Retrieving Terms and their Variants in a Lexicalized Unification-Based Framework

Christian JACQUEMIN
Institut de Recherche en Informatique de Nantes, IUT
3, rue du Maréchal Joffre
F-44041 NANTES Cedex 01, FRANCE

Jean ROYAUTE
Programme de Recherche Indexation, INIST – CNRS
2, allée du Parc de Brabois
F-54514 VANDOEUVRE-lès-NANCY, FRANCE

Abstract

Term extraction is a major concern for information retrieval. Terms are not fixed forms and their variations prevent them from being identified by a match with their initial string or inflection. We show that a local syntactic approach to this problem can give good results for both the quality of identification and parsing time.

A specific tool, *FASTR*, is developed which handles an identification of basic terms and a parser of their variations as well. Terms are described by logic rules automatically generated from terms and their categorial structure. Variations are represented by metarules. The parser efficiently processes large size corpora with big dictionaries and mixes lexical identification with local syntactic analysis. We evaluate the accuracy of results produced by these metarules and improve these results with filtering metarules.

1 A Natural Language Processing Front-End to Automatic Indexing

Complex terms represent important chunks of information in full-text documents and their identification plays a crucial role in information retrieval [1] :

- Words in multiword terms are less ambiguous than in regular syntactic structures. An *artery* can be either a road or a blood vessel while a *Coronary artery* only corresponds to the second meaning.
- The entries of a thesaurus mainly consist of multiword terms. Their detection is an important clue for assigning pointers to or postings for the documents to the thesaurus entries.
- Technical sublanguages use large lexicons of complex terms which refer to precise concepts in their domain. Their specificity makes them good candidates for the representation of the content of a text.

In this study, we focus on the identification of complex terms through natural language processing (NLP) techniques. Unfortunately, such descriptors accept a wide range of variations which have to be accounted for. These variations have three reasons: temporal evolution, language domain and style of writing. Three main categories of variations can be observed: morphology, syntax and semantics. It is reasonable to assume that semantic variation can be efficiently processed by the inference module [2]. Conversely, morphological [3] and syntactic variations are two main issues for descriptor extraction through NLP. In this study, we focus on the second one. We define a local syntactic variation of a lexical entry [4] as one of its occurrences which cannot be identified through the sole considerations of inflection or hyphenation: for example *hepatic veno-occlusive disease* is a variation of *Hepatic disease* in this sense. The lexical variations are also interesting in other fields of computational linguistics and information retrieval such as lexical acquisition. As depicted in Figure 1, the variations may change an exceptional event (the occurrence of a lexical item) into a frequent one (the occurrence of the same item or one of its variants).

The statistical approach to information retrieval shows good performances when compared with NLP techniques as reported by the TREC–2 Program Committee [5]. However, a statistical tool can take advantage of a linguistic preprocessor aimed at retrieving complex terms from documents [6]. The statistical extraction of complex lexical entries has difficulties in coping with rare but significant occurrences. This is the reason why statistical [7] approaches to lexical acquisition relate occurrences including similar content words. But a precise observation of variation such as the one proposed in this paper is much more accurate than methods only depending on the distance or relying on blind deletion of empty words. For example, *cells from her skin and peripheral blood* can be filtered out as an incorrect variant of *blood cell* through syntactic considerations. In contrast *cells from peripheral blood* is confidently accepted as a correct variant of *blood cell* as will be further explained in 3.1.

Moreover, when identified as a coordination or an insertion, a variation is an opportunity for acquiring a new lexical entry. For example, the coordination *blood and bone marrow cell* which is a variation of *Blood cell* allows for the acquisition of *Bone marrow* and *Bone marrow cell.* Variation also can help with the acquisition of noun phrase interpretation as proposed in [8]. For example, the permutations of *Blood cell* exemplified in Figure 1 mainly introduce two prepositions: *in* and *from.* They denote a semantics of spatial inclusion between both nouns schematized by *cells* INCLUDED-IN *blood* [9] where *cells* is a 'trajector' with respect to the 'landmark' *blood.*

When considering syntactic variation as a specific topic, tools have to be provided to describe and process it efficiently. In this aim, this study presents *FASTR* (a contraction of *FAST* and *PATR* that stands for *FASt Term Recognizer*) a computationally and conceptually tractable front-end to automatic indexing which is composed of a grammar generator and a parser. The generator transforms a list of terms into grammar rules with the help of an on-line dictionary. The rules are used by the parser together with a list of frequent words and a set of metarules to retrieve descriptors and their variants from untagged corpora. The estimations are illustrated through a joint experiment between the natural language laboratory of the *Institut de Recherche en Informatique de Nantes* and the documentation center of *INIST/CNRS.* A list of 80,000 multi-domain terms and two large corpora of scientific abstracts are used for this test: a 100,000-word corpus on metallurgy (METAL) and a 1.5 million-word medical corpus (MEDIC).

