
- **Variant**
  - `internal or external carotid`
  - `tumor or AN cell`
  - `(e.g. tumor or nontumorous cell)`
  - `measurement of flow`
The result of the application of the rules to term patterns and their eventual composition motivate the different filters used for:

- rejecting spurious variants (for example satellite and whole chromosome is not a correct variant of satellite chromosome),
- selecting potential candidate terms among the variants (for example labial salivary gland is a variant of labial gland which may be a candidate term),
- extracting candidate terms from the variants which may possibly involve another term (for example bone and air conduction is a variant of bone conduction which reveals the potential candidate term air conduction).

The first point concerns the extraction of term occurrences from corpora and applies to automatic indexing and text retrieval. Such rules have been applied to a medical corpus to retrieve the occurrences presented in this study. The last two points provide guidelines for a precise terminological acquisition. These applications motivate the description proposed in this study. The results of the experiments on corpora which are presented in the paper have been evaluated in [Jacquemin & Royauté, 1994] in a purpose of information retrieval.

**Base-Terms, Variants and Candidates**

Terms can be considered as a formal system composed of a core set of basic elements, the controlled terms, and an extended set of derived elements constructed through transformation rules. Our aim is to provide the reader with a set of linguistic filters selecting the constructions which are likely to be correct term variants. In other words, the constructions which are not accepted by one of these filters are, with a high probability, incorrect variants.

Before examining transformation rules that applied on controlled terms, we introduce the so-called “controlled terms”. These terms have been pointed out after a linguistic and quantitative investigation of a corpus in the domain of communications. It is presented in section 2. If we decide to characterise them thanks to their length — the length of a term is defined as the number of main items it contains1 — it appears that binary terms are by far the most frequent ones. This observation is confirmed by [Nkenti-Azeh, 1992] who reports that only 5% of the set of terms in satellite communication domain share a length greater than 3. Terms of four and more words are marginal. Moreover, most of ternary terms are created recursively from unary or binary terms and fit to the view of the binary production of terminological data.

Three main operations apply to an unary or binary term and lead to a ternary term: insertions and juxtapositions which are subdivided into “over-compositions” and modifications. The production rules applying to unary or binary terms to yield a ternary term are also variation rules. Given an unary or binary term and a production rule, a non-expert of the domain is not able to decide whether the complex nominal phrase obtained is a genuine term or just a term variant. As in section 3, production rules are shown to be a subset of variation rules, the structures produced by transformation and/or production rules are called term variants. All these occurrences, when extracted from a corpus, correspond to variants of initial terms, some of them may acquire the status of new terms.

The specific term structures which are alike to produce new term are called candidate terms. For example, in the medical domain, expression of various genes is a variant of gene expression which is not a candidate term. Conversely, autoimmune heart disease is a transformation through insertion of autoimmune disease and a good candidate term: it is even a valid term of the terminological database.

We henceforth suppose that terms are divided into two families: base-terms and term variants. Base-terms are mainly binary terms (see section 2) and term variants are terms derived from existing terms through insertion, juxtaposition, permutation and coordination. They are further divided into variant-candidate terms and variant-non-candidate terms depending on their syntactic structure and the type of rule that has been applied to the base-term, i.e. production rule or not. These families are illustrated by Figure 2. Section 5 gives a linguistic grounding to the fact that most of the candidate terms correspond

---

1 *Main items* are nouns, adjectives, adverbs, etc. Neither prepositions nor determiners are main items
to the variants constructed through insertions or juxtaposition. The choice of genuine new terms among candidate terms relies on a terminologist who has some competence in the related domain. To sum up, the following three restrictions apply to these sets of terms and their transformation rules:

1. base-terms are unary or binary terms such as leukemia or blood cell,
2. the transformation of any term variant involves at least one base-term or one term variant: for example circumflex coronary artery is constructed from both base-terms circumflex artery and coronary artery,
3. some of the syntactic rules constructing variants can yield candidate terms as well, conversely some rules generating syntactic variants such as coordination (cell metabolism and proliferation) rarely produce candidate terms—in the first case, the construction mechanism is called over-composition for its analogy with compounding, and belongs to the sub-set of production rules, in the second case, coordination belongs to variation rules. Both of them can apply to the same term structures.

The above definition of the different classes of terms is a recursive one because variants are defined partly from base-terms, partly from variants. However this possibly infinite recursion is restricted to a small finite number of iterations as indicated in section 6.

The descriptive model is associated to the observations in corpora and to their linguistic in the following three sections. Firstly, the following section 2 illustrate term variant observations through representations which rely on a structural description of terms. In this aim, the different kinds of frequent term structures observed for the English language are presented. Then, in part 3, the different kinds of variations are reviewed together with the model which will be used to describe term variation. It consists of a two-level syntactic description of base-terms and a set of rules operating on these terms. The necessity of a two-level description is motivated in part 3.2 by considerations in linguistics and terminology. The rules yield either candidate terms or term variants depending on the type of the rule (restriction 3) and on whether the transformation of the original term by the rule denotes a concept in the domain. Therefore, we will study separately variations which yield syntagmatic constructions (coordinations and permutations) in part 4 and variations which produce candidate terms in part 5 (insertions and juxtaposition). A set of associated transformation rules for internal variants are described in sections 4.3 and in 5.1.3.

2 The Most Common Types of English Two-word Terms

Before studying the different kinds of variations, it is necessary to give a picture of the different term structures. Judging from data in a terminological bank and in corpora of technical vocabulary, it appears that the majority of terms do consist of more than one word; among them the overwhelming majority are noun phrases involving two main items. # #
<table>
<thead>
<tr>
<th>Term length (in number of main items)</th>
<th>Frequencies</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>101</td>
</tr>
<tr>
<td>2</td>
<td>336</td>
</tr>
<tr>
<td>3</td>
<td>98</td>
</tr>
<tr>
<td>4 and more</td>
<td>65</td>
</tr>
</tbody>
</table>

Table 2: Frequencies of terms of different lengths in terminology bank

<table>
<thead>
<tr>
<th>Term length (in number of main items)</th>
<th>Number of terms</th>
<th>Frequencies</th>
<th>Frequency average</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>4 771</td>
<td>13 569</td>
<td>2.84</td>
</tr>
<tr>
<td>3</td>
<td>1 045</td>
<td>1 587</td>
<td>1.52</td>
</tr>
<tr>
<td>4 and more</td>
<td>168</td>
<td>203</td>
<td>1.02</td>
</tr>
<tr>
<td>Total</td>
<td>5 984</td>
<td>15 359</td>
<td>2.57</td>
</tr>
</tbody>
</table>

Table 3: Frequencies of multiword terms in corpora

First, we selected a terminological data bank: the one of the EEC, telecommunication section, which has been elaborated by experts and extracted at random 600 terms. Noun phrases constitute 500 of the 600 terms, i.e. 98.3% of the terms. Of the 10 non-NPs among these 600 terms, 4 are adjectives (adaptable), 3 are past-participles (charged), 1 is an adverb (quadratically), 1 is a verb (to reconfigure) and 1 is a prepositional noun phrases (on board). This first observation confirms that terms are mainly multiword units of nominal type. Then, extracting 10 more terms of nominal type to reach the set of terms to 600, we examine their length. This simple investigation (summarised in table 2) shows that 56% out of these 600 terms are two-words terms (83% of them have more than one word).

Within a terminological bank, it is difficult to guess the representativity of a term: a term that appears in the bank could appear only once in a corpus of the same domain. In order to evaluate the frequency of multiword terms inside technical corpora, we used the program of terminology extraction developed by [Daille, 1994] and applied it to a corpus of telecommunication of 200 000 words: The Satellite Communication Handbook. This program extracted nominal phrases whose morpho-syntax is characteristic of multiword terms form corpora. Of course, not all these nominal phrases are terms, but 85% of them have been estimated as such. The results, summarised in table 3, show that two-word terms are by far the most represented terms and, moreover, they are frequently used compare to terms of length 3 and greater than 3. Another numerical observation shows that 61% of term of length 2 appear only once in the corpus, 80% of terms of length 3 and 89% of terms of length greater than 3.

Two-word terms are by far the most frequent ones: we will call them base-terms. These base-terms could be classified following their morpho-syntactic structure such as A N, N2 N1 and N1 P N2 as presented in [Lieber, 1983] or [Nicolas, 1992].

2.1 Adjective Noun (A N)

These constructions are frequent in the database and in the corpus, being one of the most productive term types. Adjectives are invariant regardless of number, and gender of the noun they modify with a few exception which concern foreign adjectives such as blond/blonde. The determiner, the noun, freely varies in number if it is a count-noun and stays invariant if it is a non-count noun. Examples include: multiple access, solar cell(s), thermal control, and intermediate frequency, intermediate frequencies (IF)

The Adjective Noun class encompasses other Modifier Noun structures such as Present-participle Noun, Gerund Noun and Past-participle Noun which are quite rare:

1. Present-participle Noun and Gerund Noun
   modulating signal, receiving station, signalling system(s).
2. Past-participle Noun

*co-polarized component(s), gridded reflector(s), shaped beam(s).*

Adjective Gerund constructions also belong to this class of structures: *decadic signaling, component coding,* or *syllabic compounding.*

2.2 Noun Noun (N₂ N₁)

The most common type of terms of length 2 in English is composed of endocentric compounds. In such structures, the determiner, the head noun, follows the determiner, in conformity with to the Germanic composition as in *communication satellite* where *communication* determines *satellite.* Generally the determiner is not inflected (i.e. singular) although a few exceptions exist. The second noun, which is always the syntactic head (except for 2-head terms such as *input-output*) may be inflected or may be invariant.

Examples are: *data transmission(s), telephone channel(s), satellite communication(s), earth station(s), radio regulations, communication(s) equipment, radio communications/radio communications, radio relay(s)/radio-relay(s).*

We have included in this section, terms of Noun Gerund structure because form ending with “ing” are generally ambiguously Noun or Gerund: *transponder hopping, compression buffering, data processing, automatic tracking, bandwidth broadening, frequency raising.*

2.3 Noun Preposition Noun (N₁ P N₂)

There are really few terms with this structure (as well as non technical compound as stated by [Tournier, 1985]). To give an element of comparison, in our sample of terms extracted from the terminological bank, only 3 two-words terms have this structure (*assignment by reservation, line of sight and mass in orbit*). In our corpus of 200,000 words, we picked up 4,300 noun phrases # appearing at least twice. 1,817 NPs have a N₂ N₁ structure, and 2,483 have an A N structure. Only 530 NPs with a N₁ P N₂ structure have been encountered.

Among the different prepositions, the preposition *of* is the most employed as it appears in 360 noun phrases. The N₁ of N₂ structure can be composed with other variations to produce other variants such as N₁ of A N₂ structures (see section 6). # However, these variants are not good candidate terms [Chuquet and Paillard, 1989]. Similarly, the adjectival modification of terms with a N₂ N₁ compound structure produces a new term with a A N₂ N₁ structure. For example, the modification of *programme exchange* by the adjective *international* leads either to a candidate term *international programme exchange* or to a complex variant *programme of international exchange.*

Generally, an occurrence with a N₁ of N₂ structure is either an inverted form of N₂ N₁ and thus is considered as a variant (see section 4), or a noun phrase which does not represent a terminological concept such as *type of antenna* where *type of* is a classifier, *number of satellites* where *number of* is a quantifier. A few of the constructions N₁ of N₂ are terms, such as the following examples: *figure of merit, mode of operation, order of magnitude.*

Concerning other prepositions, such as *in, on,* examples of terms are rarities, at least in our corpus. The following examples are such exceptions: *carrier per transponder, # satellite(s) in orbit, transmission via satellite.*

Two-word terms are the most frequent ones and we point out in section 3.3 that the majority of terms of length greater than 2 are created recursively from base-terms thanks to production rules.

3 Observing and Representing Term Variants

3.1 An Observation of Term Variants

In order to stress the importance of variation, we have carried out two experiments: one using a corpus of medical bibliographic abstracts and the other with a set of physical bibliographic abstracts. This first sketch of the observations presents all the categories of base-terms variants observed. Then, in parts 4
and 5 we will restrict the observation to morpho-syntactic variants which constitute the core set of term variants.

