In *Information Processing and Management*, **32**(4), 445–458. 1996.

WHAT IS THE TREE THAT WE SEE THROUGH THE WINDOW: A LINGUISTIC APPROACH TO WINDOWING AND TERM VARIATION *

Christian Jacquemin
Institut de Recherches en Informatique de Nantes (IRIN)
IUT de Nantes, 3 rue du Maréchal Joffre, 44041 Nantes Cedex 01, France
Phone: +33 40 30 60 52 — Fax: +33 40 30 60 53
jacquemin@iut-nantes.univ-nantes.fr

Running title: A Linguistic Approach to Windowing

Abstract – Windowing techniques play a key role in information retrieval. Previous works have suggested that the quality of access to information relies heavily on the characteristics of the windows. This study provides a linguistic approach to text windowing through an extraction of term variants with the help of a partial parser. The syntactic grounding of the method ensures that words observed within restricted spans are lexically related and that spurious word co-occurrences are ruled out with a good level of confidence. The system is computationally tractable on large corpora and large lists of terms. Illustrative examples of term variations from a large medical corpus are given. An experimental evaluation of the method shows that only a small proportion of co-occurring words are lexically related and motivates the call for natural language parsing techniques in text windowing.

1. INTRODUCTION

The notion of text window – a span of contiguous words within a document – is crucial for several corpus-based activities such as terminological acquisition or Information Retrieval (IR). In order to be efficient, windowing techniques must assess whether the words observed within a limited span compose of a chunk of lexically and syntactically related words such as *blood mononuclear cells* and not of an occasional aggregation of words without lexical motivation such as *cell and blood* in the sentence *The study of tumor cell and blood vessel adhesive interactions becomes essential* (...). More precisely, the words included in a window must constitute a correct partial noun phrase (NP) composed only of a head noun and (some of) its arguments. The second sequence *cell and blood* does not respect this constraint for it is built from *cell* a head noun and *blood* the argument of *vessel* another noun.

Both statistical and linguistic methods are used in such shallow analyses of texts but both suffer from complementary flaws: statistical methods detect lexical affinities but do not cope with syntax while Natural Language Processing (NLP) methods lack lexical information. In this paper, we present a tradeoff of partial text observation with a restricted syntactic parse linked to knowledge of lexical selections through

*We would like to thank the French scientific documentation center *INIST/CNRS* for providing us with data. All the experiments reported in this paper have been performed on [Pascal] a list of 71,623 multi-domain terms and [Medic] a 1.56-million word medical corpus composed of abstracts of scientific papers owned by *INIST/CNRS*. Many thanks also to Jean Royauté of *INIST/CNRS* for his helpful and friendly collaboration. His contribution is detailed in (Jacquemin & Royauté, 1994).

a recycling of term lists. A precise evaluation of the method is reported on a large corpus, and guidelines for future work on lexical and thesaurial acquisition are given in the conclusion.

2. WINDOWING AND TERM VARIATION

Smadja (1993) observes words co-occurring within a five-word window in order to detect strong lexical associations, generally called *collocations*. In his study, the collocates thus extracted must be postprocessed by a shallow parser in order to separate proper syntactic relations from irrelevant word co-occurrences. This two-step selection of collocates indicates that windowing does not provide safe lexical associations on its own.

Shrinking the window is the first way of ensuring fewer false collocates. In order to avoid a precise analysis of the structures highlighted by a window, Haas & Losee (1994) have focused on the optimal size of a span of words with respect to an IR task where text windows are matched with words of a query. These authors confirm that a span of three to five words is a good value. Above this limit, there is no confidence about the kind of relationship that can hold between the words appearing inside the window. Their study points out that an enrichment of the quality of the lexical relationship holding between the words of a window leads to an improvement in the performance of the retrieval task. However, when observing related words within a restricted span, the lexical, syntactic or semantic links that can hold between these words are very heterogeneous and ought to be classified. For example, *school medicine* has little to do with a *school for aerospace medicine*, if at all. In the former occurrence, the head word is *medicine* and *school* is its argument. Semantically, *school medicine* is a *medicine* specialized for children attending schools. Conversely, the head word of the latter occurrence is *school* and its argument *aerospace medicine* (which is in turn a term also denoting a specific kind of medicine). Thus a *school for aerospace medicine* is a specific kind of school but not of medicine.

A second way of accurately tagging the relation holding between the words, consists of using a general purpose NLP tool to parse the text under study, whether it comes from corpus or query. Apart from the high computational cost of such a process, it is difficult to be sure that the words found to constitute a NP really correspond to a concept in the domain. For example, *granulocyte macrophage colony stimulating factor* is a correct term unlike *poorly deformable dense cells* which is an occasional utterance without conceptual anchoring.

Indeed, both windowing and linguistic approaches to the detection of lexical relations aim at identifying multiword terms (generally compound nouns) and their variants in full text documents. The detection of such variants is not problematical when no words are inserted inside the string of the term. *Red blood cell* is an external variant of *blood cell* which does not alter the original phrase *blood cell* because its words remain contiguous and in the same order. For the purpose of identification, external variants can be ignored. Their study becomes necessary for semantic interpretation of variants; but, as we are mainly concerned with identification tasks, we will focus on internal variation in the remainder of this paper.

The range of internal term variations is potentially very large. An observation of non-contiguous occurrences of the words *tumor* and *cell* in the [Medic] [1] corpus is reported in Table 1. The correct variants of the phrase *tumor cell* are distinguished from occasional co-occurrences where both words either have no lexical relation or have a relation different from the one of the original phrase *tumor cell*. The pure windowing methods fail to separate correct variants from spurious ones because the distance between the words is not a correct criterion. Therefore, occurrences selected through windowing must be further filtered on more restrictive linguistic criteria. This paper gives a deeper insight into the different kinds of short distance modifications that can be applied to compound nouns and multiword terms. The words composing such correct variants bear genuine lexical relationships that can be safely and efficiently used in IR and lexical acquisition.