The rest of the paper is organized as follows. Part 2 is a presentation of the formalism and the parser. Part 3 is an evaluation of the efficiency of metarules in the extraction of the different kinds of variations.

2 Coordinations	**40 Permutations**	**40 Permutations (continued)**
blood and bone marrow cells	cells (TLCs) from peripheral blood	cell frequency in the peripheral blood
blood and cerebrospinal fluid t cell	cells) (lysed whole blood	cells from blood
	cells) detected in a fetal blood	cells from her skin and peripheral blood
	cell activity of peripheral blood	cells from peripheral blood
14 Insertions	cells among circulating blood	cells from the peripheral blood
blood CD4+, CD8+ cells	cell and blood	cell homogenates of peripheral blood
blood b cell	cells available in the cord blood	cells in blood
blood b cells	cells collected from the peripheral blood	cells in bone marrow and peripheral blood
blood borne cells	cells could be detected in peripheral blood	cells in his peripheral blood
blood contained cells	cells could be enhanced in peripheral blood	cells in paired samples of peripheral blood
blood from a cell	cell count in venous blood	cells in peripheral blood
blood hematopoietic progenitor cells	cells detected in the peripheral blood	cells in the blood
	cells differed between tumour and blood	cells in the peripheral blood
blood leukemic cells	cell infiltration in the retina to blood	cell proliferative responses of peripheral blood
blood monocluclear cells	cells into the peripheral blood	cell saver for intraoperative blood
blood mononuclear cell	cell lines were prepared from peripheral blood	cell strains) and from peripheral blood
blood mononuclear cells	cell nuclear antigen and blood	cell suspensions from the peripheral blood
blood monuclear cells	cells observed in peripheral blood	cells was made in the peripheral blood
blood stem cell	cells on the blood	cells was measured in paired peripheral blood
blood t cells	cell populations were investigated in the blood	cells which are primarily derived from blood

Figure 1. The 56 different variations of *Blood cell* observed in the [MEDIC] corpus.

2 *FASTR*: From Terms to Descriptors

Two convincing arguments can be settled about the application of NLP techniques to automatic indexing. First, [10] compares statistically selected 'phrase discriminators' with syntactically selected ones for information retrieval. Fagan notes the limitations of non-syntactic methods to retrieve some good phrase descriptors such as the ones involving conjunctions. For such phrases, he has to incorporate a syntax-based phrase construction which relates semantically close constructions. Instead of such a hybrid approach, [11] use a selective NLP technique together with a thesaurus and show that it performs as well as human indexing. A second argument is that most of the NLP formalisms are adaptable enough to represent any desirable additional information such as semantic or derivational links between complex lexical entries. Both arguments have led us to the choice of a general unification-based formalism for *FASTR* stemmed from *PATR-II* [12]. *PATR-II* is used to represent various kinds of unification-grammars as well as complex lexical data [13].

2.1 Automatic Tagging of Terminological Data

The *FASTR* application needs a tagged indexing lexicon. We have applied this tagging phase to the *PASCAL* lexicon of *INIST*, but it can be extended to other terminological or indexing lexicons. This automation shortens the linguistic engineering process. We name lexicon tagging the operation which attributes to each word of this lexicon a single syntactic category. The on-line dictionary *DELAF** [14] is used in the process. Tagging is independent of the word context because it is very difficult to detect word ambiguities in the terminological noun phrase. Three principles are retained in the case of multiple ambiguities: (a) if the ambiguity is with an infinitive verb, gerund verb or past-participle verb and any other syntactic category, only the verb category is retained; this choice is justified by our desire to keep the morphological trace with a verb in order to facilitate the integration of more complex morphological properties; (b) if the ambiguity is with the noun and any other categories, it is the noun category that is retained ; (c) residual cases concern ambiguities with adjective/adverb: in this case we give priority to adjectives; (d) the most frequent words like prepositions, conjunctions, pronouns are allocated to the category which is most probable in an indexing lexicon. All the words not recognized by the *DELAF* dictionary receive a Noun category.

2.2 The Formalism: Stemming from *PATR-II*

The formalism of *FASTR* takes from *PATR-II* the decomposition of syntactic rules into a context-free portion and a set of equations. The context-free portion constrains the concatenation of constituents, while the equations constrain the information of the constituents. Rule (3) of Figure 2 describes the term *Concentration effect* which is a noun phrase composed of the concatenation of two nouns. The inflection number describes the affixes which are added to the stem for the different inflections of the lemma described by rules (1) and (2) of Figure 2.