The experiment on medicine dealt with a large corpus of 10 megabytes (1.5 million of words). The lexicon used is the PASCAL multidisciplinary lexicon which gathers 80,000 multiword and single word terms. With this corpus we identified 36,572 occurrences of terms of more than two words (pluritersms) among which 5,053 are terms subject to variation, which represents 13% compared to the total pluritersms. The distribution of the variations is the following: insertion 51.8%, permutation (with or without insertion) 38.2%, and coordination 10%. ##

In the experiment on physics, the study on variations was not the main objective. The idea was to obtain a useful tool to recover the greatest number of terms describing the content of documents. We mainly sought to link a computerised linguistic tool to an infometrical classification tool so as to develop a system for technological watch capable of working directly on the text itself (abstracts in this case). For this reason the experiment concerned a smaller corpus: # 519 bibliographic abstracts covering three physics periodicals. We used a specialised thesaurus of 21 155 terms (FIZ thesaurus, Fachinformationszentrum of Karlsruhe). With this specialised nomenclature, we identified 1,346 pluritersms of which 872 are terms which are recognised under variant forms, which represents a variation rate of 35%. These results tend to prove that results are closely linked to the degree of specialisation of the lexicon used. These results will have to be confirmed over larger corpora.

Experimentation on corpora helps in the writing of term variation rules through the observation of encountered variants but it cannot provide a secure methodology for conceiving a grammar. In order to avoid the multiplication of heuristic-based grammatical rules without clear linguistic motivation, we have found it necessary to provide the reader with a comprehensive description of a local terminological syntax which accounts for variants and embedded terminological constructions (3.2 and 3.3). In 6, these rules are applied to base-term patterns in order to generate transformed term patterns. These new patterns are used for selecting variants and candidate terms among collocates (words co-occurring inside a window).

We identify three categories of variations: 1) inflectional variation 2) syntactic variation and 3) morpho-derivational variation. Each of these variations leads to a particular problem for term recognition. We focus on syntactic variation which is very productive. The phenomena of derivational morphology are given as examples and are not processed as such. They are still to be studied.

Variants of base-terms are classified under the following categories: graphical and orthographic variants (including abbreviations), inflectional variants, syntactic variants and morpho-syntactic variants. ##

**Graphical and orthographic variants**

Graphical variants include for all kind of base terms: the use or not of capitalised letter (Domestic service or domestic service), the use of punctuation characters around a part of the base term or around the base-term itself (axial radio, “basic” parameter, [feed geometry]), # the presence of an hyphen (packet mode or paquet-mode), or the introduction of the possessive marker ‘s (children hospital or children’s hospital). #

Orthographic variants concern N2 N1 structure. For this structure, N2 is generally singular. However, this rule accepts exceptions as [Quirk and Greenbaum, 1973] have shown it. We have, indeed, encountered variants of this type in a technical corpus such as: satellite network, satellites network.

Abbreviations confirm the terminological character of the noun phrase. Most of the time, abbreviations are introduced for terms whose length are greater than 2. However, some of them are introduced for base-terms as: amplitude modulation - AM, aperture area - AE, axial ratio - AR, delta modulation - DM, energy per bit - EB, laser diode - LD, plenary assembly - PA. The abbreviation appears in the text inside parenthesis most of the time after the base-term (amplitude modulation (AM)); it is however possible to encounter the abbreviation inside the base term (electromagnetic (EM) power).

**Inflectional variations**

They allow to identify, for each term, the singular and plural forms of the nouns and their infinitive, past participle and gerund forms. For an optimal identification of the terms, it is necessary to be able to pinpoint each word of a term under its different inflected forms or to divide a word into lemma or root and ending. To each word category, there is a corresponding lemma with its various endings. It is therefore possible to
recognise the term *Vitamin deficiency* with the following textual sequences: *vitamin deficiencies*; similarly, the term *Acoustic test* can be found in the form *acoustic testing* by linking the infinitive form of the verb (which is identical in this case to the noun) to its gerund form.

**Syntactic variations**

There are three types of syntactic variations:

- **insertion and juxtaposition variations** concern each word which is not a grammatical word within a noun phrase. For example, *two dimensional tryptic peptide analysis* is associated with the term *two dimensional analysis*.

- **coordination variations** concern all coordinated forms of words (adjectives or nouns) within a noun phrase. For example, *Systolic and diastolic blood pressure* is associated with the term *systolic pressure*.

- **permutation variations** involve all words or groups of words capable of permuting around a pivot element (prepositions or verbal sequences). For example, *elements of the connective tissue matrix* is associated with the term *matrix element*.

**Morpho-syntactic Variations**

The morpho-syntactic variation integrates derivation morphological morphology in the terminological field: nominalisation of adjectives (*abnormal, abnormally*) nominalisation of verbs (*measure, measurement*), "adjectivisation" of nouns (*embryo, embryonic*). Each of these morphological derivation couples are linked to the related forms of the terms.

- Adjective nominalisation connects two base-terms of different structures, such as nominal textual sequences such as A N and N₂ N₁ (### *enzyme activities and Enzymatic activity* or N₂ N₁ and N₁ of N₂ (*abnormalities of chromosome and Abnormal chromosome*). For the latter example, since the heads are not the same in the text as in the term there is no identity of meaning but a related meaning.

- Verb nominalisation: "*promotes degradation of the cellular tumor...*” is associated with *Tumor promotion*.

###

**Focus of the study**

This paper studies more specifically syntactic variants assuming that we can rely on a correct inflectional description of terms. The syntactic variants are divided into two families: insertions, the variations which may yield a candidate term and coordinations and permutations, the variations which construct a noun phrase but which only exceptionally produce novel terms. We have included in this description of variations, juxtaposition which is added to the insertion family. Juxtapositions do not affect the structure of the base-term but often lead to the creation of new terms.

The experimental observations are modelled through a formal representation of terms and a set of rules transforming these terms into attested variants. The first step of this description consists of the following structural representation of terms.

### 3.2 A Two-level Lexico-syntactic Description of Terms

The duality of the model outlined by Figure 2 has to be mirrored by a dual representational system. For example, the term *coronary artery* must receive a double structural representation to account for its

---

2We are grateful to Philippe Barbaud for his extensive comments on a draft version of this paper. We are strongly indebted to him and many arguments in the discussion about the relevance of a two-level representation are his own.
idiomatic structure (a compound) and its syntagmatic structure (a compositional construction). The lexical
description of terms must embody their syntactic structure. Therefore term representations are syntactic
trees whose leaves are only composed of lexical items and whose root label has to be distinguished from
the noun phrase label. For example, the term coronary artery corresponds to the two trees of Figure 3.
The graphic representations given in this figure are noted as (1) for the compositional structure and as (2)
for the idiom:

\[
\begin{align*}
N, [\text{coronary}] [N \text{ artery}] & \quad \text{[sem: [is – a: artery, quality: coronary]]} \\
N, [\text{coronary}] [N \text{ artery}] & \quad \text{[sem: kind of: artery, property: coronary]}
\end{align*}
\]

The preceding formula (1) The tags of the root and the nodes correspond to the syntactic category of the
words. The level 0 of the syntactic categories such as A is the lexical level. The level 1, corresponding to
the X' categories, is the level of compounding and terms. Finally, the level 2, noted X'' (in formulae (3)
and (4), for example) corresponds to the syntagmatic level, mainly noun or adjective phrases.

The features [sem : [is – a: artery, quality : coronary]] attached to the root in Figure 3 exemplify the
kind of information that can be added to a term in order to insert it inside a thesaurus. They also denote
the semantic difference between the interpretation of a compositional term which belongs to a hierarchy
through its is-a feature and the holistic – and less transparent – interpretation of an idiom with a kind of
characterisation. As stated by [Levrat & Sabah, 1983], the link through an is-a relation is associated to a
quality in the compositional mode while the classification kind of constructed items is naturally associated
to a property which differentiates the elements of its class. 3

These features can also be used to select the syntactic transformations that are likely to operate on
a given term. For example, the semantic compatibility of the arguments can be checked to allow the
application of a coordination transformation. With a precise semantic description of the arguments of
the noun duct in the terms pancreatic duct, bile duct and electric duct, the coordination of the first two
terms would be allowed and produce pancreatic and bile duct but the coordination of pancreatic duct and
electric duct would not be accepted. The rejection of the incorrect construction comes from the difference
in the kind of feature of these terms. Both pancreatic and bile duct are parts of the human/animal body
while electric duct is a physical object. Both have the common feature to transmit a physical entity, but
this characteristic is too loose to allow a coordinated structure in any case. Such a precise description is
outside the scope of this study. In a purpose of term recognition, too permissive term transformations are
accepted, the number encountered occurrences will reduce the set of the possible combinations. In case of
a generation task, the semantics of the terms would have to be accounted for more precisely. Thus, in the
rest of this paper, we will not add any semantic feature to the term description but keep in mind that this
possibility is fully compatible with the proposed formalism and could be used for appropriate tasks.

The two-level description of terms proposed in this section is motivated by considerations in theoretical
linguistics, terminology and computational linguistics:

— **Theoretical linguistics.** The work on compounding in theoretical linguistics concerns terminology
because compounding is one the linguistic features of terms. The syntactic structure of compounds plays
an important role both in their possible syntactic variations and in their interpretation. For example,
[Selkirk, 1982] relates the interpretation of a deverbal compound such as truck-driver to the associated
verbal construction [to] drive [a] truck. More important for our study is the existence of syntactic vari-
ants involving internal transformations of term structures which call for a syntactic description of terms.
However, the specificity of this internal syntax differentiates terms from regular noun phrases and needs an
intermediate representation level between the lexical and the syntagmatic one. In the generative framework
of Government and Binding [Chomsky, 1981], [Barbaud, 1994] proposes to represent compounds at the X'
level. The main characteristic of this work is to offer an approach to compounding outside derivational
morphology contrary to the classical generative grammar or the traditional grammar. The syntax of com-
ounds and therefore of terms is asserted as a crucial component of their description. Through numerous

3 A transformation such as the one presented hereafter for term variations could be used to allow the correspondence between
both forms. Such a transformation is necessary to analyse a term when it is involved in compositional transformations where
its idiomatic structure is no longer motivated.
examples of different compound structures, [Barbaud, 1994] shows that his model is appropriate for the description of compounding.

Figure 3 depicts a two-word base-term with a $\left[ N, A \right]$ structure. This structure can be embedded within a noun phrase construction by adding a determiner and a modifier, for example. Thus, the left coronary artery is a noun phrase with the structure:

$$\left[ N', \text{[specific the]} \left[ A_N' \text{[\text{\lambda left}]} \left[ N', \text{[\lambda coronary]} \left[ N, \text{[\text{\lambda artery}]]\right]\right]\right]\right]\right]$$

This kind of construction shows the continuity between the local structure of a term and an embedding noun phrase structure. In the case where coronary artery is not a term or a compound, the left coronary artery is analyzed as a compositional construction with the following structure:

$$\left[ N', \text{[specific the]} \left[ A_N' \text{[\text{\lambda left}]} \left[ \lambda \text{coronary} \right] \left[ N, \text{[\text{\lambda artery}]]\right]\right]\right]\right]$$

# Equations (3) and (4) not only have two different syntactic structures. In equation (3), the head of the construction is the whole idiom coronary artery. Conversely, in the equation (4), artery is the head of the phrase, left and coronary are its two modifiers. A term utterance has a lexico-syntactic structure and a fine theory of variation must consider the difference between variants involving a syntagmatic compositional constructions such as equation (4) and variants involving an idiom with a restricted set of modifiers such as (3). Such a theory of # variation must make use of semantic features and selectional restrictions in order to discard compositional variants and isolate genuine ones.