As there exist large lists of terms, it is conceivable to use them as reference NPs and to search for variable occurrences of these patterns with inflectional and syntactic variations. With this aim, we have developed *FASTR*, an NLP partial parser which recycles lists of multiword terms into grammar and retrieves morpho-syntactically correct variants of these terms from raw text corpora. The tool relies simultaneously on the observation of a restricted span of words and on a local syntactic analysis.

This paper is organized as follows: Section 3 reviews trends in the study of term variation. In section 4,

[1] [Medic] is a 1.56-million word medical corpus composed of abstracts of scientific papers owned by *INIST/CNRS*.

Table 1: *Correct and spurious variants of* tumor cell *from [Medic]*

	Window size and type	
	up to 5 words/original word order	up to 6 words/reverse word order
Correct variants	*tumor biopsy cells* *tumor contained normal diploid cells* *tumor derived cell* *tumor infiltrating cells* *tumor or nontumorous hepatic cells* *tumor target cells* *tumor tissue culture cell*	*cell into a metastatic tumor* *cells and their tumors* *cells form tumors* *cells from the primary tumor* *cells from three separate tumors* *cells in all five tumor* *cells in subcutaneous tumors* *cells in unperturbed tumors* *cells of solid metastatic human tumors* *cells present within hard tumors*
Lexically unrelated words	*tumor analog of mast cells* *tumor consisting of immature cells* *tumor enhancing) cell* *tumor marker, squamous cell* *tumor of b cell* *tumor of the myeloma cell* *tumor revealed an abnormal cell* *tumor size, cell* *tumor tissues used for cell* *tumor was a small cell* *tumor was of b cell* *tumor, intraabdominal desmoplastic small cell* *tumors and cell* *tumors and cultured cell* *tumors and normal cells* *tumors are of b cell* *tumors are squamous cell* *tumors formed by polyclonal cell* *tumors from 9L cells* *tumors have a similar cell* *tumors including renal cell* *tumors invaded by b cell* *tumors of adipose cells* *tumors or k562 cells* *tumors, a mediastinal germ cell*	*cell cancer, and other tumor* *cell clones and tumors* *cell cycle phases in the tumor* *cell DNA with the ultimate tumor* *cell line (which develops tumors* *cell line) tumor* *cell lines and 15 tumor* *cell lines and primary tumor* *cell lines established from tumor* *cell lines of rat tumors* *cell lung tumors* *cell mediated immunity can alter tumor* *cell odontogenic tumor* *cell surface receptors for tumor* *cell surface tumor* *cell testicular tumors* *cells; the principal tumor* *cells (MBE) with tumor* *cells are currently used for tumor* *cells could influence the tumor* *cells failed to induce tumor* *cells induced rapid development of tumors* *cells is considered important for tumor* *cells resulted in 100 tumor* *cells subcutaneously or intramuscularly developed tumors* *cells with the tumor* *cells, colonic polyps, tumor*

the different families of term variations are described through two-level metarules in *FASTR*: *paradigmatic* metarules and *filtering* metarules. Then a precise evaluation of the productivity and the quality of this description is provided in section 5 through an experiment on the [Medic] corpus with the [Pascal][2] list of terms.

3. BACKGROUND

According to Sager (1990) the creation of concepts consists of grouping real-world and mental objects into classes. Within a restricted domain, terms are used to represent concepts or notions. The non-compositionality of terms and their lack of ambiguity simplifies their automatic processing. Their interpretation within a sublanguage requires less attention to the context than would be necessary for words within general language (Bourigault, 1994). However, this simplification has to cope with two sources of variation:

1. *Type 1 variants.* The classes of objects of a restricted domain are conceptual clouds with a certain amount of overlap. The linguistic description of non central concepts requires a compositional modification of the terms such as *tumor target cells* or *cells in subcutaneous tumors* observed for *tumor cell* in Table 1. These variants can be defined as occurrences where the content words of the original terms are not modified (except inflections) but where its structure may vary.

2. *Type 2 variants.* The second source of variation is the existence of synonyms in any technical language. Some synonyms preserve the stems of the reference term as in *arterial pressure/pressure of the arteries* while some others substitute content words by semantically related ones such as *renal/kidney* (Dunham *et al.*, 1978). These variants are defined as occurrences where some of the content words of the original term are deleted or morphologically modified.

The two preceding categories of variations can interact and call for specific computational devices to be processed properly and exhaustively.

The approach to variation through NLP tools mainly concerns the type 1 variants and relies on the notion of lexical head and related arguments. As an illustration, in the NPs *information retrieval* and *retrieval of information*, the syntagmatic head noun *retrieval* dominates its argument *information*. The domination is reversed in a structure such as *information on retrieval*. Domination relations can be extracted from raw texts by NLP parsers through the detection of phrases and their heads. With this aim, Strzalkowski (1994) uses the *TTP* parser, derived from the Linguistic String Grammar of Sager (1981), to identify NPs in full text documents. Metzler & Haas (1989) have designed the *Constituent Object Parser* to analyze queries and corpora and Sheridan & Smeaton (1991) use a shallow parser developed in the framework of *Constraint Grammar* (Karlsson, 1990). In order to parse large corpora robustly, these parsers construct underdetermined parse trees where a part of the linguistic information (e.g. the resolution of the prepositional phrase-attachments within a complex NP) is not expressed. Metzler & Haas (1989) argue that this lack of precision is acceptable when queries and corpora are parsed by the same tool. Both have the same kind of ambiguity and their pairing mainly establishes that the domination relations are compatible, to avoid incorrect associations.