2.3 The Formalism: Additional Features

The description given in 2.2 is a general description of lexical entries by syntactic rules with a flat syntactic structure. In the frequent case where lexical entries are embedded one in another, the formalism of *FASTR* allows to take advantage of a structured representation as outlined for *Lexicalized Tree Adjoining Grammars* [15]. Moreover it can be convenient to gather in a single rule several related terms with common lemmas. This corresponds to the possibility of disjunction in the structure of the rule which is exemplified by rule (4) of Figure 3. This rule represents both lexical entries *Right (pulmonary artery)* and *Left (pulmonary artery)* with a disjunction on the adjective and an embedded term *Pulmonary artery*. The sign → stands for the concatenation of constituents, while the sign = stands for the alternative between several constituents.

```
(1) Word 'concentration':
        <cat> = N
        <inflection> = 1.
(2) Word 'effect':
        <cat> = N
        <inflection> = 1.

(3) Rule N1 → N2 N3:
        <N1 lexicalisation> = 'N3'
        <N1 label> = '025972'
        <N2 lemma> = 'concentration'
        <N2 inflection> = 1
        <N3 lemma> = 'effect'
        <N3 inflection> = 1.
```

Figure 2. Rules representing *Concentration effect.*

```
(4) Rule N1 → (A2 = A3 + A4) (N5 → A6 N7):
        <N1 lexicalisation> = 'N7'
        <N1 label> = '006431'
        <A3 lemma> = 'left'
        <A3 inflection> = 1
        <A4 lemma> = 'right'
        <A4 inflection> = 1
        <A6 lemma> = 'pulmonar'
        <A6 inflection> = 3
        <N7 lemma> = 'arter'
        <N7 inflection> = 3.
```

Figure 3. Rule representing *Left (pulmonary artery)* or *Right (pulmonary artery).*

* The *DELAF* dictionary is a dictionary of inflected forms of words developed by the LADL laboratory (CNRS - University of Paris 7). The *Programme de Recherche Indexation* laboratory of *INIST* use this dictionary for their R&D application.

The main difference between the formalisms of *FASTR* and *PATR–II* is the availability of metarules in *FASTR* to represent term variations. Metarules in *FASTR* are composed of a left-hand side matching with the initial rule and a right-hand side yielding the transformed rule. As for the rules, a set of equalities can constrain the information in both these context-free portions. Metarule (5) of Figure 4 corresponds to the permutation of the first and the second constituent of a 2-constituent term and the insertion of two words (X means *any lexical category*). When applied to rule (3) of Figure 2, this metarule yields a rule describing the strings *effect* W_1 W_2 *concentration* where W_1 and W_2 are any words. It allows for the extraction of *effect of glucose concentration*. In contrast, metarule (6) of Figure 4 corresponds to permutations where the first inserted word is the preposition *of*.

```
(5) Metarule Perm( X1 → X2 X3 ) = X1 → X3 X4 X5 X2:
        <X1 metaLabel> = 'XX'.

(6) Metarule Perm( X1 → X2 X3 ) = X1 → X3 P4 D5 X6 X2:
        <P4 lemma> = 'of'
        <X1 metaLabel> = 'XX'.
```

Figure 4. Two metarules in *FASTR* used t retrieve variations.

2.4 The Parser: Lexicalization and Optimization

The organization of the application is a classic one for an NLP tool, except for the rule generator which is a specific device. First the grammar (rules of terms and lemmas) is automatically generated from the list of terms and then compiled in the application. This automation avoids errors due to human writing and ensures a very quick updating in case of modification. Then the text is parsed in two steps. A morphological step is needed for segmentation and stemming as *FASTR* is working on raw untagged corpora. The second step is the syntactic parse of the texts using the rules activated during the stemming phase. The rules which fail to be parsed are transformed through metarules and tried again.

An eye must be kept on the optimization of the application. The parser must be scalable to the size of industrial lexical resources and textual data. The first and major improvement consists of a bottom-up filtering of the grammar through lexicalization links as suggested in [15] for top-down parsing of lexicalized grammars. The parsing algorithm can take advantage of this lexicalization by only working on the rules corresponding to the lemmas in the input sentence.

Data access also has to be optimized in order to allow for short access time whatever the size of data. Stop words are memory resident and are accessed through a Hash Code table (approx. 100 words). The single words residing in the disk dictionary are accessed through a B-tree (approx. 30,000 words).

The conceptual and computational devices presented in this section are more detailed in [16]. They ensure a good computational tractability of the application as will be shown in the following section.