— **Terminology.** According to [Sager, 1990], terms are symbols which represent the concepts of a domain of knowledge. Concept formation results from the tendency to group material and immaterial objects into classes. # In order to facilitate the communication, there is a tendency to normalise terms so that the same concept is referred to by a unique linguistic expression. This tendency is simultaneously spontaneous and encouraged by institutions for terminological standardisation which list acceptable variations. When writing technical or scientific texts, an author who wants to formulate ideas into pre-existing concepts may face two difficulties:

- either the notion that she/he wants to express is slightly different from the concepts denoted by the terms she/he knows,
- or, although knowing a terminological expression of a precise concept, she/he intentionally wishes to express a slight shift of meaning of this concept.
Such circumstances lead to the expression of term variants: either different linguistic utterances referring to the same concept or different terms which correspond to separate meanings although stemming from a unique initial concept. The variants of the first category may be called synonyms in information retrieval because such utterances are conflated into the same descriptors for indexing purposes. Conversely, the second category corresponds to genuine variants and should yield separate indexes. The conceptual identity of a term and its syntactic variants should be represented by a unique lexical multiword element (the core term) and its syntactic modifications. This is illustrated by formula (3) where left coronary artery and coronary artery both share coronary artery as lexical unit. Both occurrences are linked to the same concept and their linguistic analysis is located at the second — syntagmatic — level. The analysis of a true variant should call for different lexical units: They may be different through they structure — compare equations (3) and (4) — or different through their associated meanings — compare equations (1) and (2). This conceptual distinction between linguistic elements is another motivation for a two-level representation. 4 #

— Computational linguistics. Only few unification-based formalisms have really tackled the problem of the representation of idioms and complex lexical entries. 5 The unification-based formalism with a concern for idiomaticity are roughly divided into two families: the lexicalist and the syntactic ones. The lexicalist approaches tend to consider compounds and more generally idioms as a deviant behaviour of some single words which are idiosyncratically associated with some other specific single words to construct a fixed syntactic structure with a specific meaning. For example, in Head Driven Phrase Structure Grammar [Pollard and Sag, 1987], some lexical items may have specific features to build compounds with other items having compatible features. Conversely, the syntactic approach to idiomaticity considers idioms as pieces of syntactic trees which accept syntactic transformations. [Abellé and Schabes, 1989] have used the Lexicalised Tree Adjoining Grammar formalism to represent verbal locations as unsaturated trees accepting modifiers and substructures through the native operations of this grammar: adjunction and substitution. This second approach fits the view of [Barbaut, 1994] that terms are chunks of syntactic structure accepting a restricted range of transformations. In the same mind, OLMES [Harbert, 1991] or FASTR [Jacquemin, 1994] are tools used to represent and parse locations which rely on a grammar where rules represent terms. In the latter work, a mechanism of metarules operate on the rules and yields new rules standing for term variants.

A major issue of our linguistic approach to terminology is to sketch out a description of terms which is different from the description of compounds and idioms. Due to their conceptual role in language, terms require a specific bivalent representation with syntactic and semantic constraints different from those of idioms. The computational models pointed out for the descriptions of compounds such as [Silberztein, 1994] or [Abellé and Schabes, 1989] do not account for this distinction which is, to our opinion, necessary.

Whatever the tool used to describe the syntactic transformations of terms, it is necessary to study precisely the regular mechanisms operating on terms and modifying their structure to construct variants and candidate terms. Due to the converging point of view on the lexico-grammatical status of terms, they will henceforth be represented as lexicalised syntactic tree. The precise mechanism of these transformations will be left in the dark, by only formulating them through their associated input/output. The stress is now laid on the description of the different kinds of transformations (part 3.3).

4 We are grateful to J. C. Sager for his very precise and fruitful remarks on the link between variation and conceptual denotation. This passage has been significantly inspired by his remarks. However, he should not be taken for responsible of the remaining error.

5 We intentionally ignore the Natural Language Processing approaches to Information Retrieval where terms are extracted through a general parsing method based on a grammar of dependency relations such as [Metzler and Haas, 1989] or [Strzalkowski, 1994]. These methods are essentially oriented towards query processing by matching occurrences with similar dependency relations. These authors are not really concerned with the extraction of descriptor occurrences because their systems do not include any representation of such descriptors. Such systems could be employed in acquisition tasks by extracting occurrences of maximal noun phrases (cf. also [Duille, 1994] or [Bourigaut, 1993] for studies on term acquisition in French). We call these approaches to term variation reducing approaches because general dependency relations are restricted to terms and term variants. Symmetrically, our view of variation is based on an extension of terms to syntactic variants. It is qualified as an extending approach. The main difference between these two methods is the reference to a general-purpose grammar in the reducing case and to an exhaustive list of terms in the extending approach.
3.3 Two Families of Grammatical Rules

The rules presented in this section are production rules, taking as input one or more term structures and yielding as output a transformed term structure. As multiword base-terms are supposed to be only two-word terms (section 1.2, restriction 1), the production rules only apply to two-word term structures. With such an assumption, the transformations of term variants \((n\text{-word terms with } n \geq 3)\) are performed by applying transformations to binary sub-trees of their syntactic structure (see part 6). Figure 4 depicts an example of a production rule applying to two two-word terms and producing their coordination. This rule corresponds to the following formula:

\[
\left[ \begin{array}{c}
N, \ AN_1 \ N \\
N, \ AN_2 \ N
\end{array} \right] \rightarrow \left[ \begin{array}{c}
N', \ AN_{1''} \ C \ AN_{2''} \ N
\end{array} \right]
\]

The label \(N''\) assigned to the transformed construction denotes that the coordination of both terms with a \(N'\) label is a noun phrase which cannot be considered as a candidate term (section 1.2, restriction 2). When applied to the terms \(\left[ \begin{array}{c} N, \ [\lambda \text{ internal}] \ [\lambda \text{ carotid}] \end{array} \right] \) and \(\left[ \begin{array}{c} N, \ [\lambda \text{ external}] \ [\lambda \text{ carotid}] \end{array} \right] \), this rule produces the structure \(\left[ \begin{array}{c} N', \ [\lambda \text{ internal}] \ C \ [\lambda \text{ external}] \ [\lambda \text{ carotid}] \end{array} \right] \). The only non-lexical leaf of this structure is the one with the label \(C\) which can be substituted by the item \(or\) in order to derive the term variant \(\text{internal or external carotid}\).

A structure with a final construction equal to this transformation could also be obtained by coordinating two noun phrase structures composed of a noun and a premodifier such as:

\[
\left[ \begin{array}{c}
N, \ AN_{1''} \ N \\
N, \ AN_{2''} \ N
\end{array} \right] \rightarrow \left[ \begin{array}{c}
N', \ AN_{1''} \ C \ AN_{2''} \ N
\end{array} \right]
\]

When substituting the modifiers by adjective phrases equal to \textit{internal} and \textit{external} and the head noun by \textit{carotid}, the final construction of this rule is equal to the preceding string \textit{internal or external carotid}. Such an observation on multiple derivations of syntactically sound idiomatic constructions has already been reported in the work of [Abellé and Schabes, 1989] about the application of Tree Adjoining Grammars to the representation of idioms. There exists a last possibility for constructing the preceding variant of \textit{internal carotid} through an hybrid coordination involving a term and a syntactic construction:

\[
\left[ \begin{array}{c}
N, \ AN_1 \ N \\
N, \ AN_2' \ N
\end{array} \right] \rightarrow \left[ \begin{array}{c}
N, \ [\lambda_{AN''}] \ AN_{1''} \ C \ AN_{2''} \ N
\end{array} \right]
\]

In the following examples of rules of coordination \#\# (section 4) and insertion (section 5), a formula of a transformation involving two terms and a formula of a transformation involving a term and a syntactic structure will systematically be given. This parallel illustrates the uncertainty about the terminological status of the structure (here \textit{external carotid}) participating with the base-term to the construction of its variation. Thus when encountering \textit{internal or external carotid} as a variant of \textit{internal carotid}, syntactic clues cannot convince us that \textit{external carotid} is really a term although strong presumptions exist that it is a good term. The next two parts give a picture of the transformations modifying two-word terms. They are separated into production rules and variation rules that do not lead to the constitution of a term.
<table>
<thead>
<tr>
<th>Textual sequences</th>
<th>Terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Two-word terms / coordinated structure</td>
<td></td>
</tr>
<tr>
<td>renal hemodynamics and function</td>
<td>Renal function</td>
</tr>
<tr>
<td>cell growth and differentiation</td>
<td>Cell differentiation</td>
</tr>
<tr>
<td>optic disc and nerve</td>
<td>Optic nerve</td>
</tr>
<tr>
<td>high specificity and sensitivity</td>
<td>High sensitivity</td>
</tr>
<tr>
<td>language production and comprehension</td>
<td>Language comprehension</td>
</tr>
<tr>
<td>renal cortex and medulla</td>
<td>Renal medulla</td>
</tr>
<tr>
<td>gene amplification and expression</td>
<td>Gene expression</td>
</tr>
<tr>
<td>axillary artery and vein</td>
<td>Axillary vein</td>
</tr>
</tbody>
</table>

Table 4: Head coordination

<table>
<thead>
<tr>
<th>Textual sequences</th>
<th>Terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Two-word terms / coordinated structure</td>
<td></td>
</tr>
<tr>
<td>early and late phase</td>
<td>Early phase</td>
</tr>
<tr>
<td>apical and basolateral membrane</td>
<td>Apical membrane</td>
</tr>
<tr>
<td>skin and blood tests</td>
<td>Skin test</td>
</tr>
<tr>
<td>gastrointestinal and pulmonary disease</td>
<td>Gastrointestinal disease</td>
</tr>
<tr>
<td>electron and immunoelectron microscopy</td>
<td>Electron microscopy</td>
</tr>
<tr>
<td>spinal or facial nerve</td>
<td>Spinal nerve</td>
</tr>
<tr>
<td>white and grey matter</td>
<td>White matter</td>
</tr>
</tbody>
</table>

Table 5: Coordination of modifiers

## 4 Variations Rules: Coordinations and Permutations

This part describes the two families of variations that transform a term into a phrase structure. As derivational variants are not accounted for, it is possible to assume that variants are not expanding further than the noun phrase whose head is the head of the base-term. These variations are separated into two very different families: coordinations which join two terms into a more compact structure through the elision of one or more of their common constituents and permutations which permute the linear order of the constituents and transform a right-headed construction into a left-headed one.

### 4.1 An Observation of Coordination and Permutation Variations

#### 4.1.1 An Observation of Coordination Variations

# Coordination variations link two terms with the conjunction and and or, or with a comma inside enumeration. Two types of coordinations are observed:

- The two terms share a common head and the modifiers are coordinated (modifier coordination). For example cell growth and differentiation results from the coordination of two terms: cell growth and cell differentiation.

- The two terms share common modifiers and the heads are coordinated (head coordination). For example electron and immunoelectron microscopy results from the coordination of is a head coordination of electron microscopy and immunoelectron microscopy (table 5).

Modifier coordination is much more productive than head coordination in English. #
### Table 6: Permutation variations

<table>
<thead>
<tr>
<th>Two-word terms / Permutation around of</th>
<th>Terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>mode of transmission</td>
<td>Transmission mode</td>
</tr>
<tr>
<td>analysis of variance</td>
<td>Variance analysis</td>
</tr>
<tr>
<td>effects of light</td>
<td>Light effect</td>
</tr>
<tr>
<td>duplication of chromosome</td>
<td>Chromosome duplication</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Two-word terms / Permutation around with, for, at</th>
<th>Terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>association with chromosome</td>
<td>Chromosome association</td>
</tr>
<tr>
<td>thresholds for perception</td>
<td>Perception threshold</td>
</tr>
<tr>
<td>pressure at the capillary</td>
<td>Capillary pressure</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>2/3-word terms/ Permutation around on, from, in/into</th>
<th>Terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>factors on risk</td>
<td>Risk factor</td>
</tr>
<tr>
<td>antigen from hepatitis b</td>
<td>Hepatitis B antigen</td>
</tr>
<tr>
<td>albumin in serum</td>
<td>Serum albumin</td>
</tr>
</tbody>
</table>

#### 4.1.2 An Observation of Permutation Variations

Permutation variations are more productive than coordination but produce variants of lower quality. Two facts are significantly noteworthy:

1. Permutation mainly takes place around the preposition **of**, but also around the set of prepositions which are most common and around certain verbal sequences which have a low semantic content.

2. Permutation is generally followed by an insertion variation. This insertion composition is presented in section 6.

Prepositions used in the permutation variation are divided into two categories: prepositions **of, with, for and at** which only have a relational function and the others which are locative prepositions. We will further show that some of these prepositions also have a relational function.