The cited works do not evaluate the quality and the accuracy of their parsers: Which types of domination relations are wrongly extracted? Which types of phrases are incorrectly bracketed as NPs? Which types of domination relations are ignored by such systems? Let us illustrate how difficult the extraction of such domination relations is through the following example: (...) *has been measured by examining the temperature* (...) *using a* [[$high_1$ $resolution_2$] $electron_3$ [$energy_4$ $loss_5$] $microscopy_6$]. The last NP, composed of one adjective and five nouns, is a structure typical of scientific or technical corpora. The correct bracketing of this structure is indicated in the sample sentence. An imprecision may remain about the domination relation between *electron* and *energy*: [[*electron energy*] *loss*] or [*electron* [*energy loss*]] could be a correct part of the structure. As *electron microscopy* is a correct term, we prefer the initial bracketing, it corresponds to the following dependencies between words: 1-2, 2-6, 3-6, 4-5, 5-6. Indeed there are $5 \times 4 \times 3 \times 2 = 120$ possible structures assuming that all the substructures are right-headed. It is not obvious, on pure linguistic grounds, to determine which is the correct one. Bourigault (1993) or

[2] [Pascal] is a list of 71,623 multi-domain terms used for manual indexing at the documentation center *INIST/CNRS*.

Strzalkowski & Vauthey (1992) have suggested to search in the corpus for word pairs which can help to disambiguate the structure. In our example, it would be useful to know which of the pairs *electron energy*, *electron loss*, *energy loss*, or *electron microscopy* occur most frequently in the corpus. The suggestion of these authors is a first step towards the concurrent use of syntactic information (NP structure) and subcategorization frames (Noun-Noun or Adjective-Noun preferential associations). The application of subcategorization to the disambiguation of $N_3\,N_2\,N_1$ compounds is investigated in Resnik (1993) but its extension to a systematic disambiguation of complex NPs has to be preceded by the two following tasks:

- the extraction from technical corpora of subcategorization information for any technical domain,
- the study of conflict resolution strategies in case of competing associations.

Leacock *et al.* (1993) propose a method for extracting automatically contextual representations from large corpora. Two types of contexts are extracted: local information on words immediately surrounding a word and topical information on substantive words that are likely to co-occur in the same sentence. Local context is composed of templates: short sequences of words extracted from disambiguated occurrences. For example, *telephone line* and *access line* are templates for the word *line* with the phone line meaning and *new line* is a template for the same word with the product line meaning. Such information is likely to constitute useful clues for the structural disambiguation of long NPs.

In (Sparck Jones & Tait, 1984) the second aspect of terminological variation (type 2 variants with morphologically distinct but semantically related words) is accounted for by producing as many alternative phrases of a query as possible. The variants are generated with the help of a semantic interpreter. In their opinion, a correct term variant is any syntactic construction whose interpretation is equal to the original one. Two assumptions of this method can be criticized. First, is it realistic to suppose that a semantic interpretation can be constructed for any utterance? Secondly, is the set of constructions corresponding to a given meaning small enough to consider the exhaustive generation of variants as computationally tractable? Another account for the second aspect of variation consists of grouping together semantically related words (Grefenstette, 1994). Assuming that words occurring in similar contexts (e.g. as subjects of the same verbs or as modifiers of the same nouns) tend to have a similar meaning in the considered domain, semantic similarities can be detected by grouping together words with similar contexts. Grefenstette's work mainly applies to single terms or to fixed multi-word terms and concerns synonymy links acquired through an external observation of terms. Conversely, the acquisition of hypernymy links has to cope with the internal syntax of terms. In *CLARIT* (Evans *et al.*, 1991), specialization or generalization links between candidate and reference terms are detected through partial matches between terms.

The different works on term variation or term variability can be divided into works focusing on synonymy relations of linguistically different terms covering type 2 variants (Grefenstette, 1994; Sparck Jones & Tait, 1984) and works focusing on hypernymy relations of term with linguistically related expressions covering type 1 variants or type 2 variants only with deleted content words (Evans *et al.*, 1991; Sheridan & Smeaton, 1991). Our study is related to the second family and our approach to this aspect of variation lies halfway between an NLP tool and a partial matcher. It takes from the latter a consideration for variation with respect to a list of reference terms and from the former a description of variation through (local) syntactic (meta)rules respecting the domination relations.

As indicated above, we only consider internal variants where words are inserted inside the string of the reference term. Our study could be extended to external modification – the tool is general enough for this purpose. As external modification is less constrained than internal one, it would be necessary to have a good characterization of the linguistic context of reference terms. Local contexts, extracted from large corpora (Leacock *et al.*, 1993), reveal selectional restrictions and could be exploited for selecting relevant external modifiers. Contrary to external variants, variants where content words are elided cannot be extracted in the framework proposed in this study: Without a correct identification of anaphora, the processing of elided variants is inefficient due to its low precision rate.

4. *FASTR*: A PARTIAL PARSER FOR TERM AND TERM VARIANT EXTRACTION

The parser used to extract terms and their variants from corpora is an optimized left-corner unification-based parser. It is devised for three-level grammars composed of a lexicon of single words, a grammar of terms and a metagrammar of variants. The formalism is an extension of *PATR-II* (Shieber, 1986), a

standard constraint-based framework for writing NLP-oriented grammars in a logical form. The choice of a declarative formalism has the advantage over lower level tools such as transducers to offer a comfortable and legible formalism to the user. Symmetrically, the unification-based parsers are slow and therefore must be completed with optimization devices in order to be computationally tractable. The implementation of *FASTR* is detailed in (Jacquemin, 1994a) and (Jacquemin, 1994b) and its optimization through lexicalization is reported in (Jacquemin, 1994c). The resulting application is fast enough to process large amounts of terminological and textual data. *FASTR* parses 2,562 words/minute on a Sparc 2 workstation when working with a lexicon of 38,536 single words, a list of 71,623 terms and a metagrammar of 102 metarules. This section details the formalism and the description of variation in *FASTR* before evaluating the results in the next section.