2.5 The Parser: Bench Marks

The speed of the parser in *FASTR* strongly depends on the size of the grammar of term rules and weakly depends on the size of the metagrammar. Let us study these different factors separately. First, in Figure 5, the parsing speed is depicted as a function of the logarithm of the number of rules. These results confirm that the speed remains acceptable, even for a large grammar: the speed is 18,300 words per minute with a grammar of 8,000 term rules and 2,900 words per minute with a grammar of 80,000 term rules. Secondly, the speed also depends on the number of metarules. The parser only spends a small proportion of its time on the generation of transformed rules and their application. Therefore the incidence of the number of metarules on the parsing speed is less crucial than the number of rules was (Figure 6).

By exemplifying the parsing speed on different tasks in Figure 7, we illustrate the amount of time spent by the application on the three different processes involved in parsing: the core processing (stemming and rule loading), the parsing of terms (basic rules) and the parsing of term variants (transformed rules). The three tasks illustrated by Figure 7 are (1) the general indexing (extraction of terms and their variants), (2) the indexing restricted to basic terms and (3) the extraction of term variants only.

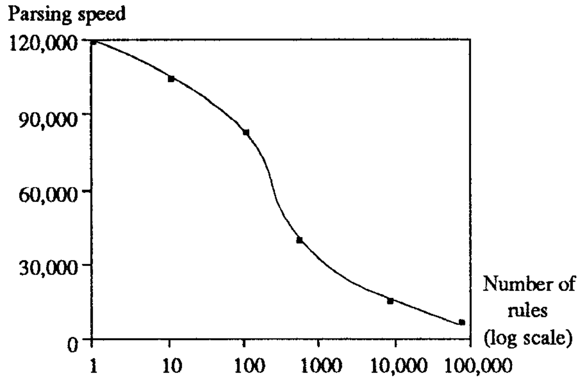

Figure 5. Parsing speed of *FASTR* as a function of the number of rules (with 110 metarules).

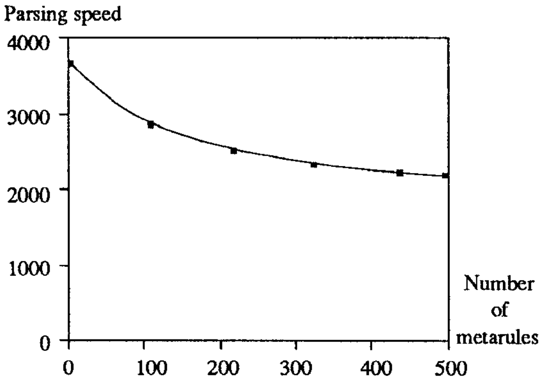

Figure 6. Parsing speed of *FASTR* as a function of the number of metarules (with 80,000 term rules).

Size of the terminological lexicon	80,000 terms	8,000 terms	1 term
Task 1: extraction of terms and variants	2,900 (2,600)	18,300	120,000
Task 2: extraction of terms	3,100 (2,800)	19,800	120,000
Task 3: extraction of variants	4,900 (5,000)	20,500	117,000

Figure 7. The parsing speed of *FASTR* on Sun Sparc-2 Workstation as a function of the tasks achieved by the parser.

The very slight difference between the values of the first and second line of Figure 7 indicates that *FASTR* only spends 10% of its time in extracting variations. The quality improvement induced by this extraction considerably makes up for its computational cost. Therefore, it is worth enhancing the number of metarules because it increases precision and only slightly reduces the parsing speed. The values between parentheses indicated in Figure 7 represent the parsing speed when giving up unification (when working only on the context-free portion of the rules). As this gain is very small, we can assume that an addition of syntactic features will have no incidence on the performances. Such an addition can be used for representing the semantic or derivational links.

In short, the parsing speed depends mainly on the size of the lexical data. The values obtained by testing the application on a workstation show that this kind of hardware is well suited for working with *FASTR* on such large corpora as [MEDIC] of *INIST*.

3 Metarules and Term variants Retrieval: An Evaluation

This part presents more precisely the extraction of term variants in two steps. First a set of paradigmatic metarules is given. The results of the induced indexing of the [MEDIC] corpus is evaluated. Then a set of more filtering metarules is proposed to rule out some of the incorrect variations of the first step. A second evaluation is realized.

The three kinds of variations studied are insertion, permutation and coordination. Elision is intentionally left aside although being an important source of variation** because it cannot be handled properly through metarules only [17]. Elision calls for a handling of the reference between sentences as well as the existence of generic links between terms.

The connection between terms and their variants is described in [4]. The authors use a syntactico-semantic parser to detect compound terms in queries and conflate terms with a similar semantic interpretation. The variations of the terms included in a query are then systematically generated in order to match them with a text database. The generation is restricted to inflections and simple permutations which do not modify the interpretation. This definition is probably too restrictive because it would not account for variations such as *cells from peripheral blood* stemming from *Blood cell.* This approach points out how important and common syntactic variation is but calls for an exhaustive semantic description of single words.