In this presentation of permutations around prepositions, we have deliberately excluded the permutations around **by** and **to**. These two prepositions create particular filtering problems. The preposition **by** most often shows a passive verbal sequence whose form is fairly unpredictable. The locative preposition **to** leads to other more complex problems since it shares its graphic form with the infinitive marker of the verb as in: **analysis to determine** the DNA ploidy value (term: **Value analysis**).

**Preposition of**

The most constant signs of permutation occur around the preposition **of**. It is with this preposition, as opposed to the others that a strict permutation, without insertion, leads to the best results, i.e., the less noise. We can verify that the terms **Variance analysis** or **Chromosome duplication** are completely equivalent to textual sequences under which they have been identified (**analysis of variance, duplication of chromosome**). In this point of view the sole permutation demonstrates a synonymous connection between the term in its basic form and its transformed form. In addition and given the semantic neutrality of the preposition **of** this synonymous connection is not altered.

Variations around the preposition **of** are the most productive, representing 51.8% of the total number of permutations observed. It is around this preposition that there are the greatest number of insertion phenomena combined with a permutation variation (see section 6). The sole permutation remains mainly limited to terms with two words (examples are given in Table 6). Permutations of terms of length 3 or greater than 3 without insertion are relatively rare.
Prepositions with, for, at

These three prepositions together represent 13.7% of the total number of permutations (with: 6.1%, for: 5.5%, and at: 21%). We were not concerned with the semantic differences between them. We merely observed that they reveal a low semantic link as shown by the examples provided in Table 6.

Prepositions on, from, in/into

These three prepositions represent 23.7% of permutations. The preceding prepositions that were examined have a low semantic link. This is not the case with on, from and in/into which are prepositions that significantly show locational properties. It is worth stressing that the type of corpus used (medicine) in which there are connections between the parts of the body, organs, cells, etc. explains the importance of locative phenomena observed. However these properties are not constant.

- **In/into.**
  The greatest number of locative phenomena (15.2% of the total number of prepositions) occur around the prepositions in and into. With into, there is still a locative effect but as it will be showed in section 6, clearly more inclusive.

- **From.**
  4.5% of the total number of prepositions.

- **On.**
  This preposition represents 4% of variations.

4.2 A typology of Coordination and Permutation Variations

In this part, we analyse the coordinated and permuted structures that we have observed and we classify them taking into account the variations of the base-term structures.

A typology of Coordination Variations

Coordination is a rather complex syntactic phenomenon and seldom generates new terms. We have so classified coordination into variation rules. Let us examine few examples of sequences of length 3 obtained by coordination:

Coordination of two base-terms sharing an identical structure

1. Coordination of two base-terms of structure $N_2 N_1$
   (a) Coordination upon $N_1$ (illustrated by preceding formulae (5) and (7) in 3.3)
   $$[N_2 N_1], [N_3 N_1] \rightarrow N_3 \text{ CONJ } N_2 N_1$$
   - modulation and demodulation equipments
   - cable or microwave links
   - command / data interface
   $$[N_2 N_1], [N_3 N_1] \rightarrow N_4 C N_3 C N_2 N_1$$
   - monitoring, alarm and control systems
   (b) Coordination upon $N_2$ (illustrated by following formulae (8) and (9) in 4.3)
   $$[N_2 N_1], [N_2 N_3] \rightarrow N_2 N_1 C N_3$$
   - packet assembly/desassembly

2. Coordination of two base-terms of structure $A N$ (illustrated by preceding formulae (5) and (7) in 3.3)
   $$[A_1 N], [A_2 N] \rightarrow A_1 C A_2 N$$
   - low and medium bit

18
- leased or purchased transponders
- conversational / interactive processes
- signalling and routing procedures

Coordination of two base-terms of different structures

1. \([N_2 \; N_1], \; [A \; N_1] \rightarrow N_2 \; C \; A \; N_1\)
   - space and terrestrial services
   - command and ranging system

A typology of Permutation Variations

Some base-terms of \(N_2 \; N_1\) structure accept the following inverted form: \(N_1 \; of \; N_2\) without determination of \(N_2\). They are illustrated by following formula (10) in 4.3.

- radio performance \(\rightarrow\) performance of radio
- antenna performance \(\rightarrow\) performance of antenna
- frequency band \(\rightarrow\) band of frequencies
- data transmission \(\rightarrow\) transmission of data

But the majority of them do not accept inverted form as for example: traffic capacity, link budget, speech signal, signal frequency, antenna feed, antenna gain, luminance signal, data rate, carrier recovery.

We take the decision to consider the base-terms which accept the two forms, namely \(N_2 \; N_1\) and \(N_1 \; of \; N_2\) as base-terms of structure \(N_2 \; N_1\) which own the property of permutation. The base-term of \(N_1 \; of \; N_2\), and more generally of \(N_1 \; P \; N_2\) which are not encountered under the \(N_2 \; N_1\) are base-terms of structure \(N_1 \; P \; N_2\) as for example curve without coding, curve with coding, period of time. We will see in section 6 that the inverted structure of a base-term can be further modified by an insertion in order to produce variants spanning over several words.

4.3 Rules for Coordinations and Permutations

As presented in 4.2, coordinations can join two terms or a term and a phrase structure either by factoring the head noun and coordinating the modifier or by factoring a common modifier and coordinating the non-common heads. The experimental observation of corpora reports a rate of coordinated modifiers much higher than the rate of coordinated heads. Equations (5) and (7) in 3.3 represent two-word term argument coordinations and Equations (8) and (9) represent two-word term head coordinations:

\[
\begin{array}{c}
\left[ \begin{array}{c} N' \\ AN \; N_1 \end{array} \right], \left[ \begin{array}{c} N' \\ AN \; N_2 \end{array} \right] \rightarrow \left[ \begin{array}{c} N'' \\ \begin{array}{c} AN \\
\begin{array}{c} A \; AN \\
N_1 \; C \; N_2 \\
N' \\
N'' \\
N'' \; C \; N \\
\end{array} 
\end{array} \right] \\
\left[ \begin{array}{c} N' \\ A \; N_1 \end{array} \right], \left[ \begin{array}{c} N' \end{array} \right] \rightarrow \left[ \begin{array}{c} N'' \\ \begin{array}{c} AN \\
\begin{array}{c} A \; AN \\
N'' \\
N'' \; C \; N \\
\end{array} 
\end{array} \right] 
\end{array}
\]

# Permutations are transformations which allow for the conflation of a non-syntagmatic construction (a term) with a noun phrase involving the words of the original terms in a reverse order but with the same dependency relation (section 4.2). This happens when a preposition is inserted between both nouns which reveals more clearly the semantic relation holding between the head noun and its argument. For example gas analysis is observed within the medical corpus of INIST as analysis by gas and serum albumin is found as albumin in serum. However, the vast majority of the observed permutations are made up of the # preposition of which is less semantically informative. A permutation can only occur when the argument has a nominal category, that is to say is a noun or a gerund but not an adjective or a past participle.
Contrary to coordination and insertion permutation requires only one structure as input: the term to be modified. It is represented by the following equation:

$$\left[ \begin{array}{c} N_2 \\ N_1 \end{array} \right] \rightarrow \left[ \begin{array}{c} N_1 P \\ N_2 \end{array} \right]$$

(10)

This transformation is generally completed with a syntactic variation of the resulting noun phrase (see section 6). In fact, permutation is motivated by a need for a syntagmatic transformation of the argument or of the head noun which is not well tolerated by compounding. For example, *resident population surveys* is better expressed as *surveys in the resident population*. The term is transformed into an equivalent syntagmatic construction by permuting its component and inserting a preposition. Romance languages which only have left-headed constructions, do not use permutations. In French, the shift from compounding to noun phrase syntax is more continuous than in English. For example the term *mesure de la tension (pressure measurement)* is modified as *mesure de la tension artérielle (measurement of arterial pressure)*.

### The English translation reveals a compound structure which has no counterpart in French.

In order to achieve the description of the different syntagmatic variations, the next section presents insertions and juxtaposition which transform a term into another one yielding a candidate term.

## 5 Production Rules: Insertion and Juxtaposition

### 5.1 Insertions

# Insertions generally operate on right-headed structures: a head word preceded by modifiers composed of nouns, adjectives, adverbs and past participles. In the study of insertion variations we make two hypotheses:

1. Elements inserted within a term have a modifying function (noun, adjective, past participle).

2. When an element inserted is a term, it also has a modifying function.

We will further see that these hypothesis restrict the geometry of the term structure and the location of the insertion.

In addition, we have observed that besides right-headed structures, some terms are right-headed: *Vitamin d, Chromosome x*. This type of structure is a metonymic expression of a noun phrase form capable of being paraphrased in the following way: X of type Y = X Y; *vitamin of type d = vitamin d.*

### 5.1.1 An Observation of Insertion Variations

**Two-word terms: insertion of a word**

The insertion of a word in the structure of a two-word term leads to two possibilities of linking the inserted element: a head link or a tail link.

- **Head link.**
  
The insertion is made according to the structure $X_2 \left[ X_3 \ N_1 \right]$. We have a structure of two-word terms ($X_2$ and $N_1$) where $N_1$ is the head word $X_2$ a modifier (adjective, noun, past participle, etc.) and $X_3$ an inserted element. In this type of structure the $X_3$ inserted element can only have a modifying position of the $N_1$ head. This clearly shows that there is a term composition [Daille, 1994]: the $X_2 N_1$ term is composed of the term $X_3 N_1$ and one can observe the $X_2 X_3 N_1$ sequence in the corpus.

# Table 7 shows the possible types of variations. For some of them, the inserted word can be omitted without altering the meaning of the term. It is the case with *Arterial pressure* when it is identified in the textual sequence *arterial blood pressure*. However, the inserted element generally brings an additional information. For the term *Transmepithelial transport, na* in the sequence *transmepithelial na* transport indicates that it is a specific type of transport. #

20
The inserted word can also be an adjective: supercritical, inserted in capillary supercritical fluid and linked to the term Capillary fluid. In this case we most often identify a structure of the term Adjective Noun. The insertion of a new adjective shows an adjectival over-modification of the base-term.

- **Tail link.** We have seen in the beginning of this section that some terms have their heads in a left position, the modifier being to the right. This is the case for a term such as Vitamin d. The structure \[
\left[ X_2 \ X_3 \right] N_1
\]
shows how a X₃ element is inserted and linked as modifier to the X₂ tail element of the term.

Hence, the textual sequence ultraviolet b radiation is the result of the term composition Ultraviolet b and Ultraviolet radiation. Table 8 gives some of the most meaningful and most recurrent examples of this type of insertion variation.

### 5.1.2 A Typology of Insertions

Insertions are variations which insert at least one constituent within the noun phrase where the term occurs. The insertion may eventually break the connectivity of the term string by introducing words inside it, but, whatever the type of the insertion, the linear order of the words is not changed. As it has been said previously, only insertions altering the term string (e.g. head substitutions given thereunder) are illustrated by transformation rules in §5.1.3. Insertions are subdivided into substitutions which transform two base-terms into another term and modifications which adjoin an adjectival or an adverbial modifier to a base-term. Substitution and modification if they vary concerning the number of base-terms employed could yield a new term. These two classes are further subdivided according to the base-term structures involved in the transformation.
Substitution

The definition of the substitution is the following: given a base-term of length 2, one of main item is substituted with a base-term whose head is this main item. For example, with the base-term of structure N₂ N₁, N₁ can be substituted by a base-term of N₃ N₁ structure to produce an overcomposed term of N₂ N₃ N₁ structure. For example, the noun network in the base-term satellite network is substituted by the base-term transit network to produce the overcomposed term satellite transit network. Substitution differs from juxtaposition (presented in section 5.2) in two points:

- it asks for two base-terms or more generally for two terms, in the above example satellite network and transit network,
- it could break the internal structure of one of the base term, in the above example the structure of satellite network is broken,
- no hyphen is introduced in the substituted structure.