Usually unification-based formalisms such as *HPSG* (Pollard & Sag, 1987) are composed of two levels: a lexical level (single words) and a syntactical level (grammar rules). However, such formalisms do not directly account for complex lexical entries e.g. compounds, terms, verbal locutions, etc. An important exception is the formalism of *Lexicalized Tree Adjoining Grammars* (*LTAGs*) described in (Abeillé & Schabes, 1989). In *LTAGs*, the lexical level is composed of single words and complex lexical entries represented by partially saturated pieces of syntactic structures. Similarly the formalism adopted for *FASTR* represents multiword terms through lexicalized grammatical rules (see rule (1)). As in *PATR-II*, rules are composed of a context-free skeleton denoting the concatenation of the constituents and a set of equations constraining the information on these constituents. Rules in *FASTR* are not restricted to immediate dependency but can represent arbitrary deep structures as in *LTAGs*. Rule (1) represents the NP *X ray analysis* as a Noun-Noun structure ($N_2\,N_5$) where the first constituent *X ray* is an embedded Noun-Noun structure ($N_3\,N_4$):

$$
\begin{array}{ll}
\text{Rule} & N_1 \rightarrow (N_2 \rightarrow N_3\ N_4)\ N_5\ : \\
& < N_1\ lexicalization > \doteq \text{`}N_5\text{'} \\
& < N_1\ label > \doteq \text{`}005223\text{'} \\
& < N_3\ lemma > \doteq \text{`}X\text{'} \\
& < N_3\ inflection > \doteq\ 5 \\
& < N_4\ lemma > \doteq \text{`}ray\text{'} \\
& < N_4\ inflection > \doteq\ 1 \\
& < N_5\ lemma > \doteq \text{`}analys\text{'} \\
& < N_5\ inflection > \doteq\ 7.
\end{array}
\qquad (1)
$$

To each single word appearing in a multiword term rule must correspond a single word rule such as the one given in (2) for the word *analysis*:

$$
\begin{array}{ll}
\text{Word} & \text{`}analys\text{'}\ : \\
& < cat > \doteq \text{`}N\text{'} \\
& < inflection > \doteq\ 7.
\end{array}
\qquad (2)
$$

One of the objectives assigned to *FASTR* is to recycle terminological data which result from human activities with concern for terminology: IR, standardization, translation, lexicography, etc. Before their integration into *FASTR*, terms have to be tagged and lemmatized.[3] Then the tagged multiword terms are automatically transformed into as many rules as terms. The single words composing the terms are transformed into single word rules. The complete process of the integration of terms into the NLP tool (rule creation from terms and rule compiling) is fully automatic. Human tuning is only required for the description of variations through metarules.

When focusing on type 1 variants, variations can be defined as local syntactic modifications of multiword terms yielding conceptually related occurrences. Only type 1 variations preserving the content words are studied here because this aspect is very productive and represents a good field of investigation on its own. More complex variations such as type 2 variations involving morphological derivation or substitution of synonymous words cannot be properly studied without acquiring firstly a good expertise in extracting pure syntactic variants.

[3] Jean Royauté from *INIST/CNRS* has used the *DELAF* lexical database for English to tag the [Pascal] list of terms. *DELAF* is developed and maintained by the *LADL* laboratory of University of Paris 7 (Courtois, 1990).

Correct processing of type 1 variation relies on the ability to relate variants, and only variants, to the original structures. Metarules in *FASTR* are a compromise between a pure two-level description that could be given by transducers and a pure syntactic framework as proposed by *TAGs*. There is however a continuity between these three tools: metarules can be easily transformed into transducers as suggested for phonology in (Kay, 1983) – but can also be seen as compiled compositions of elementary rules.

4.1 *Inflectional variation*

When preserving the content words, variation can be separated into two components: inflectional variation and syntactic variation. Inflections of words are processed due to inflectional information provided to the parser through the *inflection* feature attached to single words (see single word rule (2) and term rule (1)). The value of this feature (an integer), paired with the part-of-speech category of the word, refers to a list of affixes and features corresponding to the set of inflections of the word (more details can be found in (Jacquemin, 1994c) about inflectional morphology in *FASTR*). In the remainder of the paper, the stress is laid on syntactic variants. The morphological analyzer of *FASTR* associates to each word its possible homographs, each of them being represented as a lemma with features.

4.2 *Syntactic variation*

In order to conceive a tool for processing syntactic variation of terms, it is worth providing the user with a way of heuristically refining her/his description. We propose a two-step description where the first step is the creation of a generic set of unconstrained metarules called *paradigmatic metarules*. In this stage, a sequence of words is a variant of a given NP if, and only if, the argument(s) and the head word of the basic NP are separated by less than an arbitrary number of words. That is to say that this stage roughly corresponds to windowing. When restricted to two-word terms, this definition corresponds to the notion of *flexible collocations* of (Smadja, 1993).

Paradigmatic metarules can be grouped into three classes with a syntactic flavor: coordinations, insertions and permutations. Indeed, coordinations and insertions are both variants where the word order of the original term is preserved and where one to three words are inserted inside the term string. Coordinations are a subset of the insertions where a coordinating conjunction must either begin or end the sequence of inserted words. As coordination metarules are tried prior to insertions, it can be assumed that the inserted sequence in insertions neither begins nor ends by a coordinating conjunction. For two-word terms, metarule (3) is an example of coordination and metarule (4) an example of insertion: [4]

$$\text{Metarule} \quad Coor(X_1 \rightarrow X_2\ X_3) \quad \equiv \quad X_1 \rightarrow X_2\ C_4\ X_5\ X_3\ : . \tag{3}$$

$$\text{Metarule} \quad Ins(X_1 \rightarrow X_2\ X_3) \quad \equiv \quad X_1 \rightarrow X_2\ X_4\ X_5\ X_3\ : . \tag{4}$$

A metarule is composed of a left-hand part, the source, which matches (is unified) with a term rule and a right-hand part, the target, which yields the transformed rule. When applied to rule (1) metarule (3) yields a new rule accepting any sequence *X ray* C_4 X_5 *diffraction* such as *X ray or neutron diffraction*. Metarule (4) accepts both *tumor tissue culture cell* and *tumors are squamous cell* as variants of *tumor cell*. The first one is a correct variant while the second one is a fortuitous co-occurrence.