In order to remedy the high human cost of conceptual information retrieval, [6] completes a classic syntactic

** A study of the [METAL] corpus reported in [17] shows that 2.6% of the multi-word term occurrences are elliptic.

analysis by a statistical observation in order to extract head/modifier relations from texts. Reduction to a basic syntactic structure and word stemming account for a wide range of variations similar to the ones detected by *FASTR*. However, it requires a sentence parser which may be difficult to maintain and to modify due to the high number of interdependent rules. When using a partial parser such as *FASTR* or [11], precision is weaker and candidate terms must be filtered accurately. [11] matches candidate terms with controlled terms through substring comparisons. It yields a score of similarity and a degree of confidence which are used to classify and accept new terms. Our approach is close to the one of [11] because we look for syntactic variants of controlled terms. The variants observed in [11] are mainly elisions and embedded terms, we focus rather on permutations, insertions and coordinations.

3.1 Paradigmatic Metarules and Filtering Metarules

Metarules presented in this part are said to be paradigmatic for they do not filter the category of the inserted words, except for the requirement of a conjunction inside a coordination. Therefore the corresponding metarules are given by their context-free portion only without any additional constraints on the constituents. Metarules are grouped into classes corresponding to the rules whose roots have the same number of daughter constituents. For example, metarule (5) of Figure 4 applies to terms composed of two constituents such as *Gene expression* or *Low (melting point)*. Although being slightly noisy, these metarules which do not require any specific tuning work very well on the [MEDIC] corpus. The following is an example of metarules corresponding to 3-constituent terms. Each metarule is followed by an example of variation and the basic term :

Metarule **Coor**(X1 → X2 X3 X4) = X1 → X2 C5 X6 X3 X4:.
inflammatory and erosive joint disease [Inflammatory joint disease]

Metarule **Ins**(X1 → X2 X3 X4) = X1 → X2 X5 X3 X4:.
impaired intravenous glucose tolerance [Impaired glucose tolerance]

Metarule **Perm**(X1 → X2 X3 X4) = X1 → X4 X5 X6 X7 X2 X3:.
diseases of the central nervous system [Nervous system diseases]

Paradigmatic metarules do not constrain syntactic features of words (category, gender, number). One of the ways of increasing the precision of metarules consists in constraining the syntactic category of these words. For example, we observe that the permutation variation requires a 'pivot' element such as the preposition *of* (*fraction of cells → Cell fraction*), but also other prepositions (*cells in unperturbed tumors → Tumor cell*) or else a verbal sequence (*cell lines have been isolated → Isolated cell*). But the recognition of terms in variant forms, for reasons of robustness, uses words unknown by the terminological lexicon and it is more natural to define metarules which forbid some categories of words than metarules which constrain the syntactical category of words which can be unknown. We have named them negative metarules because they are activated prior to the other metarules. They aim at causing a spurious analysis by giving it a specific label in order to keep track of it during the tuning stage and to ignore it during result developments.

3.1.1 Negative Metarules of Insertion

Insertion marks the presence of one or several words within a term. It is probably the least constrained variation because it is difficult to prevent the presence of one or more words within a term. We define a set of negative metarules as follows:

Metarule **NIns** (X1 → X2 X3) = X1 → X2 (X6 = Pu7 + Pc8 + P9) X10 X3:
<P9 lemma> = 'of'.

This metarule identifies as spurious analysis of the X2 X3 term any sequence of text with an inserted X6 element that is a punctuation mark (Pu7), a subordinating coordination (Pc8) or the preposition *of* (P9). It identifies the following sequence *concentration ; baseline measurement* as spurious analysis of the *Concentration measurement* term. This leads us to reject the sequences with the preposition *of* because they can only be used with permutation variations: *basis of live weight* cannot be linked to the term *Basis weight*. Figures 8 and 9 show the effect of negative metarules to identify relevant terms and reject bad analyses of terms.

Textual sequences	Terms (2, 3, 4 words)
vitamin d deficiency	Vitamin deficiency
arterial blood pressure	Arterial pressure
left common coronary artery	Left coronary artery
polymerase chain amplification reaction	Polymerase chain reaction
premature rupture of the membranes	Premature rupture of membrane
farnsworth munsell 100 hue test	Farnsworth 100 Hue test
granulocyte macrophage colony stimulating factor	Granulocyte colony stimulating factor

Figure 8. Relevant term variants corresponding to insertions (Metarule Ins).