We present now a few examples of term variants obtained by substitution of the head of the term, they are modelled by the rules (11) and (12) in section 5.1.3:

1. N₂ N₁ + A N₁ → N₂ A N₁
   - noise density + spectral density → noise spectral density
   - wideband cable + submarine cable → wideband submarine cable
2. N₂ N₁ + N₃ N₁ → N₂ N₃ N₁
   - satellite antenna + receiving antenna → satellite receiving antenna
   - membrane protein + skeleton protein → membrane skeleton protein
3. A₂ N₁ + N₃ N₁ → A₂ N₃ N₁
   - arterial pressure + blood pressure → arterial blood pressure
   - dorsal ganglion + root ganglion → dorsal root ganglion
4. A₂ N₁ + A₃ N₁ → A₂ A₃ N₁
   - vertical laryngectomy + partial laryngectomy → vertical partial laryngectomy

And a few examples obtained by substitution of one of the arguments:

1. N₂ N₁ + A N₂ → A N₂ N₁
   - satellite network + geostationary satellite → geostationary satellite network(s)
   - channel signalling + common channel → common channel signalling
2. N₂ N₁ + N₃ N₂ → N₃ N₂ N₁
   - noise spectrum + intermodulation noise → intermodulation noise spectrum
3. N₂ N₁ + N₂ M₃ → N₂ M₃ N₁
   Where the M symbol is a particular modifier (digit or character) and where the term N₂ M₃ is formed with a head noun in the left position and with a modifier in the right position.
   - vitamin deficiency + vitamin d → vitamin d deficiency

# When analysing the conceptual denotation of a term and its substitution variant, two cases should be distinguished:
• The substitution variant is not a conceptual variant of the base term. For example, arterial blood pressure is not a variant of blood pressure because, implicitly, arterial pressure only concerns blood pressure. In this case, what is considered as the original term is indeed an elision variant of the pseudo-variant—here arterial pressure.

• The substitution variant is a true variant of the original term and denotes a different concept—generally a more specific one. This case is illustrated by argument substitution with vitamin d deficiency which is more specific than vitamin deficiency and by head substitution with wideband submarine cable which is more informative than wideband cable.

#

5.1.3 Rules for Substitutions

# An experimental observation of n-word terms with \( n \geq 3 \) shows that most of these terms are constructed from binary terms through substitution and adjectival modification (see [Daille, 1994] for a precise study of multiword term structures in French). Substitution is illustrated by (11) and modification by (12). The first transformation involves two terms such as abnormal chromosome and X chromosome and produces abnormal X chromosome. The second transformation take as input only one term such as plasma density and produces as output a variation such as plasma low density where low density is not a term. These two variations are labelled as insertions for they insert word inside the string of the original term without changing the order of the words. The only rules reported for insertions are the ones which modify the sequence of the terms which they apply to. #

Both insertions (11) and (12) involve two structures with a common head and are a substitution of the head in one structure by the other structure. Equations (11) and (12) represent two-word term insertions:

\[
\begin{align*}
\left[ \begin{array}{c}
N^' \\
N
\end{array} \right], \left[ \begin{array}{c}
AN_1 \\
AN_2
\end{array} \right] & \rightarrow \left[ \begin{array}{c}
AN_1 \\
AN_2 N
\end{array} \right] \\
\left[ \begin{array}{c}
N^' \\
N
\end{array} \right], \left[ \begin{array}{c}
AN_1 \\
AN_2 \ N
\end{array} \right] & \rightarrow \left[ \begin{array}{c}
AN_1 \\
AN_2 N
\end{array} \right]
\end{align*}
\]

(11) (12)

The resulting structure of both rules is labelled N' because it is a candidate term.

External modifications which do not modify the string of the initial term such as the transformation of somite malformation into complex somite malformation are given in section 5.2.1 as juxtapositions.

5.2 Juxtaposition

Juxtapositions are variations which do not break the connectivity of the base-term strings as they do not introduce words inside them. Juxtapositions are subdivided into overcompositions which transform two base-terms into another term and modifications which adjoin an adjectival or an adverbial modifier to a base-term. Overcomposition and modification if they vary concerning the number of base-terms employed could yield a new term. These two classes are further subdivided according to the base-term structures involved in the transformation.

They are illustrated by rules (14) and (15) in section 5.2.1 for juxtaposition and by rule (13) for modification.

Overcomposition

A overcomposed term obtained by juxtaposition, that we will called a juxtaposed term, is built with at least one base-term and could be characterised by the following properties:

---

6Some accidental coordinated structures are encountered for multiword terms such as Thomsen Friedreich Antigen, but they should not be considered as productive. They generally correspond to an implicit coordination where the coordinating conjunction is elided or is a '/' character.

7Coordinations defined in 4.3 are also insertions but the difference between genuine insertions and coordinations is that coordination must insert one (and only one) coordinating conjunction whereas insertion must not insert a conjunction.

23
• the items of the base-term structure do not only stay together but compounthood is also, most of the
time, reinforced by the presence of an hyphen,

• when it is a simple noun that is juxtaposed to a base-term, it is most of the time the simple noun
which appear after the base-term,

• overlaps inside the juxtaposed structure do not refer to base-terms, or if they do the presence of a
hyphen indicates clearly which is the base-term. This property is illustrated in the following examples:

1. Juxtaposition of a base term and a simple noun
[N₃(-)N₂] N₁ (length 3)

• [broadcasting satellite] space
  Neither broadcasting space nor satellite space are base-terms.

• [earth-station] antenna
  Earth antenna is not a base-term. Station antenna is a base-term but thanks to the hyphen, the
base-term is clearly identified and the ambiguity of attachment, namely earth station or station
antenna is raised.

• [radio-frequency] spectrum
  Frequency spectrum and radio frequency are base-terms but thanks to the presence of the hyphen
the ambiguity is raised.

[A N₂] N₁ (length 3)

• [low noise] amplifier(s)
  Neither low amplifier nor noise amplifier are base-terms.

• [shaped beam] antenna
  Neither shaped antenna nor beam antenna are base-terms.

• [low capacity] system(s)
  Neither low system nor capacity system are base-terms.

N₁ [N₂(-)N₁] (length 3)

• satellite [propagation delay]
  Neither satellite propagation nor satellite delay are base-terms of our corpus.

2. Juxtaposition of two base-terms
[N₃ N₂] [Adj₁ N₁] (length 4)

• [demand assigned] [multiple access]
  Neither assigned access nor demand access are base-terms. Note that the base term is demand
assignment and not demand assigned.

• [time division] [multiple access]
  Neither division access nor time access are base-terms.

[N₄ N₃] [N₂ N₁] (length 4)

• [data network] [identification code]
  Neither network code nor network identification, nor data code, nor data identification are base-
terms.

[A N₃] [N₂ N₁] (length 4)

• [high speed] [communication link(s)]
  Neither high communication nor high link, nor speed communication, nor speed link are base-
terms.
Modification

Modifiers take place before the base-term of $N_2 N_1$ or $A N$ structure and thus, do not break its internal structure. A modified base-term could lead to a new term. It is quite impossible to state about the terminological character of the modified sequence. Two main kinds of modifiers are isolated: adverbs and adjectives. These variations have no corresponding rules in the model given in section 5.1.3. Two main modifiers are juxtaposed to a base-terms: adjectives or adverbs. The criterion of no overlaps inside the juxtaposed structure that refer to base-terms is also true for modification and is illustrated in the following examples.

Adjectival Modifications

Adjectives that appear in front of the base-term modify it but remain invariant. We give now a few examples:

1. $A_2 [A_1 N]$ (length 3)
   - internal [signalling protocol]
     internal protocol is not a base-term.
   - maximum [available power]
     maximum power is not a base-term.
   - wide [shaped beam]
     wide beam is not a base-term.

2. $A [N_2 N_1]$ (length 3)
   - adaptive [delta modulation]
     Neither adaptive delta nor adaptative modulation are base-terms.
   - high [bit rate]
     Neither high bit nor high rate are base-terms.
   - interfering [earth(-)station]
     Neither interfering earth nor interfering station are base-terms.

Adverbial Modifications

Adverbs modifiers appear before the base-terms of $A N$ structure.

- relatively [low cost]
- technically [demanding installation]
- isotopically [radiated power]

5.2.1 Rules for Juxtapositions

Juxtapositions only involve one term for their creation when they juxtapose a two-word term and a single word. The only case where juxtapositions involve two two-word terms is illustrated by preceding examples such as data network identification codes. Thus rule (13) is an external modification involving one term and a modifier which is not mentioned in the left part of the rule. Rule (14) operates on one term which becomes the argument of a head word which is also not indicated in the left part of the rule. Conversely, rule (15) involve two terms. The left-most one becomes the argument and the right-most one becomes the head of the juxtaposed term:
\[ [N', AN_1 N] \rightarrow \left[ [N', AN_2 [N', AN_1 N]] \right] \]  (13)

\[ [N', AN_1 N] \rightarrow \left[ [N', [AN_1 N] N_2] \right] \]  (14)

\[ [N', AN_1 N_1], [N', AN_2 N_2] \rightarrow \left[ [N', [AN_1 N_1] [N', AN_2 N_2]] \right] \]  (15)

6 “Doing” derivations within noun phrases to generate possible term variants

In sections 3.3, 4.3, 5.1.3 and 5.2.1 ten transformations describing the basic variations of two-word terms (5), (7), (8), (9), (10), (11), (12), (13), (14) and (15) have been given. Most of them can be compounded with one another or compounded with regular syntactic transformations. We do not intend to give an exhaustive figure of all the possible compositions. More illustrative is the result of some of the most frequent and relevant non-elementary transformations. For example, the application of an insertion on the term electron microscopy results in electron scanning microscopy which, in turn, can be modified by a coordination and yield the final variant electron scanning and optical microscopy.

Any of the non-elementary variants observed in corpora is the result of such a composition with a variable degree of complexity. The description that we intend to give in this section is restricted to n-word wide windows with 3 \( \leq n \leq 5 \). The minimal size corresponds to the smallest variants of two-word terms resulting from an insertion of one word inside the two words of the base-terms (which may be eventually permuted). The maximal size of the window is more difficult to determine because the category of the inserted words is important. A more precise measure of the size of a textual window should take into account the number of inserted content words separately from the number of inserted stop words. A size of five words is usually accepted as a reasonable limit of observation beyond which spurious co-occurrences become significantly frequent. [Martin et al., 1983] have stated that most of the syntagmatic lexical relations occur between words separated by at most five words. The degradation of the results on larger windows is however not dramatic. Excellent variants are observed in say six content-word wide windows as erythroid macrophage colony forming cell, a variant of erythroid cell.

The next point focuses on the most productive variations observed in corpora within windows of three to five words. It details the transformations from which they result. Some rare compositions such as double coordinations (which are the result of a coordination composed by a coordination) are deliberately ignored. They yield variants such as animal and vegetable oils and fats. The construction mechanisms can be reversed in order to acquire new terms by eliciting the terms participating to the elaboration of the observed variants as will be shown in section 7.8

6.1 An Observation of Complex Variants

This section illustrates variations, which involve only one multiword term and which are not elementary transformations. It concerns variation insertions where two words or more are inserted and permutations and coordinations which are composed with insertions.