Permutations are insertions where the word order of the NP is reversed and where one to four words are inserted between its words. Linguistically, these variations correspond to the shift from a compound construction to a syntagmatic one. For example, metarule (5) associates the variant *cells from peripheral blood* to the term *blood cell*:

$$\text{Metarule} \quad Perm(X_1 \rightarrow X_2\ X_3) \quad \equiv \quad X_1 \rightarrow X_3\ X_4\ X_5\ X_2\ : . \tag{5}$$

The description of variation through *paradigmatic metarules* is too loose and has the drawback to accept a too wide range of variants. For example, in the sentence (...) *the liquid solid transition of alkali metals is examined* (...), the sequence *transition of alkali metals* is wrongly given by (4) as a variant of *transition metals*. An extensive interpretation of the sentence is not necessary to rule out this utterance as not conceptually related to *transition metals*. Due to the presence of the preposition *of*, a local syntactic

[4] In the following formulæ, category X stands for an undetermined category, C for a coordinating conjunction, P for a preposition and V for a verb. $\neg$ is negation.

analysis shows that *metal* dominates *transition* in the original term while the relation is reversed in the spurious variant *transition of alkali metals*.

To remedy the lack of accuracy of *paradigmatic metarules*, the first stage of the description through *paradigmatic metarules* has to be enriched with *filtering metarules*. These metarules encompass a finer syntactic knowledge and impose new constraints to the inserted elements through informational equations. The adjustment and debugging of *filtering metarules* is an exercise similar to the description of NPs in *NPtool* (Voutilainen, 1993) for partial parsing or to the description of terms in *LEXTER* (Bourigault, 1994) for terminological acquisition. Both studies propose a dual description composed of positive criteria selecting maximal constructions coupled with negative restrictions filtering out safe correct constructions. In *NPtool* two sets of rules called *NP-hostile* or *NP-friendly* compete for the interpretation of a sentence. The candidate NPs are the structures that are agreed upon as NPs by the two competing analyses. Contrary to most NLP tools, *LEXTER* describes NPs through their frontier rather than their internal structure. In this application, *NP-friendly* external segmentation rules yield maximal candidate NPs. These maximal structures are further processed by *NP-hostile* decomposition rules and broken into minimal unambiguous canonical NPs.

Similarly, *filtering metarules* in *FASTR* are dually subdivided into *positive* and *negative metarules*. The structures which are accepted as correct variants are the ones which are not selected by any of the *negative metarules* and which are accepted by at least one *positive metarule*. In this configuration, *negative metarules* are used to select among the structures accepted by too loose *positive metarules* the ones which have to be ruled out. For example, the two metarules (6) and (7) describe correct two-word term permutations with two inserted words:

$$\begin{array}{ll} \text{Metarule} & NPerm(X_1 \rightarrow X_2\ X_3) \equiv X_1 \rightarrow X_3\ X_4\ X_5\ X_2\ : \\ & \neg(< X_2\ tense > \doteq \text{'gerund'}) \\ & \neg(< X_2\ cat > \doteq \text{'V'}) \\ & \neg(< X_2\ cat > \doteq \text{'N'}) \\ & < X_4\ cat > \doteq \text{'P'}. \end{array} \quad (6)$$

$$\begin{array}{ll} \text{Metarule} & PPerm(X_1 \rightarrow X_2\ X_3) \equiv X_1 \rightarrow X_3\ X_4\ X_5\ X_2\ : \\ & < X_4\ cat > \doteq \text{'P'}. \end{array} \quad (7)$$

Positive metarule (7) adds a constraint to the *paradigmatic metarule* (5) by specifying that the first word inserted has to be a preposition. For example, *groups received homologous blood* is not granted as a variant of *blood group* by (7) because *received* is not a preposition. This *filtering metarule* is further refined negatively by (6) which rules out permutations where X_2 is neither a noun nor a verbal gerund. Thus *measurement of noninvasive continuous* which would be accepted by (7) as a variant of *continuous measurement* is filtered by the *negative metarule* (6) because *continuous* is an adjective. Through experiments, it can be observed that the couple (7)/(6) is not selective enough and could be further tuned up by differentiating prepositions possibly introducing a nominal argument such as *in*, *of* or *on* from prepositions such as *after* which can only introduce locative or temporal complements or verbal arguments.[5] Such a refinement would avoid (6) to consider wrongly *hospital after 3 days* as a correct variant of *day hospital*.

5. EXPERIMENTAL EVALUATION OF TERM VARIANT DETECTION IN *FASTR*

An extraction of term and term variant occurrences has been carried out with *FASTR* on the 1.56-million word corpus [Medic] composed of medical abstracts. No preprocessing of the corpus has been necessary. The parser has been fed with the list of 71,636 scientific terms [Pascal].

The set of the core *paradigmatic metarules* used for this experiment was composed of 36 coordinations, 18 insertions and 24 permutations. 13 metarules concerned two-word terms, 26 were for three-word terms and 39 for four-word terms. The metarules for n-word terms (with $2 \leq n \leq 4$) were restricted to windows of $n+4$ words for insertions, $n+5$ words for coordinations and $n+6$ words for permutations. The different spans for the latter two transformations are due firstly to the insertion of a coordinating conjunction into coordinations and a preposition into permutations and secondly to the syntagmatic nature of permutations.

[5] We are grateful to Owen Rambow of University of Paris 7 and UPenn for bringing this difference to our attention.

In order to evaluate the benefit of larger windows, these core metarules have been completed with 24 metarules one word wider.

To each *paradigmatic metarule* of coordination or insertion has been associated a single *positive filtering metarule* for selecting the correct variants. The filtering of permutations was carried out by assigning to each *paradigmatic metarule* a pair of *filtering metarules*: a *positive* and a *negative* one. The set of *filtering metarules* was composed of 24 *negative metarules* and 78 *positive* ones plus 6 *negative* and 24 *positive extended metarules.*

Most of the research work on *FASTR* has been devoted to the conception of the parser and it took only about one month to tune up the metarules corresponding to the results evaluated in this paper. The quickness of the writing of metarules is due to their repetitions. The constraints stated for metarules depend more on the linguistic characteristics of the variation than on their size; each constraint discovered for a specific coordination or insertion metarule can be extended to the whole set of coordination or insertion metarules. As will be pointed out in the evaluation, this generalization is less successful for permutations which require different descriptions for wider variants.