Textual sequences	Spurious analysis of terms
coronal angle , slice	Coronal slice
comparison **of** the measurements	Comparison measurement
concentration **and** gradient	Concentration gradient
concentration ; baseline measurement	Concentration measurement
hips , superolateral bone	Hip bone

Figure 9. Terms rejected through filtering insertion metarules (Metarule NIns).

3.1.2 Negative Metarules of Permutation

Generally speaking, the permutation variations are carried out around the prepositions (*of, in, with, on, from,* etc.) as *fractions of cells [Cell fraction]*, or verbal sequences as *enzyme is a serine [Serine enzyme]*. We give more examples of these in Figure 10.

The most frequent spurious analyses (see Figure 11) are permutations around conjunctions. *Identification or specific* cannot be linked to the term *Specific identification.* Another factor of spurious analyses is due to the presence of punctuation in sequences capable of leading to permutation.

Textual sequences	Terms
Preposition of	
fusion **of** tumorigenic heLa cells	Cell fusion
formation **of** insoluble proteinaceous deposits	Deposit formation
localization **of** the dural defect	Defect localization
Preposition in	
fluctuations **in** mean arterial blood pressure	Pressure fluctuation
cells **in** unperturbed tumors	Tumor cell
cell **into** a metastatic tumor	Tumor cell
cell lines **into** nude mice permits tumor	Tumor cell
Preposition with	
pressure **with** normoxic blood	Blood pressure
cell DNA **with** the ultimate tumor	Tumor cell
spectrometry **with** selected ion	Ion spectrometry
treatment **with** either sterile water	Water treatment
Verbal sequences	
cell line **have been** isolated	Isolated cell
applicator **using** microwaves	Microwave applicator

Textual sequences	Terms
Preposition from	
fractions **from** AML cells	Cell fraction
cells **from** peripheral blood	Blood cell
cultures **from** six different tissues	Tissue culture
fiber loss **from** animal	Animal fiber
Preposition for	
method **for** three dimensional measurement	Measurement method
hospital **for** sick children	Children hospital
factor **for** small cell	Cell factor
center **for** health	Health center
test **for** our hypothesis	Hypothesis test
Preposition at/on/above	
viscosity **at** varying shear	Shear viscosity
transition **at** nucleotide	Nucleotide transition
contrast **on** clinical MR images	Image contrast
units **above** control	Control unit
volume reduction **on** fetal plasma	Plasma volume
Verbal sequences	
enzyme **is** a serine	Serine enzyme
gene **may be** a negative regulator	Regulator gene

Figure 10. Relevant term variants corresponding to permutations (Metarule Perm).

Textual sequences	Spurious analysis of terms
motility in epithelial **and** carcinoma cell	Cell motility
results , **and** 1 was negative	Negative result
regurgitation , TEE identified all 14 mitral	Mitral regurgitation
factors , ie , environmental	Environmental factor
effect of body position on	On effect

Figure 11. Terms rejected through filtering permutation metarules (Metarule **NPerm**).

For this type of variation, punctuation makes it difficult to identify the term because it establishes borderlines within the sentence. Punctuation is often linked to a coordination. Thus, the sequence *cell, colonic polyps tumor* cannot be linked from a syntactic point of view to *Tumor cell* even though from a semantic point of view, any occurrence of *tumor* makes the term *Tumor cell* valid. Negative metarules reject sequences with presence of punctuation, of subordinating conjunctions and coordinating conjunctions:

```
Metarule NPerm (X1 →X2 X3) = X1 → X3 (X6 = Pu7 + C8 + Pc9) X5 X2:.
```

this metarule identifies the sequence *age and preoperative mental,* where X6 corresponds to the coordinating conjunction *and,* as a spurious analysis of the term *Mental age*. It also rejects *analysis, our data* as non-linked to the term *Data analysis* since X6 corresponds to a comma.

A few terms, such as *On effect, On line* have a Preposition–Noun structure. Since the occurrences of *effect* and *line* followed by *on* at a distance of a few words are frequent, we create the following negative metarule:

```
Metarule NPerm (X1 →P2 X3) = X1 →X3 X4 X5 P2:.
```

which identifies *effect of calcium on,* where P2 is a preposition as not being linked to the term *On effect.*