8The transformations of insertion and coordination which accept two terms as input such as (5), (8) and (11) are opportunities to acquire candidate terms by guessing the second term which has participated to the transformation. For example, the observation of the variant \( AN_1 AN_2 N \) of the term \( AN_1 N \) produced by (11) reveals the candidate term \( AN_2 N \). Thus the variant systolic blood pressure of systolic pressure shows that blood pressure is a candidate term. The interest of such an acquisition is to relate semantically the terms proposed as candidates to the modified one. This is true for both preceding terms related through an insertion as well as for terms related through a coordination: coordinated terms have closely related meanings. For example, motor nerve and sensory nerve found in motor and sensory nerve are semantically close. Variation is not only a way to enhance the recall rate of an automatic indexing preprocessor, it is also a precise tool for symbolic terminological acquisition [Jacquemin, 1995].
<table>
<thead>
<tr>
<th>Textual sequences</th>
<th>Terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Two-modifiers insertion</td>
<td></td>
</tr>
<tr>
<td>lingual mandibular salivary gland</td>
<td>Lingual gland</td>
</tr>
<tr>
<td>malignant epithelial ovarian tumors</td>
<td>Malignant tumors</td>
</tr>
<tr>
<td>digital intraoral radiographic system</td>
<td>Digital system</td>
</tr>
<tr>
<td>retaining intra urethral device</td>
<td>Retaining device</td>
</tr>
<tr>
<td>monoclonal anti ganglioside antibodies</td>
<td>Monoclonal antibodies</td>
</tr>
<tr>
<td>serum atrial natriuretic factor</td>
<td>Serum factor</td>
</tr>
<tr>
<td>Term insertion</td>
<td></td>
</tr>
<tr>
<td>papillary transitional cell carcinoma</td>
<td>Papillary carcinoma</td>
</tr>
<tr>
<td>high esophageal contraction amplitudes</td>
<td>High amplitudes</td>
</tr>
<tr>
<td>sinusoidal eye movement signal</td>
<td>Sinusoidal signal</td>
</tr>
<tr>
<td>monoclonal cell surface immunoglobulin</td>
<td>Monoclonal immunoglobulin</td>
</tr>
<tr>
<td>serum tumour necrosis factor</td>
<td>Serum factor</td>
</tr>
<tr>
<td>autologous extraaortic counterpulsation system</td>
<td>Autologous system</td>
</tr>
</tbody>
</table>

Table 9: Insertion of two words in a term structure with two words

6.1.1 Complex insertions

Most of the insertions, are basic ones. They involve binary terms and correspond to one-word insertions. We show here, that insertions can be generalised to sequences of two words and that terms of length three also accept insertions of two words. The complexity of this type of insertion is due to the fact that it is not possible to say what transformation is at stake. We show with a decomposition of terms in tree structure the different possibilities of attachment of inserted elements.

Two-word terms: insertion of two words

Insertion of two words in a structure of terms consisting of two words itself brings two possibilities of linking the inserted elements to the existing structure of the term: either the two elements are modifiers or they form a term. Formula (18), in section 6.2 describes the composition of two insertions. The formula for the insertion corresponding to the second point hereunder is not given because we have not exemplified the compositions involving juxtapositions. This insertion is a composition of the formula for external juxtaposition (14) of section 5.2.1 with the insertion given by formula (11) of section 5.1.3. Thus the one-word constituent AN₂ of the second term is in equation (11) is replaced by an embedded 2-word term.

- **Insertion of two modifiers.** The structure \( X₂ \left[ X₄ \left[ X₃ N₁ \right] \right] \) shows how the \( X₄ \) and \( X₃ \) modifiers are inserted in the \( X₂ N₁ \) term structure. With an attested term such as *Malignant tumors* there can be *epithelial* and *ovarian* modifiers in the sequence: *malignant epithelial ovarian tumors*. Table 9 gives a few examples among the most meaningful ones of a double insertion of modifiers. This double modification is most often formed with two adjectives (see example above) or of a prefix and an adjective (*infra urethral*, in the term *Retaining device*).

- **Term insertion.** A certain number of terms naturally find their place in a two-word term structure, as shown in the following structure: \( X₂ \left[ X₄ \left[ X₃ N₁ \right] \right] \). \( X₄ \) and \( N₃ \) form a term which is linked to the \( N₁ \) head element. Table 9 gives a set of examples of inserted terms such as *eye movement* inserted within the term *Sinusoidal signal* yielding the occurrence *sinusoidal eye movement signal*. It is noteworthy that all of the non attested *Adjective Noun* or *Noun Noun* forms seem to be potential candidate terms in this position.
Three-word terms: one-word and two-word insertions

In the structure of 3-word terms, the insertion point cannot be chosen at random since it has an effect on the basic structure of the term. Actually, there are two types of terms: those whose head consists of a term, and those whose tail consists of a term but in a modifying or complex modifying position. The insertion variation is therefore a mean to predict the syntactic structure of this type of term and to classify it. These variations are not illustrated by formulae because the formalisation is restricted to variations of 2-word terms.

- **Terms whose head consists of a term**
  A $S$ word insertion (one word or a sequence of two words) in the term structure $X_3 X_2 N_1$ is given by:
  \[
  \left[ \begin{array}{c}
  X_3 \\
  X_2 \\
  S \\
  S \\
  N_1 \\
  \end{array} \right]
  \] .
  $X_2 N_1$ forms a term which has a head position. The insertion point for a given term can only be located between $X_3$ and $X_2$.
  Hence the modifiers in the term *Left coronary artery* are always inserted between *Left* and *coronary* for example *left* (*common + main + circumflex*) *coronary artery* are such variants. Similarly, in the term *Pulmonary wedge pressure* the words *capillary and artery* can be inserted as indicated by the following forms: *pulmonary* (*capillary + pulmonary*) *wedge pressure*.
  Table 10 shows this particular insertion phenomenon. The inserted element ### is either a noun or an adjective, indifferently; but also a sequence of two words: *ventricle dependant* inserted into the term *Right coronary artery* in the textual sequence *right ventricle dependent coronary artery*.

- **Terms whose tail consists of a term/complex modifier**
  #Structure $\left[ \begin{array}{c}
  X_3 \\
  X_2 \\
  S \\
  N_1 \\
  N_1 \\
  \end{array} \right]$ . The examples in Table 10 confirm that a $S$ insertion (one word or a sequence of two words) near a $N_1$ head element enables to predict the term structure. The tail $X_3 X_2$ is a term in most cases. # The insertion of *jet* in the textual sequence *high frequency jet ventilation*, identified by the term *High frequency ventilation* indicates that the term *high frequency* is embedded in the base-term. Sometimes, the words $X_3 X_2$ can be complex modifiers such as *two dimensional* in *Two dimensional electrophoresis* accepting *two dimensional gel electrophoresis* for variant.

  These observations show that whatever the nature of $X_3 X_2$ (complex modifier or term) the $S$ insertion (one or two words) strongly constrains the structural forms of the term which was used for identification ends. Consequently, $S$ is relatively independent of $X_3$ and $X_2$ and can only be linked to the $N_1$ head.

  The same observations can once again be made with the insertion of two words (see examples in table 10). The $X_3 X_2$ sequence isolated through insertion will be either terms (*reverses phase* for the term *Reversed phase chromatography* with the variant *reversed phase ion pair chromatography*) or complex modifiers (*two dimensional* for the term *Two dimensional analysis* and its variant *two dimensional triplc peptide analysis*).

6.1.2 Permutations composed with insertions

More complex substitutions exist which use the inverted structure of a $N_2 N_1$ base-term. They are illustrated by formula (21) of section 6.2. The different strings produced by this rule are illustrated by Table 13.

  Generally the preposition *of* is used in such variations but other prepositions are be used as shown in the following examples:

  $N_2 N_1 + A N_2 \rightarrow N_1 P N_3(-)N_2$

  - *satellite link* + *geostationnary satellite* → *link of geostationary-satellite*

  As overcomposition is a recursive operation, substitution can occur between a base-term and an overcomposed term obtained by juxtaposition. Rule (21) given in 6.2 as a composition of elementary transformations
<table>
<thead>
<tr>
<th>Terms whose head consists of a term</th>
<th>Terms whose tail consists of a term/complex modifier</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>regional cerebral blood flow</strong></td>
<td><strong>High frequency ventilation</strong></td>
</tr>
<tr>
<td><strong>passive latex agglutination test</strong></td>
<td><strong>Hepatitis b antigen</strong></td>
</tr>
<tr>
<td><strong>scanning transmission electron microscopy</strong></td>
<td><strong>Self sustained discharge</strong></td>
</tr>
<tr>
<td><strong>right ventricle dependent coronary artery</strong></td>
<td><strong>Two dimensional analysis</strong></td>
</tr>
<tr>
<td><strong>established mouse hepatocyte cell lines</strong></td>
<td><strong>High affinity site</strong></td>
</tr>
<tr>
<td><strong>established rat glioma cell line</strong></td>
<td><strong>Intermittent mandatory ventilation</strong></td>
</tr>
</tbody>
</table>

Table 10: One / two word insertion in a three-word term structure

illustrates the following possibility:

\[ N_2 N_1 + N_4-N_3 N_2 \rightarrow N_1 P N_4-N_3 N_2 \]

- **water flow** → **flow of fire-fighting water**
- **equipment requirement** → **requirements on earth-station equipment**

Permutations most often are combined with in an insertion variation. This insertion composition is demonstrated in the presentation tables of permutation for each preposition where the inserted elements are illustrated and in bold characters.

**Preposition of**

Variations around the preposition of are the most productive. It is around this preposition that there are the greatest number of insertion phenomena combined with a permutation variation. Table 11 shows, for terms with two words, insertions that range from one word (**rodent** linked to the sequence **transformation of rodent cells**) to four words (term **Cell supernatant** linked to the sequence **supernatant of cultured peripheral blood mononuclear cells**). This variation remains mainly limited to terms with two words. Terms with three words which vary such as **Flow velocity measurement in measurements of femoral artery flow velocity** or **measurement of cerebral blood flow velocity** are relatively rare.

**Other prepositions**

Other prepositions do not have a neutral semantics like of. We have limited our work to the most usual prepositions: with, for, at, on, from and in. They show semantic links (mainly locational) to various degrees. Complex variations of permutation and insertion have been observed also with these other prepositions. The number of inserted elements is rarely above three. These syntactic constructions, like **supernatants from eleven thyroid follicular cell** derived from the term **Cell supernatant**, make the links which associate the words together more explicit and are preferred to variations for compounding for very long textual sequences.

**6.1.3 Coordination composed with insertions**

Coordination variations link together terms that share either a common head or a common modifier. The coordination can also be combined with the insertion when rules accept an additional inserted element.

Formule (16) and (17) (see section 6.2) show how the head or modifier coordinations are composed with the insertion. In the example that we give (table 12), we must notice that modifier coordinations are the
<table>
<thead>
<tr>
<th>Textual sequences</th>
<th>Terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>One-word insertion</td>
<td>Membrane interaction</td>
</tr>
<tr>
<td>interaction with hemodialysis membranes</td>
<td>Children hospital</td>
</tr>
<tr>
<td>hospital for sick children</td>
<td>Shear viscosity</td>
</tr>
<tr>
<td>viscosity at varying shear</td>
<td>Pressure effect</td>
</tr>
<tr>
<td>effects on blood pressure</td>
<td>Fluid contamination</td>
</tr>
<tr>
<td>contamination from body fluids</td>
<td>Cell transformation</td>
</tr>
</tbody>
</table>

| Two-word insertion | | |
|-------------------|-----------------|
| measurement of cerebral blood flow | Flow measurement |
| irradiation with long wavelength ultraviolet | Ultraviolet irradiation |
| factors for granulocyte macrophage cells | Cell factor |
| contrast on clinical MR images | Image contrast |
| supernatants from synovial adherent cells | Cell supernatant |

| 3/4-word insertion | | |
|--------------------|-----------------|
| supernatant of cultured peripheral blood mononuclear cells | Cell supernatant |
| effect on the extracellular glutamate concentration | Concentration effect |
| supernatants from eleven thyroid follicular cells | Cell supernatant |

Table 11: Permutation and Insertion of one to four Words

<table>
<thead>
<tr>
<th>Textual sequences</th>
<th>Terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Head coordination: two-word insertion in two-word terms</td>
<td></td>
</tr>
<tr>
<td>cell membrane structure and function</td>
<td>Cell function</td>
</tr>
<tr>
<td>image cytometric analysis and evaluation</td>
<td>Image evaluation</td>
</tr>
<tr>
<td>magnetic field inhomogeneity and susceptibility</td>
<td>Magnetic susceptibility</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Modification coordination: Two-word insertion in two-word terms</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>systolic and diastolic blood pressure</td>
<td>Systolic pressure</td>
</tr>
<tr>
<td>tumor or nontumorous hepatic cells</td>
<td>Tumor cell</td>
</tr>
<tr>
<td>premalignant and malignant oral lesions</td>
<td>Premalignant lesion</td>
</tr>
<tr>
<td>single and multiple gavage doses</td>
<td>Single dose</td>
</tr>
<tr>
<td>skin or lymph node disease</td>
<td>Skin disease</td>
</tr>
<tr>
<td>horizontal and vertical condylar axis</td>
<td>Horizontal axis</td>
</tr>
<tr>
<td>positive and negative drive pressures</td>
<td>Positive pressure</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Modification coordination: three-word terms</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>middle and posterior cerebral arteries</td>
<td>Middle cerebral artery</td>
</tr>
<tr>
<td>superior and inferior vena cava</td>
<td>Superior vena cava</td>
</tr>
<tr>
<td>peripheral and central nervous system</td>
<td>Peripheral nervous system</td>
</tr>
<tr>
<td>intrahepatic and extrahepatic bile ducts</td>
<td>Intrahepatic bile duct</td>
</tr>
<tr>
<td>visual and auditory evoked potentials</td>
<td>Visual evoked potential</td>
</tr>
<tr>
<td>cervical and thoracic spinal cords</td>
<td>Cervical spinal cord</td>
</tr>
<tr>
<td>inflammatory and erosive joint disease</td>
<td>Inflammatory joint disease</td>
</tr>
<tr>
<td>anterior or middle cerebral arteries</td>
<td>Anterior cerebral artery</td>
</tr>
</tbody>
</table>

Table 12: Two-word terms coordination composed with insertion and three-word terms coordinations
most productive, like sole coordinations. We show that coordination variations are not limited to two-word terms, but it is possible to observe this phenomenon with three-word terms: therefore the attested term Peripheral nervous system is coordinated with central nervous system in peripheral and central nervous system.