The *paradigmatic metarules* produce 10,229 *variants* and 1,683 *extended variants.* By processing them with *filtering metarules*, these variants are separated automatically into 6,247 *positive variants* and 3,982 *negative variants* and 1,462 *positive* and 221 *negative extended variants.* All these variants have been scanned by hand in order to detect the false *positive* and the false *negative* ones. These incorrect occurrences have received two labels: either *wrong* for those which could be undoubtedly qualified as incorrect or *uncertain* for those which could not be definitively qualified as incorrect without an appeal to context observation or medical expertise. For example, *accessory cell depleted normal spleen* is a *wrong positive variant* (insertion) of *accessory spleen* while *ultrasonic reflection mode CT method* is an *uncertain positive variant* (insertion) of *ultrasonic method.* Similarly *blood CD4, CD8 cells* is a *wrong negative variant* (insertion) of *blood cell* while *herpex simplex, varicella zoster* is an *uncertain negative variant* (insertion) of *herpex zoster.* In our evaluation of the quality of the results, both *wrong* and *uncertain variants* have been considered as incorrect. They are only differentiated between in the following point 6 concerning the possibility of improving the metarules. *Wrong* occurrences could be correctly extracted through more accurate metarules while *uncertain* occurrences require a different framework to be processed correctly.

The observation of the results has led us to formulate the following remarks:

1. In our medical corpus, one word in every fifteen (6 %) belongs to an occurrence of a multi-word term or one of its variants. Multi-word terms represent an important part of the text surface of technical corpora and their knowledge is crucial for NLP tasks on technical domains such as machine translation (Boesefeldt & Bouillon, 1992). Table 2 gives an evaluation of the surface covered by terms and variants in the 1.56-million word corpus [Medic]. All the multi-word term occurrences are correct ones, but term occurrences have not been manually checked and their actual rate should be slightly lower.

2. Table 2 compares multi-word term occurrences with multi-word term variant occurrences. It reveals that 15 % of the multi-word term occurrences are term variants. Variation is numerically significant and has to be accounted for in any task aiming at extracting term occurrences. There has been a debate in IR about the interest of using phrases for addressing the content of documents. Fagan (1989) reported a substantial improvement of the results through the use of phrases while Lewis & Croft (1990) were much less confident about its efficiency. The high rate of variation as well as the difficulty to select correct variants through surface co-occurrences may be responsible for this uncertainty. Point 4 confirms this opinion by reporting a high rate of co-occurrences which have to be rejected as incorrect variants.

3. An observation of the distribution of the categories of variation reveals that insertions are twice as numerous as permutations and six times as numerous as coordinations. Two-word term variations represent 90 % of the variants and three-word term variations represent 8 %. The variants of terms of four or more words can be regarded as insignificant. Table 3 details the different types of correct variants according to the size of the term and to the type of morpho-syntactic variation.

4. In a framework where collocations are defined as the co-occurrence of two words X_1 and X_2 within a restricted window, 36 % of the collocates are not variants of $X_1\ X_2$. Moreover 54 % of the permuted

Table 2: *Multi-word term occurrences and correct term variant occurrences in [Medic].*

	Term occurrences				Term variant occurrences		
	2-word terms	3-word terms	4-word terms	≥5-word terms	2-word variants	3-word variants	4-word variants
# occurrences	31,917	3,968	377	34	5,957	530	65
% occurrences	74.5 %	9.2 %	0.9 %	0.1 %	13.9 %	1.2 %	0.2 %
Total %	84.7 %				15.3 %		
text surface (words)	63,834	11,904	1,508	178	22,210	2,141	329
% text surface	4.1 %	0.8 %	0.1 %	0.0 %	1.4 %	0.1 %	0.0 %
Total %	5.0 %				1.5 %		
Total %	6.5 %						

Table 3: *Distribution of the categories of term variants according to the size of the terms or to the type of syntactic variation in [Medic].*

	Coordination	Insertion	Permutation	Total	Percentage
2-word terms	536	3,593	1,828	5,957	90.9 %
3-word terms	66	384	80	530	8.1 %
4-word terms	1	59	5	65	1 %
Total	603	4,036	1,913	6,552	
Percentage	9.2 %	61.6 %	29.2 %	100 %	

Table 4: *Rate of rejected* paradigmatic variants *in [Medic].*

	Coordination	Insertion	Permutation	Total
2-word terms	25.5 %	26.1 %	54.6 %	37.6 %
3-word terms	4.3 %	5.7 %	37.5 %	12.3 %
4-word terms	50 %	4.8 %	5.8 %	14.5 %
Total	17.3 %	24.3 %	54.1 %	35.9 %

Table 5: *Precision of term variant extraction from [Medic] as a function of the number of content words in the window.*

	Coordination	Insertion	Permutation	Total
3-word window	96.9 %	98.7 %	67.5 %	92.5 %
4-word window	92.0 %	91.2 %	76.7 %	84.1 %
5-word window	86.6 %	83.1 %	76.1 %	79.9 %
6-word window			77.9 %	77.9 %
Total	95.5 %	96.4 %	72.6 %	89.3 %

collocates (X_2 followed by X_1 within a six-word window) are spurious variants of $X_1\,X_2$. This important difference between collocates and variants suggests that the use of windowing for lexical acquisition purposes such as (Church & Hanks, 1989) has to cope with numerous lexically unrelated occurrences. The difference between collocates and variants is significantly lower for three-word collocates (12 %) than for two-word collocates (38 %). When more words are involved in a co-occurrence, the probability to observe a variant of the basic structure raises significantly. Correct term variants have been obtained by selecting among the *paradigmatic variants* (collocates) the occurrences which are conceptually related to the original term. Table 4 presents the rate of rejection of such collocates. Due to the very low number of occurrences of four-word term variants, the four-word term results are not informative.