3.1.3 Negative Metarules of Coordination

We distinguish two types of coordination (see Figure 12 the table of relevant coordinations): coordinations that concern the head of a noun phrase, and coordinations that concern the modifier part (to the right of the head–noun). The following sequence: *renal hemodynamics and function* coordinates the two head nouns *hemodynamics* and *function.* In this other sequence: *apical and basolateral membrane*, the coordination concerns the two modifiers *apical* and *basolateral.* We show in Figure 12 that coordinating elements are: comma and conjunctions *and* and *or*. The distinction of these two types of coordination makes it possible to reject all of the sequences with a plural noun in a modifier position, since a modifier noun cannot generally take on a plural form. Subsequently, we reject the following sequence: *cells or fetal cultures* as linked to the term *Cell culture* since the plural noun *cells* and the adjective *fetal* cannot be coordinated elements. The following negative metarule:

```
Metarule NCoor (X1 → X2 X3) = X1 →X2 C4 X5 X3:
          <X2 number> = plural.
```

recognizes the following sequence: *cells or fetal cultures* as not being related to the term *Cell cultures* since the plural noun *cells* and the adjective *fetal* cannot be coordinated elements. In the case of a determination, coordinations in noun phrases require that the noun sequence that follows the conjunction be not preceded by a non-possessive determiner. *Tissue or its cell culture* is a correct variant of the term *Tissue culture*. With the following metarule, we reject the sequence *relaxation and the time* as a possible variation of the term *Relaxation time*:

```
Metarule NCoor (X1 →X2 X3) = X1 →X2 C4 Dd5 X5:.
```

We give in Figure 13 the most frequent examples of bad analyses of coordination variations

Textual sequences	Terms (2, 3, 4 words)
Coordination of heads	
cell growth and differentiation	Cell differentiation
cell differentiation and proliferation	Cell proliferation
hemoglobins s and c	Hemoglobin C
interleukins 1 , 2 , and 3	Interleukin 1
Coordination of modifiers	
apical and basolateral membrane	Apical membrane
duchenne or becker muscular dystrophy	Duchenne muscular dystrophy
middle and posterior cerebral arteries	Middle cerebral artery
somatosensory and brainstem auditory evoked potentials	Somatosensory evoked potential

Figure 12. Relevant term variants corresponding to coordinations (`Metarule Coor`).

Textual sequences	Spurious analysis of terms
Coordination with non-possessive determiners	
production **and the** formation rate	Production rate
relaxation **and the** time	Relaxation time
fluids **and the** synovial fluid	Fluid fluid
Coordination with plural nouns in modifier position	
cells and a higher cloning	Cell cloning
cells and purified fractions	Cell fraction
concentrations and colonic epithelial cell	Concentration cell

Figure 13. Terms rejected through filtering coordination metarules (`Metarule NCoor`).

3.2 Indexing a large medical corpus: an evaluation

Indexing experimentation focused on the [MEDIC] corpus which is a bibliographical medical corpus of more than 9 MB of textual abstracts. We have carried out an initial evaluation over the whole corpus using paradigmatic metarules. Out of 17,304 abstracts, we have identified 31,428 pluriterms without variations, and 10720 with variations. Concerning pluriterms, there is a 34% increase in relation to pluriterms without variations. The distribution of the diverse variations is the following: Permutation (57%), Insertion (37,6%) and Coordination (5,4%). The examination of these results allows us to detect the principal causes of spurious analyses of paradigmatic metarules. We have tested the set of negative metarules defined in 3.1 on a more restrained number of abstracts (1,650). In this subset of the corpus, paradigmatic metarules are the least productive of all the corpus. We only obtain an increase of 25%. After filtering, this figure drops to 20.4%. The new distribution of the variation is as follows: Permutation (47.9%), Insertion (43.1%) and Coordination (9%). Examination of results after filtering shows that all the terms rejected with negative metarules are linguistically justified. However, they do not allow us to reject all the spurious analyses. In the long permuted sequences, results are random because there is no control on the words exterior to the terms. For insertion and coordination variations, negative metarules are very efficient and reject all the spurious analysis.

Finally, with a view to understanding why certain terms are not recognized, we have tried to qualitatively evaluate automatic indexing compared with manual indexing. In this analysis, we distinguish the terms which cannot have an occurrence in a text whose retrieval can result from an inference process in an indexing system [18] (for example generic terms formed with the head noun *disease* such as *Urinary system disease*, *Abdominal disease*, etc.). Out of 100 abstracts taken randomly, we observe 292 terms that are common to both manual and automatic indexing, which represents about 2.9 shared terms in a bibliographic reference. A large number of simple words in automatic indexing are irrelevant. This is due to two types of cases: simple words are more polysemous than pluriterms and many terms are general adjectives (such as *acute*, *low*, *high*, etc.) which modify another term in the indexing record of bibliographical references (such as *acute* which can qualify the term *infarct*). This polysemy is important because the *PASCAL* lexicon is a multidisciplinary indexing lexicon. We identify four types of 'non-analysis' of pluriterms. First, morphological variations: the term *myocardium* is not recognized in presence of the adjective *myocardial.* Secondly, elision variations: in the structure of a term, many head nouns yield little information such as in the *PASCAL* term *Pressure volume ratio* where the noun *ratio* is deleted. Thirdly, we notice that the presence of an acronym in a term (frequent in scientific texts) blocks recognition; for example, in place of the *PASCAL* term *Magnetic Resonance Imaging* we encounter *MR imaging.* The last type of non-analysis is linked to the variation of the head term within a semantic paradigm (the term *Transgenic animal* and the textual sequence *transgenic mouse*). It is possible to include in the definition of the head noun of the rule of this term a list of possible *transgenic animals.*