6.2 Composing Elementary Transformations

Any term produced by the insertion rule (11) can be reprocessed by a transformation of two-word terms by considering the term AN2 N embedded in the output of (11) as a single constituent. Thus the insertion rules can be composed by a coordination or an insertion and produce the following variants which correspond to extended coordination and extended insertion:

Composition of an insertion and an argument coordination (ultrasound imaging + magnetic resonance imaging → ultrasound and magnetic resonance imaging):

\[(5) \circ (11) \left[ \begin{array}{c}
\text{AN}_1 \\
\text{AN}_2 \\
\text{AN}_3
\end{array} \right] \rightarrow \left[ \begin{array}{c}
\text{AN}^* \\
\text{AN}_1 \\
\text{AN}_2 \\
\text{AN}_3
\end{array} \right] \] (16)

Composition of an insertion and a head coordination (cell membrane structure + cell function → cell membrane structure and function):

\[(8) \circ (11) \left[ \begin{array}{c}
\text{AN}_1 \\
\text{AN}_2 \\
\text{AN}_3
\end{array} \right] \rightarrow \left[ \begin{array}{c}
\text{AN}_1 \\
\text{AN}_2 \\
\text{AN}_3 \\
\text{C N}_1
\end{array} \right] \] (17)

Composition of two insertions (primary response + humoral immune response → primary humoral immune response):

\[(11) \circ (11) \left[ \begin{array}{c}
\text{AN}_1 \\
\text{AN}_2 \\
\text{AN}_3
\end{array} \right] \rightarrow \left[ \begin{array}{c}
\text{AN}_1 \\
\text{AN}_2 \\
\text{AN}_3
\end{array} \right] \] (18)

Composition of three insertions (standard test + oral glucose tolerance test → standard oral glucose tolerance test):

\[(11) \circ (11) \circ (11) \left[ \begin{array}{c}
\text{AN}_1 \\
\text{AN}_2 \\
\text{AN}_3 \\
\text{AN}_4
\end{array} \right] \rightarrow \left[ \begin{array}{c}
\text{AN}_1 \\
\text{AN}_2 \\
\text{AN}_3 \\
\text{AN}_4
\end{array} \right] \] (19)

The composition of permutation with noun phrase rules yields extended permutations. Let us first give a general noun phrase rule:

\[ N \rightarrow \left[ \begin{array}{c}
\text{Det AN}^* \\
\text{N Post Modifier}
\end{array} \right] \] (20)

The right modifiers of the head noun in the noun phrase rule are intentionally ignored because they do not insert any word inside the string of the term. The general extended permutation is constructed through substitution as follows (measurement of temperature → measurement of the temperature):

\[(20) \circ (10) \left[ \begin{array}{c}
\text{N}_1 \\
\text{P N}_2
\end{array} \right] \rightarrow \left[ \begin{array}{c}
\text{N}_1 \\
\text{Det AN}^* \\
\text{N}_2
\end{array} \right] \] (21)

9We do not duplicate the compositions obtained by replacing the transformations involving two terms (11), (5) and (8) by the same transformations involving only one term and a noun phrase structure (12), (7) and (9). The resulting outputs are very similar and do not produce different strings.
<table>
<thead>
<tr>
<th>Equation</th>
<th>History</th>
<th>Type</th>
<th>Output string</th>
</tr>
</thead>
<tbody>
<tr>
<td>(5)</td>
<td>Coordination</td>
<td>AN₂ C AN N₁</td>
<td></td>
</tr>
<tr>
<td>(8)</td>
<td>Coordination</td>
<td>AN₂ N₃ C N₁</td>
<td></td>
</tr>
<tr>
<td>(11)</td>
<td>Insertion</td>
<td>AN₂ AN₃ N₁</td>
<td></td>
</tr>
<tr>
<td>(10)</td>
<td>Permutation</td>
<td>N₁ P N₂</td>
<td></td>
</tr>
<tr>
<td>(16)</td>
<td>(5) ∘ (11)</td>
<td>Coordination</td>
<td>AN₂ C AN₃ AN₄ N₁</td>
</tr>
<tr>
<td>(17)</td>
<td>(8) ∘ (11)</td>
<td>Coordination</td>
<td>AN₂ N₃ AN₄ C N₁</td>
</tr>
<tr>
<td>(18)</td>
<td>(11) ∘ (11)</td>
<td>Insertion</td>
<td>AN₂ AN₃ AN₄ N₁</td>
</tr>
<tr>
<td>(19)</td>
<td>(11) ∘ (11)</td>
<td>Insertion</td>
<td>AN₂ AN₃ AN₄ N₁</td>
</tr>
<tr>
<td>(21)</td>
<td>(20) ∘ (10)</td>
<td>Permutation</td>
<td>N₁ P Det N₂ N₁ P AN₃ N₂ N₁ P Det AN₃ N₂</td>
</tr>
</tbody>
</table>

Table 13: *Strings associated to the transformations of the term X₂ X₁*

From these compositions and from the elementary transformations, the strings of potential variants within a five-word window can be shown by collecting the words on the frontier of the syntactic trees (see Table 13). In order to simplify the following strings, the inserted words are replaced by their part-of-speech categories with the paradigmatic notations. More details could be given about the words such as the necessity of singular for the nouns and gerunds in a position of modifiers. This information could be mirrored by feature values and easily added to the preceding rules. Our intention was to sketch out the framework, not to provide all the details of the resulting description.

Similarly to basic variations, composite variations are a source of terminological acquisition. For example, when observing *systolic and diastolic blood pressures*, a variant of *systolic pressure*, the following terms can be proposed as candidates: *diastolic blood pressure*, *diastolic pressure* and *blood pressure*. However, an ambiguity remains when observing these variants about the history of their construction. Thus when observing *blood and bone marrow cell* as a variant of *blood cell*, the acquisition schemata used for the preceding example are not working any more. Using them would lead to the incorrect acquisition of *bone cell* and *marrow cell* which are not involved in this composition. This example reveals the terms *bone marrow* and *bone marrow cell*. The correct acquisition relies on a disambiguated structure of *bone marrow cell*. As outlined in the introduction, such a disambiguation can be performed through selectional restrictions. A fast, but less secure way to select the correct structure of *bone marrow cell* is to search in the corpus for other occurrences of *bone cell*, *bone marrow* and *marrow cell* [Strzalkowski and Vauhre, 1992, Bourigault, 1993]. The most frequently encountered structure(s) should indicate the correct structure. In order to avoid wrong prediction, a trade-off consists of producing the higher level term *bone marrow cell* without indications about its structure. Other techniques which can supply a disambiguation would naturally enhance the quality of the extraction.

7  “Undoing” derivations within noun phrases to find possible term variants

7.1 Overall presentation

Metarules, as exemplified above, are a means to generate recognition rules for syntactic variants of a rule representing a term. These rules share some lexical material of the original one. They present an overall structure more complex than the source rule ([N[ Sentence], Noun] versus [Sentence] production rule]) or different from it ([N[ Sentence] Production, Noun] versus [Sentence] production rule]).

Another approach for isolating possible term variants consists in trying to relate on syntactic grounds the noun phrases retrieved from the corpus by a general syntactic parser with the terms of a controlled thesaurus. More precisely, for any such noun phrase, the method is to “undo” the rules which have been applied to the sequence, and to check, for each step, if the simplified sequence is equal to an existing term.
If so, the retrieved noun phrase is considered as a possible syntactic variant of this term, and is added to
the list of possible variants for this term, along with the construction steps leading from the term to the
potential variant.

For instance, when finding the noun phrase non-monotone inference rule, it is possible first to remove
the modification of the adjective monotone by the adverb non-, leading to the derived sequence monotone
inference rule, then the modification of inference rule by the adjective monotone, which eventually leads
to an existing term: inference rule. It is then possible to add non-monotone inference rule as a possible
variant of inference rule, along with the indication of the rules necessary to derive the candidate variant,
and the order in which they are applied. The same process would lead to consider the noun phrase rather
simple inference rule as another candidate variant, associated with the same ordered list of relevant rules.
On the other hand, non-monotone inference language would not be considered as a possible variant, as the
simplification of this noun phrase does not yield an already existing term.

The aim of this part is to present this method of finding new variants for controlled terms. As this
method is currently developed and tested, it is not possible to assess it.

7.2 Using the grammar as a simplification device

A very simple context-free grammar for noun phrase such as:

1. NP → Det N
2. N → A N
3. A → Adv A
4. N → N N
5. N → inference — engine — language
6. A → monotone — simple
7. Adv → non — rather

is generally used to generate noun phrases such as simple language using rules 1, 2, 6 and 5, or to recognise
and parse a noun phrase (non monotone inference language, for instance).

Here, we propose to use the grammar to prune the derivation tree corresponding to a parse. Given the
parse:

\[ [\text{N}^{\text{A}}_{\text{Adv}} \text{ non-}]_{\text{A}} \text{ monotone}] \ [\text{N}^{\text{S}}_{\text{S}} \text{ inference}]_{\text{S}} \text{ engine}] \]

the rule 3 allows to replace the Adjective phrase with a simplified version of it. The modifying Adverb
is erased, leading to the structure:

\[ [\text{N}^{\text{A}}_{\text{A}} \text{ monotone}] \ [\text{N}^{\text{S}}_{\text{S}} \text{ inference}]_{\text{S}} \text{ engine}] \]

It is then possible to split this transformation according to rule 2, and each component (\([\text{A} \text{ monotone}\]
and \([\text{S} \text{ inference}]_{\text{S}} \text{ engine}]\)) is considered in turn, and its possible relation to the controlled terms
examined.

This pruning process\(^{10}\) associates two distinct actions: in the paradigmatic dimension, replacing a
phrase with a simplified version of it (the Adjective phrase minus the modifying Adverb for instance), in
the syntagmatic dimension, cancelling the concatenation (of the leaves) of the (partial) trees corresponding
to the right hand side of the rule\(^{11}\) (here, splitting the result of rule 2).

The grammar as such represents the only necessary knowledge for this simplification procedure. Nonethe-
less, it is fruitful to write the rules so as to facilitate the kind of transformations we want to obtain. For
example, the rule 1 could be written as:

\[ 1' \text{ N → Det N} \]

\(^{10}\)Which amounts as well to backtracking in a top-down derivation.

\(^{11}\)This operation implies that at each step, several sub-trees are isolated, and afterwards simplified. This feature can be
time and space consuming.
Representing the Determiner as a modifier, and the result as another N, and not as a NP is certainly arguable, on syntactic grounds. However, if we introduce rules such as:

8  N → N PP
9  PP → P N

the new version of rule 1 permits to erase a determiner inside a N constituent, and then to relate PP including a Determiner with PP excluding it (detail of a circuit with detail of circuit, example mentioned by [Spark and Jones and Tait, 1984]).