5. With a similar effort of description, coordinations and insertions are retrieved with high accuracy, while permutations would require a better treatment of syntax. The low precision for permutations (73 %) reveals a limit of our non-contextual and weakly syntactic description of variation. Terms are compounds, therefore coordinations and insertions, which preserve the structure of the reference terms, are also compounds. Conversely, permutations are syntagmatic constructions which must be described in a syntactic framework. In *FASTR* we have chosen to give an advantage to precision over recall. For applications such as automatic indexing, it seems that silence has to be preferred to noise. Therefore, strong *filtering metarules* are used in order to keep a correct level of precision for permutations; it results in a ruling out of some correct syntactic constructions. For example *rate of progression of renal failure* is wrongly rejected as a variant of *failure rate* because it includes a two-step domination link between *rate* and *failure*. However, some $N_1\,P_2\,N_3\,P_4\,A_5\,N_6$ must be correctly rejected as variants of $N_6\,N_1$ because the domination is not necessarily a transitive relation. For example *volume of water in arterial blood* is not a variant of *blood volume*. Tables 5 and 6 report the precision and the recall rate of term variant retrieval with *FASTR*. These values have been calculated by checking manually all the accepted and rejected *paradigmatic variants*. Thus precision is the proportion of the *positive variants*[6] which are true *positive variants*. Recall is the proportion of correct variants (true *positive variants* or false *negative* ones) which are retrieved by *FASTR*.

6. As suggested previously, improvement of *FASTR* only concerns *wrong* occurrences because the detec-

[6] *Positive variants* are *paradigmatic variants* which are accepted by a *positive filtering metarule* and not ruled out by any of the *negative* ones.

Table 6: *Recall of term variant extraction from [Medic] as a function of the number of content words in the window.*

	Coordination	Insertion	Permutation	Total
3-word window	98.0 %	98.3 %	95.2 %	97.8 %
4-word window	89.7 %	73.3 %	76.4 %	76.2 %
5-word window	72.2 %	62.6 %	57.0 %	60.4 %
6-word window			32.4 %	32.4 %
Total	94.9 %	90.8 %	70.1 %	85.1 %

Table 7: *Best improvement of performances that can be expected from* FASTR *on [Medic].*

	Coordination	Insertion	Permutation	Total
Maximal precision	97.5 %	98.4 %	93.1 %	96.6 %
Maximal recall	97.2 %	97.5 %	89.0 %	94.7 %
Maximal recall improvement	4.5 %	9.3 %	50.7 %	18.8 %

tion of *uncertain*-labeled occurrences calls for a knowledge which cannot be reduced to the internal syntax of variants. Thus the best improvement that can be expected from *FASTR* is to detect correctly *wrong negative* and *wrong positive* occurrences. An important effort has to be made on permutations where up to 50 % more occurrences could be retrieved. Conversely, the description of insertions and more specifically coordinations is almost optimal and little improvement has to be expected from more accurate metarules on these categories. With "perfect metarules", neither precision nor recall would be under 90 % for any of the categories variants. Table 7 is an evaluation of the best performances that can be expected from *FASTR* assuming that all the *wrong* occurrences are correctly detected through finer metarules.

7. A count of the number of variants per term is shown in Table 8: 85 % of the terms accepting variants have less than three variants in [Medic] and 60 % have only one variant. Due to repetitive variants, the 6,552 variants observed in [Medic] are produced by 1,818 terms. It seems hard to find a correlation between the quantity of variants and their quality. For example *arterial pressure* has 36 correct variants and 3 incorrect ones while *tumor cell* has 17 correct variants and 52 incorrect ones (see Table 1).

8. An extension of the size of the window in which variants are observed would only slightly enhance the recall but would seriously degrade precision. Table 9 reports the precision and the recall that can be expected from an extension of the word span. These results have to be compared with the results from Tables 5 and 6.

Table 8: *Number and percentage of terms with a given number of variants in [Medic].*

# variations	1	2	3	4-5	6-9	10-14	15-29	30-99	100-250
# terms	1077	334	141	117	77	26	28	14	4
		1552			220			46	
Percentage	59.2 %	18.4 %	7.8 %	6.4 %	4.3 %	1.4 %	1.6%	0.7%	0.2%
		85.4 %			12.1 %			2.5 %	

Table 9: *Precision and recall with a one-word wider span of words on [Medic].*

	Coordination	Insertion	Permutation	Total
# correct variants	34	199	304	537
Precision	68.0 %	50.7 %	72.6 %	58.8 %
Recall	50.0 %	34.2 %	14.8 %	24.2 %

6. CONCLUSION AND FUTURE RESEARCH

This study has described and evaluated a method for selecting linguistically motivated textual sequences from large corpora corresponding to internal variations of attested multiword terms. The parser used for this task mixes conceptual information (term lists) with local syntactic filters (metarules).

Only short development times of metarules ensure a high quality of extraction provided that the observation is restricted to a window of five or less content words. An extension of the size of the window would reduce precision without enhancing significantly the number of retrieved occurrences. Variation has to be considered as a local phenomenon and the proposed framework is well-suited for such a local observation. Precision has been preferred over recall. An improvement of the recall rate would require a better long distance syntactic analysis through an integration of a syntactic parser.

The direct application of term variant extraction obviously concerns automatic indexing because terms and variants are good candidates for representing the content of a document. More generally, most of the applications of NLP to IR would benefit from a precise and exhaustive detection of terms and compounds. Secondly, term variation can provide a basis for term acquisition: *inflammatory and erosive joint disease* is a variant of *inflammatory joint disease* which reveals *erosive joint disease* as a candidate term and which, simultaneously, indicates that both terms are conceptually close. Research is currently being carried out on terminological acquisition through variation processing.

One of the most important conclusions of this study is that term variation and text windowing are difficult tasks which call for precise terminological and linguistic knowledge. There is still a long way to go for a correct synergy in the concurrent processing of information and language. By providing an accurate description of local linguistic variation of terms, this study is another step in this direction.