Conclusion

FASTR is a NLP front-end which links texts to descriptors in an efficient way. The representation of variations by metarules makes this parser more powerful than statistical methods and classical parsers with a pattern-matching algorithm. Analysis speed is a function of the number of terms; this makes it usable in industrial contexts with a large terminological volume. We show how the utilization of paradigmatic metarules makes it possible to obtain an acceptable indexing with variations. We adjoin to these paradigmatic metarules a set of negative metarules which aim at increasing the accuracy of results. We consider filtering results satisfying when all inaccurate analyses are justified linguistically and when linguistic engineering development time is short. The comparison made with human indexing shows which elements must be taken into account for future work.

References

1. Sparck Jones K. Assumptions and Issues in Text-Based Information Retrieval. In: Jacobs PS (ed) Text-Based Intelligent Systems, Lawrence Erlbaum Associates, Hillsdale, 1992.
2. Krovetz R, Croft WB. Word Sense Disambiguation Using Machine-Readable Dictionaries. In: Proceedings, 12th Annual ACM SIGIR Conference on Research and Development in Information Retrieval, 1989, pp 127–136.
3. Krovetz R. Viewing Morphology as an Inference Process. In: Proceedings, 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 1993, pp 191–203.
4. Sparck Jones K, Tait JI. Automatic Search Term Variant Generation. Journal of Documentation, 40(1), 1984, pp 50–66.
5. Harman D, Buckley C, Dumais S et al. Report on TREC-2 (Text REtrieval Conference). SIGIR Forum, 27(3), 1993, pp 14–18.
6. Strzalkowski T, Vauthey B. Information Retrieval Using Robust Natural Language Processing. In: Proceedings, 30th Annual Meeting of the Association for Computational Linguistics, 1992, pp 104–111.
7. Enguehard C, Malvache P, Trigano P. Indexation de textes: l'apprentissage de concepts. In: Proceedings, 14th International Conference on Computational Linguistics, 1992, pp 1197–1202.
8. Jacquemin C. A Coincidence Detection Network for Spatio-temporal Coding: Application to Nominal Composition. In Proceedings, 13th International Joint Conference on Artificial Intelligence, 1993.
9. Langacker RW. Foundations of Cognitive Grammar. Vol I. Theoretical Prerequisites. Stanford University Press, Stanford, 1987.
10. Fagan JL. Automatic Phrase Indexing for Document Retrieval: An Examination of Syntactic and Non-Syntactic Methods. In: Proceedings, 10th Annual ACM SIGIR Conference on Research and Development in Information Retrieval, 1987, pp 91–101.
11. Evans DA, Ginther-Webster K, Hart M, Lefferts RG, Monarch IA. Automatic Indexing Using Selective NLP and First-Order Thesauri. In: Proceedings, RIAO'91, 1991, pp 624–643.
12. Shieber SN. An Introduction to Unification-Based Approaches to Grammar. CSLI Lecture Notes 4, CLSI, Stanford, 1986.
13. Boguraev B, Briscoe T. Large Lexicons for Natural Language Processing: Utilizing the Grammar Coding System of LDOCE. Computational Linguistics 13(3–4), 1987, pp 203–218.
14. Courtois B. Un système de dictionnaires électronique pour les mots simples du français. Langue Française 87, Larousse, Paris, 1990, pp 11–22.
15. Schabes Y, Joshi AK. Parsing with Lexicalized Tree Adjoining Grammar. In: Tomita M (ed) Current Issues in Parsing Technologies, Kluwer Academic Publisher, Dordrecht, 1990.
16. Jacquemin C. FASTR: A Unification Grammar and a Parser for Terminology Extraction from Large Corpora. In: Proceedings of IA'94, EC2, Paris, 1994. *Forthcoming.*
17. Jacquemin C. Representing and Parsing Terms with Acceptability Controlled Grammar. In: Proceedings, Terminology and Knowledge Engineering 93, Indeks Verlag, Cologne, 1993, pp 235–244.
18. Royauté J, Schmitt L, Olivetan E. Les expériences d'indexation a l'INIST. In: Proceedings, 14th International Conference on Computational Linguistics, 1992, pp 1058–1063.