In addition, the grammar need not be restricted to a context-free one. On the contrary, it is possible to associate with the symbols of the context-free skeleton arbitrary constraints represented as equations [Shieber, 1986], and thus to use the full expressive power of unification grammars. This choice yields a greater flexibility in the matching process. For instance, a controlled term is stored in the root form (INFEERENCE ENGINE)\(^{12}\), while, during the simplification, the derived forms are replaced with the root ones (non monotone inference engines becoming non monotone inference engine)\(^{13}\). Smeaton and Sheridan [Smeaton and Sheridan, 1991] use similar techniques to care for inexact matching.

7.3 The simplification and matching process

Applying the rules of the grammar leads to a graph of construction trees. Each node of this graph is labelled with a tree\(^{14}\), the current parsing result, and each edge is labelled with the rule leading from its start node to its end node, that is from a tree to another one derived from it when applying the rule. For instance, from the construction tree \([\text{S}[\text{N}][\text{S}]])\), one edge, labelled with the rule \(A → \text{Adv } \bar{A}\), leads to the tree \([\text{S}[\bar{A}[\text{Adv } \bar{A}]][\text{S}]])\), and another one, labelled with the rule \(N → N \text{ PP}\), leads to the tree \([\text{S}[\bar{A}][\text{S}[\text{S}[\text{PP}]]])\).

The abstract trees associated with the nodes of this graph subsume the actual construction trees corresponding to terms or to syntactic variants of them. For instance \([\text{S}[\bar{A}[\text{Adv } \bar{A}]][\text{S}]])\) subsumes \([\text{S}[\bar{A}[\text{Adv } \bar{A} \text{ non-}][\bar{A} \text{ monotone}]][\text{S}[\text{N} \text{ inference}][\text{S} \text{ engine}]]\). As a matter of fact, a node in this graph is seen as an entry point to the different terms having the structure described in the node.

The process of simplification and matching is seen as \#\# moving towards the root of a derivation tree in order to simplify the current parse tree given by the analyser with the rule leading from the current node to its ancestor, possibly leading to a sequence already identified as a controlled term.

The simplification and matching procedure of a retrieved noun phrase \(x\) of structure \(y\) is described as follows:

- The retrieved noun phrase is a direct occurrence of a controlled term: Associated to the node \(y\) of the construction tree graph, there is a controlled term matching the noun phrase \(x\). The number of occurrences for this term is incremented.
- The retrieved noun phrase \(x\) is not a direct instance of a controlled term associated to the node \(y\):
  1. The current node is not the root of the construction tree graph. The grammar rules are used to simplify the current noun phrase. For each possibility of simplification, a triple is created: \([\text{original noun phrase}} \text{, simplified noun phrase}} \text{, sequence of rules leading from the original noun phrase to the simplified noun phrase}]].\) For instance, \([\text{N}[\text{A}[\text{Adv } \text{ non-}][\text{A} \text{ monotone}][\text{S}[\text{N} \text{ inference}][\text{S} \text{ engine}]]])\) would yield the three triples:

\[
\begin{array}{|c|c|c|}
\hline
\text{non-monotone inference engine} & \text{S}[\text{A} \text{ monotone}][\text{S}[\text{N} \text{ inference}][\text{S} \text{ engine}]] & (A → \text{Adv } A) \\
\hline
\text{non-monotone inference engine} & [\text{A}[\text{Adv } \text{ non-}][\text{A} \text{ monotone}]] & (N → A N) \\
\hline
\text{non-monotone inference engine} & [\text{S}[\text{N} \text{ inference}][\text{S} \text{ engine}]] & (NP → A N ) \\
\hline
\end{array}
\]

\(^{12}\)If and only if the derived forms have the same meaning, which is not the case, for instance, in standards organisation as opposed to standard organisation.

\(^{13}\)This feature is important as well to deal with typos and other spelling errors. It is in fact possible to find nominal compounds of this structure in which the modifying N is plural, violating the norm: \text{inferences engine}.

\(^{14}\)It is an abstract one, as no lexical lemmas are present.
Each such triple is then examined in turn\textsuperscript{15}, and compared with the node corresponding to the simplified tree. If a controlled term is associated with this node and matches the simplified tree, the original noun phrase is attached to it as a variant produced by the application of the sequence of rules present in the triple. In the given example, the third triple\textsuperscript{16} links at node $[s][s][s]$ \textit{inference engine} with the potential variant \textit{non-monotone inference engine}, provided that NP $\rightarrow$ AN is applied to it.

2. The current node is the root of the construction tree graph. The noun phrase $x$ is not a variant of a controlled term. It is thus attached to the node $y$ as an instance of this structure, for classification purposes (cf. below).

### 7.4 The need for other mechanisms than reverse derivation

By definition, this method can only relate sequences which are syntactic derivations of each other, and more precisely, complex sequences with less complex ones via the (reverse) application of grammar rules.

However, it does not permit to link terms to sequences which cannot be analysed as descendants of them. For instance, $[s][s$ rule] $[p$ of $[s$ inference]] does not share the structure of $[s][s$ inference] $[s$ rule]], even if the first one represents a very common paraphrase of the second one.

The metarules developed by Christian Jacquemin [Jacquemin, 1994] are used precisely for this purpose. A metarule such as $[s][s_1][p$ of $][s_2]] \Rightarrow [s][s_2][s_1]]$ is used to rewrite \textit{rule of inference as inference rule}, allowing thus to relate the first noun phrase to the controlled term.\textsuperscript{17}

The method described in the previous section must be completed:

- **Use of metarules:** if the current (possibly simplified) derivation tree can be transformed by a metarule in another tree, a new triple is created and processed. For instance, from the triple:

  \[
  \begin{array}{c|c|c}
  \text{engine of inference} & \text{[N [N engine] [p of [N inference]]]} & ()
  \end{array}
  \]

  a new one is produced by the application of the mentioned rule.

  \[
  \begin{array}{c|c|c|c}
  \text{engine of inference} & \text{[N [N inference] N engine]} & (\text{[N [N_1] [p of [N_2]]]} \Rightarrow \text{[N [N_2] [N_1]]})
  \end{array}
  \]

### 7.5 Related work

In the information retrieval field, two types of approaches trying to overcome the shortcuts of the string-based treatments and relying on parsing techniques are related to the above proposal. On the one hand, one can “simplify” the parsing results or the parsing itself so as to ease the matching process between the retrieved sequences and a controlled terminology or so as to simplify the resulting indexing. On the other hand, the matching process itself can rely on rather simple techniques. As the results of parsing are much too complex for a direct matching with a controlled terminology, for indexing or for retrieving documents corresponding to queries, the aim of the first approach is to normalise and simplify the retrieved sequences, so as to underscore dependency relationships between content words in the noun phrases, thus using these simplified representations to infer simple semantic relationships. Different techniques and representations are available: discontinuous grammars in which it is possible to skip “empty” words such as prepositions connectors [Basili \textit{et al.}, 1991], binary dependency trees with annotated nodes [Metzler and Haas, 1989] [Smeaton and Sheridan, 1991], leaving the notorious difficult prepositional attachments unresolved, inexact matching procedures [Smeaton and Sheridan, 1991]... It is then possible to look for controlled terms presenting the same semantic or conceptual relationships or to look for documents containing the same content words precisely in the same dependency relationships than in the query. However, by definition, this approach aims at ignoring the variation of terms, and hides the importance of the phenomenon. The CLARIT system [Evans \textit{et al.}, 1991] is an example of the second approach. It has focused as well on the problem of recognising morphological, lexical and semantic variants of terms. Noun phrases are retrieved by a context-free parser. Candidate terms are then matched against a first-order thesaurus of controlled

\textsuperscript{15}Note that care must be taken not to explore a simplification already examined via another path. The first triple will for instance yield the same hypothesis than the third one, if the adjective phrase is removed from it in the next step.

\textsuperscript{16}The box in the third position of the triple represents the constituent which is kept after the rule 2 has been split.

\textsuperscript{17}However, this new possibility can lead to cycles in the exploration of the construction tree graph.
domain-specific terminology. The matching process is however totally different. The controlled terms are not parsed, and the degree of matching is determined by calculating the number of windows common to each candidate term and each controlled term. A window is defined as a sub-sequence of a term. This scoring allows to partition the candidate sequences between exact terms (they are occurrences of already existing terms), general terms or possibly novel terms (for instance function-free horn-clause query related to the existing term horn-clause. In this second approach, there is a discrepancy between the richness of the input (the parsed noun phrases) and the crudeness of the matching process\(^\text{18}\). No attention is paid as well to the variation itself, thus offering no clues as to the most important syntactic rules used for it.

7.6 Assessing and classifying potential variants

The method described above represents just a rather straightforward means to relate controlled terms to sequences which could be syntactic variants of them. As it uses a grammar as only knowledge, even if this grammar is either a simple context-free one or a unification-based one, it implies only a thorough description of the noun phrase syntax. The resulting grammar can be used in different specialised sub-languages without dramatic need for adaptations. What’s more, there is no semantic knowledge involved.

However, the fact that a given noun phrase can be related to a controlled term via “undoing” the application of one or more rules does not mean that the noun phrase constitutes a variant of this term (as shown by the example rather simple inference engine below). Such an equivalence can be asserted only by an expert of the domain, on semantic and conceptual grounds. The first experiments show on the contrary that a lot of the related sequences share in fact structure and lexical material with controlled terms without having synonymy links with them. If the proportion of non relevant sequences is bothering, as far as precision is concerned, these sequences nonetheless help in enriching the controlled term variant list, possibly leading to a better recall.

Once the potential variants have been checked by a human expert, the two resulting sets are interesting result data, as to each correct or inadequate variant is attached the sequence of grammar rules necessary to derive the retrieved noun phrase starting from the term. The number of the rules which have been used, their complexity, their nature can be correlated with the status of the resulting noun phrase. The experiments mentioned above show for instance that coordination between terms scarcely leads to a new term. Observing the existing coordinations within noun phrases and their link with terms by way of this simplification method will possibly bring more evidence for this hypothesis. We can suppose as well that there is an empirical limit to the number of rules which are used successively to produce a successful variant of a term. Examining the spurious variants will permit to propose a value for this limit. On the other hand, the close examination of correct variants will underscore the syntactic operations which are most applied for term formation and variation.

8 Conclusion

Previous work [Jacquemijn & Royauté, 1994] had stated that terminological variation is a significant phenomenon in technical sublanguages, but a precise linguistic description remained to be done. The analysis presented in this study has followed the natural guideline of experimental linguistics: starting from corpus observation, making hypothesis, constructing a model respecting these assumptions and extending the model in order to contrast formal productions with observations. The results are promising. The axioms (the initial structures of base-terms) and the production rules (the rules for term variants) build an homogeneous system of formal linguistics which fairly covers the corpus utterances.

Although using notations inherited from generative grammarians, the model is very different. Instead of an architecture covering sentences and discourse, we propose a localist approach, lexicon-centred, intended at discovering the multiple aspects of variations. It aims at relating observed constructions to terms from a large knowledge base by extending soundly their local structure. Two immediate applications of this model are implicitly stated: it can be used for retrieving efficiently terms from large corpora and it can be used for

\(^{18}\)For instance, the candidate term autonomous robot navigation system yields the following list of matched terms: autonomous robot navigation system (exact match), robot navigation (general match), robot (general match), robot training device, connectionist system and robot system.
eliciting new terms from observed restricted noun phrases. More specifically, the linguistic specifications of term variation have been taken into account in:

- extraction of monolingual terminology ([Daille, 1994]) and bilingual terminology from aligned corpora ([Daille et al., 1994]),
- automatic term detection [Jacquemin & Royauté, 1994] and acquisition of thesaurial knowledge [Jacquemin, 1995].

Much work remains to be done in the field of terminological variation. Firstly, the description given in this paper is a partial description of the observed phenomena. It should be completed with more exhaustive rules and compositions. Secondly, the description is given in a purpose of analysis. Generation of term variant would require to account for selectional restrictions. For example, blood pressure and tyre pressure are two terms with a common head which cannot be coordinated *blood and tyre pressure*. Selectional restrictions would also help in disambiguating long terms in order to locate correctly the places where insertions take place. Finally, the study done for English should be extended to other languages. The Romance compounding being very different from the Germanic one, the variations observed in French are different from those reported in this study.

Whatever work remains to be done, this study was intended to provide strong basis for future works on the theme of terminological variation by proposing a formal system in which future advances can be embedded.

References