REFERENCES

Abeillé, A., & Schabes, Y. (1989). Parsing Idioms in Lexicalized Tags. In *Proceedings, 4th Conference of the European Chapter of the Association for Computational Linguistics (EACL'89)*, Manchester, 1–9.

Boesefeldt, K., & Bouillon, P. (1992). Une représentation sémantique et un système de transfert pour une traduction de haute qualité. In *Proceedings, 14th International Conference on Computational Linguistics (COLING'92)*, Nantes, 998–1002.

Bourigault, D. (1993). An Endogeneous Corpus-Based Method for Structural Noun Phrase Disambiguation. In *Proceedings, 6th European Chapter of the Association for Computational Linguistics (EACL'93)*, Utrecht, 81–86.

Bourigault, D. (1994). *LEXTER un logiciel d'EXtraction de TERminologie. Application à l'extraction des connaissances à partir de textes.* Ph.D. diss. in Mathematics and Computer Science, École des Hautes Études en Sciences Sociales, Paris.

Church, K.W., & Hanks, P. (1989). Word Association Norms, Mutual Information and Lexicography. In *Proceedings, 27th Annual Meeting of the Association for Computational Linguistics (ACL'89)*, Vancouver, 76–83.

Courtois, B. (1990). Un système de dictionnaires électroniques pour les mots simples du français. *Langue Française, 87*, Paris: Larousse, 11–22.

Daille, B. (1994). *Approche mixte pour l'extraction de terminologie : statistique lexicale et filtres linguistiques.* Ph.D. diss. in Computer Science, University of Paris 7, Paris.

Dunham, G.S., Pacak, M.G., & Pratt, A.W. (1978). Automatic Indexing of Pathology Data. *Journal of the American Society for Information Science, 29*(2), 81–90.

Evans, D.A., Ginther-Webster, K., Hart, M., Lefferts, R.G., & Monarch, I.A. (1991). Automatic Indexing Using Selective NLP and First-Order Thesauri. In *Proceedings, RIAO'91, Conference on Intelligent Text and Image Handling*, Barcelona, 624–643.

Fagan, J.L. (1989). The effectiveness of a nonsyntactic approach to automatic phrase indexing for document retrieval. *Journal of the American Society for Information Science*, *40*(2), 115–132.

Grefenstette, G. (1994). *Explorations in Automatic Thesaurus Discovery.* Dordrecht: Kluwer Academic Publisher.

Haas, S.W., & Losee Jr, R.M. (1994). Looking in Text Windows: Their Size and Composition. *Information Processing & Management*, *30*(5), 619–629.

Jacquemin, C., & Royauté, J. (1994). Retrieving Terms and their Variants in a Lexicalized Unification-Based Framework. In *Proceedings, 17^{th} Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'94)*, Dublin, 132–141.

Jacquemin, C. (1994a). Optimizing the Computational Lexicalization of Large Grammars. In *Proceedings, 32^{nd} Annual Meeting of the Association for Computational Linguistics (ACL'94)*, Las Cruces, 196–203.

Jacquemin, C. (1994b). *FASTR*: A Unification-Based Front-End to Automatic Indexing. In *Proceedings, Intelligent Multimedia Information Retrieval Systems and Management (RIAO'94)*, New York, 34–47.

Jacquemin, C. (1994c). Recycling Terms into a Partial Parser. In *Proceedings, 4^{th} Conference on Applied Natural Language Processing (ANLP'94)*, Stuttgart, 113–118.

Karlsson, F. (1990). Constraint grammar as a framework for parsing running text. In *Proceedings, 13^{th} International Conference on Computational Linguistics (COLING'90)*, Helsinki, 168–173.

Kay, M. (1983). When meta-rules are not meta-rules. In K. Sparck Jones & Y.A. Wilks (Eds), *Automatic Natural Language Parsing* (pp. 94–116). Chichester: Ellis Horwood.

Leacock, C., Towell, G., & Voorhes, E. (1993). Towards Building Contextual Representations of Word Senses Using Statistical Models. In *Proceedings, SIGLEX workshop: Acquisition of Lexical Knowledge from Text*, ACL.

Lewis, D.D., & Croft, W.B. (1990). Term Clustering of Syntactic Phrases. In *Proceedings, 13^{th} Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'90)*, Brussels, 385–404.

Metzler, D., & Haas, S.W. (1989). The Constituent Object Parser: Syntactic Structure Matching for Information Retrieval. In *Proceedings, 12^{th} Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'89)*, 117–126.

Resnik, P.S. (1993). *Selection and Information: A Class-Based Approach to Lexical Relationships.* Ph.D. diss. in Computer Science, University of Pennsylvania.

Sager, J.C. (1990). *A Practical Course in Terminology Processing.* Amsterdam: Johns Benjamins.

Sager, N. (1981). *Natural Language Information Processing.* Reading, MA: Addison-Wesley.

Sheridan, P., & Smeaton, A.F. (1992). The Application of Morpho-syntactic Language Processing to Effective Phrase Matching. *Information Processing & Management*, *28*(3), 349–369.

Shieber, S.N. (1986). *An Introduction to Unification-Based Approaches to Grammar.* CSLI Lecture Notes *4*, Stanford, CA: CSLI.

Smadja, F. (1993). XTRACT: An Overview. *Computer and the Humanities*, *26*, 399–413.

Sparck Jones, K., & Tait, J. I. (1984). Automatic Search Term Variant Generation. *Journal of Documentation*, *40*(1), 50–66.

Strzalkowski, T., & Vauthey, B. (1992). Information Retrieval Using Robust Natural Language Processing. In *Proceedings, 30th Annual Meeting of the Association for Computational Linguistics (ACL'92)*, Newark, DE, 104–111.

Strzalkowski, T. (1994). Robust Text Processing in Automatic Information Retrieval. In *Proceedings, 4^{th} Conference on Applied Natural Language Processing (ANLP'94)*, Stuttgart, 168–173.

Voutilainen, A. (1993). *NPtool*, a detector of English noun phrases. In *Proceedings, Workshop on Very Large Corpora*, Columbus, Ohio, 48–57.